NBA MVP Classifier

Carlyle Davis, Harrison Steeley, Adrian Tam, and William Zhang

Abstract

The NBA, or National Basketball Association, is a professional basketball league in North America consisting of 30 teams competing for the championship title. The Most Valuable Player (MVP) is an annual award presented to the player who had the most significant impact on their team's success during the regular season. They are selected through a vote by a panel of sportswriters and broadcasters. The MVP award holds significance as it recognizes individual excellence and contributions to team performance, often serving as a prestigious award to a player's skill, leadership, and impact on the game. This project uses history NBA data from Basketball Reference from 2018 through 2023 (note 2020 is not included due to COVID-19). We investigate various basketball player statistics, such as minutes per game, assists per game, points per game, and other advanced statistics.

Introduction

Define the Problem:
Today, NBA players face a lot of scrutiny and judgment from both the media and fans and are often pitted against each other, especially when it comes to awards. By looking at individual players' statistics alone and turning a blind eye to the media attention or the reputation of some names in basketball, we can strip the name from the player and classify them based on performance alone. Within the NBA, it is currently hard to predict who would win MVP due to many factors such as injuries, subjectivity, and team success.

Motivation for topic:
Popular sports media analysts constantly debate which players deserve awards, which player should be considered the 'GOAT' (greatest player of all time), and which players should be dropped by their teams. Every year, millions of Americans partake in the world of Fantasy Sports, which is based entirely on determining which players are better than others (or will be in
the future). When people rank players, they often cite broad, unspecific statistics as evidence of their claims

Goals and Objective:
Our goal is to develop an objective and strong model, which will be able to classify a random NBA player and determine if they are or are not an MVP candidate based on their player statistics with no bias. We aim to also determine what player statistics are most influential in selecting an MVP candidate.

Related Work

Our study explores the impact of separating a player's name from their performance to investigate its influence on regular season awards. Similar studies, like Brady Goodman from Samford University titled "The Numbers Don't Lie." He analyzes advanced metrics such as win shares and floor impact, assigning different weights to player statistics in order to determine their correlation with team success and MVP awards. Despite finding strong correlations between performance metrics and team success, bias based on player influence, as exemplified by Joel Embiid winning the MVP over Nikola Jokic in the 2022-2023 season despite statistical disparities, still persists.

Methodology

Data Acquisition
We acquired our data from Basketball-Reference because it provides various individual player statistics, such as minutes per game, assists per game, points per game, and other advanced statistics. We scraped the web data.

Data Preparation
We then cleaned the data by removing the repeating headers and change the data types of certain columns appropriately. Within the NBA, some players might not ever fill a certain statistic. For example they do not shoot free throws, so data field is stored as null, and we correct this by using a value of 0. Within the same season, certain players can also play for different teams and within Basketball-Reference their data for different teams is stored with the same name. So, we developed a method to combine a player's data if they played on multiple teams and all their data to a single row. The last thing we did in our initial EDA was to attach a unique ID to the player, allowing us to separate the name from the statistics.

Model Selection
In order to complete this classification task, we have used the SVM, KNN, and decision tree. We then selected the decision tree algorithm as the best model because it can identify important statistics and "branches," however, the model has a weakness of overfitting. In addition, the decision tree model supports nonlinear and various data types. The training will be split as follows: 80% will be used for training and 20% will be used for testing. We used hyperparameters, such as max_depth, min_samples_split, min_samples_leaf, and criterion. The evaluation of the model will include various techniques such as cross validation, accuracy, precision, recall, and F1-score.

Results and Evaluation

As seen in the classification report below, the model performed very well in classifying instances belonging to the "Not MVP" class, with high precision, recall, and F1-score. However, its performance in predicting the "MVP" class is relatively weaker, with lower precision, recall, and F1-score. This indicates that some "MVP"s may be classified as "nonMVP"s. Perhaps, we may need to provide more examples of "MPVs" to the model.

Despite this, the model has an accuracy of 0.97. This shows the model correctly classifies a significant majority of instances across both classes, showcasing its overall effectiveness. Also, the consistency between the mean train and test scores suggests that overfitting is likely minimal, and the low st dev score in both train and test scores indicate low variance in the model.

	precision	recall	f1-score	support
Not MVP	1.00	0.99	0.99	150
MVP	0.71	1.00	0.83	5
accuracy			0.99	155
macro avg	0.86	0.99	0.91	155
weighted avg	0.99	0.99	0.99	155

The Impacts

Our models aim to accomplish one main task which is to classify a player as either being worthy of the MVP award or not. This model could even help current MVP voters narrow down their pool of potential candidates down to those who are statistically worthy, followed by a personal selection using possibly their own bias in their final decision.

The model can also be capable of checking whether or not players in past seasons should have been given more consideration for the award. Often at the end of an NBA season, there is a lot of discourse regarding the player that deserves the award the most, so by using the tool, pure statistics can be used to predict and justify a winner.

Conclusion

By creating multiple machine learning models including KNN, SVM, and a decision tree model, we were able to develop more of an understanding of how the MVP is chosen and which statistics tend to be the most important. We were able to see that the decision tree model was the most accurate and predicted non-MVP winners with 100 percent precision and MVP winners with 79 percent precision. The recall score is close to 100 percent for both classifications as well.

With such strong precision and recall scores, we are confident in the decision tree model's capabilities as a predictor and classifier for NBA MVP winners and can show the multiple use cases for the model. With this model, not only can past MVPs be either validated or invalided in their MVP decision but also serve as a reliable predictor for players who might be deserving of the award.

Overall, we achieved what we set out to do in creating a model capable of classifying a given NBA player into MVP worthy or not. We were able to analyze different types of machine learning models and choose the most accurate one, the decision tree model. In the process of developing these models, we were also able to discover which features and statistics were most impactful in MVP voting. We were also able to uncover which MVP winners might have been more controversial while looking at their competitors.

In terms of future adjustments, something to consider is the competition for the award within the NBA. Our model is able to predict those worthy of the award in terms of statistics, if fed a single player. We might want to add functionality to feed the model an entire data set for a season's data. This way, the model might be able to compare players to each other and decide on the winner of the MVP award in more of a relative fashion. For example, if every player in the NBA underperforms significantly in a season, the model might tell us that no player in that year is deserving of the award, while in reality, the player who performs the best, despite everyone's statistical drop-off, would be the winner.

References

1. https://www.ibm.com/topics/decision-trees
2. https://www.datacamp.com/tutorial/decision-tree-classification-python
3. https://www.samford.edu/sports-analytics/fans/2023/The-Numbers-Dont-Lie
4. Data Source: https://www.basketball-reference.com/leagues/NBA_2023_totals.html