University of North Carolina at Charlotte

Predicting the 2024-2025 MVP With Linear and Logistic Regression

Kiefer Jenny

ITCS 3156 Intro to Machine Learning

Professor Aileen Benedict

8 December 2024

Introduction

Every year, the National Basketball Association gives out awards to its standout athletes. By far the most prestigious of these regular season awards is the Most Valuable Player award, given to the player deemed most valuable to their team and that team's success. This is voted on by reporters and sportswriters at the end of each regular season. They vote based on a variety of factors, including but not limited to: individual statistics, team success, and narratives. Since we are roughly ¼ of the way through the NBA regular season, I attempted to predict the MVP.

NBA awards are very important for the athletes involved. Accolades determine the maximum revenue on contracts players are allowed to sign for. Lifechanging amounts of money are dependent on these awards. In the case of the MVP award, this is exceptionally important when it comes to legacy. How a player is remembered is determined by accolades such as this one, and this can affect how they market themselves and make money once their playing career is over.

To predict the NBA's Most Valuable Player, I will utilize the past ten years of MVP data. The MVP race for the past ten seasons will suffice to gain an understanding of how the voters decide on who to vote for. The NBA is a shifting landscape, where teams are constantly changing their philosophy to maximize their chances of winning. Every decade that the sport has been played professionally is different than the one that came before it. I believe that going beyond ten years would not help the model's performance. Creating models trained and tested on the past ten years of MVP data will allow me to predict the next MVP.

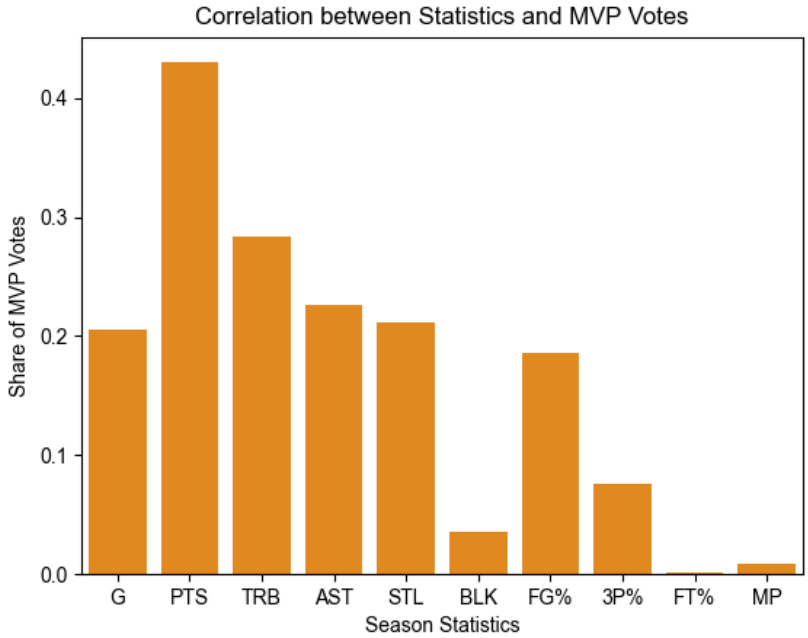
Data

The website Basketball Reference has all the data needed for this task; it contains the current data of the top ten players in the ongoing MVP race. The current Most Valuable Player race table consists of each player's name, their average statistics, games played, and the total of their team's wins. It also features each player's probability to win the MVP award, but I will not be using this since I am attempting to predict this myself. Basketball Reference also has data on previous MVP races. Each regular season has a table listing the statistics of the best players, their rank in that MVP race, and the number of votes they got. For this project, I utilized the current MVP race along with the last ten Most Valuable Player races. I downloaded the current and previous MVP race data and concatenated the last ten years of MVP data together.

Next, I decided on which statistics I will use. Unfortunately, the data in the current chart and the previous charts were not the exact same. The previous tables were lacking in the team wins category, which is very influential in the amount of votes a player will receive. Instead, they had an advanced statistic called win shares, which is an attempt at expressing an individual player's effect on team success. The current MVP race table had wins and losses, but not the win share statistic. I unfortunately decided to not use wins and losses in my model. With that out of the way, I decided to create visualizations to better understand the data. To start, I focused on the amount of games played, the share of MVP votes, the three point percentage, free throw percentage, field goal percentage (all shots that are not free throws), along with the average points, rebounds, assists, steals, blocks, and minutes over a season. I first printed the average statistics of these features for the past ten Most Valuable Players, to gain insights about the players who won MVP.

G	73.8000
PTS	28.6100
TRB	9.9100
AST	7.4500
STL	1.5000
BLK	0.8200
FG%	0.5276
3P%	0.3581
FT%	0.8239
MP	33.7400
Share	0.9326

These results above showcase the different kinds of players capable of winning the award. MVP award winners average a lot of rebounds, assists, and especially points. They also play a lot of minutes compared to most of the athletes in the league, but they do not reach the highest in the league. They are typically quite efficient from the field and on free throws, while also being capable shooters from three. They tend to average a decent number of steals and blocks, but the defensive statistics are not quite as high as the offensive when compared to the rest of the league. Another visualization was created to find the correlation between these stats and the share of the MVP vote.



This barchart was the deciding factor on which statistics I would use for my models. It showcased how important the offensive statistics of rebounds, assists, and mainly points are to the MVP voters. The efficiency with which these players score was also quite important, as field goal percentage is high on this chart. The MVP voters also care about the number of games an

athlete plays in, as the more games a star player is available the more games their team can win. Less important to voters is blocks and three-point percentage, as these statistics can be skewed based on position. Small guards tend to average a small amount of blocks, and tall centers often (but not always) avoid attempting a lot of threes. I decided to drop the three-point percentage, since some centers do not attempt them at all, along with the free throw and minutes played since they had very little to no correlation with the share of MVP votes.

To create these helpful visualizations, I first had to preprocess the data. The previous MVP races often had ties, and this was expressed in Basketball Reference with a "T" next to their rank. This created errors when trying to find the players who had the rank of one, meaning they won the MVP. This was much more complicated to solve than expected, but was accomplished by changing all the ranks whether they had a "T" or not to be a string, removing the "T", and then changing it to an int. This process combined with choosing the correct features mentioned before was the only preprocessing necessary for my assignment. There were no null values that needed to be dropped. The data from Basketball Reference was simple, and not a lot of preprocessing was required.

Methods

I decided to first create a linear model. This model would attempt to predict the share of votes that each MVP candidate would receive. I chose a linear model since it is used to predict continuous values. I assumed a linear model would be effective since I figured that most of the features would have a linear relationship with the share percentage. For example, the share of votes should increase as the amount of points averaged increases. I split the concatenated previous MVP data into a testing and training class given the shares and features I decided on before: games played, points, rebounds, assists, steals, blocks, and field goal percentage. I then

trained the linear regression model with the training data and used this model to predict the share for each player. I tested this model's effectiveness, by calculating the mean squared error and root mean squared error. The mean squared error was very low, but the root mean squared error was unfortunately 0.16. Considering the share values range from .99 to .01 and the winners average .93, this result is not ideal.

The second model chosen was a logistic regression model. This model attempts to solve the original problem of this project, which is predicting the MVP. It would predict whether a player is the MVP or not, and then find the probability of that player to be the MVP. I chose this model since it can predict binary outcomes, and this fits my criteria of whether an athlete is the MVP or not. As I mentioned for the linear model, I thought this would be successful since I assumed the relationships between the data would be linear. For this model, I used a different training and testing class. I used the same ten-year training data except this time, I highlighted the previously mentioned rank feature as the target to be predicted. I also edited the rank column, so that all numbers that were not one were set to zero, making it binary. I used the same predictor features of points, rebounds, etc. and split the data into a testing and training class. This model then predicted the probability of a player to be of rank one, and whether a player will be the MVP. I then tested the model's results. I used a confusion matrix and found the linear log regression. I chose these two because of the way that the model functioned. Since the amount of non-MVP's outweigh the number of MVP's around 10:1, and the fact most of the players in the MVP race have somewhat similar statistics, the model was always going to predict every player as not the MVP. The confusion matrix showcased this, with ten results being true negatives and one result being a false negative. To combat this, I set the model so that if a player reached over 50% probability of winning the MVP, they would be deemed to have a rank of one. This fixed

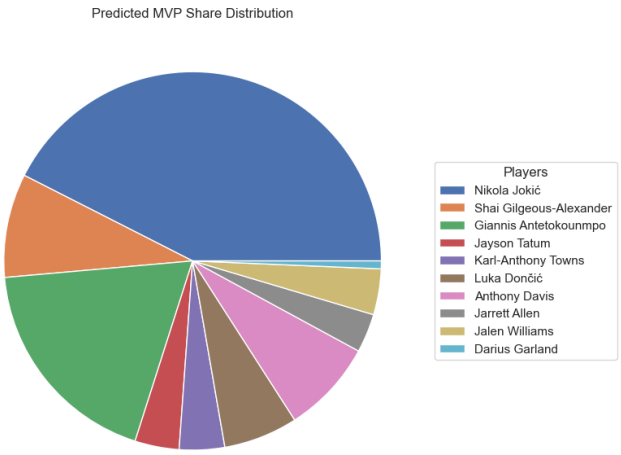
the confusion matrix, and it now showed only true results. The log-likelihood was also now -1.3, which is quite high compared to other models. Looking at these results after the changes I made, the model seems to be quite accurate.

Results

Both the linear model and logistic model had pleasing results when used on the 2024-2025 MVP race. For the linear model, I used the softmax function, that way all the results would add up to one. By using the softmax function, all the shares of the votes would accurately represent shares of an election. The softmax function has been known to create negatively skewed results. To combat this, I multiplied the results by 5 during the softmax process. The results were not properly skewed, so I did it manually. I then formatted these results, and then put them into the table and pie chart below.

Predicted Shares of MVP Votes as a Percentage

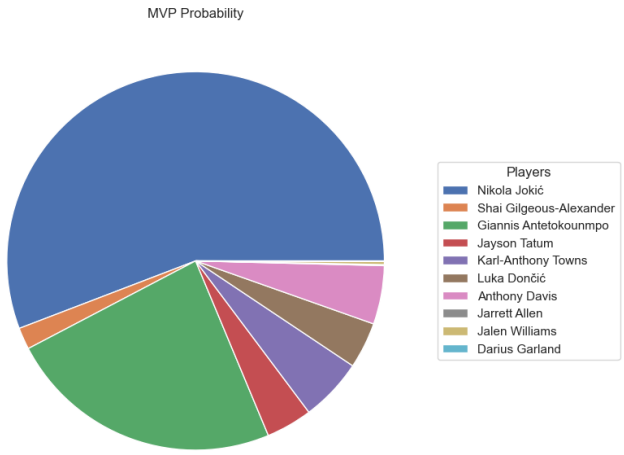
Player	Predicted Share
Nikola Jokić	42.53
Shai Gilgeous-Alexander	8.87
Giannis Antetokounmpo	18.67
Jayson Tatum	3.79
Karl-Anthony Towns	3.88
Luka Dončić	6.36
Anthony Davis	8.0
Jarrett Allen	3.31
Jalen Williams	3.92
Darius Garland	0.67



The logistic model followed a similar process. I applied the model to the 2024-2025 MVP data in order to predict the probability of each player winning MVP. Afterward, I summed the results, and then divided the results by this sum. This would ensure that the logistic model's results would sum to one, just like the linear model. It would not make sense for the probability of each athlete to sum to more than one. Finally, I created a table and pie chart for the results below.

Predicted Shares of MVP Votes as a Percentage

Player	MVP Probability
Nikola Jokić	55.8
Shai Gilgeous-Alexander	1.86
Giannis Antetokounmpo	23.63
Jayson Tatum	3.95
Karl-Anthony Towns	5.4
Luka Dončić	3.95
Anthony Davis	5.0
Jarrett Allen	0.09
Jalen Williams	0.27
Darius Garland	0.04



Both models gave interesting results that we can take various insights from. For both models, they have Nikola Jokic as the overwhelming favorite, with Giannis Antetokounmpo behind. This makes sense when checking predictions from other sources. Basketball Reference itself gives Jokic a 57% chance, although it gives Giannis a low percentage. When viewing the Vegas Sportsbook odds, they also have Jokic as by far their favorite. Still though, Giannis is only fourth in their rankings (Rogers). This is likely due to team success. My models did not have the team wins available as a statistic, and this would change the outcome. Both Shai Gilgeous-Alexander and Jayson Tatum are on two of the most winning teams in the association, and this would give them drastically more votes. What is also interesting is the difference between my models. The linear regression model predicts that Shai will have a much larger share of the votes than the logistic regression model predicts Shai's MVP chances. This could mean that Shai is likely to get second or third in this MVP race, but never outright win it. Another thing to note is the high portion of votes both models predict Anthony Davis to win, despite him not being thought to have a great chance (Rogers). My guess is that since many of the previous MVP winners were centers, Anthony Davis's high rebound and block statistics (statistics usually

dominated by centers) gave him more of an edge on the other players. On the opposite side of the spectrum of Jokic and Giannis's dominance in my models, are the players who the model gives zero chances too. Darius Garland and Jarret Allen of the Cavaliers both seem to suffer from team wins not being a statistic available, but the consensus agrees with these results regardless (Rogers). Their results being so drastically low adds some more validity to my models. Both models agree that Jokic is by far the favorite to win MVP, with Giannis firmly in second place.

Conclusion

In conclusion, the question I set out to solve was answered. I was able to predict the NBA Most Valuable Player award by finding the probability that each player will win, and the share of the votes each player is expected to receive. My linear regression model and my logistic regression model predicted Jokic as the clear favorite to win MVP. It is still early in the season, but as of right now it is his award to lose.

This project taught me a lot and challenged me. As I mentioned before, my models did not have access to the win statistics of the athletes. This distorted the results and is a possible reason for my models strong bias for Giannis being the clear second place, although I agree with this personally. This really proves how much the NBA MVP voting base cares about team success, despite there only being so much one individual can do. These models served as a fun look into a hypothetical scenario where team success did not matter for the MVP, and where the award was based on purely whose personal statistics are the most impressive. I also learned more about machine learning, data science, and programming through this assignment. I learned a new programming technique for preprocessing data in the way I edited the rank column. This was quite challenging and the solution was quite surprising, although this solution was from an external source (Stack Overflow). I also gained more experience in working with every step in

the machine learning process. This project was challenging, but it was immensely rewarding and enjoyable.

GitHub Link

https://github.com/KieferJenny/MachineLearningFinalProject/tree/main

Acknowledgement

I used the sources in my works cited page for help in this project. The basketball reference website gave me all of the data I needed through csv files. VegasInsider was used to gain an understanding of the current leaders in the NBA MVP race. I consulted the matplotlib website and Stack Overflow for programming help with this assignment.

Works Cited

Basketball Reference. “2024-25 NBA MVP Award Tracker.” *Basketball Reference*, 8 Dec. 2024, www.basketball-reference.com/friv/mvp.html.

Basketball Reference. “2023-24 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2024.html.

Basketball Reference. “2022-23 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2023.html.

Basketball Reference. “2021-22 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2022.html.

Basketball Reference. “2020-21 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2021.html.

Basketball Reference. “2019-20 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2020.html.

Basketball Reference. “2018-2019 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2019.html.

Basketball Reference. “2017-2018 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2018.html.

Basketball Reference. “2016-2017 NBA Awards Voting.” *Basketball Reference*, www.basketball-

reference.com/awards/awards_2017.html.

Basketball Reference. “2015-2016 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2016.html.

Basketball Reference. “2014-2015 NBA Awards Voting.” *Basketball Reference*, www.basketball-reference.com/awards/awards_2015.html.

“Extract Number before a Character in a String Using Python.” *Stack Overflow*, stackoverflow.com/questions/36167442/extract-number-before-a-character-in-a-string-using-python/36167504.

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp.

Rogers, Kevin. “2024-25 NBA MVP Odds.” *VegasInsider*, VegasInsider, 22 Oct. 2024, www.vegasinsider.com/nba/odds/mvp/.