MVP Predictions: Analyzing NBA Player Performance through Linear Regression

By: Howard Bell

Every April, as the NBA season comes to a close, a monumental debate ensues. Analysts like Stephen A. Smith, writers like Zach Lowe, and Twitter users across the globe engage in a heated discussion over which player truly embodies the essence of the game and deserves to be crowned the Most Valuable Player (MVP). This prestigious honor, awarded by a panel of esteemed sportswriters and broadcasters throughout the United States and Canada, holds immense significance in the world of basketball. These voters meticulously rank their picks for the MVP from first to fifth, and the player with the highest point total is bestowed with the award. However, the MVP voting process, while revered, is not without its flaws. It can sometimes be influenced by popular narratives rather than purely on-court performance.

With this project, I seek to explore the statistical factors that determine the NBA MVP and create a linear regression model to predict who should have won each year.

Executive Summary

The data was meticulously collected from basketball-reference.com, including all players who received MVP votes and their season averages. Additional columns for advanced statistics like true shooting percentage (TS%) and Value Over Replacement Player (VORP) were included to enhance the analysis. The dataset was expanded to include new players based on their top regular season averages in various categories, ensuring a robust model. Ultimately, the dataset included features such as year, player name, share of MVP votes, player statistics, and advanced metrics.

In the EDA phase, I explored how different variables related to MVP vote shares. Key findings from the correlation mapping showed that VORP, Win Shares (WS), and WS/48 had the highest positive correlations with MVP vote share, while seed also played a significant role.

I chose a linear regression model to predict MVP vote shares due to its effectiveness in modeling relationships between player statistics and vote shares. The model allowed for the identification of key features influencing MVP voting. Categorical variables like player position and seed were encoded, and the dataset was split into training and testing sets. The Mean Squared Error (MSE) was used to evaluate model accuracy, with an MSE of 0.04 indicating good performance.

The analysis of feature importance revealed that Field Goal Percentage (FG%) was the most significant predictor of MVP voting share, followed by seed and three-point percentage. Defensive contributions, represented by steals, also played an important role. Interestingly, true shooting percentage (TS%) and free throw percentage had negative importance coefficients, suggesting that other metrics might be more influential in predicting MVP vote share.

For seven out of the 20 years in the dataset, the model predicted a different MVP winner than the actual choice. Notable differences included the 2010-2011 and 2022-2023 seasons, which remain debated today. If the model were assigning the shares, LeBron James would have won 7 MVPs and Nikola Jokić 4. The model's preference for players like Jokić and James reflects their superior advanced statistics.

The discrepancies between my model's predictions and the actual MVP winners highlight the complexity and subjectivity of NBA MVP voting. While the model relies on statistical data, narratives, records, and biases also play a crucial role in the final decision. This analysis provides valuable insights into the statistical factors considered in MVP voting alongside these other elements.

Data Collection

The data for my model was collected from basketball-reference.com, a comprehensive source that provides a list of all players who received an MVP vote and their season averages, such as points, steals, blocks, and more. To enhance the depth of the analysis, I incorporated additional columns for advanced statistics like true shooting percentage (TS%) and Value Over Replacement Player (VORP). I also expanded the dataset by including new players to ensure the robustness of the model. These additional players were selected by choosing the top 2 in each of the regular season averages (points, rebounds, assists, steals, blocks) that were not already present in the dataset. Including these players with no votes will hopefully improve my model's ability to accurately account for players in various situations, such as those with very good counting stats but on losing teams. Ultimately, my dataset included the following columns:

- Year: The year that the NBA season took place (e.g., 2004-2005)
- Player: Player Name
- Share: MVP voting points a player received divided by the total possible voting points that year
- Age: The age of the player
- Team: The team the player plays for
- Seed: How highly a team is rated in their conference (1-15)
- TmWins: Wins gained by that team
- Position: Position of the player (e.g., PG)
- MP: Average minutes played per game
- TRB: Average rebounds per game
- AST: Average assists per game
- STL: Average steals per game
- BLK: Average blocks per game
- FG%: Average field goal percentage that season
- 3P%: Three-point percentage that season
- FT%: Free throw percentage that season
- TS%: True shooting percentage, a measure of shooting efficiency that accounts for field goals, three-point shots, and free throws.
- WS: Win shares, a composite basketball stat that attempts to capture an individual player's overall contribution to their team, expressed as a 'share' of the team's total wins over the course of a season
- WS/48: Win shares that account for differences in playing time
- VORP: Value Over Replacement Player, a box-score estimate of the points per 100 team possessions that a player scores over a replacement player translated to the average team over a full NBA season
- MVP_Won: Boolean, 1 if that player won the MVP, 0 otherwise

Two notes to make:

1. Changed the data so players who took no 3-point shots during the season were given a 0% 3-point percentage rather than a null value
2. For players who were traded partway through the season, I counted their wins and seed as the team with which they played the most games.

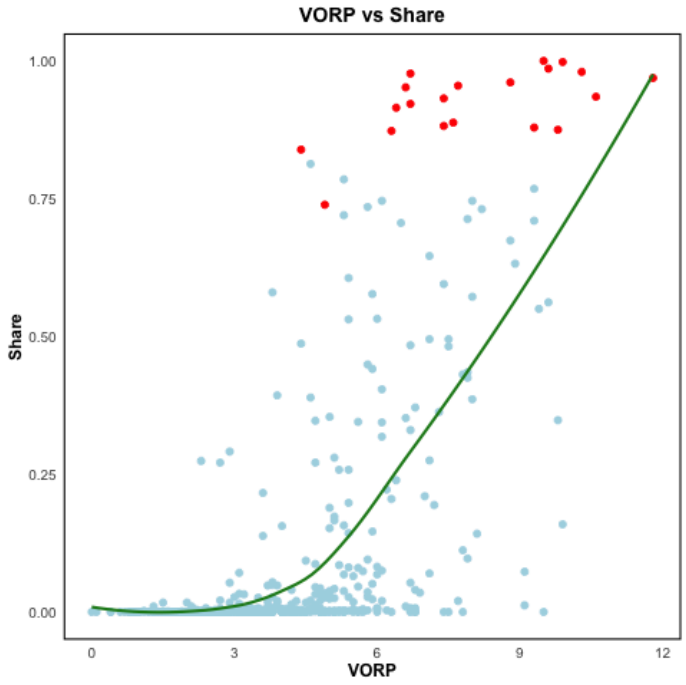
Exploratory Data Analysis

For my EDA, I wanted to explore the various variables present in my dataset and see how these variables related to the share of MVP votes. During my correlation mapping, the variables that had the most impactful correlation with the share were as follows:

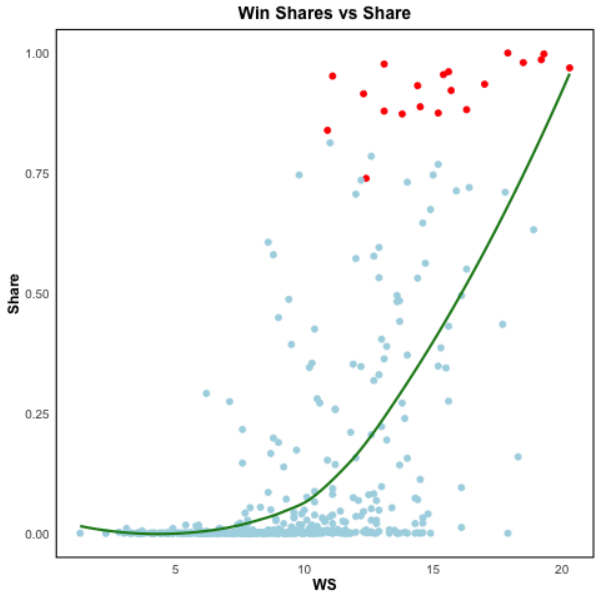
VORP	0.62
WS	0.68
WS/48	0.67
PTS	0.43
Seed	-0.38

The following visualizations are scatter plots with trend lines that analyze various variables' relationship with the Share variable to get a better understanding of these variables within the dataset. In each scatter plot, the variable we are analyzing is on the x-axis, while the Share variable is on the y-axis. The trend line is used to see the general trend of the data, with blue points representing players who did not win MVP and red points highlighting those who did.

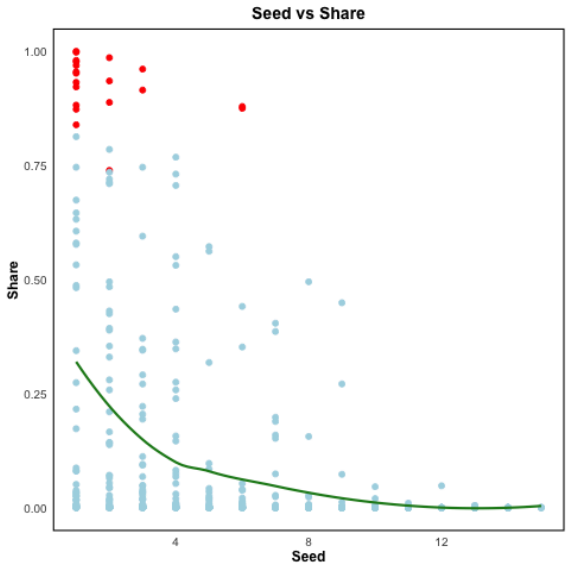
The first variable we are looking at is VORP (Value Over Replacement Player), which is a box score estimate of the points per 100 team possessions that a player contributed above a replacement player, translated to an average team and prorated to an 82-game season. As you can see, there is a positive relationship between VORP and Share. Generally, as VORP rises, Share increases as well. The data backs up this trend: in the 20 years of data present in my dataset, the player with the highest VORP won the MVP 50% of the time and was in the top 3 of MVP voting every single year.



The next variable that I examined was Win Share (WS), which also had a noticeably high correlation with Share. Win Share is defined as an estimate of the number of wins contributed by a player. The calculation for Win Share can be found [here]. Similarly to VORP, the player with the highest Win Share won the MVP 55% of the time and was in the top 3 of MVP voting every single year. The graph below shows the positive relationship between Win Shares and Share: as Win Shares increase, Share increases as well.



The final variable that I wanted to examine is more related to team success. Team success is often a significant indicator of MVP probability because voters are more likely to vote for players whose individual success has also translated to success for their teams. As you can see, the graph below supports that notion: the vast majority of players who won MVPs were within the top 3 seeds in the NBA. In fact, the only player in the last 20 years to win an MVP and not be on a top 3 team in their conference was Russell Westbrook in the 2016-2017 NBA season.



Model Creation

When creating my model, I decided to use linear regression to predict the MVP vote share because modeling the relationship between vote share and the various player statistics I collected seemed best suited to a linear approach. I especially chose a linear model because it is useful for identifying which features are most important for the vote share. Linear regression shows the relationship between the features and the vote share through a linear equation of the form:

$$\text{Vote Share} = \beta_0 + \beta_1 \times \text{Feature}_1 + \beta_2 \times \text{Feature}_2 + \ldots + \beta_n \times \text{Feature}_n$$

This allows us to examine the values of each coefficient to determine how important each feature is. To evaluate this model, we use the Mean Squared Error (MSE), which measures the average squared difference between the actual and predicted values, allowing us to determine the model's accuracy. Generally, a lower MSE indicates better model performance, while a higher MSE may show some errors.

To train the model, I encoded categorical variables like player position and seed and excluded certain variables that I did not want to be included. I then split the dataset into training and testing sets, tested the model on the testing data, and calculated the Mean Squared Error.

Findings

The mean squared error of my model was .04, which indicates that the model is performing well in predicting the MVP vote share. When looking at the features in my model we will be looking at the importance coefficient, the importance coefficient shows the strength and direction of the relationship between the variable and Share, the higher the absolute value of the coefficient is the stronger the relationship is.

Notable Features

1. Field Goal Percentage: With an importance coefficient of 0.352089, Field Goal Percentage was the most significant predictor of MVP voting share in my model. That means that players who score more efficiently are much more likely to receive MVP votes than less efficient players.
2. Seed_1: With an importance coefficient of 0.145251, being the first seed was the second most significant predictor of MVP voting share in my model. Players who played for the best seed in their conference had a much better chance of receiving MVP votes in comparison to those playing on lower seeded teams.
3. Three-Point Percentage: With an importance coefficient of 0.141933, 3 3-point percentage was another significant predictor of MVP voting share. Players who shot more efficiently from 3 were more likely to receive more MVP votes.
4. Seed_2: Similarly to the top seed, being the second seed also increases your probability of receiving MVP votes as shown by its importance of 0.102579.
5. Position (Center): The dummy variable for centers had an importance coefficient of 0.040613, which may suggest that they have an advantage in MVP voting, possibly because of their impact on both the offensive and defensive ends, which generally allows them to rack up more rebounds and blocks than other positions
6. Steals: Steals had an importance coefficient of 0.033774, showing that defensive contributions are also important when considering voting share.

Negative Influences

1. True Shooting Percentage: TS% had a negative importance coefficient of -0.454849. This could suggest that other metrics are more important for predicting MVP voter share, especially because many shooters with incredibly high TS% are role players who take relatively few shots over the course of a game. It could also be a sign that there is an issue with multicollinearity with the other features in my model.
2. Free Throw Percentage: With a negative importance coefficient of -0.179430, free-throw shooting may not be an important predictor of MVP voting share.

Year	Actual MVP	Share	Predicted MVP	Predicted Share
2004-2005	Steve Nash	0.839	Shaquille O'Neal	0.340199
2005-2006	Steve Nash	0.739	LeBron James	0.395146
2006-2007	Dirk Nowitzki	0.882	Dirk Nowitzki	0.389091
2007-2008	Kobe Bryant	0.873	Chris Paul	0.481867
2008-2009	LeBron James	0.969	LeBron James	0.666302
2009-2010	LeBron James	0.98	LeBron James	0.631479
2010-2011	Derrick Rose	0.977	LeBron James	0.429634
2011-2012	LeBron James	0.888	LeBron James	0.469406
2012-2013	LeBron James	0.998	LeBron James	0.645709
2013-2014	Kevin Durant	0.986	Kevin Durant	0.530923
2014-2015	Stephen Curry	0.922	James Harden	0.454651
2015-2016	Stephen Curry	1	Stephen Curry	0.533473
2016-2017	Russell Westbrook	0.879	LeBron James	0.455918
2017-2018	James Harden	0.955	James Harden	0.538782
2018-2019	Giannis Antetokounmpo	0.932	Giannis Antetokounmpo	0.5429
2019-2020	Giannis Antetokounmpo	0.952	Giannis Antetokounmpo	0.553308
2020-2021	Nikola Jokić	0.961	Nikola Jokić	0.51181
2021-2022	Nikola Jokić	0.875	Nikola Jokić	0.499381
2022-2023	Joel Embiid	0.915	Nikola Jokić	0.645533
2023-2024	Nikola Jokić	0.935	Nikola Jokić	0.679599

When using my model to look at each year of MVP voting, you can see some key differences. For seven out of the 20 years in my dataset, my model predicted a different MVP winner than the one who was chosen (these years are highlighted in blue). Some of these are less surprising, like the 2010-2011 and 2022-2023 MVP races, which remain hotly debated MVP races today. If my model were assigning the shares, LeBron James would have 7 MVPs, and Nikola Jokić would have 4!

My model consistently picked players like Nikola Jokić and LeBron James, who routinely have advanced statistics like VORP that are far ahead of their competition. Additionally, because of the importance of seeding, team wins were less significant. This means players who were the first seed in one conference were given the same benefit as the first seed in the other conference, regardless of which team had more wins.

Conclusion

Ultimately, the differences between the predictions of my model and the actual winners of the MVP highlight both the complexity and subjectivity of MVP voting in the NBA. While the model is based entirely on available statistics, there are still narratives, records, and biases that affect who receives MVP votes and who ultimately wins. My analysis shows what statistical factors are important alongside those other considerations when voters are deciding who is the most valuable player in the NBA.