Predicting the NBA’s Most Valuable Player

Michael Montano

Western Governors University

**Research Question:**

The National Basketball Association is widely regarded as the premier basketball organization in the world, home to the most talented players in the sport. Hundreds of the most gifted athletes on the planet compete in the NBA each season and each season only one is granted the honor of the Most Valuable Player award. It is the most prestigious regular season award a player can be granted, though there is no concrete formula for how the winning player is selected:

However, through all over the media, there are still many questions and debates regarding the definition of most valuable player. Is it a player that achieves an outstanding individual statistical success? Is it the best player on the greatest team of the regular season? or is it all about narratives/story lines that numbers cannot possibly define? (Yoo, 2022)

The award winner is selected by 100 members of the media. The members rank their five top performs from the past season, “First-place votes are worth 10 points, second-place votes are worth seven points, third-place votes are worth five points, fourth-place votes are worth three points and fifth-place votes are worth one point” (McGregor, 2022). Having 100 voters and no structure for picking the winner, there is a lot of opportunity for personal bias to enter the equation. This analysis will attempt to provide a numbers-based approach to selecting the MVP. Using 31 seasons of regular season statistics from ever play to touch the court, including all 31 MVPS, a logistic regression model will be made in an attempt to predict which player stats have a statistically significant effect on MVP voting. The study will specifically focus on points, total rebounds, assists, blocks, steals, field goal percentage and team winning percentage.

**Data Collection**

The National Basketball Association makes all of its statistics, including the individual player stats that will be used in this analysis, available on their website NBA.com and free for private, non-commercial use. This dataset, however, was retrieved from Kaggle.com. It was downloaded as a zip file, that contained five different CSV files. This analysis will only make use of the one titled player\_mvp\_stats.csv. This dataset contains 41 columns, including categorical variables such as the player’s name, position and team, as well as continuous variables like the ones mentioned in the previous section. It contains 14,092 rows, not including the column headings, which represent every player that has logged a minute in an NBA game over the last 31 years. One issue with this dataset is that there is no binary variable that represents the MVP award winners. The winners are shown in this dataset by finding the player with the highest share of MVP votes for every season. Since this analysis will require a binomial categorical variable for the logistic regression model to function properly, a new column named MVP will be added.

**Data Extraction and Preparation**

This analysis will attempt to use a logistic regression model using the dataset to predict the which player will win the Most Valuable Player award. Logistic regression is used to predict the likelihood of an event occurring, so it is a good method for us the use for the binomial ‘MVP’ variable. This analysis will be done using Jupyter Notebook running Python. Python is one of the simplest programming languages to learn and has a wide array of libraries and frameworks, such as Pandas, NumPy and SciPy, that are perfectly suited for our analysis (Bahaieva, 2020).

Before the regression model can begin to be built, the data must be prepared. The unused variable must be removed from the dataset and the data must be checked for any missing or erroneous values. The first step of this analysis will be to upload all necessary libraries.

Graphical user interface, text, application

Description automatically generated

Then the data is upload to a Pandas data frame to give the dataset access to many important functions of the Pandas library. The new column, MVP, is added to the data frame with all row values set to “No.”



To change those “No” values to “Yes” in the correct rows to indicate the MVP award winners, the correct rows most be identified for every season included in the dataset. This is done by grouping the data by year and finding the maximum value of the “Share” column for each individual year using Pandas’ .groupby() and .max() methods. The “Share” column represents the total percentage of MVP votes a player received during a given season, thus the player with the highest value in this column is the MVP winner for that year.

Graphical user interface, application, Word

Description automatically generated

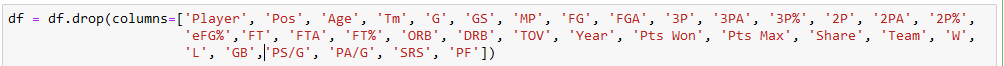
With the year and corresponding maximum share value for every season in the dataset, the proper rows can now be altered to a “Yes” value in the MVP column to acknowledge the winners.

Table

Description automatically generated

Numpy’s .where() method is used to make changes to the MVP column for the proper rows based on the year and share values shown above. The .describe() method is used to get a quick summary of the MVP column after these changes. It shows that there are 14,092 rows and only 2 different values in the column. 14,061 of these rows have a value of “No,” leaving the other 31 rows with a value of yes. This aligns with the number of MVP winners in the 31 seasons represented by this data set.

Now that the MVP column has been added, the columns that will not be used in this analysis will be removed from the dataset



A picture containing application

Description automatically generated

The dataset is now left with only the variables that will be important to analysis, our MVP column and seven numerical variables. Next the dataset is checked for missing values.

Graphical user interface, application, Word

Description automatically generated

The data shows only missing values in the field goal percentage column. This could possibly be attributed to players that may have played very few minutes during a season and didn’t manage to take a field goal attempt. The number of rows with missing values is relatively small, so they can be dropped from the dataset.

Graphical user interface, application

Description automatically generated

Statistics for the continuous variables are shown below. The table below shows the mean, standard deviation, minimum, maximum and various percentiles of each variable. Box plots for each variable have been added for visualizations of these statistics.

Graphical user interface

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart

Description automatically generatedGraphical user interface

Description automatically generated

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Nearly all the variables contain outliers, besides win/loss percentage. However, the MVP award is granted to players that preform far above the norm, so these values will be left in the dataset for the analysis. The distributions for the variables will be visualized through histograms shown below.



Chart, bar chart, histogram

Description automatically generated

As shown through the histograms, both field goal and win/loss percentage have a normal distribution while all the other variables are right skewed. Now the bivariate statistics will be shown to see how each independent variable relates to the dependent variable.

Graphical user interface, application

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

These graphs show that the mean of every variable is noticeably higher for those players that won the MVP award than those who haven’t, though it is interesting to see that both points and field goal percentage have outliers on the lower side.

**Analysis**

Now that that dataset has been cleaned and explored, the first step to building the model is to prepare the dependent variable. The MVP column currently contains yes and no values, but for this regression model, it most contain numeric values. This change is made to the column using the .get\_dummies() method from Pandas. The .describe() method is used to validate that the column values have been changed.

Graphical user interface, text, application

Description automatically generated

With the column values changed, the initial model can be constructed. The Statsmodels’ methods (shown below) will provide data about how each independent variable fits into the regression model.

Graphical user interface, application

Description automatically generated

The chart shows that some of the independent variables have rather high p-values. These variables must be removed to produce a more concise model. To do this, backwards feature elimination will be used, which is a feature selection technique used to build machine learning models that involves removing features that do not have a significant effect on the dependent variable or prediction of output (“What is Backward Elimination”). A significance level of .05 will be used for feature selection. After three iterations of backwards elimination, the model is left with the values shown below.

Graphical user interface, text, application, email

Description automatically generated

With all remaining variables meeting the significance requirements, the data can be broken up into testing and training sets and the model can be used for predictive analysis. The validity of the model will be measured by a confusion matrix. According to Sarang Narkhede of Towards Data Science, a confusion matrix is “a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values” (Narkhede, 2018).

Graphical user interface, application

Description automatically generated

The model produces an accuracy score of 99.76%. While this is accuracy score is extremely high, the confusion matrix shows that it did not correctly predict a single MVP winner.

An ROC curve will also be used to measure the model’s accuracy. An ROC curve “is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0. Put another way, it plots the false alarm rate versus the hit rate” (Brownlee, 2018). The area under the ROC curve, known as the AUC, provides a measurement of the performance across all classification thresholds and ranges between zero and one.

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

The AUC is shown to be .9957 and rounded to 1 in the ROC curve graph. While 1 is the highest possible value that the AUC can be and means that the model is nearly 100% accurate, perfect models don’t often exist, so they’re may be underlying issues with this model.

**Data Summary and Implications**

From the reduced model, the final regression equation is as follows:

Y= -37.98 + .40(TRB) + .55(AST) + .40(PTS) + 27.10(W/L%)

This shows that both defensive stats, steals and blocks, as well as field goal percentage are not statistically significant to this model. Points and rebounds have about the same impact on our model, but both are slightly outweighed by assists. Win loss percentage, however, has a massive effect on a player’s MVP chances according to this model. The model has an accuracy score of over 99%, however with a lack of a single correctly predicted MVP winner, it is hard to say that this model will be useful for predicting correct winners. This model’s issue could be attributed to the very low number of MVP winners in the dataset compared to the total number of players. Paul Allison of the website Statistical Horizons states, “…many researchers worry about whether they can legitimately use conventional logistic regression for data in which events are rare” (Allison, 2012). Small sample sizes of an event could lead to substantial bias in a logistic regression model. A different method of analysis may be better suited for predicting the NBA MVP award winner, such as linear regression or a random forest. There are also other stats that are tracked by the NBA, such as player efficiency rating and usage percentage among others, that were not included in this analysis that may be useful for future models. However, if a player were to use this model to gain an edge in the MVP race, it would be in his best interest to do what ever it takes to lead his team to a good win/loss record while improving his assists per game.

**Works Cited:**

Yoo, D. (2022, May 6). *Predicting the next NBA MVP using Machine Learning*. Medium. Retrieved June 24, 2022, from https://towardsdatascience.com/predicting-the-next-nba-mvp-using-machine-learning-62615bfcff75

McGregor, G. (2022, May 11). *2022 NBA MVP: Full voting results and breakdown for Nikola Jokic over Joel Embiid*. Sporting News. Retrieved June 24, 2022, from https://www.sportingnews.com/us/nba/news/nba-mvp-results-breakdown-nikola-jokic-joel-embiid/lkiclha7fqnjzxkoquel8cf3

Bahaieva, O. (2020, August 8). *Top 7 reasons why you need to learn python as a data scientist*. Medium. Retrieved December 27, 2021, from https://towardsdatascience.com/top-10-reasons-why-you-need-to-learn-python-as-a-data-scientist-e3d26539ec00

*Backward Elimination in Machine Learning - Javatpoint*. www.javatpoint.com. (n.d.). Retrieved December 28, 2021, from https://www.javatpoint.com/backward-elimination-in-machine-learning

Narkhede, S. (2021, June 15). *Understanding confusion matrix*. Medium. Retrieved June 24, 2022, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Brownlee, J. (2021, January 12). *How to use ROC curves and precision-recall curves for classification in Python*. Machine Learning Mastery. Retrieved January 28, 2022, from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Allison, P. (2022, May 11). *Logistic regression for rare events*. Statistical Horizons. Retrieved June 24, 2022, from https://statisticalhorizons.com/logistic-regression-for-rare-events/