Rashmika Batra
rbatra@stevens.edu

Ranking MVP Candidates

Technical Report (Research Internship)
Artificial Intelligence Laboratory
University of Nicosia

2021-08-04

1 Introduction

Inspired by Aris-Konstantinos Terzidis' MSc Thesis paper "Assessment of methods for predicting the NBA regular season MVP using Regression analysis and Classification," this paper will also aim to use machine learning models to predict the NBA MVP winner for the 2018-19 season. This report, however, will have a stronger emphasis on regression. All the models used to predict the MVP will be regression models, and they will all differ from Terzidis' original paper.

2 Background

2.1 The MVP Award

The National Basketball Association (NBA) annually presents the MVP award to the most valuable player of the season. With an average of over one million viewers tuning in to watch the 2020-21 basketball season, basketball has proven to play a huge role in modern culture[1]. The highly coveted award has been part of the NBA since 1956. Until 1980, the MVP was chosen based on a vote by NBA players. Now, the MVP is chosen by a panel of American and Canadian sportswriters and broadcasters.

Basketball analytics are crucial for coaches to assess player performance and decide which players should be drafted into a team. Analytics and predictive models also help teams strategize moves before a competition[2]. Using machine learning to predict basketball outcomes can also be useful in the massive world of sports gambling. In 2020 the sports gambling industry was reported to have a 203 million dollar market share[3], making it a lucrative and popular sector of the sports world. Being able to predict winning teams, valuable players, and useful moves is extremely beneficial when it comes to making game plans or big gambles, which is why machine learning is growing to become a huge part of the sports industry.

2.2 Predicting the MVP using Machine Learning

2.2.1 Classification

Classification is an approach in which predictions are made by sorting data into different classes. The classification approach is called a supervised learning technique because it forces the algorithm to learn through labeled data. The algorithm is then able to spot patterns to categorize unlabeled data. Some examples of classification are spam filters for email, and classifying dogs into different breeds.

The reason that this paper will not use classification as an approach is because it will lead to an imbalanced set, with 94 percent of the data falling into the majority class (players that are not the MVP), and the rest in the minority class (only one MVP per season). For this reason, we will be using a regression model.

2.2.2 Regression

Regression is a machine learning technique used to predict values. Just like classification, regression is also another form of supervised learning. Below is a description of the various methods that we will be using to test the reliability of a given regression model.

The R^2 value is measure of the correlation between the prediction model and dependent variable. The higher

[1] Christina Gough, "NBA Regular Season TV Viewers 2020-2021" (Statista, May 25, 2021), https://www.statista.com/statistics/289993/nba-number-of-tv-viewers-usa/.

[2] "How NBA Analytics Is Changing Basketball: Merrimack College," Merrimack College, June 25, 2020, https://onlinedsa.merrimack.edu/nba-analytics-changing-basketball/.

[3] S Lock, "Key Data on the Global Sports Betting Industry 2020" (Statista, May 31, 2021), https://www.statista.com/statistics/1154681/key-data-global-sports-betting-industry/.

the R^2 value, the better the correlation. For our purposes, a better model will usually have a higher R^2 value. Its calculation is shown below.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

One way to calculate variance explained by the model is by calculating the sum of squares due to regression. To calculate the total variance one can use the total sum of squares[4].

There is also another method used to quantify the effectiveness of a model, and that is Mean Square Error (MSE). It is a measure of the average difference between the actual values and predicted values. The closer the MSE value to zero, the better. Below is the formula.

$$MSE = \sum_{i=1}^{N} \frac{1}{N}(Predicted_i - Actual_i)$$

We will also be using the Root Mean Square Error (RMSE). It is the standard deviation of the difference between the predicted and actual values. A lower RMSE value indicates a better model. Below is the equation for RMSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}$$

3 Data Description

Two data sets obtained form Kaggle[5] will be used to predict the MVP winner for the 2018-19 season. One data set contains statistics dating back to 1980 all the way up to 2018. This set includes data pertaining to the performance of players in each NBA regular season that took place between those years. Metrics used to evaluate player performance include the number of field goal attempts, three point field goal attempts, free throw attempts, player efficiency rating, true shooting percent, usage percentage, box plus/minus, win percentage, award share, number of games, minutes played per game, points per game, total rebounds per game, assists per game, steals per game, blocks per game, field goal percentage, three point field goal percentage, free throw percentage, win shares, and win shares per 48 points (among a few others). The second data set contains similar statistics, but it only contains data for the players from the 2018-19 season. It also does not contain any information about award share because this is what we want our model to predict. Below we will explain what some of these statistics mean and how they are calculated.

The aim of usage percentage is to calculate an estimation of how many team plays were used by a player. It takes into consideration the amount of player field goal attempts (FGA), player free throw attempts (FTA), player turnover percent (TOV), team minutes played (TMP), team field goal attempts (TFGA), team free throw attempts (TFTA), team turnover percent (TTOV), and minutes played (MP). Below is the equation to calculate the usage percentage [6].

$$\text{Usage Percentage} = 100 \times \frac{((FGA + .0.44 \times FTA + TOV) \times \frac{TMP}{5})}{MP \times (TFGA + 0.44 \times TFTA + TTOV}$$

The award share is a key component of this paper, as it is the feature that the model will predict at the end. The award share is the amount of MVP votes a player obtained relative to the total number of MVP votes awarded. Below is the equation used to calculate award share [7].

$$\text{Award Share} = \frac{\text{Award points}}{\text{Maximum number of award points}}$$

To further explain a few more unclear terms, the player efficiency ranking is a measurement created by ESPN columnist John Hollinger that attempts to quantify how well the player preforms on a minute to minute scale. The box plus/minus is an estimation of how much a player contributes to a team [8].

[4] "R-Squared - Definition, Interpretation, and How to Calculate," Corporate Finance Institute (CFI Education Inc.), accessed July 26, 2021, https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/.

[5] Danchyy, "NBA MVP Votings through History" (Kaggle, May 14, 2019), https://www.kaggle.com/danchyy/nba-mvp-votings-through-history.

[6] "Glossary," Basketball Reference, accessed July 26, 2021, https://www.basketball-reference.com/about/glossary.html.

[7] Ibid.

[8] Ibid.

4 Experimental Setup

There are many regressors available to use, but for this project we chose to use three well known regressors which are; Random Forest Regression (FR), Deep Neural Network (DNN), and the K Nearest Neighbors Algorithm (KNN). The Deep Neural Network (DNN) attempts to emulate the way that the human brain learns. For instance, if you were to show a DNN a picture of a dog, it tries to recognize elements of it and then from that deduces what type of animal it may be. After coming up with this answer, the DNN receives feedback if it was correct or incorrect, and learns through this process so that it can eventually recognize a picture of a dog. This is similar to how the brain works because it is learning through feedback and experience. On the other hand, KNN is a supervised learning technique that runs under the assumption that similar things are closer together. It analyzes the distance between points to learn patterns and make predictions. Lastly, RF is another supervised learning technique in which many decision trees are used to make a prediction.

To train the regressors, we will use the data set with statistics dating back from 1980. The regressors will try to fill in the award share column for the other data set with information only pertaining to the 2018-19 season.

5 Results and Discussion

In Terzidis' original work[9], he used regression models such as Regression Ensemble, Random Forest, SVR and Linear Regression. Below is a table showing how well these models preformed in evaluation metrics.

Table 1: Terzidis' Regression Model Evaluation

Regression Model	R^2	MSE	RMSE	Fit time
Regression Ensemble	.614	-.025	.156	6.553
Random Forest	.592	-.027	.160	.638
SVR	.546	-.003	.169	.018
Linear Regression	-0.033	-.034	.181	.001

Below is a table representing how well the models made for the purposes of this paper preformed. Although Terzidis' also used an RF model, this one varies in the way that it was written, which ultimately yielded different results. This is evident though the differences in their R^2, fit time, and RMSE values.

Table 2: Regression Model Evaluation

Regression Model	R^2	MSE	RMSE	Fit time
DNN	.497	-.034	.182	.144
KNN	.474	-.034	.183	.002
RF	.58	-.027	.162	.303

Among the models created for this paper, it seems as though the Random Forest Regression did better than the DNN and KNN models. The strength of this model over the other two is demonstrated by the RF model having the highest R^2 value of .58. Not only does it have a higher R^2 value, it also has the best RMSE and MSE values of .162 and -.027 respectively. Overall however, it seems that the Regression Ensemble model, although slower, has the highest R^2 value, and best RMSE and MSE values. Compared to our RF model, the Regression Ensemble model has a higher R^2 value by .034 and an MSE value .002 units closer to 0 and an RMSE value lower by .006 units. Despite this, our RF model is still significantly faster by 6.25 units.

6 Conclusions

After examining all the data collected, out of the KNN, DNN and RF models the RF model preformed best. Let's test it to see how well it actually predicted the NBA MVP rankings for the 2018-19 season.

On the following page is a table comparing the ranks made by the RF model against the actual MVP rankings for the season.

[9] Aris Terzidis, "Assessment of Methods for Predicting the NBA Regular Season MVP Using Regression Analysis and Classification," Project in Data Science, University of Nicosia, 2021.

Table 3: Predicted Rank vs Actual Rank [10]

Rank	Predicted Player's Rank	Actual Player's Rank
1	Giannis Antetokounmpo	Giannis Antetokounmpo
2	James Harden	James Harden
3	Rudy Gobert	Paul George
4	Nikola Jokic	Nikola Jokic
5	Joel Embiid	Stephen Curry
6	Kawhi Leonard	Damian Lillard
7	Paul George	Joel Embiid
8	Stephen Curry	Kevin Durant
9	Kevin Durant	Kawhi Leonard
10	Damian Lillard	Russell Westbrook
11	Russell Westbrook	Rudy Gobert

The RF model was successful in predicting both the MVP and the runner up correctly, showing that it was in fact successful.

Now, let's assess how well the model predicted the actual award shares. Below is a table displaying the award share values the RF model predicted next to the actual award shares in the 2018-19 season. There is also a percent error column in order to assess how close the ranking model came to the actual values.

Table 3: Predicted Values vs Actual Values [11]

Player Name	Predicted Award Share	Actual Award Share	Percent Error
Giannis Antetokounmpo	.707	.932	24
James Harden	.582	.768	24
Rudy Gobert	.195	.001	19400
Nikola Jokic	.192	.210	8
Joel Embiid	.138	.049	181
Kawhi Leonard	.113	.013	769
Paul George	.110	.352	68
Stephen Curry	.110	.173	36
Kevin Durant	.097	.025	288
Damian Lillard	.073	.068	7
Russell Westbrook	.058	.008	625

The RF model was able to predict the award shares of Jokic and Lillard pretty accurately, with a margin of error of less than 10 percent. The award shares of Antetokounmpo and Harden were also predicted fairly accurately with a percent error of 24 percent.

Even though there are a couple of outliers, the model was in the end successful in predicting the MVP for the 2018-19 basketball season.

Notes

This work was conducted while Rashmika Batra (Stevens Institute of Technology) was on a virtual summer internship with the Artificial Intelligence Lab of the University of Nicosia. Supervisor: Ioannis Katakis.

References

[1] "2018-19 NBA Awards Voting." Basketball Reference. Accessed July 26, 2021. https://www.basketball-reference.com/awards/awards_2019.html.

[2] Chakure, Afroz. "Random Forest and Its Implementation." Medium, June 29, 2019. https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f.

[10] "2018-19 NBA Awards Voting," Basketball Reference, accessed July 26, 2021, https://www.basketball-reference.com/awards/awards_2019.html.

[11] Ibid.

[3] Danchyy. “NBA MVP VOTINGS through History.” Kaggle, May 14, 2019. https://www.kaggle.com/danchyy/nba-mvp-votings-through-history.

[4] Freire, Duarte. “Predicting 2020–21 NBA’s Most Valuable Player Using Machine Learning.” Towards Data Science. Medium, March 25, 2021. https://towardsdatascience.com/predicting-2020-21-nbas-most-valuable-player-using-machine-learning-24aaa869a740.

[5] “Glossary.” Basketball Reference. Accessed July 26, 2021. https://www.basketball-reference.com/about/glossary.html.

[6] Gough, Christina. “NBA Regular Season TV Viewers 2020-2021.” Statista, May 25, 2021. https://www.statista.com/statistics/289993/nba-number-of-tv-viewers-usa/.

[7] Harrison, Onel. “Machine Learning Basics with the k-Nearest Neighbors ALGORITHM.” Towards Data Science. Medium, July 14, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[8] “How NBA Analytics Is Changing Basketball: Merrimack College.” Merrimack College, June 25, 2020. https://onlinedsa.merrimack.edu/nba-analytics-changing-basketball/.

[9] Kurama, Vihar. “Regression in Machine Learning: What It Is and Examples of Different Models.” Built In, August 7, 2020. https://builtin.com/data-science/regression-machine-learning.

[10] Lock, S. “Key Data on the Global Sports Betting Industry 2020.” Statista, May 31, 2021. https://www.statista.com/statistics/1154681/key-data-global-sports-betting-industry/.

[11] “R-Squared - Definition, Interpretation, and How to Calculate.” Corporate Finance Institute. CFI Education Inc. Accessed July 26, 2021. https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/.

[12] Shetty, Badreesh. “An in-Depth Guide to Supervised Machine Learning Classification.” Built In, July 13, 2021. https://builtin.com/data-science/supervised-machine-learning-classification.

[13] Terzidis, Aris-Konstantinos. Tech.

Assessment of Methods for Predicting the NBA Regular Season MVP Using Regression Analysis and Classification, 2021

[14] “What Is a Deep Neural Network?” Oticoin, n.d. https://www.oticon.com/blog/what-is-a-deep-neural-network-dnn.