Assessment of methods for predicting the NBA regular season MVP using Regression analysis and Classification

Aris-Konstantinos Terzidis
MSc in Data Science
Distance Learning
University of Nicosia
aristerzidis@gmail.com

Abstract

The purpose of this project is to investigate different approaches within the Data Science discipline, that can be used to predict the NBA regular season MVP. The approaches that were explored were by using regression analysis and by converting the regression prediction problem into a binary classification problem. The algorithms used in both cases achieved acceptable results. Finding the best parameters for each model to help achieve the best possible predictive results is something beyond the scope of this project and therefore it was not explored in detail. This should be taken under consideration when evaluating the performance of the models at test. The stacking ensemble of regression algorithms achieved the best performance amongst the models tested based on regression analysis. For the classification approach the Support Vector Machine classifier achieved the best results.

1 Introduction

The National Basketball Association (NBA) is the league with the best basketball athletes in the world. Every year the best performing player is awarded the Maurice Podoloff Trophy, known as the Most Valuable Player (MVP) award. The first player in NBA history to receive this honor was Bob Pettit for the 1955-56 season. The players used to vote who would get this honorable award, but all this changed in the 1980-81 season and ever since the award is decided by people of the media such as reporters and sportscasters from the USA and Canada.

Basketball is a sport that involves the constant interaction of 10 players with each other, in the court, in numerous ways. During the last ten years basketball analytics have become an essential part of the NBA. All teams in the league have data scientists/analysts on their payroll[1] to work closely with both the coaching stuff and the players to help improve individual and team performance. Discovering undervalued players, finding the best possible draft pick, injury prevention with the use of sensors during players' sleep. These are all fields that data analytics are applied and have already influenced the way the sport is played in many ways. Making predictions such as the outcome of a game, or a player's performance, based on statistical data and other advanced features is also a big part of the game and the gambling industry.

The purpose of this project is to explore and implement different predictive methods and machine learning algorithms to predict the NBA regular season MVP.

2 Related Work

Y. Chen et. al. [1], used a mix of basic and advanced statistic metrics as input to their Neural Network model for predicting the NBA regular season MVP. M. Chen et. al. [2], also used Neural Networks to predict the MVP but prior to that, they first standardized the data they used, included the "Team Winning" factor and created an MVP Index. Hu et. al [4] in their work, used a BP Neural Network to predict the NBA MVP by considering team record, individual performance and storytelling, which is a factor that cannot be quantified in raw statistical data. The work of Hsu et. al [3], did a comparative analysis to investigate how a team's winning rate is affected by the performance of individual players. In the work of Soliman et. al. [5], the objective was to predict the players that would be selected in the NBA All-Star game. The selection of the starting 5 for each team is subjective, since they are chosen based on the number of votes they get and not based on performance merits. Apostolou et. al. [6] explored various algorithms for performance prediction, based mainly on data related to football. Data from the previous season was used to predict the individual performance of an athlete for the next season. The work of Sarlis et. al [7] focuses in exploring how Machine Learning and Data Mining techniques are used in the NBA and Euroleague games and provide information that can be used in athlete improvement, team composition and for future predictions.

3 Dataset

For the purpose of this project, public available data sets obtained from Kaggle will be used: a training data set containing historical data from the 1980-81 up until the 2017-18 NBA regular season and a testing data set that has data for the 2018-19 season. Both sets contain statistical attributes obtained from the well-known website Basketball Reference that has a lot of statistical data for the NBA [**Figure 1**]. The data sets consist of basic statistical categories, such as

[1] NBA teams that have Analytics Department. https://www.nbastuffer.com/analytics101/nba-teams-that-have-analytics-department/

points/rebounds/assist/steals etc. per game, games played, etc and advanced statistics such as Win Shares, Boxscore +/-, true shooting percentages and more[2]. The winner of the MVP award is decided from the votes each player gets from the people of the press. This is represented in two columns within the data set; the points won and the award share columns. The first one displays the amount of points each player received. The second one is the percentage of the votes that each player received.

$$\textbf{Award share} = \frac{\text{Award Points}}{\text{Max Number of Award Points}} \qquad (1)$$

The maximum number of award points is not the same each year since the number of sports casters and sports persons that are voting differ each season and that is why the award share gives us a better overview of the result through the years.

The data sets used are essentially cleaned up and ready to use. No feature engineering will be applied for the purposes of this specific project. Feature reduction was applied to reduce feature collinearity and to keep the features that, based on domain knowledge and from exploring their correlation to the prediction target, will help having more accurate predictions.

				Voting					Per Game						Shooting			Advanced	
Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%	FT%	WS	WS/48
1	Stephen Curry	26	GSW	100.0	1198.0	1300	0.922	80	32.7	23.8	4.3	7.7	2.0	0.2	.487	.443	.914	15.7	.288
2	James Harden	25	HOU	25.0	936.0	1300	0.720	81	36.8	27.4	5.7	7.0	1.9	0.7	.440	.375	.868	16.4	.265
3	LeBron James	30	CLE	5.0	552.0	1300	0.425	69	36.1	25.3	6.0	7.4	1.6	0.7	.488	.354	.710	10.4	.199
4	Russell Westbrook	26	OKC	0.0	352.0	1300	0.271	67	34.4	28.1	7.3	8.6	2.1	0.2	.426	.299	.835	10.6	.222
5	Anthony Davis	21	NOP	0.0	203.0	1300	0.156	68	36.1	24.4	10.2	2.2	1.5	2.9	.535	.083	.805	14.0	.274
6	Chris Paul	29	LAC	0.0	124.0	1300	0.095	82	34.8	19.1	4.6	10.2	1.9	0.2	.485	.398	.900	16.1	.270
7	LaMarcus Aldridge	29	POR	0.0	6.0	1300	0.005	71	35.4	23.4	10.2	1.7	0.7	1.0	.466	.352	.845	8.6	.165
8T	Marc Gasol	30	MEM	0.0	3.0	1300	0.002	81	33.2	17.4	7.8	3.8	0.9	1.6	.494	.176	.795	10.2	.182
8T	Blake Griffin	25	LAC	0.0	3.0	1300	0.002	67	35.2	21.9	7.6	5.3	0.9	0.5	.502	.400	.728	9.0	.183
10T	Tim Duncan	38	SAS	0.0	1.0	1300	0.001	77	28.9	13.9	9.1	3.0	0.8	2.0	.512	.286	.740	9.6	.207
10T	Kawhi Leonard	23	SAS	0.0	1.0	1300	0.001	64	31.8	16.5	7.2	2.5	2.3	0.8	.479	.349	.802	8.6	.204
10T	Klay Thompson	24	GSW	0.0	1.0	1300	0.001	77	31.9	21.7	3.2	2.9	1.1	0.8	.463	.439	.879	8.8	.172

Figure 1:Feature overview from www.basketballreference.com

4 Predicting the NBA MVP award winner using different techniques

Predicting an outcome using Data Science is based upon finding patterns, trends, correlations and certainties in data, extracting information out of them and finally, finding the appropriate methods and models to aid in the task.

The prediction of the NBA regular season MVP is first of all a Ranking problem. The objective is to predict the number of votes a player will get. The player that gets the most votes is awarded MVP for a given season. For the purposes of this project and taking under consideration the data sets at hand, the target for prediction can either be the award points won, or the award share. The latter has been chosen since it gives a better way of interpreting the result through the years and therefore suits the needs of the project better.
To solve this problem, two different approaches/methods will be explored and implemented. The first is to try and address this as a regression problem. The second approach will be to convert this into a binary classification problem.

4.1 Regression analysis approach

Regression analysis is the method used to determine the relationship between a dependent variable (target) and one or more independent variables. Regression analysis shows how the dependent variable is affected by the changes in the independent variables. In machine learning various algorithms are used to perform such analysis.

In this case, the dependent variable is the **Award Share** of the votes for the regular season MVP of the NBA. In order to perform this analysis, the independent variables from the statistical attributes in the data set need to be defined first. For the purposes of this project, feature selection was applied based on a combination of domain knowledge, the correlation between features and target variable and with univariate selection based on statistical tests with Select K Best method by Sklearn. The statistical test used for regression was the F measure. The final selection included both basic and advanced statistical features. More specifically, the features selected were the Player Efficiency Rate, True Shot Percentage, Win Share/ WS per 48', Box Plus/Minus, Usage percentage, Win percentage, Points/Assists/Rebounds/Steals per game, Field Goal attempts and Free Throw attempts.

4.1.1 Regression models evaluation

For the regression analysis four different models from the Sklearn [3]library have been tested. Linear Regression, Support Vector Machines regressor, Random Forest regressor and a Stacking ensemble of all the aforementioned. To evaluate the performance of the models at test, a 10-fold cross validation procedure was performed because of the small size of the data set and the limited number of instances in it. Using this procedure, the models were trained with a portion of the data on each fold and then tested with new, unseen data on each test split. This helped towards avoiding overfitting the models and selection bias.

A selection of metrics was used to evaluate the performance of all regression models. The **R-square** (R^2) and the **Mean Square Error** (MSE) / **Root Mean Square Error** (RMSE).

R-square has always a value between 0% and 100%. It is a measure of how well a model fits the dependent variables. It is the percentage of variance in the dependent variable, that can be explained by the independent variables. In other words, it shows the strength of the dependance between the prediction model and the dependent variable as well as how well the prediction is explained by the features selected to make the prediction. It is a single number metric and that is what makes it is easily interpreted by an audience. R^2 score is calculated by the following formula.

$$\boldsymbol{R^2} = \frac{Variance\ explained\ by\ the\ model}{Total\ Variance} \qquad (2)$$

[2] Statistical attributes glossary from Basketball Reference website, www.basketball-reference.com/about/glossary.html

[3] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

MSE evaluates the quality of a model. It shows the average squared difference between the values predicted and the actual values, and it is always positive. MSE is calculated by the following formula.

$$MSE = \sum_{i=1}^{N} \frac{1}{N} (Predicted_i - Actual_i) \quad (3)$$

RMSE shows the standard deviation of the difference between the predicted and the actual value that is also known as the residuals. It gives us an estimate of the amount of error the model makes in its predictions. It is a smaller number than MSE that is always positive, thus making it easier for comparison between different models. The mathematical formula used to calculate it is the following.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Predicted_i - Actual_i)^2} \quad (4)$$

R-square represents a percentage of the output variability and that is why we use it to explain how our models perform. MSE/RMSE are used to compare model performance. Values over 50% for R-square and below 0.180 for RMSE are considered good indicators. The overall performance of all models can be seen in **Table 1**.

Regression Models	RMSE	R^2 (%)
Regression Ensemble	0.156	61.4
Random Forest	0.160	59.2
SVR	0.169	54.6
Linear Regression	0.181	48.7

Table 1: Regression Evaluation

4.1.2 Prediction for 2018-19 - Regression models

As a next step the models were used to predict the 2018-19 MVP with the test data set. The winner was decided after a real struggle between James Harden and Giannis Antetokoumpo, just like it actually happened that year. Two models predicted Harden with Giannis coming closely second and the other two predicted that Giannis would win with Harden coming closely second. Based on the results, the models used did a good prediction.

4.2 Classification

Classification in machine learning is a supervised learning technique. It is the procedure of creating predictive models that categorize data points into classes. Typical examples are the classification of email as spam or not spam, if a bank transaction is fraudulent in financing and even cancer prediction in health department.

The task of predicting the NBA MVP can also be approached as a classification problem, that is whether a player is an MVP or not. Before starting with this approach, some adjustments to the data set are needed first in order to convert this into a classification problem. Since the data set did not contain class labels the appropriate field with containing them was created. The winners of the MVP award were defined as the minority class (1) and all the other players were assigned to the majority class (0), thus converting this into a binary classification problem. The winner of the award in terms of a class is the prediction target. There is only one player awarded MVP for each season. Naturally, this has resulted in a heavily imbalanced data set, where 94 % of the items are assigned in the majority (negative) class (**Figure 2**). This translates in 1:16 ratio of Class-1 to Class-0.

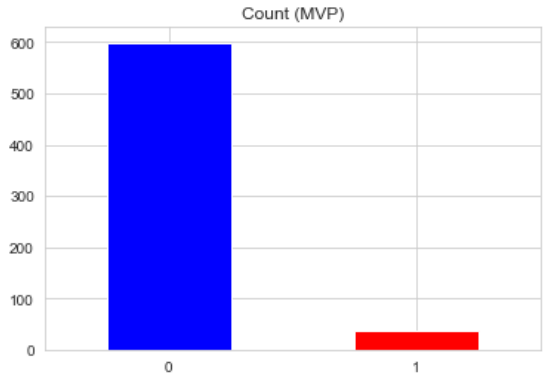

Figure 2: Class Imbalance

4.2.1 Classification models evaluation

For this approach, the probability that each player is classified in the minority class will be predicted. The player that has the greatest probability in being the MVP for a given season, will be classified as the MVP for that season. This classification problem is being built on algorithms from the Sklearn library that can interpret this probabilistic approach, such as Logistic Regression, the Linear Discriminant Analysis, Support Vector Machines classifier, Random Forest classifier, Naïve Bayes and finally, a Stacking classifier including a selection of algorithms. Feature selection combining domain knowledge and Select K Best features method based on ANNOVA test was applied. As a result, the same features that were used in regression analysis were chosen to predict the labels.

The imbalance in classes needs to be addressed when evaluating the performance of the predictive models. To deal with the imbalance different approaches were tested. Oversampling the minority class was not an option because in this problem, there can be only one MVP per season and therefore it is not possible to create more instances of the minority class per season. Resampling the data by undersampling the over-represented majority class (0) was tested. This approach did not give significant improvement and thus it did not change the best performance achieved. Since resampling the data did not give any substantial gain, it was decided to use a more appropriate metric than accuracy, to evaluate the models' performance. What is of greatest importance in this problem is the ability of each model to correctly predict the MVP. Accuracy is a poor selection of a metric for the specific task since the models tend to overpredict the majority class which in its turn results in high accuracy scores. Due to the class imbalance, Precision [4] is the better choice of a metric. The reason Precision is chosen is because what interests the most is the models' ability to correctly predict the minority class,

[4] Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly.

which in this case is the player who will win the MVP award for a given season. A formula for Precision can be seen bellow.

$$\boldsymbol{Precision} = \frac{True\ Possitives}{True\ Positives + False\ Positives} \qquad (5)$$

Apart from the precision of the models, the **Confusion Matrix** (**Figure 3**) is also used to get an overview of the overall performance and to visualize the results. The Confusion Matrix gives a good overview of the True Positives / Negatives and the False Positives / Negatives in the predictions making it easy for interpretation.

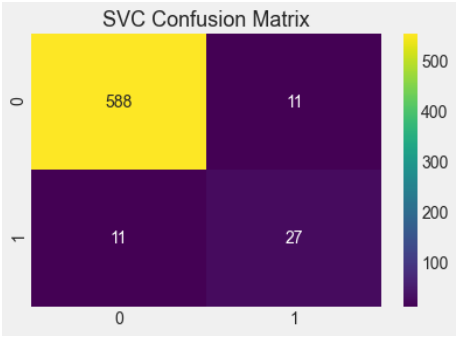

Figure 3: SVC Confusion Matrix

4.2.2 Leave-one-group-out Cross-Validation

In this binary classification problem, there is another challenge to tackle. That is the fact that there can be only one MVP per season. This means that when predicting for the minority class, there is only one instance per season. For that reason, the models at test are being evaluated using a cross validation procedure that is based on the leave-one-group-out method (LOGOCV) that is similar to leave-one-out cross validation (LOOCV). The difference is that in LOGOCV, a whole group of samples is left out instead of a single sample. In this case, the samples are split based on the time period that they represent, allowing for a cross-validation based on time splits.

For the problem of predicting the NBA MVP, all players of one season are kept out for validation and all the players from all the other seasons, are used as training data. This procedure is repeated as many times as the seasons in the data set.

During testing, almost all models achieved Precision close or above 60%, except the Naive Bayes model that was limited a bit below this point. The model that performed significantly better was the Support Vector Machine classifier which outperformed all others, achieving Precision of 71.05% in correctly predicting the MVPs for every season within the training dataset. The performance of all models at test can be seen in **Table 2**

Classification Models	Precision (%)
SVC	71.05
Random Forest	65.79
Classification Ensemble	63.16
Logistic Regression	60.53
LDA	60.53
Naïve Bayes	57.89

Table 2: Classification Evaluation

4.2.3 Prediction for 2018-19 - Classification models

Predicting the winner of the 2018-19 season using the classification models in test was again a 2-man-fight. This time 5/6 models predicted that Harden would win the award and only the Random Forest classifier predicted that Giannis would be the winner. Since we know what the result was, we can see that the prediction was once again within the 2 candidates that actually got the most votes that season, Harden and Giannis.

5 Conclusion

This project was an effort of solving a ranking problem such as the prediction of the NBA regular season MVP, by using two different machine learning techniques. The first one was with Regression analysis and the second one was converting the problem into a binary classification problem. The data set that was used was a publicly available data set from the well-known website Kaggle. This data set contained data scraped from the well-known to NBA fans, Basketball Reference. For the purpose of predicting the NBA MVP, a feature selection procedure based on own domain knowledge, combined with Select K best method from Sklearn, was performed. The same features were used in both approaches. A 10-fold cross-validation was used to evaluate the regression algorithms, based on their R-square and RMSE. Challenges were met by converting this into a binary classification problem, such as an imbalance in classes with a 1:16 ratio. This challenge was tackled by using Precision as a metric for model evaluation and by applying a leave-one-group-out cross validation procedure. This was achieved by using the data for one season as a validation set and the data of all other seasons for training. This procedure was then repeated for each season of the data set. Both approaches had one thing in common; the battle for the 2018-19 MVP was between Harden and Giannis for all models tested, just like it actually happened that season. The Regression Ensemble that was the best performing model for regression analysis, predicted that Giannis would win, whereas the SVM Classifier that was the best performing classifier, predicted Harden as the winner.

6 Future Work

From the 1980-81 season, the MVP is decided from a voting panel of reporters and people of the media. This means that the selection is not based purely on performance merits and excellence in statistical categories, but it also involves a subjective human factor. The players winning the MVP award, although in most cases they truly deserve it, are not always the ones that exceeded in the statistical categories. There can be cases that other factors may affect the criteria and judgement of those deciding. Factors like the overall expectations from a player and his team. A player's public image and how popular he is. All that is what we call "the narrative". These cannot be displayed on pure statistics. In a future attempt, feature engineering should be applied. Features that can quantify the subjective factors mentioned earlier and the narrative, should be added, in an attempt to find more data that can lead to better predictions.

REFERENCES

[1] [Yuefei Chen, Junyan Dai, and Changjiang Zhang. 2019] A Neural Network Model of the NBA Most Valued Player Selection Prediction. In Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence (PRAI '19). Association for Computing

Machinery, New York, NY, USA, 16–20. DOI:https://doi.org/10.1145/3357777.3357786.

[2] [M. Chen and C. Chen. 2020] "Data Mining Computing of Predicting NBA 2019–2020 Regular Season MVP Winner," *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2020, pp. 1-5, doi: 10.1109/ICACCE49060.2020.9155038.

[3] [P. Hsu, S. Galsanbadam, J. Yang and C. Yang, 2018] "Evaluating Machine Learning Varieties for NBA Players' Winning Contribution," *2018 International Conference on System Science and Engineering (ICSSE)*, 2018, pp. 1-6, doi: 10.1109/ICSSE.2018.8520017.

[4] [Jiazheng Hu, Haifei Zhang, and Jianlin Qiu. 2019] Prediction of MVP Attribution in NBA Regular Match Based on BP Neural Network Model. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM 2019). Association for Computing Machinery, New York, NY, USA, Article 43, 1–5. DOI:https://doi.org/10.1145/3358331.3358374

[5] [G. Soliman, A. El-Nabawy, A. Misbah and S. Eldawlatly, 2017] "Predicting all star player in the national basketball association using random forest," 2017 Intelligent Systems Conference (IntelliSys), 2017, pp. 706-713, doi: 10.1109/IntelliSys.2017.8324371.

[6] [K. Apostolou and C. Tjortjis, 2019] "Sports Analytics algorithms for performance prediction," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754.

[7] [Sarlis, Vangelis & Tjortjis, Christos, 2020] Sports analytics – Evaluation of basketball players and team performance. Information Systems. 93. 101562. 10.1016/j.is.2020.101562.