Machine learning and the NBA Game

Article in Journal of Physical Education and Sport · December 2021

DOI: 10.7752/jpes.2021.06453

2 authors:

Bojan Georgievski
Wittenborg University of Applied Sciences

Sabahudin Vrtagic
American University of the Middle East

Journal of Physical Education and Sport ® (JPES), Vol. 21 (issue 6), Art 453, pp. 3339 - 3343, December 2021
online ISSN: 2247 - 806X; p-ISSN: 2247 – 8051; ISSN - L = 2247 - 8051 © JPES

Original Article

Machine learning and the NBA Game

BOJAN GEORGIEVSKI,[1] SABAHUDIN VRTAGIC[2]
[1]Wittenborg University of Applied Sciences, Apeldorn, NETHERLANDS
[2]College of Engineering and Technology, American University of the Middle East, KUWAIT

Published online: December 30, 2021
(Accepted for publication December 15, 2021)
DOI:10.7752/jpes.2021.06453

Abstract:
The purpose of this research was to analyse changes in performance through data simulation. Machine learning has been a popular tool for prediction in the NBA. What we wanted to see how a change in performance is affected in actual points. When a club improves its game, or when a club underperforms what that means as points. Through a points-prediction simulation, we analysed how a change in performance of 1–3% will affect the top five–performing and worst performing clubs.
Key Words: - **Modelling or simulation performance drop coefficients, back propagation, NBA basketball, offensive and defensive data simulation**

Introduction

The growing need for constant improvement in sports has created the need to analyse different ways to achieve such improvement in various sports. The use of machine learning has constantly been promoted over the last several years. Despite the fact that AI/ML has been utilised in a variety of sports research, there is still a need to improve the exhibition of the models offered in this field. In this segment, research has focussed on ML procedures that can be utilised to anticipate the effects of the procedures on the results of basketball games (Cao, 2012).

Loeffelholz et al. (2009) explored ways to forecast National Basketball Association (NBA) game outcomes utilising an artificial neural network (ANN). They also analysed National Football League (NFL) data. Loeffelholz et al. outlined a subset of provisions aggregated by the creators from natural information that was utilised for contribution to the neural nets. Thereafter, different basketball sports specialists were had the opportunity to make expectations to contrast their choices and the results doled out by the ANN model. This ANN model offered a higher percentage of prescient choices than the space specialists; for example, 74.33% was predicted in his findings. The relapse strategy was built to be dependent on a memorable spread, highlighting the fostering of a technique to conjecture about the role of the spread in the NFL. Such exploration could be utilised to anticipate game outcomes. The NFL victors could thus be anticipated by utilising the estimation technique for the point spread, as seen discussed in this paper.

Purucker (1996) used back-engineering, self-assembly guides (SOM), and neural constructions to foresee the results of football match-ups in the NFL. After using different preparation systems, he inferred that back-engendering was the best construction mode for assembling a model that conveys the most precise and prescient results; this is in contrast to some other specialists' findings in the realm of football match-up expectations. Ordinary precision of 61% was accomplished in contrasted to 72% exactness by specialists, as shown in Purucker. Kahn studied the neural organisation structure of back-engendering to anticipate the results of a football match-up (NFL). Purucker broadened his work and accomplished a precision of approximately 75. Moreover, this precision rate was better compared to the majority of the specialists. Differential statistics were used by the two creators based on the crate scores and not on any crude insights.

Kopf has called attention to the fact that each NBA group has information examiners who work with mentors to augment singular players' gifts. At the start of 2009, most associations began utilising a video framework to follow the development of the ball and every player on the court 25 times per second. All of the information gathered via this framework permitted experts to utilise optimal strategies more readily to survey which players were adding to group rewards.

When analysing the use of machine learning in various sports, researchers have focussed on ML usage in terms of its role in predicting various aspects, patterns, and needs of various sports (Wang et al., 2021; Haghighat et al., 2013). Whether these authors have focussed on the possibility of predicting attendance, (King, Rice, & Vaughan, 2018), by using three machine learning algorithms (random forest, M5 prime, and extreme gradient boosting), they concluded that ML performs better than traditional methods of predicting attendance. Moreover, some authors have focussed on the possibilities of team construction and player assessment. Cwiklinski et al. (2021) have used Random Forest, Naive Bayes, and AdaBoost algorithms in order to predict player-transfer success. Additionally ML has been used as a method for constructing an injury forecaster (Rossi,

Corresponding Author: BOJAN GEORGIEVSKI E-mail: bojan.georgievski@wittenborg.eu

et al., 2018) or for analysing a specific aspect of a game (Brooks et al., 2016), perahps in terms of analysing passing characteristics.

Articles on ML in basketball have focused on identifying the drivers of victory (Miljković et al., 2010). By using cart and random-forest learning techniques, Migliorati, (2020) concluded that for the Golden State Warriors, defence is a key factor in winning a game. Another paper (Thabtah et al., 2019) concluded that the DRB (defensive rebounds) feature was deemed to be the most significant factor influencing the results of an NBA game. Furthermore, other crucial factors such as TPP (three-point percentage), FT (free throws made), and TRB (total rebounds) were also selected, which subsequently increased the model's prediction accuracy rate by 2–4%. In addition, Chen et al. (2021) used five data-mining algorithms (ELM, MARS, XGBoost, SGB, and KNN) and identified six important statistics (features) based on four game-lags — averaged defensive rebounds, averaged two-point field-goal percentage, averaged free-throw percentage, averaged offensive rebounds, averaged assists, and averaged three-point field-goal attempts.

Other articles focussing on predicting the outcomes of games (Horvat et al., 2020; Leicht et al., 2017) have analysed the relationship between match-outcome and team-performance indicators. Cheng et al. (2016) used the Maximum Entropy principle to construct the NBAME model and used the model to predict the outcome of the NBA playoffs from the 2007–2008 season to the 2014–2015 season. They concluded that the NBAME model is a good probability model for the prediction of NBA games. The prediction of NBA playoff outcomes is a difficult problem because there are many un-foreseeable factors, such as the relative strengths of the two teams, the presence of injured players, players' attitudes, and team managers' operations, that have a role in determining the winner.

Basketball research is typically performed by analysing the performance aspects of the game needed to win (Ribeiro et al., 2016) (Deshpande & Jensen, 2016). Mikołajec et al. (2013) included 52 variables in their research to analyse in-game performance. Tian et al. (2019) focussed on defensive strategies and defensive decision making. They classified two common defensive strategies (known as switch and trap) used against a popular offensive strategy (known as pick-and-roll), by considering different features from the play-and-ball tracking data. Georgievski and Labadze (2020) used defensive and offensive data in an analysis that revealed comparative advantages in the NBA. Baghal (2012) used four playing factors (effective field goal percentage, free-throw rate, turnovers per possession, and offensive rebounding percentage) to examine offensive and defensive quality.

Similarly, Dehesa et al. (2019) analysed individual performance, concluding that there are five types of individual performance during the regular season. Indivdiual performance and effectiveness led to increased game performance (Courel-Ibáñez et al., 2018). Individual aspects of the game, such as shooting and scoring probability (Zuccolotto et al., 2018) (Erčulj & Štrumbelj , 2015) and performance in the last minutes of the game. Cao et al. (2011) and Lorenzo et al. (2019) concluded that the players who performed better in so-called high-pressure moments had a greater assist and free-throw performance than other players.

Material & methods

The goal of this study was to analyse the potential of machine learning on NBA results. The data set that was used was for the NBA regular season during the period from 2010 to 2019. Every season, the number of games played was 82, except in the 2012 season, when 66 games were played. The predictability of outcomes through machine learning.

We performed points prediction on points that were made. We also noted a change in performance between 1–3% in the top five–performing teams and 1–3 % in the worst-performing teams. The situations that were analysed included both instances of both improved and worsened performances for 1–3 %.

The analysed data included points data and all game statistics, including the following:

- Field-goal data (field goals attempted, field goals made, and percentage of field goals)
- 3-point data (3-point field goals attempted, 3-point field goals made, and percentage of 3-point field goals)
- 2-point data (2-point field goals attempted, 2-point field goals made, and percentage of 2-point field goals)
- 1-point data (2-point field goals attempted, 2-point field goals made, and percentage of 2-point field goals)
- Rebounding data (offensive, defensive, and total rebounding per team)
- Turnovers per team, assists per team, blocks per team, and personal fouls

Table 1 shows the input-output parameters used to train the multilayer perceptron algorithm.

Table 1

Good/ Bad	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	43.4	91.1	0.476	13.5	38.2	0.353	29.9	52.9	0.565	17.9	23.2	0.773	9.3	40.4	49.7	26	7.5	5.9	13.9	19.6	118.1
2	44	89.8	0.491	13.3	34.4	0.385	30.8	55.3	0.557	16.3	20.4	0.801	9.7	36.5	46.2	29.4	7.6	6.4	14.3	21.4	117.7
1	43.7	92.2	0.473	10.3	29.9	0.344	33.4	62.4	0.536	17.8	23.4	0.761	11.1	36.2	47.3	27	7.4	5.4	14.8	21.1	115.4
2	42.5	88.2	0.482	11.1	29.9	0.372	31.3	58.2	0.538	20.8	27.6	0.753	11.8	35.3	47.2	25.9	9	5	15.7	20.7	116.8
			Input																		Output

The multilayer perceptron algorithm (MLP) was used to train the prediction of the PTS for different game scenarios on the basis of a total of 600 game parameters, where 70% of the data was used for training and 30% for testing. The learning process was obtained using a robust scaler for normalisation and a multilayer perceptron for machine-learning training.

RobustScaler pre-processing capacities inside the scikit-learn libraries were utilised with the goal of observing the arrangement and impact of the variance. The capacities referenced above remove the change in the dataset per quintile range by zeroing in and scaling separately on each component via processing the applicable insights from the examples in preparation to set information boundaries. The working principle of MLP is communicated as a cost function optimisation, which is mathematically expressed as follows:

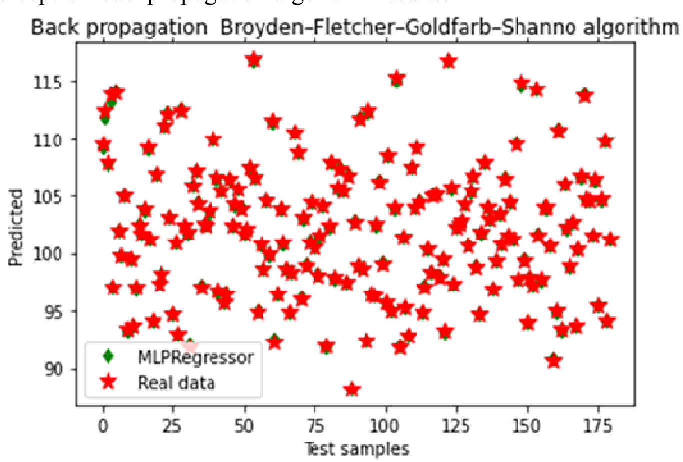
$$E_{True} = \int_{E,y} e(f(x,w),d)p(x,y)dxdy, \qquad (1)$$

where 'e' signifies a local cost function, 'f' is the BFGS work carried out by the MLP, 'x' indicates inputs, 'y' is the expected yield, 'w' indicates the loads in the framework, and p represents the likelihood dispersal. The fair of preparing is to change the boundaries 'w' with the end goal that ETrue is lessened where ETrue, means the speculation blunder. In addition, important techniques used to forestall overfitting in MLP are model decisions, early closures, weight crumbling, and pruning (Weiss & Kulikowski, , 1991); Xie et al., 2016; Zweiri et al., 2005).

Results

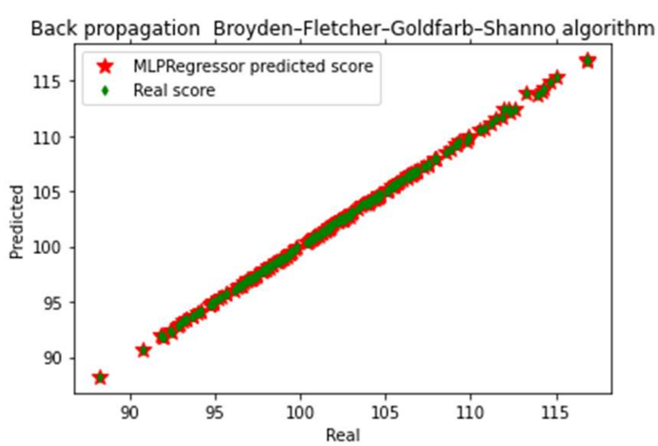
Figure 1, Figure 2, and Table 2 show the results of the MLP, which clearly indicate that the model used is reliable in terms of new input types for different match PTS prediction.

Figure 1. Multilayer perceptron-backpropagation algorithm results.



From Figure 1, it is clear that MLP prediction with the test dataset was very robust, which was confirmed by the R-squared value, as shown in Table 2. Linearised prediction results are illustrated in Figure 2, which provides a clearer image of the ML precision.

Figure 2. Multilayer perceptron-backpropagation algorithm linearised results.



The data provided in Table 2 is a summary and clear indication that the following algorithm is well trained and can be used as a tool for predicting a new game PTS:

Table 2. MLP model scores.

Score or R= 0.9996666032613732
Mean squared error 0.011203689870814933
Root Mean squared error 0.10584748400795804
MAE 0.07219751244472825

The next step was using the elaborated ML model to predict the results in the case of a 1, 2, and 3% drop in the input data parameter set. As shown in Table 3, it was clear that a drop in performance or input parameters will never appear as a drop of the same amount in the result.

Table 3. Model results of the effects on performance drops.

	Error PTS vs Real	Error Real vs Predicted	1-good/2-bad
1%	0.495	1.258	1
1%	0.507	1.249	2
2%	0.504	2.493	1
2%	0.499	2.489	2
3%	0.755	3.732	1
3%	0.753	3.73	2

As observed and shown in Table 3, the effect on a 'bad' or 'good' team was almost the same. In addition, it was clear that a greater drop in input parameters or performance will lead to a greater real effect. It was also clear that with a percentage increase, the 'Error PTS vs. Real' increases; this provides a clear indication of an actual performance drop and the effect on the PTS.

Dicussion

Further analysis of individual performance and its impact on club performance is needed. Furthermore, additional data could be added to the existing data sets aside from in-game performance, focusing on data to improve individual performance. Thus, further research can focus on determining the most beneficial minutes for an athlete to play and at which times the athlete can individually impact the game most positively.

Conclusions

In our work, we used a variety of machine-learning methods to examine the effect of in-game data on performance. The constant need to improve and to find ways to win games have pushed sport clubs to search for pathways to remain competitive.

Change in performance	Actual change for top 5 teams	Actual change for worst 5 teams
1%	1.25%	1.24%
2%	2.49%	2.49%
3%	3.72%	3.73%

In general, a drop in performance of 10% led to an additional drop of 5% in real-world terms. Based on the data, the change was smaller for lower performing teams and slightly larger for better performing teams.

Conflicts of interest - If the authors have any conflicts of interest to declare.

References:

Baghal, T. (2012). Are the four factors indicators of one factor? An application of structural equation modeling methodology to NBA data in prediction of winning percentage. Journal of Quantitative Analysis in Sports, 1–14.

Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. Wiley Online Library, 338–349.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction. [Master's dissertation, Technological University Dublin]. Retrieved from https://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis

Cao, Zheng, Joseph Price, and Daniel F Stone, “Performance under pressure in the NBA,” Journal of Sports Economics, 2011, 12 (3), 231–252.

Chen, W.-J., Jhou, M.-J., & Le, T.-S. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the National Basketball Association. Entropy, 1–14.

Cheng, G., Zhang, Z., Kyebamb, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy, 1–15.

Courel-Ibáñez, J., McRobert, A. P., Toro, E. O., & Vélez, D. (2018). Inside game effectiveness in NBA basketball: Analysis of collective interactions. Kinesiology, 218–227.

Cwiklinski, B., Giełczyk, A., & Choras, M. (2021). Who will score? A machine learning approach to supporting football team building and transfers. Entropy, 1–12.

Dehesa, R., Vaquera, A., Gonçalves, B., & Mateus, N. (2019). Key game indicators in NBA players' performance profiles. Kinesiology, 92–101.

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA player's impact on is team's chances of winning. Journal of Quantitative Analysis in Sports, 51–72.

Erčulj, F., & Štrumbelj, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. Plos One, 1–14.

Georgievski, B., & Labadze, L. (2020). Success factors and revealed comparative advantage in the NBA. Journal of Physical Education and Sport, 3420–3427.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. Advances in Computer Science: An International Journal, 7–12.

Horvat, T., Havaš, L., & Srpak, D. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. Symmetry, 1–20.

Leicht, A. S., Gómez, M. A., & Woods, C. T. (2017). Explaining match outcome during the men's basketball tournament at the Olympic Games. Journal of Sports Science and Medicine, 468–473.

King, B. E., Rice, J., & Vaughan, J. (2018). Using machine learning to predict National Hockey League average home game attendance. The Journal of Prediction Markets, 85s98.

Kopf, D. (2017). Data analytics have made the NBA unrecognizable. Quartz. https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/

Migliorati, M. (2020). Detecting drivers of basketball successful games: An exploratory study with machine learning algorithms. Electronic Journal of Applied Statistical Analysis, 454–473.

Miljković, D., Gajić, L., Kovačević, A., & Konjović , Z. (2010). The use of data mining for basketball matches outcomes prediction (pp. 309–312). 2010 IEEE 8th International Symposium on Intelligent Systems and Informatics, Subotica, Serbia. IEEE.

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. Journal of Human Kinetics, 145–151.

Purucker, M. (1996). Neural network quarterbacking. IEEE Potentials 15(3), 9–15. https://doi.org/10.1109/45.535226

Ribeiro, H. V., Mukherjee, S., & Han T.Zeng, X. (2016). The advantage of playing home in NBA: Microscopic, team-specific and evolving features. Plos One, 1–16.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, M. F., Fernandez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. Plos One, 1–15.

Tian, C., De Silva, V., Caine, M., & Svanson, S. (2019). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. Appled Science, 1–17.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 103–116.

Wang, H., Dong, C., & Fu, Y. (2021). Optimization analysis of sport pattern driven by machine learning and multi-agent. Neural Computing and Applications, 1067–1077.

Weiss, S. M., & Kulikowski, C. (1991). Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning and expert systems. Morgan Kaufmann Publishing.

Xie, J., Lu, Y., Zhu, S.-C., & Wu, Y. (2016). A theory of generative ConvNet (pp. 2635–2644). Proceedings of The 33rd International Conference on Machine Learning. PLMR.

Zuccolotto, P., Manisera, M., & Sandri, M. (2018). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. International Journal of Sports Science & Coaching, 569–589.

Zweiri, Y. H., Seneviratne, L. D., & Althoefer, K. (2005). Stability analysis of a three-term backpropagation algorithm. Neural Netw, 1341–1347.