NBA GAME SCORE PREDICTOR USING MACHINE LEARNING AND NEURAL NETWORKS

Jahin Mahbub
North South University
jahin.mahbub@northsouth.edu

Dr. Sifat Momen
North South University
sifat.momen@northsouth.edu

Abstract

Sports result projection has grown in prominence in recent years, as shown by large financial transactions in sports betting. Basketball is one of the most popular sports in the world, attracting millions of fans and attracting bettors. Particularly the National Basketball Association (NBA) of the United States. This paper makes a proposition for a new sophisticated machine learning framework for predicting NBA game outcomes by seeking to understand the combination of influential features that influence NBA game outcomes. We'd like to see whether machine learning approaches can be used to predict the outcome of an NBA game based on historical evidence, and what the major factors are that influence game scores. Several machine learning techniques that use various learning systems to extract the models were used to achieve the results. This includes Naive Bayes, Decision Tree, Random Forest, Adaptive Boost and other methods. By analyzing the performance and models produced against different sets of basketball related features, we may identify the core characteristics that lead to improved efficiency such as the prediction model's accuracy and performance. Based on the interpretation of the data, The PIR (Performance Index Rating) feature was identified as the most crucial aspect in determining the outcome of an NBA game. Other aspects of performance, such as 3-point shots, attempted 3-point shots, and offensive rebounds, were also identified.

Keywords: Machine Learning, Classification, Prediction, NBA, Sports.

1 Introduction

Sports is one of the most prominent businesses with great financial interest around the world. The money invested in sports industry all over the world is beyond imagination. It is expected that the global worth of sports will reach 440.77 billion USD in 2021. National Basketball Association (NBA) is one of the biggest shareholders of this worth.

NBA is a professional basketball league for men founded in 1946. It is the highest-level basketball league in the world with a history of more than 70 years. It is the most popular league in the world in terms of marketing, professional and attended games. For a huge amount of response and supporters all over the world, the game has multiple business in betting or anticipating results or predicting performances [1]. This process is mostly done by the supporters without any scientific basis. Rather, it entirely depends on the personal preferences and sentiments. Thus, the accuracy in prediction is very bad. But it is possible to increase the accuracy using the advance technology of data mining and machine learning algorithms. It can create a big impact on the betting industry increasing feasibility resulting an economic advancement. Hence, it is a great area for the researchers to proceed with quantitative research for its dynamic nature.

In the previous decade, this betting process followed simple statistics of each games and individual performances. Technical features like previous performances, team rank etc. were used to predict the probability of the final outcome. It follows a lower accuracy as the data is scarce and more ubiquitous [16]. Analytics records and extracts prediction considering useful information of teams and players to enhance their performances. The team itself used to give importance only to the tactics, fitness and influential players. On the other hand, machine learning can provide a lot of minor information that are to be considered and has the ability to increase precision [13]. It can help the coach, sponsors, managers and players to clarify their decision. They can also use it to predict other teams' strength and ability. Machine Learning furnishes the outcome with more crucial information that cannot be handled with simple statistics.

Machine learning has the ability to predict the outcome of a game. But there are a lot of uncertain factors that can create an impact on the result. In this paper, we try to investigate the influential features that can affect the outcome of the sports.

We apply different machine learning algorithm on a dataset of NBA games from 2014 to 2018. The aim of the paper is to explore the influencing attributes. We compare the result of different machine learning algorithm and try to investigate their performances. These models will be applied onto pre-processed data excluding unwanted factors based on their relevance to the game. The outcome predicted by the different machine learning model and the particular features that can be crucial and most significant to affect the outcome can be a great help to the managers, stakeholders, players and supporters.

The paper is outlined as follows: Section 2 provides a review of the recent works in this field. Section 3 evaluates the dataset and its pre-processing. In Section 4, we illustrate training methods of different machine learning model. Section 5 examines the results and presents the analysis. Finally, Section 6 points some limitations and future work followed by conclusion in Section 7.

2 Literature Review

In recent times, there have been a lot of works to ease the process of sports analytics with machine learning model. Here, we will discuss some of the significant works in this field.

Jain et al. [7] proposed a Hybrid Fuzzy-Support Vector Machines (HFSVM) model to predict the features and generate empirical results. It has approached a fuzzy method in SVM technique. Eventually, the paper has shown that the proposed model is better than SVM model generating better results and increased prediction accuracy.

Cao [3] showed that the Simple Logistic Classifier works far better than the Naïve Bayes algorithm, Artificial Neural Networks and SVM. The author trains the models on 5 regular seasons of NBA. This dataset was pre-processed and features like opponent statistics, players statistics and starting line-up were the most influential ones. The output showed that Simple Logistic Classifier yields 69.67 percent while Naïve Bayes only achieve 65.82 percent accuracy.

Loeffelholz et al. [11] investigated the effectiveness of neural networks for the prediction of NBA games result. The authors followed different models of neural network like radial basis, feed-forward, probabilistic and generalized regression. They have also used a fusion model following Bayes belief networks and probabilistic neural network. The dataset was of 620 NBA games outcome. The result showed that the trained model provides 74.33 percent accuracy which is better than the sports experts having a 68.67 percent accuracy on an average.

Bunker and Thabtah [2] proposed a novel framework to predict sports outcome. The authors focused on the application of Artificial Neural Network (ANN) based on data sources, means of evaluated model and functionable methodologies. They further elaborated the challenges on the field for future research.

Leung and Joseph [9] proposed a data mining approach to predict results of sports. The authors took college football results as their dataset. The analysis showed that the historical data analysis is more effective than comparing only the statistics of teams. Following the evaluation, the paper showed that their approach has a greater accuracy with 97.14 percent than the expert opinion which is less than 70 percent.

Haghighat et al. [6] presented machine learning based search method to predict sports result. It records a value to find desired object upon searching the previous outcome. They have reviewed the related literatures and different datasets like NBA and NFL (National Football League). The evaluation also showed that the proposed algorithm is better than k-random walk method.

Miljkovic et al. [12] applied machine learning methods like SVM, decision trees, k-nearest neighbor, Naïve Bayes and multivariate linear regression to predict the outcome of NBA matches. It used a dataset of 141 variables related to the game in the time period of 2009 to 2010. Some of the variables included are foul, three points, home match, away match, free throws, number of losses, number of wins, wins among others etc. The results showed that machine learning can achieve 67 percent accuracy in predicting the outcome.

Lieder [10] proposed a model to predict the significant features that are important for outcome. Different types of model like Artificial Neural Network, Logistic Regression and Linear Regression were applied. The authors used a dataset of NBA seasons from 2014 to train. The comparison showed that Logistics Regression is the best model achieving almost 70 percent accuracy.

Kopf [8] explained a different approach of the teams to predict performance of the players. There used to be a data analyst division in each team who were responsible to follow individual player to point their ability. They tracked each player's movement along with the ball 25 times per second on the court. This data eventually helps the analysts to evaluate intelligently and measure the ability of every individual.

Cheng et al. [4] proposed an entropy-based model to predict the results of NBA playoffs. The authors followed Maximum Entropy principle. It used a dataset of 10,271 records of NBA from season 2007-08 and 2014-15. The examination showed that the proposed model can be 74.4 percent accurate to predict

the outcome of the game.

3 Proposed Methodology

We have approached methodology dividing the whole research dividing into different sections.

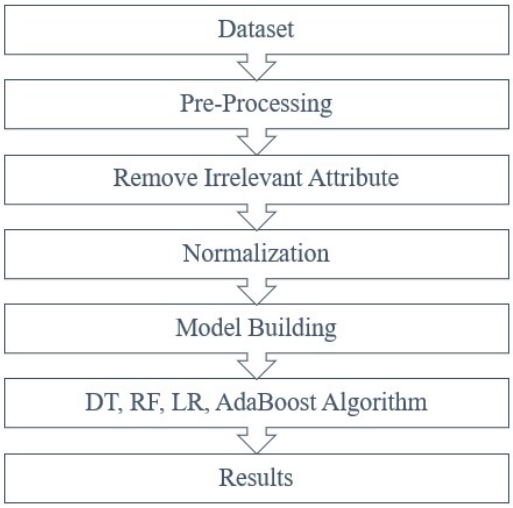

Figure 1: Proposed Methodology

Figure 1 shows our workflow of our proposed model. Here, DT means Decision Tree, RF means Logistic Regression, RF means Random Forest. These are some Classification algorithm that we will use to evaluate.

4 Dataset

The dataset we use is gathered from Kaggle named NBA Game Stats [15]. It has a total 9,840 data with 41 variables. The data holds the result of games from 2014 to 2018. The training dataset holds the crucial variables like matches played in home ground, target variable and percentage of field goals made.

Following the features engineering, we pre-process data and find the performance index rating. First, we convert the categorical data into numerical and generate new columns for rebounds. The Performance Index Rating calculates as Eq. 1:

PIR = (Points + Rebounds + Assists + Steals + Blocks + Fouls Drawn) – (Missed Field Goals + Missed Free Throws + Turnovers + Shots Rejected + Fouls Committed) (1)

Figure 2 describe the features importance following the PIR.

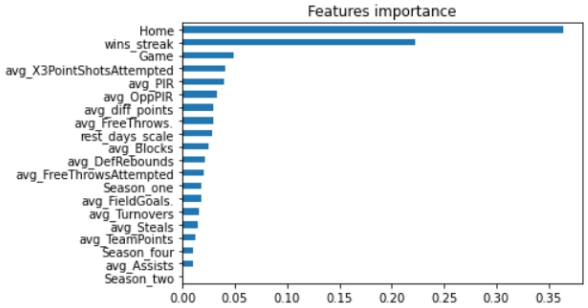

Figure 2: Features Importance

Later, we drop irrelevant columns for our training models and assign 2-class columns to Boolean. Features like team rank, season etc. were removed. It eventually reduces the dimension of input dataset [14]. We also consider the rest days and create a data frame of 9,840 data and 24 variables. A handful of key variables help to improve learning process and increase predictive accuracy of the model.

Figure 3 points the PIR of the most significant factors.

```
Season_two                  0.000000
avg_Assists                 0.010132
Season_four                 0.010153
avg_TeamPoints              0.012427
avg_Steals                  0.014215
avg_Turnovers               0.015521
avg_FieldGoals.             0.017759
Season_one                  0.017992
avg_FreeThrowsAttempted     0.019817
avg_DefRebounds             0.021793
avg_Blocks                  0.024645
rest_days_scale             0.028433
avg_FreeThrows.             0.029157
avg_diff_points             0.029187
avg_OppPIR                  0.033200
avg_PIR                     0.039738
avg_X3PointShotsAttempted   0.040839
Game                        0.048567
wins_streak                 0.222387
Home                        0.364039
dtype: float64
```

Figure 3: PIR of Features Importance

5 Models

We approach different machine learning model to train on the aforementioned data frame. The adaptability of different learning schemes is considered. The models we apply are:

- Decision Tree
- Random Forest
- Adaptive Boost
- Naïve Bayes
- Logistic Regression

5.1 Decision Tree

We have multiple numbers of variables and generating trees can be a way to find the approximate node. So, we used Decision trees based on the features gained. The ID3 algorithm is mostly used in building decision trees. For a particular feature, ID3 algorithm follows a greedy top-down approach to choose nodes. We followed Eq. 2 to measure information gain. The gain is calculated by the difference of parent node's entropy and its child nodes' average entropy. In the equation, IG represents information gain, WA is weight average, and E(P) and E(C) denotes entropy of parent node and entropy of child node respectively.

$$IG = E(P) - WA * E(C) \tag{2}$$

For measuring the entropy, we follow Eq. 3. It is the measure of disorder gathered by selecting a particular feature.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{3}$$

The ID3 algorithm chooses a variable with maximum gain and divides the data following the attributes. It goes on constructing the child nodes choosing different variables until all the features are considered. Finally, it provides a decision. Our model has several distinct attributes based of which the decision trees can build a good optimization. Hence, Decision Tree model is a good choice for our data frame.

5.2 Random Forest

Random Forest is another poplar machine learning algorithm based on ensemble method. It can combine several models to provide an optimized prediction. Several weak learners are combined to generate an ensemble method making altogether a robust model. In our case, we use decision tree as our weak model. The weaker model trains on its own newly created subset of data. Approaching this way, we can choose data subset with no replacement. This is known as Bagging or Bootstrap aggregating. The Random Forest model then calculate the predictions that matches the most trees' output. Following Random Forest, we can reduce the shortcoming of overfitting of decision trees taking the average of the results.

5.3 Adaptive Boost

Adaptive Boost or AdaBoost is also an ensemble method or classifier like Random Forest [5]. Here, we used a boosting. Boosting refers to the process when the model use decision stumps or weak learners to merge and produce strong learners. In our case, we use random forest as our ground model to generate some weak learners. These models provide predictions from sample data. Then it introduces the weight of the weak learners and generates a strong learner. In this way, the system adaptively updates. The weights are updated following the Eq. 4 as:

$$Weight = ln(x/(1-x)) \tag{4}$$

Here, x indicates the accuracy of a particular weak learner. The accuracy depends on the better model gets significant weight and the bad model gets the smaller weight. Hence, we conduct the process sequentially.

5.4 Naïve Bayes

Naïve Bayes is different from Random Forest or AdaBoost classifiers as it does not need to build a model for classification. It is a probabilistic classifier that generates joint probabilities. It is calculated by specific observations that eventually evaluates the class of a test data.

5.5 Logistic Regression

Logistic Regression is another algorithm to investigate sample predictors or different variables. It calculates the WINorLOSS occurrence and causative factors. It iterates over specific time and updates the system to reach decision.

6 Result and Analysis

The experiment has been conducted on Kaggle. We figured different matrices to show performances of each method. First, we split the dataset into training and test subset by the ration of 70:30. The matrices are F1-score, Precision, Recall, and Accuracy. Precision indicates the percentage of relevant cases among the retrieved ones. Eq. 5 shows the precision calculation.

$$Precision = (TP/(TP + FP)) \tag{5}$$

Recall indicates the percentage of relevant data that have been actually retrieved. Eq. 6 shows the Recall calculation.

$$Recall = (TP/(TP + FN)) \tag{6}$$

Here, TP and FP mean True Positive and False Positive respectively, and FN means False Negative.
Figure 4 represents the Precision-Recall graph for Decision Tree.

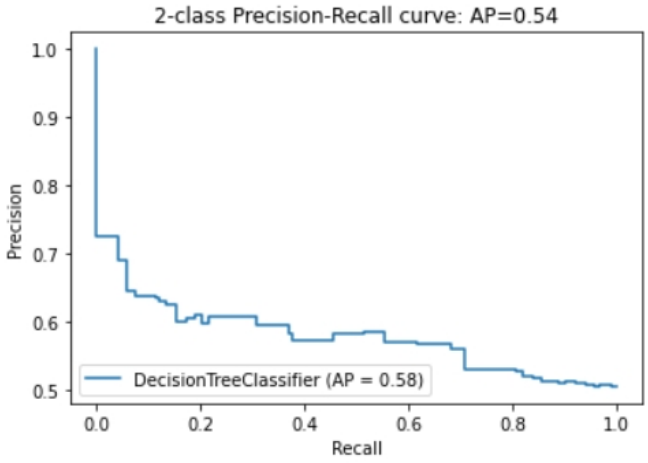

Figure 4: 2-class Precision-Recall Curve

Accuracy is the measure of how effective the model is at predicting outcomes. It is calculated as Eq. 7

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \quad (7)$$

Here, TN refers to True negative.

F1-score is the weighted average of Precision and Recall. Eq. 8 shows the calculation

$$F1Score = 2*(Recall*Precision)/(Recall+Precision) \quad (8)$$

The Accuracy we have found for Decision Tree, Naïve Bayes, Bagging, AdaBoost and Logistic Regression are 56.5 percent, 55.6 percent, 57.2 percent, 57.3 percent and 58.8 percent respectively. After optimization, we have found that Decision Trees show better result all others achieving 69.4 percent accuracy.

Another goal of this paper is to identify influential features sets, Figure 5 shows the five most predictive features after normalization.

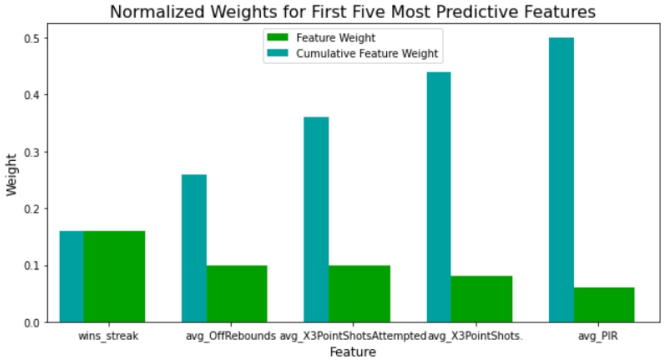

Figure 5: Normalized Weights for First Most Predictive Features

We can say from the aforementioned results and analysis that the effectiveness of feature selection is related to better results and the outcome is noteworthy.

7 Limitations and Future Work

There are a lot machine learning algorithms and the NBA is also growing newer features day by day. So, there is ample amount of chance to further research in several aspect of discussed field.

As we have seen that influencing features are significantly responsible for predicting outcome and features are related to the dataset, it is expected theoretically that better dataset with more features will provide better prediction. However, in our case, we have shown that it not the case all the time. We reduce the data frame with logical reduction resulting a handful of significant data can also be effective.

Again, more attributes and instances can help players and managers significantly. A rich dataset with minor attributes can increase model efficiency. Moreover, among the numerous machine learning algorithm, it is still rare to have explored a better model that will provide a higher accuracy in every occasion. Researches are needed to find function-based techniques. Another great gap is in the prediction of live games. Instant model adjustment can be a solution but it is still difficult formulate such algorithm efficiently. Moreover, a classification system can also perform in increasing efficiency.

8 Conclusion

Prediction in sports is one of the most entertaining competitions among the supporters as well as significantly important for the players, managers, coaches and stakeholders. These prediction in various aspects exploring most important features can uphold the sports industry and open a new dimension of entertainment. In this paper, we have shown the importance of dataset and its preprocessing to predict outcome. We have shown the prediction capability of most of the popular machine learning algorithms and found the best model for captivating. The investigation considers some crucial factors like three-point percentage, field goal percentage, free throw made, and total rebounds make a huge impact on the outcome. The intelligent normalization of dataset increases the accuracy greatly rather than regular dataset. At the same time, it provides the players and managers the features that are needed to consider most, work on, and formulate strategy. In the end, we have pointed some limitations in research on this field and their possible future solution. We hope that the paper provides the researchers a comprehensive idea based on the comparison of different machine learning model and importance of dataset configuration.

References

[1] D. L. Andrews. The (trans) national basketball association: American commodity-sign culture and global-local conjuncturalism. In *Articulating*

the global and the local, pages 72–101. Routledge, 2018.

[2] R. P. Bunker and F. Thabtah. A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1):27–33, 2019.

[3] C. Cao. Sports data mining technology used in basketball outcome prediction. 2012.

[4] G. Cheng, Z. Zhang, M. N. Kyebambe, and N. Kimbugwe. Predicting the outcome of nba playoffs based on the maximum entropy principle. *Entropy*, 18(12):450, 2016.

[5] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[6] M. Haghighat, H. Rastegari, N. Nourafza, N. Branch, and I. Esfahan. A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, 2(5):7–12, 2013.

[7] S. Jain and H. Kaur. Machine learning approaches to predict basketball game outcome. In *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall)*, pages 1–7. IEEE, 2017.

[8] D. Kopf. Data analytics have made the nba unrecognizable, 2018.

[9] C. K. Leung and K. W. Joseph. Sports data mining: predicting results for the college football games. *Procedia Computer Science*, 35:710–719, 2014.

[10] N. Lieder. Can machine-learning methods predict the outcome of an nba game? *Available at SSRN 3208101*, 2018.

[11] B. Loeffelholz, E. Bednar, and K. W. Bauer. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[12] D. Miljković, L. Gajić, A. Kovačević, and Z. Konjović. The use of data mining for basketball matches outcomes prediction. In *IEEE 8th international symposium on intelligent systems and informatics*, pages 309–312. IEEE, 2010.

[13] F. Thabtah. Autism spectrum disorder screening: machine learning adaptation and dsm-5 fulfillment. In *Proceedings of the 1st International Conference on Medical and health Informatics 2017*, pages 1–6, 2017.

[14] F. Thabtah, F. Kamalov, and K. Rajab. A new computational intelligence approach to detect autistic features for autism screening. *International journal of medical informatics*, 117: 112–124, 2018.

[15] F. Thabtah, L. Zhang, and N. Abdelhamid. Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1): 103–116, 2019.

[16] T. A. Zak, C. J. Huang, and J. J. Siegfried. Production efficiency: the case of professional basketball. *Journal of Business*, pages 379–392, 1979.