Annals of Data Science (2019) 6(1):103–116
https://doi.org/10.1007/s40745-018-00189-x

NBA Game Result Prediction Using Feature Analysis and Machine Learning

Fadi Thabtah[1] · Li Zhang[1] · Neda Abdelhamid[2]

Received: 6 October 2018 / Revised: 7 December 2018 / Accepted: 13 December 2018 /
Published online: 3 January 2019

Abstract

In the recent years, sports outcome prediction has gained popularity, as demonstrated by massive financial transactions in sports betting. One of the world's popular sports that lures betting and attracts millions of fans worldwide is basketball, particularly the National Basketball Association (NBA) of the United States. This paper proposes a new intelligent machine learning framework for predicting the results of games played at the NBA by aiming to discover the influential features set that affects the outcomes of NBA games. We would like to identify whether machine learning methods are applicable to forecasting the outcome of an NBA game using historical data (previous games played), and what are the significant factors that affect the outcome of games. To achieve the objectives, several machine learning methods that utilise different learning schemes to derive the models, including Naïve Bayes, artificial neural network, and Decision Tree, are selected. By comparing the performance and the models derived against different features sets related to basketball games, we can discover the key features that contribute to better performance such as accuracy and efficiency of the prediction model. Based on the results analysis, the DRB (defensive rebounds) feature was chosen and was deemed as the most significant factor influencing the results of an NBA game. Furthermore, others 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%.

Keywords Classification · Data mining · Features selection · Machine learning · NBA · Prediction · Sports analytics

✉ Fadi Thabtah
Fadi.fayez@manukau.ac.nz

[1] School of Digital Technologies, Manukau Institute of Technology, Auckland, New Zealand

[2] ITP, Auckland Institute of Studies, Auckland, New Zealand

1 Introduction

NBA stands for National Basketball Association, which is a professional basketball league that is one of the most premier in the world. Sports, including basketball, is an ongoing business that involves numerous financial interests worldwide. The amount of money invested in sports is beyond belief. In 2014, the global sports industry was estimated to be worth 1.5 trillion in USD [3].

Sports analytics involves the learning of useful information which can be utilised by teams as well as individual player to enhance performance. Tactics played by teams, a team's influential players, and the players' fitness levels are just some of the pieces of information club managers and coaches have great interest in. Technologies such as machine learning, which can provide such useful information, are vital to a coach or manager's planning and strategizing tactics in his team. Machine learning primarily focuses on the prediction of reliable outcomes, which can be a useful source for sports betters, sportsmen, club managers, and sponsors. Machine learning methods are used to predict the outcome of a game, which is favorable for betters, as it may increase their confidence in betting. Machine learning furnishes useful patterns which can be tapped by club managers regarding other teams as well as their own team [30, 31].

Sports analytics also goes beyond just predicting the outcomes of a game as it also processes live data when a game is played. To be more specific, teams that play in the NBA have recently adopted a tracking technology with data-streaming capabilities and cameras that record detailed data related to the players' movements and fitness levels. The collected data are then processed through a data analytics tool that is based on machine learning. The tool generates results that reveal strong and weak variables associated with the team and its players, such as player separation on the basketball court, distances travelled, possession of the balls per player, a player's speed while moving on the court, and the like.

Although machine learning techniques have been used in sports outcome prediction, there still are numerous uncertain factors that influence the outcome of games. Thus, the scope of this research is to explore only the influential features set that affects the results from predicting the outcome of a basketball game. In this study, we utilise different machine learning models to process historical data from the NBA final games from the year 1980 through 2017. The objective is to investigate the crucial factors that influence the outcome of NBA games by comparing the performances of different models and evaluating the performances after removing irrelevant features that do not contribute much information about the output variable. The prediction of the outcome of NBA games and the identification of particular features that exert the most significant effects on the results are valuable to different stakeholders such as club managers, team coaches, team players, and betters.

The paper is outlined as follows: In Sect. 1, we present the problem, objectives and the research methodology that we followed. Section 2 reviews related articles on sport analytics with focus on NBA games result prediction using machine learning. Section 3 is devoted to the data and features used and Sect. 4 discusses the experiments and results analysis. Finally, conclusions and limitations are presented in Sect. 5.

2 Literature Review

Although machine learning has been used in different sports analytics, there still is a need to improve the performance of the models offered by this intelligent technology. In this section, we focus on machine learning techniques used for predicting the outcomes of basketball games.

A predictive model that was constructed using historical data from NBA games based on Naïve Bayes Classifier by Cao [6]. The accuracy of the test dataset classification was about 65.82%.

A research for the prediction of NBA game results using artificial neural networks (ANN) was developed by Loeffelholz et al. [22]. A subset of features was compiled by the authors from the unprocessed data and was used as input for the neural nets. Afterwards, various basketball sports experts were given an opportunity to make predictions in order to compare their decisions with the outcomes assigned by the ANN model. The ANN offered higher predictive decisions than the domain experts, i.e. 74.33% [22]. The regression method was constructed based on a historic spread point to develop a method to forecast the spread point of the NFL (National Football League). This kind of research can be used to predict the game results. The NFL winners can be predicted by using the forecast method of the spread point [22].

Back-propagation, self-organizing maps (SOM), and some neural structures were used to predict the outcomes of football games in the National League of Football (NFL) by Purucker [26]. After trying various training procedures, he concluded that back-propagation was the best structure to build a model that delivers the highest predictive accuracy than any other experts in the area of football game predictions. Normal accuracy of 61% was achieved as compared to 72% accuracy by experts [26].

A neural network structure as well as back-propagation were conducted through Kahn to predict the outcome of a football game (NFL). Purucker [26] extended his work and achieved an accuracy of about 75%, which proved to be better than [26] accuracy of 61%. Furthermore, this accuracy rate was better than most of the experts. Differential Statistics was utilised by both authors from the box scores, not from any raw statistics.

Bunker and Thabtah [4] developed a data mining model for predicting the outcome of a game based on features related to the team. The authors suggested that ANN based models are useful and more accurate to forecast the outcome of the game. In the same study, the authors pinpointed to challenges associated with processing data in liver manner.

Haghighat et al. [9] has identified a number of machine learning methods that are frequently applied in sports analytics. The authors have analysed the predictive accuracy of each identified method through in depth review of past literature on the domain. In addition, various different datasets are obtained from different public sources including NBA basketball, the National Football League's (NFL's) and other websites [23, 33, 34].

Cao [6] proposed a framework based on machine learning to forecast the NBA match outcomes. Features related to NBA games such as players statistics, opponents statistics and starting line-up among others are gathered from the NBA's official website. The features are then saved in the database to pre-process and clean the collected data so it can be used to derive models by the machine learning algorithms. The authors applied different learning techniques including Support Vector Machine (SVM) [5], logistic regression [13], Naive Bayes [18] and ANN [28] on the dataset collected to produce sport analytics model for games result forecasting. The results reported that the logistic regression models were superior when contrasted with the consdiered algorithms.

Miljkovic et al. [25] investigated machine learning technology in predicting the outcome of NBA games. The authors applied different machine learning methods including k-nearest neighbor, SVM, decision trees, multivariate linear regression and Naive Bayes and to predict the NBA game result [15, 18, 20, 24, 27]. Data with 141 variables that are associated with NBA games from 2009 to 2010 were collected to evaluate the two machine learning methods. Examples of variables within the dataset are free throws, fouls, three points, field goals, blocked shots, home and away, number of wins so far, number of losses, current streak and wins among others. A number of experiments have been conducted and the results pinpointed that machine learning models can offer up to 67% predictive accuracy in classifying NBA game results.

Kopf [16] pointed out that every NBA team has data analysts who work with coaches to maximize individual player's talents. In the beginning of 2009, most leagues began using a video system to track the movement of the ball and each player on the court 25 times per second. All the data collected by this system allowed data analysts to use intelligent techniques to better assess which players were contributing to team winnings.

Lieder [21] investigated influential features related to matches to create a model for predicting the results of NBA games. The authors applied machine learning algorithms including Logistics Regression, Linear Regression, and ANN to build predictive models to evaluate the outcome of games. Data related to 2014 NBA seasons were collected for the models' training and testing. The results pinpointed that Logistics Regression achieved accuracy close to 70%.

Cheng et al. [7] developed a model for predicting the outcomes of NBA playoffs based on Maximum Entropy principle. The authors collected a total of 10,271 records of NBA games from 2007/2008 to 2014/2015 seasons. The reported results show that the derived models were able to predict the outcome of the game with 74.4% accuracy.

3 Proposed Framework

Basketball is a very prominent game and there are many available datasets for basketball games that can be utilised for sports analytics, i.e. NBA Players Stats from website https://www.kaggle.com. There is a rapid growth in the sports analytics industry in terms of improving prediction results and measuring individual player as well as team performance [29]. Experts and organizations from different domains devote time and capital towards sports analytics research to optimize game strategies by coaches and performance of players [19]. Using machine learning to predict results can offer intelligent models for accomplishing game results forecasting, game strategy, and improvement of players' fitness levels, among others [4]. Identifying the best model for obtaining NBA game outcome predictions is the key research problem. The results

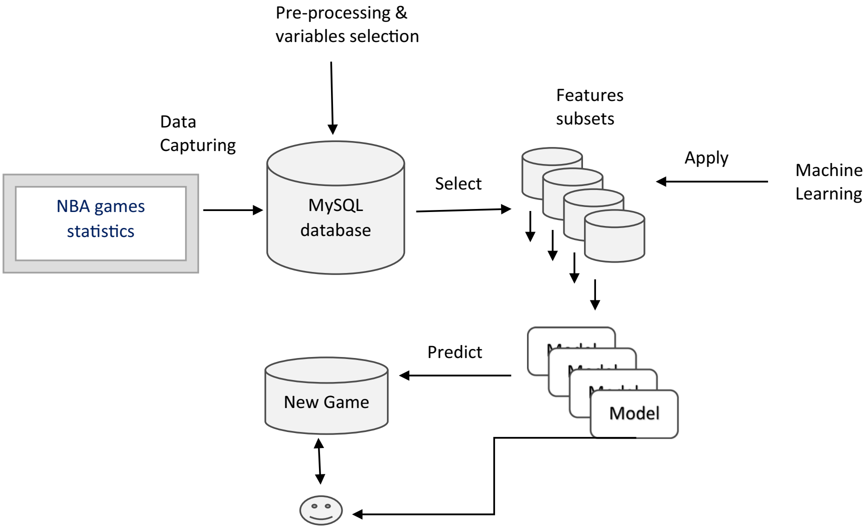


Fig. 1 The proposed framework for NBA games results classification

of this prediction can furnish useful insights for different stakeholders including club managers, betters, bookies, etc.

The objective is twofold. First, we want to detect the influential features set that impacts the outcome of games in the NBA. Second, once the features set is known, we will utilise that information and use machine learning algorithm to build a prediction model. In this case, supervised learning seems the most appropriate method for such an objective. In supervised learning, the input will include a training dataset with independent variables such as assist, steal, and free throws made. All these variables show the team's capabilities against a dependent variable (the outcome of previous games). Afterwards, the aim is to predict the outcome variables by applying a model from historical instances (dependent variables as well as true target variable values). This model will be utilised to forecast the target variable value in an unseen game (test data).

The framework proposed is shown in Fig. 1 in which statistics related to NBA games are collected from different sources using a scripting tool written in PHP. In this research project, the focus was on variables related to teams, players and opponents such as home/away, field goal attempts, three-point attempts, total number of minutes played, free throws attempts, free throws made and offensive rebounds among others. Details on the complete set of variables used and their descriptions are given in Sect. 3.1.

Once the raw data are obtained, a number of pre-processing operations were applied, including missing values replacement and discretization for certain continuous attributes. Moreover, feature selection methods including Correlation Feature Set and Multiple Regression were applied to generate distinctive features sets for further analysis used to remove features that were redundant and may have created biased results. More importantly as shown in Fig. 1 different features sets were derived for data processing using machine learning. This has been carried out to assess the intelligent models derived by the machine learning techniques in predicting the NBA games results. In evaluating the models, several different testing metrics have been utilized including predictive accuracy, error rate, recall, precision and harmonic mean (F1) among others. Section 4 gives further details on the evaluation metrics and the results analysis.

3.1 Data Capturing and Pre-processing

The dataset, named NBA Finals Team Stats used in this study, is obtained from Kaggle [14]. It contains data from the NBA Finals, including game-by-game team totals from the years 1980 through 2017. The game-by-game totals reported were from 11 different teams that participated in the NBA Finals between the year 1980 and 2017. The dataset incorporates 22 variables, including target class and 430 instances. The training dataset includes details of the independent variables such as matches played in home ground, percentage of field goals made, and the like, along with the target variable. The dataset variables and their descriptions are shown in Table 1.

Data pre-processing is an important step prior learning the model from the training dataset, and it includes handling all sort of noise [2]. For instance, dealing with missing values, discretizing continues variables, normalizing certain variables, etc. Since the task involved is classification, the format of outcome WIN should be transformed from numeric to nominal as most conventional classification algorithms do not deal with continuous class variable. The process of transforming the class variable is implemented by using the filter function of WEKA tool [11]. WEKA is a machine learning tool that consists of the implementation of different clustering, association, classification, and feature selection techniques. After transforming the class variable from continuous into categorical format, all the continuous variables have been discretized using WEKA discretization filter. The outcome is processed dataset that are free of noise (Table 2).

3.2 Feature Selection and Learning Models

Feature selection is a critical process that directly affects the performance of the models [1]. In this process, a set of influential features (variables) are identified by removing irrelevant variables, which reduces the input dataset dimensionality [32]. This in turn improves the learning process. In order to achieve good performance in terms of the model's predictive accuracy, key variables are selected during the feature selection process. To accomplish key variable selection, two different filter methods and one rule induction algorithm have been applied to the processed dataset, i.e. Multiple Regression [12], Correlation Feature Set (CFS) [10], and RIPPER algorithm [8]. By utilising these three distinctive methods, the most important features can first be identified and then processed to the models using machine learning techniques.

The results of applying the Multiple Regression method are illustrated in Fig. 2. Results clearly indicate that this method has selected seven variables (Home, TP, TPA,

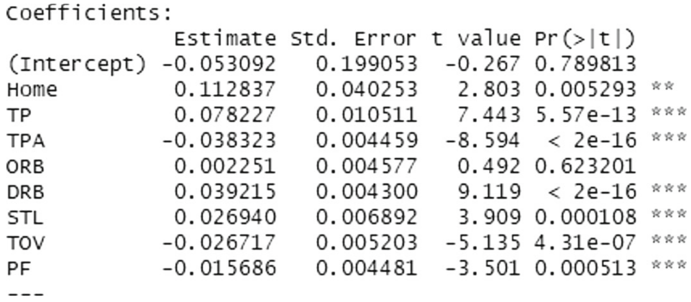
Table 1 The description of attributes

Variable name	Description
Team	The name of the team
Opponent team	A player against another player
Home	1 = home team; 0 = away team
MP	Total number of minutes played
FG	Field goals made
FGA	Field goal attempts
FGP	Field goal percentage (A statistic calculation by dividing field goals made by field goal attempts)
TP	A field goal in a basketball game made from beyond the three-point line
TPA	Three-point attempts (A field goal in a basketball game made from beyond the three-point line)
TPP	Three-point percentage (Ratio of field goals made to field goals attempted)
FT	Free throws made
FTA	Free throws attempted
FTP	Free throw percentage
ORB	Offensive rebound Statistic that measures a player’s number of offensive rebounds in proportion to the total number of offensive rebounds available during active play
DRB	Defensive rebounds Statistic that measures a player’s number of defensive rebounds in proportion to the total number of defensive rebounds available during active play.
TRB	Total rebounds Rebounds are usually grabbed by the front-court players, but guards are also expected to rebound the ball when it falls far from the rim.
AST	Assist An assist is attributed to a player who passes the ball to a teammate in a way that leads to a score by field goal
STL	Steal (A steal occurs when a defensive player legally causes a turnover by his positive, aggressive action)
BLK	Blocks (A block occurs when a defensive player legally deflects a field goal attempt from an offensive player to prevent a score)
TOV	Turnovers
PF	Personal fouls
WIN	1 = win; 0 = loose

DRB, STL, TOV, PF) and has ignored the remaining 14 variables. The CFS method selected seven variables (Fig. 3) as well, two of which were identical to the Multiple Regression method, i.e. (Home and DRB). More interestingly, when the RIPPER algorithm was applied on the complete set of variables, six rules have been generated (Fig. 4). These rules denote that five variables were selected by CFS (FGP, DRB, TRB, TPP, FT), and two variables were obtained by Multiple Regression (PF, DRB). These

Table 2 The description of feature sets

Dataset	FS method	Selected features
A	Complete dataset	Full Features in Table 1
B	M reg.	Home, TP, TPA, DRB, STL, TOV, PF
C	CFS	Home, FG, FGP, TPP, FT, DRB, TRB
D	RIPPER	FGA, FGP, TPP, FT, DRB, TRB, TOV, PF
E	C intersect D	FGP, DRB, TRB, TPP, FT



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.053092   0.199053  -0.267 0.789813    
Home         0.112837   0.040253   2.803 0.005293 ** 
TP           0.078227   0.010511   7.443 5.57e-13 ***
TPA         -0.038323   0.004459  -8.594  < 2e-16 ***
ORB          0.002251   0.004577   0.492 0.623201    
DRB          0.039215   0.004300   9.119  < 2e-16 ***
STL          0.026940   0.006892   3.909 0.000108 ***
TOV         -0.026717   0.005203  -5.135 4.31e-07 ***
PF          -0.015686   0.004481  -3.501 0.000513 ***
---
```

Fig. 2 Results from multiple regression in R

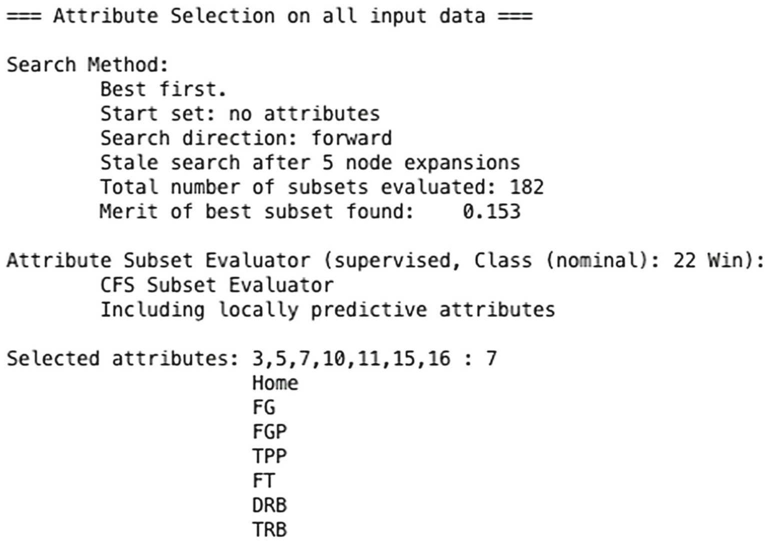
results clearly indicate that the DRB variable is highly influential, as it appeared in all rules derived by RIPPER and was also chosen by the CFS filter method. In addition, DRB was evident in all result sets derived by RIPPER, Multiple Regression, and CFS methods. Hence, DRB is a highly significant element. Surprisingly, the Home variable, which appeared in both CFS and Multiple Regression sets, did not appear in any rules derived by RIPPER algorithm.

After trimming the dataset using different feature selection methods, we managed to reduce the dataset features from 21 to 7 (dataset B), from 21 to 7 (dataset C), from 21 to 8 (dataset D), and from 21 to 5 (dataset E). These new datasets and supplementary to raw dataset A, which signifies we obtained different datasets containing different selected features using Multiple Regression, Correlation Feature Set, and RIPPER methods. The dataset size was reduced by more than half.

Different machine learning algorithms were selected for processing the aforementioned features sets as they adopt different learning schemes and have been successfully applied to other applications. These are utilised to construct the prediction models.

- ANN
- Naïve Bayes
- LMT (Logistical model OGISTIC MODEL TREE)

ANN is a machine learning approach that utilises multiple independent variables and their linked weights to build a network structure. In typical ANN algorithms, training is repeated by adjusting the weights until the desired output is reached. Naïve Bayes is a probabilistic classifier that do not require building a model for classification. Instead Naïve Bayes employs joint probabilities computed from the labelled obser-



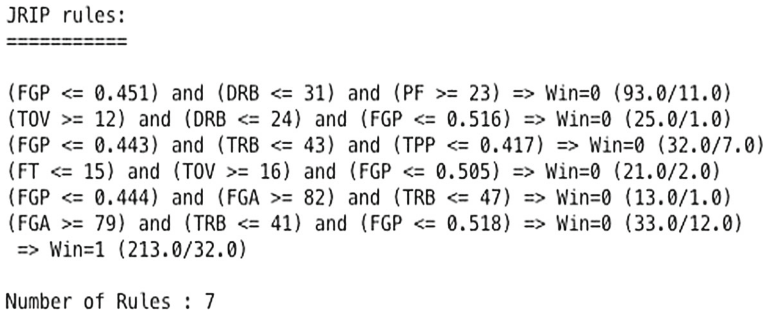
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 182
        Merit of best subset found:    0.153

Attribute Subset Evaluator (supervised, Class (nominal): 22 Win):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,5,7,10,11,15,16 : 7
                     Home
                     FG
                     FGP
                     TPP
                     FT
                     DRB
                     TRB
```

Fig. 3 The selected variables by CFS method in WEKA



```
JRIP rules:
===========

(FGP <= 0.451) and (DRB <= 31) and (PF >= 23) => Win=0 (93.0/11.0)
(TOV >= 12) and (DRB <= 24) and (FGP <= 0.516) => Win=0 (25.0/1.0)
(FGP <= 0.443) and (TRB <= 43) and (TPP <= 0.417) => Win=0 (32.0/7.0)
(FT <= 15) and (TOV >= 16) and (FGP <= 0.505) => Win=0 (21.0/2.0)
(FGP <= 0.444) and (FGA >= 82) and (TRB <= 47) => Win=0 (13.0/1.0)
(FGA >= 79) and (TRB <= 41) and (FGP <= 0.518) => Win=0 (33.0/12.0)
 => Win=1 (213.0/32.0)

Number of Rules : 7
```

Fig. 4 The derived rules by RIPPER algorithm

vations to forecast the class of a test data. Lastly, LMT combines logistic regression models and a decision tree structure to derive a single tree for classification [17].

4 Experimental Analysis

The experiments have been conducted in the WEKA 3.8.2 platform. All the models were tested based on ten-fold Cross Validation [11] and all experiments were performed on a computing machine with 2 GB RAM and 2.26 GHz processing power.

The Evaluation metrics used to assess the performance of the models include predictive accuracy, recall, precision, and harmonic mean (F1). These measures utilise the confusion matrix shown in Table 3, with their associated formulas depicted in the following Equations.

Table 3 The confusion matrix

	Predicted class	
	True	False
Actual class		
True	True positive	False negative
False	False positive	True negative

Accuracy is a measure of how effective the model is at predicting outcomes.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \tag{1}$$

Precision is a measure for positive prediction.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

Recall is a measure of correctly-predicted positive observations to all observations in the positive class.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

F1 Score is the weighted average of Precision and Recall.

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \tag{4}$$

Table 4 provides predictive accuracy rates obtained from different distinctive sets of features (A–E). Results clearly indicate that the classifiers derived by the machine learning algorithms on D features set has the highest accuracy. In addition, the performance of different algorithm models is steadier than features set A, which also obtained the highest accuracy rate. In particular, the accuracy rates obtained using the three different models (Naïve Bayes, ANN, and LMT) built on features set D revealed higher values (80%, 80%, and 83%) compared to the other features sets.

The features set E did not perform as we expected, but the prediction model built by Naïve Bayes algorithm on features set E obtained a higher accuracy rate than full features set A, as shown in above classification accuracy rate matrix. Another useful information we derived from these results is that Multiple Regression (features set A) is not a suitable method for feature selection in this case. When we built the prediction model utilising the reduced features set derived by Multiple Regression method, there was no improvement on prediction accuracy. When we looked at the performance of features set C which was derived by CFS method, prediction accuracy also declined. However, features set D, which was derived by RIPPER feature selection method, produced a better result in terms of accuracy. The accuracy rates of models based on Naïve Bayes and LMT showed obvious improvement. The accuracy rate of LMT models reached 83% as LMT combines logistic regression and decision tree learning. Typically, it obtains better performance than single decision tree or single logistic regression.

Table 4 The prediction accuracy matrix

Dataset	Algorithm	Accuracy	Precision	Recall	F1
	Naïve Bayes	0.76	0.76	0.76	0.76
A	ANN	0.83	0.83	0.83	0.83
	LMT	0.82	0.82	0.82	0.82
	Naïve Bayes	0.73	0.73	0.73	0.73
B	ANN	0.71	0.71	0.71	0.71
	LMT	0.75	0.75	0.75	0.75
	Naïve Bayes	0.77	0.77	0.77	0.77
C	ANN	0.74	0.74	0.74	0.74
	LMT	0.78	0.78	0.78	0.78
	Naïve Bayes	0.80	0.80	0.80	0.80
D	ANN	0.80	0.80	0.80	0.80
	LMT	0.83	0.83	0.83	0.83
	Naïve Bayes	0.78	0.78	0.78	0.78
E	ANN	0.76	0.76	0.76	0.76
	LMT	0.79	0.79	0.79	0.79

Among the five datasets considered, machine learning techniques achieved better performance on the D dataset than the rest. Moreover, the Naïve Bayes model built on E dataset also performed the best among all datasets.

Since the goal of this study is to identify the influential features sets, Table 4 was generated for comparison analysis to evaluate the performance of each features set. The results indicate that dataset D, which includes eight features selected by RIPPER method, exhibited the best classification performance. Three machine learning algorithms were implemented to evaluate the performance of the different features sets. By running different machine learning algorithms on the datasets, we can pinpoint the influential features set that includes the most key factors that affect the outcomes of NBA games in terms of the evaluation metrics. The above analysis signifies that the optimal model (ANN, Naïve Bayes, LMT) was built on features set D with 80%, 80% and 83% accuracy rates for each prediction model, respectively. F1, precision and recall best rates were obtained by LMT algorithm from dataset D (83%, 83%, and 83%) respectively.

Theoretically, more features should generate better results. However, we discovered that after removing several irrelevant features from the full NBA dataset (dataset A), the prediction performance improved. The recall, precision and accuracy rates improved in most models derived from dataset D, than other datasets. Only the performance of ANN algorithm declined from 83 to 80% in regards to accuracy.

From the above analysis, it is noteworthy that the utilisation of feature selection is relevant in obtaining better results. Dataset D includes the most significant features affecting outcomes of NBA matches. This is very precious information for basketball team managers, as they implement strategies to improve their team's playing capabilities.

5 Conclusions and Limitations

There is still controversy surrounding the identification of the influential features set and the best model for predicting NBA game outcomes. In this paper, an intelligent framework was developed based on machine learning and feature selection to deal with the problem of result prediction of NBA games. After investigating various machine learning techniques to build prediction models using different features sets obtained by feature selection methods, we arrive at a conclusion. From the results analysis, the DRB feature (defensive rebounds), which was selected by all the feature selection methods, should be deemed as significant factor affecting the outcomes of NBA matches. Furthermore, other crucial factors such as TPP (three-point percentage), FT (free throws made), FGP (field goal percentage), and TRB (total rebounds) were also selected and considered as influential factors to NBA game outcomes. After reducing the feature vectors through feature selection methods, there was a 2–4% increase in the prediction accuracy rate for the model.

There are still many other factors affecting the selection of influential features. First and foremost is the quality of the training data. In addition, theoretically, more features will result in a better accuracy model. However, this research indicates that it is not always the case. The reduction the dataset features may lead to more effective models and higher classification precision rates.

This paper has analysed and discussed the performances of different features sets. Although the accuracy rate of features set E (FGP, DRB, TRB, TPP,FT) did not meet our expectations, we still obtained some interesting and valuable information. The selection of the features provides some significant insights and the selected features are very useful for NBA coaches to improve their team's capabilities.

Based on what we have completed in this research study, improvements can still be made. For example, more attributes can be considered such as the players of each team and the coach of the team. Furthermore, more instances can also be collected. Although there are numerous machine learning models for sports results prediction, better models with high accuracy rates are still worth exploring especially function based techniques. This is not only because of the popularity of sports betting or for the benefit of sports clubs, but it is also quite useful for new matching strategies formulation. Hence, in near future we are going to investigate deep learners for instant model adjustment while the game is playing live. In addition, we are going to implement a classification system that has the deep learner as the core of making the final prediction with reference to the outcome of games.

References

1. Abdelhamid N, Thabtah F, Abdel-jaber H (2017) Phishing detection: a recent intelligent machine learning comparison based on models content and features. In: Proceedings of the 2017 IEEE international conference on intelligence and security informatics (ISI). Beijing
2. AlShboul R, Thabtah F, Abdelhamid N, Al-diabat M (2018) A visualization cybersecurity method based on features' dissimilarity. Comput Secur 77:289–303
3. Bradly M (2016) ABC News. https://www.abc.net.au/news/2016-01-21/bradley-corruption-inprofessional-sport-should-be-no-surprise/7101508. Accessed 18 Jan 2018

4. Bunker RP, Thabtah F (2017) A machine learning framework for sport result prediction. Appl Comput Inform. https://doi.org/10.1016/j.aci.2017.09.005
5. Burges C (1998) Tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–167
6. Cao C (2012) Sports data mining technology used in basketball outcome prediction. Dublin Institute of Technology. Retrieved from https://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis. Accessed 17 Jan 2018
7. Cheng G, Zhang Z, Kyebambe MN, Kimbugwe N (2016) Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy 18:450. https://doi.org/10.3390/e18120450
8. Cohen W (1995) Fast effective rule induction. Proceedings of the 12th International Conference on Machine Learning 115–123
9. Haghighat M, Rastegari H, Nourafza N (2013) A review of data mining techniques for result prediction in sports. In: Advances in computer science, pp 2322–5157
10. Hall M (1999) Correlation-based feature selection for machine learning. Doctoral dissertation, University of Waikato, Dept. of Computer Science
11. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA Data Mining Software: An Update. SIGKDD Explor 11(1)
12. Higgins J (2005) Introduction to multiple regression, Chapt 4, pp 111–115. Accessed 9 Feb 2018
13. Hosmer D, Lemeshow S (2000) Applied logistic regression. Wiley, New York, pp 236–269
14. Kaggle Inc (2018) Kaggle: your home for data science. Retrieved 24 July 2018, from https://www.kaggle.com/slonsky/boxing-bouts
15. Keller JM, Gray MR, Givens JA (1985) A fuzzy K-nearest neighbour algorithm. IEEE Trans Syst Man Cyberne 580(4):580–585
16. Kopf D (2017) Data analytics have made the NBA unrecognizable. Retrieved from: https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/. Accessed 25 Feb 2018
17. Landwehr N, Hall M, Frank E (2005) Logistic model trees. Mach Learn 95(1–2):161–205
18. Langley P, Iba W, Thompson K (1992) An analysis of Bayesian classifiers. In: The tenth national conference on artificial intelligence, vol. 24. AAAI Press, San Jose, pp 399–406
19. Latheef NA (2017) The number games—how machine learning is changing sports. Retrieved from https://medium.com/@nabil_lathif/the-number-games-how-machine-learning-is-changing-sports-4f4673792c8e
20. Lewis D (1998) Naive (Bayes) at forty: the independence assumption in information retrieval. In: European conference on machine learning, pp 4–15
21. Lieder NM (2018) Can machine-learning methods predict the outcome of an NBA game? 1, Mar 2018. https://ssrn.com/abstract=3208101 or http://dx.doi.org/10.2139/ssrn.3208101
22. Loeffelholz B, Bednar E, Bauer KW (2009) Predicting NBA games using neural networks. J Quant Anal Sports 5(1):1156
23. Mccabe A, Trevathan J (2008) Artificial intelligence in sports prediction. In: Fifth international conference on information technology: new generations (itng 2008). https://doi.org/10.1109/itng.2008.203
24. Meyera D, Leischa F, Hornik K (2003) The support vector machine under test. Neurocomputing 55:169–186
25. Miljkovic D, Gajic L, Kovacevic A, Konjovic Z (2010) The use of data mining for basketball matches outcomes prediction. In: IEEE 8th international symposium on intelligent systems and informatics. SISY, Subotica, pp 10–11
26. Purucker M (1996) Neural network quarterbacking. IEEE Potentials 15(3):9–15. https://doi.org/10.1109/45.535226
27. Quinlan JR (1986) Induction of decision trees. Mach Learn. https://doi.org/10.1007/bf00116251
28. Schalkoff RJ (1997) Artificial neural networks. International ed. McGraw-Hill, New York
29. Steinberg L (2015) Changing the game: the rise of sports analytics. Retrieved from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/. Accessed 15 Feb 2018
30. Thabtah F (2017) Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment. In: Proceedings of the 1st international conference on medical and health informatics. ACM, Taichung City, pp 1–6
31. Thabtah F, Abdelhamid N (2016) Deriving correlated sets of website features for phishing detection: a computational intelligence approach. J Inform Knowl Manag 15(04):1650042

32. Thabtah F, Kamalov F, Rajab K (2018) A new computational intelligence approach to detect autistic features for autism screening. Int J Med Inform 117:112–124
33. Trawinski K (2010) A fuzzy classification system for prediction of the results of the basketball games. In: IEEE international conference on fuzzy systems. Barcelona, pp 1–7. https://doi.org/10.1109/fuzzy.2010.5584399
34. Zdravevski E, Kulakov A (2009) System for prediction of the winner in a sports game. ICT Innov. https://doi.org/10.1007/978-3-642-10781-8_7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.