RESEARCH-ARTICLE

NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors

XI ZHENG, Xi'an Jiaotong University, Xi'an, Shaanxi, China

PDF Download
3538950.3538960.pdf
29 January 2026
Total Citations: 2
Total Downloads: 134

Published: 26 May 2022

Citation in BibTeX format

BDE 2022: 2022 4th International Conference on Big Data Engineering
May 26 - 28, 2022
Beijing, China

BDE '22: Proceedings of the 4th International Conference on Big Data Engineering (May 2022)
https://doi.org/10.1145/3538950.3538960
ISBN: 9781450395632

NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors

Xi Zheng*
School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049
China2194420833@stu.xjtu.edu.cn

ABSTRACT

In recent years, extensive analysis has been applied to predicting NBA game results due to the popularity of basketball and massive financial transactions in NBA betting. The primary objective of this research is to construct a predictive model to precisely forecast the outcome of NBA basketball games in the latest 2020-21 and 2021-22 NBA regular season. We designed features which incorporates both external and internal factors such as teams' Elo rating, average team performance in recent games, home court advantage and tiredness due to back-to-back games. We built up three feature sets and performed feature selection using sequential feature selection (SFS) and recursive feature elimination (RFE) to verify their effectiveness. Results show that novel features such as level of tiredness and difference of Elo ratings between home and away team improves prediction accuracy. To make fair prediction for the latest NBA season, we utilize 10-fold cross validation to train and select models with decent mean accuracy and low standard deviation for final evaluation. It is found that our best random forest model performs fairly well in predicting games in the latest 2020-21 and 2021-22 season with an accuracy of 67.98%. The prediction results and the identification of key features that exert the most significant effects on the results can be helpful and meaningful to different stakeholders in this field, such as team coaches, players and NBA betters.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Supervised learning**; • **Supervised learning by classification**;

KEYWORDS

NBA winner prediction, Sequential forward selection, Recursive feature elimination, Feature engineering, Feed forward neural network, Random Forest, Naïve Bayes

ACM Reference Format:
Xi Zheng. 2022. NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors. In *2022 4th International Conference on Big Data Engineering (BDE 2022), May 26–28, 2022, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3538950.3538960

1 INTRODUCTION

The National Basketball Association (NBA) composed of 30 teams is a professional basketball league in the United States. The NBA regular season usually runs from November to April, which is 1230-game long, with each team playing 82 games, 41 at home and 41 away from home. From April to June, the top 8 teams that stand out from the regular season participate in the playoffs where the league champion is determined.

In recent years, extensive analysis has been applied to understand what contributes to a champion team and better predict future games. Cao (2012) constructed a predictive model based on Naïve Bayes Classifier. The classification accuracy of the test dataset was about 65.82% [1]. Thabtah et al (2019) discovered influential features set that affects the outcomes of NBA games based on several machine learning methods such as decision tree, artificial neural network and naïve bayes [2]. They found that the key features that help making better predictions are DRB(Defensive Rebounds), TPP(Three-Point Percentage), FT(Free Throw), and TRB(Total Rebounds), all of which subsequently increased the prediction accuracy by 2-4%. Based on players' statistics, Nguyen et al.(2021) utilized machine learning (ML) and deep learning (DL) to predict players' future performance and whether they will be selected in All-Star game [3]. Their study shows the performance of DL algorithms is not as good as ML on structured, relatively small-scale basketball datasets. Besides scholarly articles, professional sports forecasting companies also announced their accuracies on game results prediction across various NBA seasons: FiveThirtyEight correctly predicted the winner of 66.42% games during the 2017-18 NBA season [5], while NBA Miner correctly predicted the winner of 65.30% games during the 2015-16 seasons [6]. Generally, a good performance rate is seen between 65%-70%.

Most past researches utilized historical in-game statistics of teams, players or both for prediction. However, the results of NBA games are not only determined by intrinsic team competencies, but also an interplay of external factors, such as home court advantage, tiredness due to successive games and long trips between these games. Besides, few studies have researched the optimal time range of historical data that should be used for prediction, that is, how many past games should be taken into consideration when evaluating the team's recent performance for making accurate game result predictions. To fill up the void and make better predictions, in this article, we utilized not only historical box scores and in-game statistics such as ORB(Offensive Rebounds) and FT(Free Throw), but also the location of game and level of tiredness for prediction. Besides, a series of novel features that emphasize the difference of team statistics between home and away team are also created and

*Corresponding author

BDE 2022, May 26–28, 2022, Beijing, China
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9563-2/22/05…$15.00
https://doi.org/10.1145/3538950.3538960

Table 1: Abbreviations for Basketball In-Game Statistics

Full Name	Abbreviation
Points	PTS
Total Rebounds	TRB
Offensive Rebounds	ORB
Defensive Rebounds	DRB
Assists	AST
Turnovers	TOV
Steals	STL
Blocks	BLK
All Free Throws	FT
Live Free Throws	LFT
Field Goals Attempts	FGA
Field Goals Made	FG
Field Goals Missed	FGM
2-Points Field Goals Attempts	2FGA
2-Points Field Goals Made	2FG
2-Points Field Goals Missed	2FGM
3-Points Field Goals Attempts	3FGA
3-Points Field Goals Made	3FG
3-Points Field Goals Missed	3FGM

used for prediction. Our results showed that adding these novel features into predictive models increased the prediction accuracy by 1.5%-2%. We further applied feature selection based on sequential forward selection (SFS) and recursive feature elimination (RFE) algorithm to identify the optimal feature subsets that achieve the highest prediction accuracy. Based on historical data from 2012-13 season to 2020-21 season, our optimal model gives a prediction accuracy of 67.98% in predicting the latest 2020-21 and 2021-22 NBA regular season.

This paper is outlined as follows: Section 1 presents the objectives, related articles, and the framework. Section 2 introduces data source and preprocessing process. Section 3,4 discuss feature engineering and the exploratory data analysis. Section 5,6 introduce the three feature sets and the feature selection methods. Section 7,8 present the selection of models as well as the final evaluation of top models. Section 9,10 draw conclusions and discuss possible future works.

2 DATA

2.1 Data Source

Due to the prevalence of sport data science and immense popularity of basketball and NBA games, there are various types of open-source NBA data that can be utilized for sports analytics and predictions, i.e. https://www.nba.com/, https://www.basketball-reference.com/ and https://www.espn.com/. The dataset used in this article is collected in December 2021 from Synergy Sports (https://synergysports.com/., which is a professional basketball database partners with NBA, WNBA, NCAA to create web-based, on-demand video-supported basketball analytics for the purposes scouting, development and entertainment. Synergy Sports also provides subscribers with videos of each game and an extensive range of detailed indicators including season summary, game situations, team performances, individual player behaviours, which offers valuable first-hand resources for advanced and deep investigation into the nature, laws and trends of basketball.

2.2 Data Preprocessing

Since the scraped data is stored by season, the first step of preprocessing is to combine datasets of different seasons into one dataset. The combined dataset comprises of 10364 rows and 209 columns of data. Besides, since each game comprises of 4 sections (12 minutes each) and the in-game statistics in the scraped datasets are measured on a per section basis which are too detailed for analytics, we add up the statistics of each section to obtain the statistics of the entire game and delete the original columns of section data. Finally, we obtain a clean dataset which consists of 10364 rows and 199 columns. To make it concise, the following abbreviations for basketball in-game statistics in Table 1 will be used in this article.

3 FEATURE ENGINEERING

In previous sections, a cleaned and structured dataset is obtained after scraping and preprocessing the data. However, this form of data is not of much use for future predictions because all those in-game statistics will be known only after a game is ended, not before the game, which makes them useless for predicting future game results. Therefore, in this section, some new features that are useful for predictions for future games are created based on the original datasets. In order to make better predictions, not only traditional internal factors such as teams' in-game statistics including PTS, TRB, ORB, DRB, AST and so on, but also external factors that potentially affect the outcome of game such as home court advantage and players' tiredness due to successive games are taken into consideration. The overall framework of designed features is shown in Figure 1

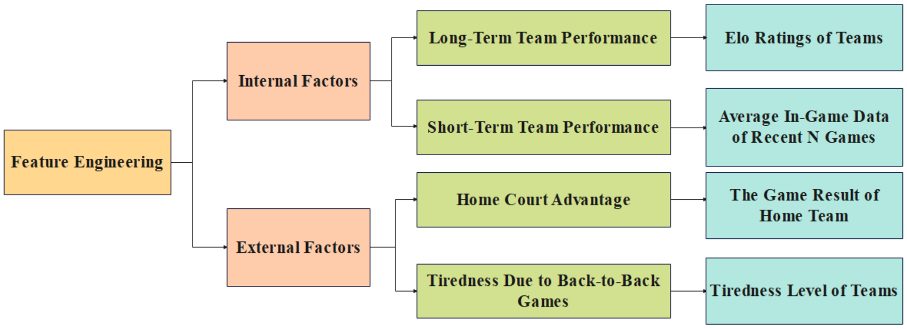
Feature Engineering
Internal Factors
External Factors
Long-Term Team Performance
Short-Term Team Performance
Home Court Advantage
Tiredness Due to Back-to-Back Games
Elo Ratings of Teams
Average In-Game Data of Recent N Games
The Game Result of Home Team
Tiredness Level of Teams

Figure 1: Framework of Feature Engineering Including Both External and Internal Factors

3.1 Internal Factors: Long-Term and Short-Term Performance

Internal factors that determine which team will win the game refer to teams' intrinsic strengths, such as the team's overall abilities of three-point shooting, rebound, block, steal and assist. Considering the fact that competitiveness of teams and players shows both short-term and long-term characteristics, we evaluate the current strength of a team by both its long-term and short-term performance. We utilize the team's Elo rating to represent its long-term performance since the 2012-13 season, and utilize the team's average in-game statistics in the recent N games to represent its short-term performance, where N is a hyperparameter.

3.1.1 Short-Term Team Performance: Average Statistics in Recent N Games. Historical data shows that even within a single season, the intrinsic strength of NBA teams often fluctuate drastically. As a result, a team's performance in recent games becomes an important predictor of its future success. In this article, we use the average performance over the past N games to represent a team's recent strength, where N is a constant. A larger value of N reflects team's performance within a relatively long period, while a smaller N reflects team's performance within a relatively short period, which is more adaptive and sensitive to latest changes. To discover the optimal N that achieves the highest predictive accuracy, we utilize N =3,4,5,. . .,10 to construct 8 sets of features and train models on them respectively. For each N, 38 features are created, 19 features for home team and 19 for away team, which includes the average PTS, TRB, ORB, DRB, AST, TOV, STL, BLK, FT, LFT, FGA, FG, FGM, 2FGA, 2FG, 2FGM, 3FGA, 3FG, 3FGM in recent N games. Note that the calculation of the above features is in fact the process of properly inserting moving averages of team statistics according to chronological order. We first rearrange the dataset by dates of games. For each game in the dataset, we then search upwards in our dataset for the most recent N games that the home team had played and calculate the mean of certain statistics. After that, we do the same for the away team.

3.1.2 Long-Term Team Performance: Elo Ratings. To evaluate the long-term performance of sports players and teams, a widely adopted methodology is the Elo rating [4], a method first proposed by the American physicist Arpad Elo for calculating the relative skill levels of players in zero-sum games such as chess. When applied to sport games such as NBA basketball games, it can measure the long-term performance of teams based on their win-loss record. The basic principle is that the winner team will gain more Elo points from the loser team when defeating stronger rivals or winning by larger margins. Take the game between Phoenix Suns and Golden State Warriors on Nov 30th, 2021 as an example, before the game started, Phoenix Suns has a rating of 1684.838 and Golden State Warriors has a rating of 1664.601. Denote the Suns as team A and Warriors as team B, then the winning probability of team A (Phoenix Suns) is

$$P_{win} = \frac{1}{1 + 10^{\frac{-(Elo_A - Elo_B + hca)}{400}}} = \frac{1}{1 + 10^{\frac{-(1684.838 - 1664.601 + 69)}{400}}} = 0.626$$

where hca represents home court advantage which approximately equals to 69 Elo points in previous researches [7]. Since team A (Phoenix Suns) defeated team B (Golden State Warriors) by 8 points (104 to 96), the Elo points that team A received from team B is:

$$\Delta Elo = k \times (1 - P_{win}) = 18.457 \times (1 - 0.626) = 6.909$$

where k is a moving constant that depends on the margin of victory and difference in Elo ratings between two teams in a certain game, which is formulated by:

$$k = 20 \times \frac{(MoV + 3)^{0.8}}{7.5 + 0.006 \times Elo_{diff}} = 18.457$$

where MoV represents the margin of victory which is the points that winner team gets more than loser team, and Elo_{diff} represents the difference of Elo rating between team A and B. It can be easily inferred that a lager MoV and a smaller Elo_{diff} can both result in a lager k, making the transfer of Elo points ΔElo from the winner to the loser team much more significant. Thus, the Elo ratings of two teams after this game are:

$$Elo_A + \Delta Elo = 1684.838 + 6.909 = 1691.747$$

$$Elo_B - \Delta Elo = 1664.601 - 6.909 = 1657.692$$

That is, after this game, the Elo rating of Phoenix Suns was increased to 1691.748, while the one of Golden State Warriors was decreased to 1657.692. Using the same method in this example, the Elo rating of home team and away team before and after each game can be calculated. Note that initial values of the Elo rating of all teams are set to 1500 ($Elo_{Initial}$) at the beginning of the

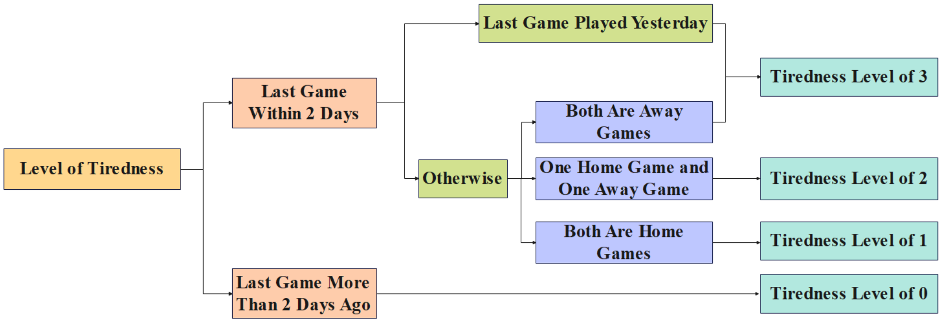

Figure 2: Level of Tiredness Determined by Dates and Locations of Games

2012-13 season. Then, teams' Elo ratings before and after each game from 2012-13 season to 2021-22 season are calculated via an iterative algorithm based on the points of two teams and the location of the game. In order to capture the uncertainty caused by team recruitments between seasons, at the beginning of each season the Elo Rating of each team is reset as:

$$Elo_{New\ Season} = 0.75 \times Elo_{Previous\ Season} + 0.25 \times Elo_{Initial}$$

where $Elo_{Initial}$ is the initial value of the Elo Rating System of NBA and is set to 1500 in this article.

3.2 External Factors: Level of Tiredness and Home Court Advantage

Only utilizing internal factors for prediction inevitably results in inaccuracy because multiple external factors that potentially affect game results are neglected, i.e. tiredness due to back-to-back games, home court advantage and players' injuries. Here we consider tiredness and home court advantage for prediction.

3.2.1 Level of Tiredness. NBA games are often scheduled on Friday and weekend nights to attract more audience. It is common occurrence that teams have to fly to another city for tomorrow night's game right after tonight's game is finished. Games played by the same team consecutively on two days or even within the same day are called back-to-back games, which along with long travels unavoidably cause stress and tiredness of players, thus affecting team performances and even skewing game results. Historical data shows that many times outcomes of games were skewed because teams were playing back-to-back. According to NBC Sports, in 2008-09 season, 22 out of 30 NBA teams had worse records in games which were back-to-back.

In our model, the tiredness of team caused by back-to-back games is categorized to 4 types. If a team played its last game yesterday or played 2 away games within the past 2 days, its tiredness is defined as 3. If a team played one away game and one home game within the past 2 days, its tiredness is defined as 2. If a team played two home games within the past 2 days, its tiredness is defined as 1. However, if a team played its last game more than 2 days ago, the team does not suffer from tiredness and therefore its tiredness is defined as 0. In this way, the tiredness level of home and away team of each game is calculated and added to our dataset as new features.

3.2.2 Home Court Advantage. Home Court Advantage refers to the benefits that the home team gains over the visitor team in sport games. From physiological perspective, playing in familiar environment make teams play better because players do not need to recover from jet lag, adapt to unfamiliar climates and foods. In addition, psychological advantages of home team also exist, for example, the support from the local audience. Previous research found that home court advantage approximately equals to 69 Elo points. Therefore, we add extra 69 Elo points to home team in each game as is shown in formula (1). Besides, labels of models are coded as 1 if home team wins and 0 if it loses, which implicitly distinguish the home and away team so that the model can discover the hidden rules and patterns of home court advantage via learning from past data.

4 EXPOLORATORY DATA ANALYSIS

4.1 Density of Elo Ratings

4 4 shows the kernel density estimation of Elo ratings of 30 NBA teams from 2012-13 season to 2021-22 season. It can be seen that in most seasons, the distribution of Elo ratings subjects to normal distribution. Seasons like the 2013-14 and 2019-20 season contain a more even distribution of teams' strengths, while seasons like the 2012-13 and 2017-18 season shows a more skewed and imbalanced distribution. In seasons where teams have similar strengths, the outcomes of games are more uncertain and difficult to predict, and are likely to rely much more on external factors such as teams' tiredness level and home court advantage.

4.2 Winning Rate by Teams

5 5 shows the accumulative winning rate of 30 NBA Teams over the past 10 regular seasons since 2012-13 season. The most successful team over the past 10 regular seasons is GSW (Gold State Warriors), followed by SAS (San Antonio Spurs), LAC (Los Angeles Clippers), TOR (Toronto Raptors) and OKC (Oklahoma City Thunder). Teams with the most upset winning rate are ORL (Orlando Magic) and MIN (Minnesota Timberwolves).

Besides, it can be inferred that none of the 30 NBA teams dominate the league over the past 10 years because even the strongest team does not achieve an accumulative winning rate over 70 percent in regular seasons. This illustrates that teams in the NBA league

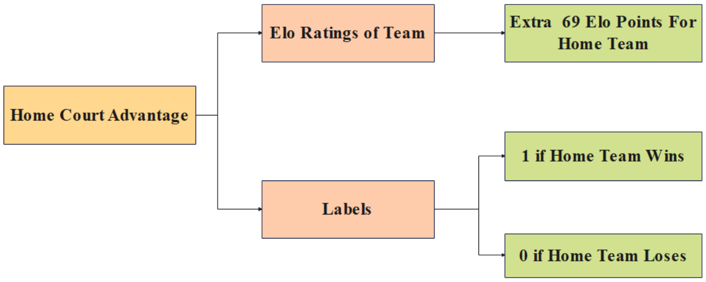

Figure 3: Home Court Advantage

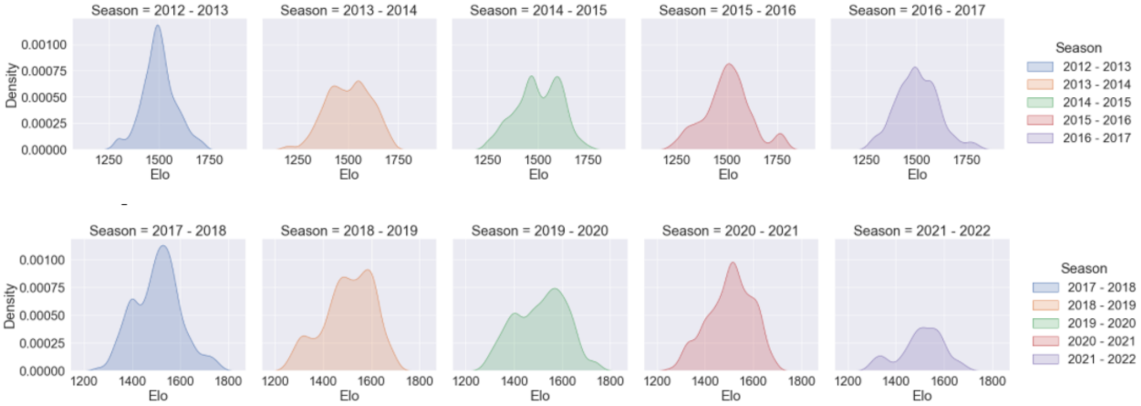

Figure 4: Density of Elo Ratings Per Season

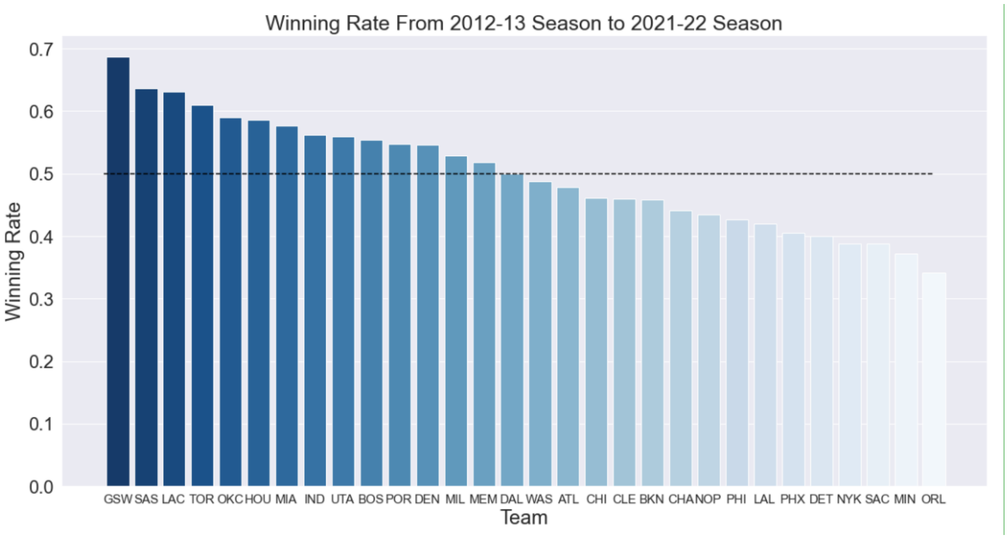

Figure 5: Winning Rate From 2012-13 Season to 2021-22 Season

are very close to one another in terms of team strength and share an even chance to win at most of the time. Thus, predicting the results of NBA games are not a trivial task. A prediction accuracy close to 70 percent is already quite satisfying.

5 BUILDING UP FEATURE SETS

After creating features for future game prediction, we verify the effectiveness of them by training and testing models respectively on three feature sets.

5.1 Feature Set A

Feature Set A contains 38 features of average in-game statistics in recent games, Elo ratings and home court advantage, which is the feature set utilized by past researches and has been proved to be effective. Based on feature set A which is treated as the baseline feature set, we further extend it to feature set B and C by proposing some new features.

5.2 Feature Set B

Feature Set B extends feature set A by adding two tiredness variables for home team and away team, respectively. If models trained on feature set B outperform the ones trained on A, the tiredness features can be proved to be effective.

5.3 Feature Set C

Feature Set C further extends feature set B by adding 21 new features that emphasize the differences between home and away team, which are calculated by home team statistics subtracting the corresponding away team statistics divided by home team statistics. We denote them by D, which stands for differences. For example, we denote Elo rating of home team subtracting Elo rating of away team divided by Elo rating of home team as “D_{Elo}” and average ORB of home team subtracting average ORB of away team divided by average ORB of home team as “D_{ORB}”. The motivation of creating these features is to make features more straightforward and understandable to models since they directly reflect to which extent the home team is performing better than the away team. If models trained on feature set C outperforms those trained on A and B, the “D” variables are proved to be effective.

6 FEATURE SELECTION

6.1 Sequential Forward Selection

Sequential Forward Selection (SFS) is an iterative feature selection method where the model starts with the best performing feature against the target, and then select another feature that gives the best performance in combination with the first selected variable. This process continues until the preset stopping condition is satisfied, for example, the number of features to be selected is reached.

6.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is an iterative feature selection method which selects features by recursively considering smaller sets of features. To begin with, the estimator is trained on the initial set of features. Then, the least important features are pruned from the current set of features. This procedure is recursively repeated until the preset number of features to be selected is reached.

7 MODEL SELECTION

In this section, we rearrange the dataset by chronological order and split it into training set (90%) and testing set (10%) without shuffling the data. Then, based on three feature sets, different machine learning models are trained using 10-fold cross validation on training set which contains 9161 NBA games from 2012-13 to 2020-21 season. After that, top models are chosen for final evaluation on testing set which contains 1036 games from 2020-21 season to 2021-22 season. After a wide selection of models, the goal is to find out the best predictive model for future games, especially games in the latest 2020-21 and 2021-22 season.

7.1 Feed Forward Neural Network

From Table 2, it can be seen that neural network models trained on feature set B, C generally outperforms those trained on feature set A for each N, demonstrating that new features are helpful for the prediction. When selecting models for further hyperparameter tuning and final evaluation, we take both mean accuracy and standard deviation into consideration. The chosen ones are indicated in bold in Table 2

7.2 SVM

From Table 3, it can be seen that SVM models trained on feature set C generally have much lower standard deviation and relatively high mean accuracy, demonstrating that new features in B and C improve models’ performance. Models indicated by bold is selected due to low variance and relatively high predictive accuracy.

7.3 Logistic Regression

Three conclusions can be drawn from Table 4. The first conclusion is that logistic regression models trained on feature set B, C generally outperforms models trained on feature set A. The second conclusion is that feature selection using SFS and RFE improves model performance by both reducing the standard deviation and enhancing the accuracy. And the third is that the SFS significantly outperforms RFE algorithm when implementing feature selection for logistic regression on our datasets.

As shown in Table 4, the best model achieves an accuracy of 67.39% and a standard deviation of 1.11%, which is the logistic regression model with SFS feature selection trained on feature set C using information of the past 3 games. Among 61 features, 21 influential features are selected by the SFS algorithm, including tiredness of home and away team, and 7 “D” features, which verifies the effectiveness of the new features we design.

7.4 Random Forest

From Table 5, we can see that SFS and RFE feature selection do not make significant improvement to random forest models. This may be due to the decision tree, the base learner of random forest, is able to identify and choose important features when deciding where to split during training. Therefore, no external feature selection algorithm is needed any more. It is also observed that random forest models without feature selection generally have a 0.1%-2% lower

Table 2: Cross Validation Results for Feed Forward Neural Network Models

Feature	Model	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Feature Set A	FFNN	66.58% (0.91%)	66.58% (0.04%)	66.76% (0.20%)	66.59% (0.90%)	66.83% (0.29%)	66.78% (0.86%)	66.75% (0.75%)	**66.80%** **(0.15%)**
Feature Set B	FFNN	66.83% (0.76%)	66.75% (0.62%)	66.65% (0.77%)	66.66% (0.44%)	66.71% (0.40%)	66.70% (0.42%)	66.82% (0.38%)	**66.88%** **(0.16%)**
Feature Set C	FFNN	66.90% (1.04%)	67.01% (0.23%)	67.01% (0.53%)	66.80% (0.42%)	66.90% (0.79%)	66.84% (0.53%)	**66.93%** **(0.09%)**	66.29% (1.40%)

*The bold indicates the selected models.

Table 3: Cross Validation Results for SVM Models

Feature	Model	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Feature Set A	SVM	66.77% (2.01%)	**66.81%** **(1.72%)**	66.65% (2.17%)	66.58% (2.04%)	66.78% (2.51%)	66.81% (2.94%)	66.69% (1.87%)	66.52% (1.86%)
Feature Set B	SVM	66.53% (2.72%)	66.58% (2.95%)	66.55% (2.51%)	66.46% (2.26%)	66.67% (2.23%)	66.65% (1.84%)	66.73% (2.52%)	**66.74%** **(0.31%)**
Feature Set C	SVM	66.66% (2.12%)	66.78% (1.18%)	**66.78%** **(0.22%)**	66.61% (0.42%)	66.72% (0.48%)	66.48% (0.35%)	66.56% (0.58%)	66.62% (0.64%)

*The bold indicates the selected models.

Table 4: Cross Validation Results for Logistic Regression Models

Feature Set	Model	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Feature Set A	SFS_LR	67.08% (1.03%)	67.07% (0.85%)	67.11% (0.72%)	67.01% (0.87%)	67.21% (1.18%)	67.09% (1.07%)	67.09% (1.10%)	67.03% (1.21%)
	RFE_LR	66.72% (2.35%)	66.86% (1.69%)	66.87% (2.24%)	66.76% (2.48%)	66.80% (2.26%)	66.66% (2.06%)	66.71% (2.15%)	66.76% (2.23%)
	LR	66.64% (2.23%)	66.86% (2.06%)	66.94% (2.12%)	66.73% (2.24%)	66.79% (2.00%)	66.89% (1.85%)	66.90% (1.89%)	66.85% (1.97%)
Feature Set B	SFS_LR	67.17% (1.03%)	67.10% (0.69%)	67.15% (0.90%)	67.11% (1.02%)	**67.28%** **(1.29%)**	67.23% (1.10%)	67.21% (1.06%)	67.00% (0.80%)
	RFE_LR	66.82% (2.59%)	66.89% (1.96%)	66.86% (2.53%)	66.74% (2.35%)	66.79% (2.19%)	66.87% (2.16%)	66.76% (2.43%)	66.83% (2.45%)
	LR	66.79% (2.50%)	66.86% (1.90%)	66.98% (2.50%)	66.78% (2.16%)	66.78% (2.85%)	66.89% (2.30%)	66.85% (2.26%)	66.74% (2.23%)
Feature Set C	SFS_LR	**67.39%** **(1.11%)**	67.35% (1.03%)	67.35% (1.17%)	67.13% (1.10%)	67.30% (0.94%)	67.14% (1.15%)	67.25% (1.24%)	67.16% (1.11%)
	RFE_LR	66.88% (2.47%)	66.86% (2.19%)	66.89% (1.87%)	66.73% (1.90%)	66.80% (1.80%)	66.77% (2.17%)	66.85% (1.97%)	66.96% (2.10%)
	LR	66.84% (2.04%)	66.78% (2.00%)	66.91% (2.52%)	66.67% (1.78%)	66.77% (1.99%)	66.84% (1.92%)	**67.04%** **(2.19%)**	66.92% (2.02%)

*The bold indicates the selected models.

standard deviation compared with other machine learning models trained on the same datasets which is due to the nature of the random forest algorithm ——it aggregates multiple high variance and low bias decision trees together to reduce variance and enhance overall performance.

7.5 Naïve Bayes

From Table 6, it can be seen that naïve bayes models with SFS feature selection outperform models without feature selection for game data from 2012-13 to 2020-21 NBA season. The best model achieves a mean accuracy of 67.48% with a low standard deviation of 1.03%, which is the naïve bayes model with SFS feature selection trained on Feature Set C using information over the past 5 games. Among 61 features, 15 influential features are selected by the SFS,

Table 5: Cross Validation Results for Random Forest Models

Feature Set	Model	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Feature Set A	SFS_RF	66.39%	66.64%	66.58%	66.39%	66.87%	66.76%	**67.12%**	66.86%
		(1.10%)	(1.05%)	(1.17%)	(1.21%)	(1.15%)	(1.29%)	**(1.07%)**	(1.63%)
	RFE_RF	65.88%	66.10%	65.98%	65.81%	65.98%	66.20%	66.41%	66.23%
		(3.20%)	(2.07%)	(2.21%)	(2.60%)	(2.51%)	(2.40%)	(3.17%)	(2.37%)
	RF	66.66%	66.75%	66.72%	66.90%	67.07%	66.89%	66.83%	66.85%
		(2.33%)	(2.11%)	(1.99%)	(1.45%)	(1.73%)	(1.92%)	(1.91%)	(2.01%)
Feature Set B	SFS_RF	66.78%	66.66%	66.41%	66.48%	66.26%	66.41%	66.65%	66.90%
		(1.02%)	(1.09%)	(1.58%)	(1.06%)	(0.29%)	(0.40%)	(1.34%)	(1.28%)
	RFE_RF	65.74%	66.08%	66.14%	66.06%	66.15%	65.93%	66.08%	66.37%
		(3.39%)	(3.42%)	(1.84%)	(3.47%)	(1.68%)	(2.67%)	(3.47%)	(2.14%)
	RF	66.58%	66.68%	66.76%	66.94%	66.97%	66.88%	66.89%	66.84%
		(2.10%)	(1.91%)	(1.68%)	(1.43%)	(2.05%)	(1.95%)	(1.88%)	(1.99%)
Feature Set C	SFS_RF	66.88%	66.65%	66.67%	66.61%	66.42%	66.68%	66.51%	66.89%
		(1.60%)	(1.29%)	(0.03%)	(0.12%)	(0.48%)	(1.29%)	(0.41%)	(0.22%)
	RFE_RF	66.35%	66.22%	66.28%	66.24%	66.50%	66.39%	66.22%	66.48%
		(1.77%)	(2.33%)	(2.33%)	(2.03%)	(2.69%)	(2.80%)	(2.27%)	(2.82%)
	RF	66.53%	66.86%	**67.13%**	66.71%	66.79%	66.73%	66.61%	**66.73%**
		(1.17%)	(0.60%)	**(0.41%)**	(0.04%)	(0.04%)	(0.18%)	(0.47%)	**(0.05%)**

*The bold indicates the selected models.

Table 6: Cross Validation results for Naïve Bayes Models

Feature Set	Model	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Feature Set A	SFS_NB	67.06%	67.30%	67.18%	67.05%	67.14%	67.14%	67.15%	67.10%
		(1.03%)	(1.21%)	(0.63%)	(1.15%)	(1.50%)	(1.20%)	(0.94%))	(0.83%)
	RFE_NB	\	\	\	\	\	\	\	\
	NB	66.82%	66.99%	66.88%	66.82%	66.97%	67.01%	66.91%	67.06%
		(2.34%)	(1.83%)	(1.91%)	(1.84%)	(2.39%)	(2.31%)	(2.63%)	(2.89%)
Feature Set B	SFS_NB	67.05%	67.16%	67.18%	67.02%	67.24%	67.17%	67.15%	67.10%
		(0.91%)	(1.20%)	(0.63%)	(1.27%)	(1.32%)	(1.34%)	(0.94%)	(0.83%)
	RFE_NB	\	\	\	\	\	\	\	\
	NB	66.82%	67.01%	66.88%	66.83%	66.96%	67.01%	66.90%	**67.06%**
		(2.36%)	(1.82%)	(1.91%)	(1.86%)	(2.41%)	(2.31%)	(2.38%)	**(2.71%)**
Feature Set C	SFS_NB	67.18%	67.30%	**67.48%**	67.16%	67.32%	67.26%	67.04%	67.26%
		(1.14%)	(0.98%)	**(1.03%)**	(1.30%)	(1.49%)	(1.31%)	(1.11%)	(1.05%)
	RFE_NB	\	\	\	\	\	\	\	\
	NB	66.59%	66.61%	**66.60%**	66.60%	66.59%	66.54%	66.55%	66.56%
		(2.17%)	(2.17%)	**(2.02%)**	(2.15%)	(2.01%)	(2.04%)	(2.23%)	(2.27%)

*RFE cannot be implemented on naïve bayes classifier since the algorithm does not have feature importance. The bold indicates the selected models.

including tiredness of home team, tiredness of away team, and 5 "D" features (tiredness, Elo rating, average 2FG, 2FGA and TOV), which indicates the effectiveness of the new features in feature set B and C.

8 FINAL EVALUATION

8.1 Fair Prediction

It is worth noting that our prediction is "fair" because we do not shuffle the dataset and stratify it by label when splitting the training and testing set. Instead, based on our 10 season's data, we utilize the first 90% data to train and select models, and then forecast game

Table 7: Fair Prediction Results of the latest 2021-22 Season

Model	Feature Set	N	Cross Validation Accuracy(Standard Deviation)	Testing Accuracy
FFNN	Feature Set A	N=10	66.80%(0.15%)	**66.50%**
FFNN	Feature Set B	N=10	66.88%(0.16%)	**67.58%**
FFNN	Feature Set C	N=9	66.93%(0.09%)	64.73%
SVM	Feature Set A	N=4	66.81%(1.70%)	64.63%
SVM	Feature Set B	N=10	66.74%(0.31%)	**67.39%**
SVM	Feature Set C	N=5	66.78%(0.20%)	64.05%
SFS_LR	Feature Set B	N=7	67.28%(1.29%)	64.44%
SFS_LR	Feature Set C	N=3	67.39% (1.11%)	64.44%
LR	Feature Set C	N=9	67.04%(2.19%)	64.73%
SFS_RF	Feature Set A	N=9	67.12%(1.07%)	**65.32%**
RF	Feature Set C	N=5	67.13%(0.41%)	64.93%
RF	Feature Set C	N=10	66.73%(0.05%)	**67.98%**
NB	Feature Set B	N=10	67.06% (2.71%)	**67.88%**
NB	Feature Set C	N=5	66.60%(2.02%)	63.16%
SFS_NB	Feature Set C	N=5	67.48% (1.03%)	63.85%

*The bold indicates accuracy above 65% in the testing set.

Table 8: Top Five Models

Model	Feature Set	N	Test Accuracy	AUC
FFNN	Feature Set A	N=10	66.50%	60.40%
FFNN	Feature Set B	N=10	67.58%	62.57%
NB	Feature Set B	N=10	67.88%	63.88%
SVM	Feature Set B	N=10	67.39%	63.06%
RF	Feature Set C	N=10	67.98%	63.96%

results of the 10% data in the latest 2020-21 and 2021-22 season, which may decline our accuracy but makes the results more useful and meaningful in real world application.

8.2 Prediction Results of the Latest 2020-21 and 2021-22 NBA Season

Based on previous analysis, 15 models are selected to conduct further hyperparameter tuning. Models with optimal hyperparameters are tested on our testing set which contains 1036 games in the latest 2020-21 and 2021-22 season. The results are shown in 7 7, where several conclusions can be drawn: First, most of the models trained on feature subsets selected by SFS algorithm do not perform as well as models trained on complete feature sets on our testing set, which is probably due to the information loss caused by feature selection. Moreover, random forest and naïve bayes are the best models for the NBA prediction task.

To highlight best models, models that give an accuracy of over 65% on testing set are indicated by bold text in 7 7 and displayed in detail in 8 8 and 6 6. It is indicated that N=10 is the most suitable choice for NBA prediction. Our best model, that is, the random forest model trained on feature set C using information over the past 10 games, achieves the highest prediction accuracy of 67.98% for the latest 2020-21 and 2021-22 NBA season, indicating the features we propose are effective predictors for future NBA games.

9 CONCLUSION

This research discovers useful new features that can be used to make better predictions for NBA basketball games, including level of tiredness and the so-called “D” variables which emphasizes the difference of statistics between home and away team. We started by calculating Elo ratings, creating features based on information in the most recent N games that a team played, and designed the tiredness variables based on dates and locations of games. After that, we built up three feature sets and implemented feature selection using SFS and RFE algorithms with the aim of verifying the effectiveness of the new features and improving models’ overall performance. Next, 10-fold cross validation is utilized to train and validate different models for a more robust model selection result. Eventually, we picked 15 top models to conduct further hyperparameter tuning and evaluate their performances on the testing set containing 1036 games in the latest 2020-21 and 2021-22 season. Our best model, which is the random forest model trained on feature set C using information in the past 10 games gives the optimal accuracy of 67.98%, which is quite a decent performance since the highest accumulative winning rate of NBA teams in the regular season is below 70%. Core findings of this research are summed up as follows:

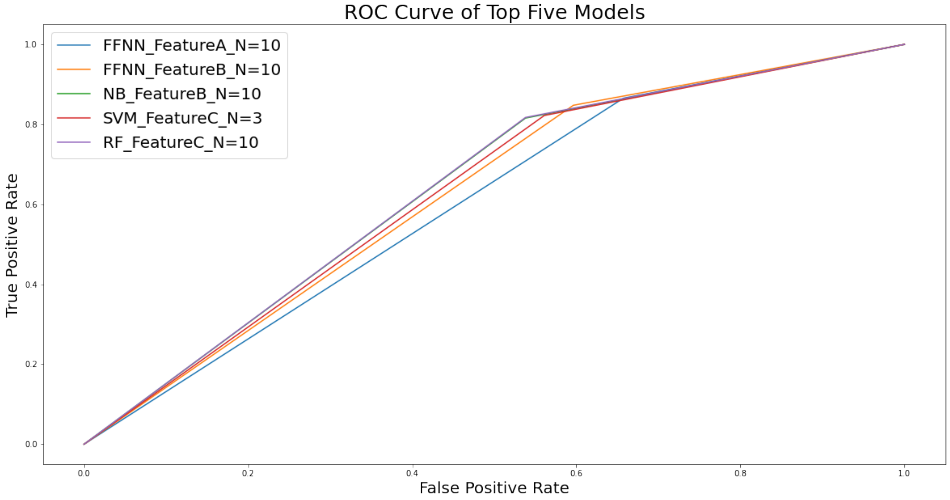

Figure 6: ROC Curve of the Top Five Models

(1) New features including level of tiredness and "D" features improve prediction accuracy, making our best models achieve accuracy that is very close to 68%.

(2) Average in-game statistics over the past 10 games achieves the best predictive performance.

(3) Feature selection improves model performance on training set by both enhancing mean accuracy and decreasing standard deviation, but worsen models' prediction accuracy when applied to new seasons.

(4) Random forest and naïve bayes models are the best models for NBA prediction among all the candidates in this article.

10 LIMITATIONS AND FUTURE WORK

Based on what we have completed in this research study, more improvements can be made. For example, more factors that potentially affect the outcome of games, both internal and external ones, can be taken into consideration. Since this study only considers team level statistics, other internal factors in future studies may include players' individual efficiencies based on player level statistics, such as the position that a player played and his abilities of shooting three-points field goals, rebounding, stealing, blocking and assisting. Besides, except for home court advantage and tiredness of teams, other external factors in future studies may also include injury, resting and return of players. Moreover, more instances can be collected to further improve prediction accuracy which may involve not only regular season games, but also playoffs and even pre-season training games. In addition, apart from the models utilized in this research, a wider collection of candidate models including both ML (machine learning) and DL (deep learning) models can be tested to identity the most effective ones for fair prediction of sports' outcomes, especially for basketball in future studies.

REFERENCES

[1] Cao C (2012) Sports data mining technology used in basketball outcome prediction. Dublin Institute of Technology.

[2] Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 6(1), 103-116.

[3] Nguyen, N. H., Nguyen, D. T. A., Ma, B., & Hu, J. (2021). The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. Journal of Information and Telecommunication, 1-19.

[4] Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. International Journal of forecasting, 26(3), 460-470.

[5] https://projects.fivethirtyeight.com/2018-nba-predictions/games/

[6] http://www.nbaminer.com/

[7] Ferrario, A. BASKETBALL ANALYTICS: THE USE OF DATA SCIENCE TO DESCRIBE AND PREDICT THE PERFORMANCE OF AN NBA TEAM.

[8] Chen, W. J., Jhou, M. J., Lee, T. S., & Lu, C. J. (2021). Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association. Entropy, 23(4),477

[9] Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. Journal of Quantitative Analysis in Sports, 5(1).

[10] Pai, P. F., ChangLiao, L. H., & Lin, K. P. (2017). Analyzing basketball games by a support vector machines with decision tree model. Neural Computing and Applications, 28(12), 4159-4167.