Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022)
IEEE Xplore Part Number: CFP22J88-ART; ISBN: 978-1-6654-8271-4

2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-8271-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECA55336.2022.10009313

Basketball Win Percentage Prediction using Ensemble-based Machine Learning

Dhruv Sikka[1]
Department of Networking and Communications,
School of Computing
College of Engineering and Technology
SRM Institute of Science and Technology, Kattankulathur-603203, India
ds7359@srmist.edu.in

Rajeswari D[2*]
Department of Data Science and Business Systems, School of Computing
SRM Institute of Science and Technology, Kattankulathur-603203, India.
drajiit@gmail.com
*Corresponding Author

Abstract — **Basketball is a prominent team sport played on a rectangular court between two teams of five players each. The goal is to shoot the ball through the defender's hoop, which is high on a backboard at each end of the court, while stopping the other team from shooting through its own hoop. Predicting how sports will turn out is an interesting sports analytics problem because it gives valuable data for running the sports market. In this work, data from NBA (National Basketball Association) basketball teams is analyzed, and a machine learning model is made that uses advanced statistical measures and performance data to evaluate and predict the winning percentage of different teams. Voting Regression is used to put together algorithms like Random Forest, Decision Trees, Linear Regression, Gaussian Regression, and Gradient Boosting into a single unit that can predict a team's win percentage over the course of a season with 93.3% accuracy.**

Index Terms — Sports outcomes prediction; Basketball game; Team win percentage prediction; Voting Ensembling; Random Forest Regression; Gaussian Process Regression; National Basketball Association

I. INTRODUCTION

In competitive basketball, the main goal is to win, and the sport changes to help achieve this goal. The goal of winning is the starting and finishing point, rendering it the game's primary objective and focus of endeavor. This work focuses on the NBA, a premier basketball league comprising 30 teams in the United States and Canada that has the most skilled basketball players in the world [1]. Sports analytics is the study of things like player performance, business processes, and recruiting that have to do with sports. Such information is useful to both athletes and businesses involved in sports [2]. As technology has improved over the past few years, it has become easier and easier to collect more and better data.

The way data is collected has also gotten better, which has helped sports analytics grow and led to progress in sophisticated statistics and machine learning [3]. By collecting and analyzing this data and analytics, players, coaches, and other staff can make better decisions before and during sports events. In the last few years, NBA teams have moved quickly to evaluate players using advanced metrics [4].

This is to take advantage of the benefits of analytics and give them an edge when making strategies. Sports betting is another sphere that has evolved due to sports analytics. Teams and businesses that let people bet on sports hire sports analysts who use data to guess what will happen in games. [5]

Adding machine learning to the current statistics can help solve the urgent need for faster and more accurate ways to analyze and predict how teams will do. Machine learning methods that can look for hidden connections in raw data are used in a lot of prediction models. To meet the need in the current situation, the work in this study tries to predict the win percentage based on how different teams have done throughout the season. It adds to the existing literature by combining advanced basketball statistics with machine learning algorithms.

The contributions of this study comprise the following:

- Usage of machine learning regression models with advanced basketball metrics.
- Prediction of the win percentage of teams based on various metrics like offensive and defensive ratings.

- An ensemble model consisting of five baseline learners was trained to predict with the utmost accuracy with minimal error.
- The strength of the team's schedule is taken as a feature for modeling.

II. LITERATURE

Jasmin A. Caliwag and Maria Christina presented work on predicting the possible winner of a specific game in a season using the cascading algorithm and machine learning techniques with a computed Four Fact Analysis as input [6]. Ilker Ali Ozkan introduced work on neural networks and basketball match prediction [7]. The study used the concurrent neuro-fuzzy system (CNFS) and artificial neural networks (ANN) to make predictions that were much more accurate than those of its predecessors. Wei-Jen Chen and Mao-Jhen Jhou talked about their research on predicting the outcome of a single game using a two-stage XGBoost model and specially designed features. and Statistical measures are inputs [8]. The study's excellent use of the quality and number of features opens up a lot of possibilities for feature engineering and shows how data mining can be used to improve the accuracy of predictions. There is a lot of uncertainty in predicting basketball games. Meihong Chen and Fuzhi Su decided to use a fuzzy logic system to make this kind of prediction [9]. This work uses both artificial intelligence and fuzzy logic theory to deal with the uncertain and variable nature of the event. Hence, it helps the prediction method be more sturdy and reliant.

These studies show that machine learning algorithms can classify data and statistical metrics in a useful way. Still, there isn't a clear way to figure out how often a team wins over the course of a whole season, not just one game. Consequently, the purpose of this endeavor is to develop a means to do so.

III. PROPOSED METHODOLOGY

For the model's training and further analysis, this work uses the NBA regular season data from 2016/17 to 2020/21 with advanced statistics and performance averages of 30 teams, which are available on the league's website [10].

A. Dataset Preparation

This study's data set is made up of different advanced statistical metrics used in basketball to measure how well a team does as a whole over the course of a regular season. The advanced metrics data of each team from 2016/17 to 2020/21 is taken from the NBA league's official website [10]. The metrics are as follows:

- **Points per game:**

Points per game (PPG) is the average amount of points earned by a team across a set of games or a full season. The total of points by the matches.

$$\frac{Total\ Points\ scored\ by\ a\ Team/Individual}{Total\ Number\ of\ games\ played} \quad (1)$$

- **Points allowed per game:**

The average amount of points conceded to the opposition each game, set of games, or season. It is calculated by dividing the number of points awarded to the opposing side by the number of games.

$$\frac{Total\ Points\ scored\ by\ the\ opponent\ \ Team/Individual}{Total\ Number\ of\ games\ played} \quad (2)$$

- **Average Points Differential:**

A Point Differential is the average differential between points scored and allowed by a team during a series of matches. This metric is computed using the following formula:

$$\frac{Total\ Points\ scored - Total\ points\ allowed}{Total\ Number\ of\ games\ played} \quad (3)$$

- **Offensive Efficiency**

An offensive proficiency rating, also called productive offensive efficiency, is a statistic used in basketball to estimate a team's offensive production at key points in the offense as a single metric.

- **Defensive Efficiency**

In basketball, the defensive rating or defensive efficiency of a team is used as a single metric to measure how well the team keeps the other team from scoring points.

- **Efficiency Differential**

The average numerical difference between a team's offensive and defensive efficiency throughout a season. This metric is computed using the following formula:

$$\frac{Total\ Offensive\ Efficiency\ -\ Total\ Defensive\ Efficiency}{Total\ Number\ of\ games\ played} \quad (4)$$

- **Strength of the Schedule (SOS)**

The Strength of Schedule (SOS) shows how hard each team's schedule was on average throughout the season. This metric considers the opponent's ranking and whether the game is played at home or away.

The machine learning models would take these metrics for training, and the win percentage of a team's season is predicted using regression-based models.

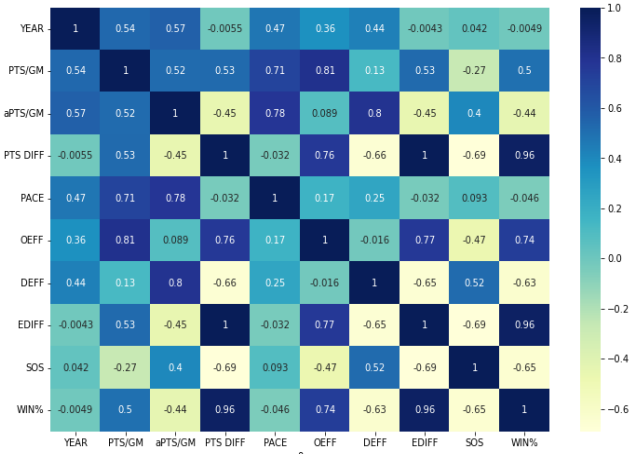

Fig 1. The above figure shows the correlation between different features of the dataset.

B. Machine Learning Model Preparation

In recent years, machine learning techniques, including classical multiple linear regression, decision trees, gradient boosting, and Gaussian Process Regression, have been often used to deal with regression problems, showing great prospects in win prediction activity [11,12]. Multiple regression models are put together and controlled by a voting ensemble model to make a strong and accurate prediction model for the same.

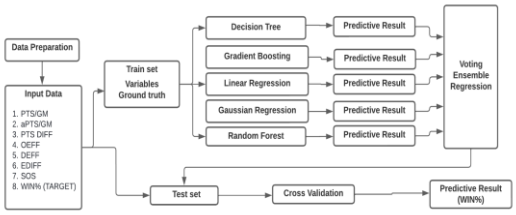

Fig 2. A flowchart of the methodology of model development

A. Multiple Linear Regression

Multiple linear regression (MLR), also called "multiple regression," employs multiple factors to explain and predict the behavior of a target variable. Two continuous variables, one of which is independent and the other dependent, are required for linear regression [13].

Algorithm 1: MLR Model
Result: Prediction of the win percentage of a team.

lm = LR()
lm.fit(P_train, q_train)
lm_cv = KFold(n_spl=10, r_s=1, shuffle=True)

Output: MLR Prediction model

B. Decision Tree Regression

A decision tree employs a tree-like structure to construct regression or classification techniques. It splits a dataset into progressively smaller subgroups while constructing a tree structure for every subset. The final output is a tree composed of decision and leaf nodes.

Algorithm 2: The Decision Tree Model
Result: Prediction of the win percentage of a team.

dt = DTR(r_s = 0)
dt.fit(P_train, q_train)
dt_cv = KFold(n_spl=5, r_s=1, shuffle=True)

Output: Decision Tree Prediction model.

C. Random Forest Regression

Random Forest Regression is a supervised learning algorithm for regression that uses a method called "ensemble learning." A random forest is an ensemble of decision trees [14].

Algorithm 3: The Random Forest Model
Result: Prediction of the win percentage of a team.

rf = RFR(n_estimators = 100, r_s = 0)
rf.fit(P_train, q_train)
rf_cv = KFold(n_spl=10, r_s=1, shuffle=True)

Output: Random Forest Prediction model.

D. Gradient Boosting Regression

The difference between the current prediction and the known accurate target value is computed through gradient-boosting regression. The discrepancy is referred to as the

residual. Gradient boosting regression is then used to train a weak model that maps features to the residual.

Algorithm 4: Gradient Boosting Model
Result: Prediction of the win percentage of a team.

```
gdr = GBR(r_s=0)
gdr.fit(P_train, q_train)
gdr_cv = KFold(n_spl=10, r_s=1, shuffle=True)
```

Output: Gradient Boosting Prediction model

E. Gaussian Process Regression

Gaussian Processes (GP) is an approach of supervised learning developed to handle regression and stochastic classification issues. GPR offers a number of advantages, including its capacity to give uncertainty measures for predictions and its ability to operate effectively with tiny datasets.

Algorithm 5: The Gaussian Regression Model
Result: Prediction of the win percentage of a team.

```
Gaus = gp.GPR(kernel=kernel, n_restarts_optimizer=10, alpha=0.1, normalize_y=True)
gaus.fit(P_train, q_train)
gaus_cv = KFold(n_spl=10, r_s=1, shuffle=True)
```

Output: Gaussian Regression model

F. Voting Ensembling Regression

A voting ensemble is an ensemble model that aggregates the forecasts of many models. It is a method for enhancing the effectiveness of the algorithm. Idealistically, the efficiency of the aggregate will be greater than any single model. This entails calculating the average of what the algorithms estimate would emerge in the scope of regression [15].

Algorithm 6: Voting Ensembling Model
Result: Prediction of the win percentage of a team.

```
final_model = VR(estimators=[('lm', lm), ('rf', rf), ('dt', dt), ('gdr', gdr), ('gaus',gaus)])
final_model.fit(P_train, q_train)
cv = KFold(n_spl=5, r_s=1, shuffle=True)
```

Output: Voting Ensembling model

In the above algorithms n_spl is the number of splits, r_s is the random state.

IV. EXPERIMENT RESULTS AND SCOPE

Using statistical measuring instruments, the model's effectiveness was computed. The prediction accuracy was calculated using the determination coefficient (R2) and the root mean square error (RMSE). R2 score is a useful metric of data fit, whereas the RMSE is an accurate measure of data fit. The greater the model's utility, the lower the RMSE. R2 spans between 0 and 1, with a higher value (closer to 1) indicating a more suited model.

In order to get the best results, a series of experiments with multiple different combinations of regression methods in the voting ensembling process. The results of the combinational experiments are given below.

Experiment 1, we consider possible combinations of two voting regression models. The results are as follows in Table 1.

Table 1: Model Combinational Results for Experiment 1

EXP NO.	MODELS USED	R^2	RMSE
EXP1	Linear Regression Random Forest	0.9298	0.0373
	Linear Regression Decision Trees	0.9262	0.0382
	Linear Regression Gradient Boosting	0.926	0.0382
	Linear Regression Gaussian Process	0.9234	0.0389
	Random Forrest Decision Trees	0.9266	0.0382
	Random Forrest Gradient Boosting	0.9254	0.0385
	Random Forrest Gaussian Process	0.9332	0.0365
	Decision Trees Gradient Boosting	0.924	0.0389
	Decision Trees Gaussian Process	0.9307	0.0371

	Gradient Boosting Gaussian Process	0.9292	0.0375

For Experiment 2, we take all possible combinations of three voting regression models. The results are as follows in Table 2.

Table 2: Model Combinational Results for Experiment 2

EXP NO.	MODELS USED	R^2	RMSE
EXP2	Linear Regression Random Forest Decision Trees	0.9318	0.0368
	Linear Regression Random Forest Gradient Boosting	0.9297	0.0373
	Linear Regression Random Forest Gaussian Process	0.93	0.037
	Linear Regression Decision Trees Gradient Boosting	0.9302	0.0372
	Linear Regression Decision Trees Gaussian Process	0.9308	0.037
	Linear Regression Gradient Boosting Gaussian Process	0.9287	0.0376
	Random Forest Decision Trees Gradient Boosting	0.9282	0.0378
	Random Forest Gradient Boosting Gaussian Process	0.9308	0.0371
	Decision Trees Gradient Boosting Gaussian Process	0.9318	0.0368
	Random Forest Decision Trees Gaussian Process	0.9334	0.0364

For Experiment 3, we consider possible combinations of four voting regression models. The results are as follows in Table 3.

Table 3: Model Combinational Results for Experiment 3

EXP NO.	MODELS USED	R^2	RMSE
EXP3	Linear Regression Random Forest Decision Trees Gradient Boosting	0.9318	0.0368
	Linear Regression Random Forest Decision Trees Gaussian Process	0.9334	0.0364
	Linear Regression Random Forest Gradient Boosting Gaussian Process	0.9311	0.037
	Linear Regression Decision Trees Gradient Boosting Gaussian Process	0.9323	0.0367
	Random Forest Decision Trees Gradient Boosting Gaussian Process	0.9325	0.0367

And finally, out of the five curated models for prediction, for Experiment 4, we take the final combination of all five models for the voting regressor. The results are as follows in Table 4.

Table 4: Model Combinational Results for Experiment 4

EXP NO.	MODELS USED	R^2	RMSE
EXP4	Linear Regression Random Forest Decision Trees Gradient Boosting Gaussian Process	0.9332	0.0364

It can be seen that Experiment No. 4 with 5 regression models (Decision Trees, Linear Regression, Gradient Boosting, Random Forest, and Gaussian Process) gives the best and most accurate results and should be used as the final model to predict win percentage which is presented in table 5.

Table 5: Model assessment for the final selected models

MODEL USED	R^2	RMSE
5-Algorithm Voting Regressor Ensemble	0.9332	0.0364

Model		
Linear Regression	0.905	0.0418
Random Forest	0.9256	0.0373
Decision Trees	0.9043	0.0436
Gradient Boosting	0.9173	0.0405
Gaussian Process	0.9282	0.0378

By plotting predicted values against actual win percentages, it can be seen how good the final 5-Algorithm Voting Regressor Ensemble Model is. This helps us understand the variability spread of predicted values and gives a deeper look into the model's accuracy. The scatter plot can be observed as follows:

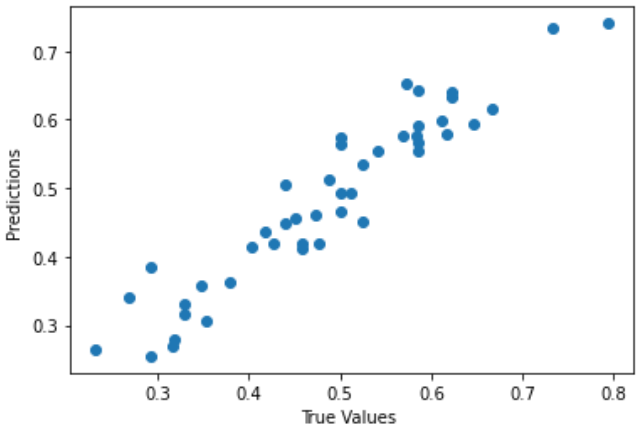

Fig 3. The above figure shows the variation between predicted and true values of win percentages.

V. CONCLUSION

This study showed a single ensemble prediction unit made up of five machine learning methods that can predict how likely a team is to win over the course of a whole season.

There are seven statistical features that are known to be important: average points per game, average points allowed per game, average points difference, offensive efficiency, defensive efficiency, efficiency differential, and schedule strength (SOS). The prediction scheme for these 5 single units is put together using voting ensemble regression. With an R2 score of 0.9332 and an RMSE value of 0.0364, the ensemble model comprising all 5 models performed the best out of all competing models at generating predictions. Future research should investigate how the recommended basketball win percentage prediction system works with more NBA seasons given that just five seasons of NBA data were used in this study. Also, using data from many seasons to come up with more reliable and useful findings about feature selection could be a topic of study in the future.

REFERENCES

[1] Torres, R. A. (2013). Prediction of NBA games based on Machine Learning, 2. Retrieved from http://homepages.cae.wisc.edu/~ece539/fall13/project/AmorimTorres_pro.pdf

[2] Morgulev, E.; Azar, O.H.; Lidor, R. (2018) Sports Analytics and the Big-Data Era. Int. J. Data Sci. Anal. 5, 213–222.

[3] Richardson L. (2015) How Predictable is NBA?, Nylon Calculus Retrieved from nyloncalculus.com/2015/02/02/predictable-nba/

[4] Cao C. (2012). Sports Data Mining Technology Used in Basketball Outcome Prediction. Retrieved from http://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis

[5] Puranmalka,K. (2012). Modelling the NBA to Make Better Predictions. Retrieved from http://hdl.handle.net/1721.1/85464

[6] Jasmin A. Caliwag, Maria Christina R. Aragon, Reynaldo E. Castillo, and Ellizer Mikko S. Colantes. 2018. Predicting Basketball Results Using Cascading Algorithms. In Proceedings of the 2018 International Conference on Information Science and Systems (ICISS'18). Association for Computing Machinery, New York, NY, USA, 64–68. https://doi.org/10.1145/3209914.3209921

[7] Ilker Ali Ozkan (2020) A Novel Basketball Result Prediction Model Using a Concurrent Neuro-Fuzzy System, Applied Artificial Intelligence, 34:13, 1038-1054, DOI: 10.1080/08839514.2020.1804229

[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. https://doi.org/10.3390/e23040477

[9] Chen, M., Su, F. A (2022) basketball game prediction system based on artificial intelligence. J Supercomput 78, 12528–12552. https://doi.org/10.1007/s11227-022-04375-w

[10] https://www.nba.com/stats/teams/advanced/

[11] Rohit, R.V.S., Chandrawat, D., Rajeswari, D. (2021). Smart Farming Techniques for New Farmers Using Machine Learning. In: Mahapatra, R.P., Panigrahi, B.K., Kaushik, B.K., Roy, S. (eds) Proceedings of 6th International Conference on Recent Trends in Computing. Lecture Notes in Networks and Systems, vol 177. Springer, Singapore. https://doi.org/10.1007/978-981-33-4501-0_20

[12] H. Agarwal, A. Singh and R. D, "Deepfake Detection Using SVM," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1245-1249, doi: 10.1109/ICESC51422.2021.9532627.

[13] Gülden Kaya Uyanık, Neşe Güler, A Study on Multiple Linear Regression Analysis, Procedia - Social and Behavioral Sciences, Volume 106, 2013, Pages 234-240, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2013.12.027.

[14] H. Soni, P. Arora and D. Rajeswari, (2020) "Malicious Application Detection in Android using Machine Learning," 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 0846-0848, doi: 10.1109/ICCSP48568.2020.9182170.

[15] F. Leon, S. -A. Floria and C. Bădică, "Evaluating the effect of voting methods on ensemble-based classification," 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2017, pp. 1-6, doi: 10.1109/INISTA.2017.8001122.