Analysis of the Principle and Models for FIFA World Cup Prediction

Haoyang Tian*

School of arts and sciences, Rutgers the State University of New Jersey, New Brunswick, United States

*Corresponding author: ht363@scarletmail.rutgers.edu

Abstract. As a matter of fact, the Word Cup predictions attracts lots of scholars on account of its complex features. In recent years, various machine learning models are proposed to realize the accurate predictions. With this in mind, in this study, the author will compare the advantages and disadvantages of five models, i.e., logistic regression, decision tree, random forest, neural network, and support vector machine, in predicting World Cup matches. The author used the results of past World Cup matches as the original dataset. After research, the author concluded that the logistic regression model was the most effective. While Neural Networks and Support Vector Machines followed closely in predictive accuracy, they also showed promising potential. Decision tree and the random forest models suffer from severe overfitting, and the author believe that these two models are difficult to apply in this field. Overall, these results shed light on guiding further exploration or World Cup prediction.

Keywords: World Cup; accuracy; overfitting; influencing factors.

1. Introduction

The FIFA World Cup stands as the world's most viewed and popular sports event. The 2022 World Cup in Qatar was watched by more than 5 billion people, 1.5 billion watched the final, and had more than 6 billion interactions on social media [1]. Many enthusiasts, companies and researchers have proposed a number of methods for predicting the results of matches in the World Cup, incorporating machine learning among the methods. This post will summarize and compare five types of models for predicting the World Cup through machine learning using data from past matches.

Predicting the outcome of World Cup matches has far-reaching implications. For sports fans, it adds an element of excitement and engagement. For analysts, accurate predictions offer insights into team and player performance [2, 3]. For bookmakers, precise forecasts are directly linked to revenue. While traditional methods like expert analysis still hold value, computational models offer a scalable and objective approach to prediction.

The main objective of this research paper is to compare and evaluate five commonly used World Cup forecasting models: logistic regression, decision trees, random forests, neural networks and support vector machines [4, 5]. The paper will explore the data collection, feature selection and model training methods required to use these five models. The paper will then delve into the complexities of each model, present its performance metrics and discuss its strengths and weaknesses. Finally, a comparative analysis will synthesize the results of these studies and suggest possible directions for improving the models in the future.

The paper is structured as follows. The Sec. 2 will describe the collection of data, the process of digitizing the results of previous World Cup matches, and the pre-processing of the selected dataset. The Sec. 3 describes each of the five models in detail, describes the training process, the resulting problems and the corresponding improvements, and shows the results of the training through graphs and charts. The advantages and disadvantages of each model in the field of World Cup prediction are summarized. The Sec. 4 describes the limitations of the current research in the field of World Cup prediction, and suggests some possible directions and methods for improvement in future research. The Sec. 5 will briefly summarize the conclusions of the article.

2. Data and Method

For the purposes of this paper, the dataset uses data from the World Cups of all years that are publicly available on Kaggle (https://www.kaggle.com/datasets/abecklas/fifa-world-cup). This dataset is presented in CSV format and contains information on all matches played in previous World Cups. After removing non-essential information, the author retained the name of the home team, the name of the away team, the score of the home team, and the score of the away team for this study.. The representation of the win/loss item in the original dataset was not uniform, so it was necessary to set up a column result_n, where 1 represents the home team's victory, 0 represents the away team's victory, and -1 represents a draw.

To enhance the accuracy of the forecasts, this study introduced additional parameters for the matches between each two countries, including the number of home team entries, the number of away team entries [6, 7], the number of home team victories, the number of away team victories, the number of goals scored by the home team, the number of goals scored by the away team, the home team's winning percentage, the away team's winning percentage, the home team's average goals per game, and the away team's average goals per game. In this study, min-max processing with standardized score processing was used for all data except home team, away team, win/loss. Where the formula for min-max processing is X_normalized = (X - X_min) / (X_max - X_min), which aims to scale the data to the range of [0, 1] and eliminate the difference in magnitude. The standard score processing formula is $z = (x - \mu) / \sigma$, which aims to convert the data into a standard normal distribution, having a mean value of 0 and a standard deviation of 1.

3. Results and Discussion

Logistic regression is a classic classification algorithm dating back to the population growth model developed by Pierre François Verhulst in the 19th century, but its modern form was developed mainly in the mid-20th century by Joseph Berkson and David Cox with other researchers [8]. Logistic regression is a special case of generalized linear models with binary outcomes. It is particularly suited to problems where the dependent variable is a categorical variable, including predicting the results of the FIFA World Cup matches. In this study, one defines the dependent variable as the "win" or "loss" of a particular team. Unlike more sophisticated machine learning algorithms, logistic regression balances interpretability and accuracy, enabling analysts to understand the significance and impact of individual characteristics such as team rankings, player statistics and past performance. This study used Python's own logistic regression function. The learning curve is shown in Fig. 1.

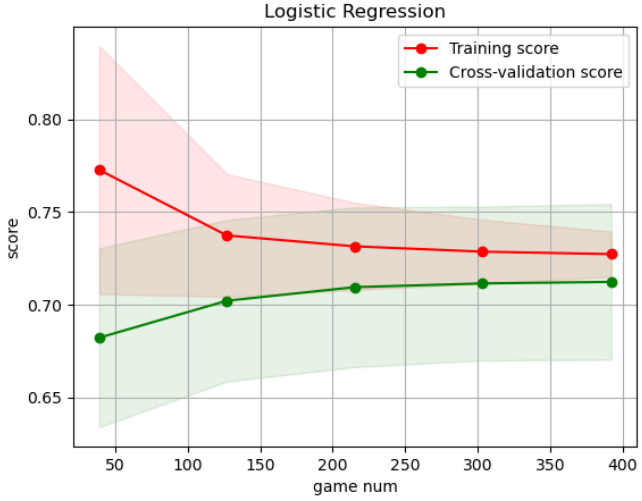

Fig. 1 Score with different game num of logistic regression.

The Accuracy of the training set is 0.727, Accuracy of the test set is 0.78, Mean Absolute Error is 0.21904761904761905, Accuracy is 0.780952380952381, Recall is 0.780952380952381 and score is 0.780952380952381. According to the Fig. 1, the learning curve converges as the number of games

approaches 400 and does not show significant under- and over-fitting. Therefore, this study concludes that the logistic regression model has relatively superior performance in World Cup match prediction.

The concept of decision trees in machine learning has its roots in operations research and decision theory, dating back to the 1960s. However, it was not until the advent of algorithms such as ID3 in the 1980s that decision trees gained prominence as a generalized machine learning tool [8]. In a decision tree, data is divided into two or more homogeneous sets based on the most important attributes at each level, making it easier to analyze. When predicting the outcome of a World Cup match, the leaves of a decision tree will correspond to the "win" or "loss" of a particular team, similar to the way dependent variables are defined in logistic regression. Decision trees are uniquely concise and interpretable, allowing the user to understand how the model arrived at a particular decision. This helps to analyze detailed factors such as home/away that affect the outcome of a game. Decision trees also model complex non-linear relationships between variables, helping to simulate the high level of uncertainty in World Cup matches. However, decision trees can easily become overly complex, capturing noise in the data and thus reducing generalizability. This study uses the decision tree algorithm that comes with Python, and the learning curve is shown in Fig. 2.

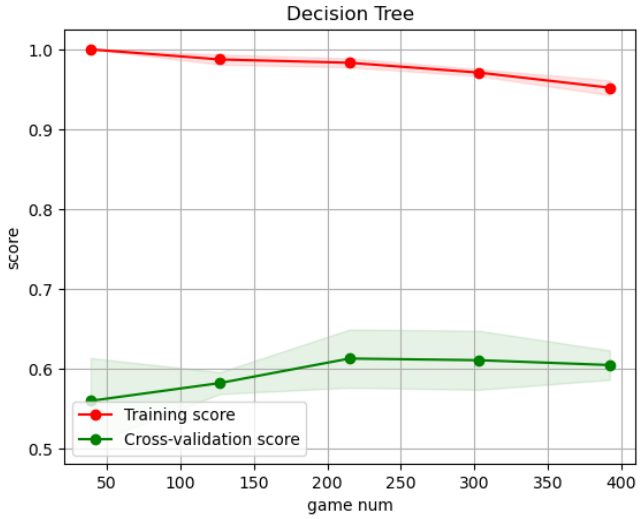

Fig. 2 Score with different game num of Decision tree.

The Accuracy of the training set is 0.943, Accuracy of the test set is 0.643, Mean Absolute Error is 0.35714285714285715, Accuracy is 0.6428571428571429, Recall is 0.6428571428571429 and score is 0.6428571428571429. According to the results, the training set has a high accuracy of 0.943 when the number of matches is close to 400, however, the test set accuracy is only 0.643. The decision tree model produces severe overfitting. This means that the generalization ability to new data is poor, therefore this study concluded that the decision tree model is not the most suitable model for predicting the outcome of World Cup matches.

The random forest model was proposed by statistician Leo Breiman in 2001 to improve on the limitations of a single decision tree, primarily its tendency to overfit when working with complex data [9]. In this model, each tree is built from a random sample of the training data, and the splitting of each node is performed by a random subset of features, thus making the model robust and generalizable. The algorithm's ensemble approach allows it to make more balanced and nuanced predictions by aggregating outputs from multiple decision trees. The disadvantage of random forests is that they are overly complex and prone to overfitting. Also, due to their ensemble characteristic, random forests are harder to interpret compared to singular decision trees or logistic regression models. In this study, Python's built-in Random Forest algorithm is used, and the learning curve is shown in Fig. 3

The Accuracy of the training set is 0.943, Accuracy of the test set is 0.729, Mean Absolute Error is 0.2714285714285714, Accuracy is 0.7285714285714285, Recall is 0.7285714285714285 and score is 0.7285714285714285. The Random Forest algorithm produces an excessively high training set accuracy of 0.943, which is the same as the Decision Tree algorithm. The test set accuracy of

0.729 is higher than the 0.643 under the Decision Tree model, but it is still far from the training set. The overfitting phenomenon of the Random Forest algorithm affects its predictive ability, but it is stronger than the Decision Tree model in terms of accuracy, recall, and F-score. Therefore, this study concludes that Random Forest is not the most suitable model for predicting the result of World Cup matches, but its performance is generally better than that of Decision Tree model.

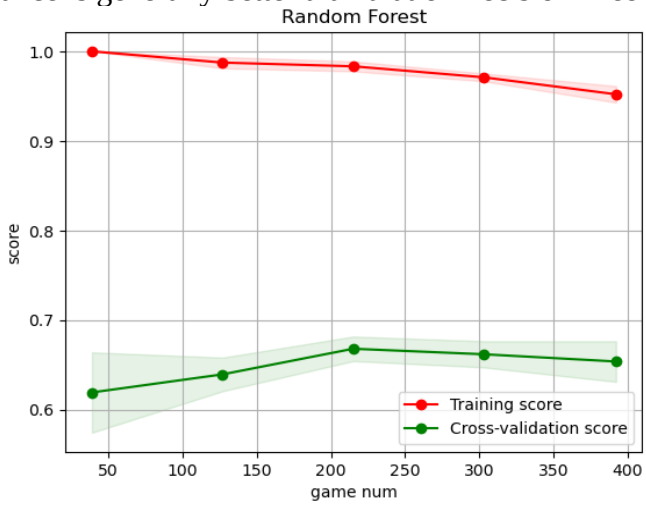

Fig. 3 Score with different game num of random forest.

Drawing inspiration from the biological neural networks present in animal brains, the idea behind Artificial Neural Networks emerged and was established as far back as the 1940s [10]. However, it was not until Alexey Grigorevich Ivakhnenko invented the first deep learning multilayer perceptron in the 1960s that research in this area began to develop rapidly. Over time, progress in deep learning has amplified the potential of ANNs. In the context of predicting World Cup matches, neural networks offer a highly flexible and powerful modeling approach. Unlike traditional statistical methods or even simpler machine learning algorithms, neural networks can learn intricate patterns from large volumes of data, offering the potential for higher accuracy. Also, Their ability to integrate and learn from diverse types of data, including time-series data of past matches, player statistics, and even unstructured data like text from news reports, provides a comprehensive approach to prediction that is hard to match with simpler algorithms. because of the limited range of variables considered in this study's data, the findings might not fully capture the capabilities of artificial neural networks. A drawback of artificial neural networks is that researchers cannot view or understand the inner workings of the algorithms, which is known as a "black box". This characteristic makes it challenging to understand how individual features affect predictions. In this study, Python's built-in Neural Networks algorithm is used, and the learning curve is shown in Fig. 4.

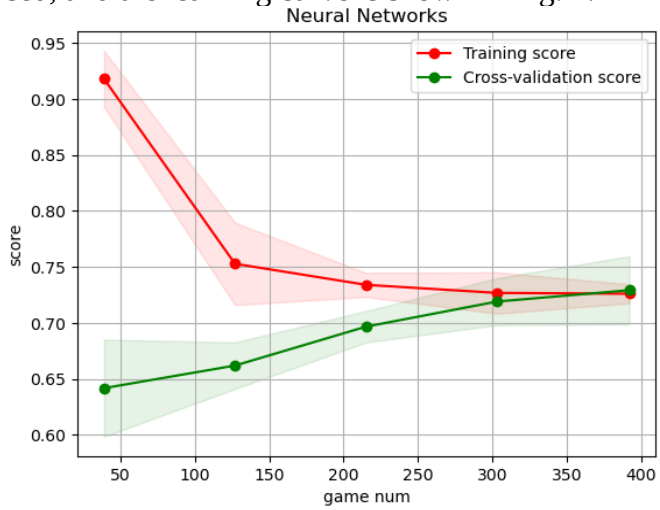

Fig. 4 Score with different game num of neural network.

The Training set accuracy is 0.719, Accuracy of the test set is 0.757, Mean Absolute Error is 0.24285714285714285, Accuracy is 0.7571428571428571, Recall is 0.7571428571428571 and score is 0.7571428571428571. According to the results, there's a minimal discrepancy in accuracy between the training and test sets, with both exceeding 0.71. Meanwhile, the accuracy, recall, and F-score are all higher than 0.75, and also higher than the Random Forest model and the Decision Tree model. Therefore, the study recognizes that the neural network model is suitable for predicting World Cup matches and outperforms the random forest and decision tree models, and slightly underperforms the logistic regression model. If more complex data are used for prediction, the performance of the neural network model will enhance further.

Initiated by Vladimir Vapnik and Corinna Cortes in the early 1990s, the Support Vector Machine algorithm quickly gained recognition as a powerful tool for classification and regression tasks [11]. Built on the foundations of statistical learning theory, SVM seeks to identify the optimal hyperplane that distinguishes data points from varying classes within a high-dimensional realm, rendering it especially proficient for scenarios with intricate decision boundaries. Similar to artificial neural networks, the full potential of the support vector machine model may not be fully realized due to the limited influencing factors included in the dataset used in this study. Support vector machines are also typically a "black box" in that their predictions are often difficult to analyze, and it is difficult for researchers to determine the impact of specific factors through adjustment. In this study, Python's built-in Support Vector Machines algorithm is used, and the learning curve is shown in Fig. 5.

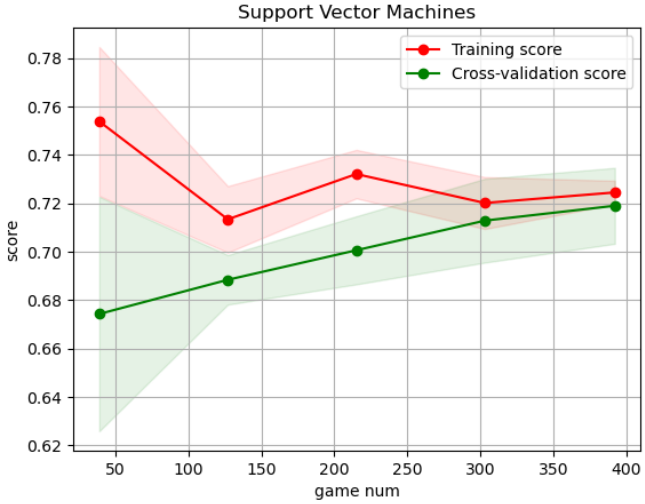

Fig. 5 Score with different game num of SVM.

The Accuracy of the training set is 0.727, Accuracy of the test set is 0.757, Mean Absolute Error is 0.24285714285714285, Accuracy is 0.7571428571428571, Recall is 0.7571428571428571 and score is 0.7571428571428571. According to the results, the training and test set accuracies converge when the number of matches reaches 400, and both accuracies show an increasing trend. The difference between the two in the support vector machine is the smallest among all five models. Therefore, this study concludes that Support Vector Machine can be considered as an excellent World Cup prediction model. It outperforms the decision tree model and the random forest model, slightly outperforms the artificial neural network model, and slightly underperforms the logistic regression model. Assuming a more complex dataset is used, then the support vector machine model performance will be further improved.

Among the five models, the logistic regression model has the highest F-score. The accuracy gap between its training and test sets is also very small, making it a very suitable model for predicting World Cup matches. Both the decision tree model and the random forest model produced severe overfitting problems and therefore are difficult to use as models for predicting World Cup matches. The neural network model and the support vector machine model perform well overall, outperforming the decision tree model and the random forest model, and slightly underperforming the logistic regression model, but they have great potential in more complex datasets.

4. Limitations and Prospects

In this study, a dataset with a very large total number of matches but fewer influencing factors was chosen to produce an intuitive comparison. This fact prevented the two models, Neural Network and Support Vector Machine, from reaching their full potential. Also, this study only compared the basic algorithms that come with Python and did not make targeted modifications to the individual models. Additionally, this study did not include prediction methods using non-competition data due to the different datasets required and the inability to make direct comparisons. Some of these methods can produce equally good predictions. A common method is to use odds set by sportsbooks as a dataset. The sportsbooks set odds based on people's predictions of the outcome of the game, and by combining the odds from multiple bookies, it is possible to reverse the prediction of the outcome of the game. Another common method is to capture keywords in major social media and forums and count the number of times and frequency of specific keyword combinations. Based on the tendency of a large number of users to speak, it is possible to derive an estimate of the outcome of the match.

There is still much that can be improved and further researched in the future for this study, as well as for the research area of using machine learning to predict the World Cup. Incorporating more variables so that the dataset responds to more aspects of the tournament is an important way to enhance prediction accuracy. A World Cup soccer match can be divided into multiple parts such as runs, passes, shots, goalkeeping, fouls, penalties, and so on. Future research could turn a team's performance in these areas into computable data to be added to the dataset. A soccer team can also be divided into the 11 players playing on the field, the substitutes and the coach. Datamining the performance of these individuals allows machine learning to be performed on individual players rather than teams, which can help to better simulate game situations while weakening the errors associated with changes in personnel and better predicting the outcome of matches.

5. Conclusion

To sum up, this study compares the performance of five common machine learning models when predicting World Cup matches. This study recommends the use of logistic regression model. Its accuracy, recall and F-score are higher. If the original dataset contains more influencing factors, the use of artificial neural networks and support vector machines can also achieve good prediction results. Decision tree and random forest models are prone to severe overfitting in prediction, so they are not recommended by this study.

References

[1] One Month On: 5 billion engaged with the FIFA World Cup Qatar 2022". FIFA. Retrieved from: https://www.fifa.com/tournaments/mens/worldcup/qatar2022/news/one-month-on-5-billion-engaged-with-the-fifa-world-cup-qatar-2022-tm (January 18 2023)

[2] Casella G, Berger R L. Statistical Inference Duxbury Press. Pacific Grove, 2002.

[3] Cramer J S. The origins of logistic regression. Springer, 2002.

[4] Quinlan J R. Induction of decision trees. Machine learning, 1986, 1: 81-106.

[5] Breiman L. Random forests. Machine learning, 2001, 45: 5-32.

[6] Eberhart R C, Dobbins R W. Early neural network development history: the age of Camelot. IEEE Engineering in Medicine and Biology Magazine, 1990, 9(3): 15-18.

[7] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks, 2015, 61: 85-117.

[8] Cortes C, Vapnik V. Support-vector networks. Machine learning, 1995, 20: 273-297.

[9] Leitner C, Zeileis A, Hornik K. Forecasting the winner of the FIFA World Cup 2010. Machine Learning, 2010

[10] Leitner C, Zeileis A, Hornik K. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. International Journal of Forecasting, 2010, 26(3): 471-481.

[11] Radosavljevic V, Grbovic M, Djuric N, et al. Large-scale World Cup 2014 outcome prediction based on Tumblr posts. KDD workshop on large-scale sports analytics. 2014.