Prediction on Olympic Medal Based on Random Forest and Logistic Regression

Yuepeng Li [1,*], Zekai Cui [2], Yanqing Zhou [2]

[1] School of Aeronautics, Northwestern Polytechnical University, Xi'an, China, 710072

[2] Honors College, Northwestern Polytechnical University, Xi'an, China, 710072

* Corresponding Author Email: liyuepeng@mail.nwpu.edu.cn

Abstract. The Olympic Games, as a highly renowned global sports event, attract significant attention, and the Olympic medal table is a focal point of public interest. The medal table not only showcases athletes' exceptional competitive abilities but also reflects a nation's overall strength. In this paper, a Random Forest model was developed to predict the medal rankings for the 2028 Olympic Games. The model innovatively considers not only economic factors but also athlete performance, the number of events participated in, and host nation advantages. By integrating these comprehensive factors, it forecasts that the United States, China, and the United Kingdom will secure the top three positions. Additionally, the study focuses on countries that have never won Olympic medals—groups that receive little attention—using Logistic Regression to predict that nations like Tuvalu have a 3.530% probability of gaining their first Olympic medals in 2028. This study not only provides forecasts for the Olympic medal table but also offers a scientific basis for countries to develop sports strategies and optimize resource allocation, which is significant for enhancing international sports.

Keywords: Olympic Medal, Random Forest, Logistic Regression, Prediction Model.

1. Introduction

The Olympic Games are an international sports festival, and the period of organizing these games is 4 years. The games are further held under two main banners, which are referred to as the summer and the winter games, with a gap of two years within the four-year cycle [1]. It contains various sports and is the focus of attention for all countries worldwide, as it is regarded as a symbol of national identity [2]. Predicting competition results is an important basis for sports organizations when making strategic decisions to allocate resources. It helps sports organizations with the increasing density of competition, as sports have a higher probability of success than others [3]. Therefore, the prediction of Olympic medals is not only helpful for sports management departments of various countries to formulate more precise development strategies but also provides data support for international sports exchanges and cooperation.

Peters et al. established Poisson regression, Random Forest, Gradient Boosting and XGBoosting models, etc., and found that population, GDP per capita, hosting the Olympics, autocratic regimes and healthcare expenditures positively influenced Olympic medal counts, while income inequality and food supply had negative effects, with XGBoosting having the highest predictive accuracy [4]. Raiyan Sayeed et al. used 13 machine learning models, including Logistic Regression and XGBoost, and found that ensemble models like XGBoost, LightGBM, and Gradient Boosting performed better, with accuracy rates of 83%, 84%, and 84%, respectively [5]. Maheswari Raj et al. employed Linear Regression, Poisson Regression, and Negative Binomial Regression models, and determined that the Log-transformed Linear Regression model performed best, with GDP and population as key predictors of medal counts [6]. Sanchez-Fernandez and Vaamonde-Liste used Linear Regression, Poisson Regression, Random Forest, and other models to predict medal counts for the Rio 2016 Olympics, finding that the Linear Regression model had the highest correlation with actual results [7]. Parveen Badoni et al. compared Decision Tree and Random Forest models, showing that Random Forest was more robust to overfitting and had higher accuracy in predicting medal counts, with athlete participation and age being important factors [8]. In addition, the Random Forest and Logistic

Regression models are widely used in sports performance prediction, demonstrating good predictive capabilities [9-12].

Previous studies have successfully predicted Olympic medal counts, but these studies often focus mostly on economic variables and lack attention to countries that have never won Olympic medals. To address these limitations, this paper comprehensively considers athlete performance, the number of events participated in, and the host-country effects of building a Random Forest model for predicting Olympic medal tables. For countries that have never won Olympic medals, a Logistic Regression model is established to predict their probability of winning their first medal.

2. Model

2.1. Random Forest

Random Forest is a machine learning algorithm based on ensemble learning, primarily used for classification and regression tasks. It improves model accuracy and robustness by constructing multiple decision trees and synthesizing their prediction results. The advantages of Random Forest include its ability to effectively handle high-dimensional data, avoid over-fitting, and exhibit good tolerance to missing values and noisy data. Additionally, it can evaluate feature importance and is suitable for fields such as medical diagnosis and financial risk control.

The working principle of Random Forest is based on two main mechanisms: Bagging and random feature selection. Bagging is the core strategy of Random Forest to reduce variance and enhance model stability. First, the model constructs T decision trees. For each tree, samples are randomly drawn with replacement from the original dataset to form the training set for that tree. Then, each decision tree is trained independently, and the prediction result is obtained by taking the average of the output results of all decision trees.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{1}$$

where $h_t(x)$ represents the prediction result of the t_{th} decision tree.

When splitting nodes in each decision tree, Random Forest further introduces a random feature selection mechanism. Traditional decision trees consider all M features during node splitting, while Random Forest randomly selects only m features and chooses the optimal splitting method from them. For regression problems, the value of m is typically $M/3$.

This paper employed grid search to optimize the Random Forest model. It traverses the preset parameter grid, conducts cross-validation for each combination of parameters, evaluates the performance of the model on the validation set, and selects the parameter combination with the best performance as the parameters for the final model. The workflow of the Random Forest based on grid search is illustrated in Figure 1.

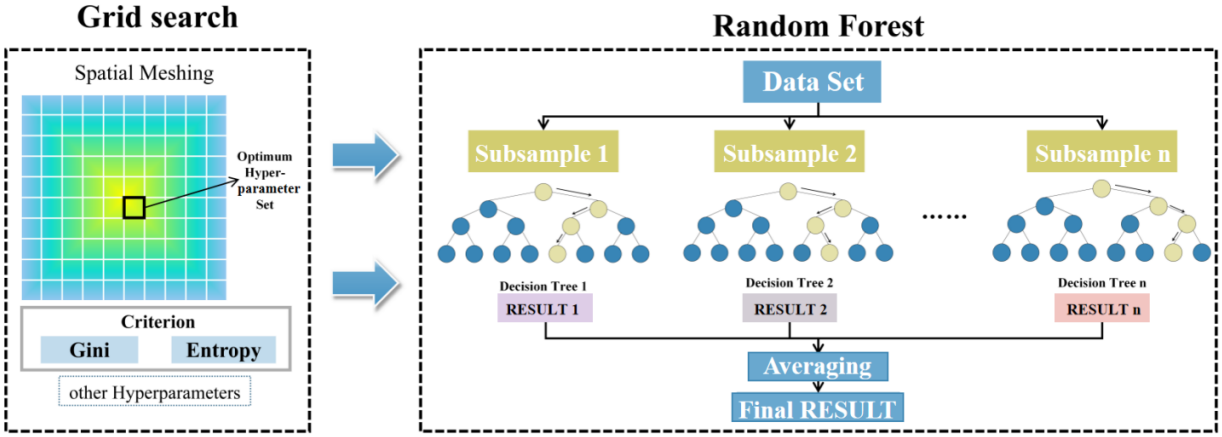

Figure 1. GSRF algorithm

For the prediction of the 2028 Olympic medal table, the article selected three features: athlete performance level, the number of events, and whether a country is the host. Among them, the athlete level is measured by the rate at which athletes gain medals.

$$athlete_i = \frac{nation_i}{\sum nation_i} \tag{2}$$

where $athlete_i$ means the medals gained rate of the athlete and $nation_i$ is total number of medals the i_{th} nation won.

Whether a country is the host is represented as a binary 0-1 variable.

$$host_i = \begin{cases} 1 & the\ country\ is\ the\ host \\ 0 \end{cases} \tag{3}$$

where $host_i$ means whether the i_{th} country is the host.

2.2. Logistic Regression

Logistic regression is a classical classification algorithm primarily used to solve binary classification problems, though it can also be extended to handle multi-classification tasks. Its core idea involves associating input features with probabilistic outputs through linear combinations and non-linear mappings. Unlike linear regression, the output of logistic regression is a probability value between 0 and 1, with class division achieved by setting a threshold. Due to its simplicity, efficiency, and strong interpretability, logistic regression is widely applied in fields such as medical diagnosis, credit scoring, and prediction.

Logistic regression mainly involves the following three steps. First, based on three features—athlete level, the number of events, and host status—the article further consider the influence of GDP, as a country's economic level is also a critical factor affecting sports performance. The features are linearly combined with weights to obtain z. Then, probability mapping is performed through the Sigmoid function, which maps z to a range of 0 and 1.

$$g(z) = \frac{1}{1+e^{-z}} \tag{4}$$

$$z = b + w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{5}$$

where represents the weights, and x_i include the number of events, GDP data, host status, and athletes' abilities.

This article used the gradient descent algorithm to update the model weights with the goal of maximizing the likelihood function. The gradient descent algorithm iteratively minimized the loss function until the op timal solution was found. The performance of the model was measured by using the logarithmic loss function, which is defined as Equation 6.

$$J(\omega) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(Z_i) + (1-y_i)\log(1-g(z_i)) \tag{6}$$

where z_i is a linear combination of the i_{th} sample, and y_i is the corresponding category label (either 0 or 1).

3. Results

The data sources are https://www.kylc.com/stats/global/yearly_overview/g_gdp.html, https://www.datapandas.org/ranking/olympic-medals-by-country, https://www.olympics.com/zh/news/ paris-2024-olympics-full-list-ioc-national-olympic-committee-codes.

3.1. The result and analysis of Random Forest

This paper used the Random Forest Model to predict the medal results for the 2028 Olympic Games, and the medal table was generated, as shown in Figure 2.

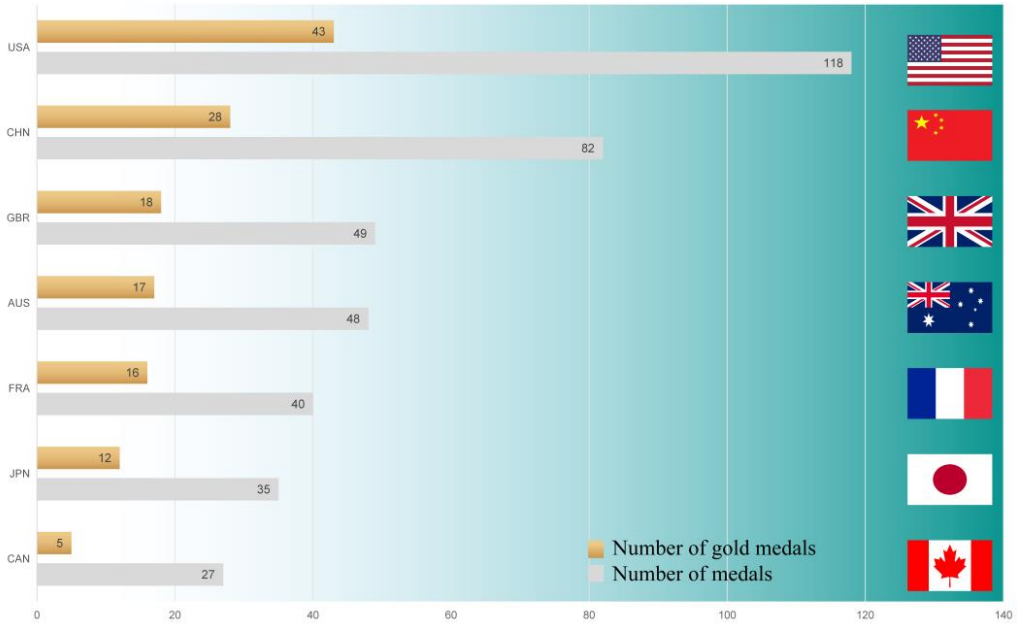

Figure 2. 2028 Forecast

As depicted in Figure 2, the United States is projected to lead the medal standings with a total of 118 medals and 43 gold medals. China is expected to secure the second position, achieving 82 medals and 28 gold medals. The United Kingdom is anticipated to rank third, earning 18 gold medals and a cumulative total of 49 medals.

This article compared the medal counts from 2024 with the predicted gold medal counts and total medal counts for 2028, resulting in Figure 3. Regarding the gold medal counts, it is evident that the United States and the United Kingdom are expected to perform better, while France remains relatively unchanged, and other countries may see a decline in their performance. As for the total medal counts, most countries are projected to see a reduction, with Canada being the only country showing a stable trend.

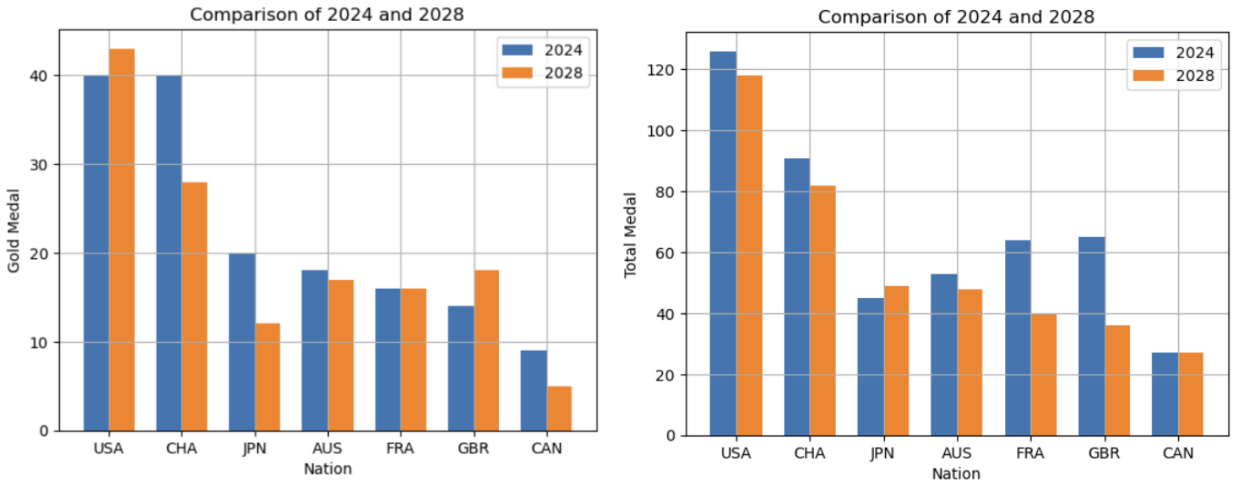

Figure 3. Comparison of 2024 and 2028

Regarding the gold medal counts, it is evident that the United States and the United Kingdom are expected to perform better, while France remains relatively unchanged, and other countries may see a decline in their performance. As for the total medal counts, most countries are projected to see a reduction, with Canada being the only country showing a stable trend.

Based on the feature importance derived from the models, shown in Figure 4, for both the gold medal prediction model and the medal prediction model, the most significant feature is the medals-gaining rate of the athlete, which reflects the athlete's level. This indicates that an athlete's level in a country has the greatest impact on the number of medals that country wins.

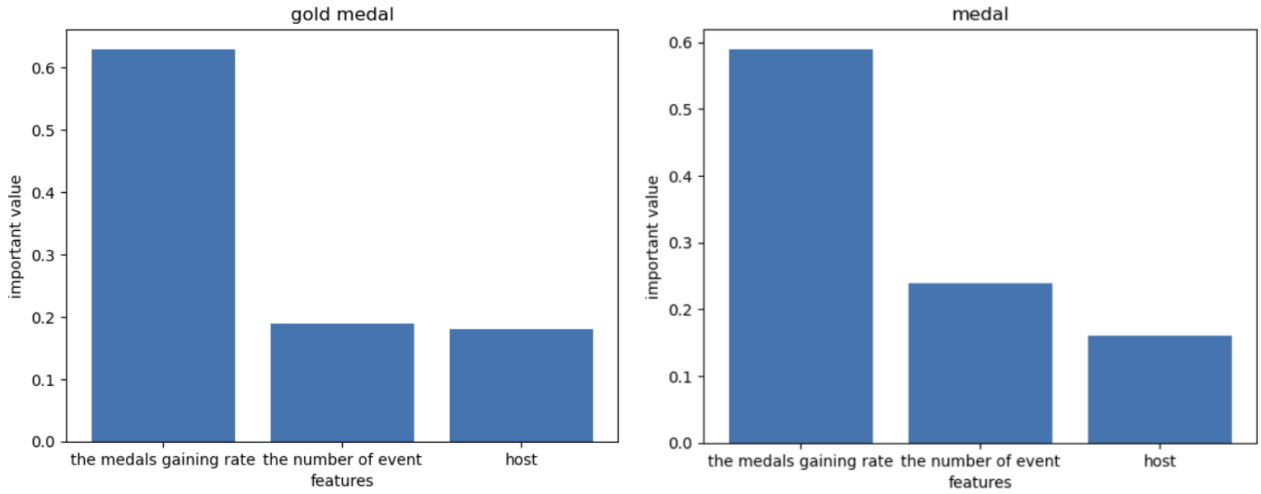

Figure 4. The importance value of each feature

The paper set the test size to 0.2 and used mean square error and mean absolute error, as well as goodness-of-fit R^2 to evaluate the medal number prediction model.

Table 1. Gold medal model's evaluation parameters

MSE	MAE	R^2
13.37	2.70	0.88

Table 2. The total medal count model's evaluation parameters

MSE	MAE	R^2
62.05	5.80	0.90

As shown in Tables 1 and 2, both the MAE and MSE are relatively small, with the R^2 approaching 1, indicating a good model fit.

Using the model to predict the outcome of the Paris 2024 Olympic Games, the predicted and actual medal counts for the seven countries are displayed in Figure 5.

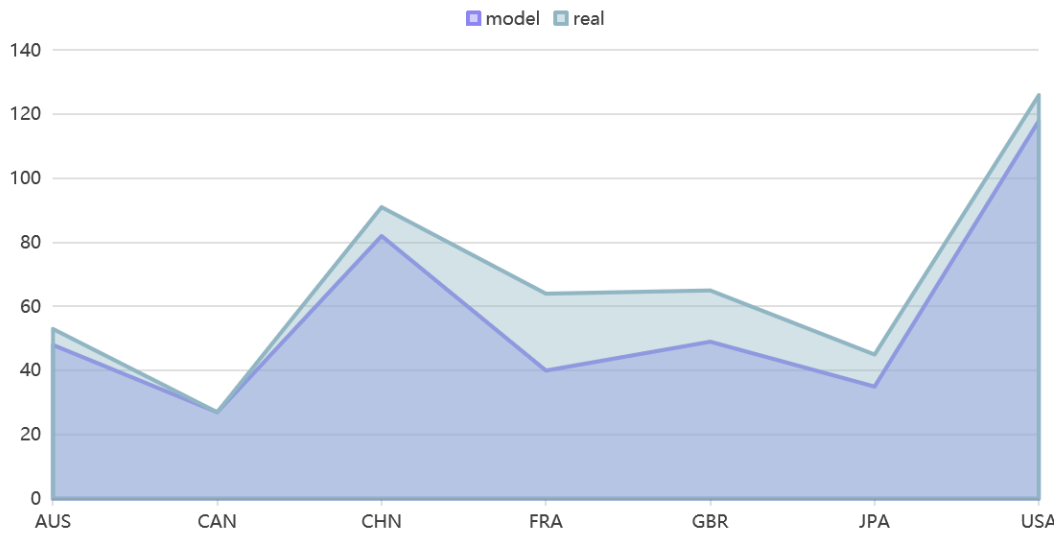

Figure 5. The difference between the true values and the predicted values

As can be seen from Figure 5, the trend of the predicted medal counts by the model aligns closely with the actual medal counts. This not only validates the model's high effectiveness in capturing the complex patterns of countries' performance in obtaining Olympic medals but also further demonstrates that the model can well account for multiple factors influencing medal counts. Specifically, for countries with strong sports capabilities and consistent active performance in the Olympics, such as the United States and China, the model's predicted values are close to the actual values, indicating that the model can accurately grasp these countries' stable performance and potential competitiveness in the Olympics. For countries with relatively fewer medals, such as Canada and Japan, the model can still reasonably predict their medal acquisition to a certain extent, reflecting the model's adaptability to countries at different sports capability levels.

3.2. The result and analysis of Logistic Regression

This article set the number of iterations to 100 and the learning rate to 0.03. After each iteration, the loss function is calculated in Figure 6, and the calibration curve is shown in Figure 7.

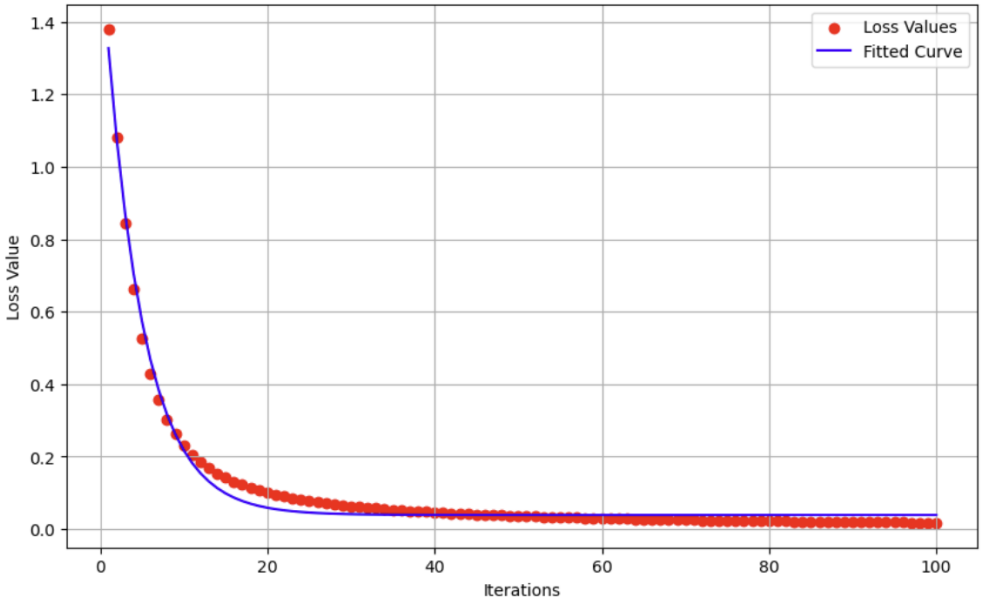

Figure 6. Loss function

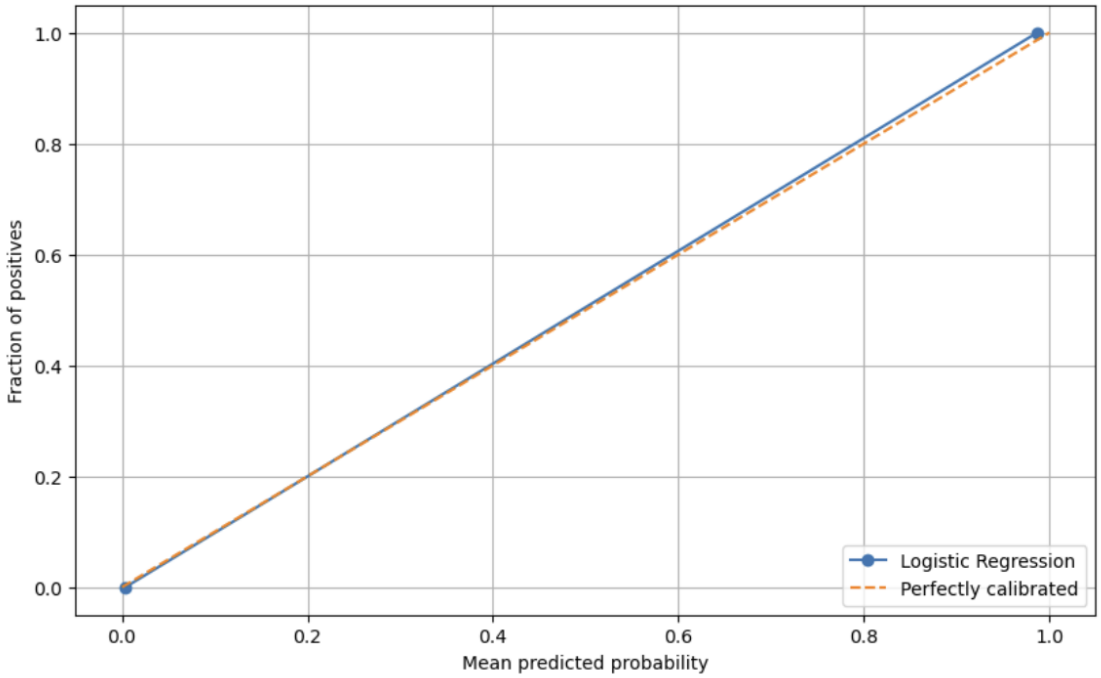

Figure 7. The calibration curve

It can be observed that the value of the loss function gradually decreases with each iteration. When the loss function reaches its minimum value, the weight set at that point corresponds to the optimal solution. The calibration curve coincides with the diagonal, indicating that the predicted probabilities by the model are roughly the same as the actual probabilities of occurrence.

After training the model, the feature vectors of the countries that have not won medals in 2028 are input into the Sigmoid function to calculate the output probabilities. The countries with the highest probability of winning medals are shown in Table 3.

Table 3. The countries most likely to win their first medal and their probabilities

Countries		Probability (%)
	TUV	3.5302
	NRU	3.5291
	MHL	3.5274
	KIR	3.5273
	PLW	3.5266
	STP	3.5232
	SAM	3.5158
	VIN	3.5158
	VAN	3.5151
	SKN	3.5146

As indicated by the data in the table, Tuvalu (TUV) ranks first with a probability of 3.5302%, suggesting a relatively higher likelihood of winning its first Olympic medal in 2028. Closely following are Niue (NRU), the Marshall Islands (MHL), Kiribati (KIR), Palau (PLW), and Vanuatu (STP), with their probabilities of winning first-time medals being closely aligned, all ranging between

3.52% and 3.53%. Although there are minor discrepancies in the predicted probabilities of different countries achieving their first Olympic medals in 2028, the overall probability levels remain relatively concentrated at low values. This indicates significant challenges for these nations in securing Olympic medals, while also offering a reference framework for the development of sports initiatives in the relevant countries.

4. Conclusions and Outlooks

This paper initially predicts the medal table of the 2028 Olympic Games, innovatively combining athlete performance, host status, and economic factors. The results show that the United States, China, and the United Kingdom rank among the top three. In addition, this paper focuses on countries that have never won a medal, and predicts that Tuvalu has a 3.530% probability of winning its first medal at the 2028 Olympics. This study provides a scientific basis for sports management departments in various countries to formulate development strategies. Additionally, the established models can be widely applied to predictive analyses of other sports events, offering new perspectives and tools for sports science research.

In the predictive model, considering only GDP as a single economic indicator may not comprehensively reflect a country's investment in sports. Further research could incorporate additional factors to enhance the model's robustness.

References

[1] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction [C]//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023: 249-267.

[2] Jia M, Zhao Y, Chang F, et al. A Random Forest Regression Model Predicting the Winners of Summer Olympic Events [C]//Proceedings of the 2020 2nd International Conference on Big Data Engineering. 2020: 62-69.

[3] Clephas C, Stergiou P, Tyreman H, et al. Predicting Olympic Success by Regression Modeling in Sport-An Analysis of the Beginning of the 21st Century [C]//World Congress of Performance Analysis of Sport & International Conference of Computer Science in Sports. Cham: Springer Nature Switzerland, 2022: 60-63.

[4] Peters T. Winning against the odds: A socio-economic analysis of Olympic success [D]. Erasmus University, 2023.

[5] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes [C]//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). IEEE, 2025, 3: 1-6.

[6] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction [C]//2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025: 822-827.

[7] Sanchez-Fernandez P, Vaamonde-Liste A. Olympic medals: Success predictions for Río-2016 [J]. South African Journal for Research in Sport, Physical Education and Recreation, 2016, 38 (3): 195-206.

[8] Badoni P, Choudhary P, Rudesh C P, et al. Predicting Medal Counts in Olympics Using Machine Learning Algorithms: A Comparative Analysis [C]//2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech). IEEE, 2023: 116-121.

[9] Groll A, Ley C, Schauberger G, et al. A hybrid random forest to predict soccer matches in international tournaments [J]. Journal of quantitative analysis in sports, 2019, 15 (4): 271-287.

[10] Zhao K, Du C, Tan G. Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm [J]. Entropy, 2023, 25 (5): 765.

[11] Gifford M, Bayrak T. A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression [J]. Decision Analytics Journal, 2023, 8: 100296.

[12] Zheng S, Man X. An improved logistic regression method for assessing the performance of track and field sports [J]. Computational Intelligence and Neuroscience, 2022, 2022 (1): 6341495.