Olympic Medal Count Prediction Model Based on Xgboost

Gaoyichou Ji *

Institute of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China, 102206

* Corresponding Author Email: 120221320121@ncepu.edu.cn

Abstract. Accurately predicting Olympic medal counts for each country is of significant value for shaping sports policy, optimizing resource allocation, and advancing the sports industry. This study presents a predictive model based on the Extreme Gradient Boosting (XGBoost) algorithm to estimate the number of gold, silver, and bronze medals awarded to each country in the Olympics. The validity of the model is demonstrated by predicting the medal counts for the 2024 Olympics. Relevant data from 1896 to 2024 were collected and preprocessed. Five key features were extracted for each Olympic edition: the host country, the number of gold, silver, and bronze medals, the total medal count, the number of athletes per country, the gender distribution of athletes, and the sports events in which each country participated. Data from 1896 to 2012 were used for training, while data from the 2016 and 2020 Olympics were used as the test set. The model was trained using the XGBoost algorithm, and optimization was performed by minimizing the root mean square error (RMSE). Four features for the 2024 Olympics—host country, number of participating athletes, gender distribution of athletes, and the sports events contested—were used to predict the medal counts for each country. By comparing the predicted results with the actual data, the RMSE values for gold, silver, bronze, and total medals were calculated to be 0.6462, 0.5547, 0.2965, and 0.3922, respectively. These results validate the exceptional performance of the XGBoost model in predicting Olympic medal counts, providing effective and forward-looking strategic insights for the optimization of sports resource allocation and the setting of competitive goals across countries.

Keywords: Olympic Games, Olympic Medal Prediction, XGBoost, RMSE.

1. Introduction

The Olympic Games are the most widely followed multi-sport event globally. The results of these competitions not only reflect the sports training systems and talent reserves of different countries but also serve as indicators of a nation's overall strength. Accurate predictions of Olympic medal counts can provide valuable insights for sports management organizations and policymakers. These forecasts can help optimize resource allocation and guide the establishment of competitive goals. The ability to make such predictions is critical for strategic planning and decision-making in the sports industry.

In the field of Olympic medal prediction, machine learning and regression models are widely adopted, with the Random Forest (RF) algorithm being one of the most common in machine learning. Dai et al. applied the Autoregressive Integrated Moving Average (ARIMA) model to perform time series analysis on historical Olympic data, predicting medal trends across different countries [1]. They then integrated historical data with ARIMA forecasts and employed RF to make the final medal count predictions. Ran et al. initially used the Fuzzy Cognitive Concept Learning model to establish a fuzzy rule base, extracting valuable granular information from historical Olympic data [2]. Subsequently, they applied the Transformer model for predicting Olympic medal counts. Yang et al. utilized DeepForest for medal prediction, leveraging feature data from Long Short-Term Memory to enhance prediction accuracy through multi-layer feature learning [3]. Li et al. first applied the Probit Regression model for binary classification to identify countries likely to win medals [4]. For the countries predicted to win, the Tobit Regression model was employed to predict the number of medals won. Despite these approaches, XGBoost has been rarely applied in the context of Olympic medal prediction.

XGBoost, as an ensemble learning algorithm in machine learning, demonstrates strong feature extraction capabilities. It excels in handling complex and high-dimensional data, nonlinear modeling scenarios, and preventing overfitting. In comparison to RF, XGBoost has superior training time and prediction speed on the same dataset, making it suitable for real-time applications [5]. He et al. employed the XGBoost regression model to predict voltage stability margin (VSM) by using operational states of power systems (such as node voltage, current, and line power) as input features, with the corresponding voltage stability margin as the regression target [6]. Bi et al. used XGBoost regression to predict four quality indicators of ore products by taking system-set temperatures, raw material parameters, and production process parameters as input features [7].

This study predicts the number of gold, silver, and bronze medals won by each country in the 2024 Olympic Games using the XGBoost algorithm. Relevant data from the Olympic Games, spanning from 1896 to 2024, were collected and preprocessed. Missing values were filled with zeros. Five key features were extracted for each Olympic edition: the host country, the number of gold, silver, and bronze medals, the total medal count, the number of athletes from each country, the gender distribution of athletes, and the sports events for each country. Data from 1896 to 2012 were used as the training set, while data from the 2016 and 2020 Olympics served as the test set. The four features, excluding the medal counts, were used to predict the 2024 medal counts. The model's ability to accurately predict Olympic medal counts has been validated. These findings have significant implications for sports policy formulation, resource optimization, and the development of the sports industry.

2. Model Development

2.1. Data Preprocessing

Data from the official Olympic website, spanning from 1896 to 2024, were initially collected. The raw data included the number of gold, silver, and bronze medals won by each country in each year, total medal counts, the number of athletes from each country, the gender distribution of athletes, the sports events in each Olympic edition, and the host countries for each edition. Missing values were filled with zeros.

2.2. Feature Engineering

After data preprocessing, five features influencing the number of medals were constructed as the model's inputs. The five features for each Olympic edition included: the host country, the number of gold, silver, and bronze medals along with the total medal count for each country, the number of athletes from each country, the gender distribution of athletes, and the sports events for each country.

The total number of medals for a given country in year i can be obtained by summing the counts of gold, silver, and bronze medals. This can be mathematically represented by Eq. (1). The calculation method for the number of athletes from a given country in year i can be mathematically represented by Eq. (2). The gender distribution is determined by classifying athletes based on their gender. This can be mathematically represented by Eq. (3) and Eq. (4).

$$\mathrm{Total}_i = \mathrm{Gold}_i + \mathrm{Silver}_i + \mathrm{Bronze}_i \tag{1}$$

where Gold_i, Silver_i, and Bronze_i represent the number of gold, silver, and bronze medals for a given country in year i.

$$\mathrm{Num\_Athletes}_i = \sum_{j=1}^{n_i} 1 \tag{2}$$

where n_i is the number of athletes from a given country in year i.

$$\mathrm{Male\_Athletes}_i = \sum_{j=1}^{n_i} 1(\mathrm{Sex}_j = \mathrm{M}) \tag{3}$$

$$\mathrm{Female\_Athletes}_i = \sum_{j=1}^{n_i} 1(\mathrm{Sex}_j = \mathrm{F}) \tag{4}$$

The m sports events participated by a given country in year i are represented as a set $\mathrm{Sports\_Events}_{\mathrm{i}}$. This can be mathematically represented by Eq. (5). The host country in year i is represented as $\mathrm{Host\_Country}_{\mathrm{n}}$.

$$\mathrm{Sports\_Events}_{\mathrm{i}} = \{ \mathrm{Sport}_{1}, \mathrm{Sport}_{2}, \ldots, \mathrm{Sport}_{\mathrm{m}} \} \tag{5}$$

After extracting the features, a specific country-year combination feature vector is generated. Each feature vector includes the number of gold, silver, and bronze medals, the total number of athletes, the breakdown of male and female athletes, sports events, and host country information. Mathematically, the n-th country-year combination is represented by Eq. (6). The total number of $\mathrm{Features}_{\mathrm{n}}$ is the product of the number of countries and the number of Olympic editions.

$$\mathrm{Features}_{\mathrm{n}} = [\, \mathrm{Gold}_{\mathrm{n}}, \mathrm{Silver}_{\mathrm{n}}, \mathrm{Bronze}_{\mathrm{n}}, \ldots, \mathrm{Sports\_Events}_{\mathrm{n}}, \mathrm{Host\_Country}_{\mathrm{n}}] \tag{6}$$

2.3. XGBoost Model Training and Optimization

The dataset is divided into training and testing sets. The training set includes data from 1896 to 2012, and the testing set includes data from 2016 and 2020. The training set contains input features $\mathrm{X_{train}}$ and target variables $\mathrm{y_{train}}$. The input features can be represented by Eq. (7). The target variable $\mathrm{y_{train}}$ represents the quantities of gold, silver, and bronze medals (Eq. (8)).

$$\mathrm{X_{train}} = \begin{bmatrix} \mathrm{Feature}_{1} \\ \mathrm{Feature}_{2} \\ \vdots \\ \mathrm{Feature}_{\mathrm{n}} \end{bmatrix} \tag{7}$$

where n is the number of countries multiplied by the total number of Olympic editions.

$$\mathrm{y_{train}} = \begin{bmatrix} \mathrm{Gold}_{1} & \mathrm{Silver}_{1} & \mathrm{Bronze}_{1} \\ \mathrm{Gold}_{2} & \mathrm{Silver}_{2} & \mathrm{Bronze}_{2} \\ \vdots & \vdots & \vdots \\ \mathrm{Gold}_{\mathrm{n}} & \mathrm{Silver}_{\mathrm{n}} & \mathrm{Bronze}_{\mathrm{n}} \end{bmatrix} \tag{8}$$

During the training process, XGBoost uses an objective function (Eq. (9)) that combines the minimization of the loss function and a regularization term [8]. The loss function measures the error between the model's predicted value $\hat{\mathrm{y}}_{\mathrm{i}}$ and the true value $\mathrm{y_i}$, using the mean squared error (MSE). MSE can be mathematically represented by Eq. (10). The regularization term $\Omega(\mathrm{f})$ is used to control the complexity of the model and prevent overfitting (Eq. (11)).

$$\mathrm{L}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \ell(\mathrm{y_i}, \hat{\mathrm{y}}_{\mathrm{i}}) + \Omega(\mathrm{f}) \tag{9}$$

$$\ell(\mathrm{y_i}, \hat{\mathrm{y}}_{\mathrm{i}}) = \frac{1}{2}(\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{i}})^2 \tag{10}$$

$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \mathrm{w}_{\mathrm{j}}^{2} \tag{11}$$

where T represents the number of leaves in the tree, $\mathrm{w_j}$ denotes the weight of the j-th leaf, γ and λ are the regularization parameters that control the model complexity.

XGBoost uses a gradient boosting approach to train trees, with the model progressively updating the prediction value $\hat{\mathrm{y}}_{\mathrm{i}}^{(\mathrm{t}+1)}$ for each round (Eq. (12)). In each iteration, XGBoost updates the model's prediction by calculating the gain of each tree [9]. During the tree splitting process, XGBoost calculates the split gain to select the optimal feature split point. The formula for calculating the split gain can be represented by Eq. (13).

$$\hat{\mathrm{y}}_{\mathrm{i}}^{(\mathrm{t}+1)} = \hat{\mathrm{y}}_{\mathrm{i}}^{(\mathrm{t})} + \eta \cdot \mathrm{f_t}(\mathrm{x_i}) \tag{12}$$

where $\hat{y}_i^{(t)}$ represents the prediction value for the t-th round, $f_t(x_i)$ denotes the prediction value of the tree model for input x_i in the t-th round, and η is the learning rate, controlling the step size for updating each round's prediction.

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in L} g_i)^2}{\sum_{i\in L} h_i+\lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i+\lambda} - \frac{(\sum_{i\in S} g_i)^2}{\sum_{i\in S} h_i+\lambda}\right] \tag{13}$$

where g_i represents the gradient, h_i denotes the second-order derivative, L and R represent the left and right subtrees of the current node, and S refers to the data set at the current node.

2.4. Model Prediction and Result Output

After the training is complete, the trained model is used to predict the number of gold, silver, and bronze medals for the 2024 Olympics. The prediction process is carried out by inputting the new features X_{2024} into the trained model f_{trained}. Finally, the predicted data for the number of medals and the total medal count for each country in 2024 is output.

3. Model Performance Evaluation

In this study, the Python programming language was utilized to implement the Olympic medal count prediction for the 2024 Games. To optimize the model, the following hyperparameters were configured during training: a learning rate of 0.1, a maximum tree depth of 6, both subsample and colsample_bytree ratios set to 0.8, and a maximum of 100 boosting iterations. The RMSE was employed as the evaluation metric [10]. During training, the RMSE for the training set gradually converged to 0.06676, while that for the validation set stabilized at 0.25165. Figure 1 illustrates the RMSE trajectories for both the training and validation sets.

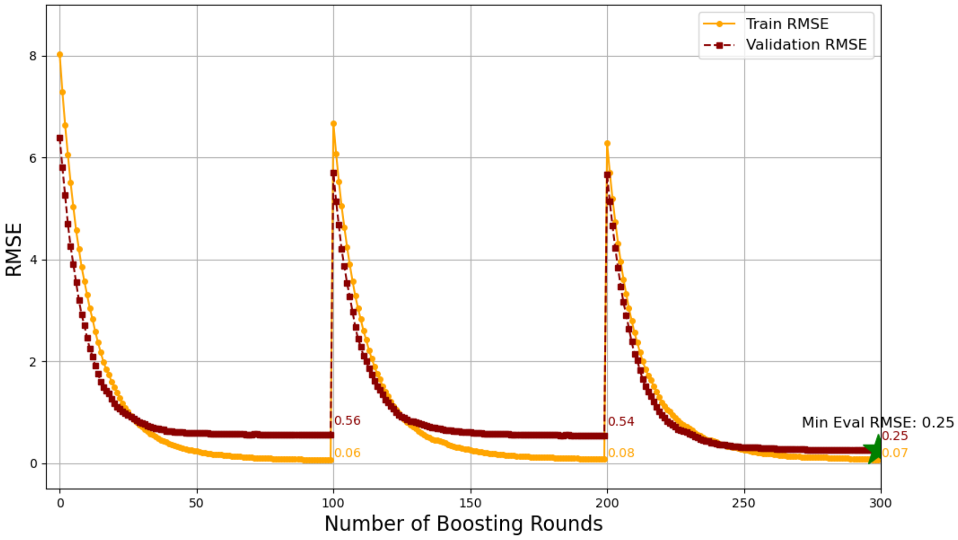

Figure 1. RMSE Curves for the Training and Validation Sets

Figure 2 presents a comparison between the predicted and actual medal counts (gold, silver, and bronze) for the top 20 countries ranked by total medal count. Predicted values for each medal type were rounded to the nearest integer and aggregated to compute the total medal counts per country. Table 1 compares the predicted and actual medal counts for the top 20 countries, based on rounded total values. The variables P_GOLD, P_SILVER, P_BRONZE, and P_TOTAL denote the predicted counts of gold, silver, bronze, and total medals, respectively, while GOLD, SILVER, BRONZE, and TOTAL represent the actual medal counts. The RMSE values for all predicted results were computed as follows: 0.6462 for gold, 0.5547 for silver, 0.2965 for bronze, and 0.3922 for total medals. Analysis indicates that predictions for bronze medals and total medal counts were more accurate, whereas those for gold and silver were comparatively less precise. Overall, the model exhibited strong predictive performance and achieved high levels of accuracy.

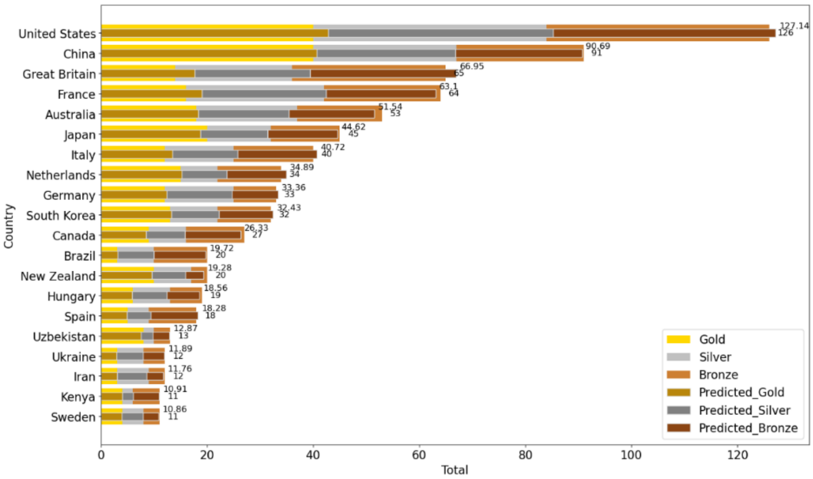

Figure 2. Comparison of Predicted and Actual Values for the Number of Gold, Silver, and Bronze Medals in 2024 for the Top 20 Countries Based on Total Medal Count

Table 1. Comparison of Predicted and Actual Values for the Top 20 Countries Based on Rounded Total Medal Counts

RANK	NOC	P_GOLD	P_SILVER	P_BRONZE	P_TOTAL	NOC	GOLD	SILVER	BRONZE	TOTAL
1	USA	43	42	42	127	USA	40	44	42	126
2	CHN	41	26	24	91	CHN	40	27	24	91
3	GBR	18	22	27	67	GBR	14	22	29	65
4	FRA	19	23	21	63	FRA	16	26	22	64
5	AUS	18	18	16	52	AUS	18	19	16	53
6	JPN	19	13	13	45	JPN	20	12	13	45
7	ITA	13	13	15	41	ITA	12	13	15	40
8	NED	15	8	12	35	NED	15	7	12	34
9	GER	12	12	9	33	GER	12	13	8	33
10	KOR	13	9	10	32	KOR	13	9	10	32
11	CAN	9	7	10	26	CAN	9	7	11	27
12	BRA	3	7	10	20	NZL	10	7	3	20
13	NZL	10	6	3	19	BRA	3	7	10	20
14	HUN	6	7	6	19	HUN	6	7	6	19
15	ESP	5	4	9	18	ESP	5	4	9	18
16	UZB	8	2	3	13	UZB	8	2	3	13
17	IRN	3	6	3	12	IRN	3	6	3	12
18	UKR	3	5	4	12	UKR	3	5	4	12
19	SWE	4	4	3	11	SWE	4	4	3	11
20	KEN	4	2	5	11	KEN	4	2	5	11

4. Conclusions

This study predicts the gold, silver, and bronze medal counts for each country at the 2024 Olympic Games using the XGBoost algorithm. First, relevant data from the Olympic Games between 1896 and 2024 were collected. After data preprocessing, five features were extracted for each Olympic Games: host country, the number of gold, silver, and bronze medals, total medal count, the number of athletes from each country, the gender distribution of athletes, and the sports events in which each country participated. Data from 1896 to 2012 were used as the training set, while data from the 2016 and 2020 Olympics were used for testing. The model ultimately predicts the medal counts for each country in 2024 based on four input features: the sports events in which each country will participate, the number of athletes from each country, the gender distribution of athletes, and the host country of the 2024 Olympics. Comparison with actual values confirmed that the model can accurately predict the Olympic medal counts.

This model can be applied to predict medal counts for future Olympic Games. For example, when predicting the 2028 Olympic medal counts, the trained model requires only these four features—sports events, the number of athletes, athlete gender distribution, and host country—for each participating nation to predict the number of gold, silver, and bronze medals. The model’s predictions for gold and silver medals are slightly less accurate than those for bronze medals and total medals, indicating room for improvement. In addition to the five features extracted in this study, other factors, such as national GDP and investments in sports, could be incorporated to further enhance the model's predictive accuracy.

References

[1] Dai L, Zhang S, Shi X Z. Design of a prediction model based on ARIMA-Random Forest algorithm[C]// International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC). Incheon, Korea, Republic of: IEEE, 2025: 820–824.

[2] Ran Z X, Ou H H, Qin X D, et al. Research on Olympic medal prediction based on the FCCL-Transformer model[C]// International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC). Incheon, Korea, Republic of: IEEE, 2025: 915–922.

[3] Yang Y K, Feng W. Research on Olympic medal prediction based on the LSTM-GWO-DeepForest model[C]// International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC). Incheon, Korea, Republic of: IEEE, 2025: 844–849.

[4] Yang H, Li X Y, Li S X. Prediction and analysis of Olympic medals based on Probit regression and Tobit model[C]// International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC). Incheon, Korea, Republic of: IEEE, 2025: 839–843.

[5] Asha V, Vasumathi M T, Chowdhury D, et al. Comparative Evaluation of Random Forest, XGBoost, and SVC Models for Fire Detection[C]. Proceedings of the Second International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS-2024). Erode, India: IEEE, 2024: 470-475.

[6] He A Q, Xiong W, Bai Y L. XGBoost-based Voltage Stability Margin Prediction for Power Systems[C]. IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE). Jinzhou, China: IEEE, 2023: 1225-1228.

[7] Bi Z B, Fu C X, Zhu J Y, et al. Control Ore Processing Quality Based on Xgboost Machine Learning Algorithm[C]. Asirope Conference on Electronics, Data Processing and Informatics (ACEDPI). Prague, Czech Republic: IEEE, 2023: 177-180.

[8] Wang Z A, Wang D Z, Li Y M. Intelligent Prediction of Short-Term Load of Distribution Network Based on XGBoost Algorithm[C]. Chinese Control and Decision Conference (CCDC). Xi'an, China: IEEE, 2024: 4617-4621.

[9] Dharani Nivash A, Ammal Dhanalakshmi M. Identification of Cyber Bullying Using XGBoost Compared to Random Forest Classifier to Improve Accuracy[C]. Chinese Control and Decision Conference (CCDC). Nitte, India: IEEE, 2025: 856-859.

[10] Sheng C, Yu H Z. An Optimized Prediction Algorithm Based on XGBoost[C]. International Conference on Networking and Network Applications (NaNA). Urumqi, China: IEEE, 2022: 442-447.