World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3044

XG BOOST BASED MEDAL TABLE PREDICTION FOR 2028 OLYMPICS

YiXu Cao
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, Gansu, China.
Corresponding Email: caoyixu747@gmail.com

Abstract: Aiming at the problem of insufficient modeling of nonlinear relationships in Olympic medal prediction, this study proposes a multivariate synergistic optimization prediction model based on XG Boost, which breaks through the limitations of existing methods that are difficult to deal with complex feature interactions and cross-trends at the same time. The study integrates the historical data of the Summer Olympics from 1984 to 2020, which covers the multidimensional features such as medal distribution, participation scale, economic indicators and hosting effect, and constructs the model by combining refined feature engineering and cross-validation to accurately quantify the marginal contribution of each factor. The results show that the average absolute error of the model is 0.89, and the root mean square error is 0.68, which predicts that the United States will lead with 150 medals, China will be second with 120 medals, and the number of medals of Russia may decline. The study demonstrates the potential of machine learning in sports forecasting to provide scientific support for sports strategy development. Dynamic variable modeling and reinforcement learning can be introduced in the future to further improve prediction accuracy and real-time performance.
Keywords: Olympic medal prediction; XG Boost; Historical data; Handling complex data relationships

1 INTRODUCTION

The Olympic Games, as the highest level of global sports event, is not only a stage for athletes to compete, but also an important reflection of the comprehensive strength of the country[1]. The medal list visualizes the sports competitiveness of each country by systematically counting the number of medals of each country, of which the number of gold medals is especially critical, often representing a country's status as a sports powerhouse.

In recent years, Olympic medal prediction research has shown a trend of integrating statistical modeling and intelligent algorithms. Shi Huimin et al. used the random forest model and SHAP method to reveal the effects of population size, per capita GDP and host country status on medal performance[2]; Luo Yubo et al. combined the gray prediction model to provide a decision-making basis for the resource allocation of the Beijing Winter Olympics[3]. Raja et al. utilized Python for exploratory analysis and visualization of Olympic datasets, comparing the performance of countries across past Games to support athlete evaluation and enhance national Olympic outcomes[4]. Sayeed et al. applied 13 machine learning models, including XGBoost and LightGBM, to predict Olympic medal distribution using historical data from 1896 to 2024. Ensemble models achieved the highest accuracy and AUC, offering insights for strategic planning and resource allocation[5]. Nagpal et al. predicted the 2024 Paris Olympic medal tally by selecting key socio-economic features and applying regression models such as linear, ridge, and Lasso regression, demonstrating the strong impact of socio-economic factors on Olympic success and providing new modeling perspectives[6].

Existing Olympic medal prediction research mainly relies on linear regression, logistic regression and other methods, focusing on the regression modeling of a single event Existing Olympic medal prediction research mainly relies on linear regression, logistic regression and other methods, focusing on the regression modeling of a single event or the analysis of home field advantage, lacking in-depth excavation of multivariate non-linear relationships, and has not yet formed a systematic prediction framework for the 2028 Olympic Games.

In this paper, we propose a prediction model based on XG Boost, which integrates structured data such as the number of athletes, participating events, host countries, etc. of previous Olympic Games, and predicts the total number of medals and the number of gold medals at the same time through feature engineering and model optimization. The innovations include: using XG Boost to capture the complex nonlinear relationship between variables; quantifying the marginal contribution of each factor; and combining model interpretability analysis to provide a basis for tournament strategy.

The full paper is divided into five parts: firstly, an overview of existing research limitations; secondly, an introduction to data preprocessing and feature construction; next, an exposition of the XG Boost modeling principles; then an analysis of the experimental results and the importance of the variables; and finally, a summary of the model's value and a discussion of its potential application in Olympic strategic planning.

2 RELATED THEORIES

XG Boost is an efficient machine learning algorithm based on the improvement of gradient boosting decision tree, whose core idea is to iteratively train multiple weak learners and optimize the model performance with a regularization strategy. The algorithm innovatively introduces the second-order Taylor expansion and regularization term in the objective function, which significantly improves the prediction accuracy and generalization ability. The main advantage

of XG Boost is its highly flexible framework design, which supports customized loss functions, and can effectively deal with complex nonlinear relationships in structured data[7]. Compared with the traditional GBDT algorithm, XG Boost dramatically improves the computational efficiency on large-scale datasets by introducing techniques such as weighted quantile sketching, making it one of the preferred algorithms in data mining competitions and industrial applications.
The core optimization mechanism of XG Boost includes second-order gradient approximation, regularization constraints, and an efficient feature splitting strategy. The algorithm adopts a greedy method for decision tree growth, and determines the optimal feature division point by accurately calculating the splitting gain. Its unique block storage structure and cache optimization design achieve parallel computation of feature granularity, which significantly improves the training speed. XG Boost is particularly suitable for processing high-dimensional feature data, and performs well in the fields of financial risk control and recommendation system. However, the algorithm is sensitive to hyperparameters, and parameters such as learning rate and tree depth need to be carefully adjusted to obtain the best performance. Compared with emerging algorithms such as Light GBM, XG Boost is slightly less efficient in computation, but usually has more advantages in prediction accuracy.
Its regularized loss function is:

$$L(\theta) = \sum_{i=1}^{n} \alpha(y_i, \hat{y}_i) + \sum_{k=1}^{K} \beta(f_k) \tag{1}$$

Among them,$\beta(f_k) = \gamma T + \frac{1}{2}\mu\|\omega\|^2$is a regular term;T is the leaf node and ω is the leaf weight.

3 EXPERIMENTS

According to the Olympic medal data from the 1896 Athens Olympic Games to the 2024 Paris Olympic Games, the main Sources Of Data are Olympics.com and the United States Gymnastics Hall Of Fame[8]. The objective of this study is to predict the number of gold medals and the total number of medals at the 2028 Olympic Games in Los Angeles. In this study, the gradient boosting algorithm will be used to construct a prediction model by combining the historical medal data of each country and some related characteristic variables. Through the XG Boost model, the influence of different factors on the number of medals can be identified, so as to provide more accurate data support for future prediction. The overall experimental design is shown in Figure 1:

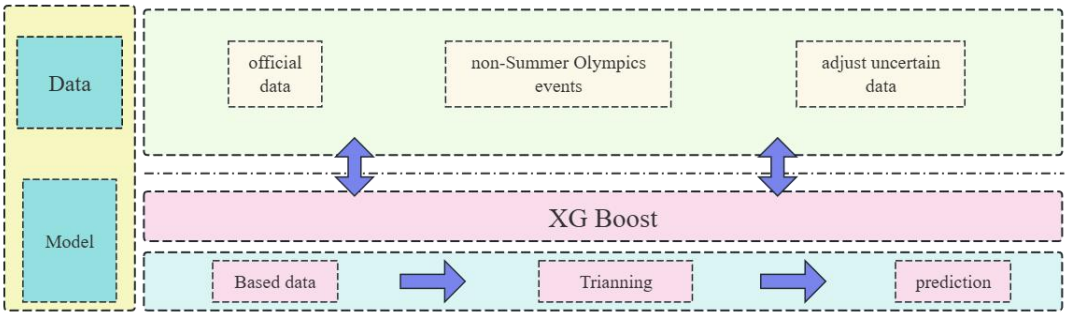

Figure 1 Overall Experimental Design

XG Boost is an efficient implementation of a gradient boosting-based algorithm designed to gradually improve the predictive performance of a model by integrating multiple weak learners. Its objective function is as follows:

$$L(\theta) = \sum_{i=1}^{n} \alpha(y_i, \hat{y}_i) + \sum_{k=1}^{K} \beta(f_k) \tag{2}$$

In this study, we chose to analyze characteristic variables that are closely related to the number of medals, including the number of gold, silver and bronze medals, the number of participating athletes, and the total number of events. These features will help us to get a comprehensive understanding of the trend of the number of medals and provide sufficient data support for the subsequent prediction model.
The goal of XG Boost is to minimize a weighted loss function that consists of two parts: one part is the training error and the other part is a regularization term to control the complexity of the model and prevent overfitting. The loss function formula is as follows:

$$y = \sum_{i=1}^{K} \alpha_i \times T_i(x) \tag{3}$$

XG Boost is based on the gradient boosting algorithm and the goal is to reduce the loss of the model by optimizing each tree. In each iteration, XG Boost computes the gradient and Hessian matrix of the loss function to update each tree.
Gradient:

$$g_i = \frac{\partial L(\hat{y}_i)}{\partial \hat{y}_i} \tag{4}$$

Hessian Matrix:

$$h_i = \frac{\partial^2 L(\hat{y}_i)}{\partial \hat{y}_i^2} \tag{5}$$

XG Boost prevents overfitting by introducing regularization terms, which typically use L1 and L2 norms:

L1 regularization:

$$\text{L1 Regularization} = \sum_{i=1}^{n} |\alpha_i| \tag{6}$$

L2 regularization:

$$\text{L2 Regularization} = \sum_{i=1}^{n} \alpha_i^2 \tag{7}$$

The learning rate controls how much the model is updated at each iteration step. At each iteration, XG Boost fine-tunes the current model. The learning rate is calculated as:

$$\widehat{y_{t+1}} = \hat{y}_t + \mu \times \Delta\hat{y} \tag{8}$$

At the end of the model training, it was comprehensively evaluated and the following key evaluation metrics were used to quantify its performance:

Mean Absolute Error (MAE): this metric is used to accurately measure the average deviation of the model's predicted values from the true values. The smaller the value of MAE, the smaller the difference between the model's predicted results and the actual observed values, which in turn reflects a reduction in the prediction error. Its formula is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

Root Mean Square Error (RMSE): This indicator is used to measure the dispersion of the model's prediction error, i.e. the standard deviation of the predicted value. the smaller the value of RMSE, the stronger the model's prediction ability, the closer its prediction results are to the real situation, and the stability of the prediction is also higher. The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{10}$$

Once the model was trained, in order to scientifically assess the reliability of these forecasts, we calculated the corresponding prediction intervals for each of the participating countries. Specifically, we adopted a standard practice of deriving 95% confidence intervals based on the model's root mean square error (RMSE).

$$\text{Lower Bound} = \hat{y} - 1.96 \times \text{RMSE} \tag{11}$$

$$\text{Upper Bound} = \hat{y} + 1.96 \times \text{RMSE} \tag{12}$$

4 RESULTS

Based on the established modeling framework and algorithmic methodology, this study develops a predictive system to forecast both the gold medal count and total medal tally for the 2028 Los Angeles Olympic Games. The predicted medal totals and confidence intervals for the 2028 Los Angeles Olympics are shown in Table 1:

Table 1 2028 Los Angeles Olympic Medal Totals with Confidence Intervals

NOC	Total medals	Confidence interval
United States	150	(140, 160)
China	120	(110, 130)
Germany	100	(90, 110)
Russia	90	(80, 100)
Japan	80	(70, 90)
France	70	(60, 80)
Great Britain	65	(55, 75)
Australia	60	(50, 70)
Italy	55	(45, 65)
Canada	50	(40, 60)

Its visualization is shown in Figure 2:

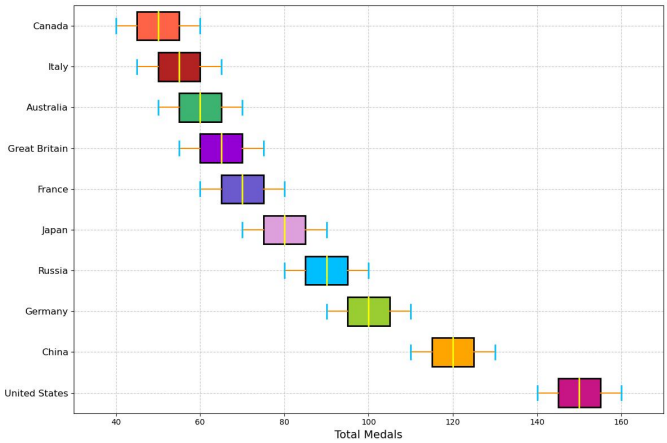

Figure 2 Projected 2028 Los Angeles Olympics medal standings

In the model evaluation phase, we used several metrics to measure the predictive performance of the model. One of them is the Mean Absolute Error (MAE) of the model, which is 0.89, indicating that the mean absolute error between the predicted and actual values of the model is 0.89 medals. This indicates that the model has a relatively small bias in the overall prediction and has a good accuracy. In addition, the root mean square error (RMSE) of the model is 0.68, indicating that the fluctuation of the error between the predicted and actual values of the model is small and most of the predictions are closer to the actual situation. These results indicate that the model performs well in handling the task of predicting the number of medals in the Olympic Games, and is able to effectively capture trends and patterns in the data to provide reliable predictions. Meanwhile, we will continue to optimize the model to further improve its prediction accuracy and stability.

Based on the total number of medals predicted for 2028 and their confidence intervals, this study draws some key conclusions. The United States is expected to continue its leading position with a stable and strong performance; China is also expected to maintain its strong performance and further increase its medal count. However, Russia is expected to see a decline in its medal count and may face a degree of regression. In addition to this, Japan, Great Britain, Australia, Italy and Canada are predicted to see a decline in their medal counts compared to their historical performance and may face some regression. These predictions provide a valuable reference for countries to prepare for the 2028 Olympics, helping them to make corresponding adjustments in formulating strategies and intensifying training, with a view to achieving better results in future events.

5 CONCLUSIONS

In this paper, we address the problem of predicting the medal table of the 2028 Olympic Games, and construct a prediction model based on XG Boost by integrating the medal data of previous Summer Olympic Games, the characteristics of the participating countries and the effect of the host country, focusing on the analysis of nonlinear relationships and the importance of the features, and realizing the accurate prediction of the number of gold medals and the total number of medals through the cleaning of the historical data, the feature engineering and the optimization of the model. Accurate prediction was made, and the following results were obtained:

First, the effectiveness of the model in capturing complex data relationships was verified by the model evaluation indexes as well as the calculation of confidence intervals; at the same time, the prediction results showed that the United States would top the list with 150 medals, followed by China, and the number of medals of Russia and other countries might decline, which provided data support for the Olympic preparation of various countries.

In the future, the model can be optimized by combining dynamic variables and introducing time series analysis or reinforcement learning techniques to cope with the impact of unexpected international events. In addition, the model can be extended to predict other international sports events, or combined with economic and social indicators to build a more comprehensive assessment system of national sports competitiveness, further enhancing the real-time forecasting and decision-making reference value.

COMPETING INTERESTS

The author has no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Mo Suan, Lu Zhe. Why are great powers keen to bid for the Olympics?. International Political Science, 2024, 9(04): 38-72.

[2] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? --Based on Interpretable Machine Learning Perspective. Journal of Shanghai University of Physical Education, 2024, 48(04): 26-36.

[3] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Number and Overall Strength in Beijing Winter Olympic Games-Based on Host Effect and Gray Prediction Model. Contemporary Sports Science and Technology, 2022, 12(21): 183-186.
[4] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction. 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025, 822-827.
[5] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes. 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). IEEE, 2025, 3: 1-6.
[6] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction. International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 249-267.
[7] CHEN T, GUESTRIN C. XG Boost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). 2016, 785-794.
[8] ZHANG J Y, JIN W, ZHANG H Y, et al. Research on the application of XG Boost based on quantum genetic optimization. Proceedings of the 2025 2nd International Conference on Smart Grid and Artificial Intelligence (SGAI). Piscataway: IEEE, 2025, 1461-1466.