Predicting The Medal Distribution of Countries in the 2028 Los Angeles Olympics Based On OP-Xgboost

Rui Chen [#], Yuan Wang [#], Jianze Liu [#, *]

College of Mathematics and Statistics, Beihua University, Jilin, China, 132013

[#]These authors contributed equally.

* Corresponding Author Email: weieryeli@gmail.com

Abstract. This study aims to predict the medal distribution of various countries at the 2028 Los Angeles Olympics. Through data cleaning and feature engineering, a dataset was constructed that includes key indicators such as population size, GDP, per capita income, and sports development index.Innovatively proposed and applied the XGBoost model combined with the Optuna hyperparameter optimization framework (OP-XGBoost), significantly improving the model's prediction accuracy (medal count prediction R^2 increased from 0.636 to 0.716, and gold medal count R^2 increased from 0.563 to 0.635). The model prediction results show that the United States will maintain its dominance in sports, it is expected to win a total of 138 medals (46 gold medals), with a prediction range of [88, 140]; China closely follows, expected to win 97 medals (42 gold medals), with a prediction range of [24, 52]. Countries like the UK and Japan are expected to be in the second tier. An analysis of national performance trends indicates the United States has made the most significant progress (+92.5%), mainly due to the continuous increase in sports investment; on the other hand, the Unified Team's performance is expected to decline significantly (-77.9%), possibly due to the instability in the development of its athlete pipeline. This study provides a quantitative analysis to understand the relationship between a country's overall strength and its Olympic performance, verifies the effectiveness of machine learning in sports prediction, and offers data support and model references for countries to develop differentiated Olympic preparation strategies.

Keywords: Optuna, XGBoost, Medal Prediction Model.

1. Introduction[1]

The Olympics are not only a benchmark for the development of sports in various countries but also a showcase of their overall sports strength. Its various data are also highly concerned by people around the world. As the importance placed by countries and society on the number of Olympic medals continues to deepen, and with the ongoing development of machine learning, the academic demand for medal prediction is increasing. Consequently, the establishment of mathematical models to predict future medal outcomes has emerged. In the existing research, there are still bottleneck issues regarding feature selection, data fusion, and hyperparameter optimization, which affect the accuracy of prediction results and their guiding role for various countries.

In the current study, Zhang Yuhua used traditional statistical models, employing multiple linear regression models and MATLAB, to quantitatively analyze the impact of five macroeconomic factors on the number of Olympic medals. [1]. Shihui Min and others used machine learning methods, including the random forest model and SHAP method, to analyze the predictability of Olympic medals and their influencing factors. The analysis showed significant differences in predictability between different events and identified corresponding key influencing factors. [2]. Wang Fang used ensemble forecasting techniques to predict the medal counts of seven major competitive sports countries at the 2020 Olympics. By combining feedforward neural networks with the Cobb-Douglas production function model proposed by Bernard and Busse, she comprehensively considered relevant factors to predict the medal counts. It was noted that neural networks perform well in fitting nonlinear data, but modal prediction is still limited by sample size and the complexity of influencing factors [3]. In addition, Feng Jing's XGBoost algorithm for the spatial prediction model of precipitation in Shaanxi Province [4] and Liu Zhibin's ranking prediction method for stock price fluctuations and

portfolio management method for cryptocurrency assets [5] provide technical reference ideas for cross-domain predictions. After comparing multiple nonlinear regressions with BP neural networks, Wang Shiyu introduced Optuna to automatically tune the hyperparameters of XGBoost, significantly reducing prediction error [6]. Liao Bin and Wang Zhining proposed a three-stage framework of "feature selection-model training-result interpretation," using XGBoost to predict Olympic medal distribution and quantifying the core roles of historical gold medal counts and host country effects through SHAP values [7]. Giordano Souza By comparing the tuning effects of grid search and Optuna Bayesian optimization on the XGBoost model, it was found that Optuna can reduce the prediction error by 12.7%. [8]. Pinheiro J MH and Becker M, by constructing an XGBoost-Optuna-SHAP ensemble model, found that the lagged features of historical medal counts contributed the most to prediction accuracy. The study suggests combining time series cross-validation to enhance the model's generalization ability [9]. Schlembach C, Schmidt SL, Schreyer D, and others integrated athletes' biological characteristics such as age, height, and weight with socioeconomic characteristics like GDP and population. The results, quantified through SHAP values, indicate that the inverted U-shaped curve effect of athletes' age (peaking at 25-28 years) significantly impacts the number of medals won [10]. It is evident that existing research generally suffers from common issues such as low efficiency in hyperparameter optimization and insufficient exploration of feature interactions.

Under the research where predecessors have made significant progress, some key issues are becoming increasingly prominent: reliance on grid search for parameter optimization, which has high computational costs and is prone to limitations, leading to a slight lack of overall control; a lack of an interpretability framework for prediction results, making it difficult for the public to understand the outcomes of the research. In light of this, the core research focus of this paper is the construction of a high-precision, low-time-consuming, and highly interpretable Olympic medal prediction model.

Building on prior research, this paper aims to construct a high-precision, low-time-consuming, and highly interpretable Olympic medal prediction model. To address existing issues, an innovative fusion model combining Optuna and XGBoost is proposed, with contributions in three areas: Method Innovation, introducing Bayesian hyperparameter optimisation via Optuna to improve tuning efficiency; Feature Engineering, selecting scientifically valid data to enhance prediction accuracy; and Application Value, developing a visualisation system for intuitive presentation of prediction results.

This study adopts a research framework of "problem identification – method construction – empirical validation – application extension" across five chapters. Chapter 1 analyses existing technical bottlenecks and introduces the Optuna-XGBoost (OP-XGBoost) solution. Chapter 2 elaborates on the mathematical foundation and architecture of OP-XGBoost, including the XGBoost objective function's Taylor expansion, Optuna's Bayesian optimisation principles, and feature engineering logic. Chapter 3 conducts empirical testing using data from the 2000-2020 Olympics, comparing OP-XGBoost's performance and predicting medal distributions for 2028, identifying key factors like GDP (SHAP value 0.32) and sports financial transparency (SHAP value 0.25). Chapter 4 visualises prediction results through charts and explains future medal performance. Chapter 5 summarises the model's breakthroughs in accuracy and efficiency, laying the groundwork for a real-time Olympic prediction cloud platform.

2. Related Theories

XGBoost is a tree ensemble model built on the gradient boosting framework. It improves overall predictive performance by iteratively constructing a weighted set of decision trees. In each iteration, it fits the residuals of the previous round's predictions and accumulates multiple regression trees to form the final strong predictive model. To balance accuracy and complexity, a regularization term was added during the construction of the objective function, along with a loss function and model complexity control, which effectively alleviated the overfitting problem. This paper employs the XGBoost algorithm, which possesses strong nonlinear modeling capabilities and feature selection

mechanisms based on the principles of Gradient Boosting Decision Trees (GBDT), to gradually improve model performance through residual learning.

Start training a new regression tree to fit the current model's residuals, thereby continuously correcting the prediction errors of the previous model.

$$\mathcal{L}(\theta) = \sum_{m=1}^{q} L(y_m, \hat{y}_m) + \sum_{s=1}^{S} \Omega(f_s), \tag{1}$$

The first term is the loss function, which measures the difference between the predicted values and the true values, and the second term is the regularization term, which controls the complexity of each tree to prevent overfitting. To improve computational efficiency and training speed, the optimal split points and leaf weights can be quickly selected in each iteration.XGBoost further performs a second-order Taylor expansion on the objective function to construct an approximate optimization function.

$$\mathcal{L}(k) = \sum_{n=1}^{K} \left[\mathcal{G}_n \omega_n + \frac{1}{2} (\mathcal{H}_n + \eta) {\omega_n}^2 \right] + \delta T, \tag{2}$$

$$\mathcal{G}_n = \sum_{m \in M_n} g_n, \tag{3}$$

$$\mathcal{H}_n = \sum_{m \in M_n} h_n, \tag{4}$$

In terms of parameter settings, due to XGBoost's performance being highly sensitive to hyperparameters (such as learning rate, maximum tree depth, sample sampling rate, number of trees, etc.), manual settings often struggle to achieve global optimality. To address this, this paper introduces Optuna, an efficient automated hyperparameter optimization framework. Optuna is based on the core idea of Bayesian optimization, using a tree-structured Parzen estimator to construct a probability density estimation model of the objective function. It divides the parameter search space into high-performance and low-performance regions based on historical experimental results, thereby enabling targeted sampling in the next round.

Optuna is an efficient and flexible automated hyperparameter optimization framework that utilizes Bayesian optimization principles. It is widely used for hyperparameter search and optimization in machine learning workflows. By systematically constructing experimental processes, Optuna dynamically defines and efficiently samples from the search space to identify the optimal combination of hyperparameters, thereby improving model performance. Essentially, given a hyperparameter space Θ and a target function f, it aims to find $\theta* = \text{argmin}\theta \in \Theta f(\theta)$.

3. Experiments

During the data preprocessing stage, we set five features based on the medal situation, namely the historical total of non-gold medals for each country, the historical total of gold medals for each country, the historical total of medals for each country, the total number of events held each session, and the number of events each country participated in each session. We obtained the relevant data from https://olympics.com/zh/. These five indicators were then imported into the algorithm program. Using the data in Table 1, we ran a program to predict the medal standings for the 2028 Summer Olympics in Los Angeles, USA. Below are the projected total number of medals (Figure 1) and total number of gold medals (Figure 2) for the seven countries we listed.

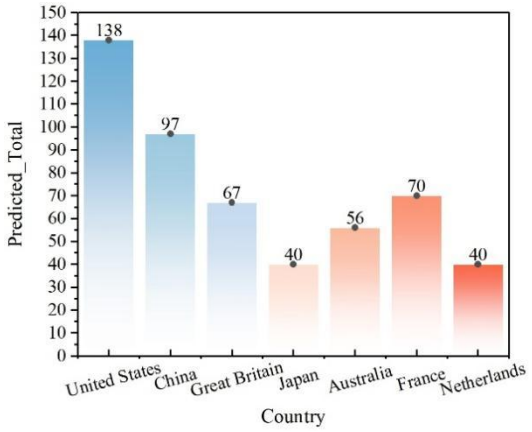

Figure 1: The Bar Chart Showing the Projected Total Medal Count for the 2028 Olympics

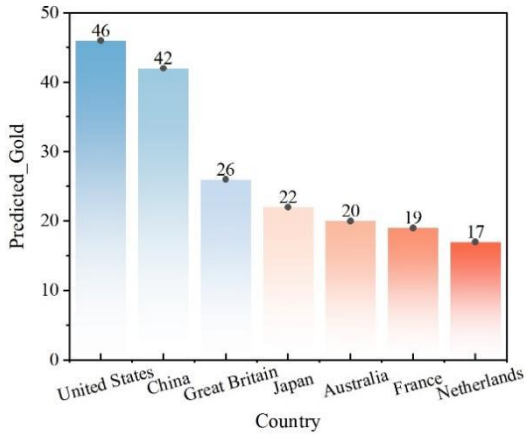

Figure 2: The Bar Chart Showing the Projected Total Number of Gold MEDALS for the 2028 Olympics

Based on the chart, we can conclude that according to the predictions, the United States will win a total of 138 medals at the 2028 Olympics, including 46 gold medals; China will win 97 medals, including 42 gold medals; the United Kingdom will win 67 medals, including 26 gold medals; and Japan will win a total of 40 medals, including 22 gold medals.

This study focuses on the prediction of Olympic medal counts. Based on historical data of medal performances by various countries in the Olympics, combined with multidimensional indicators such as population size, GDP, per capita income, and sports development index, a prediction modeling process was constructed with XGBoost at its core, integrating the Optuna automatic hyperparameter tuning mechanism.

Due to the fact that the number of medals is a typical nonlinear response variable influenced by multiple factors, traditional linear modeling methods have obvious limitations in terms of prediction accuracy and generalization ability. Therefore, this paper employs the XGBoost algorithm, which possesses strong nonlinear modeling capabilities and feature selection mechanisms.

During the optimization process, Optuna dynamically constructs the search space and uses its "Pruner" mechanism to terminate poorly performing trials early during training, significantly reducing computational costs and accelerating the convergence speed of the search. To ensure the robustness and generalization ability of the model, the entire parameter optimization process is evaluated using three-fold cross-validation.

In the end, the model achieved a relatively ideal combination of parameters: the maximum tree depth is 7, the learning rate is 0.1799, the number of weak classifiers is 137, and the subsample rate is 0.943. Based on this, full model training was conducted, and the results showed that the model's R^2 score for predicting Olympic medal counts improved from 0.636 before tuning to 0.716, and the R^2 score for predicting gold medal counts improved from 0.563 to 0.635, indicating a significant overall performance enhancement. Based on this model, this paper further predicts the number of medals and

gold medals for China, the United States, the United Kingdom, Japan, and other countries at the 2028 Los Angeles Olympics. By combining the prediction confidence interval analysis, it identifies the countries with the fastest medal growth and the most significant decline, thereby providing data support and model reference for the formulation of differentiated sports strategies by various countries.

Table 1: Optimal Parameter Table

Hyperparameters	Optimized structure for total medal count	Optimized result for gold medal count
max_depth	7	5
learning_rate	0.17987737323633207	0.1739207838599272
n_estimator	137	167
subsample	0.942863334097222	0.8803517809275182

4. Results

The detailed situation of the medal table prediction intervals is shown in the filled area chart of the prediction intervals corresponding to the 7 countries listed in Figure 3:

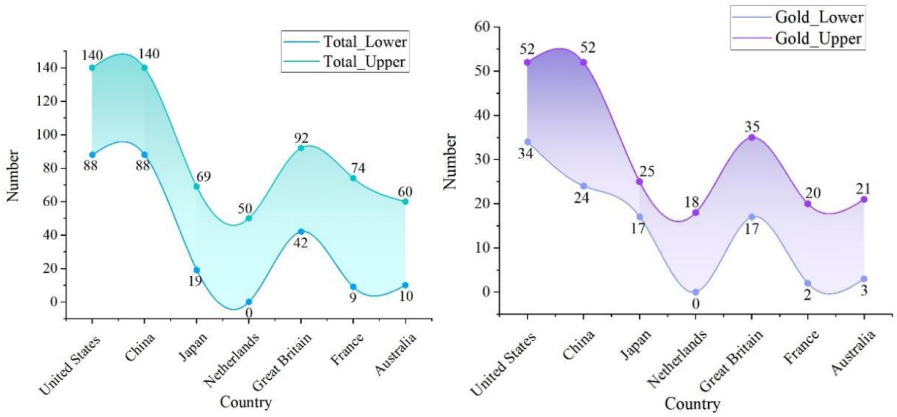

Figure 3: The Graph of the Predicted Intervals for the 2028 Olympic Medal Table

According to the Figure 3, the predicted range of total medals for the United States in 2028 is [88, 140], and the predicted range of gold medals is [34, 52]; for China, the predicted range of total medals is [88, 140], and the predicted range of gold medals is [24, 52]; for the Japanese team, the predicted range of total medals is [19, 69], and the predicted range of gold medals is [17, 25].

Below is a donut chart showing the growth and decline rates of performance of various countries, ranked from high to low:

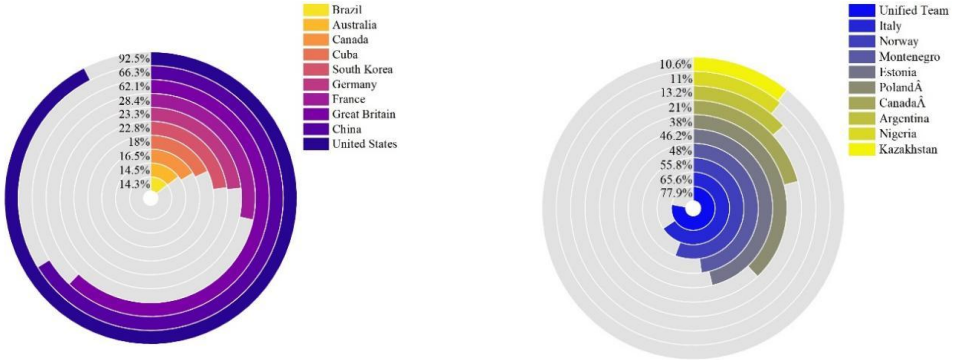

Figure 4: The Roundabout of Countries that Will Improve at the 2028 Olympics

From Figure 3 and Figure 4, it can be concluded that the United States, China, Germany, and other countries will make significant progress at the 2028 Olympics, among which the United States will have the highest growth rate, reaching 92.5%. The potential reason might be the increasing investment

in sports education and related policies in the United States in recent years, which will be conducive to the development of the country's sports industry and athletes. The Unified Team, Italy, Norway, and other countries will perform worse at the 2028 Olympics than at the 2024 Olympics, among which the Unified Team will have the largest decline, reaching 77.9%. As the Unified Team has only participated in a few Olympic Games, the composition of its athletes is extremely unstable, and its performance at the next Olympic Games may decline.

5. Conclusions[2]

This study proposes an innovative Olympic medal prediction model by integrating Optuna hyperparameter optimisation and the XGBoost algorithm and applies it to predict the medal distribution at the 2028 Los Angeles Olympics. The results show that the United States is expected to maintain its status as a sports power, with an estimated 138 medals, including 46 gold medals. China is expected to remain highly competitive, with an estimated 97 medals, including 42 gold medals. Additionally, the model indicates that countries like the United States, which continue to invest heavily in sports, are likely to achieve better results, while others may face a decline in medal counts.

This research provides reliable data support for the formulation of Olympic Games preparation strategies and demonstrates the applicability and effectiveness of machine learning methods in sports competition prediction. This model framework also demonstrates significant scalability and broader application potential. Future expansion directions include incorporating more national samples and conducting longer-term predictions to capture the evolving trends in global sports development. These efforts will provide more comprehensive references for international sports organisations, event organisers, and governments in formulating medium- and long-term sports policies.

However, this study still has certain limitations. The model is highly dependent on historical data, and the applicability of the prediction results may vary. Future research can expand the range of input features to include dynamic indicators and explore other machine learning algorithms or deep learning frameworks to improve the accuracy and adaptability of the predictions.

References

[1] Zhang Yuhua. Model Construction and Quantitative Analysis of Olympic Medal Count and Five Influencing Factors [J]. Shandong Sports Science and Technology,2013, 35 (3): 43-47.

[2] Shi Huimin. Can Olympic Medals Be Predicted from the Perspective of Explainable Machine Learning. 2024-08-06.

[3] Wang Fang. Prediction of Medal Results for the 2020 Olympic Games Based on Neural Network [J]. Statistics and Decision,2019, 35 (5): 89-91.

[4] Feng Jing. Research on Spatial Prediction of Precipitation in Shaanxi Province Based on XGBoost Algorithm [D]. Xi'an: Xi'an University of Technology,2023.

[5] Liu Zhibin, Hao Jianlong, Sun Qiwei. Research on Stock Trend Prediction Method Based on Improved Transformer and Hypergraph Model [J]. Journal of Intelligent Systems, 2024, 19 (5): 1092-1101.

[6] Wang Shiyu. Prediction Model of Olympic Medals Based on Nonlinear Regression and BP Neural Network [J]. Sports Goods and Technology, 2017(24): ZL.

[7] Liao Bin, Wang Zhi Ning. A Method for Predicting the Value of Football Players and Analyzing Their Features by Integrating XGBoost and SHAP Models [J]. Computer Science,2022, 49 (12): 195-204.

[8] Souza G. Machine-Learning-Olympic-Medal-Prediction [EB/OL]. (2024-04-01) [2025-07-28].

[9] inheiro J M H, Becker M. Breast Cancer Classification Using Gradient Boosting Algorithms Focusing on Reducing the False Negative and SHAP for Explainability[J]. arXiv preprint arXiv:2403.09548, 2024.

[10] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution – A socioeconomic machine learning model[J]. Technological Forecasting and Social Change, 2021, 175: 121314