Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 1, 2025

Predicting the 2028 Los Angeles Olympic Medal Table: A Machine Learning Approach with Gradient Boosting and Random Forest Models

Wei Gao*

School of Computing, Beijing Institute of Technology, Zhuhai 519088, China
* **Corresponding author**: Wei Gao (Email: ab2639150579@126.com)

Abstract: This study predicts gold and total medal counts for the 2028 Los Angeles Summer Olympics, analyzing key factors such as the "star coach" effect and event settings to provide strategic insights for National Olympic Committees (NOCs). Traditional Olympic medal predictions rely on near-event data, with limited consideration of historical trends, while recent machine learning models leverage past data but face challenges with overfitting. Using data from the 1896–2024 Summer Olympics, countries are categorized as traditional, emerging, and potential sports powerhouses. Tailored features are selected, applying Gradient Boosting Regression Tree (GBRT) to the first two categories and Random Forest Regression to the latter. The prediction models are evaluated through R^2, MSE, and MAE, achieving an R^2 of 0.65, MSE of 24.2, and MAE of 24.8. The United States and China are projected to lead with 92 total medals and 39 golds each, followed by Great Britain, Australia, and Japan. This framework offers reliable predictions, highlights the impact of historical trends and coaching influence, and aids NOCs in effective resource allocation and strategy development. The novelty of this approach lies in its combination of historical data with machine learning techniques, offering a more comprehensive method for Olympic medal forecasting.

Keywords: Olympic medal prediction, GBRT, Random Forest, NOCs.

1. Introduction

The prediction of Olympic medal tables has long been a hotspot in sports research Traditional prediction methods are mostly carried out when the event is approaching and participating athletes are confirmed, with relatively low reliance on historical medal data However, historical data may actually hide many factors influencing medal counts, such as the development trends of sports in different countries and the inheritance of advantages in specific disciplines Meanwhile, factors like the setting of Olympic events and the mobility of coaches may also affect the number of medals won by each country For example, renowned coaches such as Lang Ping and Belak Karolyi have led teams from different nations to achieve outstanding results through their exceptional coaching abilities Whether this "star coach" effect is significantly reflected in medal counts is worth exploring in depth

Currently, scholars at home and abroad have conducted research on Olympic medal prediction models, with research methods and steps generally falling into the following categories [1]: first, selecting influencing factors for medal counts to analyze and discuss, then establishing prediction models based on the selected factors For instance, Zihan Lu [2] and others conducted in-depth research using a combination of multiple models and algorithms to predict the medal rankings for the 2028 Los Angeles Olympics They used logistic regression models to forecast countries that would win their first medals, identified potential medal-winning nations, revealed the laws of medal distribution, and clarified the impact mechanisms of various factors QianFeng Jin [3] and others constructed a weighted fusion model combining regression and time-series models to achieve higher-precision predictions for the 2028 Olympic medal table DingShu Yan [4] utilized historical data from the 1896–2024 Summer Olympics to build a Long Short-Term Memory (LSTM) model, while also employing decision tree models to study the impact of "excellent coaches" on medal performance, providing valuable strategies for optimizing sports resource allocation and enhancing global competitiveness Schlembach Christoph [5] and others applied a two-stage random forest model, which for the first time outperformed more traditional simple predictions in four consecutive Olympic Games from 2008 to 2020

Today, Olympic medal prediction models have entered a stage of integrating deep learning Compared with traditional mathematical prediction models, deep learning is more effective in processing large volumes of historical data, better at extracting features of relevant factors, and capable of handling more complex problems However, it is crucial to prevent overfitting, which may lead to inaccurate predictions and extreme forecast results After analyzing the research of domestic and foreign scholars, this paper proposes a prediction method different from previous approaches First, based on the historical medal data of all countries from the 1896–2024 Summer Olympics, countries are classified into three types Different features are selected for each type, and Gradient Boosting Regression Tree and Random Forest models are used for prediction Relevant statistical indicators are then employed to validate the feasibility of the model predictions

This paper aims to construct medal count prediction models for different types of countries using medal, event, and athlete data from past Summer Olympics The models will not only predict the medal rankings for the 2028 Los Angeles Summer Olympics and assess the models' uncertainty and performance but also deeply analyze the relationship between event settings and medal counts, explore evidence for the "star coach" effect, and uncover other unique insights into Olympic medal counts revealed by the models, with the goal

of providing valuable references for National Olympic Committees (NOCs)

2. Related Theory

The prediction framework for the 2028 Los Angeles Olympics relies on machine learning algorithms and statistical theories tailored to handle complex Olympic medal data. This section outlines the theoretical foundations of the Gradient Boosting Regression Tree (GBRT) and Random Forest Regression models, as well as the statistical methods used for feature selection and uncertainty quantification.

2.1. Gradient Boosting Regression Tree (GBRT)

GBRT, rooted in ensemble learning and boosting theory, builds a predictive model by combining multiple weak learners to minimize a loss function, typically mean squared error. The algorithm iteratively fits trees to the residuals of prior predictions, updating the model to reduce errors. Mathematically, for a target variable y, the model is expressed as:

$$y = F(x_1, x_2, , x_k) + \varepsilon \quad (1)$$

where F represents the cumulative prediction from all trees, xi are input features , and ε is the residual error. GBRT's strength lies in capturing non-linear relationships and interactions, making it suitable for predicting medal counts for traditional and emerging sports powerhouses.

2.2. Random Forest Regression

Random Forest Regression, rooted in bagging and ensemble learning, builds multiple decision trees using random subsets of data and features, combining their predictions to deliver a robust and stable estimate. By averaging these predictions, the model minimizes overfitting, making it ideal for predicting medal counts for potential countries with inconsistent performance. Features such as historical medal statistics and event participation are randomly sampled to create diverse tree structures, enhancing the model's generalization and reliability.

2.3. Feature Selection Theory

Feature selection is grounded in statistical learning theory, aiming to identify predictors that capture Olympic performance trends. Features such as mean, median, and standard deviation of historical medals, growth rates, and host country status are chosen based on their correlation with medal counts. The "star coach" effect, inspired by human capital theory, is modeled as a potential feature to reflect coaching impact, though its inclusion depends on data availability.

3. Experiment

To forecast the number of gold medals and total medals for the 2028 Los Angeles Olympics, a classification and regression-based framework was implemented, categorizing countries into traditional sports powerhouses, emerging sports powerhouses, and potential countries based on quantitative indicators Each category utilized distinct classification criteria, regression models, and feature sets tailored to their unique performance profiles The framework was further extended to include prediction interval estimation and model evaluation to ensure robust and reliable predictions Historical Olympic data were used for training and validation, ensuring the models' applicability across diverse performance patterns[6].

Traditional sports powerhouses were identified by calculating the cumulative sum of medals and gold medals across all Olympic Games Countries were ranked based on this total, with the top-ranked nations classified as traditional powerhouses A Gradient Boosting Regression Tree (GBRT) model was employed to predict medal counts, iteratively fitting residuals to minimize prediction errors The model is expressed as:

$$y = F(x_1, x_2, , x_{k+3}) + \varepsilon \quad (2)$$

The feature set included the mean, median, and standard deviation of medals (gold and total) from previous Olympic Games, capturing historical performance trends Additional features incorporated the historical medal growth rate, a binary indicator for host country status, and the number of Olympic events, reflecting the scope of competition

Emerging sports powerhouses were classified using two quantitative indicators: the sum of medals won in the last five Olympic Games and the average medal growth rate over the same period Thresholds were applied to these metrics to select top-ranked countries The prediction model also utilized GBRT, expressed as:

$$y = F(x_1, x_2) + \varepsilon \quad (3)$$

The features included the total number of medals won in the last five Olympic Games and the average medal growth rate, focusing on recent performance and growth trends to reflect the dynamic nature of these countries

Potential countries were identified based on three quantitative indicators: a binary indicator of historical medal wins, the decay rate of medals over the last five Olympic Games, and the standard deviation of medal counts to measure performance stability Thresholds were applied to these indicators to select potential countries A Random Forest Regression model was used, aggregating predictions from multiple decision trees to produce a robust estimate, expressed as:

$$y = F(x_1, x_2, x_3, x_4, x_5, x_6, x_7) + \varepsilon \quad (4)$$

The feature set included the mean, median, and standard deviation of medals from the past five Olympic Games, the fluctuation level of medals, the highest number of medals won in specific events, the number of events participated in, and the host country factor, capturing both historical performance and potential in specific disciplines

Prediction intervals were estimated to quantify uncertainty in the forecasts For the Random Forest Regression model used for potential countries, prediction intervals were calculated using quantile regression to determine the upper and lower bounds of a 95% confidence interval, leveraging the distribution of predictions across multiple decision trees, as illustrated in Figure 2, which outlines the calculation principle of quantile intervals For the GBRT models applied to traditional and emerging sports powerhouses, prediction intervals were computed based on the assumption of a normal distribution, also at a 95% confidence level, using the following formulas:

$$\text{upperbound} = pv + 2 \times \sigma \quad (5)$$

$$\text{lowerbound} = pv - 2 \times \sigma \quad (6)$$

Here, pv represents the predicted value, and σ is the standard deviation of the predictions, derived from the variance across boosting iterations and adjusted for residual errors

Model evaluation was conducted using two key performance metrics: the coefficient of determination R^2 and

Mean Square Error (MSE) The R^2 metric measures the proportion of variance in the dependent variable explained by the model, with values closer to 1 indicating a better fit, calculated as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (7)$$

where y_i is the actual medal count, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual medal counts The MSE quantifies the average squared difference between actual and predicted values, with lower values indicating better predictive performance, computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (8)$$

where n is the number of observations These metrics were applied to evaluate the performance of the GBRT models for traditional and emerging sports powerhouses and the Random Forest model for potential countries The combination of tailored feature selection, robust regression models, prediction interval estimation, and comprehensive evaluation ensures that the proposed framework delivers accurate and reliable predictions for the 2028 Los Angeles Olympics

4. Results

The predictive framework for the 2028 Los Angeles Olympics generated detailed forecasts for gold and total medal counts across countries cate go rize dast raditional sports power houses, emerging sports powerhouses, and potential countries Using Gradient Boosting Regression Tree (GBRT) and Random Forest Regression models, the results are summarized in Table1, which presents the predicted total and gold medal counts for selected countries, and Table 2, which provides the 95% confidence intervals for these predictions to capture the uncertainty in the forecasts. As shown in Table 1 and Table 2.

Table 1: Predicted Medal Counts for Selected Countries

NOC	Predicted Total Medals	Predicted Gold Medals
United States	92	39
China	92	39
Great Britain	67	29
Australia	50	17
Japan	46	19

Table 2: Prediction Intervals for Selected Countries (95% Confidence Interval)

NOC	Total Upper	Total Lower	Lower Gold	Upper Gold
United States	32	92	15	39
China	32	92	15	39
Great Britain	2	67	1	29
Australia	2	50	0	17
Japan	2	46	0	19

The predicted medal counts indicate that traditional sports powerhouses, such as the United States and China, are expected to lead with 92 total medals each, including 39 gold medals Great Britain is projected to secure 67 total medals and 29 golds, followed by Australia with 50 total medals and 17 golds, and Japan with 46 total medals and 19 golds The 95% confidence inter- vals in Table2 reflect the range of possible outcomes, with wider intervals for countries like Great Britain, Australia, and Japan, indicating greater uncertainty due to variability in historical performance or event participation

To highlight countries with potential improvements in performance, the predicted change in total medal counts from the 2024 to the 2028 Olympics is visualized in Figure 1This chart illustrates national progress and decline, offering insights into emerging trends in Olympic per- formance. As shown in Figure 1and Figure 2.

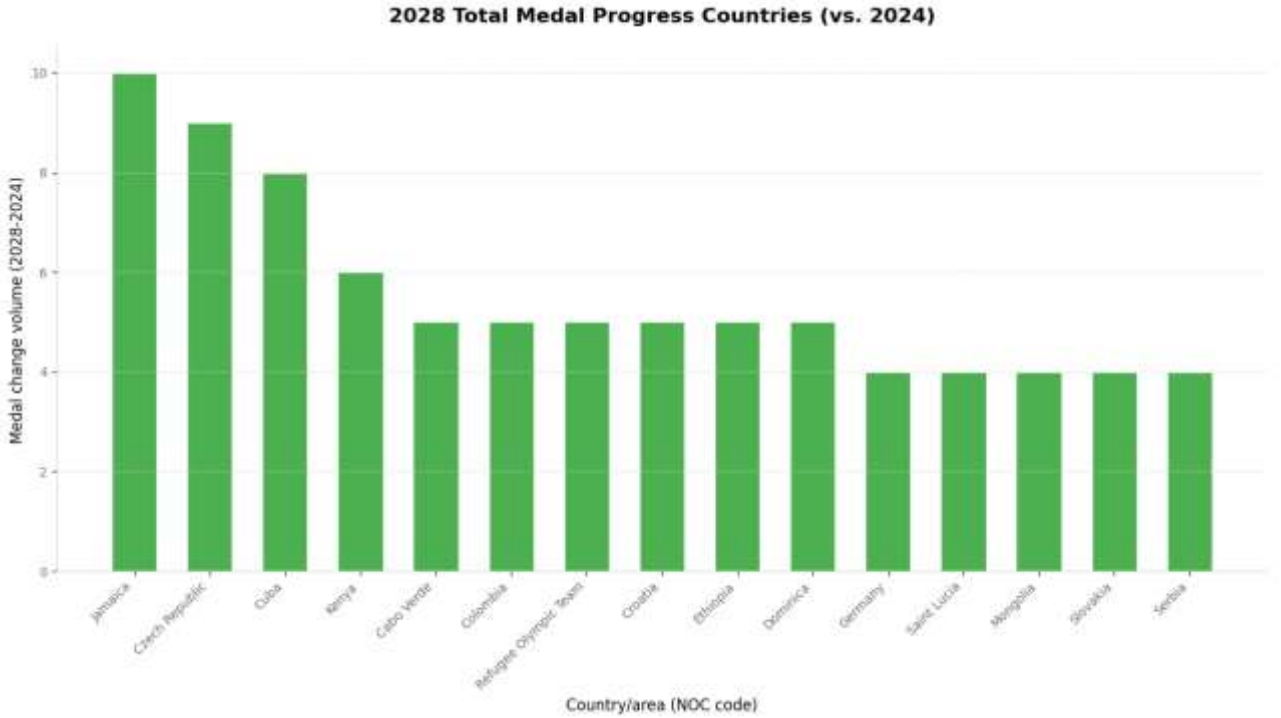

Figure 1: Analysis of National Progress

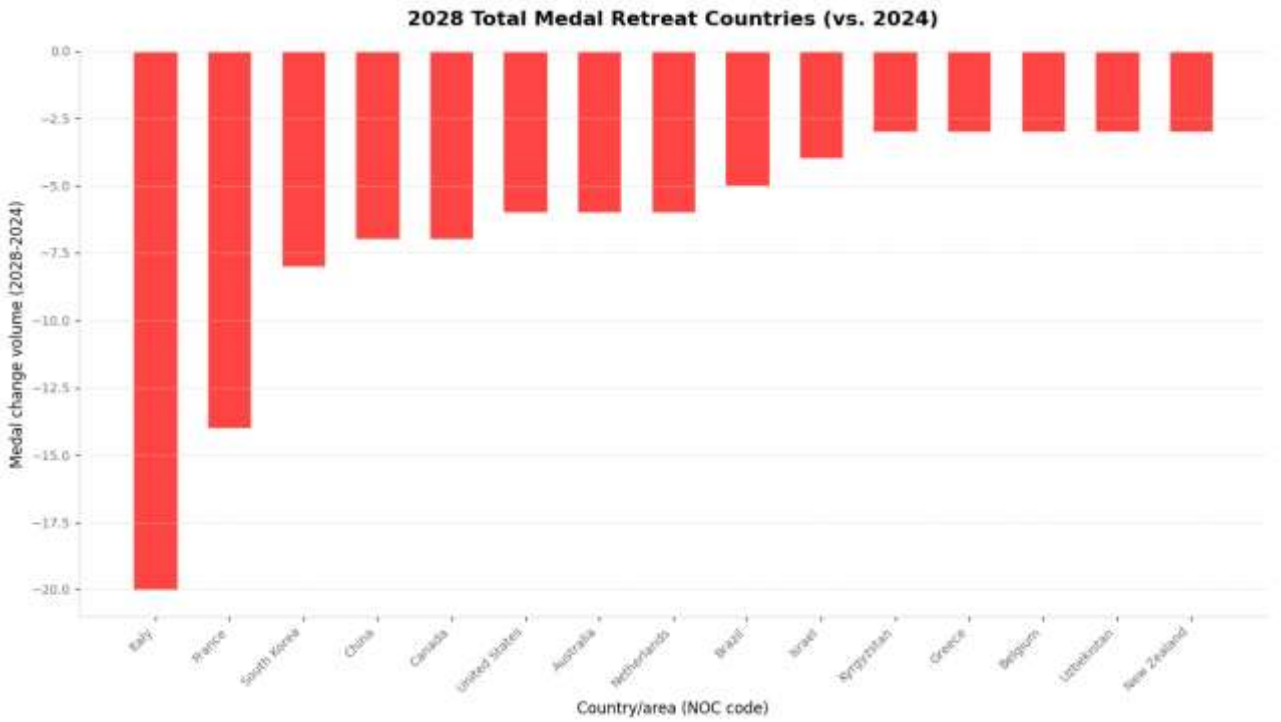

Figure 2: Analysis of National Decline

Additionally, the uncertainty in the predictions is visualized in Figure 3, which presents the 95% confidence interval plot for medal counts across selected countries, providing a clear rep- resentation of the range within which the true medal counts are expected to fall. As shown in Figure 2.

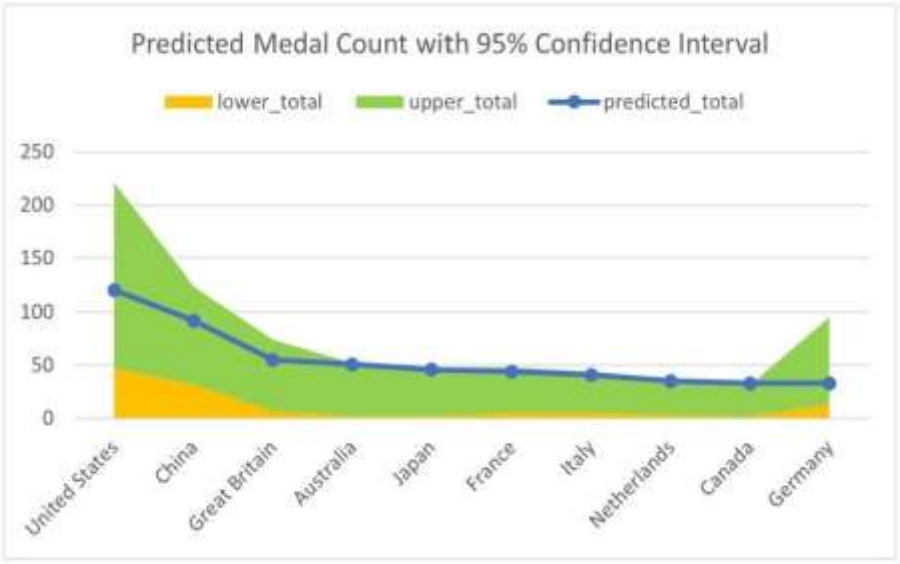

Figure 3: 95% Confidence Interval Plot of Medal Counts

Model performance was evaluated using three key metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R2) The results are presented in Table 3

Table 3: Model Evaluation Metrics

Metric	Value
Mean Squared Error (MSE)	242
Mean Absolute Error (MAE)	248
Coefficient of Determination (R2)	065

The MSE of 242 indicates the average squared difference between predicted and actual medal counts, suggesting reasonable predictive accuracy The MAE of 248 reflects a low average abso- lute deviation between predicted and actual values The R2value of 065 demonstrates that the models explain approximately 65% of the variability in gold medal counts, indicating a generally good fit to the data These results validate the effectiveness of the tailored feature selection and regression models in capturing the dynamics of Olympic performance for the 2028 Los Angeles Olympics

5. Conclusions

The proposed framework effectively predicts medal counts for the 2028 Los Angeles Olympics by classifying countries into traditional, emerging, and potential sports powerhouses. Using Gradient Boosting Regression Tree (GBRT) for traditional and emerging powerhouses and Random Forest

Regression for potential countries, the models achieved an R^2 of 0.65, MSE of 24.2, and MAE of 24.8. The United States and China are projected to lead with 92 total medals and 39 golds each, followed by Great Britain, Australia, and Japan. Key predictors include historical trends, event participation, host status, and the "star coach" effect, though more data is needed to refine this factor. However, the framework has limitations. It relies heavily on historical data, which may not fully reflect recent changes in performance or unexpected shifts like athlete injuries. Additionally, the "star coach" effect remains difficult to quantify due to incomplete coaching data. Real-time variables and more dynamic metrics could improve prediction accuracy.

Future work could involve incorporating real-time data such as athlete performance or injuries. Integrating deep learning models like Long Short-Term Memory (LSTM) could better capture temporal trends. Expanding the dataset to include more granular factors like athlete-specific data and coaching effects would further improve accuracy and robustness, providing a more reliable tool for NOCs in planning and resource allocation.

References

[1] Luo Yubo, Cheng Yanfang, Li Mengyao et al. Prediction of China's Medal Count and Overall Strength at the Beijing Winter Olympics —— Based on the Host Effect and Grey Prediction Model [J]. Contemporary Sports Science & Technology, 2022,12(21):183-186.

[2] Lu Z, Li S, Sun J. Prediction of Olympic Medal Based on Multiple Linear Regression and Logistic Regression[J]. Frontiers in Computing and Intelligent Systems, 2025, 12(1): 17-21.

[3] Jin Q, Yao R. PREDICTION STUDY OF 2028 OLYMPIC MEDAL TABLE BASED ON WEIGHTED FUSION MODELING[J]. World Journal of Management Science, 2025, 3(1).

[4] Yan D. OLYMPIC MEDAL PREDICTION AND ANALYSIS BASED ON LSTM AND TOPSIS MODELS[J]. Journal of Computer Science and Electrical Engineering, 2025, 7(3).

[5] Christoph S, L. S S, Dominik S, et al. Forecasting the Olympic medal distribution – A socioeconomic machine learning model[J]. Technological Forecasting & Social Change, 2022, 175

[6] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? —— From the perspective of explainable machine learning [J]. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.