World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3099

PREDICTING OLYMPIC MEDALS: UNVEILING THE INFLUENCES

ChaoYin Liu, JinLing Chen, Jin Lu*
School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou 510000, Guangdong, China.
**Corresponding Author: Jin Lu*

Abstract: This study focuses on the prediction of Olympic medal data and the quantitative analysis of its influencing factors, aiming to address the complexity and nonlinearity in performance data. The research primarily employs the Autoregressive Integrated Moving Average (ARIMA) model to forecast time-series trends in medal counts, capturing the inherent periodicity of medal acquisition. Furthermore, by integrating linear regression with the Gradient Boosting Tree (GBT) model, the study quantifies correlations among event structures, medal attainment, and the host effect. The primary innovation lies in establishing a multi-level comprehensive model that overcomes the limitations of traditional methods in handling time-series dependence. The research findings provide quantitative support for predicting future Olympic trends, offer a research reference for quantifying the host country's advantages, and provide a new perspective for sports policy-making and related academic research.
Keywords: Olympic medals; ARIMA; GBT; Host effect

1 INTRODUCTION

As the world's largest international sports event, the Olympic Games not only showcase the competitive levels and sports systems of athletes from various countries, but also the rankings on the medal table directly reflect a country's overall strength and international image [1]. Therefore, this quantitative indicator has received extensive attention worldwide. Based on a series of research hotspots of the Olympic Games in recent years, we have found that people are particularly concerned about the specific presentation of the number of MEDALS in the next Olympic Games and the quantitative confirmation of the host effect. However, due to the strong nonlinearity and complexity of Olympic data [2], the industry has not yet formed a unified and convincing model or mechanism.

At present, there are three mainstream methods for predicting Olympic MEDALS in the industry. The first one is to use ensemble learning, especially gradient boosting tree models, for prediction, such as the XGBoost model [3]. This method can often effectively handle multi-dimensional features to obtain accurate predicted values. The second type is to explore by using Neural networks, especially the optimized BP Neural Network, which is suitable for mining deep nonlinear relationships in the data. For example, the BP neural network model is used for medal prediction [4]. The third type is classical econometrics and regression models. Due to their relatively low accuracy, these models are often used as benchmark models for the interpretability analysis of socio-economic factors. For instance, a study on the prediction and analysis of MEDALS for the 2028 Olympic Games using multiple linear regression models in 2025 [5]. The host effect represented by a significant increase in the number of MEDALS won by the host country has long been a common perception, but this phenomenon still lacks significant quantitative confirmation. At present, some researchers have preliminarily confirmed the medal enhancement effect brought by the host effect by modifying the panel Tobit estimator using the Mundlak transform [6]. Meanwhile, Urban T L also proposed the quantifiability of the host effect in his 2025 theory [7].

This study focuses on the prediction of Olympic medal data and the quantitative analysis of its influencing factors, aiming to address the inherent complexity and significant nonlinear characteristics of Olympic performance data.The research work primarily revolves around two core objectives: Firstly, employing the Autoregressive Integrated Moving Average (ARIMA) model to accurately forecast the time-series trends in medal counts for participating nations, thus capturing the inherent "inertia" and periodic characteristics of medal acquisition over time [8]. Secondly, utilizing a combined methodological approach of linear regression and the Gradient Boosting Tree (GBT) model, the study meticulously analyzes the quantitative correlations among event program structures, medal attainment, and the Host Effect. The primary innovation of this research lies in establishing a multi-level, strongly correlated comprehensive model that effectively overcomes the limitations of traditional prediction methods in handling time-series dependence. This research aims to go beyond the traditional research direction , hoping to provide some reference ideas and trial and error for the subsequent research of personnel. Recently, reinforcement-learning techniques have begun to be applied to Olympic-medal analysis [9]; We will continue to integrate new methods in our ongoing work on this topic.

2 MODEL

2.1 ARIMA Model

2.1.1 Model introduction

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

The autoregressive moving average model, denoted as (p, d, q), is an important linear and parametric prediction model in the field of time series analysis and is widely used to capture trends, seasonality and randomness in sequence data. The core of this model lies in its three components: autoregression (AR, p order), difference (I, d order), and moving average (MA, q order). Among them, the difference term is the key for the model to handle non-stationary time series. The original sequence is transformed into a new sequence that satisfies the stationarity assumption through D-order difference. The autoregressive term (p) aims to utilize the linear dependency between historical observations and their current values; The moving average term (q) is used to model the linear combination relationship between the current value and the historical prediction error term. The ARIMA model, by effectively integrating these three mechanisms, can comprehensively describe the internal structure and dynamic characteristics of univariate time series, and is a powerful tool for modeling, parameter estimation and short-term prediction following the Box-Jenkins methodology [10].

Differential process (I): The ARIMA model requires that the sequences it processes must be stationary sequences. For the original non-stationary time series Y_t, it needs to be transformed into a stationary sequence W_t through an integrated process. The order of difference d is the minimum number of differences required to make the sequence stationary. The first-order difference is defined as $Y_t - Y_{t-1}$, while the d -order difference can be expressed in the form of the lag operator L:

$$\Delta^{\mathrm{d}}\mathrm{Y_t} = (1-\mathrm{L})^{\mathrm{d}}\mathrm{Y_t} \tag{1}$$

Autoregressive process (AR): describes the linear relationship between the current value W_t and its past p observations. This process is usually expressed in terms of the lag operator polynomial $\Phi(L)$:

$$\Phi(L)W_t = \delta + \epsilon_t \tag{2}$$

Here, $\Phi(L) = 1 - \phi_1 L - \cdots - \phi_P L^P$.

The moving average process (MA): describes the linear relationship between the current value W_t and the current and past q white noise error terms ϵ_t. These two processes are usually expressed by the lag operator polynomials $\Phi(L)$ and $\Theta(L)$:

$$W_t = \delta + \Theta(L)\epsilon_t \tag{3}$$

Here, $\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$.

The ARIMA model combines the above three processes to form the final comprehensive expression, which describes the dynamic characteristics of the original sequence Y_t after a d order difference:

$$\Phi(L)(1-\mathrm{L})^{\mathrm{d}}\mathrm{Y_t} = \Theta(L)\epsilon_t \tag{4}$$

The advantages of the ARIMA model in prediction lie in its strong adaptability, simple and understandable model, high prediction accuracy, low data requirements, reliable prediction results, easy interpretation, high computational efficiency, good scalability, and wide application support. These characteristics enable the ARIMA model to effectively capture trends in time series, be applicable to various types of data, and ensure the accuracy and reliability of predictions through statistical tests and model diagnostics.

2.1.2 Model construction

To predict the changes in a country's performance at the 2028 Olympics, first, collect the medal count data of various countries from previous Olympic Games and organize it in a time series format. Then, through trend and periodic analysis, identify the trends and periodic patterns in the medal count changes. Next, analyze the influencing factors such as sports policies, training systems, and financial investments, and assess their impact on the medal count. Finally, time series analysis methods, such as using the ARIMA model to model and predict historical data, can be adopted to capture the trends and periodic changes in historical data, and thereby predict the medal counts of various countries in 2028.

First, we carried out data preprocessing. We standardized the data, sorting the medal data of each country by year and establishing a separate time series model for each country. The ARIMA model requires the input data to be stationary. Therefore, we differenced some non-stationary data to make it stationary. The following is the difference formula:

First differencing:

$$\Delta y_t = y_t - y_{t-1} \tag{5}$$

Second differencing:

$$\begin{aligned} \Delta^2 y_t &= \Delta y_t - \Delta y_{t-1} \\ \Delta^2 y_t &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ \Delta^2 y_t &= y_t - 2y_{t-1} + y_{t-2} \end{aligned} \tag{6}$$

We fitted an ARIMA model (AutoRegressive Integrated Moving Average), choosing the number of gold medals, silver medals, bronze medals and total medals as the target variables for modeling. The medal data of each country should be sorted by year and input into the model as a time series.

In the fitting of the function, we adopt the maximum likelihood estimation (MLE) or minimize the AIC (Akaike Information Criterion) to determine the optimal parameters: the order of differentiability (d), the order of MA (p), and the order of AR (q), thereby predicting the performance changes of the country in the 2028 Olympic Games.

After performing a second-order difference, the result was not much different from the first-order difference. Therefore, we can set d to 1. After the data is made stationary through a first-order difference, we can continue to use the ACF and PACF graphs to assist in determining the p and q parameters in the ARIMA model. The detailed exploration process of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots is presented in the model results section below.

2.2 GBT MODEL

2.2.1 Model introduction

Gradient boosting tree is a powerful ensemble learning model that adopts an additive model structure and is optimized through the strategy of gradient descent in the function space. The core mechanism is to iteratively train a series of weak learners, enabling each new tree to focus on improving the shortcomings of the previous round of the model.

GBT constructs the final prediction function as the weighted sum of M decision trees. The learning objective of the model is to find an optimal additive function to minimize a predefined differentiable loss function. This kind of optimization is different from the traditional algorithms that are carried out in the parameter space. Instead, it iterates in the function space, that is, at each step, an optimal new function is found and added to the existing model.

In each iteration, the GBT algorithm first evaluates the performance of the current integrated model $\widehat{F_{m-1}}(x)$. To determine the optimization direction of the next weak learner $h_m(x)$, the model calculates the negative gradient of the loss function relative to the current prediction result. In regression tasks, this negative gradient is precisely what we often refer to as the residual, and thus it is also called a pseudo-residual. This pseudo-residual precisely indicates the extent and direction to which the current model "makes mistakes" at each data point. The training objective of the newly generated decision tree $h_m(x)$ is to fully fit these pseudo-residuals, thereby accurately capturing and correcting the errors of the existing model.

After the new decision tree $h_m(x)$ is trained, the model needs to determine an optimal step size γ_m. This step size is determined through Line Search, with the aim of maintaining the directionality of $h_m(x)$ while ensuring that the overall loss function can be minimized when it is added to the existing model. Ultimately, the model is updated in a regularized manner, introducing a key parameter ($\mu \in (0, 1]$):

$$\widehat{F_m}(x) = \widehat{F_{m-1}}(x) + \mu * (step\ size) * h_m(x) \tag{7}$$

The learning rate μ limits the contribution of each tree by reducing the small step size, which greatly enhances the generalization ability and robustness of the model and is an important means to prevent overfitting. Through this gradient descent iteration, GBT gradually increases and eventually converges into a high-precision prediction model.

2.2.2 Model construction

To measure the intrinsic correlation between event selection and the total number of medals, constructing a linear regression model is a suitable choice. This model effectively extracts the core features of the relationship between the two, simplifying the understanding and analysis of this complex phenomenon, and providing a clearer theoretical framework and quantitative basis for further research. We extracted the number of events and types of events for each Olympic year as feature variables $X1$ and $X2$, and set the total number of medals as the target function y. The parameters θ0, θ1, and θ2 are the regression coefficients corresponding to each feature vector.

Based on the above data analysis, we derived the linear regression model expression:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \tag{8}$$

The loss function for the linear regression model is the mean squared error, which is expressed as:

$$L(\theta) = \frac{1}{2m}\sum_{i-1}^{m} (y^{(i)} - (\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}))^2 \tag{9}$$

Next, we used gradient descent to minimize the model's mean squared error. In each iteration, the parameters θ are updated based on the gradient (partial derivatives) of the loss function with respect to θ. The update formula is as follows (α is the learning rate, and θ_j represents θ_1 or θ_2):

$$\theta_j := \theta_j - \alpha \frac{\partial L(\theta)}{\partial \theta_j} \tag{10}$$

By iteratively updating the parameters θ of the linear regression model, we obtained the final regression coefficients.

To investigate the host country effect, we found that using a gradient boosting tree model is a better choice. This model has strong feature selection capabilities and can continuously select split features to maximize information gain during the construction process, which helps identify which features (such as event quantity, types, etc.) have the greatest impact on the host country's medal count.

First, an initial model $F0(x)$ is chosen, where we use the mean of the training set as the initial prediction value:

$$F_0(x) = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{11}$$

Second, iterative training is performed. The residuals are calculated first ($r_i^{(m)}$represents the residuals):

$$r_i^{(m)} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{12}$$

Third, the residuals are used as the target to train a new decision tree, and the newly trained tree is added to the current model to update the model's prediction values ($\gamma_m$is the learning rate):

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{13}$$

Finally, after M iterations, the final model is obtained:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x) \tag{14}$$

3 RESULTS AND ANALYSIS

3.1 Analysis of ARIMA Model Prediction Results

Following the model setup described above, we first construct the specific Autocorrelation Function (ACF) and Partial

Autocorrelation Function (PACF) plots to facilitate subsequent model application. the ACF and PACF graphs is shown in Figure 1.

ACF plot: After lag 1, the autocorrelation coefficients rapidly decay and fluctuate within the confidence interval, which typically indicates that there are no significant autoregressive components in the data, or that the order of the autoregressive components is low.

The PACF plot: The partial autocorrelation coefficient is significant at lag 1, then rapidly decays and fluctuates within the confidence interval. This typically indicates that there is a significant autoregressive component in the data.

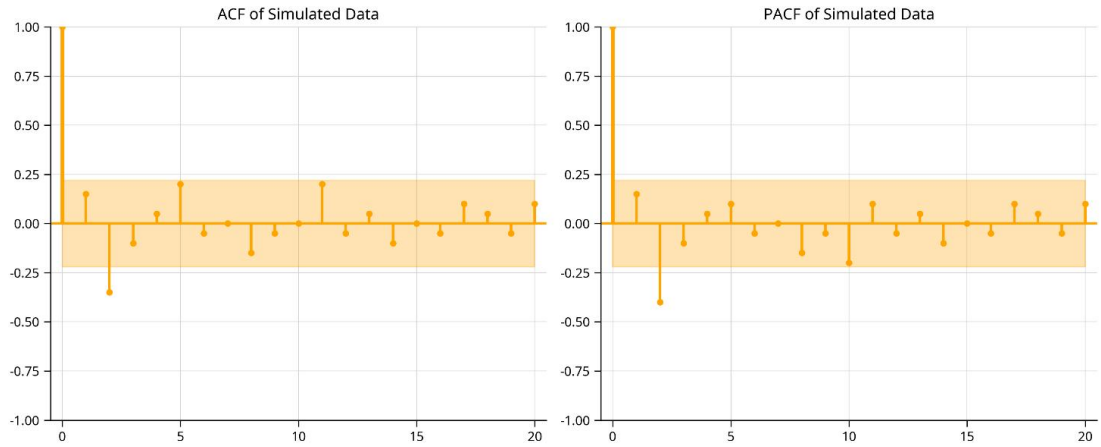
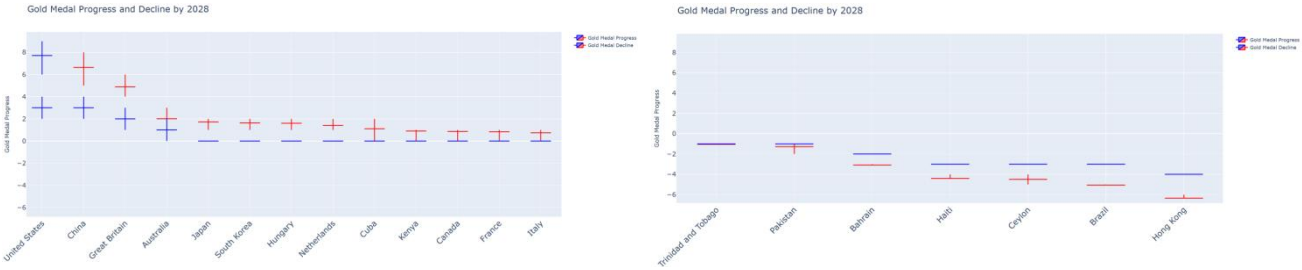


Figure 1 ACF and PACF Graphs of the Simulated Data

Since the PACF plot decays rapidly after lag 1, this indicates that there might be a significant autoregressive component. Therefore, a possible value for p is 1. This means that the model can consider an AR (1) component, that is, there is a linear relationship between the current value and the previous value. The ACF plot does not show significant autocorrelation after lag 1, suggesting that a significant moving average component might not be necessary. However, as the ACF plot has a significant peak at lag 1, this might indicate the presence of an MA (1) component, that is, there is a linear relationship between the current value and the previous error term. Therefore, a possible value for q is also 1.

Based on the above analysis, a possible ARIMA model is ARIMA (1,1,1). This model assumes that there is a significant autoregressive component and a significant moving average component in the data, and that the data has been made stationary through first-order differencing.

Firstly, in order to predict the possible progress or regression of each country, we should apply the ARIMA (1,1,1) model just constructed above. We need to read the preprocessed medal data set of each country into the model and verify the model to obtain the residual as white noise. Visualize and output the above aggregated data to represent and reveal the relationship between the event project settings and the number of medals. Prediction of the upward or downward trend of medal counts for all countries is shown in Figure 2.



(a) (b)

Figure 2 Prediction of the Upward or Downward Trend of Medal Counts for All Countries

From this, we can sort out the top 10 countries that are most likely to make progress or decline in the overall medal table at the 2028 Los Angeles Olympics, and visualize the data to obtain the relevant bar charts. the charts is shown in Figure 3.

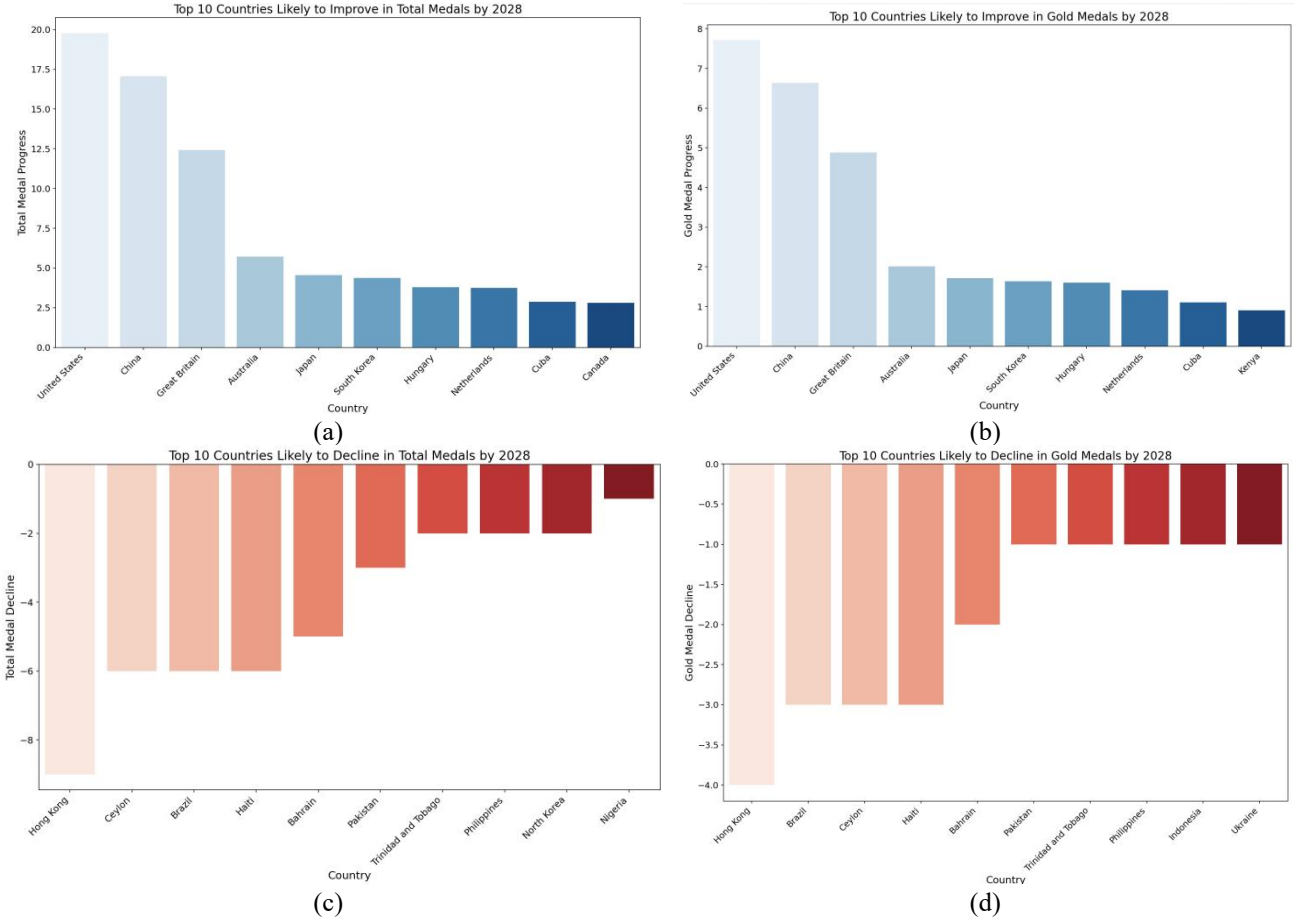


Figure 3 The Top 10 Countries Where the Number of MEDALS is most likely to Increase or Decrease

Based on the same train of thought, we only input the number of gold medals of each country into the model, perform differencing, and use ACF and PACF graphs to obtain key parameters. After visualizing the results and organizing them, we have identified the 10 countries that are most likely to improve and decline in the gold medal tally at the 2028 Los Angeles Olympics.

3.2 GBT Model Results and Practical Significance

To explore the relationship between event selection and medal counts, we first applied the previously constructed linear regression model. The preprocessed Olympic event dataset was fed into the model, and through event aggregation, we obtained the event quantity and type for each Olympic Games. Subsequently, we imported and processed the Olympic medal dataset to count the total medal count for each Games. Finally, by integrating features from the event, medal, and athlete datasets, we successfully predicted the medal counts using the gradient boosting tree model. The visualization and output of the results revealed the intrinsic relationship between event selection and medal counts. The relationship curve is shown in Figure 4.

Figure 4 The Relationship between the Project Setting and the Number of Medals (1896-2024)

After analyzing the macro-level relationship between event selection and medal counts, we focused on the specific

impact on countries. To investigate the different strong events in different countries and reveal their underlying causes, we considered various factors and conducted strict preprocessing of the data. By analyzing the event weights of several representative countries, we obtained the following event distribution charts for four countries. The distribution of project importance among the four countries is shown in Figure 5 and Figure 6.

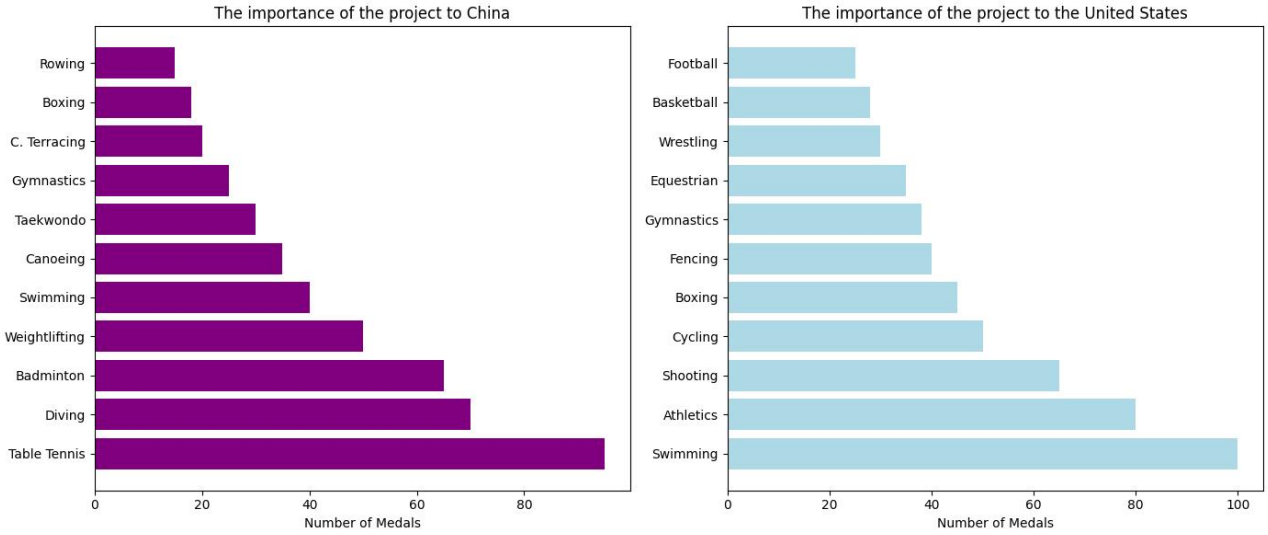


Figure 5 The Ranking of Project Importance in China and the United States

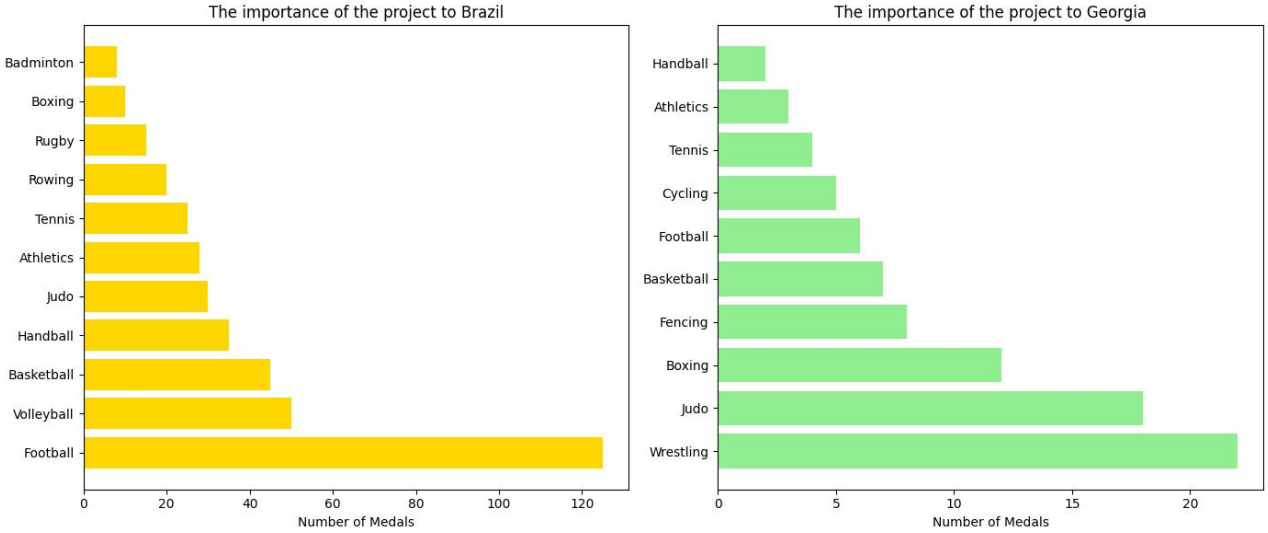


Figure 6 The Ranking of Project Importance in Brazil and Georgia

Model results indicate that for traditional sports powerhouses, a higher number of participating events correlates positively with gold medal counts, and long-established, stable events (e.g., swimming and athletics) have a significant positive impact on total medals. Conversely, newly introduced events (such as 3x3 basketball and rock climbing) contribute relatively little to the medal distribution of these powerhouses. Furthermore, we observed an inertia effect in the historical medal counts of countries with smaller economic scales, where past successes tend to carry over into future competitions. In these nations, medal counts from strong events are significantly higher than others, underscoring the critical importance of sustained investment in specialized, high-performing events.

Using the gradient boosting tree model, we predicted the medal growth rates for host countries. By comparing the medal growth rates during the host years and the four Olympic Games before and after, we excluded the influence of the announcement timing for the host city. As shown in Figure 7, host nations consistently win more medals in their hosting Games than in the four-Olympiad window centred on that event; limiting the comparison to this four-year span mitigates confounding from longer-term changes in athlete cohorts and national sports investment.

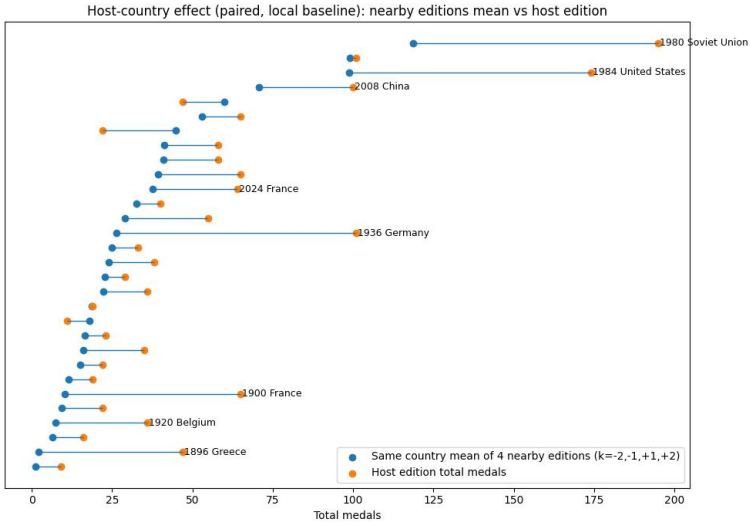


Figure 7 Host-country Medal Counts: 4 years before vs. 4 after

Figure 8 shows that, after standardizing medal counts across the three Games before and after the host edition, the host-nation effect yields a 2.4-fold increase over the baseline, indicating a statistically significant contribution to medal growth.

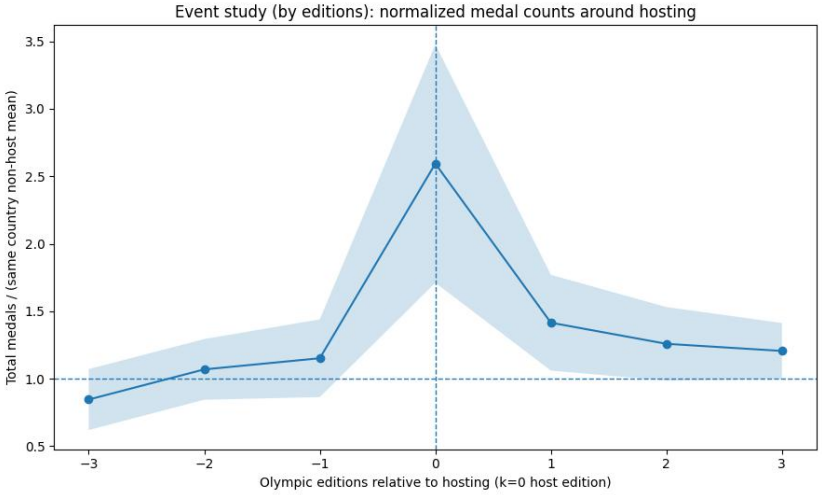


Figure 8 Standardized Line Graph of Medal Counts for the Host Edition and its Proximate Editions

4 CONCLUSIONS AND OUTLOOKS

This study developed a multi-level comprehensive modeling framework to address the complexity of Olympic medal prediction. By employing the ARIMA (1,1,1) model, we successfully captured the "inertia" and periodic trends of historical performance to forecast the 2028 Los Angeles Olympics. Simultaneously, the integration of Linear Regression and Gradient Boosting Tree (GBT) models allowed for a precise quantification of the "Host Effect"—revealing a 2.4-fold growth baseline—and the identification of key event-performance drivers. The primary innovation lies in this hybrid approach, which overcomes the limitations of traditional single-method models by simultaneously handling time-series dependencies and high-dimensional non-linear features. Despite these results, the current model lacks integration of macro-external factors such as national GDP and sports investment policies. Future research will focus on incorporating these socio-economic indicators and exploring advanced deep learning architectures, like LSTM networks, to further enhance the model's robustness and predictive accuracy against the volatile nature of international sports.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] LeBlanc P J. Sweeping The Soft Power Podium: a Quantitative and Qualitative Analysis of Olympic Soft Power's Impact on The Host Nation’s International Image. Monterey, CA; Naval Postgraduate School, 2021.

[2] Zhang T L, Chen J J, Cheng C, et al. Nonlinear Prediction Based on 2028 Olympic Events and Medals. World Journal of Information Technology, 2025, 3(4): 51-57. DOI: 10.61784/wjit3053.
[3] Chen X, Sun X, Zheng Q, et al. Olympic medal prediction model based on XGBoost. Proceedings of the 2025 5th International Conference on Applied Mathematics, Modelling and Intelligent Computing. 2025: 406-413.
[4] Li N, Li J, Fang H, et al. A Hybrid Intelligent Model for Olympic Medal Prediction Based on Data-Intelligence Fusion. Technologies, 2025, 13(6): 250.
[5] Zheng X, Liu Y, Zhang T. Unlocking Olympic Success: Predictive Modeling and Strategic Insights for the 2028 Games. Advances in Engineering Technology Research, 2025, 14(1): 1045-1045.
[6] Rewilak J. The (non) determinants of Olympic success. Journal of sports economics, 2021, 22(5): 546-570.
[7] Urban T L. Does the Olympic home advantage extend beyond a country's borders?. Managing Sport and Leisure, 2025: 1-9.
[8] Jiao Y. Medal Prediction Analysis Based on ARIMA Time Series Model with Random Forests. 2025 IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA). IEEE, 2025: 1301-1306.
[9] Wang Y, Wang J, Huang T Y, et al. STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization. arXiv preprint, 2025. DOI: 10.48550/arXiv.2501.17711.
[10] Stellwagen E, Tashman L. ARIMA: The Models of Box and Jenkins. Foresight: The International Journal of Applied Forecasting, 2013(30).