Prediction for Olympic Medal Tables by Machine Learning

Jianzhang Li [1, a], Yueran Zhang [2, b]

[1] Applied Mathematics, School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

[2] Computer Science and Technology, School of Advanced and Technology, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

[a] Jianzhang.Li23@student.xjtlu.edu.cn, [b] Yueran.Zhang22@student.xjtlu.edu.cn

Abstract. The goal of this article is to predict the gold and overall medal rankings for 2028. ARIMA is used to analyze cyclical fluctuations, followed by XGBoost optimization for time-series predictions. The model forecasts gold medal standings and visualizes the top 20 countries. Further, this paper pay attention to predicting the likelihood of countries without Olympic gold medals winning their first in 2028. Our study create binary classification labels and countries that have never won gold are identified and relevant features are extracted. SVM is employed for classification, with an AUC score close to 1, indicating high accuracy. Eventually, the relationship between event selection and medal counts is analyzed. The data is preprocessed and event types are converted into categorical variables. Random Forest regression is used, revealing that host country event selection affects medal performance but has minimal impact on overall rankings. The model's performance is validated using MAE, MSE, and other metrics.

Keywords: ARIMA Model; XGBoost; SVM; Random Forest.

1. Introduction

The Olympic Games are a significant global sporting event that facilitates cultural exchange and international understanding. Audiences not only focus on individual competitions but also the overall medal table, which has become a point of interest, especially regarding gold medals and total counts. As the Games near their conclusion, fluctuations in medal standings generate discussion about potential last-minute upsets or historic breakthroughs. In the 2024 Paris Summer Olympics, countries like Albania, Cape Verde, Dominica, and Saint Lucia achieved their first Olympic medals, while over 60 countries remain without any. This underscores the need for prediction models to forecast future Olympic medal standings. Bernard and Busse examined the link between Olympic medal counts and factors like population and GDP, finding that larger populations and higher GDPs correlate with more medals, while host countries and resource-rich nations, such as the Soviet Union, also have advantages [1]. They used these models to predict the 2000 Sydney Olympics medal counts. Guo and Zhao applied the GM(1,1) grey theory model, based on data from the last six Olympic Games, to predict the 2016 Rio Olympics gold medal standings [2]. Their model predicted that China would win 55 gold medals, with the UK and US securing 48 and 41 respectively, and confirmed its accuracy through residual tests. Shiyu Wang combined the methods of Bernard, Busse, Guo and Zhao, integrating multivariate nonlinear regression with BP neural networks to forecast medal counts [3]. This approach produced predictions similar to the regression model after training and testing.

The main contributions of this study can be summarized as: 1) combines ARIMA time series analysis with machine learning methods (XGBoost, SVM, and Random Forest) to enhance the accuracy of medal predictions for the 2028 Olympics; 2) Analyzes the impact of event setups and athlete performance on medal counts while optimizing the model to avoid overfitting; 3) Provides insights for countries to make informed decisions regarding athletic investments and strategies for future Olympics.

The structure of this paper is as follows: the first part is the introduction, which presents the research background, the current status of Olympic medal predictions, and key contributions of the study; the second part is the methodology, detailing the ARIMA model, XGBoost, Support Vector

Machine (SVM), and Random Forest techniques employed for prediction; the third part focuses on results and analysis, showcasing the predictions for the 2028 Los Angeles Olympics, along with visualizations and evaluations of model performance; the fourth part discusses the implications of the findings, reviewing the insights gained from the predictions for emerging countries and the significance of various factors influencing medal counts; and the fifth part concludes the paper, emphasizing the overall contributions and future directions for research in Olympic medal forecasting.

2. Related Theories

ARIMA, or Auto-Regressive Integrated Moving Average, is a widely used statistical method for time series analysis and forecasting. ARIMA is well-suited for univariate time series data, particularly when trends and seasonality are present after transformation. The model is defined by three parameters: (p, d, q), where p corresponds to lag observations, d indicates the degree of differencing, and q refers to the moving average window size. Its effectiveness in capturing temporal structures makes it a staple in forecasting applications.

XGBoost, which stands for Extreme Gradient Boosting, is an advanced implementation of the gradient boosting framework that excels in speed and performance, particularly in supervised learning tasks such as classification and regression. The core idea behind XGBoost is to build models in a stage-wise manner, sequentially combining multiple weak learners, typically decision trees, to create a more robust predictive model. It employs gradient boosting to optimize a loss function through gradient descent.

Support Vector Machine (SVM) is another powerful supervised learning algorithm primarily used for classification, though it can also handle regression problems. The fundamental principle of SVM is to find the optimal hyperplane that separates different classes within the feature space. This separation is achieved by maximizing the margin between the hyperplane and the nearest data points from each class, known as support vectors. A notable feature of SVM is its use of the kernel trick, which allows it to project data into higher dimensions to handle non-linear separations without explicitly calculating the coordinates.

Random Forest is an ensemble learning technique that builds a multitude of decision trees during training and outputs the mode of their predictions for classification tasks or the average for regression tasks. The algorithm employs bagging, or bootstrapped aggregating, to create diverse models using random subsets of the training data. Each decision tree is constructed based not only on these random samples but also on randomly selected subsets of features at each split, which enhances diversity and improves performance.

In conclusion, ARIMA, XGBoost, SVM, and Random Forest represent fundamental approaches to predictive analytics, each addressing unique challenges. ARIMA delivers powerful forecasts in time series analysis, while XGBoost stands out in competitive domains for its high performance and accuracy. SVM provides strong boundary-based classification capabilities, especially beneficial in high-dimensional contexts, and Random Forest excels through its ensemble approach, providing robustness and interpretability. Collectively, these algorithms equip data scientists and analysts with essential tools for tackling diverse forecasting and prediction challenges.

3. Experiments

3.1 ARIMA and XGBoost

To more accurately predict the gold and total medal counts for the 2028 Los Angeles Olympics, this study chooses to combine the ARIMA model and XGBoost model. The ARIMA model is a classic time series analysis method, particularly suitable for capturing trends and seasonal variations in data [4][5]. Since the Olympics are held every four years, medal counts are heavily infuenced by periodic and seasonal fuctuations. Thus, the ARIMA model can efectively capture this temporal

dependency and forecast future medal counts based on historical data. The basic formula for the ARIMA model is as follows:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (1)$$

In this formula, by selecting the appropriate autoregressive (AR) order p, moving average (MA) order q, and diferencing order d, the ARIMA model can fit historical medal data and capture seasonal fuctuations and periodic changes. This allows the ARIMA model to predict future medal counts based on past medal data, reflecting the trend of how medal counts change over time.

XGBoost is a powerful ensemble learning method that solves nonlinear problems through the combination of multiple decision trees. In this case, the variation in medal counts is influenced not only by historical data but also by external factors, which may have complex nonlinear relationships with each other. XGBoost excels at capturing these intricate relationships, providing supplementary information for more accurate medal count predictions [6]. The basic objective function for XGBoost is as follows:

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

In this function, the loss function measures the diference between the model's predicted values and the true values, while the regularization term controls the complexity of the model. In this way, XGBoost is able to iteratively adjust the model during the fitting process, capturing the nonlinear relationships between medal counts and external factors [7].

3.2 SVM

In this phase, our study will use Support Vector Machine (SVM) for classification prediction. The core idea of SVM is to find an optimal hyperplane (or, in the case of non-linearity, find an optimal high-dimensional space mapping) in the feature space that separates the data points into different categories [8][9]. Since the data is complex and nonlinear, this research can choose a complex kernel function, such as the Radial Basis Function (RBF) Kernel, to map the data into a higher-dimensional space to aid in classification. The specific formula and method for the RBF Kernel are as follows:

(1) linear kernel:

$$K(x_i, x_j) = x_i^T x_j, \quad (3)$$

(2) Gaussian Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}), \quad (4)$$

(3) polynomial kernel:

$$K(x_i, x_j) = (x_i^T x_j + c)^2, \quad (5)$$

After using the kernel function, the objective function of the optimization problem becomes:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \quad (6)$$

Next, this paper use the training dataset to train the SVM model, adjusting model parameters (such as the C parameter, kernel function parameters, etc.) to maximize the classification performance.

Finally, with the trained SVM model, our study predict the countries that have not won medals, and determine whether each country is likely to win a medal in the next Olympic Games. Based on the model's predictions, the probability of each country winning a medal can be calculated, and specific predictions can be made.

3.3 Random Forest

The random forest model performs regression predictions by integrating multiple decision trees, effectively handling high-dimensional data and complex nonlinear relationships. This approach helps

avoid the overfitting issues that may arise with a single model, making it particularly suitable for analyzing the impact of event setups and the influence of athletes from various countries on medal counts [10]. In the construction of decision trees, the split at each node is typically determined by either information gain (used in classification tasks) or mean squared error (MSE) (used in regression tasks). In classification tasks, our model evaluate the selection of a particular feature by calculating the information gain:

$$IG(X) = H(D) - \sum_{i} \frac{|D_i|}{|D|} H(D_i), \tag{7}$$

For regression tasks, decision trees use mean squared error (MSE) to evaluate the effectiveness of each split. The formula for calculating MSE is as follows:

$$MSE(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} (y_i - \hat{y})^2, \tag{8}$$

After training, the final prediction of the random forest is made by aggregating the results of all decision trees. In regression tasks, the predicted value is the average of the predictions from all the trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t, \tag{9}$$

For classification tasks, the prediction result is obtained through voting, where the final category is the one that appears most frequently:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}, \tag{10}$$

First, our paper divide the data into a training set and a test set, typically using 70%-80% of the data for training and the remaining portion for model validation. Then, the model is trained on the training set and continuously adjust hyperparameters (such as the number of trees, maximum depth, etc.) to ensure the model fits the data more effectively. To achieve optimal model performance, this section also employ techniques like grid search or random search for hyperparameter tuning.

4. Results and Analysis

4.1 Results and Analysis of ARIMA and XGBoost

This study predicts the gold and total medal counts for each country in the 2028 Los Angeles Olympics using historical Olympic data and the distribution of past Olympic host countries. Our study apply both the ARIMA method and XGBoost regression model for prediction and analysis.

First, this study uses the ARIMA model for time series forecasting of medal counts. For countries with fewer than 5 data points, this paper use ARIMA(1, 1, 1); for 5-10 data points, we use ARIMA(2, 1, 1); and for more than 10 points, this paper apply ARIMA(2, 1, 2). This approach allows us to make initial predictions for the 2028 Olympics. Next, our research refine these predictions using the XGBoost model, incorporating the host country distribution data. The trained XGBoost model gives detailed predictions for the gold and total medal counts for each country in the 2028 Olympics. The results are shown in Table 1:

In terms of visualization, the predicted results are displayed using two horizontal bar charts: one chart shows the predicted ranking of gold medals by country, and the other displays the predicted ranking of total medals by country. These charts provide a clear view of which countries may perform exceptionally well in the 2028 Olympics and which countries may maintain their leading positions. The results are shown in Figure 1:

Table 1. Predicted Top 20 Medal Table for the 2028 Los Angeles Olympic Games (Gold Medals)

Ranking	Country	Predicted Gold	Predicted Total
1	United States	39	124
2	China	30	92
3	Japan	23	48
4	Russia	20	54
5	Great Britain	14	64
6	France	14	59
7	Australia	14	46
8	Netherlands	14	36
9	Korea Republic	14	33
10	Italy	10	41
11	Germany	10	32
12	Uzbekistan	7	11
13	Canada	6	26
14	Brazil	6	20
15	New Zealand	6	19
16	Hungary	6	19
17	Spain	4	17
18	Kenya	3	10
19	Sweden	3	10
20	Mixed team	3	3

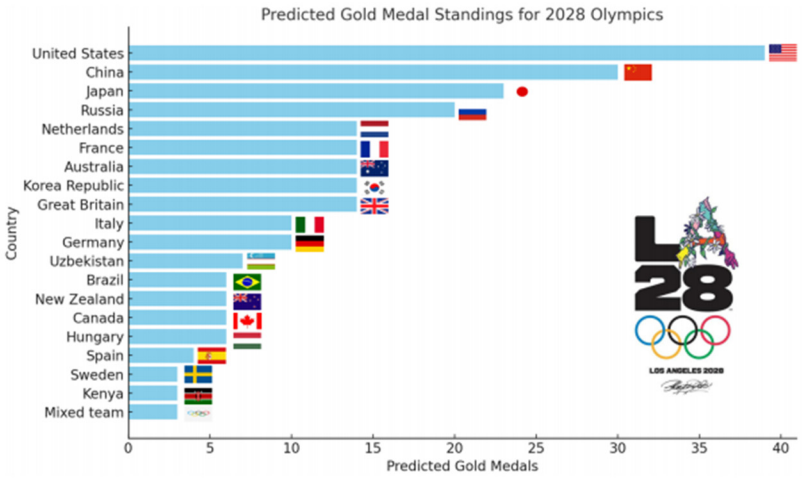

Figure 1. Predicted Gold Medal for 2028 Olympics

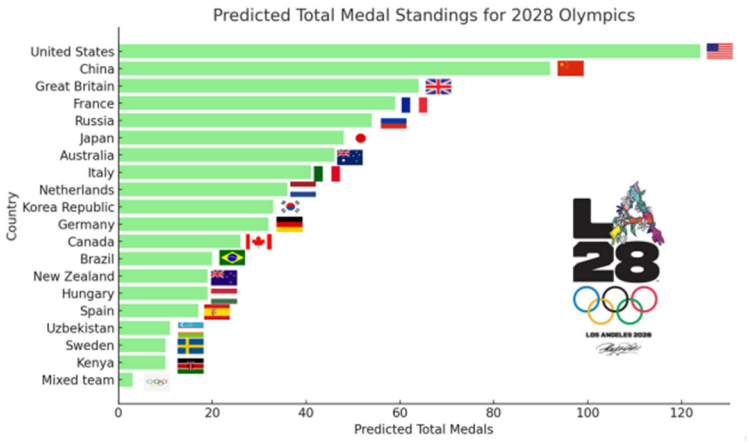

Figure 2. Predicted Total Medal for 2028 Olympics

Based on the combination of the prediction model and historical data, this paper forecasted the progress or decline in the number of gold medals and total medals for all countries in the 2028 Olympics. Our study selected the top fve countries with progress and decline and visualized the results using a stacked bar chart. The results are shown in Figure 3:

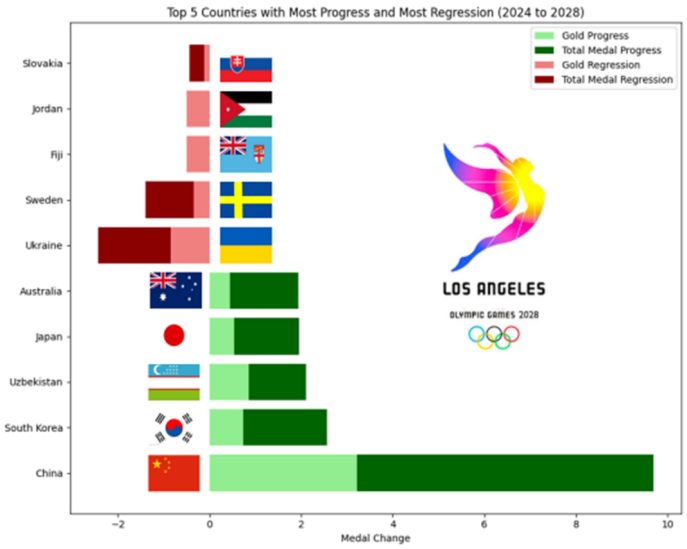

Figure 3. Top 5 Countries with Most Progress and Most Regression (2024 to 2028)

In summary, the overall trend shows that China, South Korea, Uzbekistan, and Japan have made signifcant progress in their medal counts. China, in particular, shows the highest progress, especially in the increase of total medals and gold medals. China's medal growth is substantial, far exceeding that of other countries. South Korea, Uzbekistan, and Japan also show notable medal growth, with their progress being relatively balanced. Although their progress is not as signifcant as China's, they still demonstrate positive growth, especially in total medals. On the other hand, some countries like Slovakia and Jordan are facing a decline in medal counts. Finally, this section evaluated the predictive performance of the models using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coeffcient of Determination (R^2). The computed coefcients for the gold medal ranking prediction and the total medal ranking prediction are shown in the Table 2.

Table 2. Model Evaluation Results

Model	MSE	RMSE	R^2
Gold Medal Model	0.217978213	0.466881369	0.9927223
Total Medal Model	0.528503217	0.726982267	0.9980740

4.2 Results and Analysis of SVM

First, this paper preprocess the historical Olympic data, focusing on gold medals and medal table information. Since the dataset did not include countries that had never won a medal, our study extracted relevant information such as country names, gold, silver, bronze medals, and total medals. After cleaning the data and filling missing values, our model created a binary classifcation label to indicate whether a country has won a gold medal. This process revealed 43 countries that had never won a gold medal. After excluding countries that no longer exist and those with incomplete data, this study narrowed down to 8 countries.

Next, our research applied the Support Vector Machine (SVM) model with the RBF (Radial Basis Function) kernel. Our paper optimized the model using GridSearchCV to fine-tune the penalty parameter (C) and kernel coefcient (gamma), ensuring the best prediction results. After training, the model predicted the probability of these countries winning their first gold medal in the 2028 Olympics. The predicted probabilities and rankings are shown in the Table 3 and Figure 4:

Table 3. Probability table for first-time gold medal winners at the 2028 Olympics

Country	Predicted Probability
Barbados	0.02279
Eritrea	0.02279
Cyprus	0.02279
Cabo Verde	0.02279
Burkina Faso	0.022784
Afghanistan	0.021144
Albania	0.021078
Djibouti	0.021078

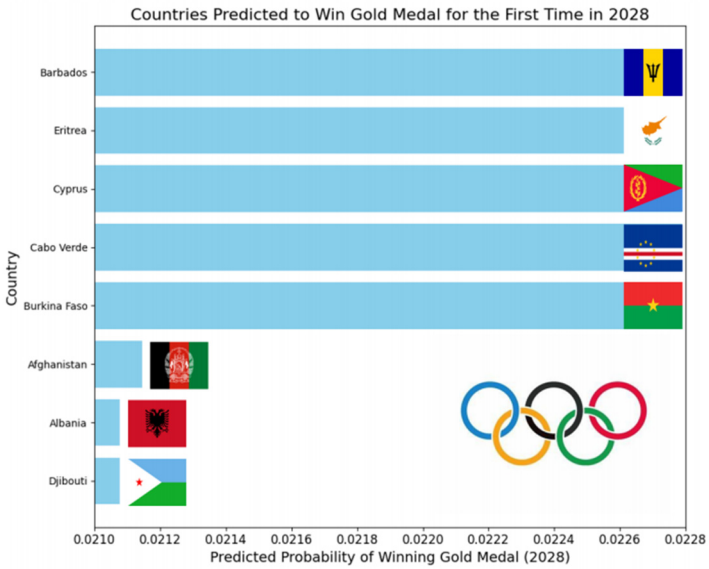

Figure 4. Countries Predicted to Win Gold Medal for the First Time in 2028

Through the chart analysis, it is found that the gold medal-winning probabilities of these countries are generally distributed between 0.0210 and 0.0228. This indicates that the probability of these countries winning their first gold medal in the current Olympics is quite low. Among them, Barbados has the highest predicted probability, which is 0.022790, though the probability is still very small. Finally, this research conducted a test on the model, and the test results are shown in Figure 5. The AUC value obtained was 0.9356246590289143, which is close to 1, proving that the performance of our model is very good.

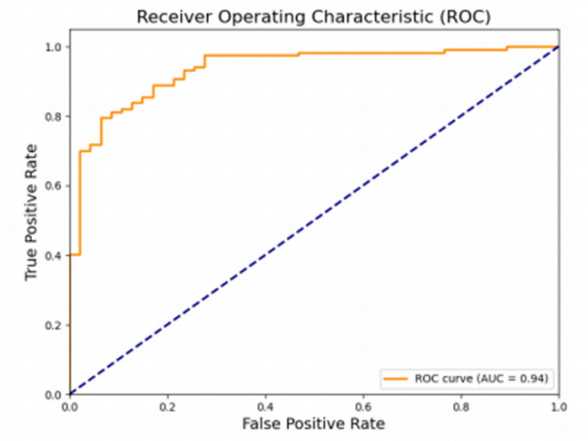

Figure 5. Receiver Operating Characteristic (ROC)

4.3 Results and Analysis of Random Forest

In this analysis, our paper employed a random forest model to train data from three datasets provided and optimized the model's hyperparameters through cross-validation to achieve ideal results. Our research first converted event types into categorical variables, while the number of events and medal counts for each sport were treated as continuous variables for model input. To improve model accuracy and reduce overfitting, this paper selected key features, including the number of events in each Olympic Games, the performance of athletes from each country (including gold, silver, and bronze medal counts), event types, and the host country's historical performance and geographical location.

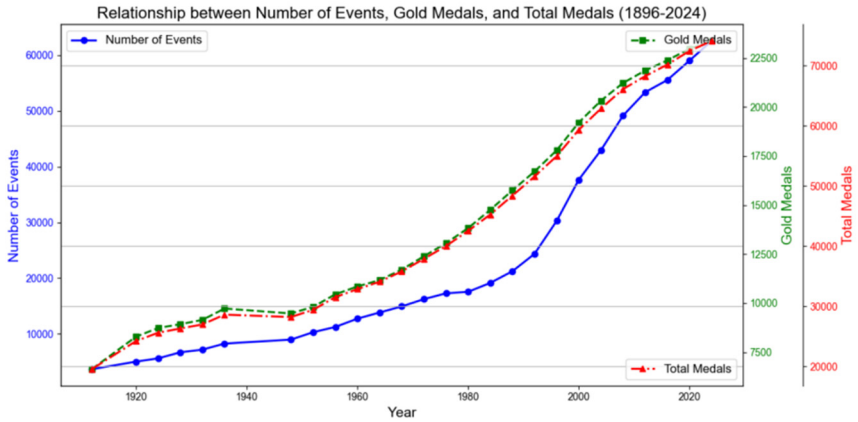

Figure 6. Relationship between Number of Events, Gold Medals, and Total Medals (1896-2024)

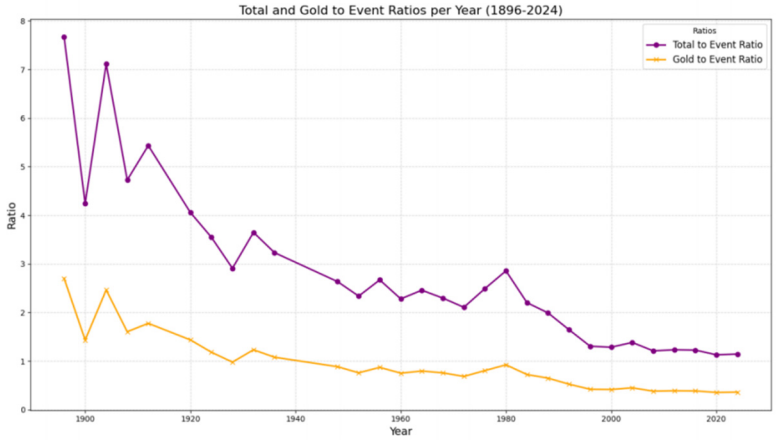

Figure 7. Total and Gold to Event Ratios Per Year(1896-2024)

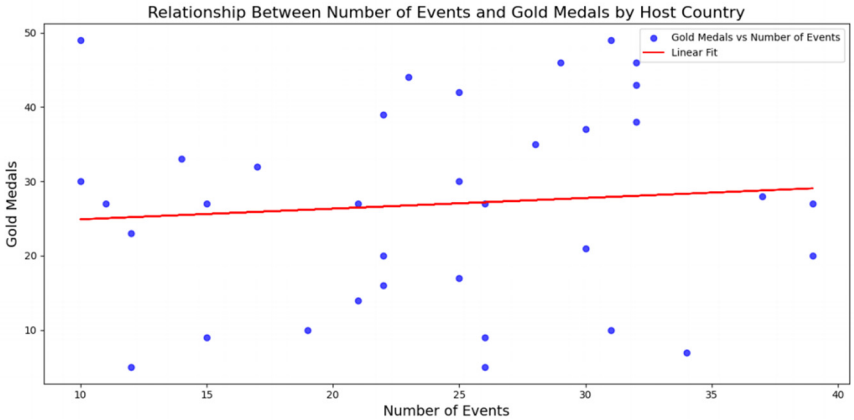

Figure 8. Relationship Between Number of Events and Gold Medals by Host Country

Finally, through feature selection, the model focused on the factors most closely related to medal counts. Our study then trained the data using the random forest regression model and further optimized hyperparameters such as the number of trees and tree depth through cross-validation and during the training process, resulting in improved model performance. The final data visualization results are shown in the Figure 6.

From the Figure 7-8, it is observed a positive correlation between the number of events and both the gold medal count and the total medal count. Additionally, certain events, such as athletics and swimming, contribute signifcantly to the total medal count, while other events, like football and synchronized swimming, have a smaller contribution. Furthermore, the model reveals the distribution characteristics of medals across different events, indicating that some events dominate the medal distribution due to their high participation and larger scale. The events selected by the host country also have a certain impact on the total medal count for that country. Host countries typically choose their strongest events and invest more in them, a strategy that helps improve their performance on the medal tally.

However, upon analyzing the actual situation, it can be easily found that the major sports powerhouses tend to remain consistent over the years, and the identity of the host country does not have a signifcant impact on this. Among the top sports countries, the host nation enjoys a slight advantage, but for smaller countries, even as the host, the impact on their gold medal count is minimal.

Finally, this paper evaluate the performance of the model designed in this study using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coeffcient of Determination (R^2). The calculated results and corresponding coefcients are shown in the Table 4:

Table 4. Model Evaluation Results

Model	MSE	RMSE	R^2
Gold Medal Model	15.23513442	3.22760948	0.850763296
Total Medal Model	13.98372654	2.85477457	0.876452098

5. Conclusion

This study combines ARIMA time series analysis with machine learning methods (XGBoost, SVM, and Random Forest Regression) to predict the medal distribution for the 2028 Los Angeles Olympics. The method captures Olympic medal count trends and considers external factors like GDP, population, and athlete performance. Data preprocessing and model optimization ensure prediction accuracy and reliability. The ARIMA model captures seasonal fuctuations in medal distribution, while XGBoost handles complex factors like GDP and event numbers, improving prediction accuracy. For countries potentially winning their first gold medal, the SVM model predicts the likelihood, identifying promising nations despite limited data.

References

[1] Bernard A B, Busse M R. Who Wins the Olympic Games: Economic Resources and Medal Totals[J]. Review of Economics Statistics,2006,86(1).

[2] Aimin Guo, Mingfa Zhao. “Prediction of the 2016 Summer Olympics Gold Medal Rankings Based on Grey Theory.” China Science and Technology Information, 2013(9).

[3] Shiyu Wang. “Olympic Medal Prediction Model Based on Nonlinear Regression and BP Neural Network.” Physical Observation and Sports Equipment Technology, 2017(24):3.

[4] Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). Time Series Analysis: Forecasting and Control (4th ed.). Wiley.

[5] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[7] Liu, H., & Yang, X. (2022). Research on the Prediction of Housing Prices Based on Random Forest Regression Model. Journal of Data Science and Analytics, 4(3), 45-56.

[8] Zhang, Y., & Li, Z. (2021). Avocado Price Prediction Using Random Forest Algorithm. International Journal of Data Science, 3(2), 89-102.

[9] Wang, L., & Zhang, J. (2023). Performance Prediction Model of Cryogenic Cooling Machine Based on Random Forest Regression. Optics and Precision Engineering, 31(5), 126-134.

[10] Chen, J., & Zhang, W. (2023). Spatiotemporal Evolution and Driving Mechanisms of Urban Construction Land Structure in the Yellow River Basin: A Random Forest Approach. Progress in Geography, 42(7), 1105-1118.