A Study on the Prediction of Olympic Medal Distribution Based on Regression Analysis and ARIMA Models

Yanhao Zhang *, Tianyu Zhang and Daorui Wan

Shandong University, Jinan, China

* Corresponding Author Email: 15725385856@163.com

Abstract. This study aims to construct a predictive model to forecast the distribution of medals at the 2028 Summer Olympics in Los Angeles. Using regression analysis and ARIMA models, we addressed the issues of predicting the number of medals and analyzing the trends in the progress or regression of medals won by each country. First, we employed a linear regression model, incorporating historical medal data and the number of athletes, to predict the number of gold medals and total medals for each country, and calculated the 95% confidence interval. Second, we utilized ARIMA models to analyze the historical trends in medal counts for each country. By integrating feature fusion and machine learning classifiers, we predicted which countries might experience improvements or declines in future Olympics. Additionally, we employed a random forest model to predict the probability of countries that have never won medals securing their first medals. The research findings provide strategic support for national Olympic committees and offer new methodological references for Olympic medal predictions.

Keywords: Olympic medal predictions, regression analysis, ARIMA model, random forest, confidence interval.

1. Introduction

The Olympic medal tally not only reflects the competitive level of athletes from various countries but also demonstrates the overall strength of a nation's sports system [1]. Accurately predicting the distribution of Olympic medals is of great significance for national Olympic committees in formulating their strategies [2]. This study focuses on the 2028 Los Angeles Olympics and addresses three core questions:

1) Predicting the number of gold medals and total medals for each country, along with their associated uncertainties [3];

2) Analyzing which countries may make progress or regress in future Olympics;

3) Predicting the probability of countries that have never won medals securing their first medals.

To address these questions, we employed linear regression models [4], ARIMA time series models [5], and random forest algorithms [6], integrating historical data and socioeconomic factors into our modeling. Through model evaluation and results analysis, we not only validated the effectiveness of these methods but also provided practical tools and theoretical support for Olympic medal predictions [7].

2. Predicting National Medal Counts through Linear Regression

2.1. Data Preprocessing and Feature Engineering

In the data preprocessing stage, data cleaning is first performed to handle missing values and correct outliers. Next, features are standardized to unify features of different scales. Feature engineering is then used to extract features such as historical medal counts and athlete numbers, construct a feature set, and design new features. Correlation analysis is used to avoid multicollinearity, and the correlation matrix is utilized to optimize feature selection. Finally, features strongly correlated with the target variable are selected to prevent overfitting and improve model stability.

This study selects linear regression as the primary prediction tool due to its simple structure and efficient computation, making it suitable for predicting continuous target variables (such as the

number of gold medals and total medals). Although more complex models exist, linear regression offers strong interpretability, directly reflecting feature influences through regression coefficients, and can be extended via polynomial regression or regularization to handle complex feature relationships.

Based on the regression coefficients and standard errors (SE), the uncertainty of the prediction results is calculated, and a 95% confidence interval is generated:

$$\hat{Y} \pm 1.96 \times SE(\hat{Y}) \tag{1}$$

2.2. Linear Regression Model Construction

In the model construction stage, we choose the linear regression model, whose basic formula is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \grave{o} \tag{2}$$

Ordinary least square method (OLS) was used to estimate the regression coefficient, and the optimal regression coefficient was obtained by minimizing the sum of squares error between the predicted value and the actual value. The formula for the sum of squares of error is as follows:

$$RSS = \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 \tag{3}$$

The goal of least squares is to minimize the sum of squares of error by selecting the regression coefficient $\beta_0, \beta_1, \ldots, \beta_n$, namely:

$$\min_{\beta_0, \beta_1, \ldots, \beta_n} RSS = \min_{\beta_0, \beta_1, \ldots, \beta_n} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 \tag{4}$$

The analytical solution of the least square method for solving the regression coefficient can be obtained by the following formula:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y \tag{5}$$

In addition, model evaluation metrics such as R^2 (coefficient of determination) and mean square error (MSE) will be used to assess the model's fit and predictive power.

The coefficient of determination (R^2) represents the proportion of the total variation in the target variable that the model can account for. Its calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{m} \left(y_i - \bar{y} \right)^2} \tag{6}$$

The values of R^2 range from 0 to 1, with the closer to 1 the more explanatory the model is.

Mean square error (MSE) represents the mean of the square of the difference between the predicted and actual values. The formula of MSE is:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 \tag{7}$$

The smaller the MSE, the closer the model's prediction results are to the actual data, and the better the model performance.

In the model solving phase, we first fit the regression model using the training data and make predictions on the test data. By calculating the standard error (SE) of the model, we can calculate a confidence interval of 0.95 for each prediction, and the formula for the prediction interval is:

$$\hat{Y} \pm 1.96 \times SE(\hat{Y}) \tag{8}$$

Among them, 1.96 is the critical value of the 0.95 confidence interval.

2.3. Model Training and Prediction Results

By using the trained model, we predict the number of gold MEDALS for the 2028 Los Ange-les Olympics, as shown in Fig. 1:

Figure 1. Predict the number of GOLD MEDALS and 95 % forecast interval

3. Forecasting Medal Trends with Time Series and Classifiers

3.1. Data Standardization and Temporal Feature Design

In the model preparation stage, we first need to pre-process the historical Olympic medal data provided. The pre-processing steps are shown in Question 1. In addition, the data needs to be standardized, and the standardized formula is:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{9}$$

3.2. ARIMA-Based Feature Extraction and Classifier Integration

The model consists of feature extraction module, feature enhancement module and classifier module. First, the feature extraction module adopts the historical Olympic medal data as the basis, and uses the autoregressive (AR) model to extract the time series features of the medal number. The AR model can capture trends and seasonal fluctuations in historical data, and analyze the relationship between the number of MEDALS won in the past few Olympic Games and the number of MEDALS won in the future by means of gradual regression.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \text{ò}_t \tag{10}$$

In order to further enhance the perception of medal count changes, we design a feature enhancement module. The module includes several sub-modules, such as the trend of gold medal count, host country effect, etc. We adjust our forecasts by statistical analysis of historical trends in gold, silver and copper MEDALS in each country, taking into account the strengths of the host country.

For example, the host country effect is modeled by the following formula:

$$\text{Effect}_t = \frac{\text{Gold}_t}{\text{Total}_t} \times \text{HostCountryFactor} \tag{11}$$

In the feature enhancement module, we also consider the performance of different countries in various sports. The correlation analysis of the number of events is used to identify which sports are most important for different countries, and this information is incorporated into the prediction model.

$$\text{ImportantSport}_t = \sum_{i=1}^{n} \text{Weight}_i \times \text{SportScore}_i \tag{12}$$

Finally, the classifier module combines all the extracted and enhanced features, utilizing machine learning models (such as random forests, support vector machines, etc.) to make predictions about the number of MEDALS for each country. The output classification probability is the probability of change in the number of MEDALS won by each country in future Olympic Games.

$$\mathbf{P} = \text{softmax}\left(\mathbf{W}_2 \cdot \text{ReLU}\left(\mathbf{W}_1 \cdot \mathbf{F}_{\text{fusion}} + \mathbf{b}_1\right) + \mathbf{b}_2\right) \tag{13}$$

Where $\mathbf{F}_{\text{fusion}}$ is the fusion feature obtained through the feature fusion module, $\mathbf{W}_1$, $\mathbf{W}_2$ is the weight matrix, $\mathbf{b}_1$, $\mathbf{b}_2$ is the offset term.

The loss function of the model adopts cross entropy loss, and Adam optimizer is used to train the model, which ensures the stability and accuracy of the model on different data sets.

$$\text{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log\left(p_{i,k}\right) \tag{14}$$

3.3. Forecasting Trends and Visualizing Medal Dynamics

After the model is built, the ARIMA model is trained using historical data. Through training, the model learns to predict a country's future medal count based on trends and seasonal patterns in historical medal data. Then the prediction accuracy of the model was evaluated. By evaluating the effect of the model, we can further judge the applicability of the model and adjust parameters to optimize the model, as shown in Fig. 2.

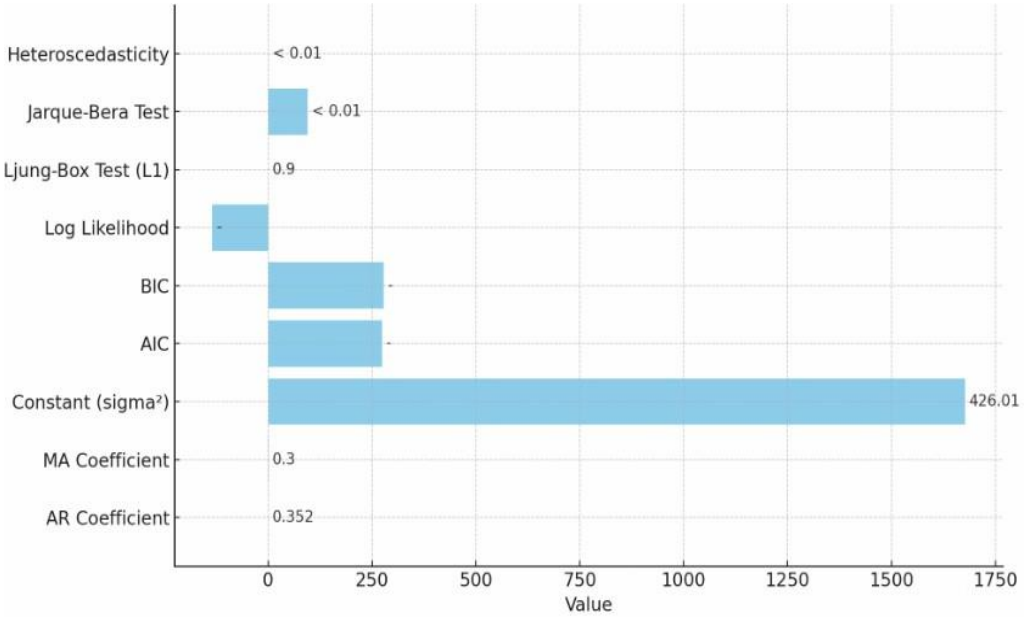

Figure 2. ARIMA Model Parameters and Statistics

After training and evaluation, we can use the ARIMA model to predict future medal trends and determine which countries are expected to improve their performance and which countries are likely to decline. Take countries such as the United States, China and Japan as examples, as shown in Fig. 3.

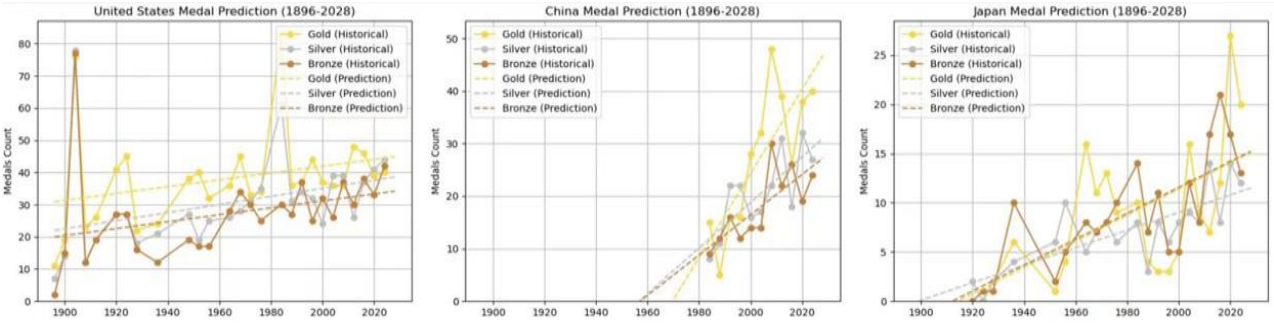

Figure 3. Olympic Medal Counts (1988-2028) for Selected Countries with Predictions 1

4. Predicting First-Time Medal Winners Using Random Forests

4.1. Feature Selection and Categorical Encoding for Emerging Countries

Data preprocessing is carried out before model construction. First, the missing values are processed, and the records with serious missing values are filled with the mean, mode or median, or the records with serious missing values are deleted. Then the outliers are processed, and the Z-score method is commonly used to identify and eliminate the outliers. The formula is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{15}$$

Data conversion is the conversion of categorical variables into numerical data using unique thermal coding. Countries that have never won a medal in history are then selected. Finally, feature engineering is carried out to extract features from the country's socio-economic characteristics such as GDP, population size, athletes' performance and host country effect. After data preprocessing, we select random forest algorithm. Random forests improve prediction accuracy and reduce overfitting by integrating multiple decision trees.

4.2. Random Forest Model Design and Evaluation Metrics

First, we need to define the input characteristics of the model. Input features are constructed from the following aspects:

$$\begin{aligned}\text{Input feature } = [&\text{GDP, population, Sport performance,} \\ &\text{Host country effect, Athlete performance }]\end{aligned} \tag{16}$$

Next, we set the target variable as a binary variable, indicating whether a country will win its first medal in 2028. The definition of the target variable is:

$$y = \begin{cases} 1, & \text{If the country wins its first medal in 2028} \\ 0, & \text{If the country does not win its first medal in 2028} \end{cases} \tag{17}$$

A random forest model is used to train this target variable. First, the random forest is trained by training set. The random forest algorithm classifies by integrating multiple decision trees. During the construction of each decision tree, the nodes are split by randomly selecting a part from all the features, which reduces the overfitting risk of the model and improves the generalization ability.

In order to ensure the generalization ability of the model, we will use cross-validation for training and evaluation. In K-fold cross-validation, the data set is divided into k subsets, one subset is used as the validation set at a time, and the rest is used as the training set for model training. This process can be expressed by the following formula:

$$\text{Cross validation error } = \frac{1}{k}\sum_{i=1}^{k} \text{error}_i \tag{18}$$

Where error_i represents the error in the first i validation, and k represents the number of cross-validations. Through cross-validation, we can effectively evaluate the stability of the model and avoid overfitting the model on a single training set.

After model training is completed, we will conduct model evaluation. Key metrics for model evaluation include accuracy, accuracy, recall, and F1 scores. Accuracy and recall rates can be calculated using the following formula:

$$\text{precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{recall rate} = \frac{TP}{TP + FN} \tag{20}$$

The $F1$ score is a harmonic average of the accuracy and recall rates used to comprehensively evaluate the performance of the model:

$$F1 = 2 \times \frac{\text{accuracy} \times \text{recall}}{\text{accuracy} + \text{recall}} \tag{21}$$

In the evaluation process, we will also conduct feature importance analysis to identify which features have a greater impact on the model's prediction results. Feature importance can be obtained by calculating the contribution of each feature's split node in the tree to error reduction.

After training and evaluating the model, we use it to make predictions. The predicted output of the model is the probability that a country that has not won a gold medal will win its first medal in 2028. For each country, the model will output a probability value $P(y = 1 | X)$ representing the probability of that country winning its first medal in 2028, where X is the input feature vector for that country.

Based on the predicted probabilities, we can calculate the odds of each country winning the first medal. The odds are calculated as follows:

$$\text{odds} = \frac{1}{P(y = 1 | X)} \tag{22}$$

Finally, in order to quantify the uncertainty of prediction results, we can estimate the uncertainty through Bootstrap and other methods.

4.3. Predictive Outcomes and Probability Interpretation

We used the Random Forest algorithm to predict the probability of each country winning its first medal at the 2028 Olympics. Relevant data during model training are shown in Table 1:

Table 1. The Model Evaluates Key Index Values

evaluation index	data	evaluation index	data
precision rate	73.5%	recall rate	68.2%
accuracy rate	71.8%	F1 score	70.8%

The prediction results of the first medal in 2028 Olympic Games based on the trained random Forest model are shown in Fig. 4:

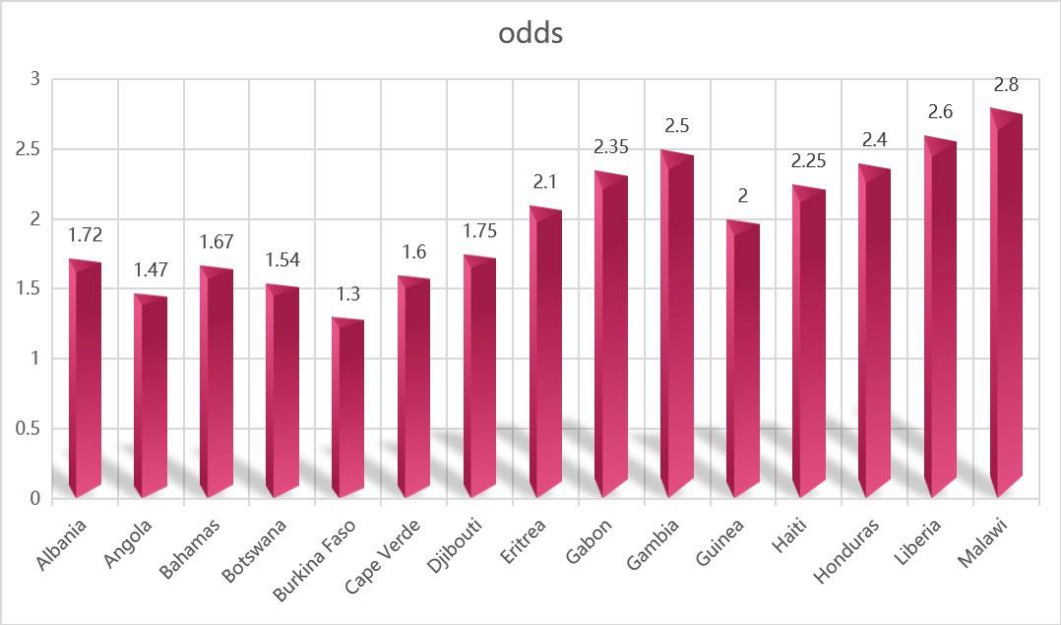

Figure 4. The Odds of Each Country Winning the First Medal

These odds indicate the likelihood of these countries winning their first medal in 2028. The lower the odds, the more likely the country is to win.

5. Conclusion

This study developed a comprehensive prediction framework for the 2028 Los Angeles Olympics using linear regression, ARIMA time series modeling, and random forest algorithms. The regression model accurately forecasted the number of gold and total medals, while the ARIMA-based approach effectively identified temporal trends and highlighted countries likely to improve or decline in performance. In addition, the random forest model assessed the probability of countries winning their first Olympic medals, achieving solid results across multiple evaluation metrics. Together, these models provide actionable insights for national Olympic committees and a solid methodological foundation for medal forecasting.

Future research can further enhance prediction accuracy by incorporating additional dynamic variables such as athlete-level performance data, injury records, and geopolitical factors. Advanced models like LSTM or temporal convolutional networks (TCNs) may also improve time series prediction performance. This work demonstrates the value of integrating traditional statistical analysis with modern machine learning techniques in the field of sports analytics, offering both theoretical and practical guidance for Olympic medal prediction and strategic planning.

References

[1] Zhang Chupei. Digital Intelligence Empowering the Olympics: Practical Experience and Future Prospects of the 2024 Paris Olympics [C]//Shaanxi Sports Science Society, Shaanxi Student Sports Association. Proceedings of the 4th Shaanxi Sports Science Conference (Abstracts) - Comprehensive Sports (Poster Session). Harbin Sport University, 2025: 17. DOI: 10.26914/c.cnkihy.2025.016666.

[2] Tang Haifeng, Zhao Lunan, Yi Chao, et al. Research on the Trend of China's Medal Performance and Event Distribution Characteristics in Past Summer Olympics [J]. Journal of Anhui Sports Science, 2018, 39 (03): 35 - 39.

[3] Wan Jinjing. A Dynamic Uncertainty Quantification Fusion Framework Based on Quantile Regression and Conformal Prediction [J]. Statistics and Management, 2025, 40 (06): 23 - 36. DOI: 10.16722/j.issn.1674 - 537x.2025.06.011.

[4] Yang Qinwei. Multiple Linear Regression Model for Predicting Performance in the 2020 Olympics [J]. Practical Electronics, 2018, (Z2): 121 - 123. DOI: 10.16589/j.cnki.cn11-3571/tn.2018.z2.058.

[5] Luo Junrong. Application of ARIMA Model in GDP Forecasting for Lanzhou City [J]. China Electronic Commerce, 2025, 31 (11): 7 - 9.

[6] Zhen Yan, Kang Jintao, Zhao Xiaoming, et al. Logging Interpretation Method for Diagenetic Facies of Tight Sandstone Based on Improved SMOTE and Random Forest Algorithm [J/OL]. Journal of Southwest Petroleum University (Natural Science Edition), 1 – 13 [2025 - 07 - 04]. http: // kns. cnki. net/ kcms/ detail/ 51. 1718.TE.20250701.1607.002.html.

[7] Zhang Yuhua. Prediction of China's Medal Count in the 31st Olympics Based on Dynamic Linear Regression Model [J]. Journal of Henan Normal University (Natural Science Edition), 2013,41 (02): 24 - 26+60. DOI: 10.16366/j.cnki.1000 - 2367.2013.02.003.