Olympic Medal Count Prediction Research Based on A Hybrid ARIMA-Xgboost-Lightgbm Model

Zijun Du [*, #], Yirui Zheng [#]

College of Science, Beijing University of Civil Engineering and Architecture, Beijing, China, 102627

* Corresponding Author Email: 202208010203@stu.bucea.edu.cn

[#]These authors contributed equally.

Abstract. This study proposes an integrated ARIMA–XGBoost–LightGBM model to predict both medal counts and medal-winning probabilities for the 2028 Los Angeles Olympic Games. The framework begins by applying the ARIMA model to forecast time-series features, such as the number of athletes and participating countries. These features are then fed into an XGBoost regressor to estimate gold, silver, bronze, and total medals per country. Additionally, a LightGBM classifier is utilized to predict the probability that a nation will win at least one medal. Model performance was evaluated using ten-fold cross-validation, with R^2 values exceeding 0.84 across all medal categories, demonstrating high accuracy and robust generalization ability. Notably, six countries are projected to win their first Olympic medals based on the predicted probabilities. This ensemble approach effectively combines time series forecasting with machine learning algorithms, showcasing its potential in supporting strategic sports planning and medal outcome prediction. It provides valuable forecasting capabilities for complex, dynamic events such as the Olympics.

Keywords: Olympic medal prediction, ARIMA, XGBoost, LightGBM.

1. Introduction

The Olympic Games serve as a premier international sporting event that not only showcases athletic excellence but also reflects a nation's comprehensive strength in sports development, international engagement, and public investment in talent cultivation [1]. Accurate prediction of Olympic medal counts has important strategic value: it can assist national Olympic committees in planning resource allocation, identifying emerging competitive fields, and setting realistic performance goals [2-4]. Especially in light of the upcoming 2028 Los Angeles Olympics, the task of forecasting medal outcomes is of growing interest, given the potential influence of the host-country advantage, evolving athletic capabilities, and changes in Olympic event composition.

Traditional medal prediction methods often rely on linear regression models or simple statistical extrapolation based on historical rankings [2, 5, 6]. While these approaches may provide a rough estimation, they often fail to incorporate nonlinear dynamics, inter-country disparities, and temporal trends in participation and performance [7]. Moreover, few models account for uncertainty in predictions or offer interpretable insights into which features drive success. These limitations hinder their applicability to complex multi-nation, multi-event competitions like the Olympics, particularly under changing global conditions and evolving athlete pools [8-10].

To address the above challenges, this study proposes a hybrid ARIMA–XGBoost–LightGBM modeling framework. The approach leverages the strength of ARIMA for time-series forecasting of key indicators (e.g., number of athletes), XGBoost for regression-based medal count prediction, and LightGBM for probabilistic classification of medal-winning likelihoods. This integrated pipeline effectively combines sequential trends with nonlinear learning and offers both quantitative predictions and probabilistic assessments. By applying this ensemble model to historical Olympic data and forecasting the 2028 Games, the proposed method demonstrates both high predictive accuracy and strong generalizability, while providing interpretable insights through SHAP values and model decomposition.

2. Construction of the ARIMA-XGBoost-LightGBM Medal Prediction Model

This study presents an integrated ARIMA–XGBoost–LightGBM framework, as illustrated in Figure 1, for predicting medal outcomes at the 2028 Los Angeles Olympics. ARIMA forecasts key variables such as athlete counts, which serve as inputs to an XGBoost regressor for estimating each country's medal counts. A LightGBM classifier further predicts the probability of winning any medal.

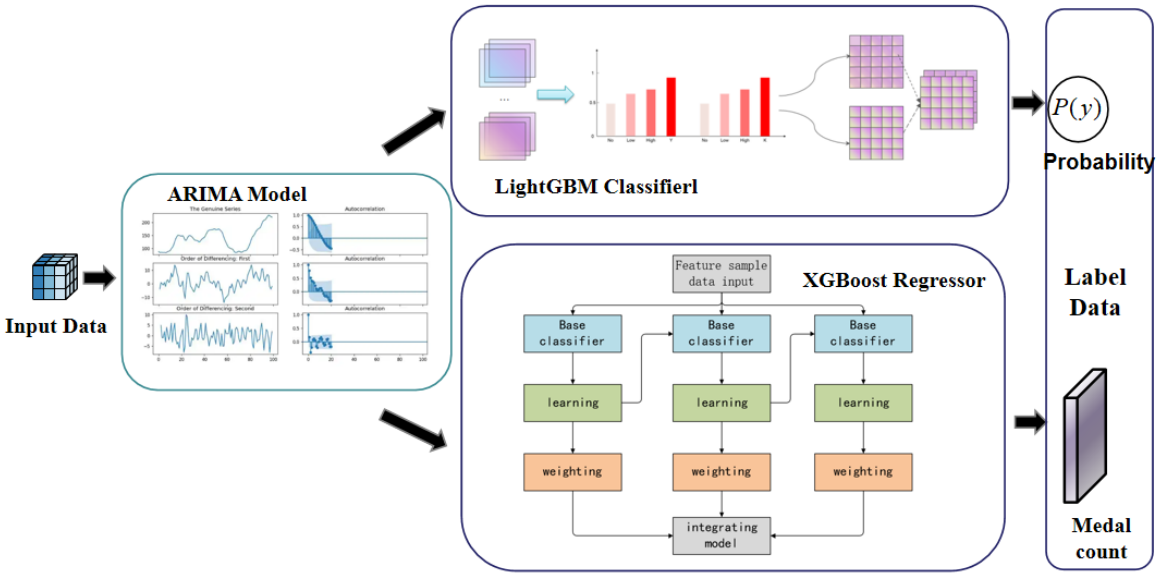

Figure 1. ARIMA–XGBoost–LightGBM model framework diagram

2.1. Construction of the ARIMA Model

The ARIMA model is characterized by three parameters: p, d, and q, which represent the order of the autoregressive terms, the degree of differencing, and the order of the moving average terms, respectively. The general form of the model is:

$$y_t=\mu+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\cdots+\phi_p y_{t-p}+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q}+\varepsilon_{t-p} \quad (1)$$

where, y_t denotes the observed value of the time series at time t; μ is a constant term (i.e., the mean); $\phi_1,\phi_2,\cdots,\phi_p$ are the autoregressive coefficients; $\theta_1,\theta_2,\cdots,\theta_q$ are the moving average coefficients; and ε_t represents the error term.

By modeling and analyzing time series data, the ARIMA model effectively forecasts key features—such as the number of athletes—for the 2028 Los Angeles Olympics. These predicted features serve as valuable inputs for the subsequent prediction and classification tasks performed by the XGBoost and LightGBM models.

2.2. Construction of the XGBoost Model

XGBoost (Extreme Gradient Boosting) is an ensemble algorithm based on gradient-boosted decision trees. It improves prediction accuracy by sequentially fitting new trees to the residuals of previous models. To enhance generalization and prevent overfitting, XGBoost incorporates regularization into its objective function, which consists of a loss term and a regularization term, defined as:

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

where $L(y_i, \hat{y}_i)$ denotes the loss function for the *i*-th sample, typically the mean squared error (MSE); y_i is the true value, and $\hat{y}_i$ is the predicted value. $\Omega(f_k)$ represents the regularization term, which controls model complexity and helps prevent overfitting.

The training process of the algorithm can be summarized as follows:

1) In each iteration, a new decision tree is added to the existing model.

2) Before each iteration, gradient statistics are computed to guide the tree structure optimization.

$$g_i = \frac{\partial L\left(yi, \hat{y}^{(s-1)}\right)}{\partial \hat{y}^{(s-1)}} \tag{3}$$

$$h_i = \frac{\partial^2 L\left(yi, \hat{y}^{(s-1)}\right)}{\partial \hat{y}^{(s-1)}} \tag{4}$$

where g_i represents the first-order gradient of the objective function, and h_i represents the second-order gradient.

3) A complete decision tree $f_s(x)$ is constructed based on the exact greedy algorithm and gradient information.

During node splitting, the optimal feature and split point are selected by evaluating the gain in the objective function.

$$Gain = Obj_{\text{split}} = -\frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right) - \gamma \tag{5}$$

The weights of the leaf nodes in the final tree are given by:

$$\omega_j^* = \frac{G_j}{H_j + \lambda} \tag{6}$$

4) The newly generated tree is added to the existing model:

$$\hat{y}_i^{(s)} = \hat{y}_i^{(s-1)} + \eta f\left(x_i\right) \tag{7}$$

where the weight parameter η functions similarly to a learning rate by controlling the influence of each newly added tree on the overall model. This ensures that each iteration makes only a small improvement, allowing the model to gradually converge toward the optimum and effectively reducing the risk of overfitting.

2.3. Construction of the LightGBM Model

The core idea of LightGBM is to reduce prediction error iteratively using a gradient boosting framework. For multi-class classification tasks with *K* categories, LightGBM builds multiple binary classifiers, each estimating the probability of a specific class. The final predicted class is the one with the highest probability. The procedure is as follows:

1) Objective Function: At each iteration, LightGBM aims to minimize a loss function. In this study, the multi-class cross-entropy loss is adopted:

$$L(y, p) = -\sum_{k=1}^{K} y_k \log(p_k) \tag{8}$$

where y_k is the indicator that the true label of sample y belongs to class k, and p_k denotes the predicted probability that the sample belongs to class k.

2) Gradient Boosted Decision Tree: In each iteration, a new decision tree is constructed based on the residual errors from the previous round. The prediction is updated as follows:

$$F_{m+1}(x) = F_m(x) + \alpha T_m(x) \tag{9}$$

where $F_{m+1}(x)$ is the output of the m-th tree, $T_m(x)$ is the newly trained decision tree, and α is the learning rate.

3) Class Prediction: For multi-class classification tasks, LightGBM trains a separate classifier for each class. The predicted probability for class k is calculated as:

$$p_{\dot{\kappa}} = \frac{\exp(F_{\dot{\kappa}}(x))}{\sum_{j=1}^{K} \exp(F_j(x))} \tag{10}$$

LightGBM incorporates several optimization techniques to enhance its efficiency and performance, particularly on large-scale datasets. Key among these are the histogram-based decision tree learning algorithm and the leaf-wise tree growth strategy.

3. Results

3.1. Datasets

The data in this paper are sourced from the official website of Nielsen. This study integrates multi-source historical data from the Summer Olympics, including athlete participation records, national medal statistics, host country information, and event classifications, as released by the International Olympic Committee. Key preprocessing steps included: (1) aggregating athlete data by country, year, and sport to calculate participation numbers and gender ratios; (2) merging medal statistics with zero-value imputation for non-winning countries; (3) introducing a host country indicator variable to quantify the host nation effect. For the 76-dimensional high-dimensional features, principal component analysis (PCA) was applied to extract the top 10 principal components (PC1-PC10), achieving a cumulative variance contribution rate of 92.4% to reduce data complexity. Feature engineering derived critical indicators such as medal growth rate and medal efficiency, with standardization ensuring model input consistency.

3.2. Analysis of ARIMA Model Prediction Results

To forecast the number of participants and other key indicators for the 2028 Olympic Games, this study employed the ARIMA model. First, a stationarity test was conducted on variables such as PC1 using the Augmented Dickey-Fuller (ADF) test, which confirmed that the data became stationary after first-order differencing. Subsequently, the ARIMA (0,1,0) model was selected based on the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. As shown in Figure2, the fitted model captures the historical trend effectively, providing a reliable basis for forecasting future medal-related indicators. Based on this fitted model, predictions for the number of participants and other relevant features at the 2028 Olympics were obtained.

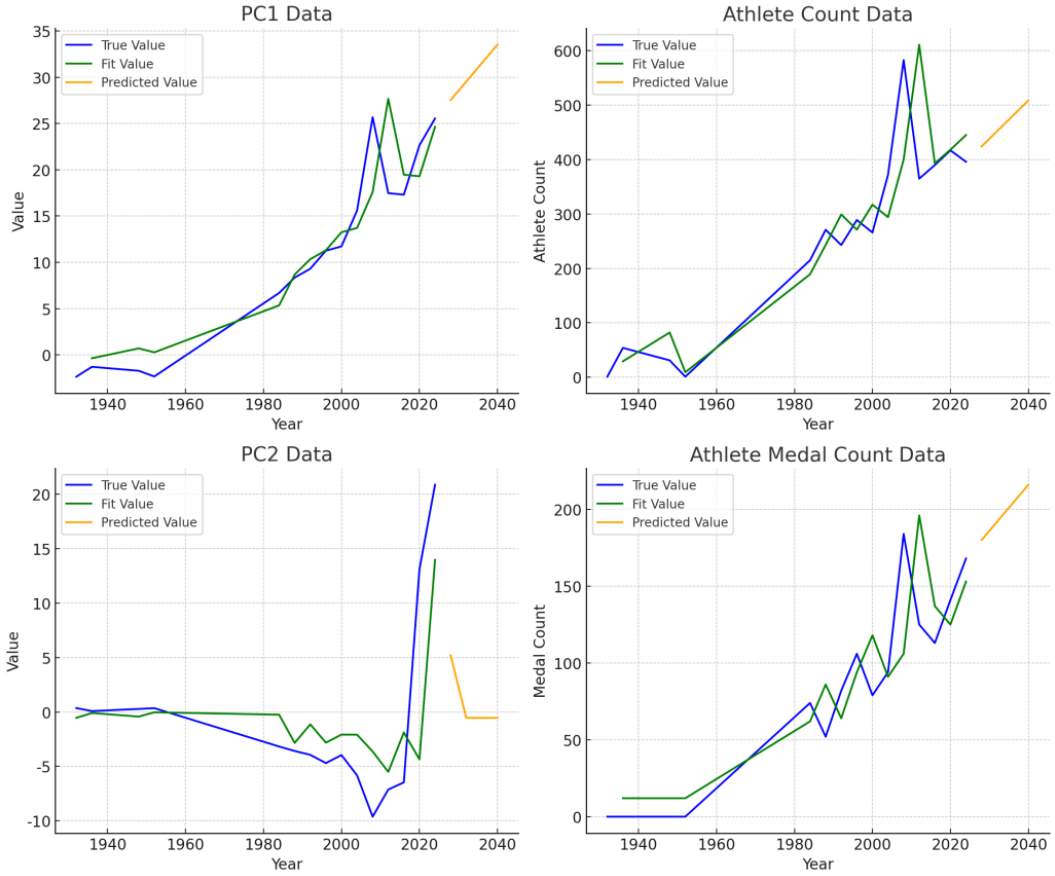

Figure 2. Time Series Analysis for Olympic Data

Figure 2. illustrates the ARIMA model's prediction results for the PC1 data, including the actual values, fitted values, and forecasted values. For some feature columns, the ARIMA model exhibited suboptimal performance, which may be attributed to the loss of temporal information during PCA-based dimensionality reduction. Therefore, for features with a coefficient of determination (R^2) below 0.6, the average values from the previous two Olympic Games were used as the predicted values for the 2028 Olympics.

3.3. Analysis of XGBoost Prediction Results for Medal Counts

During the modeling process, the extracted and dimensionally reduced features were used as training inputs. The data were first standardized, and the XGBoost model was then trained on this processed dataset. Through iterative learning and optimization of the feature space, the model generated fitted results for gold, silver, bronze, and total medal counts, as shown in Figure 3.

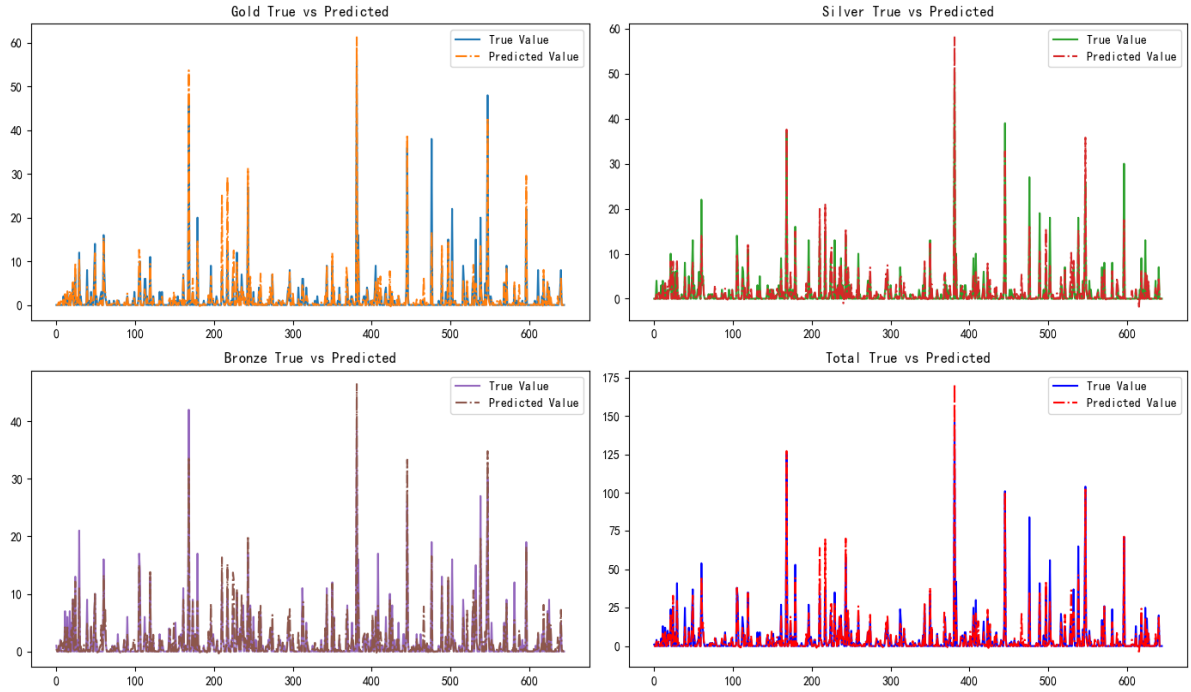

Figure 3. Comparison of XGBoost Model Predicted Values and Actual Values

From these comparison plots, it can be observed that the XGBoost model fits the actual values well in most cases, particularly in the high-value range, where it accurately captures the peaks in medal counts. To assess the uncertainty of the model's predictions, ten-fold cross-validation was conducted. During this process, performance metrics were recorded, and each trained model was saved for subsequent prediction tasks. The performance metrics are summarized in Table.1.

Table.1. XGBoost Model Ten-Fold Cross-Validation Results and Evaluation Metrics

Type	Mean	Std
Gold Explained Variance	0.8421	0.0781
Gold Mean Absolute Error	0.7549	0.0838
Gold Mean Squared Error	4.9746	1.9713
Gold R-squared	0.8416	0.0782
Silver Explained Variance	0.8791	0.0585
Silver Mean Absolute Error	0.6585	0.0868
Silver Mean Squared Error	2.8670	1.1410
Silver R-squared	0.8790	0.0586
Bronze Explained Variance	0.8461	0.0640
Bronze Mean Absolute Error	0.7256	0.0806
Bronze Mean Squared Error	3.7554	2.3033
Bronze R-squared	0.8456	0.0644
Total Explained Variance	0.9180	0.0471
Total Mean Absolute Error	1.3082	0.2084
Total Mean Squared Error	18.530	10.175
Total R-squared	0.9177	0.0472

The 2028 forecasted values of the feature variables obtained from the ARIMA model were used as inputs for the ten trained XGBoost models. Based on these inputs, the expected number of medals for each country in the 2028 Olympics was estimated. The prediction intervals are illustrated in Figure 4.

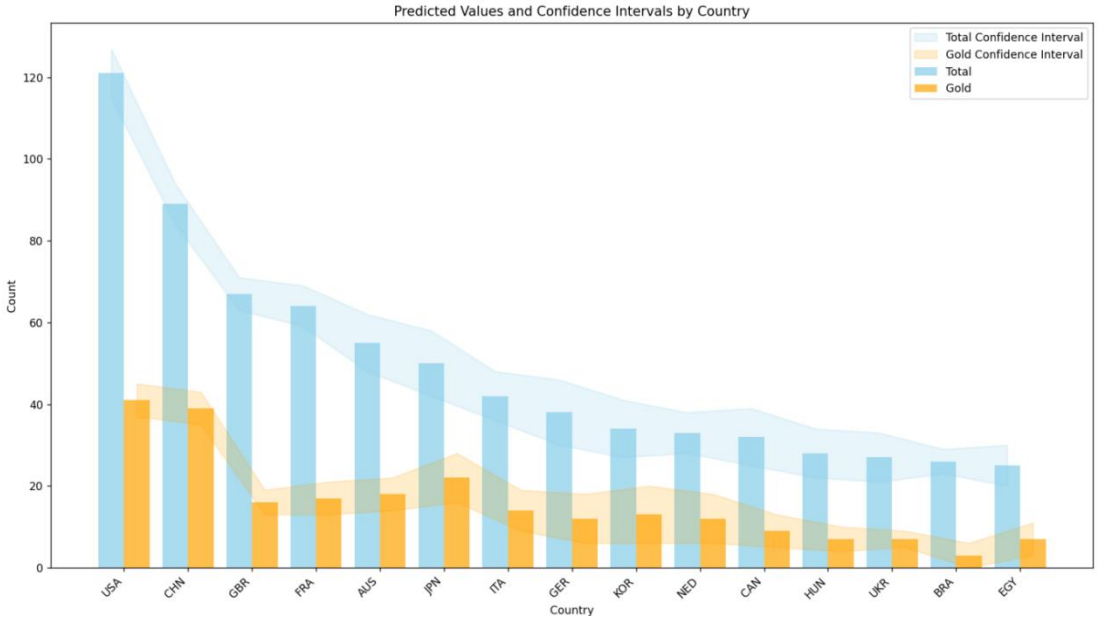

Figure 4. Comparison of XGBoost Model Predicted Values and Actual Values

The U.S. leads in both gold and total medals with stable predictions, followed by China and the UK; countries like Egypt and Brazil show greater uncertainty.

3.4. Analysis of LightGBM Model Results for Medal-Winning Probability Prediction

We employed the LightGBM model to perform a binary classification task on the "Medal_pro" variable. Using the features extracted and dimensionally reduced in, the model was trained to predict whether a country would win at least one medal in a given year. Instead of outputting a discrete class label, the model returns the probability of class "1," which represents the likelihood that a country will win a medal. The classification results are shown in Figure 5.

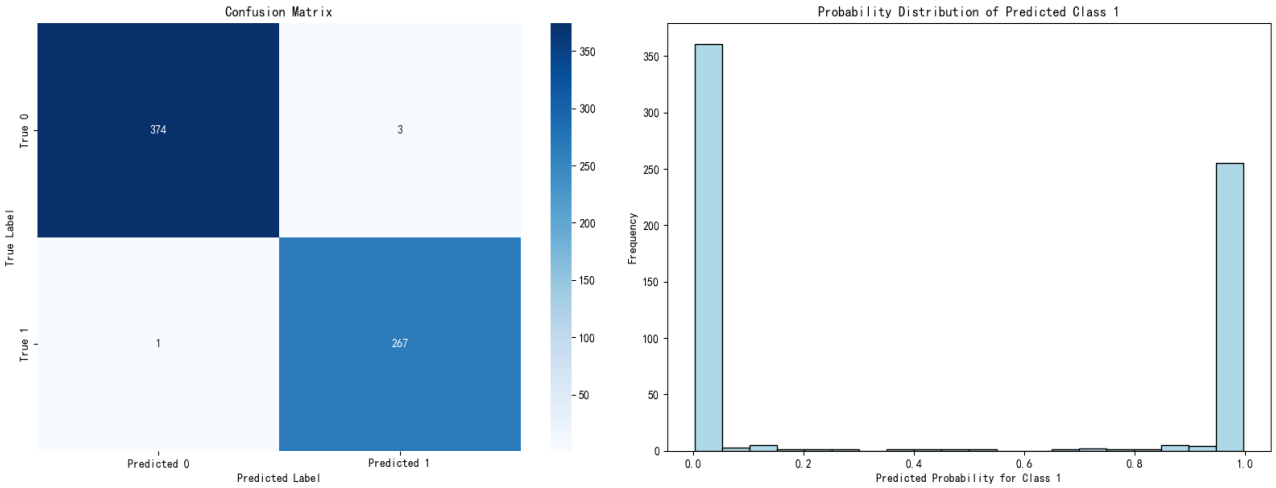

Figure 5. LightGBM Model Prediction Results: Confusion Matrix and Predicted Probability Distribution for Class 1

Figure 5 presents the model's confusion matrix and the predicted probability distribution for class 1. The confusion matrix summarizes the classification performance on the test set, including both correct and incorrect predictions. The probability histogram illustrates the distribution of predicted probabilities for class 1.

Using the 2028 feature forecasts generated by the ARIMA model as inputs to the trained LightGBM model, we estimated the medal-winning probabilities for countries that did not win medals previously. The results are shown in Table 2.

Table.2. Table of predicted national award probabilities

Year	NOC	Medal_pro_prob
1932	AUS	0.2032
1932	MEX	0.3590
1948	IRL	0.2193
1956	GER	0.3084
1964	GER	0.3299
1968	TPE	0.3268
1992	IOA	0.3072
1996	SCG	0.2067
2000	SCG	0.2189
2012	MDA	0.3428
2016	IOA	0.4043
2020	ROC	0.4495
2024	AIN	0.4382

We set 0.2 as the probability threshold; countries with predicted medal-winning probabilities above this value are considered to have a relatively high likelihood of earning a medal in the upcoming Olympics. Based on this criterion, six countries are projected to win their first Olympic medal: PHI, VNM, KEN, FJI, BAH, and ARE.

4. Conclusions

This paper presents a novel ensemble framework integrating ARIMA, XGBoost, and LightGBM to predict Olympic medal outcomes. The ARIMA component enables the capture of time-series patterns in key features, while XGBoost and LightGBM enhance predictive accuracy and classification capabilities. The model demonstrates strong performance in both medal count prediction and medal-winning probability estimation, with high consistency observed through cross-validation. Forecast results suggest that the United States will maintain its lead, while emerging nations such as the Philippines and Kenya show potential for earning their first Olympic medals. The proposed framework not only offers high accuracy but also improves interpretability through the use of SHAP analysis. Future work may incorporate more contextual variables (e.g., economic or policy factors) and extend the framework for multi-Games comparative prediction.

While the model has performed well in medal prediction, incorporating additional contextual variables (e.g., economic and policy factors) could further improve accuracy. Future work could extend the model's application to other international events, enhancing its adaptability and real-time forecasting ability. As data and algorithms evolve, the model's generalizability is expected to improve.

References

[1] DE BOSSCHER V. The global sporting arms race: An international comparative study on sports policy factors leading to international sporting success [M]. Meyer & Meyer Verlag, 2008.

[2] SCHLEMBACH C, SCHMIDT S L, SCHREYER D, et al. Forecasting the Olympic medal distribution–a socioeconomic machine learning model [J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[3] Moolchandani J, Chole V, Sahu S, et al. Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics[C]//2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS). IEEE, 2024: 1987-1992.

[4] WANG Y, WANG J, HUANG T-Y, et al. STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization [J]. arXiv preprint arXiv:250117711, 2025.

[5] ZHAO S, CAO J, STEVE J. Research on Olympic medal prediction based on GA-BP and logistic regression model [J]. F1000Research, 2025, 14: 245.

[6] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction[C]//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023: 249-267.

[7] Fazlollahi P, Afarineshkhaki A, Nikbakhsh R. Predicting the medals of the countries participating in the Tokyo 2020 olympic games using the test of networks of multilayer perceptron (MLP)[J]. Annals of Applied Sport Science, 2020, 8(4): 0-0.

[8] SUN A W. Medal count disparities at the Olympic Games: An econometric analysis of the determinants of national Olympic success using an economic growth framework [D]; Master thesis]. Department of Economics: Copenhagen Business School, 2020.

[9] REIS F J, ALAITI R K, VALLIO C S, et al. Artificial intelligence and machine-learning approaches in sports: Concepts, applications, challenges, and future perspectives [J]. Brazilian Journal of Physical Therapy, 2024: 101083.

[10] LECKEY C, VAN DYK N, DOHERTY C, et al. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis [J]. British Journal of Sports Medicine, 2025, 59(7): 491-500.