ISSN: 2688-8645

Study on Medal Prediction Based on XGBoost and ARIMA Models

Jiayi Ma*, Yunyi Chen, Xiang Lin

School of Electronic Information Science, Fujian Jiangxia University, Fuzhou, China

*1462397599@qq.com

Abstract

Focusing on predictive modelling research, this paper constructs an Olympic medal prediction model and evaluates its performance. XGBoost combined with Monte Carlo simulation is used to predict the number of Olympic medals. Optuna optimises the hyperparameters and evaluates the model in terms of mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE), and the results show that the model performs well in predicting the number of gold medals and the total number of medals; Monte Carlo simulation is used to determine the prediction intervals and to evaluate the model uncertainty. For the first medal prediction, the ARIMA model was used to process the data with time series characteristics, and the ADF test was used to judge the stability of the data, construct and solve the model to predict the trend of the number of gold medals; at the same time, the logistic regression model was used to convert the medal count labels into binary labels for training, and ROCAUC was used to assess the classification performance of the model, and the code outputs its value as 0.7403825829634327 , indicating that the model has the ability to distinguish whether awards.

Keywords

Olympic Medal Prediction; XGBoost Model; ARIMA Model; Monte Carlo Simulation.

1. Introduction

At a time when big data and artificial intelligence are deeply integrated, prediction models are becoming increasingly important in many fields. Taking sports events as an example, Olympic medal prediction can not only satisfy the public's keen attention, but also has key significance for sports resource planning and so on.

In the past, the commonly used linear regression forecasting methods, such as time series methods, analytical methods, and pattern recognition methods, have obvious deficiencies when dealing with Olympic medal data. These methods are limited by linear assumptions, which makes it difficult to explore the complex patterns and potential relationships in the data, leading to poor reliability of forecasting results. For example, it is impossible to accurately consider the impact of economic development, sports policy changes and other factors on the ability of countries to obtain medals.

In recent years, machine learning technology has been booming, and new methods such as XGBoost, Monte Carlo simulation, ARIMA model, and logistic regression show great potential: XGBoost can efficiently process complex data, Monte Carlo simulation can quantify prediction uncertainty, ARIMA is good at capturing time series patterns, and logistic regression is good at classification problems. Therefore, this paper innovatively applies these methods to Olympic medal prediction, hoping to construct a more accurate and stable model, which opens up a new way of thinking for the prediction research in the field of sports, and at the same time provides a reference for the prediction work in other fields.

ISSN: 2688-8645

2. Medal Prediction Model

When predicting Olympic medal counts, XGBoost is suitable for processing large amounts of complex data [1], which can be combined with Monte Carlo simulations to estimate prediction intervals [2]. Logistic model is used to deal with the classification problem, and after classifying the medal situation, predict the possibility of each country winning the prize, and solve the problem of sample imbalance [3]. The ARIMA model captures and predicts the change trend of the medal number based on the time series characteristics [4]. The overall prediction process is shown in Figure 1.

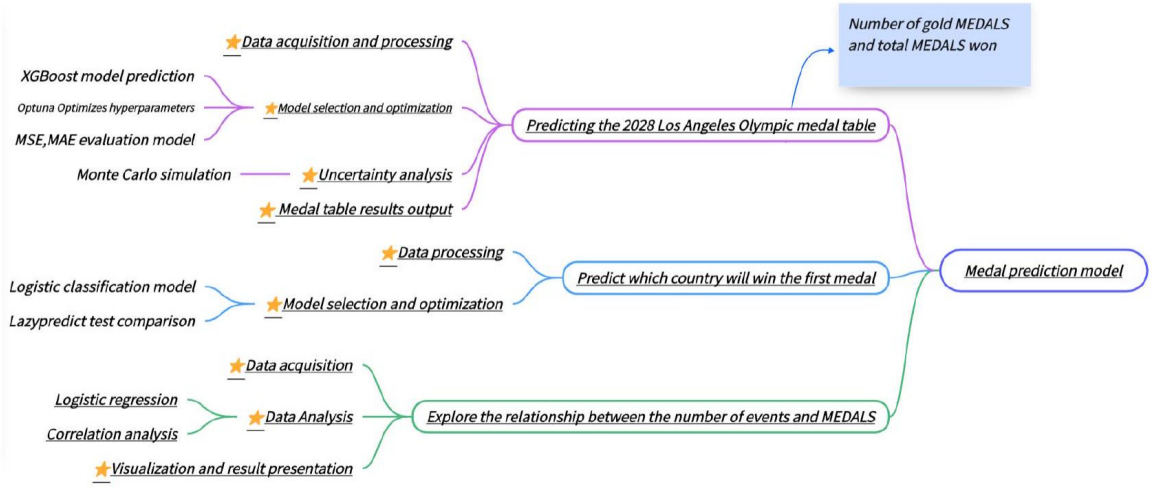

Figure 1. Forecasting process

2.1. Medal Table Prediction

2.1.1. Model Building

XGBoost is based on the gradient Lift decision Tree (GBDT), which is suitable for handling the large data sets in this study and can cope with uneven medal numbers, and it can naturally handle numerical and typing variables [5]. There are a variety of features involved in medal prediction, such as economic indicators, historical achievements, etc. XGBoost can automatically assign weights to these features so that these features are balanced and easy to calculate later. Combined with Monte Carlo simulation, the prediction results of XGBoost can be further used to estimate the prediction interval of medal number and increase the reliability of the results. In XGBoost, the objective function usually consists of two parts: the loss function and the regularization term. The objective function formula is as follows:

(1) objective function:

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

Where n is the data set sample, $l(y_i, \hat{y}_i)$ is the loss function (mean square error). K is the number of trees in the model, the complexity regularization term of the kth tree f_k (used to control model complexity and prevent model overfitting).

Using Optuna for hyperparameter optimization, the objective function is defined as mean square error, that is, the error between the predicted number of medals and the real number of medals (MSE), and the optimal combination of hyperparameters is found by minimizing MSE. MSE visually measures the square of the average error between the predicted and true values. Create a study object and run 50 experiments to optimize the hyperparameters, and finally return the optimal combination of hyperparameters.

ISSN: 2688-8645

2.1.2. Model Predicts:

The prediction function of XGBoost is to sum the prediction results of k trees, and the formula is as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (2)$$

Where $\hat{y}_i$ is the predicted value of the model for the ith sample, $f_k(x_i)$ is the predicted value of the k tree for the ith sample x_i.

2.1.3. Model Evaluation:

XGBoost is commonly used to predict continuous values (predicting Olympic MEDALS). These tasks require evaluating the error between the model output and the actual continuous value, so the prediction effect of the XGBoost model is evaluated by means of the mean square error (MSE), the root mean square error ($RMSE$), and the mean absolute error (MAE) :

Mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the prediction effect of the XGBoost model:

Mean squareerror:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

where $y_i(i = 1,2,\cdots n)$ is the actual value of the medal and $\hat{y}_i(i = 1,2,\cdots n)$ is the predicted value. MSE gives greater weight to larger errors because the errors are squared, which makes it sensitive to outliers.

Root mean square error:

$$RMSE = \sqrt{MSE} \quad (4)$$

It has the same dimension as the predicted value and the true value, so it is more intuitive in interpretation and can more directly reflect the average error size between the predicted value and the true value.

Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)|. \quad (5)$$

MAE is less sensitive to outliers than MSE because it does not square the error and better reflects the average absolute deviation between the predicted value and the true value.

Through these indicators, we evaluate the predictive accuracy and stability of the model. The closer the values of MSE, RMSE and MAE are to 0 , the smaller the gap between the prediction results and the real data, and the better the performance of the model.

We show the XGBoost model evaluation results and the evaluation of the total medal number model in Table 1.

Table 1. Results of XGBoost model evaluation

Metric	MSE	RMSE	MAE
Gold Medals	0.5056	0.7110	0.1735
Total Medals	4.6803	2.1634	0.4245

From Table 1, it can be seen that XGBoost model has a better performance in predicting the number of gold medals and the total medal number.

2.2. Monte Carlo Simulation

Monte Carlo simulation is a numerical method to estimate the results of mathematical problems by random sampling. In this study, we simulate the change of gold medals and total medals in future Olympic Games by generating a large number of random samples. In this way, Monte Carlo simulation is performed to determine prediction intervals based on percentiles. We can

ISSN: 2688-8645

evaluate the uncertainty of the model by obtaining the distribution of gold medals and total medals, as well as confidence intervals.

Assuming that the gold medal prediction follows a normal distribution,

$$\text{LowerBound} = \text{Percentile}(\text{S imulatedPredictions}, 2.5)$$
$$\text{UpperBound} = \text{Percentile}(\text{SimulatedPredictions}, 97.5)$$

Where, Percentile represents the percentile function that calculates the given data.

lower_bound: The lower bound of the calculated forecast range, which is the 2.5 th percentile of all simulated forecasts.

upper_bound: The upper end of the calculated forecast range, which is the 97.5 th percentile of all simulated forecasts.

Historical data from the United States were selected for prediction, and Monte Carlo simulations were used to generate 95% confidence intervals. Simulation results as the chart, we can see that the data presented certain central tendency, most of the simulation results between 40 to 44, confidence interval of the upper and Lower, respectively, marked by red line (95% of the Lower Bound is 39.24 , 95% Upper Bound is 44.87) The Mean of the simulation results is marked with a solid green line, which is 42.03 .

The Monte Carlo simulation of medal counting is shown in Figure 2.

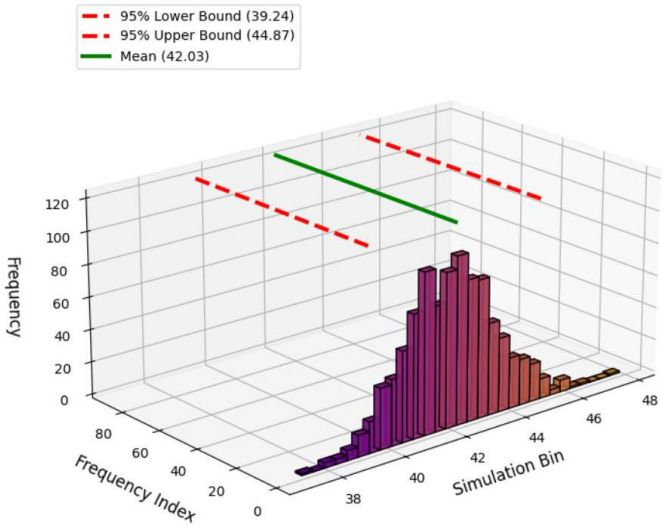

Figure 2. Monte Carlo prediction

3. First Medal Prediction

3.1. ARIMA Model

Gold medals and total medals have obvious time series characteristics in the history of Olympic Games. Using ARIMA model to predict time series data must be stable, if the data is unstable, it is impossible to capture the rule. We use ADF test for stability testing and the unsteady sequences are treated by difference analysis.

3.1.1. ADF Test:

In the ADF test, the T -value is the statistic obtained by t test on y_{t-1} in the test regression equation. By calculating the T -value, we judge the stationarity of the time series and select the appropriate model. Taking the ADF test equation

$$\Delta y_t = \alpha + \beta_1 y_{t-1} + \sum_{i=1}^{p} \beta_{i+2} \Delta y_{t-i} + \epsilon_t \quad (6)$$

ISSN: 2688-8645

With intercept terms and no trend terms as an example, the t-value is calculated by dividing the estimated coefficient by its standard error, $t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$, where $\hat{\beta}_1$ is the estimate of β_1 and $\mathrm{SE}(\hat{\beta}_1)$ is $\hat{\beta}_1$ standard error.

The t value is calculated to measure whether the difference between coefficient and 0 is significant. Under the null hypothesis $H_0: \beta_1 = 0$, the t -value is used to determine whether there is enough evidence to reject the null hypothesis. If the absolute value of the t-value is large enough, it indicates that β_1 is significantly non- 0, and the null hypothesis that the sequence has a unit root (non-stationary) may be rejected.

The critical value is a value determined at different significance levels based on a specific distribution (ADF distribution). In the ADF test, the calculated T-value is compared with the critical value. If the T -value is less than the critical value (usually the left side test), the null hypothesis is rejected and the sequence is considered stationary; On the contrary, the null hypothesis cannot be rejected.

The time series of the original historical data may be non-stationary, and after the first-order difference, a large probability has reached the stationary state, and the second-order difference further ensures the stationarity of the sequence, but some data will be lost. In order to retain the original law of the data, we use first-order difference for subsequent modeling and calculation.

3.1.2. ARIMA Model Establishment

The ARIMA(p,d,q) model can be expressed as:

$\phi(B)$ is an autoregressive polynomial. $\theta(B)$ is the moving average polynomial, where the autoregressive polynomial coefficients ($\phi(B)$) and moving average polynomial coefficients $\theta(B)$) are determined by the maximum likelihood estimation (MLE) method. The goal of maximum likelihood estimation is to find a set of parameter values that maximize the probability that a given data will occur. c is a constant term. ϵ_t is a white noise sequence, approximated by residuals (the difference between the actual value and the model's predicted value). p is the order of the autoregressive term, indicating that observations from the past p period are used to predict the current value. d is the difference order used to convert a non-stationary sequence to a stationary sequence.q is the order of the moving average term, indicating that the error term of the past q period is used to predict the current value.

3.1.3. Model Solving

In this study, ADF test is performed on the input time series data first, and whether the data is stable is judged according to the T -value of the test result. If the T -value is greater than 0.05 , the data is considered non-stationary, suggesting that differential conversion can be considered. If the T -value is less than or equal to 0.05, the data is considered to be stable and suitable for time series modeling. Then the ARIMA model is constructed, and the ARIMA model object is created according to the input time series data and the order of (p, d, q). Then the fit method is called to fit the model, and the parameters such as the autoregressive polynomial coefficient, moving average polynomial coefficient and constant term are determined by the maximum likelihood estimation. Finally, the trained ARIMA model object model_fit is used to predict the data of the future steps period, and the prediction result is returned.

According to Figure 3, the ARIMA model predicts that the number of gold medals will continue to rise after 2020, and the upward trend is more obvious. From the comparison between the prediction line and the historical data curve, the prediction line meets the historical data around 2020, and the growth rate in the subsequent years is relatively large. This suggests that based on the time-series characteristics of the historical data, the ARIMA model believes that the number of gold medals will continue to increase in the future.

ISSN: 2688-8645

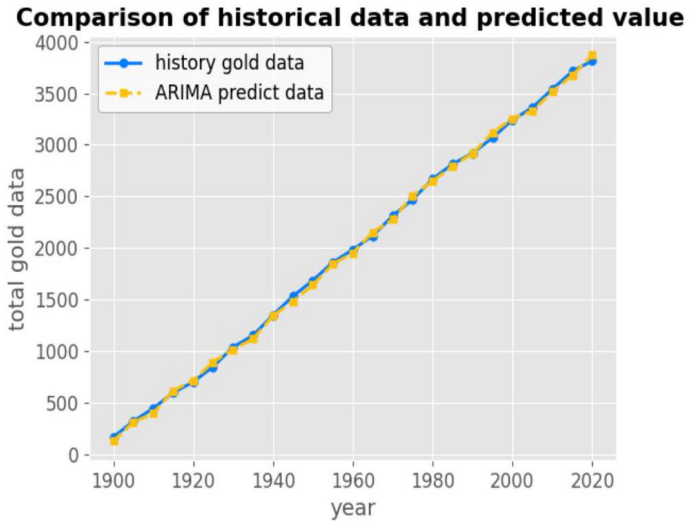

Figure 3. ARIMA trend forecast

3.2. Logistic Regression Model

3.2.1. Model Establishment

Converts the consecutive medal count labels to binary labels (medals won or not), with actual medals won (gold medals or total medals greater than 0) marked as 1 and no medals won as 0.

$$y = \begin{cases} 0(\text{ No prize }) \\ 1(\text{ Win a prize }) \end{cases} \quad (7)$$

Taking the medal data of historical Olympic Games as the data set, taking whether won the gold medal as the label, the year, the number of medals of various types as the characteristics, using logistic regression model for training, the following formula can be obtained:

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}} \quad (8)$$

Where Y indicates whether the gold medal was won (1 indicates that the gold medal was won, 0 indicates that the gold medal was not won), $X = [X_1, X_2, \ldots, X_n]$, is a feature vector, containing the year, historical gold medal number, silver medal number, bronze medal number and total medal number and other information, $b_0, b_1, \ldots b_n$ is the parameter of the model and the regression coefficient is estimated by the maximum likelihood estimation method:

Taking into account the fact that the number of gold medal winners and non-gold medal winners may vary significantly, the scikit-learn library is used to deal with class imbalances and determine weights: The core idea of class_weight='balanced' is to adjust the loss function according to the number of class samples in the training data, so that the model pays more attention to a few class samples during the training process.

3.2.2. Model Evaluation

The goal of Logistic regression is to distinguish between different categories, so its output is probability (predicting the probability value of each category). This requires the use of metrics that can measure classification performance, so ROCAUC is used to evaluate the model.

The ROC curve is composed of a set of points that represent the true positive rate (TPR) and false positive rate at different classification values(FPR). Each point is calculated by the following formula:

$$TPR = \frac{TP}{TP+FN} \quad (9)$$

TP is the number of samples of countries that actually won gold medals and were correctly predicted to be "gold", FN is the sample that actually should have been gold (positive class, but was incorrectly predicted by the model to have no gold medals (negative class)

ISSN: 2688-8645

$$FPR = \frac{FP}{FP+TN} \tag{10}$$

FP is the number of samples that the model incorrectly predicted as "gold medals" when in fact those samples did not win gold medals

TN is the sample that did not actually win a gold medal and was correctly predicted by the model to have no gold medals.

By adjusting the medal values predicted by the model, a series of TPR and FPR values can be obtained and eventually plotted into the ROC curve as shown in Figure 4.

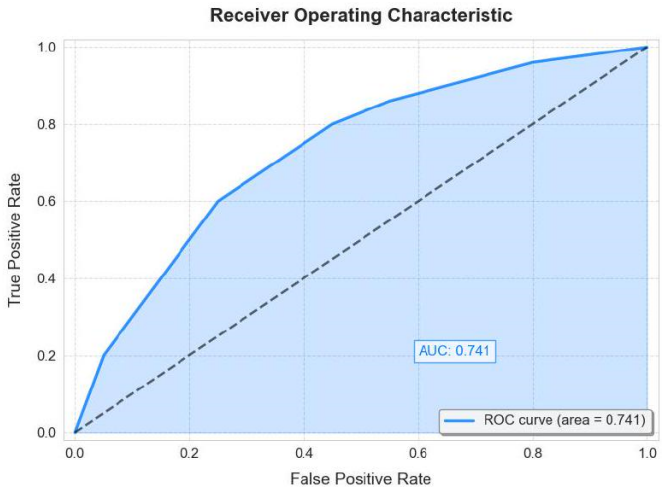

Figure 4. ROC curve

ROCAUC is the area under the ROC curve, which is formulated as the difference between two integrals, or by a classifier The rate is calculated to approximate the numerical integration. The formula is defined as:

$$AUC = \frac{\sum_{i \epsilon Pos} R\mathrm{Rank}(i) - \frac{n_{\mathrm{Pos}} * (n_{\mathrm{Pos}} + 1)}{2}}{n_{\mathrm{Pos}} * n_{Neg}} \tag{11}$$

Where n_{Pos} represents the number of samples that actually won gold medals.

Where n_{Neg} represents the number of samples that did not actually win a gold medal.

Rank (i) represents the ranking of the model's confidence in the prediction of positive samples, and the sample with the highest prediction probability ranks first.

The code outputs the ROC AUC value 0.7403825829634327 of the ARIMA model. The closer the value of ROC AUC is to 1, the stronger the ability of the model to distinguish between positive and negative samples, indicating that the model has the ability to distinguish between medals or not.

4. Conclusion

In this study, a multi-model fusion Olympic medal prediction system is successfully constructed, which effectively improves the accuracy and reliability of Olympic medal prediction through the synergy of different models. In the medal list prediction, the XGBoost model shows powerful data processing ability, which can effectively integrate multi-dimensional features such as economic indicators and historical achievements, and after combining with Monte Carlo simulation, it not only gives the predicted value of the number of medals, but also evaluates the uncertainty of the results by determining the prediction interval, which makes the prediction results more practical.

For the prediction of the first medal, the ARIMA model makes full use of the time series characteristics of medal data to accurately capture the trend of the number of gold medals; the

ISSN: 2688-8645

logistic regression model transforms medal counting into a classification problem, and by reasonably dealing with the sample imbalance problem, the model shows a good classification performance in terms of ROCAUC evaluation, and effectively determines the likelihood of winning a medal for each country.

However, there is still room for optimisation of the model. In the future, we can further explore the potential factors affecting medal counts, such as changes in athletes' status and the environment of the event venue, to expand the data dimensions; we can also try to integrate the deep learning model into the existing system, to explore more complex non-linear relationships, and continuously improve the performance of the prediction model, so as to provide more accurate results for the prediction of Olympic medals, and to provide stronger support for related research and decision-making.

Acknowledgments

The authors would like to thank Dr. Cai Shuting for her important and useful suggestions to im prove our paper.

References

[1] JIANG Dong, LUO Yayuan, JIN Haibo, et al. Prediction of building settlement based on combined ARIMA-LSTM-XGBoost model[J/OL]. Civil Engineering Information Technology,1-6[2025-02-12]. http: //kns.cnki.net/kcms/detail/11.5823.tu.20250212.1053.002.html.

[2] Jingrun Zhang. Research on optimisation method of complex electronic equipment spare parts based on Monte Carlo simulation[J]. Daily Electric Appliances, 2024, (09):6-12.

[3] XIAN Yurui, WANG Jing, ZHANG Min. Comparison of risk prediction performance based on decision tree model and logistic regression model in the occurrence of multidrug-resistant bacterial infections in neurocritical care patients[J/OL]. Chongqing Medicine,1-11[2025-02-12].http:// kns. cnki. net/kcms/detail/50.1097.r.20250208.1025.012.html.

[4] Chen Xuemei. Forecasting and analysing the incidence of HIV/AIDS in Yunnan Province from 2010 to 2020 based on ARIMA model[J]. Modern Medicine and Health,2025,41(01):11-17.

[5] HANG Yanchun, LI Yang, WANG Shijun, et al. Fault classification method for distribution network information review centre based on improved gradient boosting decision tree[J/OL]. Automation Technology and Application,1-6[2025-02-12]. http: // kns. cnki. net/ kcms/ detail/ 23. 1474.TP. 20241230.1543.205.html.