Olympic Medal Prediction Model Based on Existing Data Set

Panniankuan Chen [1, *], Lu Xu [2], Siwei Zhang [2]

[1] School of Electronic Science and Technology, Hohai University, Changzhou, China, 210024

[2] School of Communication Engineering, Hohai University, Changzhou, China, 210024

* Corresponding Author Email: 2262910232@hhu.edu.cn

Abstract. The success of the Paris Olympics revealed to the world the meaning of the Olympic motto, highlighting unity and excellence in global sports. The most eye-catching part of the Olympics is obviously the“medal table”, which also stimulates interest among researchers and policymakers in accurately predicting the number of medals. The study developed a medal prediction system to quantitatively forecast medal acquisition and establish comparative benchmarks for national sports programs. The computing system was initially established using a multiple linear regression model, with missing values of independent variables determined through an ARIMA model, enabling precise completion of medal table predictions and determination of prediction intervals. TOPSIS and the Entropy Weight Method were then employed to construct a complete project evaluation system, which provides reliable statistics for different countries while offering constructive advice for further development of sports in each country, thereby assisting more countries in winning medals.

Keywords: Olympic Medal Predictions, Multivariable Linear Regression Model, ARIMA, TOPSIS.

1. Introduction

The significance of Olympic medals is multi-dimensional. The medals are not only a recognition of the excellent performance of the Olympians themselves, but also a reflection of the overall strength of the country. Countries that rank high on the medal table always attract more attention, which in turn gives them greater influence in both competitive sports and cultural fields. While it is true that the uncertainty of results is one of the attractions of competitive sports, accurate predictions of the number of medals to be won, as well as the programs to be won, can often help different countries to focus their resources on different programs, and thus achieve greater success!

Research on the factors influencing the production of Olympic medals usually adopts discriminant analysis, regression analysis, etc. Bernard and Busse innovatively applied the Cobb-Douglas production function to the analysis of medal acquisition, and came to the conclusion that the larger the country's population and the higher the per capita GDP of the country serving as the host of the current Olympic Games, the more medals the country would win, and it aroused widespread international attention [1]. Wang Guofan and Tang Xuefeng, on the other hand, systematically categorized the Olympic medal prediction methods into time series model, empirical model and neural network model on this basis, which provided an attempted practice for reference [2]. However, all the above studies focus on the influence of socio-economic indicators on the overall number of national medals.

In recent years, this research has been heating up, and the most common prediction model is based on randomized decision forests. Schlembach greatly improved the prediction accuracy of Olympic medals by synthesizing the dataset with a two-stage randomized forest model [3]. Shi Huimin, on the other hand, used a machine learning approach to focus on the differences in predictability between different sports, thus providing a feasible solution for countries to explore the extent to which different sports are associated with the acquisition of medals in their own countries [4]. And more models are applied to this problem such as GA-BP and logistic regression model, the socioeconomic machine learning model, the LSTM and TOPSIS model, the weighted fusion model, the SEQ2SEQ model, the Interpretable Machine Learning and so on [5-10].

2. Medal prediction modeling

The data used in this article is sourced from www.contest.comap.com.

2.1. The Two-Stage ARIMA-Regression Hybrid Model

Exploring the impact of various data information on the number of medals won by a country as a result of its participation in the Olympic Games is clearly a multivariate, single-output modeling problem. In order to capture more accurately the complex effects of the various factors on the final number of medals obtained, and ultimately to provide an accurate prediction of the medal table for the next Olympic Games, a multivariate linear regression model was developed.

The research team first conducted a preliminary analysis of the available data files and identified the following five variables by comparing data between countries and searching for relevant literature. They are: the total number of athletes participating in each sport (X_1), the total number of athletes who have won awards (X_2), whether hosted (X_3), the total number of Olympic events (X_4), and the total number of projects involved (X_5).

Among the independent variables, four quantitative variables were available, while the qualitative variable X_3 required value assignment before model input. The following valuation protocol was applied:

$$X_3 = \begin{cases} 1, & Yes \\ 0, & No \end{cases} \tag{1}$$

The dependent variable (Y) is the number of medals won by the country. In this study:Y_1 represents the aggregate count of gold medals and Y_2 denotes the summation of all medals.

Based on these definitions, the research team establishes the following multivariate linear regression model.

$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i X_i + \varepsilon \tag{2}$$

Where β_0 and β_i are the model parameters, and ε is the error term.

Although the multivariable linear regression model allows the researchers to predict the number of medals from the independent variables for the required solution, the only independent variables that can be identified for the 2028 Summer Olympics in Los Angeles, CA as of today are the host country and the total number of events. Therefore, the ARIMA model - one of the most popular linear models in contemporary time series forecasting - was employed to construct a more accurate mixed model for estimating the remaining unknown independent variables.

ARIMA model is fully known as Autoregressive Integrated Moving Average Model, which consists of the following three main components: autoregressive model (AR), differential process (I) and moving average model (MA).

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \xi_t \tag{3}$$

The above equation is the mathematical expression of the AR model, where c is a constant term, φ_1 to φ_p are the parameters of the AR model, which are used to describe the relationship between the current value and the values at the past p time points. And Y_t is the time series data we are considering. It can be seen that the AR part is the autoregressive part used to deal with the time series, which takes into account the effect of observations from several periods in the past on the current values.

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{4}$$

The above equation is the mathematical expression of the MA model, where μ is a constant term and θ_1 to θ_p are the parameters of the MA model, which are used to describe the relationship between the current value and the error at the past q time points. It can be seen that the MA part is used to deal with the moving average part of the time series, which takes into account the effect of past forecast errors on the current value.

The differencing process (I) is used to smooth a non-smooth time series and the order of differencing is recorded as d. In general, one order is sufficient.

The combination of the above three models is known as the ARIMA (p, d, q) model, which is able to extract the time series patterns hidden behind the data by means of autocorrelation and differencing of the data, and then use these patterns to predict future data.

2.2. Construct Project Evaluation System Based on Entropy Weight Method of TOPSIS

The entropy weight method is an objective assignment method that relies on the data itself to derive the weights, thus providing a basis for the comprehensive evaluation of multiple indicators. However, the application of the entropy weight method requires the initial normalization of the data matrix. Therefore, the TOPSIS algorithm is adopted for data preprocessing. Based on the entropy weighting evaluation algorithm model incorporating data preprocessing via the TOPSIS algorithm, the importance of different sports to the country is assessed. Furthermore, the impact of sports selection on the final medal outcomes is explored using the entropy weighting evaluation algorithm model.

Data preprocessing using the TOPSIS model begins with the normalization of the raw data. The raw data can be categorized into three types of indicators, as shown in Table 1.

Table 1. Indicator Category Table

Name of indicator	Indicator characteristic
Benefit-type indicator	The bigger the better.
Interval-type indicator	The closer to a certain value, the better.
Cost-type indicator	The smaller the better.

Normalization is the uniform conversion of all indicator types into Cost-type indicators according to the relevant formulas, so that they can be processed and compared in a uniform manner, thus providing the basic conditions for subsequent standardized processing.

The purpose of standardization is to eliminate the effects of different indicator scales. If the original standardization matrix is denoted as X and the normalization matrix is denoted as Z, the following equation is satisfied:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{5}$$

Where x_{ij} is the value in the original standardization matrix, where z_{ij} is the value in the normalization matrix.

That is, divide each element by the sum of the squares of the elements in the column under the root sign to get the normalized data.

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij}\ln\left(p_{ij}\right)(j=1,2,\ldots,m) \tag{6}$$

Where e_j is the j-th metric information entropy, where p_{ij} is the probability.

$$\mathrm{d}_j = 1\text{-}\mathrm{e}_j \tag{7}$$

Where d_j represents the information utility value.

Finally, the information utility value is normalized to obtain the entropy power.

$$W_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{8}$$

Where W_j means the entropy power.

3. Results

3.1. The result of the Two-Stage ARIMA-Regression Hybrid Model

Since there are numerous countries participating in the Olympics, China is selected as a representative example. By substituting historical data from previous years into the model for fitting, the relevant parameters of the linear regression model can be obtained. The calculated model parameters are presented in Table 2.

Table 2. Statistical table of model parameters

Model parameter	β (Gold)	β (Total)
β_0	-18.7965	-23.7113
β_1	-0.1309	-0.1951
β_2	0.1284	0.3111
β_3	21.6916	18.9951
β_4	0.1439	0.1779
β_5	0.2809	0.5198

After substituting the model parameters, the resulting final fitted regression model is

$$Y_1 = -18.7965 - 0.1309X_1 + 0.1284X_2 + 21.6916X_3 + 0.1439X_4 + 0.2809X_5 \tag{9}$$

$$Y_2 = -23.7113 - 0.1951X_1 + 0.3111X_2 + 18.9951X_3 + 0.1779X_4 + 0.5198X_5 \tag{10}$$

At this stage, model checking is carried out. R^2 is called the coefficient of determination of the model, and the closer its value is to 1, the better the regression model fits the data. Model Y_1 has an R^2 of 0.9880 and model Y_2 has an R^2 of 0.9597, indicating that both models fit the data reasonably well and therefore produce more accurate predictions. This is shown in Figure 1.

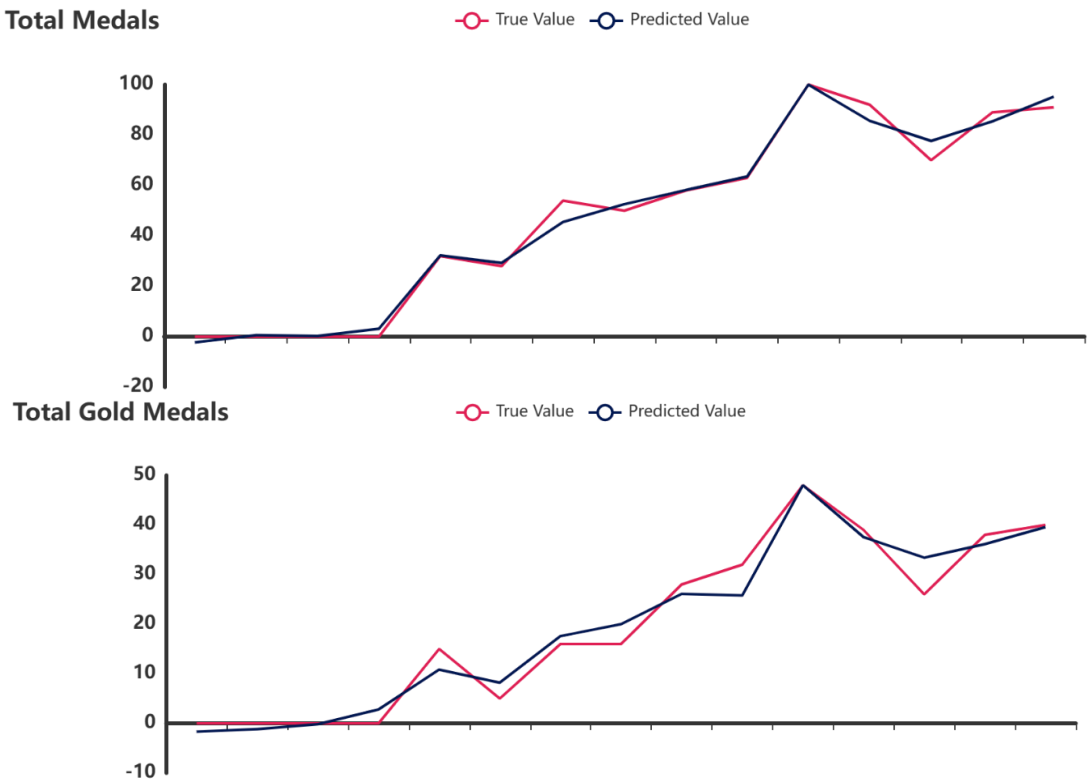

Figure 1. Plot of predicted medals fitted to actual medals awarded

Figure 1 shows that our fitting results are very similar to the actual results, illustrating the accuracy of our model.

As for the ARIMA model, the time series must first be smooth, which requires us to perform stability tests on the model. The results are shown in Figure 2 and Figure 3.

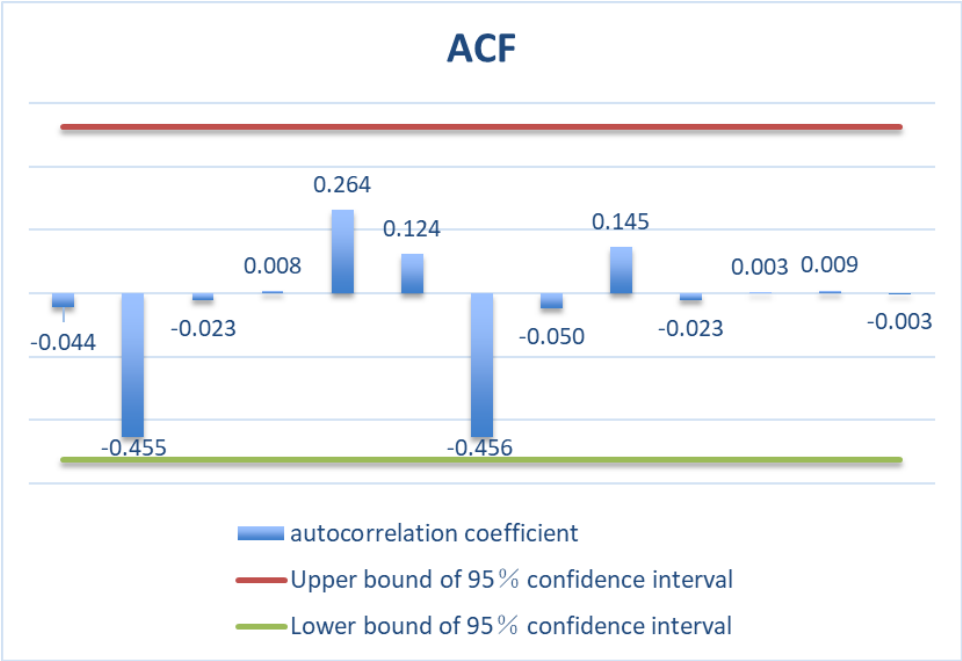

Figure 2. The ACF of ARIMA Model Chart

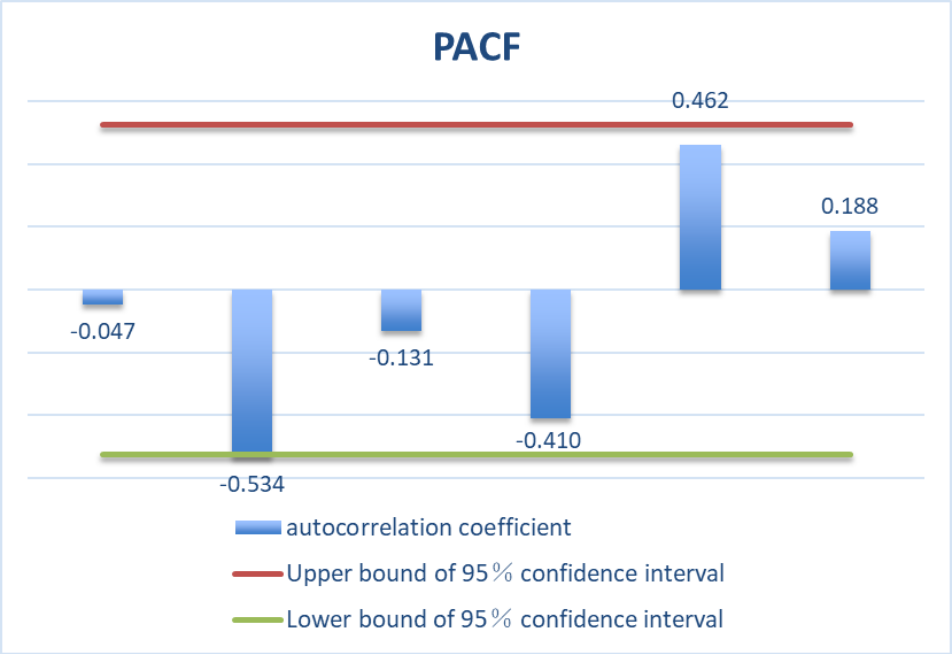

Figure 3. The PACF of ARIMA Model Chart

Figure 2 and 3 are the ADF test table, the autocorrelation chart of final difference data (ACF), and the partial autocorrelation chart of final difference data (PACF) drawn to predict the total number of participating events in the new Olympic Games in the case of China. The parameters p and q were initially estimated using ACF and PACF plots, while the order of differencing d was determined through analysis of stationarity. Model optimization was performed by considering both the Akaike Information Criterion (AIC) and the statistical significance of the parameters. Finally, the residuals were tested and confirmed to exhibit white noise characteristics. After finding the optimal parameters based on the AIC information criterion, the final model result is established as an ARIMA (0,1,0) model. At this point, the model's goodness-of-fit R^2 is 0.803, which proves that the model performs well, meets expectations, and is able to make more accurate predictions, as shown in Figure 4.

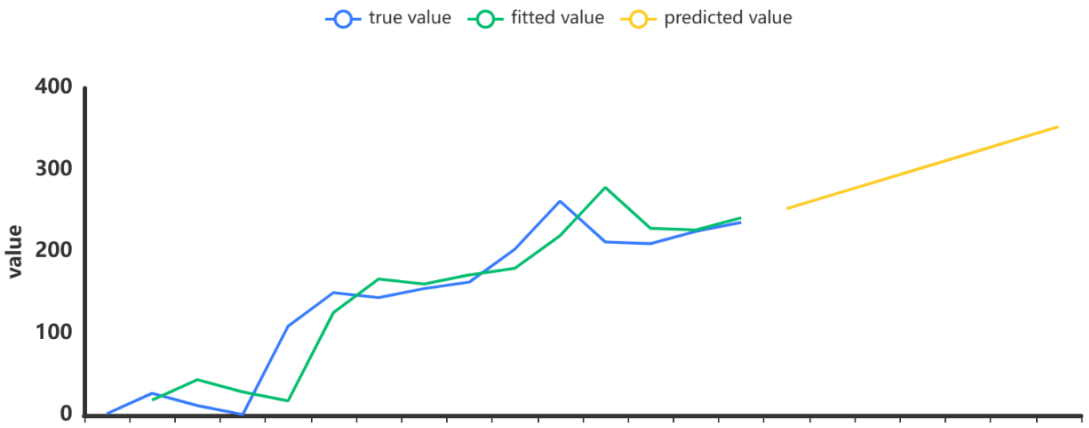

Figure 4. Schematic comparison of actual and projected values

Based on the ARIMA model described above, the required independent variables were accurately predicted. Subsequently, a multivariable linear regression model was employed to project the number of medals for each country in the 2028 Summer Olympics in Los Angeles, USA. The prediction results are presented in Table 3. (Due to space constraints, only the top 16 countries are listed here).

Table 3. 2028 Prediction Medal Table

Ranking	Country	Gold	Total	Ranking	Country	Gold	Total
1	United States	50	109	9	Netherlands	11	32
2	China	41	100	10	France	9	32
3	Australia	19	55	11	New Zealand	9	20
4	South Korea	18	39	12	Canada	6	25
5	Great Britain	16	52	13	Brazil	6	24
6	Germany	15	39	14	Hungary	6	19
7	Japan	12	26	15	Romania	5	12
8	Italy	11	34	16	Spain	4	22

3.2. The result of entropy weight method of the TOPSIS model

By multiplying the entropy weights with the standardized data, the corresponding scores for each sport in the country can be obtained. And the higher the scores, the more important the sport is to the country. Taking China as an example, the following Figure 5 shows the corresponding scores for each of China's major competing sports.

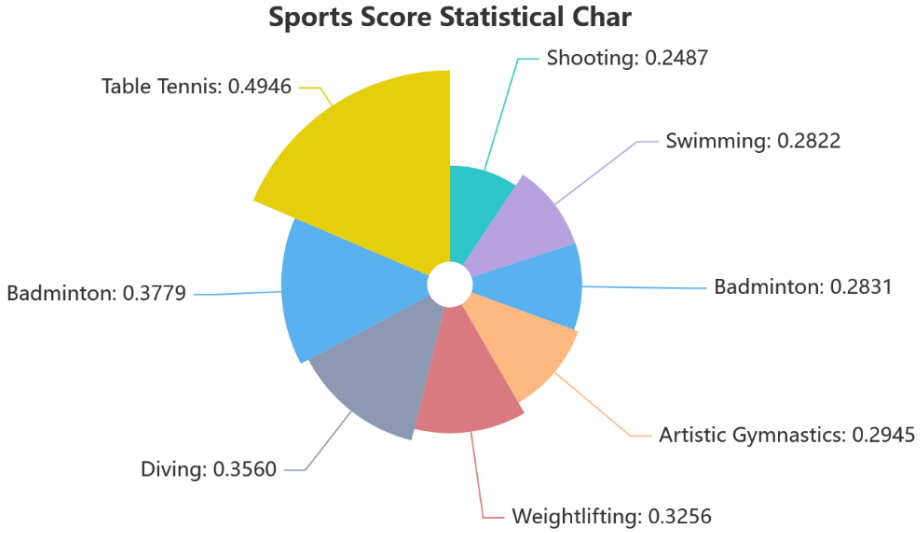

Figure 5. Sports Score Statistical Chart

From Figure 5, it can be seen that table tennis, badminton, diving and weightlifting are more important to China than the other four major sports. These sports also tend to bring more medals to the country. So China should pay more attention to these competitions. In the same way, other countries can calculate the most important sports for themselves.

4. Conclusion

Olympic medals have multifaceted significance and require accurate predictive modeling to guide resource allocation and strategic planning in participating countries. This study innovatively integrates multivariate linear regression and ARIMA hybrid models to cope with time trends and multifactorial influences. By utilizing the ARIMA model to predict variables over time and the regression model to capture the complex interactions between key predictors, the proposed system achieves a high level of accuracy. In addition, a robust evaluation framework based on the entropy weighting method of TOPSIS is developed, enabling countries to prioritize sports with the highest strategic value, as demonstrated in the case study of China. The experimental results validate the reliability and generalizability of the model, not only in predicting the medal results of the 2028 Olympic Games in Los Angeles, but in fact, the research methodology is applicable to any similar

sport and competition and can be dynamically adjusted by parameters to adapt to different real-life situations. Thus, it provides a scalable framework for strategic decision-making in competitive sports. In the future, the research team will explore the integration of more advanced machine learning methods with existing models to enhance the predictive capability for nonlinear relationships and long-term dependencies. Additionally, the model's applicability will be validated across more countries to examine its transferability under different sports development frameworks.

References

[1] Bernard A B, Busse M R. Who wins the Olympic Games: Economic resources and medal totals [J]. Review of economics and statistics, 2004, 86 (1): 413 - 417.

[2] Wang Guofan, Tang Xuefeng. Domestic and foreign research dynamics and development trend of Olympic medal prediction [J]. China Sports Science and Technology, 2009, 45 (06): 3 - 7+135.

[3] 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.

[4] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? --Based on Interpretable Machine Learning Perspective [J]. Journal of Shanghai University of Physical Education, 2024, 48 (04): 26 - 36.

[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] 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.

[7] Yan D S. Olympic Medal Prediction and Analysis Based on LSTM And TOPSIS Models [J]. 2025.

[8] Jin Q F, Yao R X. Prediction Study Of 2028 Olympic Medal Table Based on Weighted Fusion Modeling [J]. 2025.

[9] Yao Q L, Cheng F, Guo Y P, et al. Olympic Medal Prediction Based on SEQ2SEQ Model and Tpe Optimization [J]. 2025.

[10] Huimin S H I, Dongying Z, Yonghui Z. Can Olympic Medals Be Predicted? Based on the Interpretable Machine Learning Perspective [J]. Journal of Shanghai University of Sport, 2024, 48 (4): 26 - 36.