Research on Olympic Medal Prediction Based on PLSR

Zhikang Yuan*

School of Economics and Management, Tongji University, Shanghai, China, 201804

*Corresponding author: yzk18306159695@163.com

Abstract. The medal table at every Olympic Games can always get keen attention from all over the world. However, due to the dynamic nature of competitive sports, which leads to the low reference value of early historical data and the small amount of recent data, it is difficult to predict the number of medals in the next Olympic Games. In order to solve this problem, this article plans to establish a comprehensive evaluation system of national sports strength, using the PLSR algorithm suitable for small sample size, combined with the bootstrap method to predict the number of medals of some countries in the next Olympic Games. At the same time, for countries that have not won Olympic medals in history, the logistic regression algorithm is used to estimate the probability of the number of medals exceeding 0 in the next Olympic Games. By integrating both macro and micro dimensions of sports competitiveness, this article not only addresses data limitations but also pioneers a systematic approach to capture national strengths, providing a missing link between traditional socio-economic analysis and sport-specific dynamics.

Keywords: PLSR, Prediction Model, Bootstrap Method, Logistic Regression.

1. Introduction

The Olympic Games, as the pinnacle of international sports competition, serve not only as a platform for athletic excellence but also as a reflection of national prestige and socio-economic development. In terms of medal prediction for the Olympic Games, existing studies mainly focus on the impact of macro factors such as socio-economic factors on the medals won by countries. Bernard et al. [1] used the Logit model to analyze the Olympic medal table and found that the number of Olympic medals won by an Olympic team is related to the population of the country or region it represents, the GDP per capita, and whether it is the host country of the Olympic Games. Schlembach et al. [2] used a random forest model to predict the performance of each team at the Olympic Games and evaluated the contribution of different characteristic variables to the prediction. SHI Huimin et al. [3] used the random forest model to predict the performance of each Olympic team in each sub-event, and compared the differences in the predictability of the performance of different Olympic events. However , these articles ignored sports-specific factors (such as the discipline dominance and the presence of elite athletes) and the difficulty of traditional models to deal with small and heterogeneous datasets.

This article addresses these gaps by proposing a dual innovation in data utilization and methodology. First, this article integrates macro socio-economic indicators with micro-level sport-specific metrics—such as historical medal trends, athlete participation intensity, and national dominance in events—to construct a comprehensive evaluation system. Second, this article introduces a hybrid framework: PLSR handles multicollinearity and small samples, Bootstrap quantifies prediction uncertainty, and logistic regression estimates breakthrough probabilities for emerging nations. Additionally, by classifying countries into distinct categories (e.g., high participation/high medals vs. low participation/low medals), the model tailors predictions to heterogeneous competitiveness levels, a dimension overlooked in homogeneous approaches.

2. The preliminary work for prediction

2.1. Indicator System Establishment

The following indicators have been selected for the prediction of the number of medals for a given country in a given sport, as illustrated in Table 1.

(1)Total Nations in the Sport (TNS)

TNS denotes the total number of nations participating in the sport, reflecting the global participation in the sport. More countries participating means more competition in the sport, so TNS is an important indicator for predicting the number of medals a country will win in the sport.

(2)Total Athletes in Olympics/ Total Athletes in the Sport (TAO/TAS/TES)

TAO represents the total number of athletes from a country registered for a specific sport, TAS refers to the total number of athletes from a country entered in the Olympic Games, while TES indicates the total number of events in which a country has fielded athletes for the Olympic Games. These indicators reflect a country's commitment to both the sport and the Olympics, as well as its competitive strength in the relevant events. As such, they serve as key factors in predicting medal outcomes.

(3)Host Nation (H)

In order to quantify the host nation effect, i.e., the country hosting the Olympics usually performs much better at that Olympics and is often able to exceed its usual level of competition[4], the article sets it as a categorical variable, which helps to more accurately capture the potential performance enhancement of host countries[5].

(4)Historical Medals of Previous THREE/ TWO/ ONE Year(s) (HM3/HM2/HM1)

Since the medal counts from the previous three editions offer a more stable performance trend, the data from the last two editions better reflect the Olympic cycle and preparation rhythm, while the data from the most recent edition is the most directly relevant to the upcoming Olympic Games. By using medal counts from these varying time spans as indicators respectively, the article can provide a more accurate and comprehensive prediction of medal outcomes.

(5)Strength (S)

Some countries have a significant dominance in particular events, and this dominance may have a significant impact on their medal totals. In order to quantify this phenomenon, this article lets SA be number of "dominant event" in the sport for that nation with the following categorization criteria:

When a country wins more than 50% of the medals in a particular sport, that is considered to be the country's "dominant event", while others are categorized as "common event", which can demonstrate more explicitly the core competitiveness of countries in different fields and provide a more targeted reference for medal forecasting.

(6)Superior Athletes (SA)

This article defines a "superior" athlete as one who has won two or more consecutive medals. SA is the total number of "superior" athletes in the sport for that nation.[6]

Table.1. Indicator System

Number	Indicator	Description
1	TNS	Total Nations in the Sport
2	TAO	Total Athletes in the Olympics
3	TAS	Total Athletes in the Sport
4	TES	Total Events in the Sport
5	H	Host Nations
6	HM_3	Historical Medals of Previous Three Years
7	HM_3	Historical Medals of Previous Two Years
8	HM_3	Historical Medals of Previous One Years
9	S	Strength
10	SA	Superior Athletes

2.2. Nation Classification

Given the significant disparities in Olympic medal competitiveness among nations, this article classifies the participating countries based on factors such as the number of times they have competed and the average number of medals won per participation. For each category, this article develops customized prediction models to enhance the accuracy and effectiveness of our forecasts.

After careful consideration and adjustments, this article classifies countries into four groups: High participations and high medals (HH), High participations and low medals (HL), Low participations and low medals (LL), and No participation in the last seven editions (N), as shown in Figure 1.

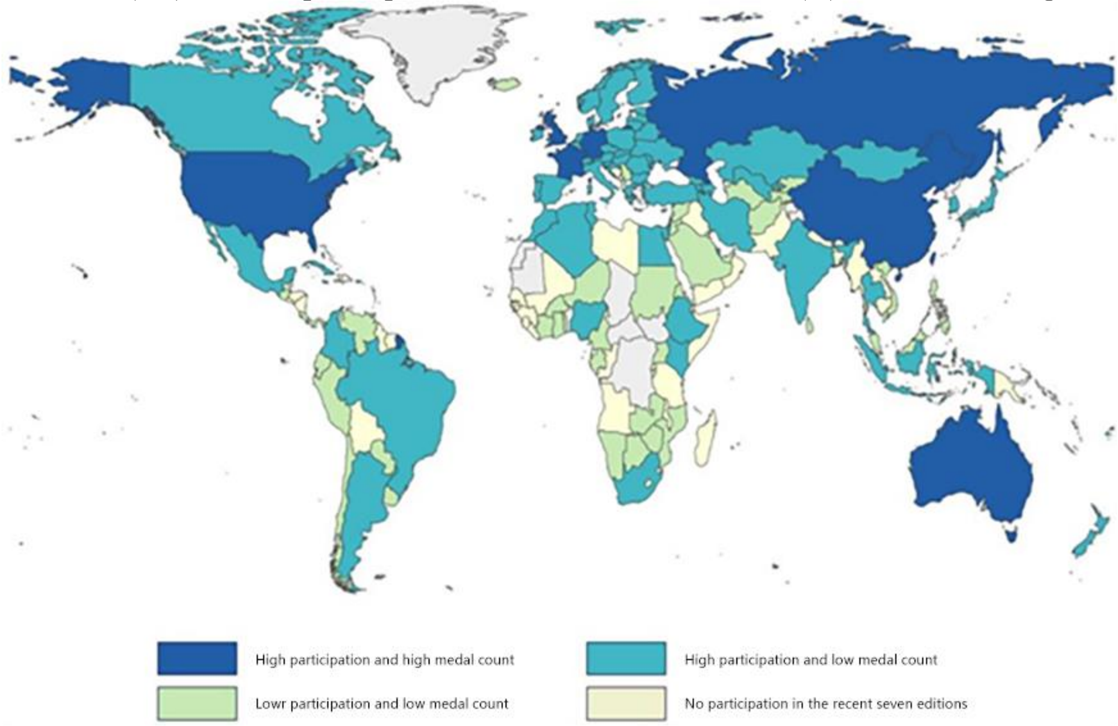

Figure 1: Nation Classification (Map data from Google Maps)

Assuming participating countries remain relatively stable, this article hypothesizes that countries in category N will not participate in the upcoming 2028 Olympic Games and therefore do not predict their medal counts. For the other three categories, this article builds separate models to predict their medal counts, including both total medals and gold medals.

3. Establishment of predictive models

3.1. Prediction of the number of medals

3.1.1 Partial Least Squares Regression (PLSR)

Partial Least Squares Regression (PLSR) is a regression method that can handle multicollinearity and effectively manage high-dimensional data, which is well suited to our chosen indicators[7].

As mentioned above, this article uses x'_{ij} to represent the normalized value of the j-th indicator for a certain country and sport in the i-th year, and y_i denotes the number of medals in the i-th year (j $\in$ {1,2,...,8}, represents the eight indicators in Figure 2, and i $\in$ {1,2,...,8}, represents the eight Olympic years selected (1996-2024)). In addition, $X \in R^{8\times 8}$ and $Y \in R^{8\times 1}$ represent the matrices formed by the two respectively.

PLSR extracts latent variables (also known as components) by maximizing the covariance between the predictor matrix X and the response variable matrix Y. The latent variables T $\in$ $R^{8\times t}$ are linear combinations of the input matrix, defined as:

$$T = XP \tag{1}$$

Where $P \in R^{8\times t}$ is the loading matrix representing the contribution of each input feature to the latent variables, and t is the number of latent components. This process is iterative, with each step selecting components that maximize the covariance between T and the response variable Y.

After extracting the latent variables, a linear regression model is established between the latent variables and the response variable. The model is given by:

$$Y = TB + E \tag{2}$$

Where $B \in R^{t\times 1}$ is the regression coefficient matrix representing the relationship between the latent variables and the response, and E is the error term. The regression coefficients B are estimated by minimizing the sum of squared errors:

$$\hat{B} = (T^T T)^{-1} T^T Y \tag{3}$$

Once the regression coefficients $\hat{B}$ are estimated, the model can be used to predict new response variables $\hat{Y}$. And to avoid overfitting, the optimal number of latent components t is typically chosen through cross-validation, which is eventually determined to be 5. Using the trained model, the predicted medal number $\hat{y}$ is obtained as:

$$\hat{y} = X\hat{P}B \tag{4}$$

where $\hat{P}$ is the trained loading matrix, and $\hat{B}$ is the estimated regression coefficient.

3.1.2 Deriving Confidence Intervals with Bootstrap Method

After the trained PLSR model has made the medal count predictions, we use the Bootstrap method to estimate the confidence intervals of these predictions.[8] This approach enables us to quantify the uncertainty inherent in the model's predictions. The specific modeling steps are as follows:

First, randomly sample with replacement from the training dataset D to create multiple new training datasets $D_b^* = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \dots, (x_n^*, y_n^*)\}$, each likewise containing n samples. After each sampling, some data points may be repeated in D_b^*.

Then, for each generated dataset D_b^*, the PLSR model is retrained to obtain a new set of parameters, resulting in B retrained PLSR models. Each of these B trained PLSR models is used to predict the future medal count, yielding B sets of predicted values $\hat{y}_b^*$. For this series of predictions, we calculate the standard deviation (SD) of the predicted numbers of medals.

Finally, the confidence interval upper and lower bounds are:

$$\hat{y} \pm z_{\frac{\alpha}{2}} \cdot \mathrm{SD} \tag{5}$$

where $z_{\alpha/2}$ is the critical value corresponding to the $1-\alpha/2$quantile of the normal distribution, taken as 1.96 for a 95% confidence interval (i.e., $\alpha = 0.05$).

Through these steps, the Bootstrap method provides an uncertainty estimate for the BKA-SVR's predictions, giving a confidence interval for the predicted values and helping assess the reliability of the model's predictions on new data.

3.1.3 Forecast Results

Based on the previous nation classification illustrated in Figure1 the article selected 14 nations including the United States, China, Australia and United Kingdom from the HH category, Japan, Netherlands, Italy, Spain and Egypt from the HL category, and Singapore, Malaysia, Philippines, Syria and Dominica from the LL category. Using the model, the article predicted the total number of medals and the number of gold medals for these countries in various events.

After that, the predicted results of all the sports of all these countries were summed up respectively to get the predicted values and confidence intervals of the total number of medals and gold medals, and compared with the results of the 2024 Olympic Games, so as to judge whether they are progressing or regressing as a whole, with part of the results presented in Figure 2, Figure 3.

Nation	2028
USA	120.7523
China	95.2594
Australia	44.3893
UK	56.4536
Japan	56.9207
Netherlands	45.9612
Italy	42.5647
Spain	36.8631
Egypt	25.9073
Singapore	8.7842
Malaysia	13.8421
Philippines	7.0719
Syria	0.8943
Dominica	0.5601

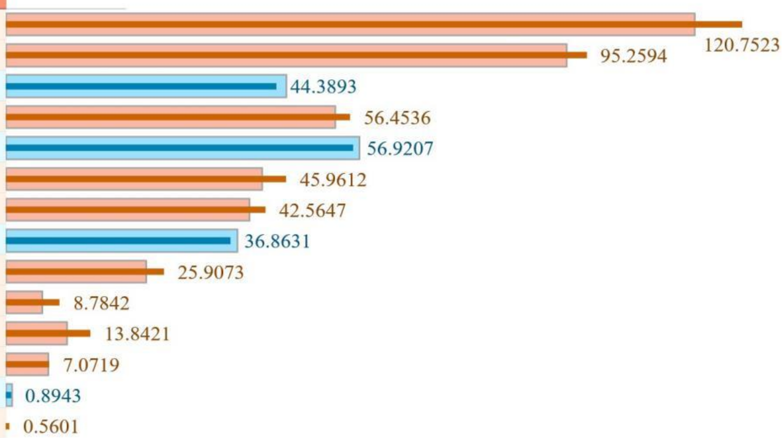

Figure 2: 2028 Total Medals compared to 2024

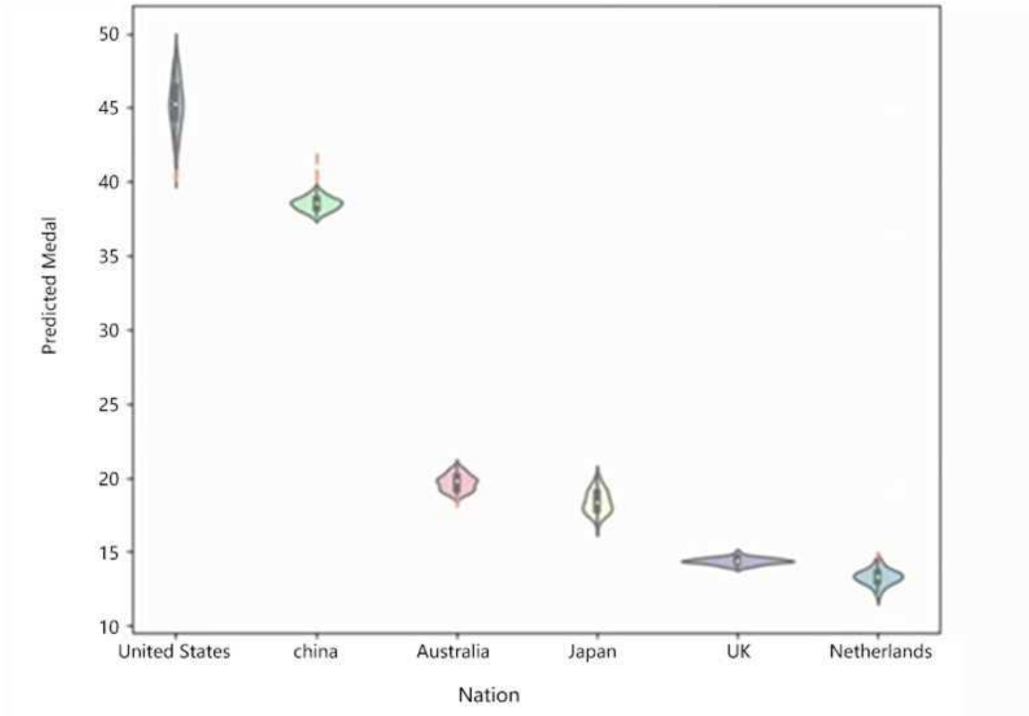

Figure 3: Gold Medal Predictions for Several Nations

3.1.4 Sensitivity Analysis

Given that the model incorporates various indicators, the article performed a sensitivity analysis to assess the model's robustness, taking USA, China, Japan and Singapore as examples. In this analysis, this article adjusted each indicator by ±10% from its current normalized value and observed the corresponding changes in medal predictions, as shown in Figure 4.

The results show that these changes are relatively small, with the largest being about 7.6%. Overall, these results indicate that changes in individual metrics do not lead to significant changes in predictions, suggesting that our model is robust.

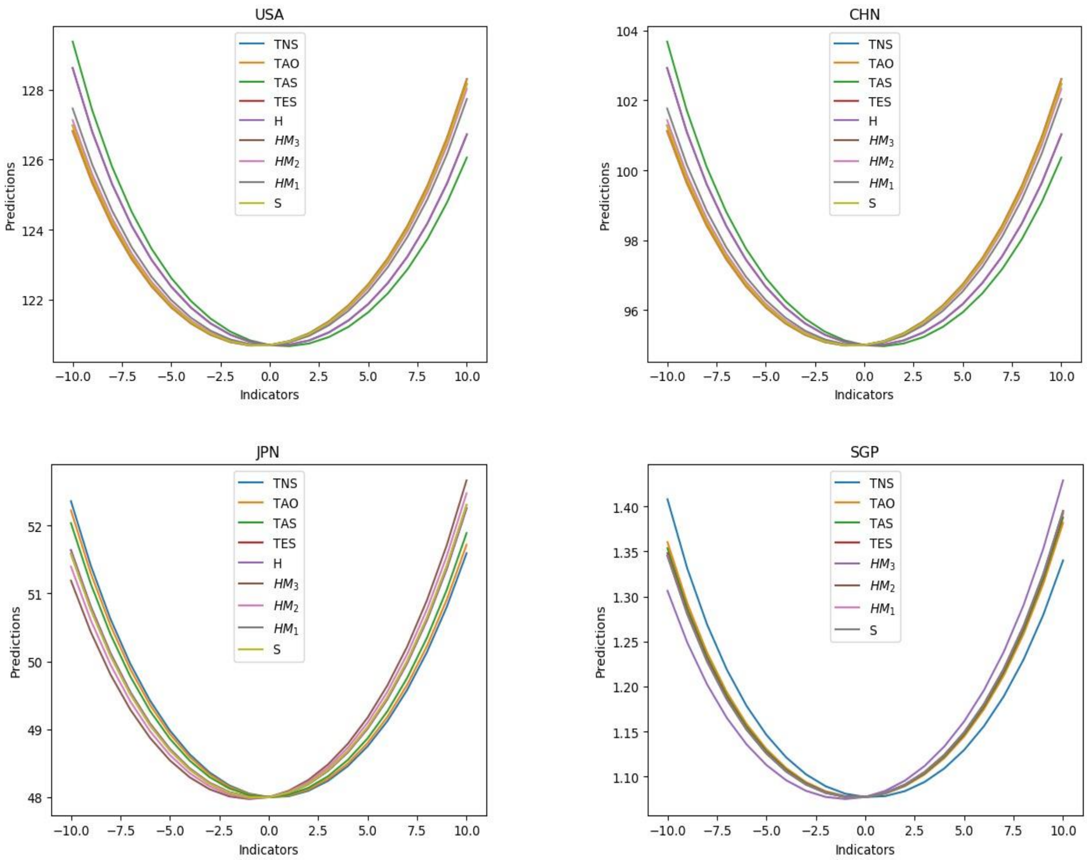

Figure 4: Sensitivity Analysis

3.2. Medal Breakthrough probability

3.2.1 Forecast Results

Binary Logistic Regression is a widely used statistical model for solving binary classification problems[9].It is designed to estimate the probability of an event occurring, where the event has only two possible outcomes. As such, it can be used to predict the probability of a nation making a breakthrough in a particular sport category and winning a medal (with 1 representing winning a medal and 0 representing not winning one).

The goal of logistic regression is to learn a function h(X) that outputs the probability of y being 1.

This function takes the linear combination $z = \omega^T X + b$as input, which is then transformed by the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$, which converts the output of the linear mode into a probability, indicating the likelihood of $y = 1$, i.e.,

$$h(x) = \sigma(\omega^T X + b) = \frac{1}{1 + e^{-(\omega^T X + b)}} \tag{6}$$

Next, the article used log loss to assess the difference between the model’s predicted result and the reality, i.e.,

$$L(\omega, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \log\left(h(X_i)\right) + (1 - y_i)\log\left(1 - h(X_i)\right)\right] \tag{7}$$

where m is the number of groups of data used, i.e., the number of years for which the data were used.

Then, the article employed gradient descent to minimize the loss function, updating w and b, i.e.,

$$\frac{\partial L}{\partial \omega_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h(X_j) - y_i\right) x_{ij} \tag{8}$$

$$\frac{\partial L}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\left(h(X_j) - y_i\right) \tag{9}$$

$$\omega_j \coloneqq \omega_j - \alpha \frac{\partial L}{\partial \omega_j}, \quad b \coloneqq b - \alpha \frac{\partial L}{\partial b} \tag{10}$$

where α is the learning rate, which controls the step size of each update and is selected by cross-validation.

Eventually, after several iterations of training, we obtain the optimal parameters ω and b for new data for prediction. For the feature vector X for the new year of this sport in this country, the probability of winning a medal is:

$$\hat{y} = \sigma(\omega^T X + b) \tag{11}$$

This article selected several countries' sports that had not won any medals in Olympics between 1996 and 2024 and applied the model to make predictions. Based on the predictions of all the countries and sports selected, this article determined that the most likely country and event to experience a medal breakthrough in 2028 is Mali in football, with a forecasted probability of 0.0071. The top four in the probability of a breakthrough are shown in Figure 5.

Figure 5: Top 4 of Predicted Medal Breakthrough Probability

3.2.2 Sensitivity Analysis

To assess the sensitivity of the model, we utilize the ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve)[10]. The procedure is as follows:

We select a threshold T and convert the predicted probability $\hat{P}$ into binary classification results. If $\hat{P} \geq T$, the prediction is positive (1); otherwise, if $\hat{P} < T$, the prediction is negative (0). We then compute the recall rate based on the confusion matrix, using the formula:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Here, TP (True Positives) refers to the number of correctly predicted positive samples, while FN (False Negatives) indicates the number of positive samples incorrectly classified as negative.

Next, we plot the false positive rate (FPR) against the true positive rate (TPR) for various threshold values, resulting in the ROC curve. The formulas for FPR and TPR are as follows:

$$FPR = \frac{FP}{FP + TN}, \qquad TPR = \frac{TP}{TP + FN} \tag{13}$$

This allows us to generate the ROC curve for different thresholds, with the X-axis representing the FPR and the Y-axis representing the TPR. We also compute the AUC (Area Under the Curve) to quantify the model's sensitivity.

The AUC value we obtained is 0.99, which is very close to 1, indicating that the model we trained demonstrates excellent sensitivity.

4. Conclusions

Olympic medal prediction has always been a difficult problem due to its comprehensiveness and specificity. By utilizing a variety of models, including PLSR and logistic regression ,this article not only effectively predicts trends in medal counts but also assesses the potential for emerging countries to win medals. The experimental results show that the models have good predictability and robustness, and has a certain practical application value. Although they also have certain limitations in practical applications, it is highly possible that with future research and improvements, these models can be further refined and applied to a broader range of sports fields.

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] 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] SHI Huimin, ZHANG Dongying, ZHANG Yonghui. Can Olympic medals be predicted? ---Based on the perspective of explainable machine learning[J]. Journal of Shanghai Sport University, 2024, 48(04):26-36.

[4] Luo Yubo, Cheng Yanfang, et al. Prediction of the number of medals and overall strength of China in the Beijing Winter Olympics: Based on the host effect and grey prediction model[J].Contemporary Sports Science and Technology, 2022, 12(21):183-186.

[5] Tian Hui, He Yiman, et al. Medal prediction and participation strategy of Chinese athletes at the 2022 Beijing Winter Olympics: Based on the analysis of the home field advantage effect of the Olympic Games[J]. Sports Science, 2021, 41(02):3-13+22.

[6] HOOG ANTINK C, BRACZYNSKI A K, GANSE B. Learning from machine learning : Prediction of age-related athletic performance decline trajectories [J]. GeroScience, 2021, 43(5): 2547-2559.

[7] Zhang Chao, Zhang Yanhua. Experimental error analysis based on bootstrap method [J]. Quality and Certification, 2025(02):110-112.

[8] Xiao Song. Research on biased estimation of several classes in binary logistic regression models[D]. Guizhou University for Nationalities, 2024.

[9] Wang Menghan. Sensitivity analysis of symmetric eigenvalue problem[D]. Nanjing University of Aeronautics and Astronautics, 2022.

[10] Ren Huan. Application and improvement of partial least squares regression algorithm[D]. Tianjin Polytechnic University, 2019.