Research on Olympic Medals Prediction Model Based on Linear Regression and Logistic Regression

Ziyuan Hao [1, †, *], Weijie Feng [2, †] and Xuezhen Zhou [3, †]

[1] SWUFE-UD Institute of Data Science, Southwestern University of Finance and Economics, Chengdu, China

[2] School of Finance, Southwestern University of Finance and Economics, Chengdu, China

[3] School of Mathematics, Southwestern University of Finance and Economics, Chengdu, China

* Corresponding Author Email: 18769267100@163.com

† These authors also contributed equally to this work

Abstract. This paper proposes a linear regression model, a logistic regression model, and a counting model with fixed and random effects for Olympic medal prediction, focusing on the application of different models in predicting the number of medals of each country in the 2028 Olympic Games, the possibility of winning a medal for a country that has not yet won a medal, and the influence of excellent coaches on the number of medals. First, a linear regression model was used to predict the number of gold, silver and bronze medals of the 2028 Olympic Games for 15 countries by estimating the regression coefficients through the least squares method with the number of historical medals and the number of athletes as the characteristic variables. Secondly, logistic regression model is applied to extract the characteristics such as the number of athletes and the number of events, and a binary classification model is built through maximum likelihood estimation to predict the probability of winning for the countries that have not won any medals, and the 10 countries that are most likely to win gold medals are given. Finally, a counting model containing fixed effects and random effects is constructed, and the parameters are estimated with the help of Bayesian method to quantify the influence of excellent coaches on the number of medals. The model system can accurately predict Olympic medals from different dimensions, and the combination of multiple algorithms effectively improves the accuracy and comprehensiveness of the prediction, providing scientific model support for Olympic medals prediction and related analysis.

Keywords: OLS; linear regression; logistic regression; maximum likelihood estimation; feature variables.

1. Introduction

This paper focuses on the problems related to medal forecasting for the Olympic Games, aiming to deeply analyze the medal acquisition potentials and influencing factors of each country by constructing multiple models [1]. Firstly, considering the linear correlation between the number of medals and multiple characteristics, a linear regression model is adopted, using historical medal data and athlete size as characteristic variables, and optimizing the regression coefficients through the least squares method to achieve the prediction of the number of gold, silver and bronze medals of the target countries [2][3]. Secondly, for the potential assessment of countries that have not won medals, a logistic regression binary classification model is constructed to extract the characteristics such as the number of athletes, the number of competitions, the economic level, etc., and with the help of maximum likelihood estimation method, the probability of each country to win medals is calculated and ranked [4] [5]. Finally, in order to investigate the influence of good coaches on the number of medals, a counting model with fixed and random effects of countries and sports is constructed, and Bayesian method is used to estimate the parameters and quantify the actual effect of "great coaches" on the U.S. volleyball team's medals [6].

The experimental results show that the constructed model system can effectively capture the key factors of medal prediction from different dimensions, which provides a quantitative tool with practical value for analyzing the medal pattern of the Olympic Games.

2. Predictive Model for Number of Medals Based on Linear Regression

2.1. Linear Regression Modeling

To predict the number of gold medals and total medals for each country in the 2028 Olympic Games, a linear regression model is used. The linear regression model assumes that the number of medals is linearly related to a series of features. The regression model is defined as:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \text{ò} \tag{1}$$

Where: Y is the predicted number of gold medals or total medals? $X_1, X_2, \ldots, X_n$ Are the feature variables, such as historical medal counts, number of athletes, number of events, and the host flag? β_0 Is the intercept, which represents the baseline number of medals when all feature variables are zero? $\beta_1, \beta_2, \cdots, \beta_n$ Are the regression coefficients, reflecting the impact of each feature variable on the number of medals. ϵ is the error term, representing the random fluctuations and unexplained parts of the regression model.

The size of the regression coefficients reflects the extent to which each feature influences the number of gold medals or total medals. The goal is to estimate the regression coefficients by training the model on a dataset and minimizing the error between predicted and actual values. In the model, it is assumed that the number of medals (Y) is determined by a linear relationship between various features, such as historical medal counts, the number of athletes, and the number of events. To estimate the regression coefficients, Ordinary Least Squares (OLS) is used.OLS estimates the regression coefficients by minimizing the sum of squared errors between the predicted and actual values. Suppose there are N training samples, and each sample contains n feature variables. The number of medals for each sample is denoted by y_i, with corresponding feature values $x_{1i}, x_{2i}, \ldots, x_{ni}$. The objective is to minimize the following objective function:

$$\min \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} \right) \right)^2 \tag{2}$$

By minimizing the above objective function, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_n$ can be estimated. To solve this optimization problem and find the minimum value of the objective function, gradient descent or the normal equation is typically used.The solution to the normal equation is:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y \tag{3}$$

Where: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)^T$ are the estimated regression coefficients, X is a $N \times (n+1)$ matrix, where each row represents the feature vector of a training sample, with the first column being 1 (corresponding to the intercept), Y is a $N \times 1$ vector containing the actual number of medals for all training samples, X^T is the transpose of matrix X.

Model Evaluation The performance of the regression model is typically evaluated using the following metrics:

R-squared (R^2): Represents the proportion of variability explained by the model. The closer R^2 is to 1, the better the model fits the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2} \tag{4}$$

Where $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the sample.

Mean Squared Error (MSE): Represents the average squared difference between the predicted and actual values. A lower MSE indicates better model performance.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{5}$$

Residual Analysis: Examining the residuals (the differences between actual and predicted values) to see if they follow a normal distribution and if any systematic errors are present.

2.2. Linear Regression Model Solving and Analysis of Results

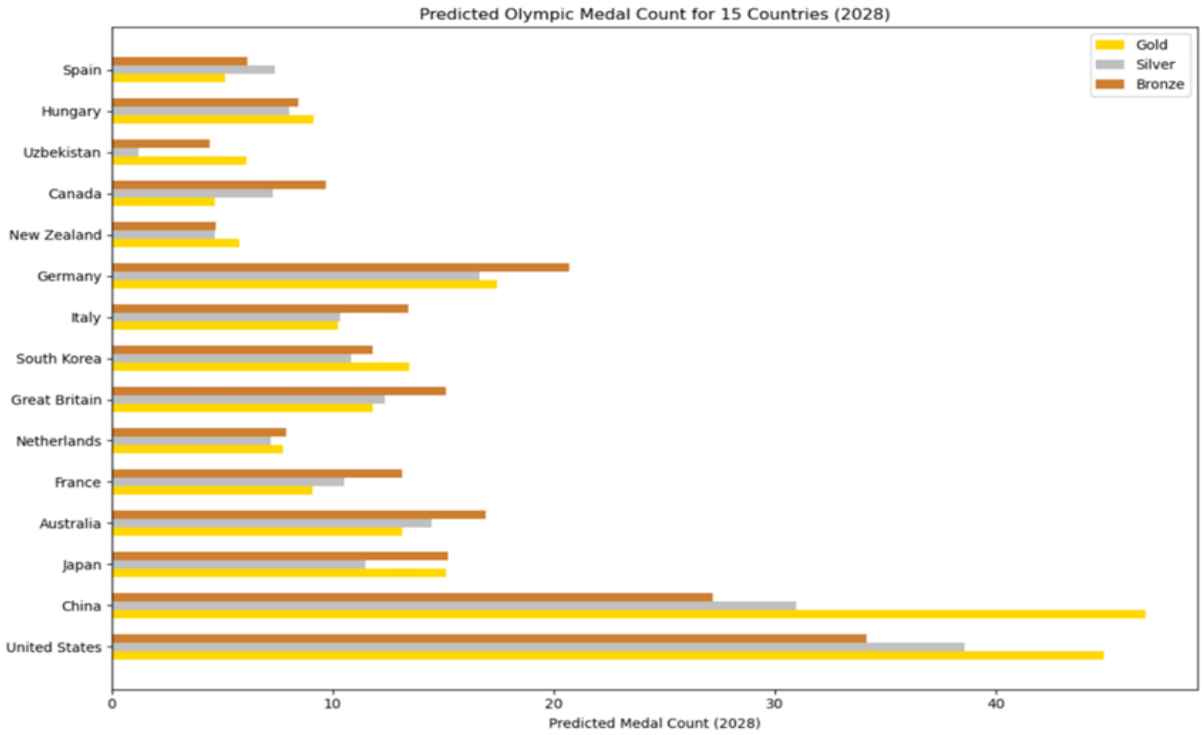

Fig 1. Predicted Olympic medal count for 15 countries in 2028.

As can be seen in Fig. 1, from the projected medal counts for the 15 countries in the 2028 Olympics, the United States and China are projected to be the leaders in the medal standings, with a clear dominance in the number of gold medals in particular. The United States has a far superior total medal count, especially in gold and silver medals. China follows closely behind, also showing a significant advantage in the number of gold and total medals. Other countries such as France, Germany, Japan and Australia remained at the top, albeit in smaller numbers compared to the former two. In addition, countries like Spain, Hungary and Uzbekistan have a more balanced distribution of gold and silver medals, demonstrating their overall competitiveness in a number of sports. The results of the medal distribution shown in the charts reveal the potential performance and competitive advantages of each country in future Olympic Games.

3. Classification Model for Award Probability Based on Logistic Regression

3.1. Feature Extraction

First, the countries that have not yet won a medal need to be identified and selected, and features from the data that may influence their potential to win a medal should be extracted. These features may include: the country's athlete population, the number of participating events, athletes' historical performance, the country's economic level, historical sports infrastructure, and so on. Using logistic regression models, a binary classification model can be established, where the target variable is "whether a medal is won," thereby predicting whether a particular country is likely to win a medal.

3.2. Logistic Regression Model Feature Engineering

$X_1, X_2, \ldots, X_n$ Are the feature variables that influence whether the country can win a medal? These features can include the number of athletes, the number of sports participated in, historical performance, economic level, etc. These features together determine the probability of the country winning a medal;

3.3. Logistic Regression Modeling

In order to perform classification prediction, the logistic regression model was chosen. Logistic regression is a commonly used binary classification model that predicts the probability of an event occurring. The mathematical expression of the logistic regression model is as follows:

$$P(\text{ Medal }) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \tag{6}$$

In this model, $P(Medal)$ represents the probability of a country winning a medal, i.e., the likelihood of the country winning a medal in the 2028 Olympics. Since it is a probability value, $P(Medal)$ will be between 0 and 1;

Maximum Likelihood Estimation:

In logistic regression, the estimation of the regression coefficients is done by maximizing the likelihood function. The likelihood function represents the probability of observing the current data given the feature data. Suppose there are mmm samples, with labels y_i and features X_i for each sample. Then, the likelihood function can be expressed as:

$$L(\beta_0, \beta_1, \cdots, \beta_n) = \prod_{i=1}^{m} P(y_i | X_i) \tag{7}$$

Where $P(y_i \mid X_i)$ represents the probability that the label y_i of sample i is 1. Since this is a binary classification problem, the label y_i can take values 0 or 1, indicating whether the country won a medal. Therefore, $P(y_i \mid X_i)$ can be written as:

$$P(y_i | X_i) = P(\text{ Medal })^{y_i} (1 - P(\text{ Medal }))^{(1-y_i)} \tag{8}$$

If the label $y_i = 1$ (i.e., the country won a medal), the probability of the sample is$P(Medal)$; If $y_i = 0$ (i.e., the country did not win a medal), the probability is $1 - P(Medal)$. To simplify the calculations and improve numerical stability, the logarithm of the likelihood function is typically taken, resulting in the log-likelihood function. The log-likelihood function is expressed as:

$$\ell(\beta_0, \beta_1, \cdots, \beta_n) = \sum_{i=1}^{m} \left[y_i \log(P(\text{ Medal })) + (1 - y_i) \log(1 - P(\text{ Medal })) \right] \tag{9}$$

By maximizing the log-likelihood function, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_n$ can be obtained. The goal of maximizing the log-likelihood function is to make the model's predicted probabilities as consistent as possible with the actual labels. This process typically uses optimization algorithms (such as gradient descent) to find the optimal regression coefficients.

AUC value is commonly used to evaluate the overall performance of classification models, especially in cases of data imbalance, where AUC as a performance evaluation metric has high reliability.

$$\text{AUC} = \sum_{i=1}^{n-1} \frac{1}{2} (\text{FPR}_{i+1} - \text{FPR}_i)(\text{TPR}_{i+1} + \text{TPR}_i) \tag{10}$$

3.4. Logistic Regression Model Solving and Screening of Potential Countries

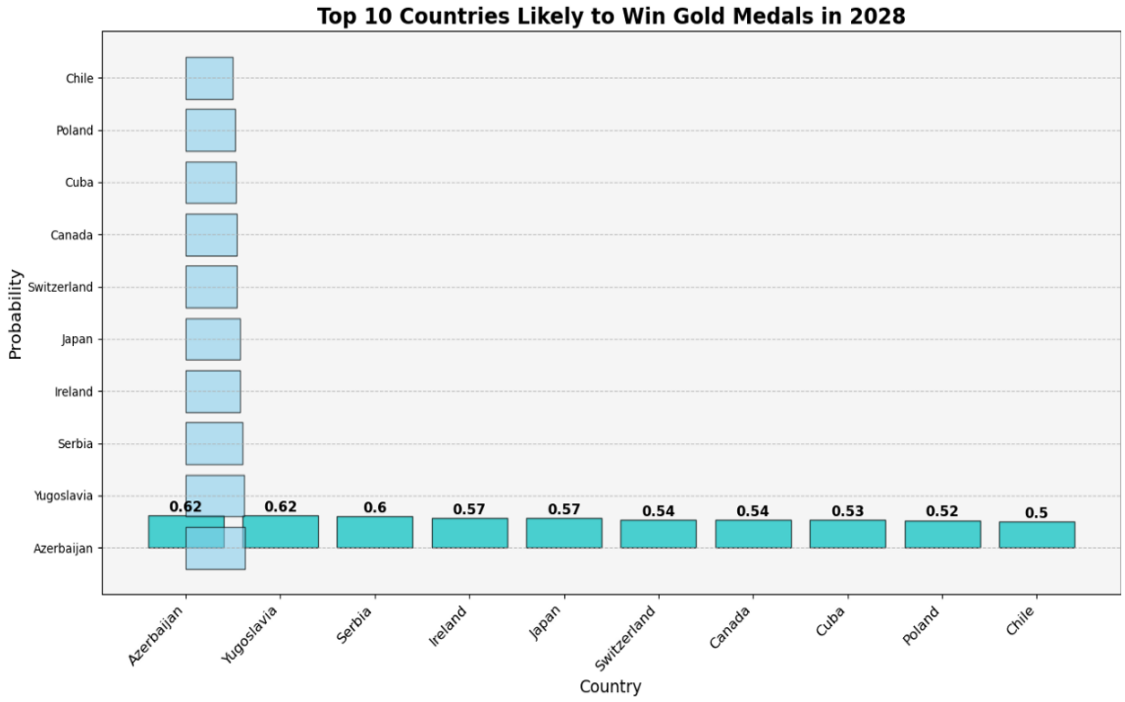

Fig 2. Top 10 countries likely to win gold medals in 2028.

As can be seen in Fig. 2, Chile, Poland and Cuba are the most likely countries to win a gold medal in terms of the 2028 Olympic Games gold medal forecast, with Chile having the highest probability of winning a gold medal at 62%. They are closely followed by Serbia and Ireland, both with a high gold medal probability of 57%.

4. Coaching Impact Analysis Based on Fixed Effects Counting Models

4.1. Coaching Impact Quantitative Problem Analysis

This variable takes the value 1 if country c had a "great coach" for sport s during the t-th Olympic Games, and 0 otherwise.

$$\text{Coach}_{c,s,t} = \begin{cases} 1, & \text{if country } c \text{ had a "great coach" in sport } s \text{ in the } t \text{ th Olympic Games;} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

4.2. Fixed Effects Counting Model Construction Based on Bayesian Estimation

First, let $M_{c,s,t}$ denote the number of medals won by country c in sport s during the t-th Summer Olympics. Considering that the number of medals is a non-negative count variable, it is assumed that:

$$M_{c,s,t} \sim \text{Poisson}\left(\lambda_{c,s,t}\right) \tag{12}$$

Where λ_c,s,t represents the expected number of medals for country c in sport s during the t-th Olympic Games.

To relate the logarithm of the expected number of medals to the explanatory variables, a log-link function is used, modeling it as a linear combination of the predictors:

$$\log\left(\lambda_{c,s,t}\right) = \alpha + \beta \cdot \text{Coach}_{c,s,t} + \gamma_c + \delta_s + \text{ò}_t + \eta_{c,s} \tag{13}$$

In the above expression:

α represents the global intercept, indicating the baseline number of medals in the absence of any influencing factors; β represents the regression coefficient of the "great coach" effect, measuring the impact of having a "great coach" on the number of medals; γ_c represents the fixed effect of country c, controlling for characteristics of different countries that do not vary over time; δ_s represents the fixed effect of sport s, controlling for characteristics that are constant across different sports; $?_t$ represents the fixed effect of the tth Olympic edition, capturing systematic influences specific to that edition; $\eta_{c,s}$ represents the random effect of country c in sport s, reflecting unobservable characteristics specific to the combination of country and sport.

The model parameters α, β, γ_c, δ_s, and $\eta_{c,s}$ need to be estimated using Maximum Likelihood Estimation (MLE) or Bayesian methods. Given that the model includes numerous fixed and random effects, Bayesian methods leveraging Markov Chain Monte Carlo (MCMC) sampling can more effectively estimate parameters and their uncertainties. The estimated β coefficient quantifies the specific contribution of a "great coach" to the medal count. Specifically, the exponentiated value of β, e^{β} represents the multiplicative effect on the medal count when a "great coach" is present compared to when one is not. For example, if $\beta = 0.5$, then $e^{0.5} \approx 1.6487$, indicating that having a "great coach" increases the medal count by approximately 64.87%.

$$e^{\beta} = \begin{cases} > 1, & \text{indicates an increase in medal count} \\ = 1, & \text{indicates no effect;} \\ < 1, & \text{indicates a decrease in medal count} \end{cases} \tag{14}$$

Select United States and their volleyball. For the country-sport combination, identify the editions of the Olympics where "great coaches" were present and estimate their impact on medal counts using the model.

Let β represent the regression coefficient for the "great coach" effect. If, in a given Olympic edition, country c in sport s has a "great coach," the expected medal count is:

$$\log\left(\lambda_{c,s,t}\right) = \alpha + \beta + \gamma_c + \delta_s + \grave{o}_t + \eta_{c,s} \tag{15}$$

By exponentiating this expression, the expected medal count becomes:

$$\lambda_{c,s,t} = \exp\left(\alpha + \beta + \gamma_c + \delta_s + \grave{o}_t + \eta_{c,s}\right) \tag{16}$$

Compared to the expected medal count without a "great coach," the multiplicative effect of having a "great coach" is:

$$\frac{\lambda_{c,s,t}(\text{ Coach } = 1)}{\lambda_{c,s,t}(\text{ Coach } = 0)} = \exp(\beta) \tag{17}$$

This ratio directly reflects the impact of a "great coach" on the medal count. Through parameter estimation of the model, the impact of a "great coach" on medal counts for specific countries and sports can be quantified. As the Fig. 3 shows the change in medals since the great coaches joined USA Volleyball.

4.3. Model Solution

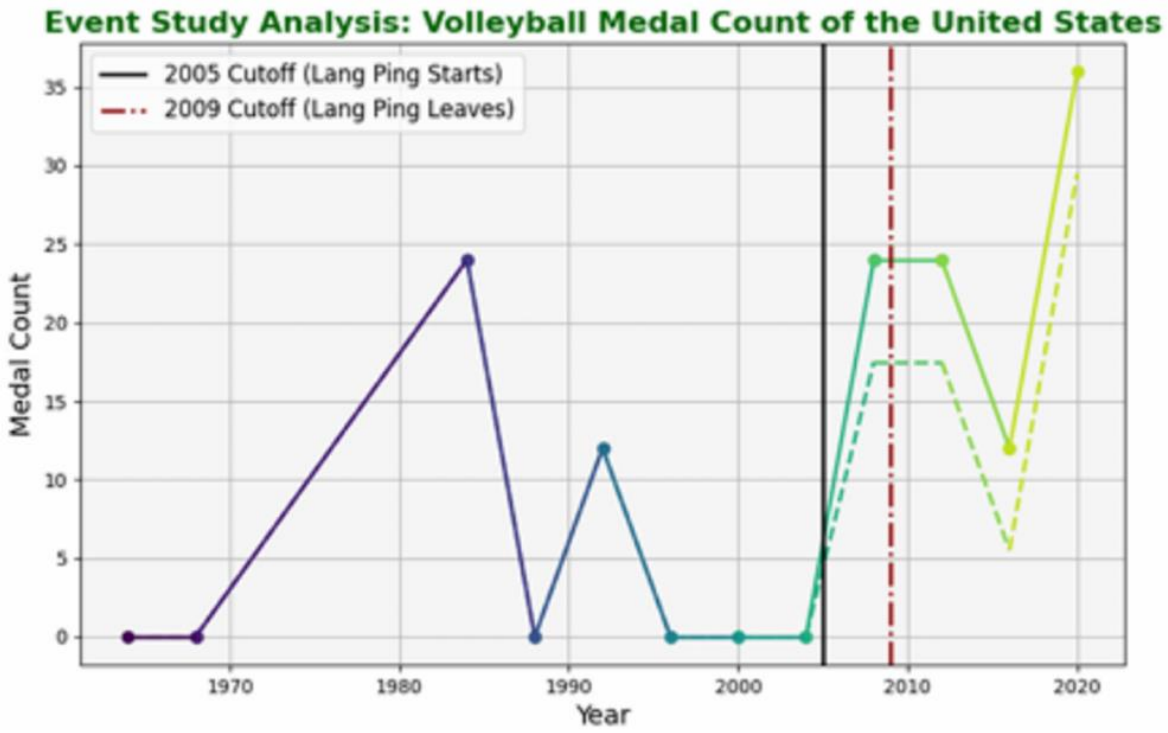

Fig 3. Volleyball medal count of the United States.

As shown in Fig 3, the trend in the number of U.S. volleyball medals shows that the addition of Coach Lang Ping in 2005 significantly boosted the number of medals, especially between 2005 and 2009, when the number of medals rose rapidly. However, after the departure of Lang Ping in 2009, there was a significant drop in the number of medals. The chart demonstrates the impact of coaching changes on team performance, showing that Lamping's leadership was critical to improving the overall performance of the U.S. volleyball team.

5. Conclusion

In this paper, linear regression, logistic regression and counting models with fixed and random effects are proposed, which have the ability of integrating multi-dimensional features to accurately capture the complex associations in medal prediction, and effectively improve the prediction accuracy and depth of analysis through the complementary advantages of different algorithms. First, the linear regression model takes the historical medal count and the number of athletes as features, and optimizes the parameters through the least squares method to realize the quantitative prediction of the medal count of the 2028 Olympic Games for 15 countries, which has the advantage of visually showing the linear influence of features and medal count. Secondly, the logistic regression model extracts multi-dimensional features such as economic level and number of participating events, and constructs a classification model with the help of maximum likelihood estimation, which provides a probabilistic analytical framework for assessing the potentials of countries that have not won any medals, and can clarify the competitive advantages of high-potential countries. Then, the coach impact counting model, by including fixed and random effects such as countries and programs, and estimating parameters with Bayesian method, successfully quantifies the actual contribution of "great coaches" to the number of medals of the U.S. volleyball team, which reflects the ability of in-depth excavation of specific influencing factors. Finally, the three models form a systematic analysis framework, covering the needs of medal number prediction, potential assessment and influencing factors analysis, which provides a scientific methodological support for Olympic medal research. Future research can further expand the dimensions of features, such as incorporating dynamic

variables such as athletes' age structure and training technology input, or exploring the correlation model of medal competition across sports, in order to enhance the foresight and universality of prediction.

References

[1] Yang Jinghan. Design of Olympic medal ranking prediction based on univariate and multiple regression models [J]. Information System Engineering, 2018, (02):149-152.

[2] Zhang Yuhua. Prediction of China's 31st Olympic Games medal count based on linear regression dynamic 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.

[3] Yang Qingbao. Linear regression optimized least squares prediction model for tactical fit success [J]. Science and Technology Bulletin, 2017, 33(04):88-91.DOI:10.13774/j.cnki.kjtb.2017.04.020.

[4] Zhou Wenqing, Zhou Da, Kang Jianjun. Research on nuclide identification algorithm based on logistic regression binary classification [J]. Nuclear Electronics and Detection Technology, 2023, 43(01):12-17.

[5] Zhao Lijiao. Research on image matching method based on maximum likelihood estimation [D]. Dalian University of Technology, 2014.

[6] Liu Ting, Chen Qingrong, Yang Hongfeng, et al. A multi-label number estimation method based on Bayesian inference [J]. Communication Technology, 2021, 54(05):1179-1183.