World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3066

PREDICTION OF OLYMPIC MEDAL DISTRIBUTION BASED ON LOGISTIC REGRESSION MODELS

RunMo Liu[1*], YiWen Gu[2], Jing Zhang[2]
[1]*College of Resources Environment and Tourism, Capital Normal University, Beijing 100048, China.*
[2]*School of Mathematical Sciences, Capital Normal University, Beijing 100048, China.*
[*]*Corresponding Author: RunMo Liu*

Abstract: The Olympic Games serve as a global platform to showcase athletic performance and national competitiveness. Accurate forecasting of medal outcomes not only provides scientific support for sports policy and resource allocation but also contributes to understanding the dynamics of international competition. This study employs logistic regression to address two key research problems: (1) predicting the distribution of total and gold medals across countries in the 2028 Los Angeles Olympic Games, and (2) estimating the probability of countries historically without medals achieving their first Olympic success. The models integrate variables such as historical performance, number of participants, number of events, and geographical proximity to the host nation. Results indicate that the proposed framework achieves high predictive accuracy, with strong model fit and low error values, while also identifying emerging countries with significant potential for breakthroughs. The findings not only enhance medal prediction methodology but also provide broader insights into the evolving landscape of global sports competitiveness.

Keywords: Olympic games; Logistic regression; Medal prediction; Sports analytics; Forecasting models; International competitiveness

1 INTRODUCTION

The Olympic Games, as the most influential international sporting event, not only embody the pursuit of athletic excellence but also reflect the economic, cultural, and political power of participating nations. With the growing scale of participation and the diversification of competition events, the prediction of Olympic medal counts has become an important interdisciplinary research topic, involving statistics, machine learning, and data-driven decision-making. Accurate medal forecasting can provide insights for sports policy makers, training institutions, and scholars in sports economics [1-2].

In recent years, scholars have applied traditional econometric approaches to forecast Olympic medal counts. Bernard and Busse employed a log-linear regression relating medal totals to GDP per capita, population, host advantage, and political system, and even produced out-of-sample predictions for Sydney 2000 [3]. Johnson and Ali likewise used regression across post-war Summer and Winter Games to quantify how income, population, and political factors drive participation and medals, including estimating the GDP “cost” of an extra medal. However, these models are limited in handling complex nonlinear relationships and high-dimensional data. With the development of artificial intelligence, machine learning models such as random forests, support vector machines, and deep learning networks have been applied to sports analytics [4-5], achieving more robust predictive performance. Logistic regression, in particular, has been widely used in classification and probability estimation tasks, making it suitable for medal prediction and the assessment of a nation’s likelihood of achieving specific competitive milestones.

Despite the progress, existing studies still face several challenges. First, most research has emphasized aggregate medal predictions [6], while less attention has been paid to identifying emerging countries likely to obtain their first Olympic medal. Second, the uncertainty of prediction results has not been sufficiently quantified, leading to difficulties in policy applications. To address these issues, this paper applies a logistic regression framework to two problems: (1) predicting the distribution of total and gold medals among countries in the 2028 Los Angeles Olympic Games, and (2) estimating the probability of countries without historical medals winning for the first time. By integrating historical performance, participation data, and geographical factors, our work provides not only accurate forecasts but also insights into the developmental trends of global sports competitiveness.

2 LOGISTIC REGRESSION MODEL

Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model that belongs to supervised learning in machine learning. Its derivation process and computation are similar to the process of regression, but in fact it is mainly used to solve binary classification problems (can also solve multi-categorization problems) [7]. The model is trained with a given n sets of data (training set) and at the end of the training the given set or sets of data (test set) are classified. Each of these data sets is composed of several metrics [8].

2.1 Model Prediction for the Number of Medals

Step 1:For data processed by logistic regression

For each country, the paper have given the following indicators: historical performance, number of participants, number of events, total number of events, and neighboring countries, from which these characteristics are used to predict the countries' gold and total medals won at the 2028 (i.e., 34th) Olympic Games in Los Angeles. The paper refers to and summarize the data related to historical performance, number of participants, number of events, total number of events, etc. of each country in the 2016 2024 (i.e., 31st-33rd) Olympic Games, in which the item of historical performance is obtained by summing up the number of gold, silver, and bronze medals of the session, and the item of leader country assigns a value of 2 to the host country, a value of 1 to the neighboring countries of the host country, and a value of 0 to the other countries.

Step 2: Reading data and extracting features

From the processed data, all data columns except the total number of medals and the number of gold medals of each country were extracted as feature variables X_i, and the total number of medals and the number of gold medals of each country were taken as target variables Y_1, Y_2. Then the mean and variance of each data were calculated, and the data of the feature values were normalized.

Step 3:Model training and evaluation

The collected dataset is first divided by dividing the data in the ratio of 80% as training set and 20% as test set. A logistic regression model was trained using the training set data. Next, it was assumed that the total number of medals and the number of gold medals of each country obeyed a normal distribution. The parameters in the logistic regression model are estimated by maximum likelihood estimation. The goal of maximum likelihood estimation is to find a set of parameters $(\beta_0,\beta_1,\cdots,\beta_n)$, such that the probability that the model produces them is maximized given the data. Specifically, for each sample in the training set(x_i,y_i), the likelihood function is as follows:

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

where $\beta_1,\cdots,\beta_n$ is coefficients of each characteristic variable, β_0 is intercept, x_i is the historical results and number of participants of the ith sample, and y_i is the actual number of gold medals or total medals won for the i-th sample.

Step 4:Evaluation of the model

The trained model is used to predict the training and test sets and the coefficient of determination R^2 is calculated to evaluate the model fit. The coefficient of determination is calculated as follows:

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

where R^2 means coefficient of determination, $\hat{y}_i$ is the predicted value of the target variable for the ith sample, and $\bar{y}$ represents the average of the actual values of the target variables.

The above equation gives the training set R^2 for the total number of medals is 0.9949, the test set R^2 is 0.9935; the training set R^2 for the number of gold medals is 0.9688 and the test set R^2 is 0.9796, then it shows that the model has a good fit to the data. The paper also calculates the value of MSE by using the following formula to detect the magnitude of error between the predicted and actual values in our model:

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

where *MSE* is Mean Squared Error, and N means total sample volume.

The paper ends up with a value of 0.5553. This means that the model is more accurate in predicting medal counts.

Step 5: Forecasting and obtaining outcome data

The total number of medals and the number of gold medals of each country in the 34th Olympic Games in Los Angeles are predicted using the trained model, and the predicted results are rounded to the nearest whole number as the predicted outcomes z,y. Meanwhile, prediction intervals are set in order to assess the uncertainty of the predictions:

$$[z-6,z+6] \tag{4}$$

where z is the total number of medals predicted by the model.

$$[y-5,y+5] \tag{5}$$

where y is the number of gold medals predicted by the model.

Thus, it is possible to obtain the total number of gold medals and total number of medals that will eventually be won by each country at the 2028 Olympic Games in Los Angeles.

2.2 Model prediction for How Many Countries will Earn Their First Medal

Step 1: Data cleaning

The paper screened the list of 206 countries that have not won any awards up to 2024, totaling 64 countries that are still participating in the Olympic Games continuously. The number of athletes from these 64 countries who have participated in the Olympics more than two times was filtered.

In terms of model selection, since the final output is the probability of winning, plus it is a multivariate problem, we continue to choose the logistic regression model to do the calculation.

Step 2: Feature normalization

In order to avoid the influence of different feature scales on the model, the paper standardizes the training data and prediction data in the first step. The standardization method the paper chooses to use Z-score standardization, the standardized data has a mean of 0 and a standard deviation of 1.Equation below is the formula for Z-score:

$$z=\frac{x-\mu}{\sigma} \tag{6}$$

where z is the Z-score value, which represents the distance of a particular data point X from the data set mean μ, measured in units of standard deviation σ, x is the value of the data point to be normalized, i.e., a specific observation in the original data set, and μ is the mean of the data set, i.e.
The average of all data points, calculated as:

$$\mu=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{7}$$

where n is the number of data points, x_i is the value of the ith data point, and σ is the standard deviation of the data set, which is used as a measure of the degree of dispersion of the data, and is calculated as:

$$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{8}$$

Step 3: Partition the training set and test set
The paper divides the training data into 80% as the training set and 20% as the test set by the function for c-v partition. The model will be evaluated on the test set after training.
Step 4: Train the logistic regression model
The function for fitglm is used to train the logistic regression model. The paper specifies the binomial distribution and logit link function because this is a binary classification problem and the target variable Awarded has a value of 0 or 1. Equation below is the formula for logistic regression:

$$p(\text{medal})=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}} \tag{9}$$

where P is the probability of gaining the first medal, β represents the regression coefficient, x represents the feature independent variable, and e is the natural logarithmic base which is used to ensure that the output result is between 0 and 1, in accordance with the definition of probability.

3 RESULTS AND ANALYSIS

Data source statement: The data in this article is sourced from:
https://www.zyzw.com/yday/yday012.htm
https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/
https://olympics.com/en/paris-2024/medals

3.1 Logistic Regression Model for Predicting the Number of Medal in 2028

Due to the large number of countries participating in the Olympics and the possibility that some countries will have no medals, it is not possible to show all of them here, but the paper has based on the countries participating in the 2024 Paris Olympics. The result 1 is shown in Figure 1 and Figure 2:

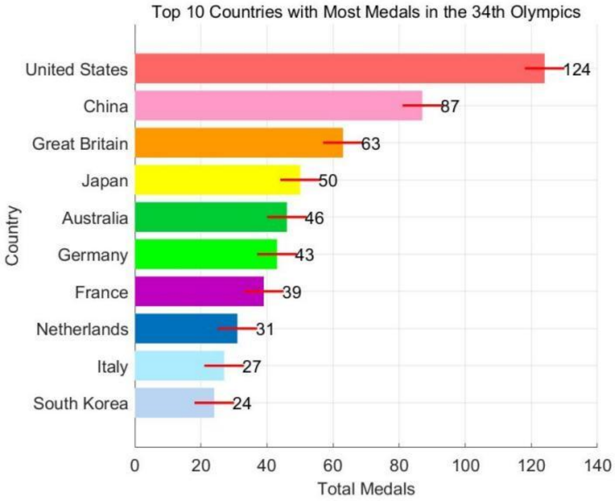

Figure 1 Top 10 Countries with Most Total Medals in the 34th Olympics

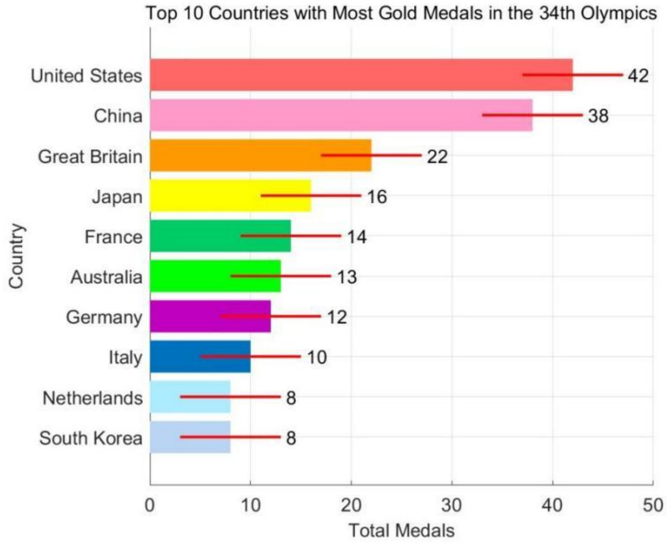

Figure 2 Top 10 Countries with Most Gold Total Medals in the 34th Olympics

The paper has shown the countries that are predicted to potentially win the top ten in the overall medal standings as well as the gold medal standings in the 2028 Los Angeles Olympics. From this the paper can get the United States is expected to win the most medals and gold medals in the 2028 Olympics, the total number of medals is about [118,130], the number of gold medals is about [37,47], China is expected to get the second place in the medal table, the current prediction is that the total number of medals that can be won [81,93], the number of gold medals in the [33,43], the medal table top ten is expected to be the United States, China, Britain, Japan, France, Australia, Germany, Italy, the Netherlands, South Korea contracted. Japan, France, Australia, Germany, Italy, Netherlands, and South Korea contracted. However, since Russian and Belarusian athletes are banned from international competitions in almost all Olympic sports, they cannot be accurately assessed.
The result 2 is shown in Figure 3 and Figure 4:

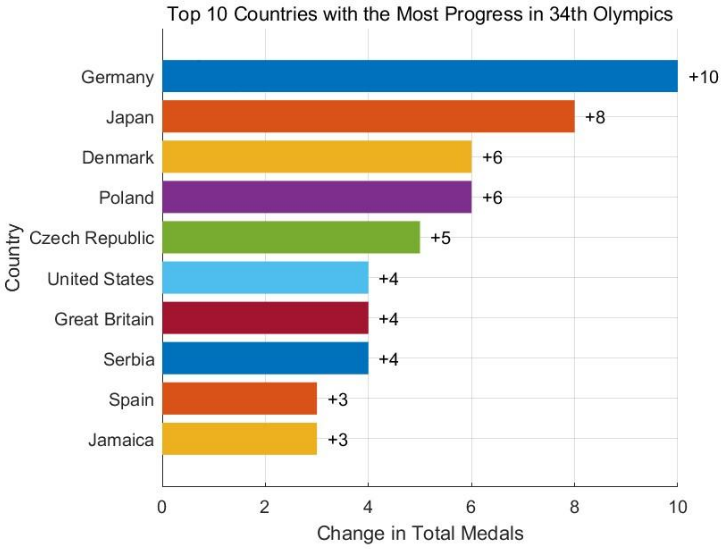

Figure 3 Top 10 Countries with the Most Progress in 34th Olympics

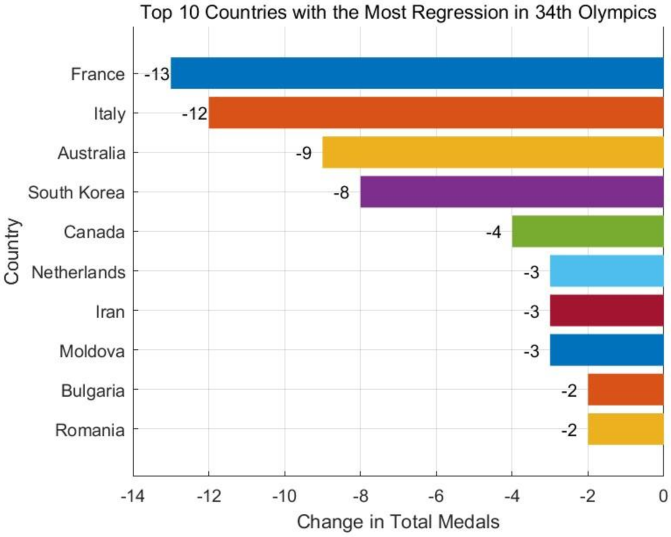

Figure 4 Top 10 Countries with the Most Regression in 34th Olympics

By looking at the predicted total number of medals won and comparing it to historical performance, the paper have obtained the difference in change between the two for each country and ranked them in order, with the countries most likely to improve being Germany, Japan, Denmark, Poland, Czech Republic, United States, Great Britain, Serbia, Spain, Jamaica. States, Great Britain, Serbia, Spain, Jamaica. The countries most likely to regress are France, Italy, Australia, South Korea, Netherlands, Canada, Iran, Moldova, Bulgaria, Romania. Romania.

The result 3 is shown in Figure 5 and Figure 6:

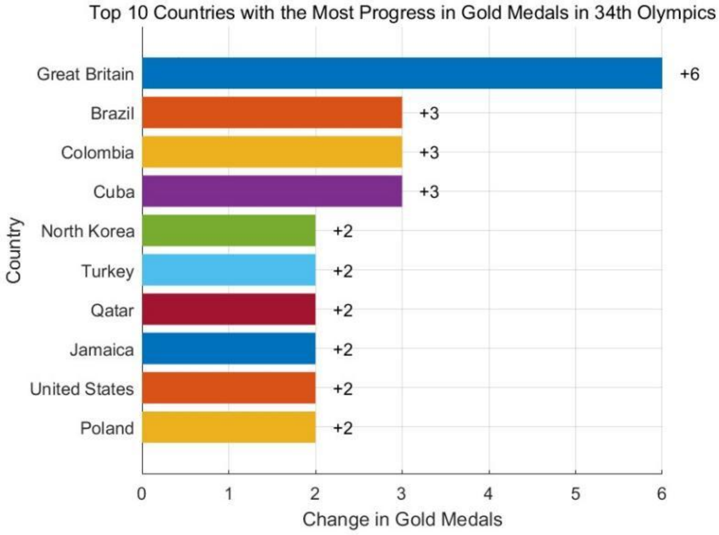

Figure 5 Top 10 Countries with the Most Progress in Gold Medals in 34th Olympics

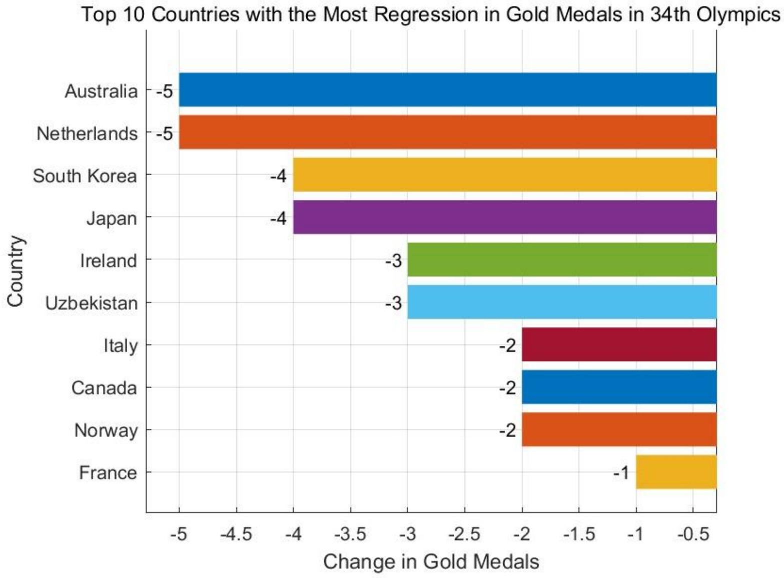

Figure 6 Top 10 Countries with the Most Regression in Gold Medals in 34th Olympics

By looking at the predicted total number of gold medals won and comparing it to historical performance, the paper has obtained the difference in change between the two for each country and ranked them in order, with the countries most likely to improve being Great Britain, Brazil, Colombia, Cuba, North Korea, Turkey, Qatar, Jamaica, United States, Poland. Qatar, Jamaica, United States, Poland. The countries most likely to regress are Australia, Netherlands, South Korea, Japan, Ireland, Uzbekistan, Italy, Canada, Norway, France.

3.2 Logistic Regression Model Prediction for How Many Countries will Earn Their First Medal

The results of the logistic regression modeling process are shown below:

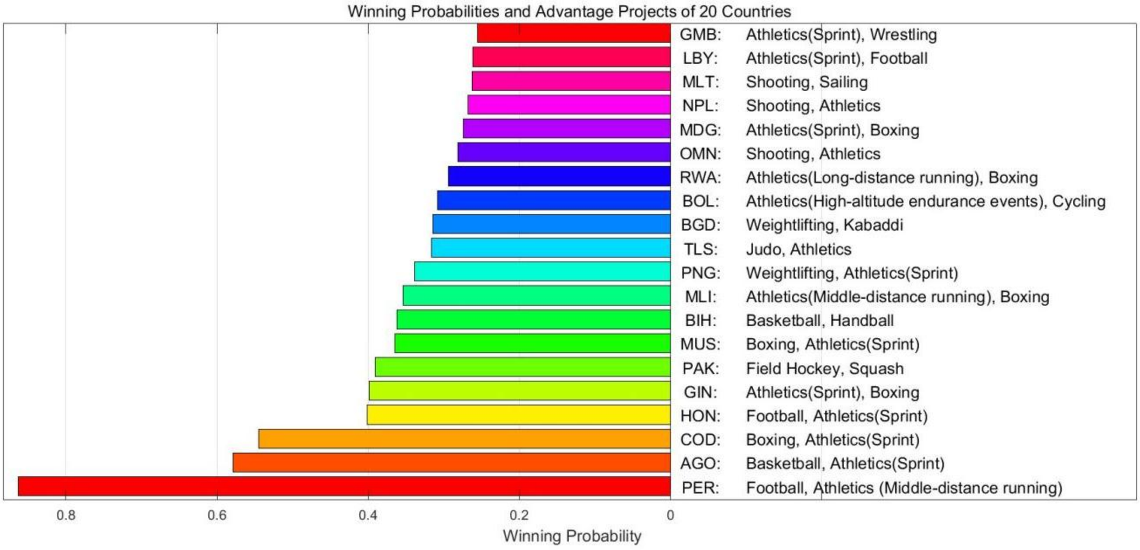

Figure 7 Winning Probabilities and Advantage Projects of 20 Countries

From Figure 7 the result is that Peru has the highest probability of winning the first medal, 86.33%, where the most likely events to win are presumed to be soccer and middle-distance running. Soccer in Peru has deep roots, the domestic league is more mature, has appeared in the World Cup and other tournaments, the youth team also has a certain degree of strength, the Olympic Games soccer competition is expected to achieve good results. Peru in track and field (middle and long-distance running, etc.) occupies the geographical advantage of being located in the plateau, which is conducive to the athletes to improve cardiorespiratory fitness and endurance. The figure also shows that Angola and Congo have a higher probability of winning, with 57.86% and 54.44%, respectively. Angola's men's basketball team is a strong team in Africa, with many good results in the African Championships, and has strong strength and experience in the game. That's why it has a chance to win in the Olympic basketball competition. Congo has a certain boxing atmosphere and athletes with good physical fitness, and if they receive better training and resource support, they have a chance to win in the Olympic boxing program. Below 40 per cent and above 20 per cent, a total of 37 countries have an award probability of 30 per cent, with the majority of countries concentrating on 30 per cent. There are 19 countries with a probability of winning less than 20%. Due to the large amount of data, the paper will only show the top 20 countries in terms of winning percentage, and the projects that are most likely to win according to common sense and the analysis of the competition programs in these 20 countries. For the remainder, see appendix.

4 CONCLUSIONS AND OUTLOOKS

This study developed and applied logistic regression models to forecast medal outcomes for the 2028 Los Angeles Olympic Games, addressing two key research problems: the distribution of total and gold medals across nations, and the probability of countries winning their first Olympic medal. The findings demonstrate that our proposed method delivers high predictive accuracy, supported by strong model fit and low mean squared error. Moreover, the analysis reveals not only the nations likely to dominate upcoming Games but also emerging countries with substantial breakthrough potential.

Building on these results, the framework proposed here can be readily adapted to other arenas of sports analytics—such as World Championships or continental competitions. Enhancing the model with additional data dimensions—including athlete-level performance metrics, training environment variables, and macroeconomic indicators—may further boost its predictive power.

Despite its strengths in interpretability and computational efficiency, logistic regression is limited in capturing complex, nonlinear relationships among predictors. Future research could overcome this by exploring ensemble learning methods or deep neural networks, thereby improving the model's capacity and robustness.

Overall, this research contributes a transparent, data-driven approach to sports performance forecasting, laying the groundwork for deeper insights into the evolving competitive dynamics of international sports.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhao S, Cao J, Lu K, et al. Research on Olympic medal prediction based on GA-BP and logistic regression model. F1000Research, 2025, 14: 245.

[2] Zhang Z, Ma T, Yao Y, et al. Predicting Olympic Medal Performance for 2028: Machine Learning Models and the Impact of Host and Coaching Effects. Applied Sciences, 2025, 15(14): 7793.

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

[4] Vayadande K, Kalshetti A, Kelzarkar T, et al. Olympic Medal Prediction Using Linear Regression and Data Analytics. 2025.

[5] Song X, Liu X, Liu F, et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. International journal of medical informatics, 2021, 151: 104484.

[6] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction//2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025: 822-827.

[7] Schober P, Vetter T R. Logistic regression in medical research. Anesthesia & Analgesia, 2021, 132(2): 365-366.

[8] Zhou Y, Song L, Liu Y, et al. A privacy-preserving logistic regression-based diagnosis scheme for digital healthcare. Future Generation Computer Systems, 2023, 144: 63-73.