Journal of Social Systems and Policy Analysis
http://dx.doi.org/10.62762/JSSPA.2025.975234

RESEARCH ARTICLE

Medal Prediction LASSO-Logistic and Bayesian Probit Models

Yuxiao Zhu[1,*], **Chenhao Wang**[2] **and Lemei Qin**[2]

[1] School of Economics and Management, Northeast Normal University, Changchun 130024, China
[2] School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China

Abstract

In Olympic competitions, the dynamic changes in the medal standings are influenced by a variety of factors. We predict the medal distribution for the 2028 Los Angeles Summer Olympics by analyzing historical data. In Task 1.1, we formulated predictive models for the overall medal count, as well as the individual counts for gold, silver, and bronze medals. These models integrated lagged variables and employed multivariate linear regression to ascertain the weightings of various indicators, thereby enabling the projection of the 2028 Olympic Games' medal table. The United States was projected to secure the top position with a total of 145 medals, including 50 gold medals. In Task 1.2, we deployed an integrated model that combined the Lasso regression and logistic regression techniques to calculate the indicator weightings, which were then used to predict the likelihood of certain nations achieving their inaugural gold medal, such as Montenegro in the water polo event. In Task 1.3, we conducted an in-depth analysis of the impact of dominant sports and the host country effect on the medal standings, quantifying their respective influences. The final aspect of our methodology involved visualizing the parameters derived from both the training and validation phases of our models. In Task 2, our primary investigation was to determine whether the presence of a Great Coach influences the likelihood of winning awards. We innovatively incorporated priors into the ordered Probit model, updating the posteriors in real-time and employing the Gibbs sampling method to segment threshold values and examine the nonlinear relationships of the latent variables. Through the analysis of two classic case studies, we concluded that at a 95% confidence level (with significance $p < 0.1$), the Great Coach index exceeds 1 for all cases, indicating a significant effect. Additionally, we visualized the weight of each parameter. Finally, a one-page memo with suggestions for country Olympic committees is also produced to help them formulate guidelines aimed at advancing the development of the Olympic Games for 'Faster, Higher, Stronger-Together'.

Keywords: multiple linear regression, bayesian ordered probit model, LASSO-Logistic regression, host country effect, gibbs sampling.

1 Introduction

1.1 Problem Background

The Olympic Games, as the world's most influential and authoritative comprehensive sports event, serves not only as a platform for athletes to showcase

Submitted: 28 May 2025
Accepted: 29 June 2025
Published: 16 August 2025

Vol. 2, **No.** 3, 2025.
10.62762/JSSPA.2025.975234

***Corresponding author:**
✉ Yuxiao Zhu
zhuyuxiao@nenu.edu.cn

Citation
Zhu, Y., Wang, C., & Qin, L. (2025). Medal Prediction LASSO-Logistic and Bayesian Probit Models. *Journal of Social Systems and Policy Analysis*, 2(3), 133–148.

themselves and pursue excellence but also as a concentrated manifestation of a nation's sports capabilities and overall national strength. Since the revival of the modern Olympic Games in 1896, the Olympic medal table has consistently been one of the focal points of global attention. The number of medals reflects a nation's achievements in sports and symbolizes national honor and ethnic pride. For example, at the 2024 Paris Olympics, the United States topped the medal table with 126 medals, while China tied with the US for the most gold medals (40). The host country, France, also achieved remarkable results with 16 gold medals and a total of 64 medals.

Studies have shown that economic resources strongly predict Olympic medal totals [1]. However, the composition of the Olympic medal table is not static. In addition to the long-standing rivalry between traditional sports powers, some countries and regions are breaking through and achieving major breakthroughs from "zero to one". For example, in Paris 2024, Albania, Cape Verde, Dominica and St. Lucia are among the countries and regions winning Olympic medals for the first time, reflecting the dynamic distribution patterns captured in Figure 1.

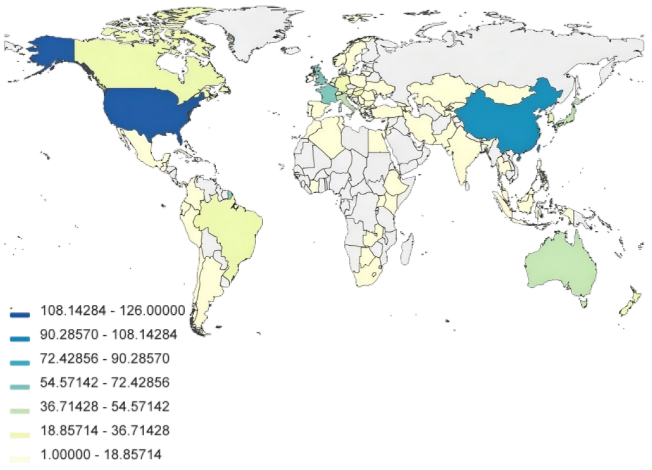


Figure 1. Map of medal distribution for countries at the 2024 Paris Olympics.

Given the dynamic and complex nature of the Olympic medal table, predicting the medal standings for the upcoming 2028 Summer Olympics in Los Angeles requires consideration of multiple factors beyond a nation's historical performance. Advanced modeling techniques such as two-stage sparse logistic regression [7] and robust adaptive LASSO [8] provide powerful frameworks for handling the high-dimensional, correlated predictors inherent in sports performance data. These factors include the characteristics of athletes, the host country effect, and the role of coaching teams, all of which can be effectively captured through modern regularization methods.

1.2 Restatement of the Problem

Given the background information and constraints outlined in the problem statement and based on the provided data, we need to address the following issues:

- **Development of Medal Count Prediction Models:** Develop a predictive model for estimating the medal counts of each country and provide metrics to assess the performance of the model. Utilize this model to forecast the medal standings for the 2028 Summer Olympics in Los Angeles, USA.
- **Model Analysis:** Based on the model, analyze and predict the performance of different countries in the 2028 Olympics. For countries that have never won a medal, predict which ones might secure their first medal in 2028 and estimate the associated probabilities.
- **Analysis of the "Host Country Effect":** Investigate the relationship between the number and types of Olympic events and the medal counts of participating nations. Analyze how the selection of events by the host country influences the distribution of medals.
- **Analysis of the "Great Coach Effect":** Examine the potential impact of highly effective coaches on medal counts when they coach in different countries. Estimate the contribution of the 'Great Coach Effect' to the medal results.

1.3 Our work

Building upon the aforementioned analysis, we have conducted our work. The framework of our work is illustrated in Figure 2. Our modeling approach builds upon the sparse statistical learning framework [5], particularly leveraging the LASSO's properties for high-dimensional prediction problems in sports analytics.

2 Assumptions and Justification

Some fundamental assumptions are listed below.

- **Assumption 1:** Assuming that the future performance of countries and athletes can be predicted based on historical data, and medal data

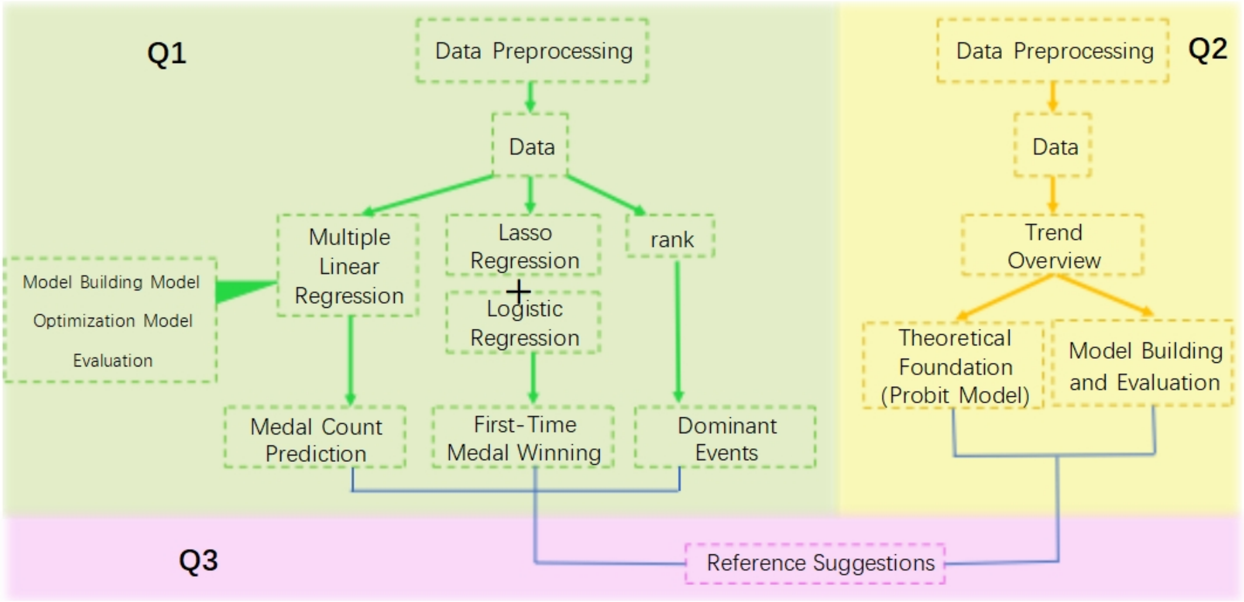


Figure 2. Our work.

from 2000 onwards is used. Justification:Historical medal data can reflect long-term performance trends and competitive levels of countries and athletes at the Olympic Games. Earlier outdated athletic programs are less relevant to the current Olympics. The use of data from 2000 onwards reduces the interference of outdated information and thus improves the accuracy and interpretability of forecasts.

- **Assumption 2:** Ignoring the complexities of GDP, policy and rare programs, etc.
Justification:Introducing complex macro variables can be highly correlated with other variables leading to multicollinearity problems, and to simplify the model and eliminate the overall impact of rare events in a given year, our model focuses on more directly sport-related factors (historical medal counts, host effects, etc.)

- **Assumption 3:** Countries with an annual average of the total number of medals won in Olympic competitions greater than 1 are assumed to be 'high medal' countries. In contrast, a country with an average value of less than 1 was considered a 'low medalist' country. Justification:In the regression analysis for the prediction of the medal table, the mean data of large-scale 'low medal countries' have a large impact on the parameters of the entire regression model and need to differentiate between countries that have not won any medals, so it is reliable to divide it into these two categories.

- **Assumption 4:** Assuming that an athlete's participation career is considered to be four Olympic cycles, predicting the number of medals for the new edition uses the participation data from the previous three editions.

Justification:Idealize an athlete's career as 16 years, or four Olympic cycles. That is, most athletes compete in a maximum of four Olympic Games during their careers. Using data from the first three will cover the majority of an athlete's career, thus providing more comprehensive and representative data.

3 Notations

As shown in Table 1, the following symbols are used throughout this paper to formalize our Olympic medal prediction framework. Key dimensions include:

- Temporal indices (x) for Olympic editions since 2000
- Sport-specific indices (i) for event-level analysis
- Medal-type superscripts (g, s, b) denoting gold, silver, and bronze

4 Data visualizations

According to the given data, we transform the complex statistical information into intuitive charts, analyze the situation of typical countries' winning patterns, and compare the changes in the number of national medals at different times and under different conditions. Mining the features related to the number of medals. We employ the LASSO regression method [3] for

Table 1. Summary of key symbols and their descriptions.

Symbol	Description
M	Total number of sports
i	Sequential numbering of sports $(1 \leq i \leq M)$
x	Number of Olympic Games editions since 2000
Md_x^g	Number of gold medals at the xth Olympic Games
Md_x^s	Number of silver medals at the xth Olympic Games
Md_x^b	Number of bronze medals at the xth Olympic Games
$m_{x,i}^g$	Number of gold medals in the ith sport at the xth Olympics
$m_{x,i}^s$	Number of silver medals in the ith sport at the xth Olympics
$m_{x,i}^b$	Number of bronze medals in the ith sport at the xth Olympics

variable selection in our predictive model, which effectively handles high-dimensional data by shrinking less important coefficients to zero.

- **Effect of number of participants:** According to the results presented in Figures 3 and 4, as the number of parameters increases, the number of medals won in each event also increases. There is a positive correlation between the number of participants and the number of medals.
- **Host effect:** According to the trend in Figure 4, there was a surge in the number of medals won by the United States as the host at the 1984 Los Angeles Olympics, suggesting that there is usually some correlation between host countries in terms of the number of medals won. Consider the host-energy feature.
- **Good Athlete Effect:** According to Figure 5, some of the best performing "star athletes" have had a positive impact on the number of U.S. medals. Therefore, the correlation between good athletes and the number of medals is considered.
- **Analysis of the "Great Coach Effect":** Examine the potential impact of highly effective coaches on medal counts when they coach in different countries. Estimate the contribution of the 'Great Coach Effect' to the medal results. Select three countries and analyze which sports they should consider hiring 'great coaches' and estimate the potential impact of such hires on their medal performance.

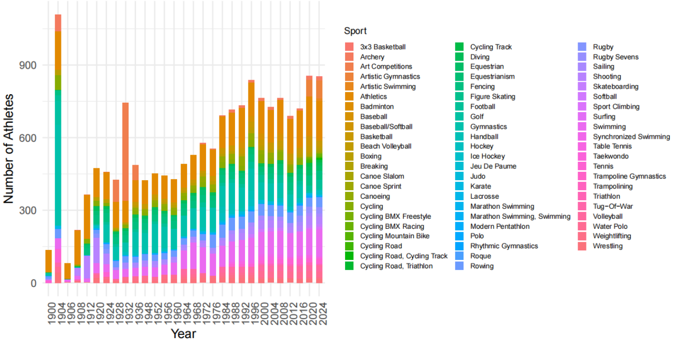


Figure 3. Previous U.S. participation in each sport in Olympics.

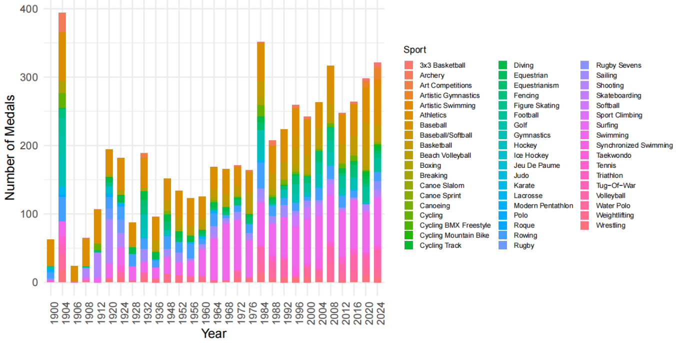


Figure 4. U.S. Olympic medal counts in each sport in previous Olympics.

5 Data preprocessing

5.1 Indicator selection and interpretation

Based on the above analysis, we considered in the model indicators such as the number of participants, the number of prizes won, whether the Games were hosted or not, and an indicator of the gender ratio based on demographic characteristics.

Based on the assumption of four cycles for the athletes, n is set as the n-th Olympic Games in which the athletes have participated in the past, either 1, 2 or 3. We have expressed these indicators in the following form.

5.1.1 Number of entries and awards

Key performance indicators are defined in Table 2, quantifying historical participation $(X_{t-n,i}^{\text{sum}})$ and medal achievements $(T_{t-n,i}^g, T_{t-n,i}^s, T_{t-n,i}^b, H_{t-n,i}^{\text{sum}})$ per sport i over n prior Olympics.

5.1.2 Hosts

The host country advantage is captured through a binary indicator variable, as formally defined in Table 3. This accounts for documented hosting effects including additional athlete quotas, home-field advantage, and preferential scheduling observed in prior Olympic studies.

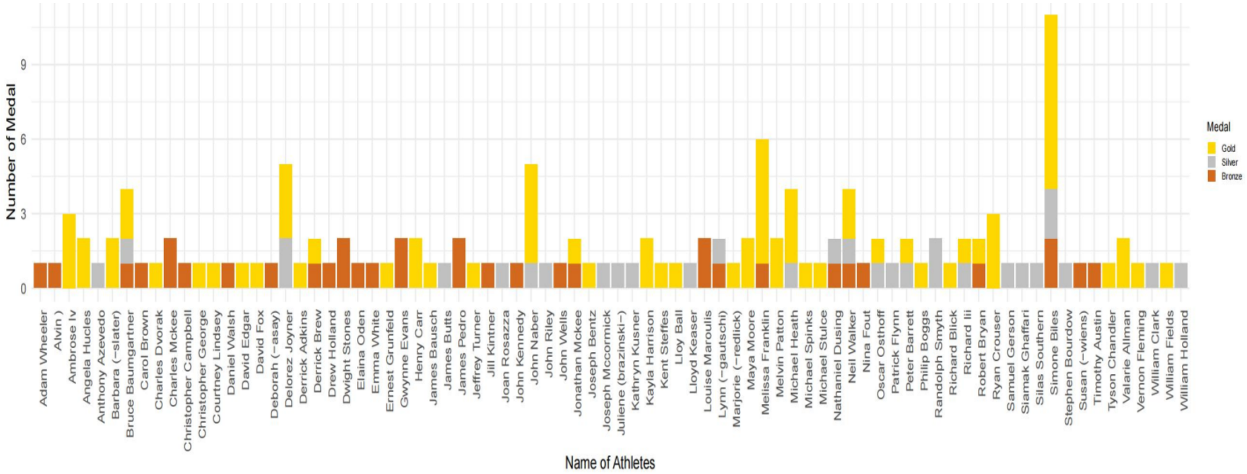


Figure 5. Typical U.S. 'Star Athletes' and medal counts.

Table 2. Symbols for the number of entries and awards and their descriptions.

Symbol	Description
$X^{\text{sum}}_{t-n,i}$	Number of participants in the i-th sport over the past n Olympic Games
$T^{s}_{t-n,i}$	Number of silver medals in the i-th sport over the past n Olympic Games
$T^{g}_{t-n,i}$	Number of gold medals in the i-th sport over the past n Olympic Games
$T^{b}_{t-n,i}$	Number of bronze medals in the i-th sport over the past n Olympic Games
$H^{\text{sum}}_{t-n,i}$	Total number of medals in the i-th sport over the past n Olympic Games

Table 3. Symbols for hosts and their descriptions.

Symbol	Description
Z	$Z = 1$ indicates that the country is the host and $Z = 0$ that it is not the host

5.1.3 *Total statistics*

Global performance benchmarks are defined in Table 4, capturing sport-level medal distributions ($G^{g}_{t-n,i,j}$, $G^{s}_{t-n,i,j}$, $G^{b}_{t-n,i,j}$) and gender compositions ($F^{g}_{t-n,i,j}$, $F^{s}_{t-n,i,j}$, $F^{b}_{t-n,i,j}$, $F^{sum}_{t-n,i,j}$) across all countries. These metrics provide context for evaluating a nation's relative performance in each sport i over n prior Games.

Table 4. Symbols for total statistics and their descriptions.

Symbol	Description
$G^{g}_{t-n,i,j}$	Number of gold medals in the i-th sport over the past n Olympic Games for all countries
$G^{s}_{t-n,i,j}$	Number of silver medals in the i-th sport over the past n Olympic Games for all countries
$G^{b}_{t-n,i,j}$	Number of bronze medals in the i-th sport over the past n Olympic Games for all countries
$F^{g}_{t-n,i,j}$	Proportion of male gold medals in the i-th sport over the past n Olympic Games for all countries
$F^{s}_{t-n,i,j}$	Proportion of male silver medals in the i-th sport over the past n Olympic Games for all countries
$F^{b}_{t-n,i,j}$	Proportion of male bronze medals in the i-th sport over the past n Olympic Games for all countries
$F^{sum}_{t-n,i,j}$	Proportion of male medals in the i-th sport over the past n Olympic Games for all countries

5.2 Standardisation of indicator

When studying the prediction of the total number of medals and the number of gold, silver and bronze medals for a country, we choose different explanatory variables to build the model respectively.

Taking the gold medal as an example, we chose a total of 9 categories such as $X^{\text{sum}}_{t-n,i}$, $T^{g}_{t-n,i}$, $H^{\text{sum}}_{t-n,i}$ as explanatory variables, with a total of 28 subdivided explanatory variables. The distribution of these 28 variables is shown in Figure 6 (taking the United States as an example).

Prior to normalization (Original Feature Distribution), the range of values for each feature shows significant heterogeneity. The distribution pattern also tends to be more homogeneous, the difference in the distribution of values between features is reduced, and the dispersion of the data is well controlled.

- The values of some features are highly

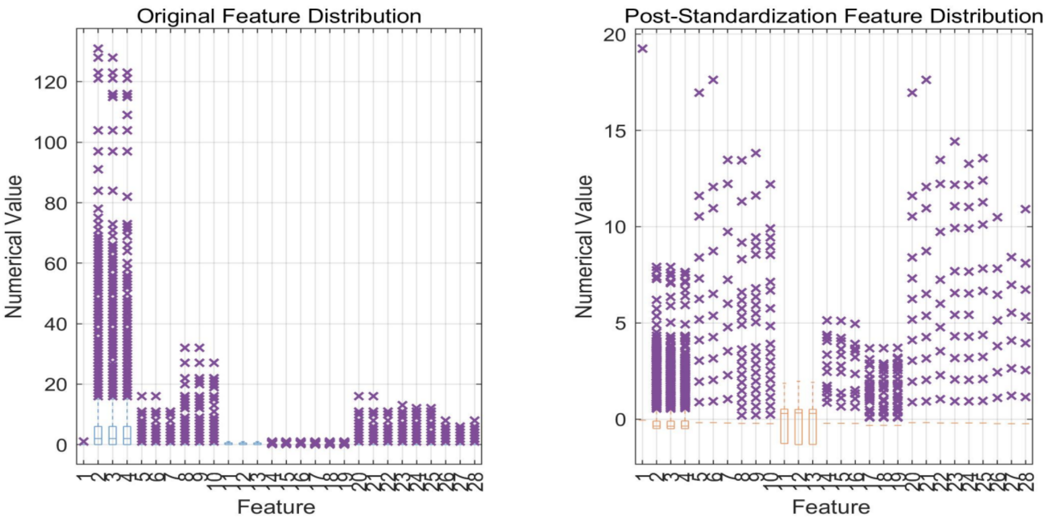


Figure 6. Distribution of 28 explanatory variables for the U.S. Gold Medal.

concentrated in the smaller interval of 0-20, while the values of some other features can be as high as about 120, showing a large span. In terms of the distribution pattern, the distribution of the values of each feature is extremely uneven, with the three variables 1, 2, and 3 being more discrete, and the rest being relatively compact, reflecting the complexity of the original data in terms of feature dimensions.

- After the standardization process, the numerical range of the features has changed significantly, and the whole is significantly smaller and more concentrated, effectively reducing the numerical span.

Meanwhile, based on the source data, we divided 2/3 of the data as the training set and 1/3 of the data as the test set. In turn, we obtained four indicators, namely, Mean Square Error, Root Mean Square Error, Coefficient of Determination and Mean Absolute Error, between each sport and the total number of medals on the training set and test set, and accordingly filtered out 3 × 3 Basketball,Badminton, Basketball, Beach Volleyball, Boxing, and 48 other sports categories that have a greater impact on the total number of medals per country, as a way to conduct a follow-up study.

6 Question 1: Forecast and analysis of the medal situation at the 2028 Olympics

6.1 Forecast of total number of medals

6.1.1 Multiple linear regression modeling and solution

To predict a country's medal total for the 2028 Olympics, we use two methods.

- **Method 1:** Direct modeling for the total number of medals, choosing appropriate explanatory variables in the set indicators for direct modeling for prediction; the following multiple linear regression model was developed:

$$y = \beta_0 + \sum_{n=1}^{3} \beta_{1,n} X^{\text{sum}}_{t-n,i} + \sum_{n=1}^{3} \beta_{2,n} H^{\text{sum}}_{t-n,i} + \sum_{n=1}^{3} \beta_{3,n} R_{t-n,i} + \sum_{n=1}^{3} \beta_{4,n} N^{\text{sum}}_{t-n,i} + \sum_{n=1}^{3} \beta_{5,n} F^{\text{sum}}_{t-n,i} + \beta Z \tag{1}$$

- **Method 2:** The number of gold, silver and bronze medals in the 2028 Olympic Games of that country is predicted separately, which is then summed and aggregated to give the total number of medals. According to the classification of sports, the following multiple linear regression model was established with the number of gold medals in each sport as the explanatory variable:

$$G = \beta_0 + \sum_{n=1}^{3} \beta_{1,n} X^{\text{sum}}_{t-1,i} + \sum_{n=1}^{3} \beta_{2,n} T^{g}_{t-n,i} + \sum_{n=1}^{3} \beta_{3,n} H^{\text{sum}}_{t-n,i} + \sum_{n=1}^{3} \beta_{4,n} R_{t-n,i} + \sum_{n=1}^{3} \beta_{5,n} N^{g}_{t-n,i} + \sum_{n=1}^{3} \beta_{6,n} N^{\text{sum}}_{t-n,i} + \sum_{n=1}^{3} \beta_{7,n} G^{g}_{t-n,i} + \sum_{n=1}^{3} \beta_{8,n} F^{g}_{t-n,i} + \sum_{n=1}^{3} \beta_{9,n} F^{\text{sum}}_{t-n,i} + \beta Z \tag{2}$$

Similarly, the number of silver (S) and bronze (B) medals can be derived from a similar model . The resulting model for the total number of medals is:

$$Y = G + S + B \tag{3}$$

6.1.2 Regression weighting factor (β)

According to the preprocessing data, the parameters are obtained. Due to too much data, only the regression coefficients ofthe total number of gold, silver, bronze and medals in the U.S. boxing program are taken as an example, and some of the regression weight coefficients are shown in Figure 7.

6.1.3 Model Optimization

In order to improve the generalizability of the model, the historical data of other countries were brought into the set of models for 50 iterations to obtain the root mean square error of the training and test sets, as shown in Figure 8.

As the number of iterations of the model increases, the RMSE values of both the training and validation sets decrease and converge to 1. In some iterations, the RMSE values of the training set are significantly lower than those of the validation set, which indicates that the model performs well in the training set but poorly in the validation set, and there is a certain degree of overfitting phenomenon. It is noted that the RMSE value fluctuates less at 46 iterations, when the model is more stable and the RMSE value of the training set is higher than that of the validation set, so the model at this point is selected as the final model.

Subsequently, we obtain a residual plot for the validation set, as shown in Figure 9.

- Distribution of remnants: The distribution of the residuals is more randomly distributed around the red reference line, indicating that there is no significant systematic bias in the model's prediction error, whereas ideally the residuals should be randomly distributed with no recognizable pattern, which is consistent with the model's assumptions.
- Residual size: Most of the residual values are between -2 and 2, which indicates that the predictions of the model are usually close to the actual values; the RMSE value is 0.88, which is relatively low, indicating that the model's prediction error on the validation set is small and the model performs well.

For the comparison of the model's residual distribution with the normal distribution on the training and validation sets, we obtain Figures 10 and 11.

The distribution of residuals on both the training set and the validation set is roughly normal, and the test set is closer to the normal distribution than the training set, and they are all concentrated around 0. This indicates that the model fits better on the training set, with smaller residuals and a more centralized distribution, which is in line with the normality assumption of the linear regression model, and the assumption conditions of the model are satisfied.

6.1.4 Model Evaluation

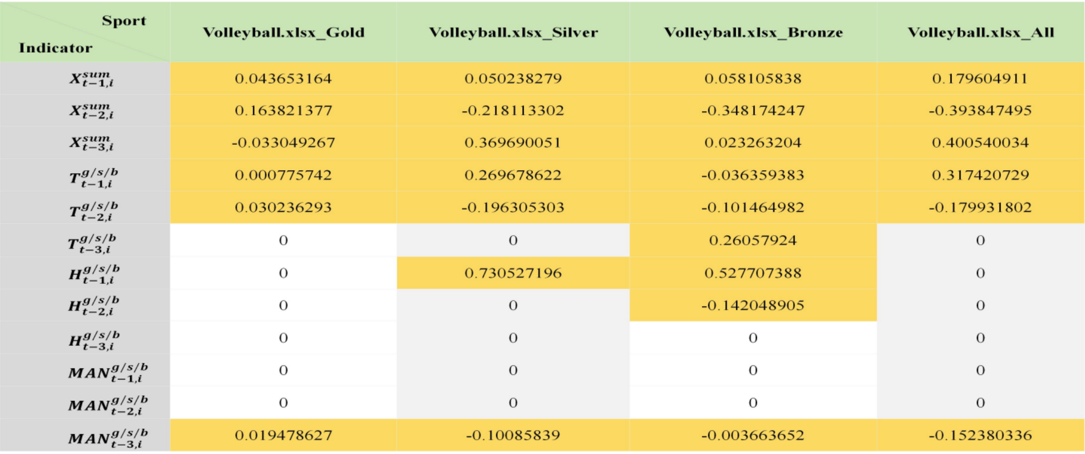
In order to evaluate the generalization ability and prediction accuracy of the model, the comparison between the number of predicted medalsand the number of actual medals under the training and validation sets are obtained, as shown in Figure 12 and 13.

- **Figure 12:** The RMSE value is 0.93, and from the scatter distribution, when the actual number of medals is small (≤ 15), the numerous scatters are closely clustered near the dotted line representing the predicted value is equal to the actual value. It means that the model is able to fit the data accurately on the training data, the deviation between the predicted and actual values is small, and the model is capable of capturing the features of the training data.
- **Figure 12:** The RMSE value is 0.88, which is slightly reduced compared to the RMSE value of the training set, and in the scatter distribution, the scatters are more discrete compared to the training set, and are no longer highly concentrated near the dotted line, which reflects that the model's prediction effect on the validation set is slightly insufficient. There is a slight overfitting phenomenon of the model on the training set, but overall the generalization ability of the model is still better, and it performs better on both the training and validation sets, and the difference between the predicted values and the actual values is smaller, which indicates that the model has a better prediction ability.

In conclusion, after a comprehensive assessment of the model, it can be concluded that the predictive performance of this model is relatively excellent.

6.1.5 Model Conclusions

In the case of the United States, based on the Method 1 model (1) above, with the appropriate data brought in, the total number of U.S. medals predicted for 2028 turns out to be 145. Model 2 predicts about 2 gold medals for U.S. boxing in 2028. Total U.S. gold medals are estimated at 50, with silver and bronze medals at



Sport / Indicator	Volleyball.xlsx_Gold	Volleyball.xlsx_Silver	Volleyball.xlsx_Bronze	Volleyball.xlsx_All
$X^{sum}_{t-1,i}$	0.043653164	0.050238279	0.058105838	0.179604911
$X^{sum}_{t-2,i}$	0.163821377	-0.218113302	-0.348174247	-0.393847495
$X^{sum}_{t-3,i}$	-0.033049267	0.369690051	0.023263204	0.400540034
$T^{g/s/b}_{t-1,i}$	0.000775742	0.269678622	-0.036359383	0.317420729
$T^{g/s/b}_{t-2,i}$	0.030236293	-0.196305303	-0.101464982	-0.179931802
$T^{g/s/b}_{t-3,i}$	0	0	0.26057924	0
$H^{g/s/b}_{t-1,i}$	0	0.730527196	0.527707388	0
$H^{g/s/b}_{t-2,i}$	0	0	-0.142048905	0
$H^{g/s/b}_{t-3,i}$	0	0	0	0
$MAN^{g/s/b}_{t-1,i}$	0	0	0	0
$MAN^{g/s/b}_{t-2,i}$	0	0	0	0
$MAN^{g/s/b}_{t-3,i}$	0.019478627	-0.10085839	-0.003663652	-0.152380336

Figure 7. Regression full weight coefficients for US boxing programs.

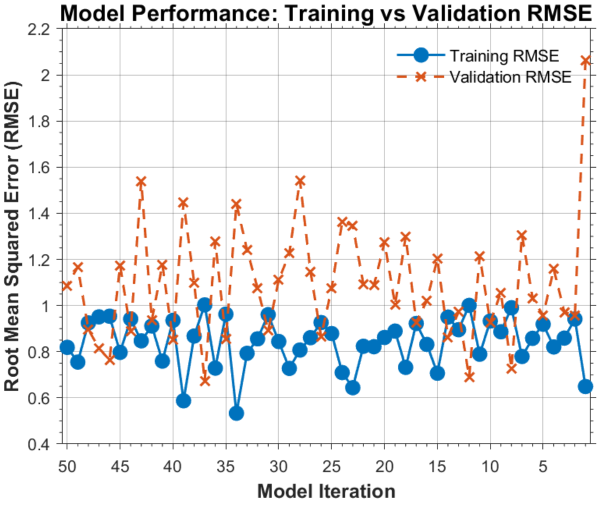


Figure 8. Training vs Validation RMSE.

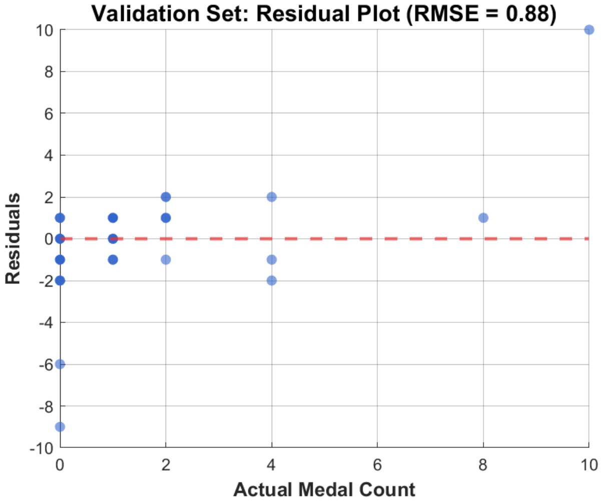


Figure 9. Residual plots for the test set.

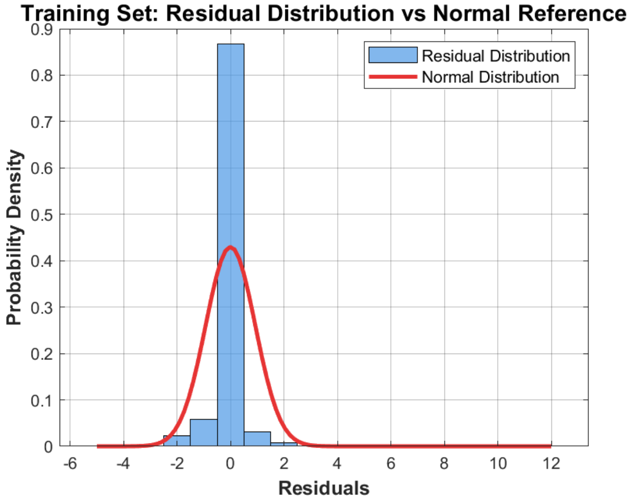


Figure 10. Training set.

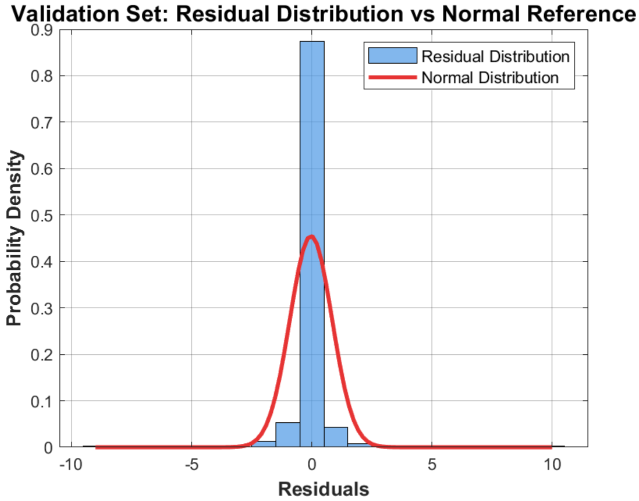


Figure 11. Validation set.

52 and 46, respectively, leading to a projected total of about 126 medals.

Using both models to predict medals for other countries, the results from both methods are similar. The top 10 countries in the 2028 medal table, as predicted, are shown in Figure 14.

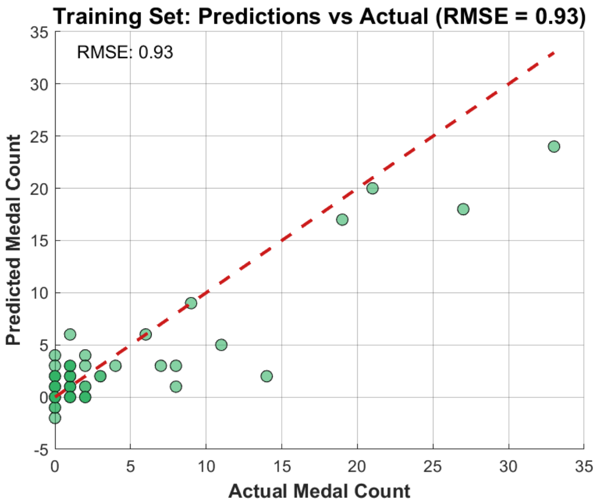


Figure 12. Training Set.

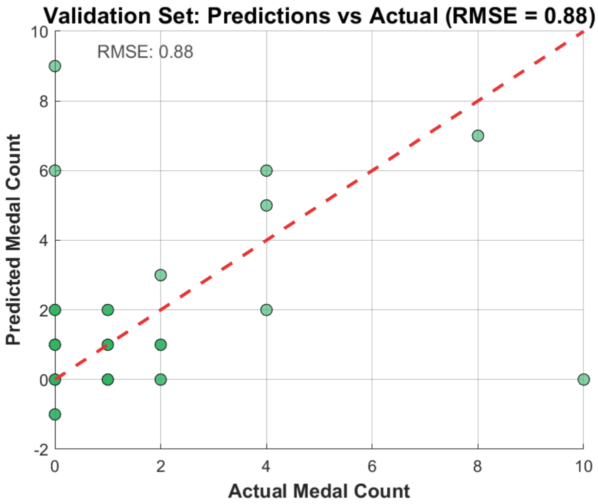


Figure 13. Validation set.

Geographically, the U.S. is leading with the most medals, including the highest counts of gold, silver, and bronze, followed by China, New Zealand, France, Japan, Great Britain, Russia, Australia, Italy, and Spain. Most are expected to improve their medal totals from 2024 to 2028, except for Great Britain, which may experience a slight decline. Countries likely to improve include Russia, Argentina, Brazil, Canada, China, Denmark, France, Germany, India, Japan, Jamaica, Mexico, New Zealand, Norway, Poland, and Romania. Some countries might see a decrease in medals from 2024 to 2028, like the U.K., Hungary, South Korea, Israel, Cuba, Turkey, and others.

6.2 First Medal Prediction

6.2.1 Data selection

- Separate screening of countries that have won and those that have not won in previous Olympic competitions.
- Dividing the time period among the countries that have received the award is 20002024, 2004-2024, 2008-2024 respectively ......
- Take every two neighboring datasets in turn, e.g., take 2000-2024 and 2004-2024, take 2004-2024 and 2008-2024, and so on.
- Each time, the set with the larger year span is taken as the full set, the complement of the intersection of these two datasets is taken, and the country that wins the first medal (denoted as dataset 1) is filtered in this complement to build the model.
- The non-winning countries were used in the prediction set (denoted as dataset 2) to predict whether or not the first medal would be won in 2028.

6.2.2 LASSO Regression and Logistic Regression Combined Modeling

Combining LASSO regression with logistic regression, the modeling process is shown in Figure 15, from which it is determined whether these countries that have never won a medal in the past will obtain the first medal in the program, and the same can be obtained for other program awards.

LASSO regression is performed and the corresponding model, based on the first question, is obtained with the explanatory variables shown in Figure 16.

The LASSO regression selected λ is 5.15e-01, there are three explanatory variables that have the greatest impact on the explanatory variables, here we will set them in order as X1, X2, X3, and then logistic regressions for countries that have not won a medal so far, in which winning the first medal is recorded as 1, and 0 otherwise. Model the following:

$$y = \frac{1}{1 + e^{a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3}} \tag{4}$$

$$p(y \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}} \tag{5}$$

6.2.3 Model Conclusions

In the end, we predict 2 countries that could win their first medal in 2028, UA Emirates and Montenegro, whose corresponding events are Judo and Water Polo, respectively.

- Judo is not a dominant sport in the UAE, and the Olympic competition is highly competitive, typically dominated by traditional judo

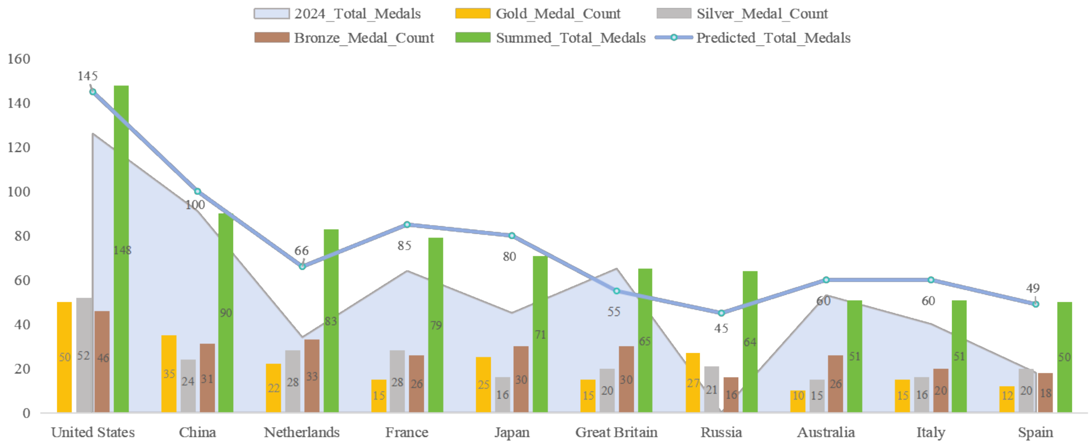


Figure 14. Top ten countries in terms of total number of medals and their medal situation.

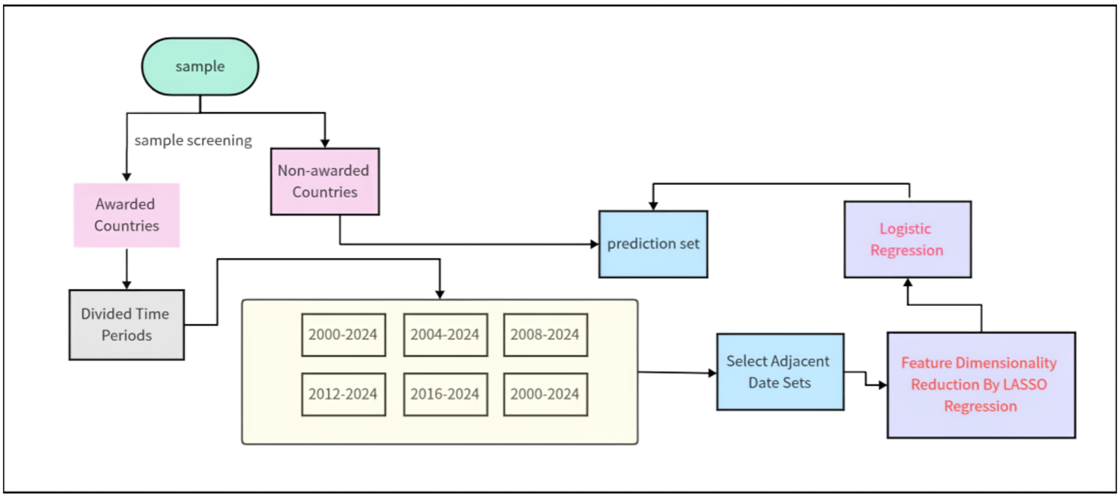


Figure 15. LASSO regression and logistic regression modeling flowchart.

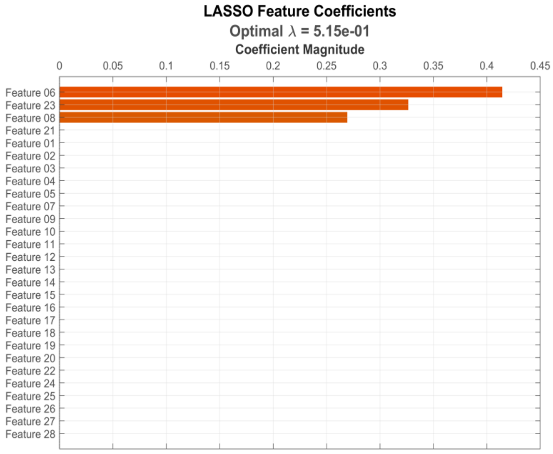


Figure 16. Explanatory variables for LASSO regression.

powerhouses like Japan, France, and Russia. However, considering the UAE's hosting of the 2024 Abu Dhabi World Judo Championships indicates a desire and capability to win their first medal. Historical data suggests a slim chance of achieving this feat.

- Montenegro's water polo team is strong and could win their first medal at the 2028 Los Angeles Olympics if conditions are favorable.

6.3 Great Coach's Analysis of Medal Level Increases

To summarize the total number of prizes won by each country in each sport over the years and to get the top of the list in that sport, we have intercepted nine traditional and three emerging sports, as shown in Table 5.

- **Advantageous programs:** Athletics, Swimming, Wrestling, Diving, Shooting, Sailing, Boxing are all advantageous programs in the United States, especially Athletics.
- **Emerging sports:** The 2024 Paris Olympics has new emerging sports such as surfing, rock

Table 5. Medal count by sport and NOC.

Sport	Medal_Count	NOC
Athletics	861	USA
Swimming	578	USA
Fencing	136	ITA
Wrestling	63	USA
Diving	149	USA
Shooting	56	USA
Boxing	54	USA
Table Tennis	66	CHN
Sailing	126	USA
Surfing	2	BRA
Climbing	4	FRA
Breaking	1	JAP, CAN

climbing and breakdancing, and Brazil, Japan, France and Canada are more dominant in emerging sports.

- **Host effect:** The host country's advantage is significant due to the number of events. The host can introduce beneficial events or remove unfavorable ones, and influence rules, affecting medal outcomes.

7 Question 2:Great Coach's Analysis of Medal Level In-creases

7.1 Data preprocessing

7.1.1 Data Screening and Preprocessing

We illustrate the preprocessing methodology using US and Chinese women's volleyball as a case study. After extracting sport-specific historical data, all variables were standardized via Z-score normalization to ensure comparability. The operational definitions of these transformed variables are detailed in Table 6, which includes:

- The ordinal medal outcome Y_i (gold=3 to no medal=0)
- Binary indicators for host status (x_1) and elite coaching (x_2)
- Contextual controls for total medals (x_3) and national programs (x_4)

7.1.2 Data Overview

The number of gold, silver, bronze, and total medals won by the U.S. vs. China women's volleyball team and the U.S. vs. Romania women's gymnastics in the corresponding time periods are shown in Figures 17 and 18. (PS: The red line on the timeline indicates the period of time the coach has been coaching in the country of his/her nationality, and the blue line indicates the period of time the coach has been coaching abroad.)

Table 6. Variable definitions.

Code	Instruction
Y_i	Medal level: divided into no medal (0), bronze (1), silver (2), gold (3)
x_1	Host: 1 for being the host, 0 otherwise
x_2	Great Coach: 1 for having a great coach, 0 otherwise
x_3	Total medals: standardized continuous variable
x_4	Country: 0 for USA, 1 for China

Regardless of whether a coach coaches in the country of his nationality or overseas, the number of medals for the country he coaches changes significantly after he joins.

For example, during Lang Ping's tenure in the United States from 2005-2008, the United States won a medal atthe 2008 Olympics, ending the regret of not winning a medal in previous years. Romania won more medals than ever before during Béla Karolyi's tenure, and when Karolyi joined the U.S., the number of U.S. women's gymnastics medals increased dramatically and showed a trend of growth, while Romania's medals in the program showed a trend of decline. This tells us that the coaching effect plays a positive role in the total number of national awards.

7.2 Bayesian Ordered Probit Modeling

We employ the Bayesian ordered Probit model following [2], where the latent variable Z is assumed to follow a normal distribution with thresholds T_j dividing the medal categories (no medal, bronze, silver, gold). The Gibbs sampling procedure in our Bayesian model follows the foundational work of [6], who first established the convergence properties of stochastic relaxation algorithms for high-dimensional posterior distributions.

7.2.1 The segmentation of the threshold

Using τ_1 to denote the threshold from no medal to bronze, τ_2 to denote the threshold from bronze to silver, and τ_3 to denote the threshold from silver to gold.

The regression coefficient β affects the value of the latent variable Z, which indirectly affects the category

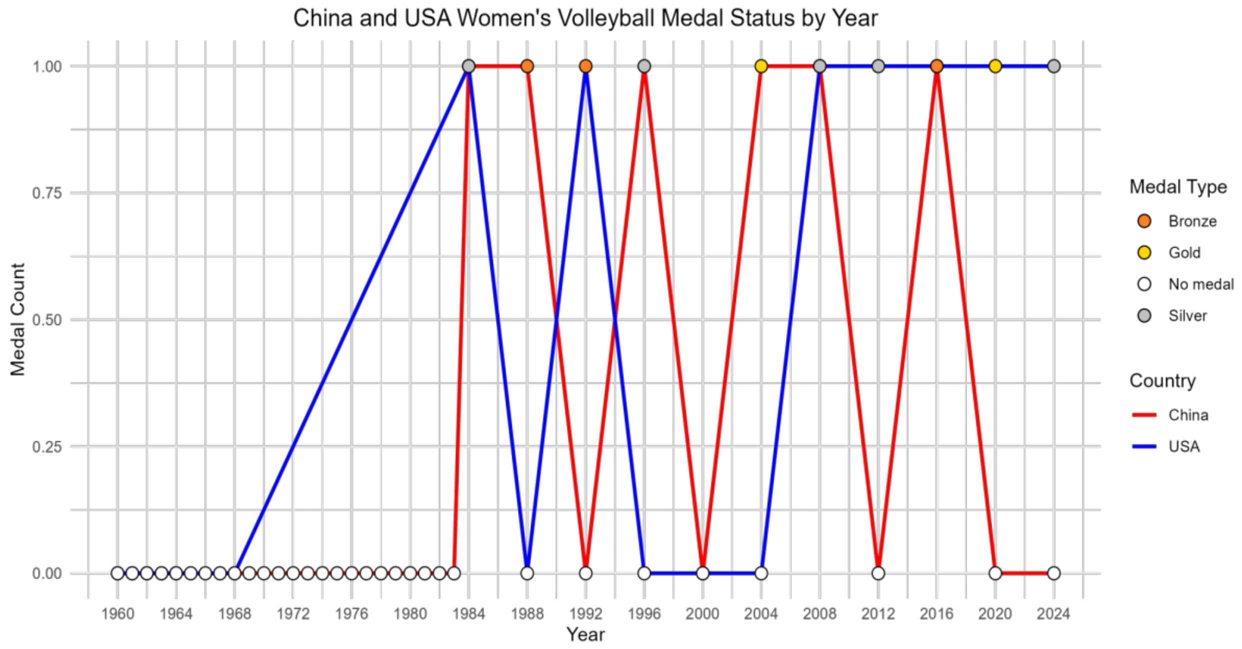


Figure 17. China and USA women's volleyball medal status by year.

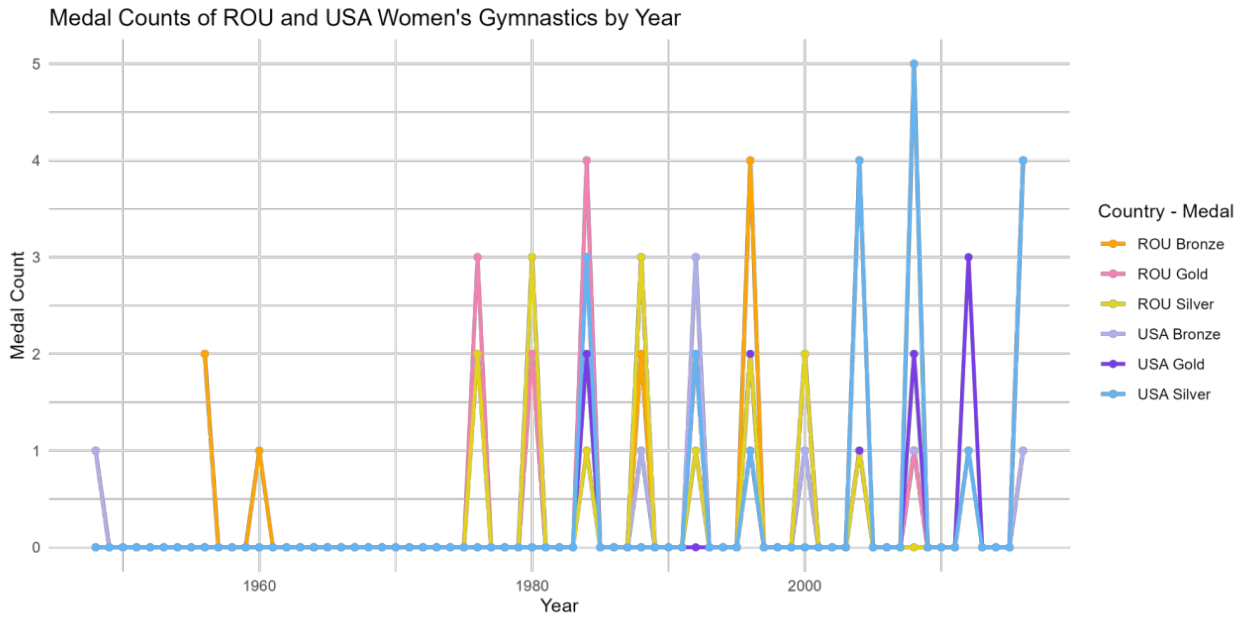


Figure 18. Medal counts of ROU and USA women's gymnastics by year tional awards.

of Y. The specific formula is as follows:

$$Y = \begin{cases} 0 & \text{if } Z \leq T_1 \\ 1 & \text{if } T_1 < Z \leq T_2 \\ 2 & \text{if } T_2 < Z \leq T_3 \\ 3 & \text{if } Z > T_3 \end{cases} \quad (3)$$

where:

- Z is the latent variable, usually assumed to follow a normal distribution $Z \sim \mathcal{N}(\eta, 1)$.
- $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ is the linear predictor.
- T_j are the threshold values, satisfying $T_0 = -\infty$ and $T_4 = +\infty$.

7.2.2 Initialization parameters

We need to set initial values for the regression coefficients β and the threshold values τ. These initial values can be randomly selected or determined based on prior knowledge or the results of preliminary data analysis. The choice of initial values may affect the convergence speed of the algorithm, but in most cases, Gibbs sampling is relatively insensitive to the choice of initial values.

7.2.3 *Gibbs Sampling*

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method used to draw samples from multivariate probability distributions. In this model, we alternately sample the latent variable Z, regression coefficients β, and threshold values τ.

The specific steps are as follows:

1. Sample the latent variable Z: Given the current values of β and τ, draw a new value of Z from the conditional distribution.
2. Sample the regression coefficients β: Given the current values of Z and τ, draw a new value of β from the conditional distribution.
3. Sample the threshold values T: Given the current values of Z and β, draw a new value of τ from the conditional distribution.

This process is repeated until a predetermined number of iterations or convergence criteria are met.

7.2.4 *A posteriori inference*

After Gibbs sampling is completed, we perform a posteriori inference based on the sampling results. Specifically, we compute the posterior means of the parameters with 95% and 90% confidence intervals, which help us assess the stability of the model.

7.2.5 *Model equation*

The medal hierarchy is an ordered categorical variable, and the Probit model has a natural advantage in dealing with such data. Moreover, the Probit model effectively captures the complex relationships between variables through its link function, introducing nonlinear transformations.

In the case of small sample sizes, the Bayesian method helps alleviate overfitting issues and enhances model robustness by introducing prior distributions.

The Probit model is given by the formula:

$$P(Y = j \mid \mathbf{X}) = \Phi(T_j - \eta) - \Phi(T_{j-1} - \eta) \tag{7}$$

- $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ is the linear predictor.
- T_j are the threshold values ($T_0 = -\infty, T_4 = +\infty$).
- Φ is the cumulative distribution function of the standard normal distribution.

Among them, the Bayesian method updates the posterior distribution by introducing prior distributions and combining them with observed data:

$$P(\beta, T \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \beta, T) \cdot P(\beta) \cdot P(T) \tag{8}$$

- $P(Y \mid \mathbf{X}, \beta, T)$ is the likelihood function.
- $P(\beta)$ and $P(T)$ are the prior distributions of the regression coefficients and threshold values, respectively.

The posterior distribution of regression coefficients illustrates the effect size and direction of each variable. The posterior distribution of the split thresholds verifies the model's reasonable classification of medal grades. After inputting the preprocessed data sets into the model and outputting their weight coefficients at 90% and 95% confidence intervals, controlling the significance level ($\text{p} < 0.1$), the results are shown in Figures 19 and 20.

- **Host Country (X1)**: 90% HDI interval is [0.25, 1.39], does not include 0, indicating a significant effect ($p < 0.1$).
- **Great Coach (X2)**: 90% HDI interval is [0.78, 1.70], does not include 0, indicating a significant effect ($p < 0.1$).
- **Total Medals (X3)**: 90% HDI interval is [0.10, 0.80], does not include 0, indicating a significant effect ($p < 0.1$).
- **Country (X4)**: 90% HDI interval is [−0.82, 0.20], includes 0, indicating no significant effect.

7.3 Model conclusions

The results align with our expectations. We can conclude that within different confidence intervals, being the host country, having a great coach, and total medals have significant effects on the performance of China's and the USA's women's volleyball teams. The lower bounds of their confidence intervals do not include zero, and the great coach indicator has the highest weight, with an expected value above 1. The total medals indicator plays a positive marginal role, reflecting the overall level of the country, while the country indicator (X4) is not significant, highlighting the substantial impact of great coaches and host status on medal counts.

When applying the model to the historical medal wins of Nigeria and the USA in gymnastics, we find that the country itself also has a significant effect, indicating different historical backgrounds. We reasonably

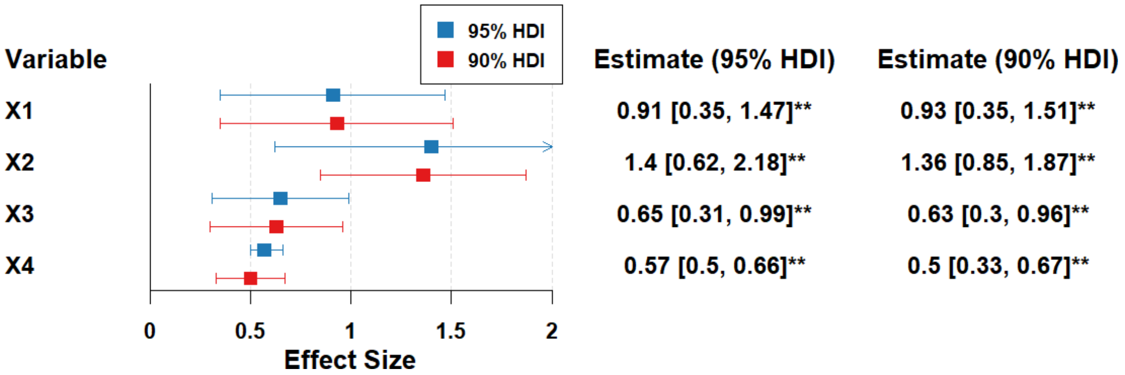


Figure 19. Comparison of USA and Nigerian gymnastics.

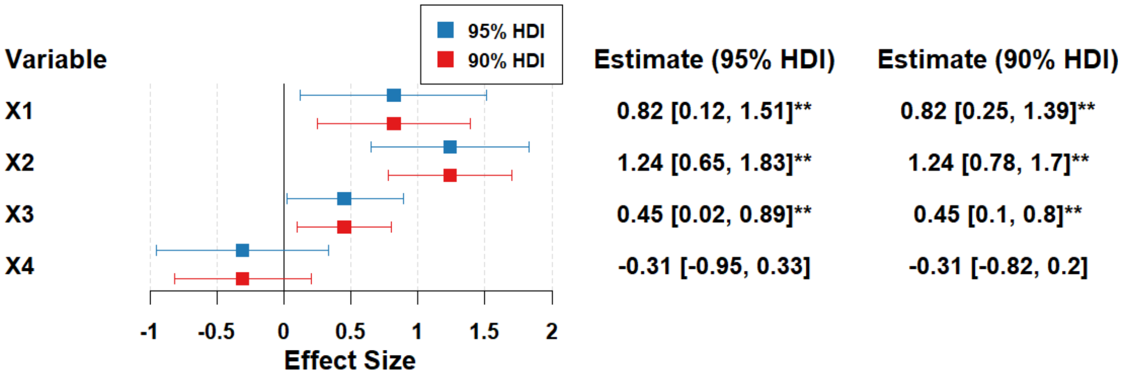


Figure 20. Comparison of USA and China volleyball.

speculate that economic levels may influence their performance in the events.

7.4 Robustness analysis of the model:

7.4.1 Posterior Distribution of Split Values

The estimated threshold parameters T_j dividing medal categories (no medal, bronze, silver, gold) are reported in Table 7. These posterior distributions were obtained through 10,000 iterations of Gibbs sampling [6], with convergence confirmed by $\hat{R} < 1.01$ for all parameters. The ordering $T_1 < T_2 < T_3$ (-0.74 < 0.15 < 1.08) validates our proportional odds assumption, consistent with Bayesian ordered probit theory [2].

Notably, the 90% highest density intervals (HDI) show:

- Clear separation between categories (T_1 HDI [-1.10, -0.38] vs T_2 HDI [-0.20, 0.50])
- Significant gap between silver and gold thresholds ($T_3 > T_2 + 2\sigma$)

This structure reflects the increasing difficulty of medal upgrades in elite competition.

7.4.2 Model Robustness Analysis: Posterior Predictive Checks

By comparing the consistency between observed data and model predictions. The results in Table 8 demonstrate exceptional model performance:

Table 7. Posterior distribution of split values.

Threshold	Posterior Mean	90% HDI Interval
T1	-0.74	[-1.10, -0.38]
T2	0.15	[-0.20, 0.50]
T3	1.08	[0.65, 1.51]

$$\mathrm{PPC} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{4}$$

Table 8. Posterior distribution of split values.

Statistic	Value
Mean Absolute Error	0.23
Maximum Error	0.67
Error Standard Deviation	0.12

7.4.3 Posterior Predictive Checks

- The value represents the average absolute difference between the model's predicted values and the actual observed values. A lower MAE indicates higher prediction accuracy of the model, which can effectively capture the trends and patterns in the data. An MAE of 0.23 indicates that the model's prediction error is relatively small and possesses a higher degree of accuracy.

- The standard deviation measures the dispersion of the prediction error. A lower standard deviation (0.12) indicates that the model's predictions are more stable, with less variability in error, further confirming the reliability of the model.

8 Model Evaluation and Outlook

8.1 Model 1

Our model leverages historical data and factors such as participant numbers, host effects, and elite athletes for comprehensive predictions. It demonstrates strong generalization across countries and exhibits low RMSE, signifying high accuracy. However, it has limitations, including reliance on post-2000 data, potential oversight of earlier insights, and exclusion of macro factors like GDP. Its complexity and computational cost may impede real-time forecasting and decision-making. The optimization problem is solved via coordinate descent [4], providing computational advantages when dealing with high-dimensional predictors in Olympic medal data.

8.2 Model 2

This model's non-linear processing, combined with Bayesian methods, presents a significant advantage. It mitigates overfitting risks under small sample conditions and captures complex variable relationships. Notably, its temporal integration mechanism enhances time-ordered data modeling by tracking lagged effects and temporal evolution. Despite improved predictive performance, there is room for further optimization. The model's construction may affect cross-prediction stability, and Bayesian sensitivity to prior distribution choices requires attention, particularly with limited data. Sensitivity analysis and robustness tests are recommended to strengthen conclusions.

9 Memorandum to the Olympic Committee

After an in-depth study of the influencing factors related to the Olympic Games, in order to contribute to the better development of the Olympic Games, we would like to provide you with the following reference suggestions.

First, being the host nation significantly influences the medal count in the Olympics. Historically, countries like the US, UK, China, Australia, Japan, and Brazil have shown a notable advantage when hosting the Games. This is due to factors like extra participation slots, possible referee bias, and favorable scheduling. To reduce this disparity, non-host countries can prepare thoroughly and offer psychological support to ensure athletes perform optimally.

Second, economic development is closely linked to the number of Olympic medals. Wealthier countries like the US, China, and the UK invest more in sports, leading to better athlete performance and higher medal counts. For less developed countries, effective resource allocation and strategic project development can boost Olympic success.

Additionally, the variety and quantity of events significantly affect medal counts. Some nations excel in traditional events, like China in table tennis and the US in athletics, winning many medals. New events also offer opportunities for less dominant countries. For instance, skateboarding and rock climbing are gaining prominence in the 2024 Olympics. Thus, countries should strategically allocate resources to leverage their strengths and explore new events to boost their medal tally.

Moreover, China's "Wolf Cultivation Plan" for table tennis, initiated in 2009, has not only produced top players like Fan Zhendong and Chen Meng but also raised the sport's global level through international exchanges and training camps, aligning with the Olympic values of "Faster, Higher, Stronger, Together." It's hoped that the Olympic Committee will encourage more international coaching to enhance participation and performance.

We hope that the above content can provide useful references for the work of the Olympic Committee, and sincerely wish that the Olympic Games can continue to be successfully held and continuously promote the vigorous development of the global sports cause!

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

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

[2] Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association, 88*(422), 669-679. [CrossRef]

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 58*(1), 267-288. [CrossRef]

[4] Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22. [CrossRef]

[5] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. [CrossRef]

[6] Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence, 6*(6), 721-741. [CrossRef]

[7] Algamal, Z., & Lee, M. H. (2009). A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. *Advances in Data Analysis and Classification, 13*, 753–771. [CrossRef]

[8] Basu, A., Ghosh, A., Jaenada, M., & Pardo, L. (2024). Robust adaptive LASSO in high-dimensional logistic regression. *Statistical Methods & Applications, 33*(5), 1217-1249. [CrossRef]

Yuxiao Zhu (Bachelor of Finance Northeast Normal University). (Email: zhuyuxiao@nenu.edu.cn)

Chenhao Wang (Bachelor of Northeast Normal University). (Email: 1729921528@qq.com)

Lemei Qin (Bachelor of Northeast Normal University). (Email: lmqin365@nenu.edu.cn)