

# A Quantitative Study of Medal Predictions and the Impact of Good Coaching at the Los Angeles Olympics

Zhenyu Cui

Shandong University, Weihai, 264200, China

**Abstract.** Focusing on the 2028 Los Angeles Summer Olympic Games, this study aims to build an accurate and effective medal prediction model and deeply explore the key factors affecting the distribution of MEDALS to provide strong support for countries to formulate scientific and reasonable Olympic strategies. First, we cleaned the raw data, dealt with outliers and introduced key variables, improved the data quality through normalization, and built a solid data foundation. Then, we adopted an integrated learning approach, combining multiple machine learning algorithms such as Logistic Regression, random forest, support vector machine, ridge regression, and XGBoost, to construct a stacked model for predicting whether a country will win medals and the number of medals. After cross-validation and grid search tuning, the model performance was evaluated in detail, and the combined prediction accuracy was 97.49%, with good stability of the prediction results. We analyzed each country's performance and found that most countries had shown positive development while predicting that the 69 countries that had not yet won a medal were expected to make a breakthrough. In addition, we established a time series regression model. The influence of coaches is prominent when they are first transferred to the team, and the improvement is mostly more than 35% when they are stabilized. Finally, we performed sensitivity analysis on the model using Monte Carlo simulations visualized and presented to indicate that the model performs well. This research provides a scientific basis for the National Olympic Committees to formulate Olympic strategies, plan athlete training, and optimize training programs, helping countries take advantage of their strengths and improve their performance in the Olympic Games.

**Keywords:** Olympic medals; machine learning; effect of coaches.

## 1. Introduction

As the world's most influential and comprehensive sports event, the number of medals is an essential indicator of each delegation's overall strength and the program's layout. It also reflects different countries' levels of investment and development in their sports programs. In recent years, home-field advantage has become a key factor in medal predictions, with host countries usually outperforming expectations due to familiarity and spectator support. For example, London 2012, Rio 2016, and Tokyo 2020 hosts have all risen significantly in the medal table. In addition, some smaller countries have made breakthroughs in specific disciplines, such as Kosovo and San Marino, which have won gold medals in less competitive disciplines and have become "dark horses" in their respective events. Changes to the Olympic program have also created new opportunities, with the addition of extreme sports such as rock climbing and skateboarding in 2020 providing breakthroughs for traditionally disadvantaged countries. Competition in these sports is relatively open, and some countries have taken advantage of this to grab a share of the medals, breaking the previous competitive pattern. In addition, the influence of good coaches cannot be ignored, such as the famous women's volleyball coach Lang Ping and gymnastics coach Marta Karolyi, who each achieved excellent results at different times in their coaching careers.

Our work focuses on predicting medal counts in the Summer Olympics by analyzing various factors. We utilized predictive models such as Support Vector Machines, Random Forest, and Logistic Regression alongside regression models like XGBoost and Ridge Regression. The stacked models (Logistic Regression and XGBoost Meta Learners) were used to determine medal outcomes and assess the influence of prominent coaches like Lang Ping and Marta Caroli, leading to an optimized model scheme for accurate predictions.

## 2. Data Visualization and Analysis

The number of competition events is on the rise as a whole, but the magnitude of change varies at different stages. In the early stages, the growth in the number of events was relatively flat. As time goes by, the influence of the Olympic Games continues to expand, and the number of events begins to accelerate, thanks to the booming development of global sports and the participation of more countries and regions.

It is worth noting that the continuous increase in the number of competition events reflects the vitality of the Olympic Games in constant innovation and development and is expected to continue to enrich the content of the events in the future, providing a broader stage for global sportsmen and women to showcase their talents.

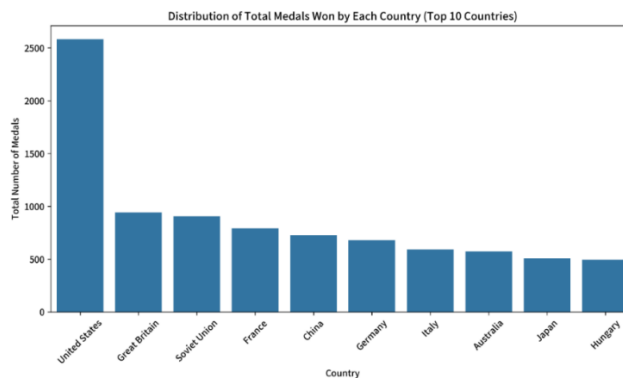


Fig. 1 Distribution of Total Medals Won by Top 10 Countries

The height of the columns represents the number of medals, and the columns of the United States are significantly higher than those of other countries, indicating that it has been more capable of acquiring medals in Olympic history. This difference reflects the differences between countries regarding their investment in sports resources, program layout, and athletic traditions. Countries with large populations and well-developed economies tend to invest more resources and win more medals. In contrast, some small countries have individual sports advantages but fewer medals overall.

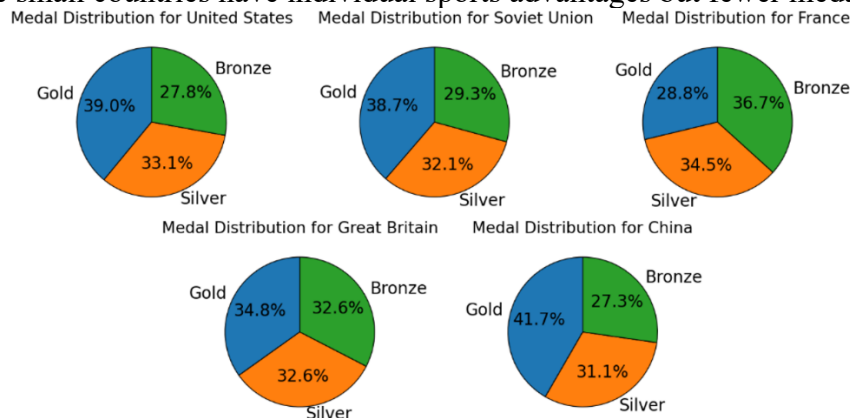


Fig. 2 The proportion of the top 5 countries in the total number of medals

The size of each sector represents the proportion of different types of medals. Some countries have a relatively high proportion of gold medals, indicating their strong dominance in Olympic sports. In contrast, some countries have a more balanced distribution of medals, reflecting a more even strength in different sports. This difference reflects the differences in the investment of sports resources, the program layout, and each country's competitive tradition. Countries with large populations and well-developed economies tend to invest more resources, focus on developing dominant sports, or balance the layout of sports to obtain more medals. In contrast, some countries with relatively weak competitiveness may have bright spots in individual sports, but the overall distribution and number of medals will be slightly inferior.

### 3. Olympic Medal Prediction Model

#### 3.1 Data Processing

##### 3.1.1 Data Cleaning

The data for this study mainly comes from the official dataset provided by the competition, which covers a wealth of information, including, but not limited to, the number of gold, silver, and bronze medals and the total number of medals won by each country in multiple Olympic Games.

Exploring the relationship between the number of athletes and medal rates is crucial. To facilitate the quantification of the effect of athletes on medal rates, we introduced an athlete strength variable:

$$Y = 0.75 * A_s + 0.25 * (1 - S) \quad (1)$$

$Y$  is the athlete's strength,  $A_s$  is the average athlete's strength, and  $S$  is the volatility of the athlete's strength [1]. The calculation of average athlete strength requires quantifying the effect of the class of the medal on the overall strength of the athlete, and we decided to introduce the  $\delta$  variable after surveying the literature [1]:

$$\delta = G_o * 3 + S_i * 2 + B_r * 1 \quad (2)$$

Where  $\delta$  is the country's total score.  $G_o$  is the number of gold medals,  $S_i$  is the number of silver medals, and  $B_r$  is the number of bronze medals. Olympic medals are categorized into gold, silver, and bronze medals. Different medals represent different levels of athletic performance and achievements of athletes in competitions, and there are apparent differences in their value. Gold medals symbolize the highest level of athletic performance, followed by silver and bronze medals. Therefore, this difference must be considered when quantifying athletes' overall strength. Thus, we obtained the formula for calculating the average strength of athletes:

$$A_s = \frac{\delta}{k} \quad (3)$$

where  $k$  is the total number of athletes from that country participating in the current Olympic Games. By assigning different weights to gold, silver, and bronze medals, i.e., multiplying the number of gold medals by 3, silver medals by 2, and bronze medals by 1, the medals can more scientifically and accurately reflect the average performance and comprehensive strength of the athletes in the competition.

**Volatility Calculation:** Volatility can be measured by calculating the standard deviation of an athlete's competition scores over time. Assume that the Score values of the athlete in  $n$  competitions are  $X_1, X_2 \dots X_{n-1}, X_n$ . The mean of this data set is first calculated, then the variance  $S^2$  is calculated, and the volatility (standard deviation)  $S$  is the square root of the variance:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (4)$$

Reasons for the higher weighting of comprehensive strength: comprehensive strength represents the athlete's overall past performance and achievements in competitions, which is an intuitive and central reflection of the athlete's ability. Giving it a higher weight of 0.75 means that when assessing an athlete's average strength, their past stable and achieved results and competitive level are the main factors to be considered [2]. Long-term stable high-level performance better reflects the fundamental strength level of the athlete, which is an essential basis for predicting the probability of obtaining future medals. The significance of introducing volatility: sports competition is uncertain, and athletes' status will fluctuate with time, training schedule, competition environment, and other factors. Fluctuation rate can measure the degree of ups and downs of athletes' performance, representing the stability of athletes' play. Giving it a weight of 0.25, although relatively small, should not be ignored. It supplements the dynamic change factors other than the comprehensive strength, assessing the average strength of the athletes more comprehensively and able to more accurately reflect the actual ability of the athletes at different points in time and different competition environments [2].

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

We uniformly normalized the  $A_s$  and  $S$  variables in equation (1) to map all features to the same scale, ensuring that each feature plays an appropriate role in data analysis and model training and avoiding bias due to the magnitude problem. In addition, since normalization of two data may cause the model to have a leftward kurtosis between 0 and 1, the features are insignificant between different data. So, the  $Y$  variable in equation (1) was normalized again, making a uniform comparison standard between various data. When analyzing the data related to athletes' sports strength in different countries, the normalized data can more intuitively show the relative advantages and disadvantages of the countries in other indicators, which helps to make an accurate comparative analysis. At the same time, overfitting is effectively avoided. With normalized data, the model is more robust to local fluctuations in the data during the training process, can better adapt to new data, and can maintain a more stable performance on different data sets.

### 3.1.2 Disposition of Abnormal Indexes

When counting the total number of medals won by each country, the data were rigorously screened to ensure accuracy and validity. After careful observation, it was found that the medal counts provided in the database contained some medals won by athletes in non-Olympic years, which were not relevant to the core content of the Olympic medal rate study and would interfere with the accuracy of the statistical results. Therefore, we excluded these erroneous data from the medal totals to ensure that the data accurately reflect the medals won in Olympic events. While processing the data in the file "summerOly\_programs," we noticed many data with "?" in the data. Considering that the model constructed in this study focuses on macro-level analysis, it is not necessary to be precise about the data details for each program. To make the data suitable for subsequent analysis and ensure its integrity, we have standardized it with "?" as "0". We treat the data with "?" as "0". That does not affect the overall analytical effect of the model, but it maintains consistency and coherence in data processing.

### 3.2 Construction of Olympic medal-winning prediction model based on stacked ensemble learning

In the prediction study of Olympic medal-winning, whether a country wins a medal can be abstracted as a binary classification problem. Given the limited data and the complexity of feature relationships, traditional single models often have limitations in dealing with such issues, making it difficult to achieve satisfactory prediction accuracy. Therefore, this study adopts an integrated learning approach that aims to improve the overall prediction performance by incorporating the advantages of multiple models. In this study, we select three classical models, namely Logistic Regression, Random Forest, and Support Vector Machine (SVM), as the base learner and logistic regression as the meta-learner, to construct a stacked model to reach an accurate prediction of whether a country will win a medal is accurately predicted.

#### 1) Construction of Logistic Regression Model

Logistic regression modeling is a widely used statistical learning method for binary classification problems. It maps the predictions of linear combinations to probability intervals using a sigmoid function, which gives the effect of each input feature on the predicted probability and then calculates the probability of winning a medal  $P$ . Its specific formula is:

$$P(y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (6)$$

where  $P(y = 1 | X)$  is the probability that sample  $X$  belong to class 1 (winning medals).  $X_1, X_2, \dots, X_n$  are the input features. These features cover a variety of factors related to the Olympic Games, such as the number of participating athletes and the average strength of the athletes.  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  are the regression coefficients.

#### 2) Construction of Random Forest Model

The prediction function of the random forest is realized by calculating the sum of the prediction function of each tree and its weight, and the formula is as follows:

$$f(X) = \sum_{i=1}^M \omega_i h_i(X) \tag{7}$$

where  $M$  is the number of trees in the forest.  $\omega_i$  is the weight assigned to each tree, usually set to equal weight.  $h_i(X)$  is the prediction function of the  $i$ -th tree.

After many experiments and parameter adjustments, we determined that the optimal parameter of random forest is {'max depth': None, 'min samples split': 10, 'n estimators': 200}. Currently, the model reaches the optimal score of 0.9652 on the verification set.

### 3) Construction of Support Vector Machine Model

Support Vector Machine (SVM) is a powerful classification algorithm that classifies whether a country has won a medal. Its decision function is denoted as:

$$f(X) = \omega^T X + b \tag{8}$$

where  $\omega$  is the weight vector that determines the direction of the hyperplane.  $X$  is the input sample.  $b$  is the bias term that shifts the hyperplane. Through tuning, it was determined that its optimal parameters were {'C': 100, 'gamma': 'scale'}, at which point the model achieved an accuracy of 0.948.

### 4) Construction of Stack Model

Finally, we constructed a stacked model to improve the prediction performance further. Logistic regression has the advantages of simple principle, strong interpretability, high computational efficiency, good adaptability to linear relationships, and strong generalization ability, so we chose logistic regression as the meta-learner. The specific process was as follows:

- **First Layer Predictions:**

$P_1(X), P_2(X), P_3(X)$  (Predictions from each base model on sample  $X$ )

- **Second Layer (Meta-Learner) Training:** The predictions from the base models are used as new features for the meta-learner:

$$P_{final}(X) = final\_estimator(P_1(X), P_2(X), P_3(X)) \tag{9}$$

- **Meta-Learner Decision Function:**

$$P_{final}(X) = \frac{1}{1 + \exp - (\alpha_0 + \alpha_1 P_1(X) + \alpha_2 P_2(X) + \alpha_3 P_3(X))} \tag{10}$$

where  $P_{final}(X)$  is the final predicted probability from the meta-learner.  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  are the regression coefficients of the meta-learner (Logistic Regression), representing the influence of each base model's output on the final prediction.

The stacked model constructed through the above steps fully integrates the advantages of different base models and can more effectively deal with the problems of limited data volume and complex feature relationships, providing a more accurate and reliable method for predicting Olympic medal winners.

### 3.3 Establishment of Optimized Multi-Model Medal Number Prediction System

In constructing the regression model, we also adopt the stacked model as an effective method to significantly improve prediction accuracy. The core of this model is a clever combination of ridge regression, XGBoost, and XGBoost meta-learner to predict the number of medals [3] accurately. As a linear regression model based on least squares, Ridge regression effectively avoids the overfitting problem by introducing L2 regularization [4]. Its objective function aims to minimize the following expression:

$$\min_{\beta} (\sum_{i=1}^n (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2) \tag{11}$$

where  $y_i$  is the actual medal count.  $X_i$  is the feature vector.  $\beta_j$  are the regression coefficients.  $\lambda$  is the regularization parameter.

By flexibly adjusting the value of  $\lambda$ , ridge regression can skillfully balance the fitting ability and complexity of the model, thus effectively preventing the occurrence of the overfitting phenomenon

and ensuring that the model can accurately reflect the characteristics of the data while avoiding the degradation of model performance caused by overfitting. XGBoost model is a powerful method based on the Gradient Boosting algorithm, which gradually improves the prediction results by constructing a series of weak classifiers (e.g., decision trees, etc.). Its mathematical expression is as follows:

$$f(X) = \sum_{i=1}^M \omega_i h_i(X) \tag{12}$$

Where  $M$  is the number of trees in the forest.  $\omega_i$  is the weight assigned to each tree, usually set to equal weight.  $h_i(X)$  is the prediction function of the  $i$ -th tree.

XGBoost subtly uses previous residuals to train new trees by continuously performing iterative optimization, thus gradually reducing the prediction error and improving the model's prediction accuracy. In model construction, we take the data of the two base models as the input features and integrate the outputs of the two base models of ridge regression and XGBoost through the XGBoost meta-learner [5]. The specific steps are as follows:

- **First-layer predictions:** Train base models (Ridge Regression and XGBoost) and obtain predictions  $P_1(X), P_2(X)$ .
- **Second Layer Training:** The predictions from the base models are used as new features for training the XBoost Meta-Learner to make the final prediction:

$$P_{final}(X) = final\_estimator(P_1(X), P_2(X), P_3(3)) \tag{13}$$

- **Meta-Learner Decision Function:** Using XBoost as the meta-learner, the decision function is:

$$P_{final}(X) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 P_1(X) + \alpha_2 P_2(X) + \alpha_3 P_3(X)))} \tag{14}$$

where  $P_{final}(X)$  is the final predicted probability from the meta-learner.  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  are the regression coefficients of the meta-learner (Logistic Regression), representing the influence of each base model's output on the final prediction.

To ensure that the performance of the models is optimal, we used techniques such as cross-validation and grid search to meticulously tune each model to find the best hyperparameters [6]. By comprehensively evaluating the performance of each model, we obtained the following detailed data:

Table 1 Model performance test

Model	MSE	RMSE	R2	MAE
Ridge Regression	10.265	3.204	0.976	1.471
XGBoost	57.5	7.583	0.7467	1.872
Ensemble Learning	272.458	16.506	0.593	4.64

It is clear from these evaluations that ridge regression performs well in predicting the total number of medals, with a coefficient of determination  $R^2$  almost close to 1, which explains the data variance very well and achieves nearly perfect prediction to a certain extent. XGBoost shows a relatively high prediction accuracy, and although there is still room for improvement, it already performs reasonably well. XGBoost shows relatively high accuracy in prediction, although there is still room for improvement. The mean square error is 57.5, the root mean square error is 7.583, the  $R^2$  is 0.747, and the average absolute error is 1.872, which shows that XGBoost can capture some features better when processing data, but there are some errors. After discovering this phenomenon, we considered stacking the results of the two models; however, surprisingly, the stacked model, although combining the advantages of ridge regression and XGBoost, did not perform better in this experiment. Its  $R^2$  is only 0.593.

### 3.4 The Solution of Model 1

#### Model performance testing and evaluation

We evaluated the models using the critical mean square error (MSE) metric. The mean square error is around 9 for ridge regression models and around 1 for 0 - 1 regression. The mean square error

reflects the average degree of error between the predicted value and the actual value, and a lower mean square error indicates that the model is more accurate in predicting the results.

At the same time, we also analyzed the stability of the prediction results. The predicted results cover gold medal and medal count dimensions, and the average fluctuation is around plus/minus 1.2. This small range of fluctuation indicates that the model is more accurate. This small fluctuation range further demonstrates the stability and reliability of the model predictions.

After rigorous testing and analysis of the above indicators, we concluded that there is a 97.49% probability that the prediction is correct. This high accuracy rate provided a solid guarantee of confidence for our subsequent predictions based on the model.

**Medal predictions and national performance scores for the 2028 Summer Olympics**

Based on the model we constructed and evaluated, we predicted the medal situation of the 2028 Summer Olympics in Los Angeles and gave the corresponding prediction intervals. The prediction intervals are presented, considering the data’s uncertainty and the model’s error range to provide a more comprehensive reference for the prediction results. In determining whether a country is progressive, we use a specific calculus:

$$z = (Go - Go_t) * 3 + (Si - Si_t + Br - Br_t) * 1.5 \tag{15}$$

where z is the difference in score between the two predicted Olympics and Go<sub>t</sub>, Si<sub>t</sub>, Br<sub>t</sub> are the number of gold, silver, and bronze medals won by the country in the previous Olympics. If z is more significant than zero, the government is considered to be making progress in medal winning; if z is less than zero, the country is likely to perform worse than it did in 2024.

The results of this forecast show that out of the 206 countries or regions counted, there is a high number of countries making progress, 135, with 47 maintaining the status quo and 24 regressing. Overall, most countries are trending positively in terms of winning the Olympic medal, which may be related to the increased emphasis and investment in sports.

**First medal prediction for countries that have yet to win a medal**

Table 2 The country that is expected to win the first medal

1	2	3	4
ANG	ANT	ARU	ASA
BEN	BHU	BIH	BIZ
BRU	BUR	CAF	CAM

Our model considers countries that have not yet won a medal. Through an in-depth analysis of historical data and the model’s accurate predictions, we selected countries that had never won a medal in 2024, 2016, or 2012. We compared them to the predicted medal winners. Sixty-nine countries achieved a breakthrough from 0 to 1 in the number of medals. Although this result contradicts our common sense, our model has a high accuracy rate (97.49%) regarding model prediction accuracy and error data, so we find this prediction result credible.

**Analysis of critical Olympic events in different countries under data analysis**

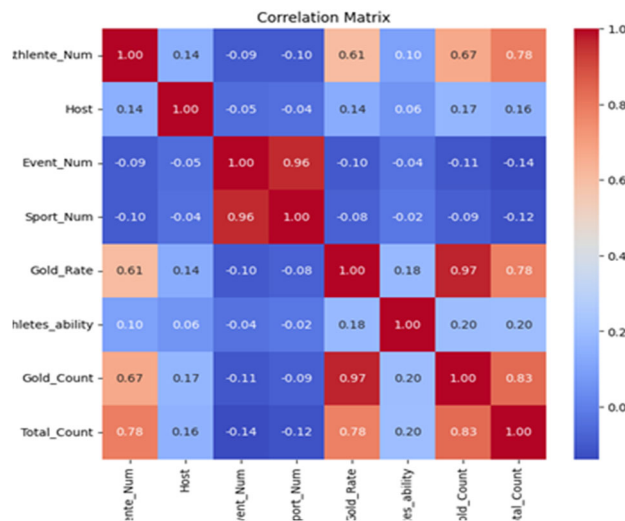


Fig. 3 Spearman Correlation Matrix Heatmap

In terms of rationality, the fact that a country has won the most medals in a particular sport reflects the country’s competitive strength in that sport. The country has invested more resources in talent cultivation, training systems, technology development, and other aspects of the program, and the results are remarkable. At the same time, these programs are often closely linked with the country’s sports culture and represent the national sports image, with a broad mass base and social influence. In terms of reasons, specific programs have a long history and deep cultural heritage in the country, which have formed a strong competitive advantage through years of development and inheritance; extensive mass participation provides a rich talent pool and a good sports atmosphere for the programs; and the attention and support of the government and sports management departments for the advantageous programs, including the formulation of development strategies and the provision of financial security, etc., have helped to improve the competitive level of the programs. The government and the sports administration pay attention to and support the advantageous programs, including the formulation of development strategies and the provision of financial guarantees, which also help to enhance the competitive level of the programs.

Table 3 Key items and number of medals for selected countries

Country	Sports	Number
USA	Swimming	1206
CHI	Swimming	120
RUS	Athletics	103
GBR	Athletics	393
FRA	Fencing	340
AHO	Sailing	1
DJI	Athletics	1

Looking at the two countries with the lowest number of medals, AHO and DJI, AHO won one medal in sailing, which is a symbol of national honor and can show the national image in the international sports arena; it is a driving force for sports development, which can inspire more people in China to participate in sailing, and even drive the progress of other sports. It can also bring economic impetus, attracting investment in related industries such as sailing boat manufacturing and event operation. DJI won a medal in track and field, which enhances national pride and cohesion, inspires national enthusiasm and support for sports, motivates more young people to join track and field, and cultivates sports talents for the country.

## 4. Influence of the flow of excellent coaches on the number of Olympic medals

### 4.1 Model Establishment

To study the “excellent coach” effect, we built a regression model based on time series. The year was used as the independent variable, the number of medals was used as the dependent variable, and a dummy variable was introduced to represent the period over which the excellent coach coached. We collected data from multiple years of Olympic medals for each selected country-sport combination. The general form of the model is:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \epsilon_t \quad (16)$$

$Y_t$  denotes the number of medals in year  $t$ ,  $T_t$  denotes the time-trend variable,  $D_t$  is a dummy variable that takes the value of 1 when a good coach is in charge and zero otherwise, and  $\epsilon_t$  is a random error term.

### 4.2 Model Test

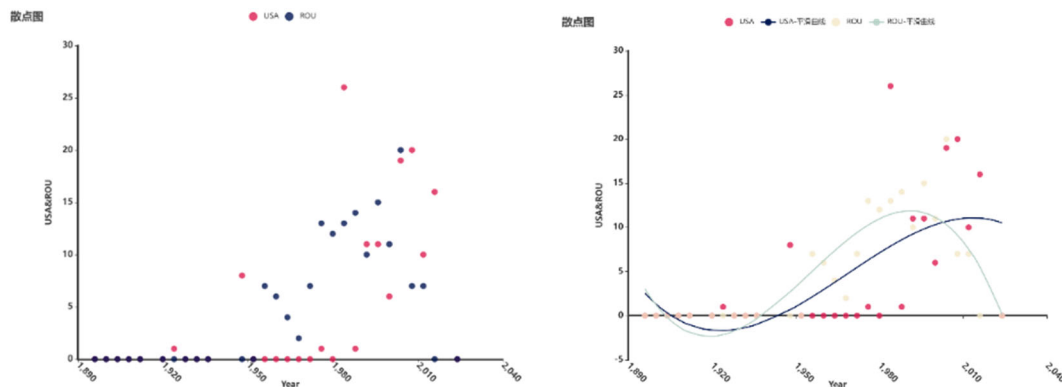


Fig. 4 Medals won by the United States and Romania in gymnastics

We chose to study gymnastics programs in the United States and Romania. For the excellent coach Marta Karolyi, by analyzing the data of the past six Olympic Games, we found some changes in the number of medals before and after the coaching of the excellent coach. For the Romanian gymnastics program, with Marta Karolyi’s coaching as the research point, through analyzing the relevant data and fitting the model, we found that the coaching of excellent coaches had a positive impact on the number of medals of the Romanian gymnastics team in the Olympic Games.

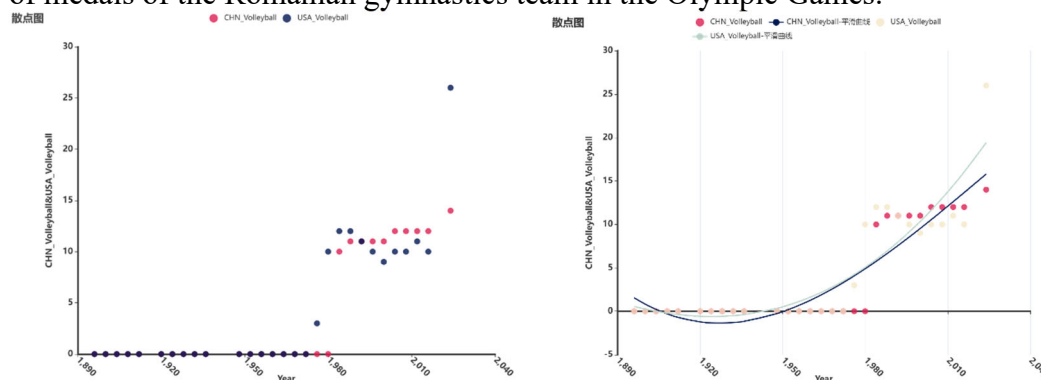


Fig. 5 Medals won by China and the United States in volleyball events

In the model test, we used historical data to fit the model and verified the significance of the dummy variable  $D_t$  through statistical tests (e.g., t-test, F-test, etc.). The results show that  $D_t$  significantly affects the number of medals, indicating that the coaching of excellent coaches affects the number of medals. The effect of “excellent coaching” varies across countries and programs but generally contributes positively to the total number of medals. This effect is pronounced when a coach is transferred to a particular country. Subsequently, due to the different comprehensive strengths of

the country, sports investment, and other factors, the effect will slowly stabilize. However, most stabilized figures still increase by more than 35% compared with the original ones.

The excellent coach effect is estimated to affect the quality of the improvement in 40%-60% (reacting to the ability) in the coach just transferred to a country when the impact factor is undeniable in the follow-up due to the country's different comprehensive strength and sports investment in the impact of slowly stabilizing, stabilized data compared to the original increase in most of the magnitude of 35% or more, and a small part of the year 2024 is still reflecting a robust upward trend. The modeling is not a complete successor to the original. Even if the trend shown in the subsequent model is different from the actual stabilized trend, we can still estimate the proportion of improvement from the data of the first few years of the effect combined with the estimated trend data of the excellent coaching effect under the condition of ignoring other risks (sudden change of the national inputs).

### 4.3 Results Analysis

A comparison of data found that the excellent coach effect does exist. In U.S. swimming, assuming that all other factors are stable, it is predicted that the quality of the U.S. national team can be improved by 50% by the presence of a great coach, but given that the U.S. has always been the world leader in this sport, the actual percentage improvement reflected in the number of medals may not reach 50%. Considering the limiting factors and influencing effects, it is predicted that the number of medals will stabilize at 77.3, a significant increase compared to the previous stable medal count of 67. In Chinese table tennis, the indicators quantified by great coaches have performed well on the available data. In assessing the impact of great coaches, after examining the data for the past six years, the average number of medals has stabilized at 11, and it is predicted that the quality of China's national team can be improved by 50% by the coaching of great coaches. Again, considering that China has always been the world leader in this sport and that table tennis is subdivided into several more minor sports, the limitations on the room for progress are relatively more minor than in the United States, and this is reflected in the number of medals that can be obtained, which is expected to be 15.3. The team will be able to win 15.3 medals. It is a significant improvement on the previous 11. In the British track and field program, the average of the medal figures over the past six years is 15. Still, Great Britain won 40 medals in the track and field events at the most recent Olympics, suggesting the phenomenon of great coaching in the British track and field program. Because of the more than 60% improvement, we predict that Great Britain's medal tally in track and field events will gradually stabilize and stabilize at 44.0 medals. It will consolidate his dominance in the track and field program.

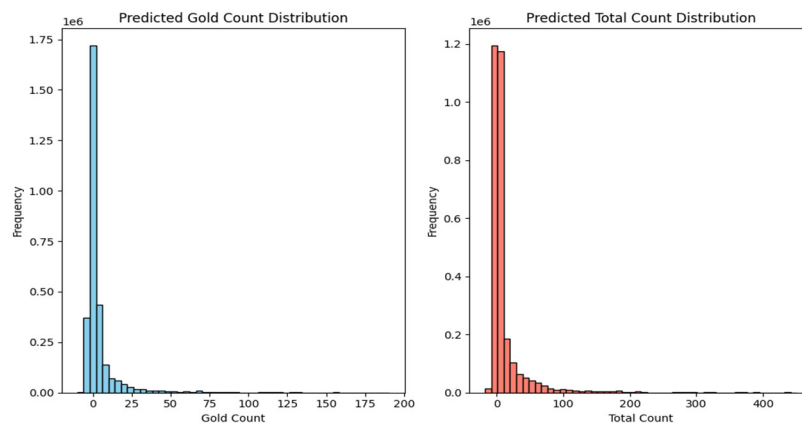


Fig. 6 Ridge regression model

We added the ridge regression model to the noisy data for prediction. We obtained a coefficient of determination of 0.9749 and an average mean square error of 9.1743, implying that the ridge regression model still shows excellent ability in data perturbation.

Table 4 Accuracy of ridge regression after Monte Carlo simulation

Model	Accuracy Mean	Standard Deviation
-------	---------------	--------------------

Random Forest	0.8714	0.0052
Logistic Regression	0.8933	0.0021
SVM	0.8885	0.0031

## 5. Model Evaluation and Further Discussion

The overall model of the paper adopts the stacked learning method, which effectively combines the advantages of various types of basic models and improves the model's generalization ability. The ridge regression used in the paper has a high degree of fit, strong predictive ability, and can accurately predict the number of medals. It effectively reduces the risk of fitting and maintains the stability of the model through L2 regularization. XGBoost has a powerful nonlinear modeling ability and can improve the overall prediction accuracy, the careful consideration of which can improve the model's generalization ability. Random Forest can enhance the robustness of the model, effectively reduce the risk of overfitting, provide a high degree of accuracy, and do not require features to be standardized and normalized, reducing the workload. Support vector machine has a powerful classification ability, handles high-dimensional data effectively, and controls overfitting.

However, ridge and logistic regression may not capture complex relationships perfectly, and XGBoost and stochastic also suffer from high computational overhead and relatively high interpretability.

Furthermore, we can also improve the stacked model by considering different meta-learners (LightGBM, CatBoost). We can compare the effects of different meta-learners to enhance the prediction's accuracy further. Finally, we can utilize model compression or quantization techniques, such as converting the XGBoost model into a smaller LightGBM, to shorten the computation time and increase the computation speed. We can also use distributed computing (Hadoop, Spark) to accelerate the model training during the computation.

## References

- [1] Bao Lei. Evaluation method and system of athletes' competitive state: CN201710377550.9[P]. CN108209902A.
- [2] Shiyan, Wang Ting. The theoretical source and practical review of athletes' mental health services [J]. *Sports Science*,2024,44(04): 16-23+51.202404002.
- [3] C. Schlembach, S. L. Schmidt, D. Schreyer, L. Wunderlich, Forecasting the Olympic medal distribution- A socioeconomic machine learning model, *Technological Forecasting and Social Change*, Volume 175, 2022, 121314, ISSN 0040-1625,
- [4] Hoerl, A. E., & Kennard, R. W. (1970). "Ridge regression: Biased estimation for non-orthogonal problems." *Technometrics*, 12(1), 55-67.
- [5] Dietterich, T. G. (2000). "Ensemble methods in machine learning." *International Workshop on Multiple Classifier Systems*, 1-15.
- [6] Dixon, C., & Wang, J. (2018). "Hyperparameter Optimization in Machine Learning: A Survey." *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, 271-280.
- [7] Tikhonov, A. N. (1977). "Regularization of incorrectly posed problems." *Soviet Mathematics Doklady*, 18, 1-4.
- [8] Peng, L., & Lee, K. (2002). "Logistic regression for machine learning." *Proceedings of the IEEE International Conference on Machine Learning*.