Assessment of Happiness Measurement: A critical review - investigation of existing happiness measurement tools

Junjie Wang[1],Yiyang Liu[2],Zhiyuan Zhang[3]

[1]University of Shanghai for Science and Technology,Shanghai,200093,China;

[2]Tianjin Chengjian University,Tianjin,300384,China;

[3]The University of Sydney,Sydney,2006,Australia;

Junjie Wang,Yiyang Liu, Zhiyuan Zhang are co-first authors

Abstract. With the development of Olympic events, countries hope to achieve good results in these events by investing their efforts effectively. Therefore, it has become particularly important to explore the relationships among Olympic events, athletes, and other relevant factors. We developed a prediction model for the total number of medals for each country. It was found that countries like the United States and China will get the largest number of medals, and countries like France will get more medals in the next Olympics. Then, we built a BP Neural Network Model to predict the countries that can win the first medal in the Olympics. We found that countries such as Andorra and Bolivia have a probability of 0.35 and 0.32, respectively, of winning their first medal. The indicators, such as the accuracy rate of the model, all reach above 0.94, showing good prediction results. Furthermore, by further considering the number and types of sports events and the advantageous events, and using methods such as the independent sample T-test for analysis, we found that six countries, including China, have their advantageous events, such as table tennis. At the same time, the status of the host country also has an important impact on the medal distribution. We still analyzed the impact of introducing excellent coaches on sports performance. We found that China, the United States, and Romania need to introduce coaches in football, archery, and track and field events, respectively, to effectively improve their performance. Through further analysis, we obtained some other insights on the number of Olympic medals. First, we used the Random Forest Algorithm to analyze the correlation between the number of medals of each country and sports events. Then, we analyzed the gender ratio of athletes. Finally, by calculating the medal proportion of each country in different events, we identified the events with the least competition. The insights can help the National Olympic Committees optimize their Olympic strategies, allocate resources reasonably, and make breakthroughs in events with less competition. Finally, we conducted a sensitivity analysis and evaluated the advantages and disadvantages of the model. After analysis and verification, our model is not sensitive to parameter changes, is relatively stable, and has certain practical significance.

Keywords: Olympic games; Multiple Linear Regression; BP Neural Network; Random Forest Algorithm.

1. Introduction

The Olympic Games, the world's largest and most influential comprehensive sports event held every four years, not only reflect a country's overall sports strength but also reveal various aspects such as its social economy and culture. As the 2028 Los Angeles Olympics approaches, predicting the medal performances of different countries can help formulate and optimize Olympic strategies for various nations in advance, thereby enhancing their performances in the next Olympics. This paper aims to explore the factors influencing medal performances by establishing a medal prediction model, providing a basis for the Olympic strategies of different countries.

This study is based on rich historical Olympic data and uses scientific analysis and modeling methods to achieve the following three research objectives: First, by constructing an accurate prediction model based on data from previous Olympic Games, estimate the number of gold medals and the total number of medals for each country at the 2028 Los Angeles Summer Olympics. Secondly, explores the mechanism by which excellent coaches affect athletes' medal-winning performance, quantitatively analyzes the differential contributions of this effect to different countries

and specific sports events, and reveals the key value of coaching factors in Olympic competition. Finally, we will delve into the novel insights contained in the prediction model for the total number of medals and explore the feasible paths that these findings can help National Olympic Committees optimize their future Olympic strategies, providing a theoretical basis and practical guidance for enhancing the Olympic competitiveness of various countries.

2. Empirical research

We begin by performing a visual analysis of the Olympic data. Figure 3 presents a stacked bar chart that illustrates the medal distribution of the top 20 countries with the highest total medal count at the 2024 Paris Olympics.

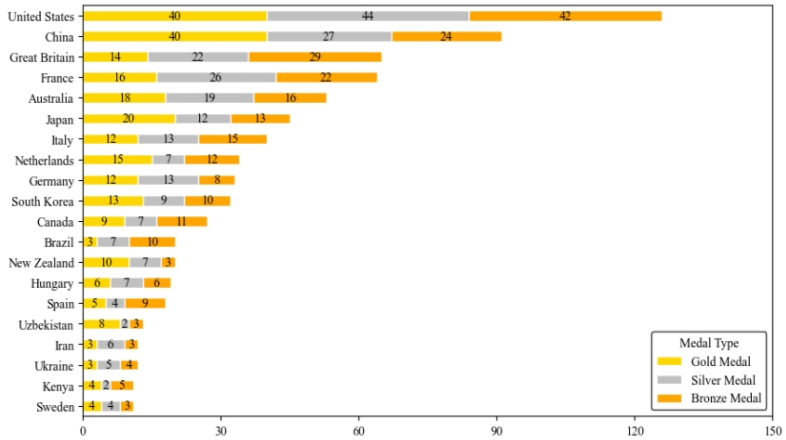


Fig. 1 The Top 20 Countries with the Highest Total Medal Count at the 2024 Paris Olympics

2.1 Prediction of Medal Count for the Los Angeles Summer Olympics

To predict the performance of each country at the 2028 Olympics, we have decided to develop two multiple linear regression models to describe the effect of various indicators from each country's participation in past Olympic Games on the total number of medals and gold medals. Since the Olympic program has undergone significant changes after 2000, with many events being added and removed, and the Olympics before 2000 were potentially more influenced by political and economic factors, such as political confrontations during the Cold War, we have decided to exclude data from the Olympics before 2004 when predicting the medal count for the 2028 Olympics. The input variables X1-X6 are defined as follows:

Change in Gold Medal Count (X1): This is a quantitative variable that represents the change in the number of gold medals a country has won compared to the previous Olympic Games. This variable reflects fluctuations in a country's performance over time, including changes in overall strength and preparation. It can be either positive (indicating an increase in gold medals) or negative (indicating a decrease). The value of this variable is a continuous integer.

Host Country Variable (X2): This is a qualitative variable used to indicate whether a country is the host of the current Olympic Games. Host countries typically invest more resources and enjoy a certain home-field advantage, making this an important part of the predictive model. The variable can be specifically classified into two categories: the host country and the non-host country.

Total Number of Events (X3): This is a quantitative variable that represents how many events a country participates in during the current Olympic Games. The value is a non-negative integer.

Total Number of Disciplines (X4): This is a quantitative variable indicating the total number of disciplines (major sports categories) a country participates in at the current Olympic Games. It is a non-negative integer.

Total Number of Sports (X5): This is a quantitative variable representing the number of sports a country participates in during the current Olympic Games. More sports generally imply more opportunities to win medals. The value is a non-negative integer.

Total Number of Elite Athletes (X6): This variable represents the total number of elite athletes from a country participating in the current Olympic Games. An elite athlete is defined as one who has participated in two consecutive Olympic Games and earned a medal. The number of elite athletes in a country directly correlates with its potential for winning medals. This variable is an integer.

However, since X2 (the host country variable) is a qualitative variable, it needs to be encoded:

$$X2 = \begin{cases} 0 & not\ host \\ 1 & host \end{cases} \quad (1)$$

Next, we can use data from the 2004 to 2024 Olympics for various countries to build models for both gold medal counts and total medal counts. A multiple linear regression model can be formulated as follows:

$$Y = \beta_0 + \sum_{i=1}^{6} \beta_i \cdot X_i + \varepsilon \quad (2)$$

Where the dependent variable Y represents the number of gold medals won by a country in a specific Olympic Games, substituting the variable data into the model, the final fitted regression model is as follows:

$$Y_{Gold} = -6.919 + 0.471 \cdot X_1 + 1.705 \cdot X_2 + 0.055 \cdot X_3 - 0.678 \cdot X_4 + 0.685 \cdot X_5 + 0.464 \cdot X_6 \quad (3)$$

Where Y_{Gold} represents the total number of medals won by a country in a specific Olympic Games, substituting the variable data into the model, the final fitted regression model is as follows:

$$Y_{Total} = -30.18 + 0.505 \cdot X_1 + 3.267 \cdot X_2 + 0.223 \cdot X_3 - 1.841 \cdot X_4 + 1.482 \cdot X_5 + 1.370 \cdot X_6 \quad (4)$$

The significance test results of the regression parameters for the independent variables in the total medal count model and the gold medal count model are summarized in Table 1 and Table 2, respectively.

Table 1 Significance Test Results of Parameters in the Gold Medal Count Model

Variable	t-value	P-value
X1	9.554	0.000
X2	0.858	0.391
X3	1.585	0.113
X4	-3.898	0.000
X5	2.528	0.012
X6	46.992	0.000

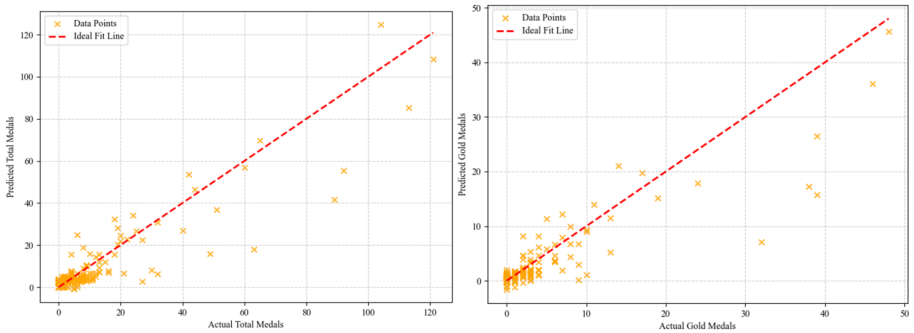
For the gold medal prediction model, among the six independent variables, the P-values of X1, X4, X5, and X6 are all less than 0.1, indicating that the regression coefficients of these four variables are significant. This demonstrates that these variables have a notable impact on predicting the number of gold medals. However, the P-values of X2 and X3 are 0.391 and 0.113, respectively, both greater than 0.1, suggesting that their regression coefficients are not significant.

Table 2 Significance Test Results of Parameters in the Total Medal Count Model

Variable	t-value	P-value
X1	3.727	0.000
X2	1.149	0.251
X3	1.910	0.056
X4	-4.123	0.000
X5	2.541	0.011
X6	51.730	0.000

For the total medal count prediction model, among the six independent variables, the P-values of X1, X3, X4, X5, and X6 are all less than 0.1, indicating that these five variables have a significant impact on predicting Olympic medal counts. However, the P-value of X2 is greater than 0.1, suggesting its influence is relatively weaker.

Additionally, scatter plots were created for both models to illustrate the relationship between actual and predicted medal counts. By adding a red prediction line, we can assess the accuracy of the models. As shown in Figure 2(a) and Figure 2(b), the scatter points are relatively close to the prediction line, demonstrating that both models exhibit good predictive performance.



(a)Total Medals (b)Gold Medals

Fig. 2 Scatter Plots of Actual and Predicted Medal Counts

We evaluate the effectiveness of the model using two indexes: R^2 and MSE. The results, as shown in Table 3, indicate that the R^2 is close to 1, suggesting an excellent fit of the model.

Table 3 Evaluation of Models

Model	R^2	MSE
Total Medals	0.820	54.418
Gold Medals	0.774	10.516

Based on the multiple linear regression model established above, we predicted the number of gold medals and total medals that each country would earn in the 2028 Los Angeles Olympic Games. These predictions were compared with the medal outcomes of the 2024 Paris Olympics to assess whether the countries have progressed or regressed. The top 20 countries in the ranking are presented in Table 4.

Table 4 Top 20 Countries and Their Predicted Medal Counts

2024 Rank	NOC	2024 Total	2028 Total	2028 Gold	2028 Rank	Status	
1	USA	126	143.3493	49.2942	1	No Change	-
5	FRA	64	79.54894	28.56126	2	Up	↑
2	CHN	91	72.04786	24.81707	3	Down	↓
7	GBR	65	69.73727	21.03498	4	Up	↑
4	AUS	53	44.13923	15.0486	5	Down	↓
9	ITA	40	41.90397	14.59005	6	Up	↑
6	NED	34	39.30872	14.609	7	Down	↓
3	JPN	45	35.98794	9.884542	8	Down	↓
15	ESP	18	32.31273	11.336	9	Up	↑
8	KOR	32	26.61714	10.90272	10	Down	↓
11	NZL	20	24.59677	9.017993	11	No Change	-
10	GER	33	24.09167	8.54681	12	Down	↓
12	CAN	27	22.7215	8.081945	13	Down	↓
14	HUN	19	20.34113	6.674715	14	No Change	-
18	NOR	8	18.97095	6.209851	15	Up	↑
20	BRA	20	18.32075	4.789984	16	Up	↑
16	SWE	11	9.884813	3.426981	17	Down	↓
13	UZB	13	9.164836	4.381982	18	Down	↓
17	KEN	11	6.639365	2.026069	19	Down	↓
19	IRL	7	4.9092	2.038705	20	Down	↓

Furthermore, in this study, we used a BP neural network to classify and predict the medal counts of these countries. The inputs for the model include the medal count from the previous year, the number of participants in the last three Olympic Games, the total years of participation for the athletes in the current Games, and the average years of participation of the athletes. The output is the predicted number of medals. The BP algorithm typically involves the following steps:

Initialize weights and biases: The weights and biases are initialized, typically with small values.

Forward propagation: For each layer, the output is calculated as $a^l = \sigma(z^l)$, where $z^l = W^l \cdot a^{l-1} + b^l$. Here, σ is the activation function, W^l and b^l are the weights and biases of the current layer, with a^{l-1} being the output of the previous layer.

Calculate output error: The error is computed based on the difference between the output layer's prediction and the true labels.

Back-propagate error: The output error G for each layer is calculated, often involving the gradient of the loss function concerning to the output at that layer.

Compute gradients: Using the chain rule, the gradients of the loss function concerning to each weight and bias are calculated.

Update weights and biases: The weights and biases are updated using gradient descent or other optimization algorithms.

Repeat the above steps: The process is repeated until the desired performance is achieved or the predefined number of iterations is reached. To assess the accuracy of the prediction model, this paper evaluates its performance using three metrics: RMSE, MAE ,and R^2. We set the following parameters for the BP neural network: activation function: ReLU, number of hidden layer neurons: [2, 2], number of iterations: 1`000, error threshold: 10-6, and 90% of the data used for training, with the sample shuffling method employed for training. To select the optimal number of neurons, we performed a sensitivity analysis on the prediction model by testing neuron counts from 3 to 10, resulting in the following LOSS values for different numbers of neurons, as shown in Figure 3 below:

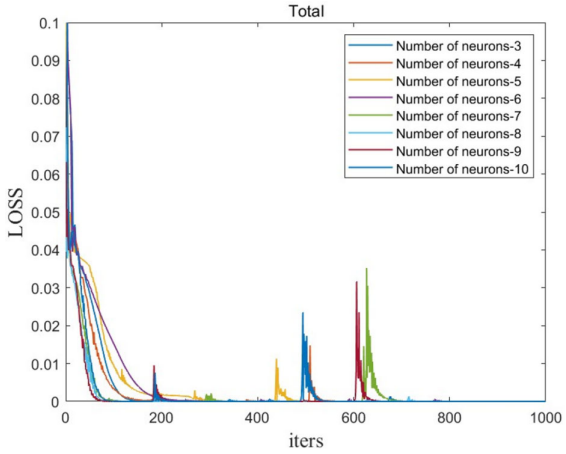


Fig. 3 LOSS Values for Different Numbers of Neurons

From this, it can be observed that when the number of neurons is 9, the LOSS value converges first. Based on this result, we can adjust the neuron parameters of the total medal prediction model. We obtained the classification evaluation metrics, as shown in Table 5.

Table 5 Classification Evaluation Metrics of the Prediction Model

Accuracy	Precision	Recall	F1 Score
0.96	0.94	0.98	0.96

From the table, we can see that the classification accuracy is performing well. We have predicted the probability of countries that have not won medals in previous Olympic Games winning their first medal at the 2028 Summer Olympics in Los Angeles. The six countries with the highest and lowest probabilities are shown in the following figure.

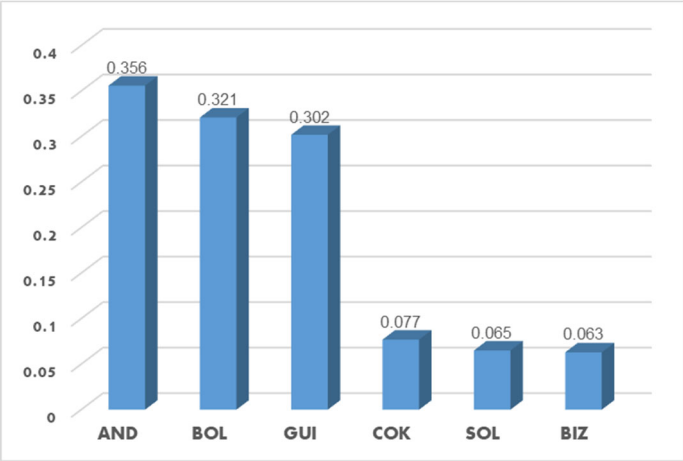


Fig. 4 Probability of Winning the First Medal for Some Non-Medalist Countries at the 2028 Olympics

We assume that the more elite athletes participate in a sport, the higher the likelihood of winning a medal in that sport. Thus, for the past three Olympic Games, if in any two of these editions, the number of elite athletes in a particular sport accounts for 30% or more of the total athletes participating in that sport, we classify it as the country's specialty sport.

We have identified the specialty sports for fourteen countries, and the results are shown in Table 6.

Table 6 Fourteen Countries and Their Specialty Sports

NOC	Specialty Sports	NOC	Specialty Sports
USA	Basketball, Swimming	GER	Table Tennis
CHN	Table Tennis, Diving	GBR	Triathlon
JPN	Judo, Table Tennis	SRB	Water Polo
FRA	Judo	CRO	Rowing
GRE	Shooting	SMR	Shooting
AUS	Swimming	THA	Taekwondo
CAN	Football	KOR	Archery

In order to explore how the selection of events by host countries influences medal counts, we selected countries that have hosted the Olympics multiple times. The United States hosted the Games in 1904, 1932, 1984, and 1996, while France hosted in 1900, 1924, and 2024. Therefore, we take the United States and France as examples for conducting an Independent Samples T-test.

Table 7 Mean Values of the Independent Sample T-Test

Country		Means
United State	Host	154.0
	Not Host	85.92
France	Host	68.333
	Not Host	22.667

As shown in Table 7, there is a significant difference in the mean total medal count when the United States and France serve as host countries compared to when they do not. The difference magnitude for the United States is 2.141, and for France, it is 3.167 (where 0.20, 0.50, and 0.80 correspond to small, medium, and large effect sizes, respectively). These differences indicate a very large effect size. Additionally, the mean total medal count is significantly higher when these countries act as host nations.

2.2 Modeling the "Great Coach" Effect

Since the "Great Coach" effect has an impact on Olympic sports performance, we select three countries and establish three linear regression models as follows. By controlling variables, we can observe the impact of the introduction of coaches on sports performance. The equation is as follows:

$$y = a \cdot t + b \cdot \beta + \varepsilon \tag{5}$$

Where t represents time, and β is a binary variable indicating whether the coach is in place (0 or 1). When there is no coach, $b \cdot \beta = 0$, and when there is a coach, $b \cdot \beta = b$.

To estimate the contribution of the "Great Coach" effect on the medal count, we consider that having a coach will result in a greater fluctuation in the number of medals. Thus, we introduce a curve after adding the excellent coach, calculate its standard deviation, and use it as a threshold to determine whether a coach is present. To identify the sports projects where countries should invest in "great" coaches, and to better highlight the influence of excellent coaches, we set different score weights for gold, silver, and bronze medals as follows:

$$S_{medal} = 5 * N_{gold\_medal} + 3 * N_{silver\_medal} + N_{bronze\_medal} \tag{6}$$

Where Smedal represents the medal score, Ngold_medal represents the number of gold medals, Nsilver_medal represents the number of silver medals, and Nbronze_medal represents the number of bronze medals.

When certain countries' specific sports events show extreme fluctuations in annual medal scores (either too low or too high), the introduction of a "great" coach may not have a noticeable effect on improving the medal count. Therefore, the analysis focuses on sports events that maintain an average annual medal count in the 1-2 digit range (ensuring that the medal count is not too low) and have a score between 5 and 10. These events are then subject to standard deviation analysis to identify those that meet the criteria for investing in a "great" coach. The host country's status can significantly inflate the medal count, which may distort the analysis. Hence, the analysis avoids the periods when countries are hosting the Games.

It is known that Lang Ping coached from 1995 to 1999 and from 2013 to 2021. Therefore, we calculated the medal scores during this period to explore the relationship between the great coach effect and the number of medals. After removing data related to host countries, we obtained the linear regression model fitting results for the Chinese women's volleyball team, as shown in Figure 10.

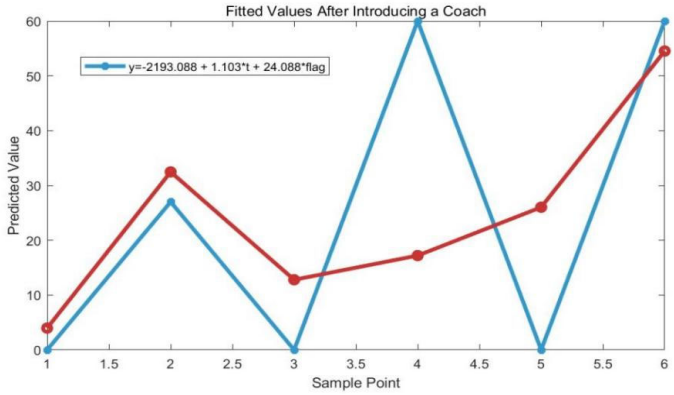


Fig. 5 Fitted Values After Introducing a Coach(China)

Using the standard deviation of 27.2685 for the Chinese women's volleyball team as a threshold, we identified the sports in which China needs to introduce coaches, including Football, Archery, Badminton, and a total of 10 sports.

The data shows that gymnastics coach Bela Karolyi successfully coached both the Romanian and the U.S. women's gymnastics teams. To explore the relationship between the effect of great coaches and medal scores, we calculated the U.S. medal scores around the time Bela Karolyi coached, while removing data from the host country periods. Based on the model analysis, we can conclude that the model's R^2 is 0.883, which significantly rejects the null hypothesis of a zero overall regression

coefficient ($P < 0.05$). This indicates a significant linear relationship. The coefficient of β is 44.086, meaning that a great coach can enhance the medal score. When β changes from 1 to 0, the medal weight score for gymnastics in the U.S. would decrease by 44.086 points. If converted entirely to gold medals, this would result in a loss of approximately 8-9 gold medals. The fitting result of the U.S. gymnastics linear regression (OLS) is shown in Figure 6:

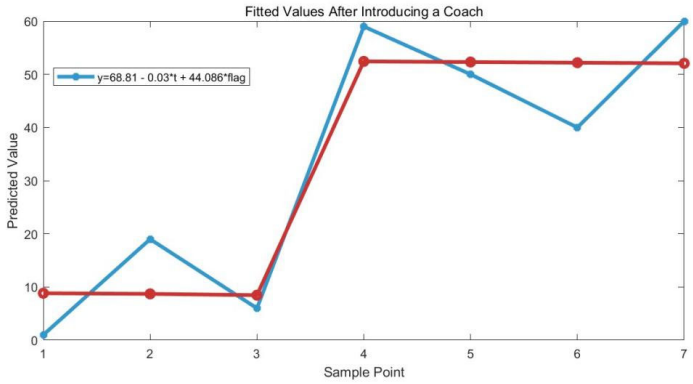


Fig. 6 Fitted Values After Introducing a Coach (U.S.)

With a standard deviation of 23.4445 for U.S. gymnastics as the threshold, it was found that the U.S. needs to introduce coaches in 18 events, including Archery, Judo, Boxing, and others. According to the data, gymnastics coach Bela Karolyi successfully coached both the Romanian and U.S. women's gymnastics teams. Therefore, by calculating the medal score data for Romania during this period and excluding host country-related data, the linear regression (OLS) fitting results for Romanian gymnastics are shown in Figure 7 below:

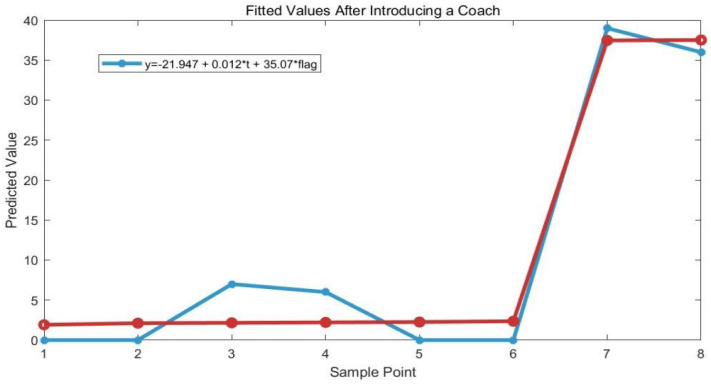


Fig. 7 Fitted Values After Introducing a Coach (Romanian)

When β changes from 1 to 0, the medal score for Romania in gymnastics will decrease by 35.07 points. If converted entirely to gold medals, this would result in a loss of approximately 7 gold medals. Using the standard deviation of 16.6218 for Romanian gymnastics as the threshold, we identify 8 sports in which Romania should introduce coaches, including Athletics, Rowing, and Shooting.

2.3 Other insights on the number of Olympic medals

2.3.1 Random Forest Importance Evaluation

Feature importance evaluation is a key characteristic of random forests, which helps us understand the contribution of each feature to the model's results. The visualization of the random forest feature importance evaluation for the medal counts and events of China, United States is shown in Figure 8.

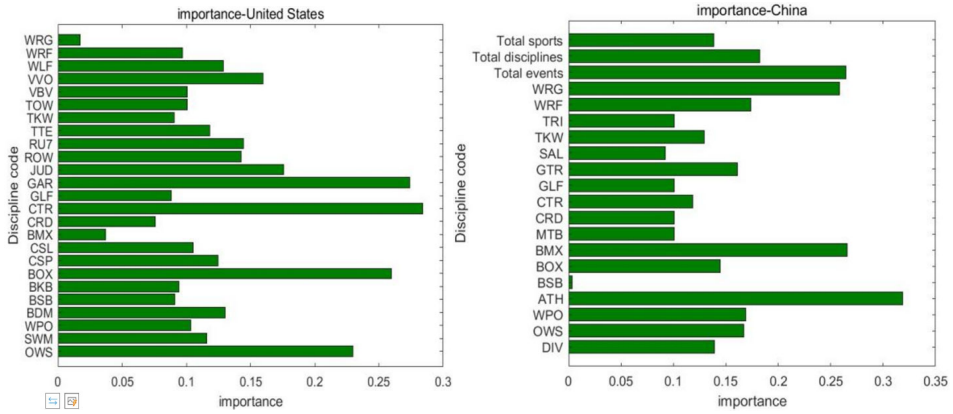


Fig. 8 Random Forest Feature Importance Evaluation of Medal Counts and Events for Four Countries.

The event types ATH, BMX, and OWS contribute significantly to China's medal count. The GLF event type has the smallest contribution to the United States medal count, while the GAR event type has the largest contribution.

2.3.2 The Gender Ratio of Athletes

Through analysis, we found that historically, the gender ratio of athletes has been close to 7:3. In the early years of the Games, they were predominantly male participants. It wasn’t until 1932 that the balance started to shift, and only by 2012 did this ratio approach 1:1.

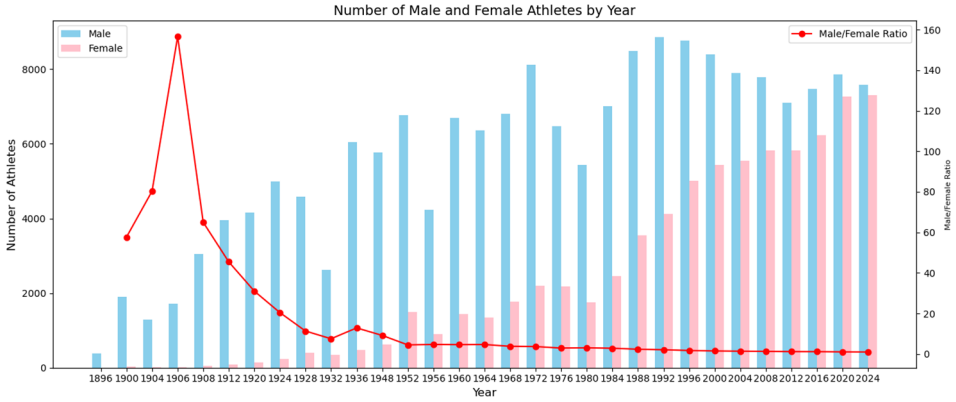


Fig. 9 Number of Male and Female Athletes by Year

2.3.3 Non-Dominated Events

To identify non-dominated events, we followed these steps: Calculate the medal share for each country in every event. Rank the events by the country with the highest medal share, identifying the highest share for each event. The less dominated an event is, the lower the highest share value. Therefore, we can sort the events by this value in ascending order to display the non-dominated events, as shown in Figure 10.

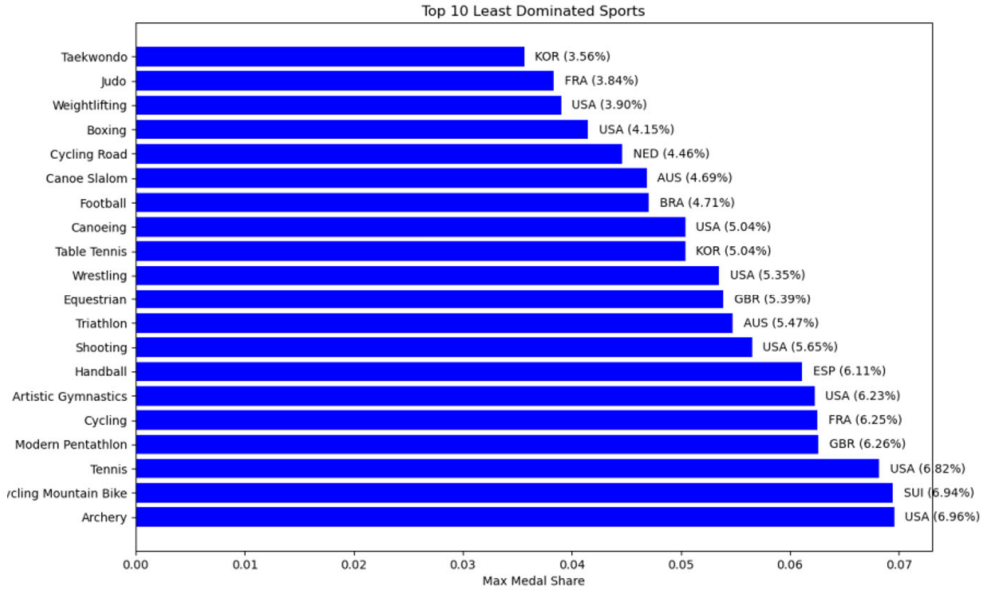
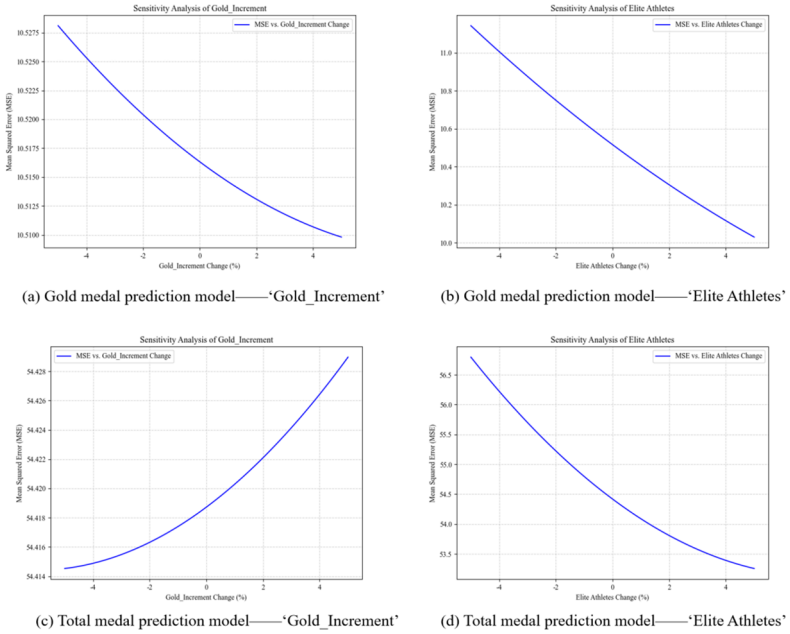


Fig. 10 Top 10 Least Dominated Sports

3. Sensitivity tests

We conducted sensitivity tests on the gold medal prediction model and the total medal prediction model. We chose the independent variables X1 and X6 for the sensitivity analysis. We evaluated the sensitivity of each feature to the model outcome by varying each independent variable within a ±5% range and observing the changes in MSE. Upon examining the chart, we can see that the MSE of the gold medal prediction model remains stable within a small range of 10.5, and the MSE of the total medal prediction model remains stable within a small range of 54.4. This indicates that the model is not sensitive to parameter changes and is relatively stable.



(a) Gold medal prediction model——'Gold_Increment' (b) Gold medal prediction model——'Elite Athletes'

(c) Total medal prediction model——'Gold_Increment' (d) Total medal prediction model——'Elite Athletes'

Fig. 11 Sensitivity Analysis

4. Conclusion

With the development of Olympic events, countries hope to achieve good results in these events by investing their efforts effectively. We developed a prediction model for the total number of medals for each country. It was found that countries like the United States and China will get the largest

number of medals, and countries like France will get more medals in the next Olympics. Then, we built a BP Neural Network Model to predict the countries that can win the first medal in the Olympics. We found that countries such as Andorra and Bolivia have a probability of 0.35 and 0.32, respectively, of winning their first medal. Furthermore, by further considering the number and types of sports events and the advantageous events, we found that six countries, including China, have advantageous events. At the same time, the status of the host country also has an important impact on the medal distribution. We still analyzed the impact of introducing excellent coaches on sports performance. We found that China, the United States, and Romania need to introduce coaches in football, archery, and track and field events, respectively, to effectively improve their performance. Through further analysis, we obtained some other insights on the number of Olympic medals. First, we used the Random Forest Algorithm to analyze the correlation between the number of medals of each country and sports events. Then, we analyzed the gender ratio of athletes. Finally, by calculating the medal proportion of each country in different events, we identified the events with the least competition. The insights can help the National Olympic Committees optimize their Olympic strategies, allocate resources reasonably, and make breakthroughs in events with less competition.

References

[1] Olympics.com, https://olympics.com/en/paris-2024/medals

[2] Olympics.com Biography, Lang Ping, https://olympics.com/en/athletes/ping-lang

[3] USA Gymnastics Hall of Fame,https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/Information on:

[4] Zhang Q. On relationships between Chatterjee's and Spearman's correlation coefficients[J].Communications in Statistics - Theory and Methods,2025,54(1):259-279.

[5] V A C. Understanding the independent samples t-test in nursing research[J].British Journal of Nursing,2025,34(1):56-62.

[6] Shi H, Zhang D, Zhang Y. Can Olympic medals be predicted? ——Based on the perspective of interpretable machine learning [J]. Journal of Shanghai Sport University, 2024, 48 (04): 26-36.

[7] Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests[J].Research Papers, 2017, 8(6):1831-45.

[8] Cheng H, Lv J, Yuan T. Predicting China's Track and Field Performance at the Tokyo Olympics from the 2018 World Top 20 Athletics Rankings [J]. Sports Science and Technology Literature Bulletin, 2020, 28 (04): 4-8.