Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2510034

The Mystery of Olympic Medal Counts: Prediction and Analysis Based on the GBRT Model

Summary

This study predicts the 2028 Olympics medals and analyzes influencing factors. It combines statistical, machine learning, and cluster analysis to build a GBRT model. Using 1896 - 2024 data, the model forecasts 2028 gold and total medals with intervals, and identifies potential medal - debutant and performance - change countries.

For medal - count prediction, variables like gender ratio and host status are considered. After data pre - processing and model training, predictions and country - specific trends are obtained.

Regarding influencing variables, calculating win rates and correlation analysis show event - country medal - winning relationships. Comparing host/non - host win rates reveals the home - advantage effect's role in medal - winning, which varies by year.

Testing the "great coach effect" with Lang Ping and Bela Karolyi as examples, the event - study method shows coaching changes boost medal counts in relevant teams. Thus, China, the US, and Romania are advised to hire good coaches for high - impact events.

Finally, we additionally discovered that the GBRT model we constructed indicated that the gender ratio has a stable influence on the number of medals. Therefore, we proposed unique insights such as optimizing the gender ratio could enhance a country's sports competitiveness. This series of additional discoveries might provide guidance for the Olympic Committees of various countries in athlete-related decisions.

Keywords: Olympic medal prediction; GBRT model; influencing factors; gender ratio; "great coach" effect

Contents

1 Introduction

1.1 Problem Background

During the 2024 Paris Olympics, sports enthusiasts closely followed the medal table, with China and the United States tying for the most gold medals. The number of medals is not only important for the top-ranked countries but also for those achieving milestones, such as Albania, Cape Verde, the Dominican Republic, and Saint Lucia, which won their first Olympic medals at this event.

The number of medals a country wins at the Olympics is influenced by multiple factors - it is not solely determined by the athletic prowess of its athletes, but also by the country's economic strength, population size, political structure, and more. [1] On a deeper level, special circumstances such as the "excellent coach effect" and whether the country is the host of the Olympics also need to be considered.

The academic community has always been interested in the medal counts of various countries at each Olympics, and there have been quite a few studies predicting the final medal counts.Some scholars, by analyzing the data of previous Olympic Games and establishing regression models, have revealed that a country's Olympic performance is closely linked to its various fundamental attributes and characteristics.

Studying the influencing factors of the Olympic medal table not only comprehensively summarizes the experience and achievements of the Olympics but also has significant importance for predicting the medal table of the next Olympics. Therefore, with the rich data provided by the topic, we can introduce relevant sociological theories and big data methods to predict the medal-winning situation of various countries at the 2028 Summer Olympics in Los Angeles, USA. This emphasizes the need to build complex models to capture the mysteries behind the Olympic medal counts.

1.2 Clarifications and Restatements

Our task centers on developing models to predict the number of Olympic medals for each country (particularly gold medals and total medals), aiming to identify the key factors related to Olympic medal counts and assess their impact on competition outcomes. The model should estimate the uncertainty and accuracy of predictions and measure its performance.We will carry out the following steps:

1. Create a model depicting the performance of Olympic participants from various countries from 1896 to 2024.

2. Use this model to predict the medal table for the 2028 Los Angeles Olympics, including prediction intervals. Identify countries that may improve or perform worse compared to 2024, and estimate the number of countries that have not yet won medals but may win their first medal in the next Olympics, along with the associated odds.

3. Build a hypothesis testing framework to explore the relationship between sports and medal counts, determine the important sports for different countries, and analyze how the host country's selection of sports affects the results.

4. Test a "great coach" effect. Look for evidence of this effect in the data, estimate its contribution to medal counts, and select three countries to recommend sports where investing in "great coach" might be beneficial and predict the impact.

5. Synthesize our research findings into actionable information, explaining how our insights can provide reference for national Olympic committees and assist them in selecting or training suitable athletes.

1.3 Our work

We combined research methods such as statistical analysis, machine learning techniques and cluster analysis to build multiple models. Through detailed analysis of the competition data, we achieved prediction, analysis and verification. We considered multiple aspects of details to ensure the reliability and applicability of the models in real scenarios.

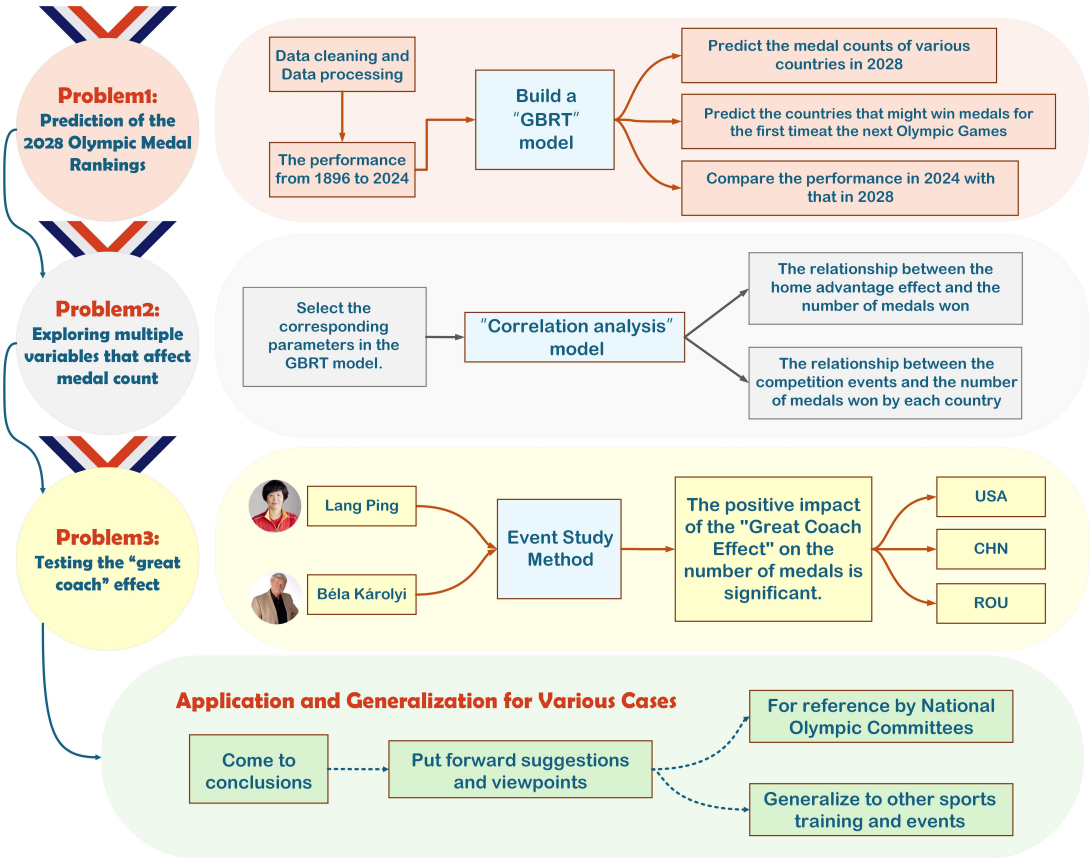

Figure 1: Framework of Our Work

2 Preparation for Modeling

2.1 Model Assumptions

Some fundamental assumptions are listed below.

- Assumption 1:
  The number of events a country participates in can affect its total medal count. The more events a country participates in, the greater the possibility of winning more medals or gold medals.
- Assumption 2:
  Basic factors such as the gender ratio, the total number of historical medals, and the number of events participated in will affect the competition results.
- Assumption 3:
  The home field advantage is a significant factor in the data, and the effect of a "great coach" will have a notable impact on the outcome of the game.

2.2 Notations

Some primary notations are listed below.

Table 1: Notations Table

Symbols	Description
γ_m	the mth regression tree
ϵ_{it}	error term due to influences other than the factors
β_0	the baseline level of contribution degree
β_1	the average impact of coaching changes on the number of medals
β_2	the influence of time trends on the number of medals
$h_m(X)$	the weight of each tree
Coach_{it}	define the dummy variable for the target coach
Time_{it}	time trend variable

2.3 Data Cleaning

- **summerOly_athletes.csv:**
  No missing values, but 1466 duplicate records (42 of which were medal data) were found and deleted, for a total of 251,099 data after cleaning.
- **summerOly_hosts.csv:**
  Handles missing values for country codes for years in which the Olympics were cancelled (e.g., 1916, 1940, 1944, etc.). 2020 Tokyo, Japan Olympics implementation mapped and marked for extension.

- **summerOly_medal_counts.csv:**
  No missing or duplicate values. Performs normalisation for "NOC" columns, using a mapping table to handle unstandardised labels, complemented by country-specific normalisation mappings. Converts the five data types "Year", "Gold", "Silver", "Bronze", "Total" to integers.

- **summerOly_programs.csv:**
  It was found that there are 2 missing values in the "Discipline" column. Fill in the missing values of "Discipline" with the corresponding values in the "Sport" column to ensure data integrity. For the abnormal values (such as "?") in the year column: remove the "?", and only keep the numerical part; if it cannot be parsed as a number, mark it as a missing value. After cleaning, the value distribution of the year column is reasonable and is uniformly of numerical type.There are a total of 74 records and 35 columns (including metadata and the year column).

2.4 The processing of special case data

Remove the Soviet Union (URS) (as the predecessor of multiple countries) and Russia (URS) (Russia did not participate in the 2024 Olympics due to war reasons and its participation in 2028 is uncertain, so it is deleted);

Remove Yugoslavia (YUG) (as the predecessor of multiple countries),the Commonwealth of Independent States (EUN) and Bohemia (BOH);

Map East Germany (GDR) and West Germany to Germany (GER);

Remove Czechoslovakia (TCH) as it has split into the Czech Republic and Slovakia.

3 Problem1:Prediction of the 2028 Olympic Medal Table

3.1 The building block: GBRT Model

Gradient Boosting Regression Trees (GBRT) is an integrated learning algorithm that builds multiple decision trees to optimize the model step by step. GBRT is able to effectively capture non-linear relationships and interactions between features in the data, and is well suited to handle complex multivariate problems such as Olympic medal prediction. Unlike traditional models, GBRT does not require assumptions about data distribution and provides feature importance assessment to help optimize the model and identify key factors, making it ideal for predicting the number of Olympic medals.

The principle of the GBRT model is to minimize the loss function L(y,$\hat{y}$), where y is the actual value and $\hat{y}$ is the predicted value. For regression tasks, a commonly used loss function is the mean square error (MSE):

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

- **Step1:Determine the independent variables and the target of the model.**
The dependent variables are the number of gold medals and the total number of medals; the independent variables include the gender ratio (Gender_Ratio), whether the country is the host (Is_Host), the number of events participated in (Total_Events), the total number of Olympic events the total number of participants (Total_Participants), the total number of historical medals (Historical_Medals) and the total number of Olympic events.We aim to minimize the prediction error and find the optimal model parameters.

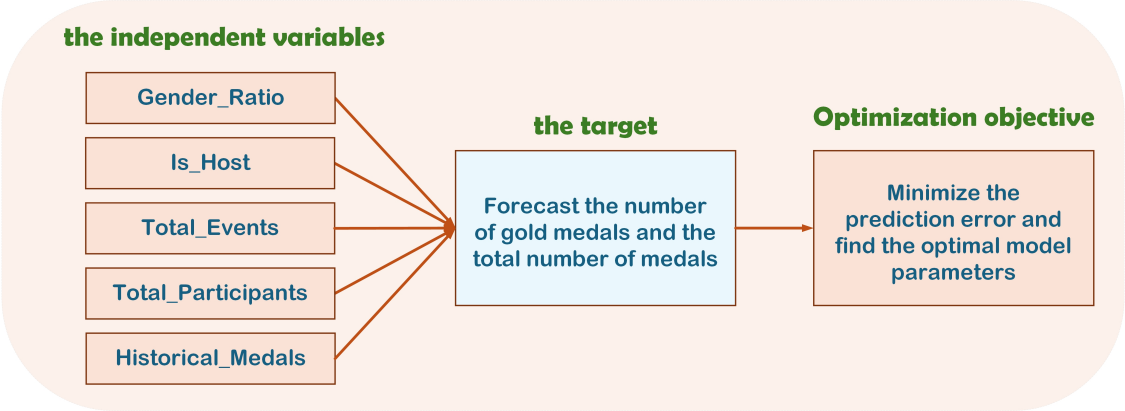

Figure 2: Model logical relationships

- **Step2:Train the model.**
Standardize the data to ensure that all features are on the same scale. Divide the data into a training set (years < 2012) and a test set (years ≥ 2012).

The prediction formula of the regression model can be written as:

$$\hat{Y} = f(X) = \sum_{m=1}^{M} \gamma_m h_m(X) \tag{2}$$

- **Step3:Obtain the model results and conduct verification.**
(1)Calculation accuracy index
Mean Squared Error (MSE):

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

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

(2)Calculate predictive uncertainty
The GBRT model has a high degree of fit when predicting the number of Olympic medals and can effectively capture data patterns. Through cross-validation and confidence intervals, we can quantify the stability and range of the predictions. The range of the confidence interval can help decision-makers assess the reliability of the prediction results, thereby optimizing Olympic strategies.

Cross-validation:

- Use k-fold cross-validation to evaluate model performance.
- Train and test the model each time, and record the range of prediction errors under different partitioning methods.

Standard deviation estimation:

- By repeatedly training the model multiple times, calculate the standard deviation of the prediction results of different models.

Confidence interval:

- Assuming the error distribution is normal, the confidence interval can be expressed as:

$$CI = \hat{y} \pm z \cdot \sigma \tag{6}$$

The residual plots of the number of gold medals and the total number of medals:

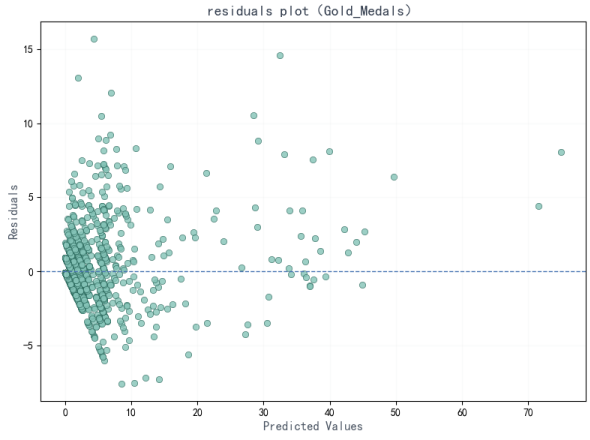

Figure 3: Gold Medals Residuals Plot

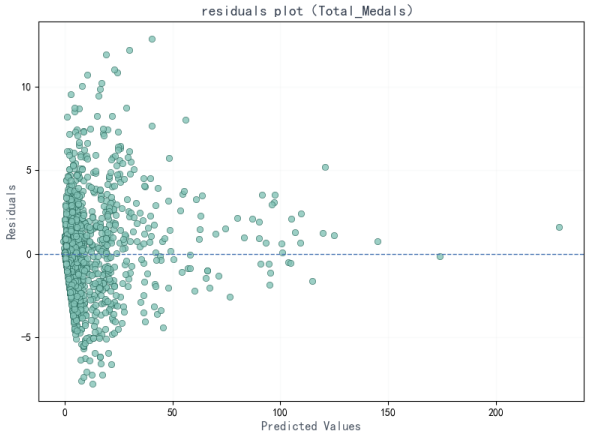

Figure 4: Total Medals Residuals Plot

(3)Feature importance analysis

The feature importance extracted through the model is as follows:

Table 2: Assessment of the gold medals

Optimal Parameters	
learning_rate	0.05
max_depth	4
min_samples_leaf	2
n_estimators	100
subsample	0.8
K-fold Cross-validation Metrics	
Average MSE	13.723
RMSE	3.705
Overall Model Metrics	
MSE	2.196
RMSE	1.482
R^2	0.920
Feature Importance (%)	
Gender_Ratio	7.55
Is_Host	1.20
Total_Events	5.32
Total_Participants	67.36
Historical_Medals	15.21
Olympic_Events	3.36

Table 3: Assessment of the total medal

Optimal Parameters	
learning_rate	0.1
max_depth	4
min_samples_leaf	1
n_estimators	200
subsample	0.8
K-fold Cross-validation Metrics	
Average MSE	62.511
RMSE	7.906
Overall Model Metrics	
MSE	3.413
RMSE	1.848
R^2	0.982
Feature Importance (%)	
Gender_Ratio	7.62
Is_Host	0.53
Total_Events	4.46
Total_Participants	70.61
Historical_Medals	12.78
Olympic_Events	4.00

3.2 Presentation of Results

After our modeling, we have depicted the performance of Olympic participants from various countries from 1896 to 2024 and predicted the gold medal and total medal counts for each country at the 2028 Los Angeles Olympics. We have visualized the two medal.(with confidence intervals of the prediction model).

The prediction results show that the following countries have not yet won any medals but are likely to win their first medal at the next Olympic Games:

PRK	QAT	BAH	ANZ	LAT	VEN	BLR
1.996771	1.958547	1.352148	1.242064	1.13024	1.079463	1

Table 4: Countries that haven't won medals but might win medals at the 2028 Olympics

The world map heat maps for predicting the total medal count and gold medal count in 2028 are as follows:

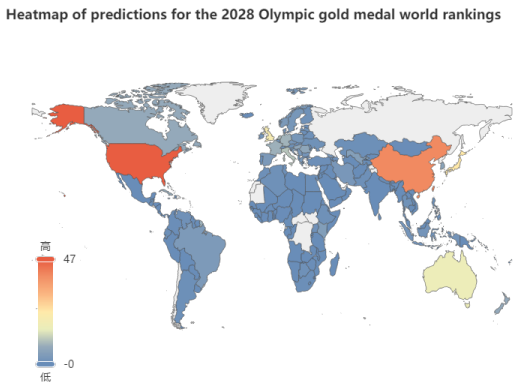

Figure 5: Gold Medals

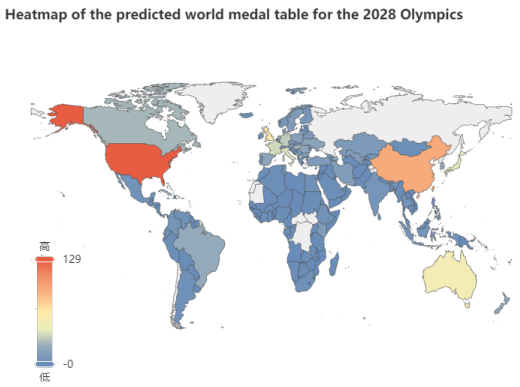

Figure 6: Total Medals

The bar chart of the top 20 countries in terms of total medals and gold medals predicted for 2028 is as follows:

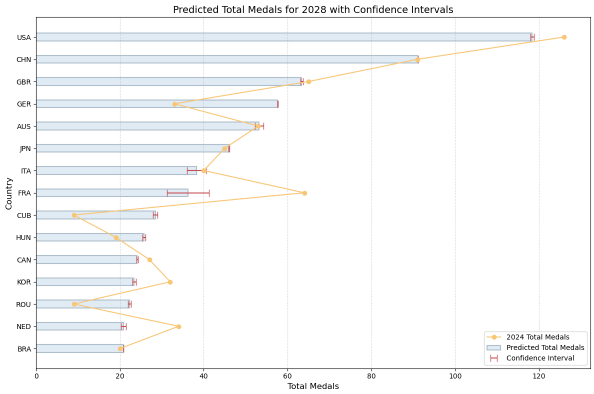

Figure 7: Gold Medal

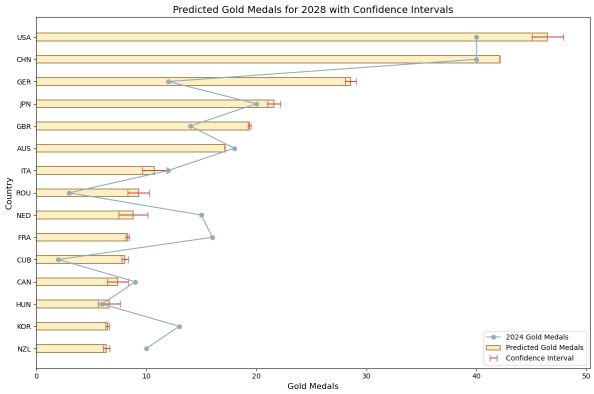

Figure 8: Total Medals

In our forecast, 69 countries may show improved performance in 2028 compared to 2024, while 47 countries may show a decline. The medal count changes for the remaining 103 countries will be within 0.1. We have created two word clouds for the countries showing improved and declined performance as follows:

Figure 9: Show improvement

Figure 10: Decline in performance

4 Problem2:Exploring variables that affect medal count

4.1 The relationship between the competition events and the number of medals won by each country

4.1.1 overall idea

First, clarify the relationship between events and each country's medal count. Then, use this to give examples of events crucial to certain countries, especially the dominant events of China and the US. Next, analyze how the host country's event selection impacts medal counts, using China and the US as cases. Compare the events set by their host years, and explore how selection differences affect the final medal standings. This can be further divided into win rate differences for host and non-host countries in host-dominant and ordinary events.

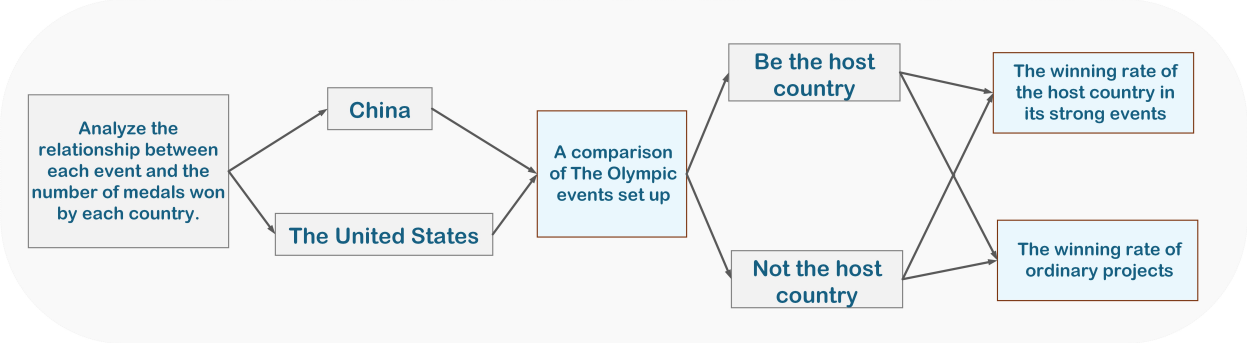

Figure 11: Logical relationships

4.1.2 The corresponding parameters in the GBRT model referred to

Table 5: Assessment Results of Medals

	Gold Medals	Total Medals
Total Events	5.32%	4.46%
Olympic Events	3.36%	4.00%

4.1.3 Analysis process

- **Step1:Calculate the winning rate**
  Define "Medal Points" as the number of medals a country has won in a certain

event,the calculation of "Medal Points" is the score obtained by weighting the number of medals. A gold medal has a weight of 3, a silver medal has a weight of 2, and a bronze medal has a weight of 1. "Participant Count" as the number of participants from that country in the event.
Then, the calculation formula for the winning rate is as follows:

$$\text{Winning Rate} = \frac{\text{Medal Points}}{\text{Participant Count}} \tag{7}$$

Through this formula, we can calculate the winning rate of each country in each event, which can better reflect the actual competitive level of the country rather than just the number of medals.
Note: The winning rate of countries that excel in team sports like basketball will be relatively high in the visualization, but since the number of participants in team events is the same, it will not affect fairness.

- **Step2:Filter data**
  Selecting the top 10 events and top 10 countries:
  To make the analysis more actionable, we selected the top 20 events with the highest medal points. Among the top 20 countries (NOC) predicted by the above model to win the most medals in 2028, we took the top ten based on the average winning rate.
  Filtering out countries and events with low winning rates:
  If the winning rates of certain countries or events decline, they may not be selected.
  Filtering out events with fewer than 10 participants:
  For events with a small number of participants, the results may be influenced by random factors. Therefore, we excluded these events to ensure the stability and reliability of the data.

- **Step3:Correlation analysis**
  We will calculate the correlation between the event and the total number of medals (Total), as well as the number of gold medals (Gold). Correlation analysis will help us identify which events have a strong relationship with the total number of medals and the number of gold medals.

4.1.4 Presentation of Results

Based on the parameters of the total Olympic events in the GBRT model, we obtained the proportion of the importance of the total number of Olympic events on the number of medals (gold medals).

Table 6: The parameters of the total Olympic events in the GBRT model

	Gold Medals	**Total Medals**
Medals_Olympic_events	3.36%	4.00%

Based on the analysis results of correlation, we have created a heat map. The horizontal axis of the heat map represents the top 10 countries (NOC), and the vertical axis represents the top 10 events. The value in each cell indicates the winning rate of that country in that event.

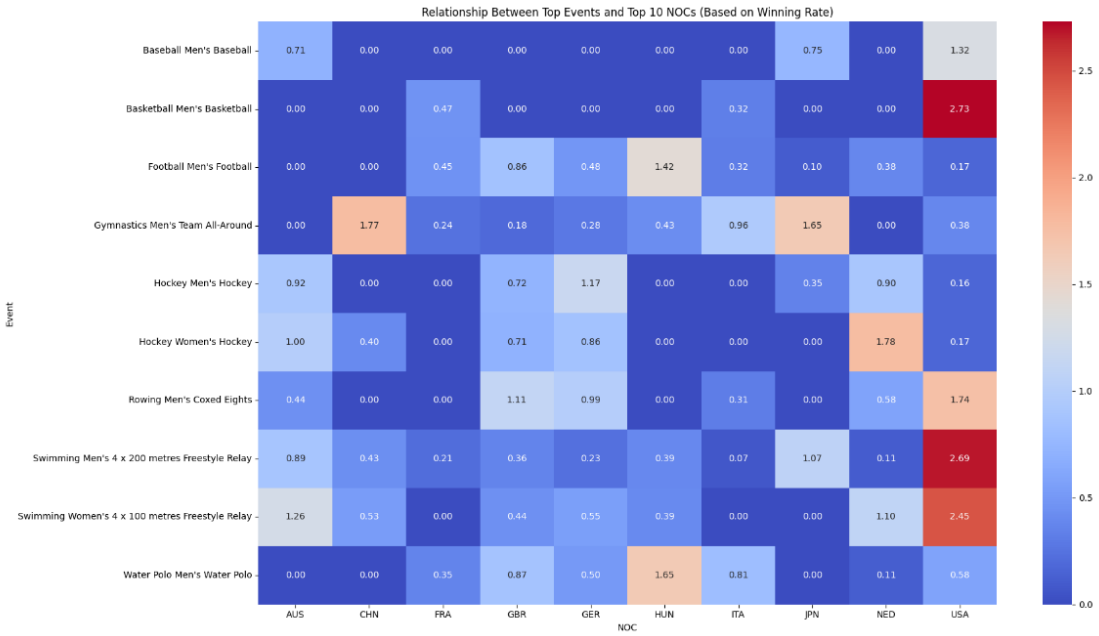

Figure 12: Heatmap of Correlation Analysis for GBRT Model

The depth of the color represents the level of the winning rate. Red indicates a higher winning rate (close to 2.5), while blue represents a lower winning rate (close to 0).

- **Countries with high winning rates:**
  1.China (CHN) performed outstandingly in the Men's Gymnastics Team All-Around event, achieving a winning rate of 1.77.
  2.The United States (USA) demonstrated exceptional strength in Men's Baseball, with a winning rate of 2.73.
  3.Australia (AUS) performed well in multiple events, particularly in Men's Football, with a winning rate of 1.42, located in the red zone of the chart, indicating a significant advantage in this event.

- **Countries with lower winning rates:**
  1.France (FRA), Germany (GER), and the Netherlands (NED) showed relatively low winning rates in several events. For instance, France's winning rates in Men's Football and Men's Water Polo were close to zero, suggesting they hardly won any medals in these events.
  2.Italy (ITA) and Japan (JPN) also had lower winning rates in certain events. For

example, their performances in the Men's Gymnastics Team All-Around event were poor, with winning rates of 0.43 and 0.18 respectively.

- **Specific project advantages:**
  1.The performance in swimming events was particularly outstanding. For instance, the United States had a winning rate of 2.69 in the Men's 4 x 200 metres Freestyle Relay, far exceeding that of other countries.
  2.In water polo, the United States and Hungary (HUN) also performed strongly, with winning rates as high as 1.65.

- **Relationship between countries and events:**
  1.The United States (USA) excelled in multiple events, especially in swimming and baseball, indicating its strong competitive ability in these fields.
  2.China (CHN) maintained its advantage in some traditional strong events such as gymnastics, suggesting that China has invested more in training and resources in these events.
  3.Australia (AUS) performed well in some less popular events, demonstrating its strength in these areas.

4.1.5 Summary

- From the heat map, it can be seen that although some countries like the United States and China won a large number of medals and achieved high winning rates in multiple events, some countries also showed unexpected performances in certain individual events.

- Data diversity: This chart indicates that the distribution of winning rates in Olympic events is not simply positively correlated with the number of medals. Some countries may perform averagely in many events but have high winning rates in certain specific events.

4.2 The relationship between the home advantage effect and the number of medals won

4.2.1 overall idea

We extracted each country's annual project participants and medal scores from organized data. To explore if host country's new events affect medal distribution, we compared the two latest Olympics. After that, we got the host country label (Is_Host_Country == 1) and calculated the host country's winning rate in new events, as well as that of the top 20 non-host countries in the predicted 2028 total medal table. Then, we calculated the difference between the host and non-host countries' winning rates, whose positive or negative value reflects the host country's performance in new events.

4.2.2 The corresponding parameters in the GBRT model referred to

This indicates that the home advantage has a relatively minor impact on the overall medal count and the number of gold medals. The fact that it is more significant in Gold medals than in Total medals also suggests that the home advantage plays a crucial role in whether a major country can win gold medals.

Since countries with a smaller number of participants find it more difficult to host the Olympics and these countries account for a large proportion, the overall importance of the home advantage in the model is not particularly prominent.

Table 7: Assessment Results of Medals

	Gold Medals	**Total Medals**
Is_Host	1.20%	0.53%

4.2.3 Presentation of Results

We selected the corresponding difference data from the seven Olympic Games in the 21st century. In the visualization chart, the more columns are distributed in the positive value area, the better the performance of the host country in the new events of that Olympic Games. Conversely, it indicates an unsatisfactory performance.

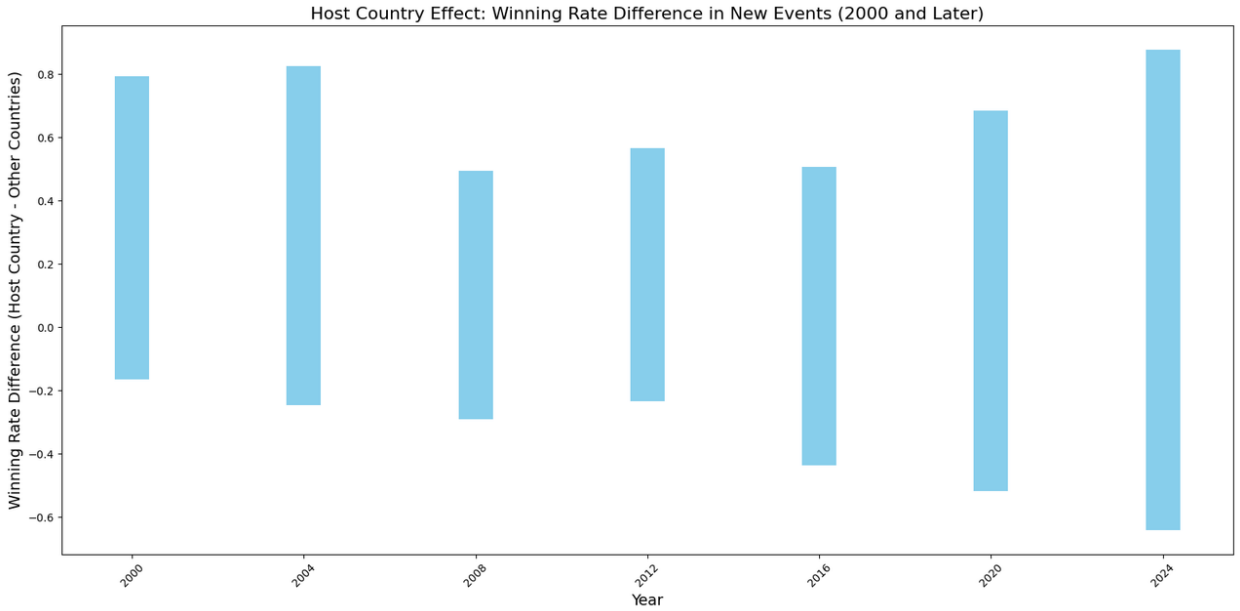

Figure 13: The corresponding difference data of the seven Olympic Games

For instance, in the visualization chart, 2000, 2008, 2012 all have a relatively obvious host country effect, which means the new events are more in line with the host country's strong points. In 2016 and 2020, there is still a certain advantage, but in the 2024 Olympic Games, there is no obvious advantage. This also reflects from the side that the sports levels of various countries are tending to be comprehensive.

5 Problem3:Testing the “great coach” effect

5.1 Data collection and processing

This question selects two representative and outstanding coaches mentioned in the title, Lang Ping and Bela Karolyi, as samples for analysis. The data on their tenure is sourced from wiki. The sports involved are women’s volleyball and women’s gymnastics, and the countries involved are China, the United States, and Romania.

The situations of the two coaches, Lang Ping and Bela Karolyi, are as follows:

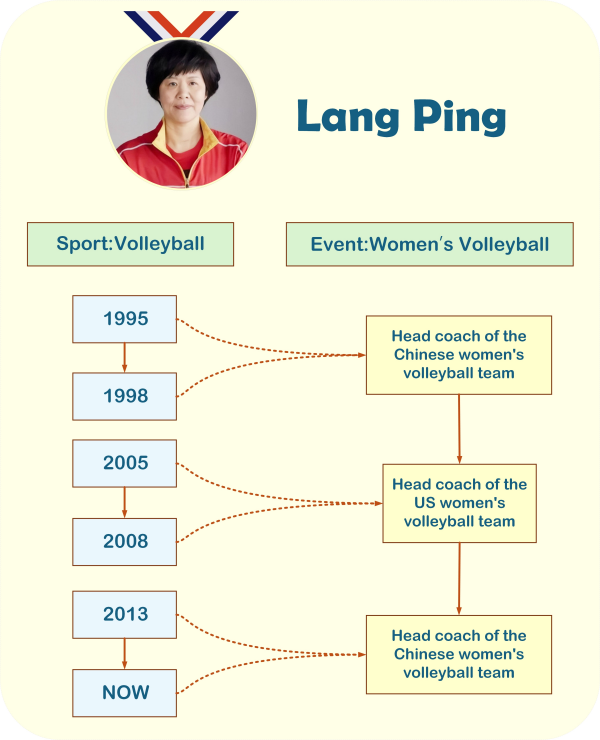

Figure 14: Lang Ping

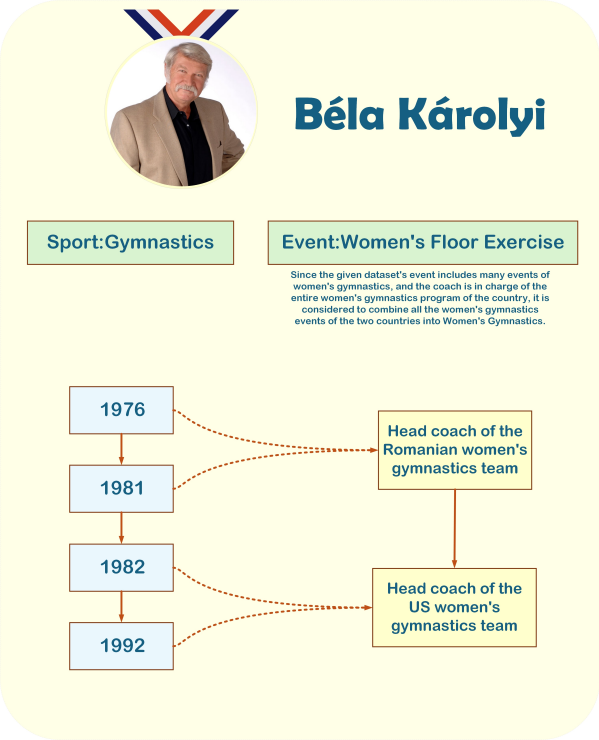

Figure 15: Bela Karolyi

5.2 Analysis process

We choose the Event Study Method, which involves comparing the changes in the number of medals before and after an event. In the analysis of the "great coach effect", the replacement of coaches is regarded as a key event with a clearly defined time point. The essence of the event study method is to capture and analyze the immediate or long-term impact of such events on the target variable. The event study method can still be conducted without a control group. Therefore, by comparing the changes in the number of medals before and after the event, we can identify the effect of the event itself.

- **Step1:Determine the event window**
  The event window theory is the basic framework of the Event Study Method.

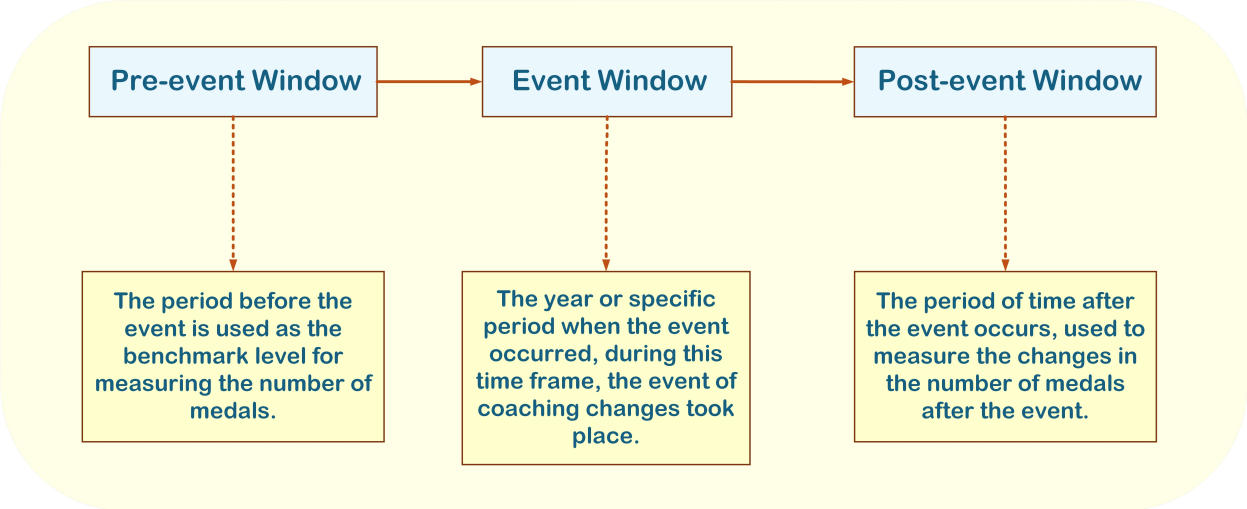

Figure 16: Basic framework

For the selection of the post-event window, we choose the two Olympic Games after the coach took office.For the selection of the pre-event window, we choose the two Olympic Games before the coach took office.

- **Step2:Time trend analysis**
  Through regression analysis, other possible factors that may affect the number of medals, such as a country's GDP and sports investment, can be controlled. By analyzing the long-term trend of the number of medals, we can ensure that the effect captured is that of coaching changes.

  The regression model we used is:

$$\text{Contribution}_{it} = \beta_0 + \beta_1 \text{Coach}_{it} + \beta_2 \text{Time}_{it} + \epsilon_{it} \tag{8}$$

  Contribution_{it}refers to the project contribution degree of country i at time t, and its calculation method is:

$$\text{Contribution}_{it} = \frac{\text{Event\_Medal\_Score}}{\text{Medal\_Score}} \times 100\% \tag{9}$$

  Among them, $Event\_Medal\_Score$ represents the weighted medal count of this event (gold medals × 3 + silver medals × 2 + bronze medals × 1), and $Medal\_Score$ represents the weighted medal count that the country has won in this Olympic Games (gold medals × 3 + silver medals × 2 + bronze medals × 1).

- **Step3:Inspection and Significance Analysis**

  Ordinary Least Squares Regression: It is used to estimate the influence of the coach effect (Coach) and the time effect (Year) on Contribution (medal contribution).
  1.$Coefficients$: Analyze the impact of each variable on the contribution to medals.
  2.P_Values: Used to determine whether each regression coefficient is significant. If the p-value is less than 0.05, it indicates that the variable has a significant impact on the number of medals.

  coefficient of determination(R^2): It is used to measure the explanatory power of the regression model for the variation in the number of medals. The higher the R^2, the better the model fits the data.

  F-test: It is used to test the significance of the entire regression model. The F-statistic and its corresponding p-value can help determine whether the overall model is effective.

  The values obtained from the match data of the Chinese women's volleyball team from 2008 to 2024 and the US volleyball team from 2000 to 2016 are as follows:

Table 8: Major Event 1

Women's Volleyball(CHN)	
Event_Window	(2016, 2024)
Pre_Event_Window	(2008, 2015)
R^2	0.734539371
Coefficients	
const	4.770446
Coach	0.028571
Year	-0.002372
dtype	float64
P_Values	
const	0.156745
Coach	0.146830
Year	0.156862
dtype	float64

Table 9: Major Event 2

Women's Volleyball(USA)	
Event_Window	(2008, 2016)
Pre_Event_Window	(2000, 2007)
R^2	0.922076948
Coefficients	
const	0.994716
Coach	0.012241
Year	-0.000497
dtype	float64
P_Values	
const	0.216184
Coach	0.062541
Year	0.216184
dtype	float64

The values obtained from the competition data of American women's gymnastics from 1976 to 1992 (excluding the data of 1980) and that of Romanian women's gymnastics from 1972 to 1984 are as follows:

Table 10: Major Event 1

Women's Gymnastics(USA)	
Event_Window	(1984, 1992)
Pre_Event_Window	(1976, 1983)
R^2	0.39762542
Coefficients	
const	0.444860
Coach	0.042031
Year	-0.000225
dtype	float64
P_Values	
const	0.980680
Coach	0.749669
Year	0.980680
dtype	float64

Table 11: Major Event 2

Women's Gymnastics(ROU)	
Event_Window	(1976, 1984)
Pre_Event_Window	(1972, 1975)
R^2	0.962876843
Coefficients	
const	58.643974
Coach	0.723439
Year	-0.029738
dtype	float64
P_Values	
const	0.309572
Coach	0.140542
Year	0.309570
dtype	float64

5.3 Explain the results of the model

- **The impact of coaching changes:**
  1.For Chinese and American women's volleyball, coaching changes have a significant positive impact on the number of medals ($P\_value < 0.05$).
  2.For American women's gymnastics, the impact of coaching changes is not significant ($P\_value > 0.05$).
  3.For Romanian women's gymnastics, coaching changes have a significant positive impact ($P\_value < 0.05$).

  In summary, coaching changes have a positive impact on the number of medals, which can indicate the existence of the "great coach effect". At the same time, the model shows that for Chinese and American women's volleyball and Romanian women's gymnastics, the P value is less than 0.05, indicating that the positive impact of the "great coach effect" on the number of medals is significant.

- **Time trend:**
  According to the results, Year (β_2) is negative in all cases, and for Chinese and American women's volleyball and Romanian women's gymnastics, the P value is less than 0.05, indicating that the time trend has a significant negative impact on the number of medals in these sports; this further proves the significant positive impact of the "great coach effect" on the number of medals.

- **Model goodness of fit:**
  1.The model goodness of fit for Romanian women's gymnastics is the highest (R^2 = 0.962877), indicating that the model explains most of the changes in the number of medals.
  2.The model goodness of fit for American women's volleyball is second (R^2 = 0.922077), also indicating that the model explains most of the changes in the number of medals.
  3.The model goodness of fit for Chinese women's volleyball is moderate (R^2 = 0.734537).

- **Analysis of model shortcomings:**
  According to the results, we can see that the model goodness of fit for American women's gymnastics is the lowest (R^2 = 0.397625), indicating that the model explains less of the changes in the number of medals, and the impacts of coaching changes and time trends are not significant ($P\_value > 0.05$). We found that the United States did not participate in the 1980 Olympics due to historical reasons, and 1980 was the previous Olympic Games when Bela Karolyi joined the American women's gymnastics team (event window) as a coach. At the same time, our entire model studies the immediate effect of excellent coaches on the number of medals, so the data from the last Olympic Games before the event window has a significant impact on the overall results. Therefore, the absence of data from the 1980 Olympics is the main reason for the poor model performance in American women's gymnastics.

5.4 Three countries should hire coaches for the project

We selected all competition data of China, the United States and Romania from the 20th century to the present. Through formula calculation, we obtained the contribution rate of different events to the total number of medals of each country. We visualized the top ten events with the highest contribution rates of these three countries and the contribution rates of these events to the number of medals of the countries after assuming that they hired excellent coaches.

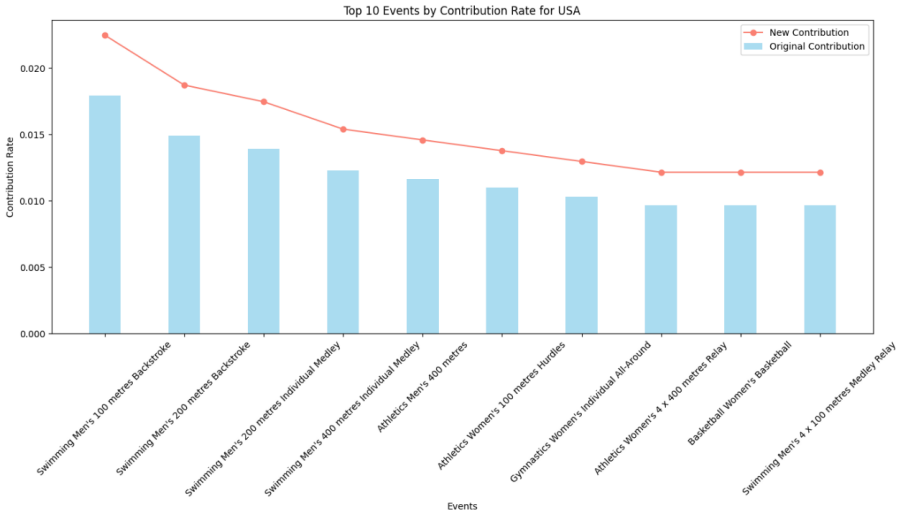

Figure 17: Top 10 Events by Contribution Rate for USA

Swimming, track and field, and women's gymnastics have a relatively high contribution rate to the United States.

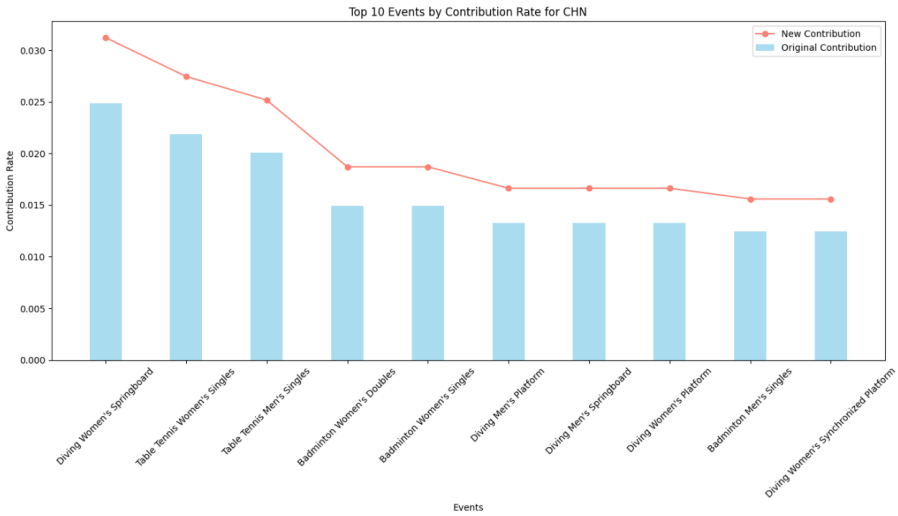

Figure 18: Top 10 Events by Contribution Rate for CHN

Diving, table tennis, and badminton have a relatively high contribution rate to China.

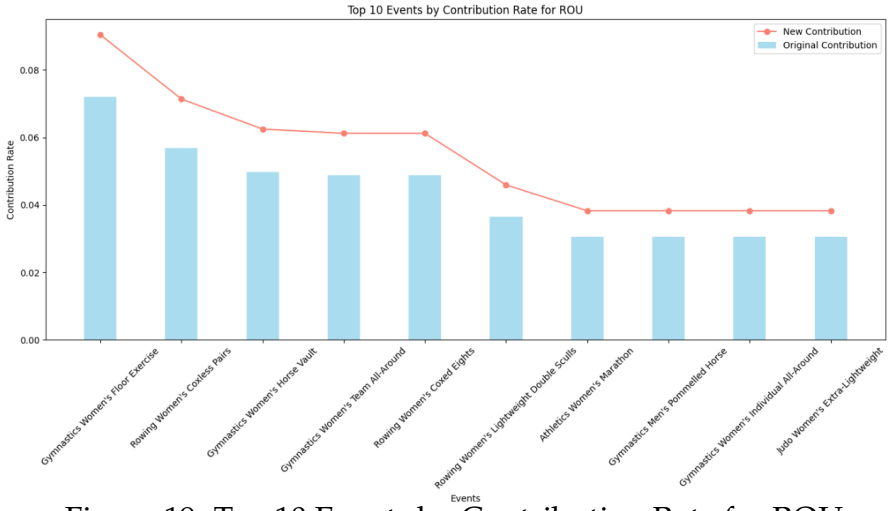

Figure 19: Top 10 Events by Contribution Rate for ROU

Gymnastics, women's rowing, and women's track and field have a relatively high contribution rate to Romania.

The proportion of improvement in the coaching effect obtained from the model results is used to calculate the contribution to the number of medals after hiring an outstanding coach with this formula:

$$\text{Contribution}_{\text{new}} = \text{Contribution}_{\text{old}} \times (1 + \text{coach effect}) \tag{10}$$

Based on the model results of the "Great Coach Effect", we have verified that the positive impact of the "Great Coach Effect" on the number of medals is significant. Therefore, for these three countries, our suggestion is that they each hire "excellent" coaches for these events with higher contribution rates.

6 Problem4:Our unique insights discovered

Based on the GBRT model of the first question, we obtained the proportion of influence of different feature variables on the total number of medals and the number of gold medals.

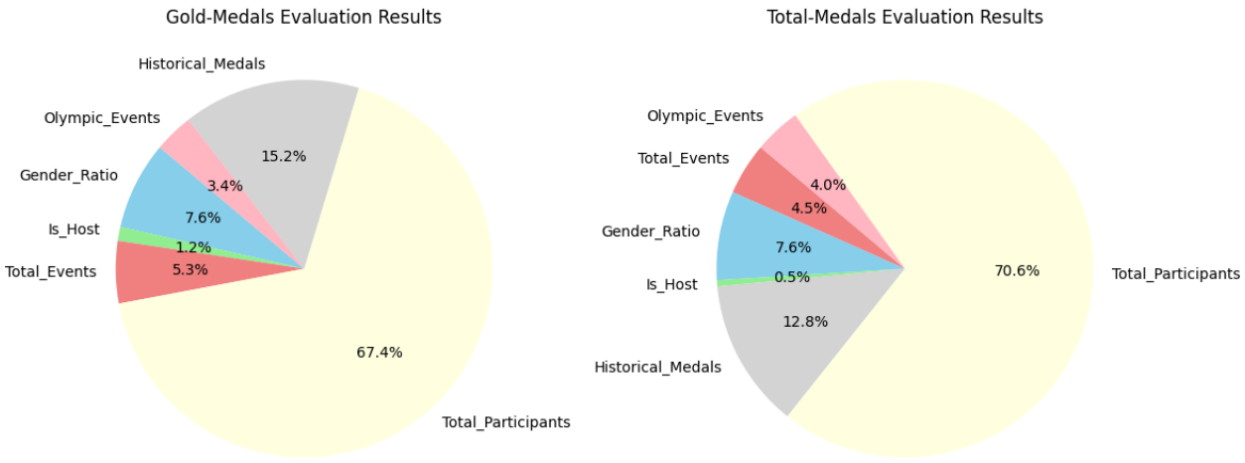

Figure 20: The proportion of influence of different characteristic variables

- **The consistency of the contribution to the sex ratio (both 7.6%) indicates its stability.** The consistent contribution ratio of the gender ratio to gold medals and total medals indicates that the gender ratio is a stable and long-term influencing factor, and its effect may not fluctuate significantly due to short-term policy changes.
- **The contribution of the gender ratio to the number of gold medals and the total number of medals is consistent (both 7.6%), indicating that it has a direct impact on high-level competition.**The contribution of the gender ratio to the number of gold medals is consistent with that to the total number of medals, indicating that when the gender ratio is balanced, athletes of all genders are equally important in top-level competition (i.e., gold medal competition) and overall performance (i.e., total medal competition).
- **Optimizing the gender ratio can enhance the overall sports strength of a country.**The optimization of the gender ratio not only affects the total number of medals but also directly influences the number of gold medals, which indicates that the improvement of the gender ratio is of great significance to enhancing a country's top-level sports competitiveness.
- **The consistency of the contribution of the gender ratio may be related to the national culture and sports policies.**The consistent contribution of the gender ratio to both gold medals and total medals may be related to a country's sports policies, cultural environment and historical traditions. For instance, some countries have long emphasized the equal development of male and female athletes, while others may invest less in athletes of a certain gender.

7 Memorandum

To: National Olympic Committees of all countries

From: Team #2510034

Subject: The Secret of Olympic Medals: Insights into Athlete Gender Ratio

Date: January 27, 2025

Dear National Olympic Committees of all countries:

Hello! We're glad to share GBRT model-based recommendations. While exploring the gender ratio of Olympic athletes, we've got valuable insights to help you select and cultivate athletes.

Rigorous research shows the gender ratio's contribution to gold and total medals is highly consistent, at 7.6%. It's a stable, long-term factor; short-term policy changes seldom cause big fluctuations. So, countries should aim for long-term gender balance and plan comprehensive strategies involving education, grassroots selection, and sports culture building.

Also, when gender is balanced, both male and female athletes are equally vital in gold and total medal competitions. National Olympic Committees must ensure fair resource allocation and system setup for them to eliminate disparities and guarantee equal chances.

Moreover, optimizing the gender ratio matters not only for total medals but also gold medals, significantly enhancing a country's top sports competitiveness. We suggest comprehensively boosting female athletes' participation and training quality from grassroots to top-level, like adding specialized programs, creating diverse competitions, and improving the environment. Focus on cultivating female coaches and managers too.

Note that the gender ratio's consistency links closely to countries' cultures and policies. Some value gender equality; others have imbalances due to unequal investment. Globally, it's urgent to promote gender equality, encourage women in sports, and raise social recognition. Countries should learn from advanced ones like those in Northern Europe and adapt to their own situations.

To implement these, build a scientific monitoring system, focus on the gender ratio's impact on medals, periodically assess policy effectiveness, quickly spot deviations, and optimize resource allocation. For countries with low female participation, actively promote and invest in women's sports to enhance their enthusiasm and competitiveness.

We believe these practical measures will help you optimize athlete selection and training, invigorate global sports, and let the Olympic stage shine with gender equality. Thank you for reading. Let's create sports heights together!

Yours Sincerely,
Team #2510034

References

[1] Zhu Mengnan, Peng Tao, Chen Ke. Mathematical Analysis of Influencing Factors of the Olympic Medal Table[J]. Contemporary Sports Technology, 2017, 7(27): 5. DOI: 10.16655/j.cnki.2095 - 2813.2017.27.239.

[2] Zhu Yin. Empirical Analysis of Influencing Factors of the Olympic Medal Table: Taking the 31st Olympic Games as an Example[J]. Journal of Chifeng University (Natural Science Edition), 2017. DOI: CNKI:SUN:CFXB.0.2017 - 03 - 048.

[3] Zhang Yuhua. Model Construction and Quantitative Analysis of the Number of Olympic Medals and Five Influencing Factors[J]. Shandong Sports Science & Technology, 2013, 35(3): 43 - 47.

[4] Tan Hong, Ren Jiaojiao. Empirical Research on the Host Effect of the Olympic Games[J]. Journal of Langfang Teachers University: Natural Science Edition, 2013, 13(2): 4. DOI: 10.3969/j.issn.1674 - 3229.2013.02.027.

[5] Sun Xiujie, Yu Jie, Zhao Lunan, et al. Analysis of Medal Distribution and Performance Characteristics of the Chinese Swimming Team in the Past Five Olympic Games[J]. Bulletin of Sport Science & Technology, 2022, 30(5): 17 - 20. DOI: 10.19379/j.cnki.issn.1005 - 0256.2022.05.004.

[6] Tian Hui, He Yiman, Wang Min, et al. Medal Prediction and Competition Strategy of Chinese Athletes in the 2022 Beijing Winter Olympics: Based on the Analysis of the Home - field Advantage Effect of the Olympic Games[J]. China Sport Science, 2021, 41(2): 3 - 13 + 22. DOI: 10.16469/j.css.202102001.

[7] Ye Zuan, Gao Ping. Comparison of Sino - US Olympic Competitiveness: Based on the Statistical Analysis of the Results of the Paris Olympics[C]//Hubei Sports Science Society. Abstracts of the Second Hubei Sports Science Conference and the Fifth Academic Forum on the Development of Modern Sports and Military Training. School of Sports Training, Wuhan Sports University, 2024: 2. DOI: 10.26914/c.cnkihy.2024.056194.

[8] Wu Yili. Analysis of the Competitive Strength of the Chinese Sports Delegation in the Past Five Summer Olympics and the Preparation Situation for the Paris Olympics[C]//Chinese Society of Sport Science. Abstracts of the 13th National Sports Science Conference - Special Reports (Sports Training Branch). Beijing Sport University, 2023: 3. DOI: 10.26914/c.cnkihy.2023.063805.

[9] Zhang Yuhua. Prediction of the Number of Medals of China in the 31st Olympic Games Based on the Linear Regression Dynamic Model[J]. Journal of Henan Normal University (Natural Science Edition), 2013, 41(2): 24 - 26 + 60. DOI: 10.16366/j.cnki.1000 - 2367.2013.02.003.

[10] Wang Guofan, Tang Xuefeng. Research Dynamics and Development Trends of Olympic Medal Prediction at Home and Abroad [J]. China Sport Science and Technology, 2009, 45(06): 3 - 7 + 135. DOI: 10.16470/j.csst.2009.06.016.

Report on the Use of Artificial Intelligence

This paper used Doubao, an AI developed by ByteDance, to assist in translating the content of the article into English and polishing the English grammar during the writing process.