

# The MCM Thesis of Team 2517105

## Summary

The Olympic Games have always been a global spectacle, capturing the imagination and attention of millions. For enthusiasts and analysts alike, forecasting the medal tally for each participating nation in the 2028 Olympics has emerged as a compelling area of interest.

**In Task 1**, we utilized a Random Forest model to predict the Olympic performance of various countries. We evaluated the contribution of different feature variables to these predictions and explored the myriad factors that influence Olympic outcomes. To gain deeper insights into the model's results, we employed SHapley Additive exPlanations (SHAP), an interpretable machine learning method. This allowed us to analyze the trained model's outcomes comprehensively. We also projected the medal outcomes for 2028, identifying which countries might first break through the zero-medal barrier.

**For Task 2**, we applied a linear regression model from statistics to analyze the underperforming events of three specific countries. By examining the medal outcomes of dominant countries in these events, we pinpointed instances where medal counts experienced significant shifts. This served as a benchmark to investigate the impact of coaching on athlete performance and to assess the extent of the "Great Coach Effect." Ultimately, we provided investment recommendations for different countries based on these findings.

**In Task 3**, building on the models from the previous tasks, we analyzed additional information from the dataset. We focused on the gender ratio of participants in various events and examined the host country effect. By reviewing the gender ratios in popular events over four Olympic cycles and visualizing this data, we were able to offer valuable insights and recommendations for event organizers.

The results of the model can provide data support for the formulation of sports policies, helping governments and sports organizations understand which areas require more investment and attention. By identifying potential projects and athletes, relevant departments can allocate resources more effectively, thereby enhancing the country's competitiveness in international events.

**Keywords:** Random Forest, SHapley Additive exPlanations, Linear Regression, Statistics, Great Coach Effect.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Background . . . . .	3
1.2	Restatement of the Problem . . . . .	3
1.3	Assumptions . . . . .	4
<b>2</b>	<b>Analysis of the Problem</b>	<b>4</b>
2.1	Analysis of the Problem 1 . . . . .	4
2.2	Analysis of the Problem 2 . . . . .	5
2.3	Analysis of the Problem 3 . . . . .	6
<b>3</b>	<b>List of Notation</b>	<b>7</b>
<b>4</b>	<b>Model Construction and Problem 1</b>	<b>7</b>
4.1	Algorithm Design . . . . .	7
4.2	Data Preprocessing and Variable Selection . . . . .	8
4.3	Model Analysis and Results . . . . .	9
4.3.1	Medal Table Forecast for the 2028 Summer Olympics in Los Angeles, USA . . . . .	10
4.3.2	Predicted number of countries with zero MEDALS by 2028 . . . . .	10
4.3.3	Analysis of Project and Country Features . . . . .	10
<b>5</b>	<b>Model Construction and Problem 2</b>	<b>12</b>
5.1	Algorithm Design . . . . .	12
5.2	Data Analysis and Variable Selection . . . . .	13
5.3	Model Analysis and Results . . . . .	14
5.4	Investment Recommendations . . . . .	15
5.4.1	Investment Recommendations for the United States . . . . .	15
5.4.2	Investment Recommendations for China . . . . .	16
5.4.3	Investment Recommendations for Great Britain . . . . .	16

<b>6</b>	<b>Model Construction and Problem 3</b>	<b>17</b>
6.1	Host effect . . . . .	17
6.2	Gender Ratio . . . . .	18
6.3	Recommendations . . . . .	19
<b>7</b>	<b>Evaluation and Generalization</b>	<b>20</b>
7.1	Advantages of the Study . . . . .	20
7.2	Limitations of the Study . . . . .	21
7.3	Application of the Study . . . . .	21
	<b>References</b>	<b>22</b>
	<b>Appendices</b>	<b>22</b>

# 1 Introduction

## 1.1 Background

The Olympic Games are the largest and most influential comprehensive sports events in the world, held every four years, and divided into the Summer Olympics and the Winter Olympics. Since the first modern Olympic Games were held in Athens, Greece in 1896, the Olympics have become an important stage for global athletes to showcase their talents and pursue excellence, as well as an important platform for cultural exchanges and friendship enhancement among countries.



Figure 1: The Olympic Games

In recent years, with the application of new technologies such as big data and artificial intelligence, the organization and management of the Olympics have also become more scientific and efficient. Through data analysis, countries can better formulate training plans, optimize resource allocation, and improve the competitive level of athletes. At the same time, these new technologies have also provided tools for audiences to predict match results. For fans of the Olympics, predicting the overall "medal table" of each country has become a highly anticipated matter.

## 1.2 Restatement of the Problem

This study aims to address the following research problems based on the data given in the question:

- **Complete the Medal Count Prediction Model**
  - Based on the model, predict the medal table for the 2028 Los Angeles Summer Olympics, including prediction intervals for all results, and predict which countries may advance or regress.
  - The model should include countries that have not yet won medals and predict the likelihood of these countries winning medals for the first time in the next Olympics.

- Consider the Olympic sports (quantity and type), analyze the relationship between sports and the number of medals won by each country, determine the most important sports for each country, and analyze how the host country's choice of sports affects the results.
- **Explore the Impact of the "Great Coach" Effect on Medal Counts:** Select three countries and identify sports projects that should invest in "great" coaches, and estimate their impact.
- **Reveal Other Original Insights of the Model on Olympic Medal Counts:** Explain how these insights can inform National Olympic Committees.

### 1.3 Assumptions

To simplify the model, we propose the following assumptions:

- **International Competitive Environment:** Assume that the international competitive environment is relatively stable in the short term, that is, the competitive relationships between countries will not undergo drastic changes.
- **Interaction Effect of Coaches:** Assume that the training level of coaches in various countries is relatively stable in the short term and can be superimposed.

## 2 Analysis of the Problem

### 2.1 Analysis of the Problem 1

Despite fluctuations in the number of medals won by various countries in recent Olympic Games, overall, the changes in the medal counts of most countries are not significant and show relatively clear development trends, as shown in Figure 2a. After several editions of the Olympics, a relatively stable competitive pattern has essentially taken shape: the "first group" represented by the United States and China, the "second group" represented by the United Kingdom, France, and Australia, and the "third group" composed of many other countries. To prove this stability, we selected data from all countries that participated in the 2020 and 2024 Olympics and won medals for relevant analysis and drew a heatmap, as shown in Figure 2b, which proves the speculation of this paper.

In previous studies, Schlembach et al.<sup>[7]</sup> used the random forest model to predict the performance of various countries in the Olympics and evaluated the contribution of different feature variables to the prediction. However, the focus was on the overall performance of various countries in the Olympics, that is, the number of gold medals or total medals, without exploring the differences under the influence of different events. In fact, there are significant differences in the performance of different countries in different events. For example, as a sports powerhouse, the United States has long been at the top of the Olympic medal table, but American athletes have never won medals in events such

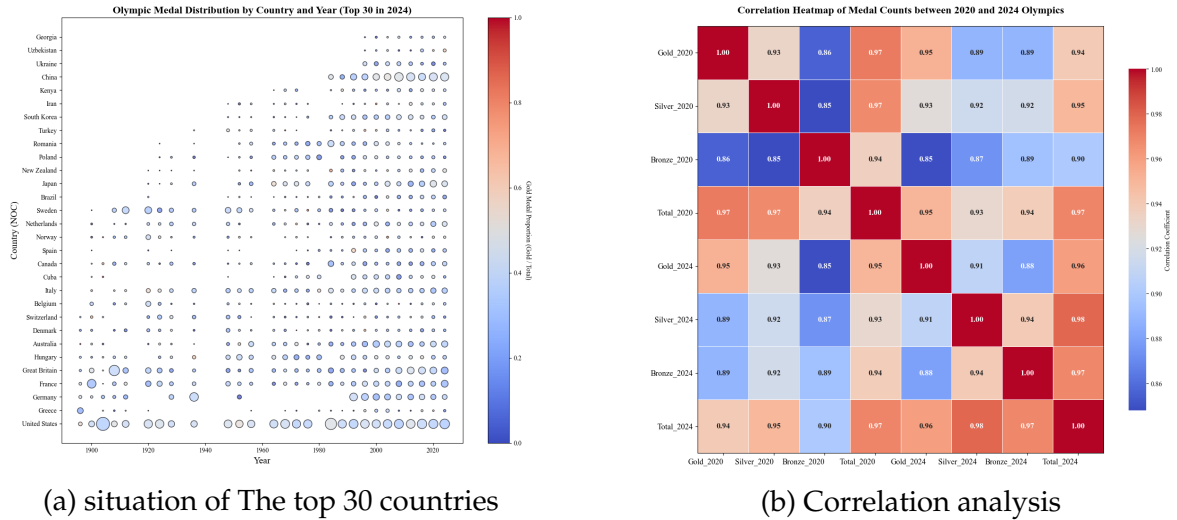


Figure 2: Analysis of Olympic Medal Counts

as table tennis and badminton, and the overall data cannot explain this difference. In addition, from the perspective of improving sports performance, it is necessary to discuss the factors affecting performance on a per-event basis.

This paper adopts the method of SHI Huimin et al.<sup>[2]</sup> and selects the random forest as the main model for analysis. The random forest is composed of multiple independent learners through linear combination, which can reduce the prediction bias as a whole and has strong robustness. For example, Kellie et al.<sup>[1]</sup>'s study examined the effectiveness of random forest variable importance measures in identifying real predictive variables through extensive simulation studies. This paper trains the model at the country-event level, and then predicts the number of medals and gold medals won by various countries in events and sub-events in the 2028 Summer Olympics, estimates the number of countries that may break through 0 medals, and on this basis, evaluates the main effects of each feature in sub-events.

## 2.2 Analysis of the Problem 2

In the history of sports, outstanding coaches often have a profound impact on the careers of athletes and the development of national sports projects. For competitive sports, a great coach can enhance the performance of a country's athletes in a particular Olympic cycle and even achieve breakthrough progress in a certain event. Moreover, unlike athletes, coaches face fewer restrictions when coaching across countries, as they do not need to change their nationality to coach. Against the backdrop of globalization, this kind of international sports cooperation and exchange has become increasingly frequent, and the large number of cross-country coaching phenomena has provided possibilities for "great coaches".

A great coach refers to the impact of a coach's coaching experience in different countries on the performance of sports projects in that country. Successful coaches can significantly improve the competitive level of athletes through professional knowledge, tactical

vision, and management ability, thereby driving the overall performance of the national team. For example, Lang Ping<sup>[6]</sup> once coached the volleyball teams of China and the United States, leading the Chinese women's volleyball team to win gold medals in the 1984 and 2016 Olympics, and the US team to win a silver medal in 2008. Bella Karolyi also coached the women's gymnastics teams of Romania and the United States, achieving great success in both cases. These cases have put forward a hypothesis that great coaches can bring significant medal increments.

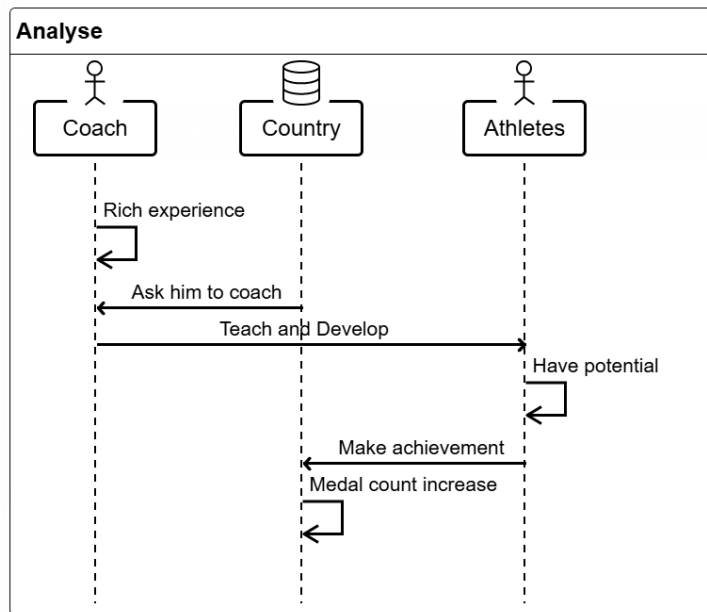


Figure 3: the Impact of a Coach's Coaching Experience

Given this, this paper aims to explore whether great coaches have significantly changed the number of medals, especially when some countries introduce coaches with world-class successful experience, whether they can drive their performance in related sports projects. By analyzing historical data, we will focus on some key countries and projects, identify which sports may benefit from "great coaches", and further assess the marginal contribution of this effect on the number of medals, providing a scientific basis for the strategy of building a sports power.

### 2.3 Analysis of the Problem 3

Based on the developed model, this paper proposes the following two insights and provides information for National Olympic Committees:

- **Host Effect:** The host effect refers to the phenomenon in which the host of a large-scale sports event has advantageous resources, policy support, and home-field advantage, which positively impacts the organization of the event and the competitive state of athletes.
- **Gender Differences:** The historical and cultural backgrounds of certain sports may lead to gender differences. For example, traditional sports such as football and

hockey are often regarded as male sports, and this perception is deeply rooted in society, affecting women's willingness to participate. The degree of social support for women's participation in sports directly affects their participation rate. In some societies, recognition and support for female athletes are still insufficient, resulting in a lower participation ratio of women in some sports, which indirectly affects the country's medal count.

### 3 List of Notation

Table 1: Notation and Brief Definitions

Notation	Brief Definition
$\phi_i(f(x))$	Contribution of feature $i$ to prediction
$F$	Set of all features
$S$	Subset of features excluding $i$
$ F $	Number of features
$ S $	Number of features in subset $S$
$f(S)$	Prediction on subset $S$ without feature $i$
$\alpha_0$	Baseline medal count
$\alpha_1$	Influence of baseline characteristics
$\alpha_2$	Influence of coach effect
$\epsilon_i(t)$	Unexplained model error

## 4 Model Construction and Problem 1

### 4.1 Algorithm Design

Due to their inherent complexity, machine learning methods are often regarded as "black box" models, making it difficult to intuitively interpret their underlying economic implications. Therefore, when interpreting the results derived from machine learning methods, additional explanatory means are usually required. Common explanatory methods include partial dependence plots, accumulated local effects, and model-agnostic local interpretability methods, among others. However, research by Lundberg et al.<sup>[3]</sup> has pointed out that for interactive modeling methods such as random forests, the variable weights obtained by the above methods may show significant differences between different evaluation systems and may even lack comparability. In addition, the built-in independent variable importance screening mechanism of the random forest model tends to favor discrete variables and continuous variables with more values when selecting independent variables. Given the limitations of these methods, this paper chooses to use the SHAP method as the interpretability analysis tool.

The SHAP method decomposes the contribution of each feature variable to the model prediction and quantifies its contribution degree using SHAP values. Studies by SHI Huimin et al.<sup>[2]</sup> have shown that when using the SHAP method to predict the number of



medals in various Olympic events, Shapley values of different variables can be obtained, which can provide a basis for detailed analysis of each individual sport. By summarizing the SHAP values obtained from decomposing all sample predictions, the SHAP values of feature variables can be obtained.

In this paper, we use SHAP values to measure the contribution of each feature to the model prediction. The basic principle is as follows: Consider all possible feature combinations. For a model containing  $n$  features, there are  $2^n$  possible feature combinations; Calculate the marginal contribution of each feature in each combination, that is, the change in model output before and after adding the feature; Take the weighted average of the contributions in all possible combinations to obtain the SHAP value of the feature. Its calculation formula can be expressed as:

$$\phi_i(f(x)) = \sum_{S \subseteq F \setminus \{i\}} \frac{|F|!}{|S|!(|F| - |S| - 1)!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

In equation 1,  $f(S \cup \{i\})$  is the model's prediction value on the feature subset  $S$  plus feature  $i$ .  $\frac{|S|!(|F| - |S| - 1)!}{|F|!}$  is the weight of the Shapley value, used to ensure that the contribution of each feature is fairly calculated.

At the same time, by considering the marginal contribution of the features, this paper evaluates the independent impact of each feature, calculates the difference between the prediction value with the feature and the prediction value without the feature, and can understand the independent impact of each feature on the model prediction; at the same time, the interaction between features is evaluated,  $\phi_i(f(x))$  not only considers the impact of a single feature, but also the interaction between features. Specifically, the higher the  $\phi_i(f(x))$ , the greater the contribution of the feature variable to the model prediction, the stronger the correlation with the prediction result, and the more likely it is to be a key factor affecting the predicted variable.

## 4.2 Data Preprocessing and Variable Selection

This paper optimizes the selection of feature variables and the data processing process, constructing a random forest model for Olympic medal prediction. The feature variables of the model include the following categories: national feature variables, sport historical performance variables, and national potential trait variables. First, dummy variable encoding is used to distinguish the inherent differences between different countries. To capture the dynamic changes in the host advantage, host-related variables are set, including whether it is the host of the current session, whether it is the host of the previous session, and whether it is the host of the next session. Then, to reflect the traditional and potential advantages of different countries in a certain sport, this paper adds two key historical performance variables based on the year of the Olympic Games: the number of gold medals and the total number of medals won by the country in the specific sport in previous sessions. Finally, considering that countries that have not yet won medals also have potential for development, this paper adds variables to identify whether the country is winning medals for the first time in each Olympic Games, which helps to capture the

rising trend of emerging countries in Olympic sports.

Based on detailed "sport-athlete" data, this paper constructs data at the "country-sport" level. The data covers Olympic performance indicators such as the number of gold medals and medals. For some sports that are team events and have tied medals, this paper adjusts the actual number of medals based on the actual number of gold medals and total medals won by each country in each Olympic Games, ensuring the accuracy and consistency of the data, and performs box-cox transformation on the data to make it more in line with statistical assumptions. Secondly, the teams of various countries are processed. In 2024, Russia, Belarus and other countries were banned from participating in the Summer Olympics, so they were removed from the dataset. At the same time, based on the host information from 2004 to 2028, the national feature variables of various countries were supplemented, and a dummy variable representing the host was established. Considering that there is a strong correlation between the performance of two adjacent Olympic Games, and the reference value of distant competition records for current medal prediction is low, this paper only selects the data of the last five Olympic Games as the training sample to reduce the interference of historical data. Finally, the data of the 2008, 2012, 2016, 2020 (actually held in 2021) and 2024 Olympic Games were selected.

For sport data, the sport settings of the 2024 Olympic Games were taken as the standard, and the sports that had changed in history were processed. The 1 new sport added in 2024 (breakdancing) and the 5 new sports added in 2020 (skateboarding, surfing, climbing, baseball/softball and karate) were removed because these sports lack sufficient historical data support. Golf, baseball, trampoline gymnastics and other sports with missing years were also removed because they are difficult to effectively incorporate into model analysis. After the above processing, this paper finally obtained the competition records in the 2008-2024 Olympic Games, providing a high-quality data basis for model training.

### 4.3 Model Analysis and Results

The establishment process of the random forest model is as follows:

1. **Train-Test Split:** Divide the training set and the test set, divide the data set into training set and test set, using the proportion of 80% training set and 20% test set;
2. **Random Forest Model Construction:** Build the random forest model, predict the number of gold medals and the total number of medals respectively, and adjust the model parameters to optimize the model performance, check whether the model is over-fitting or under-fitting, and finally set the number of subtrees to 400, use the mean square error to measure the quality of splitting, and set the maximum number of explanatory variables to 1/3 of the total number of explanatory variables;
3. **Prediction for the 2028 Olympics:** Use the trained random forest model to predict the data of the 2028 Olympic Games, and output the predicted number of gold medals and the total number of medals for each country;
4. **Analysis of First-Time Medal-Winning Countries:** Add a new target variable, that

is, the number of countries that break through 0 medals each year, train the random forest model to estimate the number of new countries that may appear in 2028;

5. **Calculation of  $\phi_i(f(x))$ :** The SHAP values are calculated to assess the independent impact of each feature and the interaction between features, providing a deeper level of interpretation.

#### 4.3.1 Medal Table Forecast for the 2028 Summer Olympics in Los Angeles, USA

Based on the sub-event fitting model, the prediction results of the 2028 Summer Olympic Games in Los Angeles, USA are summed up. In this paper, the top 30 countries in the total number of MEDALS in the prediction results are selected and displayed, as shown in the table. The mean square error of the gold medal number prediction model is 0.093, and the mean square error of the total medal number prediction model is 0.338, indicating that the model fits well.

Table 2: Predictions for Medal Counts in Future Olympics

Rank	NOC	Total	Gold	Rank	NOC	Total	Gold	Rank	NOC	Total	Gold
1	USA	102	37	11	CAN	23	6	21	DEN	10	2
2	FRA	75	24	12	JPN	19	3	22	UZB	10	6
3	CHN	70	35	13	NZL	17	6	23	KEN	10	3
4	AUS	66	12	14	HUN	15	5	24	BRA	9	2
5	GBR	43	11	15	LTU	13	0	25	POL	9	1
6	GER	42	22	16	ROC	13	4	26	BEL	8	3
7	NED	38	22	17	UKR	12	3	27	SUI	8	2
8	ITA	34	9	18	IRI	11	3	28	CZE	8	3
9	KOR	27	11	19	CUB	10	4	29	CRO	8	3
10	ESP	24	3	20	SWE	10	3	30	JAM	7	2

To more intuitively reflect the progress and decline of various countries, a histogram was used to show the ranking changes of the top 30 countries in the 2024 Paris Olympics at the 2028 Los Angeles Olympics, as shown in Figure 4. It shows countries with rising rankings such as Ukraine and Denmark, countries with declining rankings such as Japan and Ireland, and a stable group of countries including the USA and China.

#### 4.3.2 Predicted number of countries with zero MEDALS by 2028

Based on the model fitting the number of countries breaking through 0 medals, it is predicted that 6 countries will break through at the 2028 Summer Olympics in Los Angeles, USA. The mean square error of the prediction model is 32.40.

#### 4.3.3 Analysis of Project and Country Features

Based on the random forest model built earlier, the importance of each variable was calculated using the SHAP method. First, for each sample observation, its SHAP value

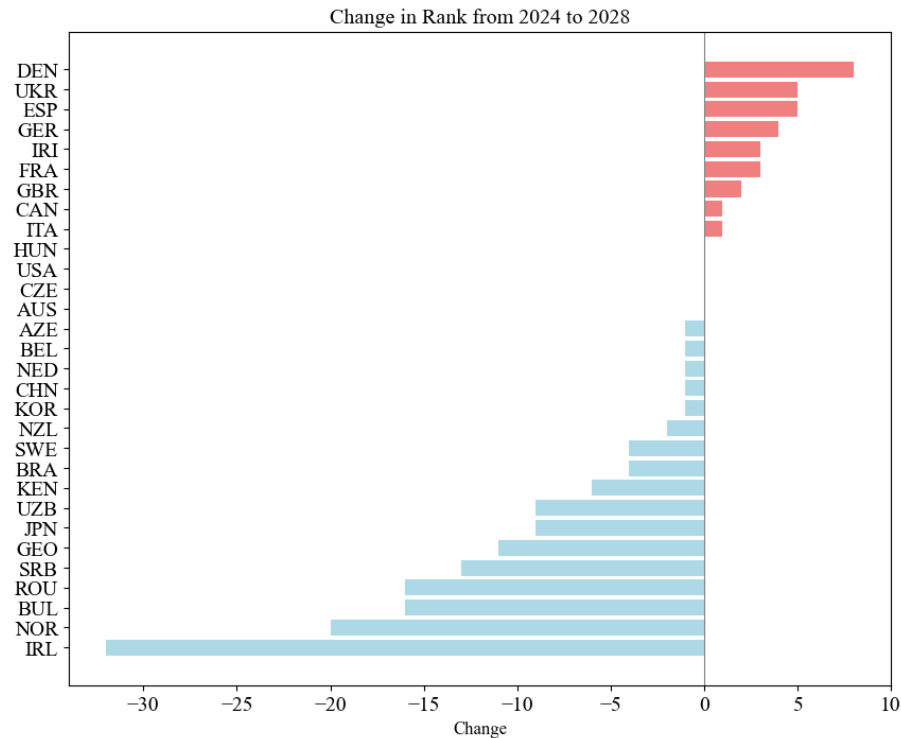


Figure 4: Medal situation of the top 30 countries

was calculated. It is worth noting that the figure shows the average value, and the SHAP value of the dummy variable has the characteristic of sparsity, that is, only a few are not zero, so the value is small. The top ten variables in terms of importance measured by the mean value are listed, as shown in Figure 5, and the following conclusions are drawn:

- **Country Dummy Variables:** Among the country dummy variables, Germany, the United Kingdom, and South Korea have a significant impact on model prediction; in Figure 5a on the right, the SHAP value distribution of the United Kingdom is wide, indicating that Germany has different impacts on model output under different feature values.
- **Sport Dummy Variables:** Among the sport dummy variables, athletics, boxing, swimming, and other sports also have a great impact on model prediction; the SHAP value of athletics is the highest, indicating that athletics has the greatest impact on model output. In Figure 5b on the right, the wide distribution of the SHAP value of athletics indicates that it has different impacts on model output under different feature values.

For each country, athletics, as a basic major event, often significantly affects the overall performance of a country; boxing, as a popular and competitive sport, may have a greater impact on model prediction; swimming has a relatively high popularity, but its performance may be affected by various factors, such as the personal state of athletes and

the competition environment, which may lead to its prediction stability in the model not being as good as other sports.

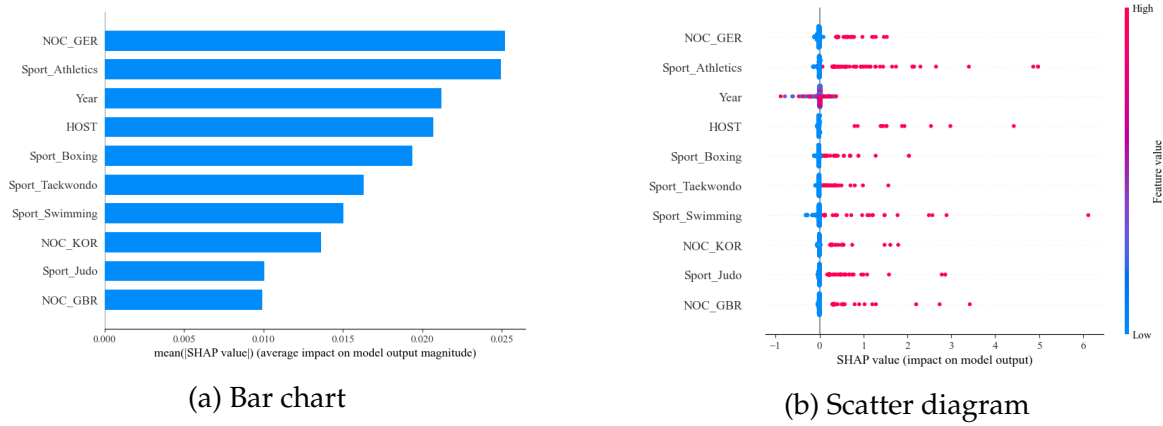


Figure 5: Features of the top ten SHAP values

In the sports where they have an advantage, most countries' total number of medals basically remains stable, and they have a high contribution value to the final medal count. To quantify this advantage, considering the host advantage and historical performance, we selected the top 18 sport-country combinations with the highest interaction effect values and listed them in Table 3. It can be seen that in the combinations with higher interaction values, most countries have traditional advantages in the corresponding sports, which are important sports for them to improve their medal performance, such as swimming in the United States and table tennis in China, etc., which verifies our view.

Table 3: Top Performing Countries in Various Sports Based on Shapley Values

Sport	NOC	SHAP	Sport	NOC	SHAP
Gymnastics	Belarus	0.37	Sailing	Great Britain	0.36
Swimming	USA	0.50	Judo	Uzbekistan	0.36
Athletics	USA	0.44	Cycling	Great Britain	0.33
	Belarus	0.32	Diving	China	0.33
Weightlifting	Kazakhstan	0.41		Germany	0.31
	Georgia	0.36	Belarus	Kazakhstan	0.32
Archery	South Korea	0.41		Ukraine	0.32
Taekwondo	South Korea	0.40	Table Tennis	China	0.31
Canoeing	Czech Republic	0.37	Equestrian	Germany	0.30

## 5 Model Construction and Problem 2

### 5.1 Algorithm Design

In statistics, linear regression is a regression analysis that models the relationship between one or more independent variables and a dependent variable using a linear regression equation, which is a least squares function. In this paper, we perform time series

analysis, using the increase or decrease in the number of gold medals to measure the impact of the coach effect, where an increase represents a positive impact and a decrease represents a negative impact. We use a linear regression model to evaluate the coach effect:

$$y_i(t) = \alpha_0 + \alpha_1 X_i(t) + \alpha_2 C_i(t) + \epsilon_i(t) \quad (2)$$

The coach influence index is the marginal gain brought by the coach. For each country  $i$  that coach  $j$  has coached, we define the coach's influence index  $C_j(t)$  as the growth rate of gold medals in the project coached by coach  $j$  in country  $i$  in year  $t$ .

## 5.2 Data Analysis and Variable Selection

To accurately assess the historical achievements and competitive strength of countries in specific sports, this paper, based on "athlete-sport" data, screened out the top ten countries with the most award-winning athletes in the Olympic Games over the years, and counted the sports in which these countries' athletes won the most gold and silver medals, as shown in Table 4. The more athletes a country has won medals in a sport, the more likely that sport is to be the country's strength, and the athletes or coaches of that country are more likely to have the potential "great coach" qualities. Based on this, this paper selects the United States, China, and the United Kingdom for the analysis of investable "great coach" projects.

Table 4: Medal Counts by Sport for Selected Countries

NOC	Sport	Gold	Sport	Silver
USA	Swimming	698	Athletics	350
CHN	Diving	78	Swimming	47
GBR	Rowing	124	Athletics	150
FRA	Fencing	135	Fencing	107
AUS	Swimming	155	Swimming	184
JPN	Gymnastics	65	Gymnastics	47
GER	Rowing	106	Rowing	75
RUS	Synchronized Swimming	54	Athletics	45
KOR	Archery	68	Handball	57
ITA	Fencing	155	Fencing	151

For priority investment projects, this paper screens from two aspects: development prospects and coaching resources. First, the paper counts the 15 sports with the least number of gold medals won by athletes from each country in the Olympic Games, which have the potential to win more medals and show good development prospects, and can increase the chances of winning awards due to the addition of coaches. Second, by evaluating the coaching strength of each country based on the countries' strong sports (Table 4) and the sports with high contribution to the number of medals (Figure 5), the future

development potential of specific sports is determined, thus providing a basis for strategic investment. This paper selects three pairs of countries with excellent coaches and countries without excellent coaches but with investability: France and the United States in fencing, South Korea and China in archery, and China and France in diving. Collect the number of medals won by each group of countries over a certain period of time, as shown in Figure 6.

To fully reflect the "great coach" effect, this question only evaluates the coach's influence within the range below the upper limit of gold medals in the project. Considering the Olympic cycle, historical cumulative effect, and coach background, data from various projects between 2000 and 2020 were finally selected for linear regression prediction.

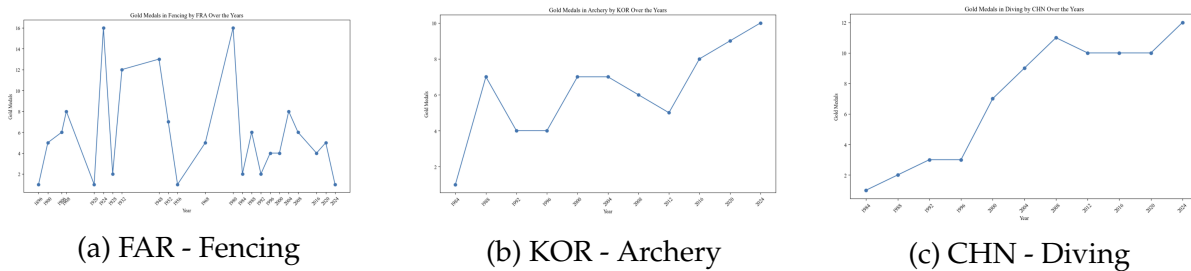


Figure 6: Gold Medals by Sport for Selected Countries

### 5.3 Model Analysis and Results

The process of establishing the linear regression model is as follows:

1. **Data preprocessing:** Divide the dataset into training and test sets, using a ratio of 80% for the training set and 20% for the test set.
2. **Quantification of coach influence:** To quantify the impact of coaches on the number of medals, first calculate the gold medal growth rates for French fencing, Korean archery, and Chinese diving to comprehensively reflect the influence of coaches.
3. **Model construction:** Build a linear regression model to predict the number of medals, and adjust the model parameters to optimize model performance, checking whether the model is overfitting or underfitting.
4. **Prediction results:** Use the trained random forest model to predict the data of the Olympic Games from 2000 to 2020, and output the predicted number of gold medals.

Comparing the prediction results with the actual data (as shown in Figure 7) to verify the prediction effect of the model. Finally, this paper draws the following conclusions:

- Countries with excellent coaches show a significant upward trend in the number of medals.

- The data shows that athletes with excellent coaches have an accelerated increase in the number of medals from 2000 to 2020, while those without excellent coaches have a slower increase.
- There is a clear positive correlation between coaching experience and the number of medals, with a significant increase in the number of medals obtained as coaching experience increases.

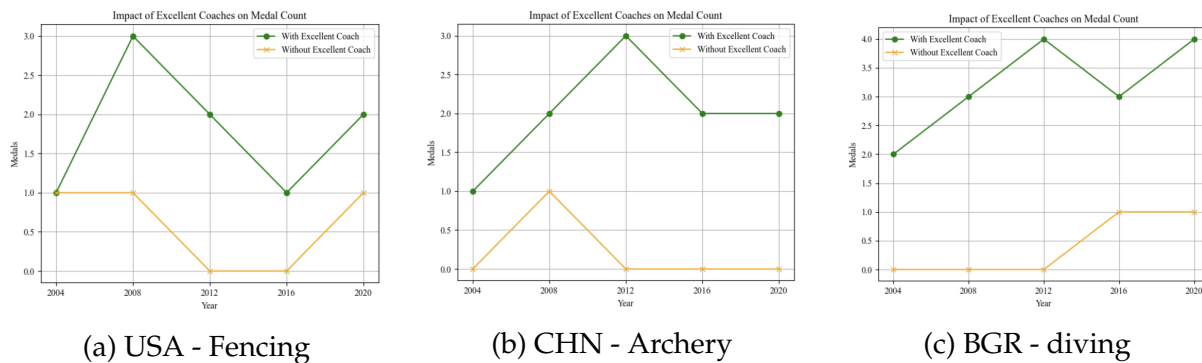


Figure 7: The Impact of Coach Effect on Selected Countries

## 5.4 Investment Recommendations

### 5.4.1 Investment Recommendations for the United States

In Figure 6a, we plotted the trend of gold medals won in French fencing from 1896 to 2024. The data shows that France had outstanding results in fencing between 1920 and 2004, particularly with abnormally high gold medal counts in 1924 and 1980. This may be related to the famous athlete and later coach, Yuka Aubry<sup>[5]</sup>.

During his athletic career, Yuka Aubry won the gold medal in the men's team épée at the 2004 Athens Olympics and secured two silver medals in both individual and team events at the 2000 Sydney Olympics. He also had an impressive record at the World Championships, accumulating 3 golds, 2 silvers, and 1 bronze, including winning the men's individual épée championship at the 1998 World Championships in La Chaux-de-Fonds, Switzerland. After retiring, Yuka Aubry left the French team in September 2016 to coach the Chinese épée team.

Following his departure, French athletes' medal counts in fencing fluctuated over the next few decades, showing an overall downward trend; meanwhile, the Chinese team began to rise, with the women's épée team consistently reaching the podium at the World Championships from 2017 to 2019, even winning gold in 2019. This was the third time the Chinese women's épée team captured the World Championship gold medal, following victories in 2006 and 2015.

Given Yuka Aubry's status as an outstanding coach and athlete, with exceptional professional skills, international perspective, and cross-cultural communication abilities, his



appointment to the U.S. fencing team would have a profound impact on American fencing. His successful experiences on the international stage would greatly enhance the overall strength and competitiveness of the U.S. team. Yuka's extensive competition experience and tactical wisdom would help American athletes reach new heights both technically and psychologically.

### 5.4.2 Investment Recommendations for China

According to Figure 6b, South Korea's archery performance saw a significant improvement in 2012, largely thanks to the guidance of excellent coaches. For instance, Park Cha-sun, who became the coach of the South Korean archery team in 2016, led the team to win four gold medals at the previous Tokyo Olympics, with his protégé An San winning three golds individually<sup>[4]</sup>.

The strength of South Korean archery has attracted attention from around the world, with many countries inviting South Korean coaches to lead their teams (reference material 2). For example, Coach Kwon Yong-seok led the Chinese team, Coach Woo Sun-taek coached the French archery team, Coach Park Cha-sun guided the Vietnamese team, Coach Lee Kyung-su led the Indonesian team, and Coach Kim Sang-hun coached the Japanese men's team. Under their guidance, various national archery teams have achieved commendable results, showcasing a thriving archery scene globally.

Bringing in South Korean coach Lee Wang-woo to the Chinese archery team (reference material 3) is expected to have a profound impact on the sport in China. As an experienced coach who has developed seven Olympic champions, his inclusion would not only provide high-level technical guidance but also introduce systematic training philosophies and scientific training methods to the Chinese team.

Firstly, Lee Wang-woo's extensive experience and successful track record will offer valuable learning opportunities for Chinese archers. Secondly, his broad network and influence in international competitions will facilitate more opportunities for international exchange and collaboration for the Chinese archery team. He possesses deep technical and tactical knowledge in archery, enabling him to develop personalized training plans tailored to each athlete's strengths.

Under Lee Wang-woo's guidance, the Chinese archery team is expected to achieve better results in future international competitions, particularly in major events like the Olympics and World Championships, aiming for more gold medals. As China's archery performance improves, it will also enhance China's standing in the international archery community, positioning it as a strong competitor globally.

### 5.4.3 Investment Recommendations for Great Britain

In Figure 6c, we plotted the trend of gold medals won by the Chinese diving team from 1896 to 2024. The data shows that there was a turning point in China's diving performance in 1996, after which the Chinese diving team consistently improved and won a significant number of gold medals. This success is not solely attributable to the athletes' talent but is also closely linked to the dedicated coaching of the Chinese diving coaches.

It is understood that Zhou Jihong graduated in 1990 and has served as the head coach of the Chinese national diving team since 2000. Under her leadership, the team won 5, 6, and 6 gold medals at the 2000, 2004, and 2012 Olympics, respectively, and secured 7 gold medals at both the 2008 and 2012 Olympics. As a result, the Chinese national diving team has been dubbed the "Dream Team."

If the UK were to invite Zhou Jihong to coach their diving team, it would undoubtedly have a profound impact on the sport in the UK.

As a highly experienced and accomplished coach, her involvement would bring significant technical improvements and systematic training methods to the GBR diving team. Zhou Jihong's extensive experience and successful track record would provide British athletes with unique learning opportunities. Under her guidance, the Chinese diving team has repeatedly won gold medals in international competitions, and her successful training methods and competition strategies would help British athletes improve both technically and mentally. .

## 6 Model Construction and Problem 3

### 6.1 Host effect

To better reflect the host effect, the main effect values of host-related features were extracted, as shown in Figure 8.

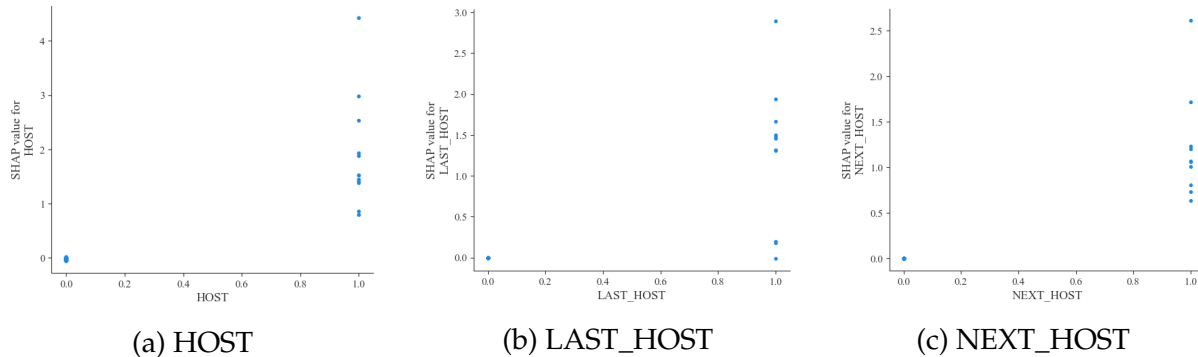


Figure 8: SHAP Value of the Host Country

Figure 8 shows the higher contribution of the host, with *HOST* being the first, followed by *NEXT\_HOST*, and then *LAST\_HOST*. Among them, having or about to obtain the host status has a certain positive effect on the number of medals, but this effect is not decisive. Studies by Crusllia et al. have shown that if the socio-economic factors remain unchanged, the host country of the Olympic Games will not win more medals because of the host status. However, the host country may achieve better results in some events due to factors such as home-field advantage and audience support.

## 6.2 Gender Ratio

Based on the random forest built for Problem 1, we comprehensively selected the top 10 sports with higher feature values and more participants, which are judo, hockey, canoeing, football, rowing, shooting, cycling, gymnastics, swimming, and athletics, and visualized the gender ratio of each sport from 2012 to 2024.

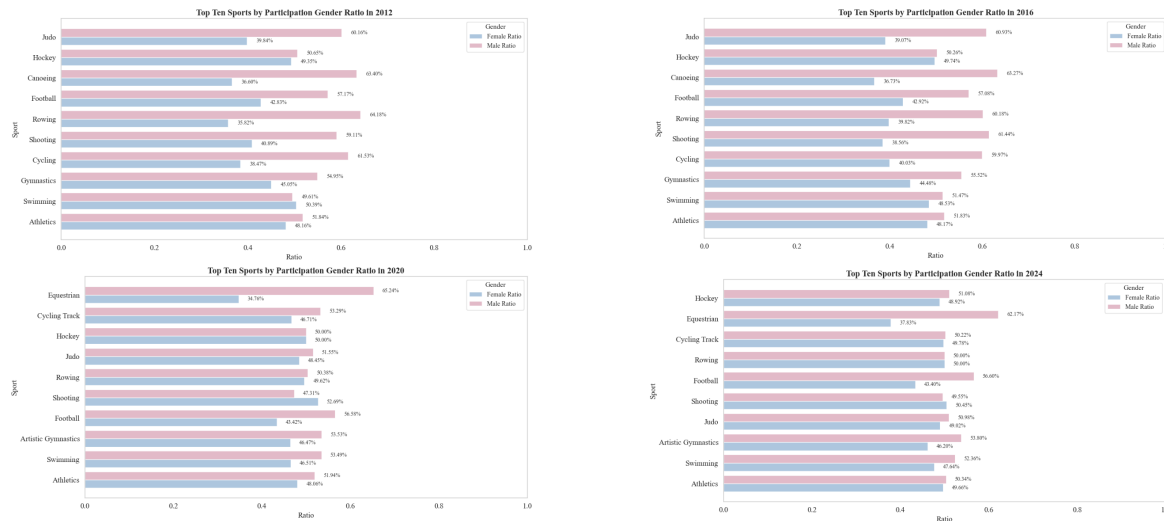


Figure 9: Gender Ratio Data for between 2012 and 2024

### Gender Ratio Data for 2024

- **Positive Correlation:** In several sports (such as equestrian and swimming), the proportion of female participants is relatively high, indicating a gradual increase in women's participation in these events, aligning with societal emphasis on gender equality.
- **Non-Linear Relationship:** In certain sports (such as hockey and soccer), male participation still dominates, suggesting that gender differences remain evident in some traditional sports, potentially linked to cultural and historical factors.

### Gender Ratio Data for 2020

- **Ongoing Gender Differences:** Although some sports have seen an increase in female participation (such as equestrian), overall, male participation remains higher in most events, indicating that gender balance still needs to be strengthened in certain areas.

### Gender Ratio Data for 2016

- **Gender Ratio Changes:** Data from 2016 showed that in multiple sports, the participation rate of women was generally lower than that of men. For example, the participation rates of women in swimming and athletics were 48.57% and 40.92%, respectively. This indicates that men still dominate in these sports.

- **Historical Trends:** Compared to data from 2012 and 2020, the 2016 chart shows that female participation rates remain low, reflecting a slow advancement of gender equality in some traditional sports.
- **Event Characteristics:** Certain sports (such as judo and rowing) exhibit closer gender ratios, possibly related to the physical requirements and technical characteristics of these events. A relatively high female participation rate may suggest greater inclusivity in these sports.
- **Cultural Factors:** Some traditional sports (such as hockey and football) still show significant gender differences, which may be related to the cultural and social perceptions and encouragement of participation in these sports.

### Gender Ratio Data for 2012

- **Historical Trends:** Compared to the 2024 chart, the 2012 chart shows generally lower female participation rates, indicating a significant change in gender participation over the past decade.

## 6.3 Recommendations

- **Strategic Adjustments and Optimization**
  - **Diversified Goals:** National Olympic committees should consider adjusting their strategies to not only focus on gold medals but also emphasize the competition for silver and bronze medals, especially in highly technical and competitive events.
  - **Project Prioritization:** Identify projects with potential for winning silver and bronze medals and provide athletes with more support and training resources to enhance overall medal counts.
- **Targeted Training and Support**
  - **Specialized Training Plans:** Develop specialized training programs aimed at different types of medals, helping athletes progress on all fronts, particularly in the competitions for silver and bronze.
  - **Coaching Team Development:** Strengthen the coaching team to ensure coaches can develop appropriate training and competition strategies tailored to the characteristics of different events, enhancing athletes' overall performance.
- **Continuous Monitoring and Feedback**
  - **Data-Driven Decision Making:** Establish a regular data analysis mechanism to monitor the acquisition of various types of medals, allowing for timely adjustments to strategies and resource allocation.

- **Feedback Mechanism:** Continuously optimize training and competition strategies based on feedback from athletes and coaches to maintain competitiveness in the pursuit of all types of medals.
- **Public Promotion and Support**
  - **Enhancing Public Awareness:** Increase public awareness of the importance of silver and bronze medals through promotional activities, encouraging societal attention and support for all types of medal competitions.
  - **Incentive Measures:** Develop incentive measures for athletes who excel in winning silver and bronze medals to boost their motivation to compete and improve overall performance.

## 7 Evaluation and Generalization

### 7.1 Advantages of the Study

- **Comprehensiveness:** The model integrates multiple key factors, such as competition events, host country effects, and excellent coaches, providing a comprehensive depiction of the major factors influencing medal distribution. Through multidimensional analysis of these variables, the model can deeply explore the potential of different countries in various events, helping to understand their competitive strengths and weaknesses.
- **Flexibility:** Based on regression analysis and machine learning methods, the model can flexibly adjust parameters or expand variables to meet the predictive needs of different countries and scenarios. This flexibility allows the model to quickly respond to changing environments and data, ensuring high predictive accuracy under varying conditions and times.
- **Practicality:** Through visual analysis and quantitative predictions, the model not only provides precise estimates of medal counts but also offers specific investment recommendations for potential countries in various sports. This practicality enables decision-makers to allocate resources more effectively and formulate targeted training and development strategies, thereby enhancing the overall competitive level of the country.
- **Innovation:** Building on traditional medal prediction, the model introduces the "exceptional coach effect," providing national Olympic committees with meaningful strategic references. This innovation not only considers the individual capabilities of athletes but also emphasizes the crucial role of coaches in athlete development and competitive performance, offering new perspectives for strategy formulation.
- **Scalability:** The model is based on publicly available data, making it versatile for use in predicting and analyzing medal counts in other international events. This scalability provides extensive application potential, supporting performance analysis of different countries and regions across various international competitions.

Additionally, the random forest method has broad applications in social science verification analysis; for instance, Athey et al. have applied it to strategy research, and Li Yong et al. have found its performance particularly outstanding in economic and financial predictions.

## 7.2 Limitations of the Study

- **Unconsidered Factors:** In practical applications, political factors, cultural background, education levels, and other elements may also significantly impact medal distribution. This paper has not comprehensively incorporated these complex variables, which may affect the accuracy of the model's predictions. Therefore, future research needs to consider these external factors to enhance the model's comprehensiveness and accuracy.
- **Limited Applicability:** The model is built on existing Olympic medal data and has not yet been validated in other contexts (such as the Winter Olympics or regional events), which may limit its applicability. This restricts the model's general applicability, necessitating testing and adjustments in different types of events in the future to verify its suitability.

## 7.3 Application of the Study

The model developed in this study has several potential applications and areas for future development:

- **Application to Different Sports:** The model can be extended to other international sports events (such as the Winter Olympics, Asian Games, World Championships, etc.) by adjusting specific variables (such as competition events, number of participants, etc.) to assess the medal distribution patterns in these events. This will help achieve a more comprehensive understanding of the characteristics of different events and promote the optimization of countries' preparation strategies for various competitions.
- **Expansion of Variable Dimensions:** By incorporating more external data (such as education investment, sports infrastructure development, athlete psychological qualities, etc.), the model's predictive capabilities can be further optimized, allowing for the exploration of additional potential influencing factors. This expansion will enhance the model's depth and breadth, making it more practical and adaptable to the ever-changing sports environment.
- **Policy Formulation and Resource Allocation:** The results of the model can provide data support for the formulation of sports policies, helping governments and sports organizations understand which areas require more investment and attention. By identifying potential projects and athletes, relevant departments can allocate resources more effectively, thereby enhancing the country's competitiveness in international events.

## References

- [1] ARCHER K J, KIMES R V. Empirical characterization of random forest variable importance measures[J/OL]. *Computational Statistics & Data Analysis*, 2008, 52(4): 2249-2260. <https://www.sciencedirect.com/science/article/pii/S0167947307003076>. DOI: <https://doi.org/10.1016/j.csda.2007.08.015>.
- [2] HUIMIN S, DONGYING Z, YONGHUI Z. Can Olympic Medals Be Predicted? A Perspective Based on Explainable Machine Learning[J/OL]. *Journal of Shanghai University of Sport*, 2024, 48(4): 26-36. <https://qikan.cqvip.com/Qikan/Article/Detail?id=7111825433>.
- [3] LUNDBERG S M, ERION G G, LEE S I. Consistent Individualized Feature Attribution for Tree Ensembles[J/OL]. *ArXiv*, 2018, abs/1802.03888. <https://api.semanticscholar.org/CorpusID:3626364>.
- [4] Of SPORT OF CHINA G A. Chinese Archery Team Adjusts Vigorously for World Championships[EB/OL]. General Administration of Sport of China. 2024. <https://www.sport.gov.cn/n20001280/n20745751/n20767274/c21836584/content.html>.
- [5] Olympics.com. Hugues Obry Discusses the Plan and Aim of the Chinese Epee Fencing Team for the Tokyo Olympics[EB/OL]. 2024. <https://www.olympics.com/zh/news/hugues-obry-talks-about-the-plan-and-aim-of-chinese-epee-fencing-team>.
- [6] Olympics.com. LANG Ping[EB/OL]. 2025. <https://olympics.com/en/athletes/ping-lang>.
- [7] SCHLEMBACH C, SCHMIDT S L, SCHREYER D, et al. Forecasting the Olympic medal distribution – A socioeconomic machine learning model[J/OL]. *Technological Forecasting and Social Change*, 2022, 175: 121314. <https://www.sciencedirect.com/science/article/pii/S0040162521007459>. DOI: <https://doi.org/10.1016/j.techfore.2021.121314>.

## Appendices

### Report on use of AI

1. Kimi (Jan 28, 2025 version)

Query1: Can you explain the difference between SHAP values and SHAP interaction values?

Output: SHAP values (SHapley Additive exPlanations) are a method for interpreting machine learning model predictions, based on the Shapley values from cooperative game theory, used to explain the contribution of each feature to the model's predictions. SHAP interaction values further consider the impact of interactions between features on the model's predictions. In simple terms, SHAP values focus on

the contribution of individual features, while SHAP interaction values focus on the interactions between pairs of features.

2. (Jan 28, 2025 version) Query1: Finally, dataset's 6558090 rows  $\times$  938 columns too large.

Output: I understand that the shape is 6,558,090 rows  $\times$  938 columns, which is indeed a very large dataset. Datasets of this scale can easily exceed the system's memory limits when processed in memory, leading to a MemoryError.

Problem Analysis

Large Data Scale: A dataset of 6,558,090 rows  $\times$  938 columns occupies a significant amount of memory, especially when the data type is float32. Each row of data occupies approximately 3.75 KB ( $938 \times 4$  bytes), and the entire dataset occupies about 24.6 GB of memory. This clearly exceeds the memory capacity of most ordinary computers.

Possible Reasons:

The data generation process may have produced too many combinations, leading to an explosive increase in data volume.

The model input features may contain too many columns, especially after one-hot encoding. Solutions

Batch Processing of Data:

Divide the data into multiple small batches, predict each batch individually, and then merge the results. This method can effectively reduce memory usage.

Reduce the Number of Features:

If there are too many features, you can try the following methods to reduce the number of features:

Feature Selection: Choose only the features that are most helpful for model prediction. Dimensionality Reduction: Use Principal Component Analysis (PCA) or other dimensionality reduction techniques to reduce the number of features.

Optimize Memory Usage:

If the data volume is still large, you can try the following methods to optimize memory usage:

Use More Efficient Data Types: For example, replace float32 with float16, and int32 with int16.

Reduce Data Precision: If the data precision requirement is not high, you can convert the data to a lower precision type.

Summary

By processing data in batches, reducing the number of features, and optimizing memory usage, you can effectively solve the MemoryError problem.



## Code

## Python source code:RandomForestRegressor

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error,
    mean_absolute_error
import shap
import matplotlib.pyplot as plt
import itertools
from tqdm import tqdm
import lightgbm as lgb
from skopt import BayesSearchCV
data = pd.read_csv('output_new\competition_df_2008.csv')
df = data.drop(columns=["Event"])
df['Gold'] = df['Gold'].apply(lambda x: 1 if x != 0 else 0)
new_countries_per_year = df[df['IS_NEW'] == 1].groupby('Year')['NOC'].unique()
merge_columns = ["Year", "Sport", "NOC", "HOST", "LAST_HOST", "NEXT_HOST", "IS_NEW"]
df = df.groupby(merge_columns, as_index=False).agg({"Bronze": "sum", "Gold": "sum", "Silver": "sum", "total": "sum"})
data=df.drop(columns=['Silver', 'Bronze'])

data = data[~data['NOC'].isin(['RUS', 'BLR'])]
data
df_2024 = data[data['Year'] == 2016]

total_by_country_2024 = df_2024.groupby('NOC')['total'].sum().reset_index()
total_by_country_2024_sorted = total_by_country_2024.sort_values(by='total', ascending=False)

print(total_by_country_2024_sorted.head(20))
unique_years = df['Year'].unique()
unique_countries = df['NOC'].unique()
unique_sports = df['Sport'].unique()

all_combinations = list(itertools.product(unique_years, unique_countries, unique_sports))

complete_df = pd.DataFrame(all_combinations, columns=['Year', 'NOC', 'Sport'])
```

```

    NOC', 'Sport'])
complete_df['HOST'] = 0
complete_df['LAST_HOST'] = 0
complete_df['NEXT_HOST'] = 0
complete_df['IS_NEW'] = 0
complete_df['Gold'] = 0
complete_df['total'] = 0

final_df = pd.merge(complete_df, data, on=['Year', 'NOC', 'Sport'],
                    how='left')

final_df['HOST'] = final_df['HOST_y'].fillna(final_df['HOST_x'])
final_df['LAST_HOST'] = final_df['LAST_HOST_y'].fillna(final_df['LAST_HOST_x'])
final_df['NEXT_HOST'] = final_df['NEXT_HOST_y'].fillna(final_df['NEXT_HOST_x'])
final_df['IS_NEW'] = final_df['IS_NEW_y'].fillna(final_df['IS_NEW_x'])
final_df['Gold'] = final_df['Gold_y'].fillna(final_df['Gold_x'])
final_df['total'] = final_df['total_y'].fillna(final_df['total_x'])

final_df = final_df.drop(columns=['HOST_x', 'HOST_y', 'LAST_HOST_x', 'LAST_HOST_y',
                                'NEXT_HOST_x', 'NEXT_HOST_y', 'IS_NEW_x', 'IS_NEW_y',
                                'Gold_x', 'Gold_y', 'total_x', 'total_y'])

print(final_df)

final_df['new_num'] = final_df['Year'].map(new_countries_per_year)
final_df.to_csv('final_test_data.csv')
expanded_df = pd.get_dummies(final_df, columns=['NOC', 'Sport'], drop_first=True)
data = expanded_df.replace({True: 1, False: 0})
data
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import randint, uniform
features = data.drop(columns=['Gold', 'total', 'new_num']).copy()
targets = data[['Gold', 'total', 'new_num']]

```

```
X_train, X_test, y_train, y_test = train_test_split(features,
    targets, test_size=0.2, random_state=42)

rf_gold = RandomForestRegressor(n_estimators=400, random_state
    =42)
rf_total = RandomForestRegressor(n_estimators=400, random_state
    =42)
rf_new = RandomForestRegressor(n_estimators=100, random_state=42)

print("Training models...")
rf_gold.fit(X_train, y_train['Gold'])
rf_total.fit(X_train, y_train['total'])
rf_new.fit(X_train, y_train['new_num'])

print("Making predictions...")
y_pred_gold = rf_gold.predict(X_test)
y_pred_total = rf_total.predict(X_test)
y_pred_new = rf_total.predict(X_test)

mse_gold = mean_squared_error(y_test['Gold'], y_pred_gold)
mse_total = mean_squared_error(y_test['total'], y_pred_total)
mse_new = mean_squared_error(y_test['new_num'], y_pred_new)
print(f"MSE for Gold: {mse_gold}, MSE for Total: {mse_total}, MSE
    for New_num: {mse_new}")

combinations = list(itertools.product(unique_countries,
    unique_sports))
data_2028 = pd.DataFrame(combinations, columns=['NOC', 'Sport'])

data_2028['Year'] = 2028
data_2028['HOST'] = 0
data_2028['LAST_HOST'] = 0
data_2028['NEXT_HOST'] = 0
data_2028['IS_NEW'] = 0

data_2028.loc[data_2028['NOC'] == 'USA', 'HOST'] = 1
data_2028.loc[data_2028['NOC'] == 'FRA', 'LAST_HOST'] = 1
data_2028.loc[data_2028['NOC'] == 'AUS', 'NEXT_HOST'] = 1

data_2028 = pd.get_dummies(data_2028, columns=['NOC', 'Sport'],
    drop_first=True)
data_2028 = data_2028.replace({True: 1, False: 0})
data_2028

batch_size = 10000
```

```
predictions = []
data_2028_gold=data_2028.copy()
for i in tqdm(range(0, len(data_2028), batch_size), desc="
Predicting batches"):
    batch_data = data_2028.iloc[i:i + batch_size]
    batch_predictions = rf_total.predict(batch_data)
    predictions.extend(batch_predictions)

data_2028['predictions'] = predictions

predictions_gold = []
for i in tqdm(range(0, len(data_2028_gold), batch_size), desc="
Predicting batches"):
    batch_data = data_2028_gold.iloc[i:i + batch_size]
    batch_predictions_gold = rf_gold.predict(batch_data)
    predictions_gold.extend(batch_predictions_gold)
data_2028['predictions_gold'] = predictions_gold
predictions_new = []
for i in tqdm(range(0, len(data_2028_gold), batch_size), desc="
Predicting batches"):
    batch_data = data_2028_gold.iloc[i:i + batch_size]
    batch_predictions_new = rf_new.predict(batch_data)
    predictions_new.extend(batch_predictions_new)

predictions_new

array = np.array(predictions_new)

array_non_zero = array[array != 0]

mean_value = np.mean(array_non_zero)

mean_value
NOC_columns = [col for col in data_2028.columns if col.startswith
('NOC_')]
sports_columns = [col for col in data_2028.columns if col.
startswith('Sport_')]
data_2028['NOC'] = data_2028[NOC_columns].idxmax(axis=1).str.
replace('NOC_', '')
data_2028['Sport'] = data_2028[sports_columns].idxmax(axis=1).str
.replace('Sport_', '')
data_2028.drop(columns=NOC_columns + sports_columns, inplace=True
)
print(data_2028)
total_by_country_2028 = data_2028.groupby('NOC')['predictions'].
```

```
    sum().reset_index()
gold_by_country_2028 = data_2028.groupby('NOC')['predictions_gold
    '].sum().reset_index()
result = pd.merge(total_by_country_2028, gold_by_country_2028, on
    ='NOC', how='outer')
result = result.sort_values(by='predictions', ascending=False)
print(result)
result.to_csv('total_by_country_2028_sorted.csv')
print(result)
result.to_csv('total_by_country_2028_sorted.csv')
noc_abbreviations = {
    "United States": "USA",
    "China": "CHN",
    "Japan": "JPN",
    "Australia": "AUS",
    "France": "FRA",
    "Netherlands": "NED",
    "Great Britain": "GBR",
    "South Korea": "KOR",
    "Italy": "ITA",
    "Germany": "GER",
    "New Zealand": "NZL",
    "Canada": "CAN",
    "Uzbekistan": "UZB",
    "Hungary": "HUN",
    "Spain": "ESP",
    "Sweden": "SWE",
    "Kenya": "KEN",
    "Norway": "NOR",
    "Ireland": "IRL",
    "Brazil": "BRA",
    "Iran": "IRI",
    "Ukraine": "UKR",
    "Romania": "ROU",
    "Georgia": "GEO",
    "Belgium": "BEL",
    "Bulgaria": "BUL",
    "Serbia": "SRB",
    "Czech Republic": "CZE",
    "Denmark": "DEN",
    "Azerbaijan": "AZE"
}

result['Rank_2028'] = result['predictions'].rank(ascending=False,
    method='min').astype(int)
```

```

file_path = 'five_years_medals (2).xlsx'
rank = pd.read_excel(file_path, sheet_name='Sheet1')
rank['NOC'] = rank['NOC'].map(noc_abbreviations)
rank = rank.merge(result[['NOC', 'Rank_2028']], on='NOC', how='
    left')
rank['Rank_2028'] = rank['Rank_2028_y']
rank.drop(columns=['Rank_2028_x', 'Rank_2028_y'], inplace=True)
rank['Change'] = rank['Rank_2024'] - rank['Rank_2028']
print(rank)
import matplotlib.pyplot as plt

data = rank[['NOC', 'Change']]
data = data.sort_values(by='Change')
colors = ['lightcoral' if change > 0 else 'lightblue' for change
    in data['Change']]
plt.rcParams['font.family'] = 'Times New Roman'
plt.figure(figsize=(10, 8))
plt.barh(data['NOC'], data['Change'], color=colors)
plt.axvline(0, color='grey', linewidth=0.8)
plt.title('Change in Rank from 2024 to 2028')
plt.xlabel('Change')
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.show()
explainer_total = shap.TreeExplainer(rf_total, X_test)

shap_values_total = explainer_total(X_test)

shap_values_total
shap_values_array = shap_values_total.values

X_test_data = shap_values_total.data

if isinstance(X_test_data, np.ndarray):
    X_test_data = pd.DataFrame(X_test_data, columns=X_test.
        columns)

import matplotlib.pyplot as plt

plt.rcParams['font.family'] = 'Times New Roman'
plt.rcParams['font.size'] = 12
shap.dependence_plot('HOST', shap_values_array, X_test_data,
    interaction_index=None)
shap.dependence_plot('LAST_HOST', shap_values_array, X_test_data,

```

```

        interaction_index=None)
shap.dependence_plot('NEXT_HOST', shap_values_array, X_test_data,
        interaction_index=None)
shap.summary_plot(shap_values_total, X_test, plot_type="bar",
        max_display=10)
shap.summary_plot(shap_values_total, X_test, max_display=10)

host_shap = shap_values_total[:, ['HOST', 'LAST_HOST', 'NEXT_HOST'
    ']]

shap.summary_plot(host_shap, X_test[['HOST', 'LAST_HOST', '
    NEXT_HOST']], max_display=10)

```

---

### Python source code:TopMedal

---

```

import pandas as pd
data = pd.read_excel('D:/output/filtered_athletes.xlsx')
countries = ['USA', 'CHN', 'GBR', 'FRA', 'AUS', 'JPN', 'GER', '
    RUS', 'KOR', 'ITA']
filtered_data = data[data['NOC'].isin(countries)]
filtered_data['Medal'] = filtered_data['Medal'].replace({
    'Gold': 1,
    'Silver': 2,
    'Bronze': 3,
    'No medal': 0
}).astype(int)
gold_count = filtered_data[filtered_data['Medal'] == 1].groupby([
    'NOC', 'Sport']).size().reset_index(name='Gold_Medals')
gold_max = gold_count.loc[gold_count.groupby('NOC')['Gold_Medals'
    ].idxmax()]
silver_count = filtered_data[filtered_data['Medal'] == 2].groupby(
    ['NOC', 'Sport']).size().reset_index(name='Silver_Medals')
silver_max = silver_count.loc[silver_count.groupby('NOC')['
    Silver_Medals'].idxmax()]
bronze_count = filtered_data[filtered_data['Medal'] == 3].groupby(
    ['NOC', 'Sport']).size().reset_index(name='Bronze_Medals')
bronze_max = bronze_count.loc[bronze_count.groupby('NOC')['
    Bronze_Medals'].idxmax()]
result = gold_max.merge(silver_max, on='NOC', how='outer').merge(
    bronze_max, on='NOC', how='outer')
result['NOC'] = pd.Categorical(result['NOC'], categories=
    countries, ordered=True)
result = result.sort_values('NOC')
print(result)

```

---