Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2503361

The Road to Gold: Cracking the code of Olympic Medals Prediction and the Impact of Great Coaches

Summary

The Olympic Games have long been a global stage for national competition, with medal tables as a significant indicator of a country's sporting success. To address the challenges of building robust predictive models for future Olympic medal counts and assessing the great coach effect on medal counts, we make an **Integrated Learning Model** and a **Multiple Linear Regression Model** to tackle these problems and share the insights from the models to inform country Olympic Committees.

For Task 1, an **Integrated Learning Model** is built to develop predicted medal counts for each country, both for gold and total medals, and estimate the precision of the model. First, through the **Auto-Regressive Integrated Moving Average (ARIMA) Model**, the various indicators and winning status of the athletes are analyzed and converted into medal characteristics. Then, the medal characteristics are input into the **Stacked Regression**, which belongs to the category of integrated learning algorithms. The base learner is the **Random Forest** and the meta learner is a **Linear Regression Model**. Using data analytics, we predicted the number and range of medals for all countries and predicted which countries would win the first medal ever since with an **accuracy rate of 77%**. Third, we use the **Pearson correlation coefficient** to get the most important sports for each country and how the choice of the home country affected the results.

For Task 2, the "great coach" effect is analyzed using a **Multiple Linear Regression Model**. In this model, the "great coach" effect, along with historical performance, the number of athlete participants, the intensity of medal competition, and the host effect, are independent variables, while the number of medals is the dependent variable. This approach enhances the explanatory power of the results. After performing the multiple linear regression, we obtain different β values corresponding to the various independent variables. Taking the results of volleyball as an example, the β values are **-0.1820, 4.4222, 0.0463, -0.0201, 0.3098, and 0.7351**, respectively, with $\boldsymbol{R^2 = 0.933}$. This indicates that the number of medals is expected to increase by 4.42 for each "great coach" introduced on average. Moreover, we identify three countries and sports where investment in coaching could yield high returns, quantifying the potential medal gains in **volleyball for Japan, athletics for the US, and swimming for China**.

For Task 3, we delve into the Olympic medal count prediction model, uncovering several insights: the gender factor, the efficiency of sports organizations, the attributes of the sport that affect medal competition intensity, and the different types of countries. These insights have implications for the National Olympic Committees (NOCs). NOCs should allocate resources based on model forecasts and consider sports traditions and changes to the Olympic program.

Finally, we analyze the sensitivity of the established models, evaluate their advantages and disadvantages, and propose further improvements.

Keywords: Stacked regression; ARIMA; Random forest; Integrated learning; Multiple linear regression; Great coach

Contents

1 Introduction

1.1 Background

The Olympic Games, held every four years, is one of the largest and most influential sports events, and its medal table is a key indicator of both the competitive level of athletes and the overall sports strength of a country. However, the distribution of medals has become more complex over time, with shifting competition patterns among traditional powers and the rise of emerging countries. For example, countries like China and the United States have increased their dominance in certain sports, while smaller countries and emerging economies have improved their performance through strategic investments and adjustments.

The distribution of Olympic medals is influenced by factors such as athletes' quality, national training systems, the setup of Olympic sports, the scale of the event, and the geographic advantages of the host country. Therefore, establishing a mathematical model to analyze and predict medal distribution is important for national Olympic committees and sports organizations to formulate athlete selection and training strategies and optimize resource allocation [1].

Figure 1: 2028 Olympic in Los Angles

1.2 Restatement of the Problem

Considering the background information and limiting conditions identified in the problem statements, we are supposed to address the following issues:

- Develop a model to predict the number of gold medals and the total number of medals each country might win in future Olympic Games and to assess the uncertainty and accuracy of the model's predictions. Specifically, forecast the medal standings for the 2028 Summer Olympics in Los Angeles, predicting whether individual countries will improve or not in terms of awards; for countries that have not yet won a medal, predict how many countries will win at the next Olympics; and explore the relationship between specific events and the number of medals won by a country, which sports are the most important to each country, and how the activities of the host country will affect the results.

- By analyzing the data, explore the impact of a great coach on the number of medals won. In addition, three countries were selected to identify sports in which they should consider investing in great coaches and to estimate their impact.
- Based on the model that has been developed, analyze what other unique insights about Olympic medal counts the model explains and explain how these insights inform the NOCs.

1.3 Our Work

To avoid complicated description, intuitively reflect our work process, the flow chart is shown as Figure 2.

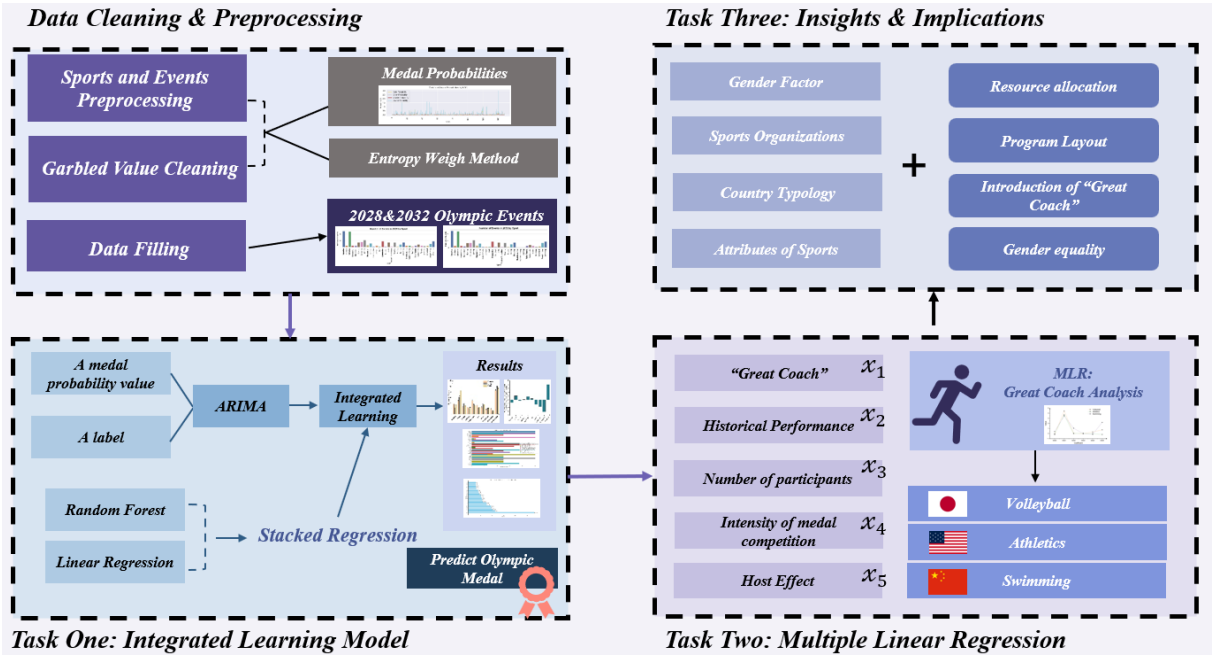


Figure 2: Flow chart of our work

2 Preparation of the Model

2.1 Assumptions and Explanations

In this section, we need to make reasonable assumptions to simplify the problem and facilitate the modelling, and each hypothesis is closely followed by its appropriate explanation.

***Assumption* 1:** The probabilities of gold, silver, and bronze medals awards are independent, which means they do not affect each other.

Explanation: While in reality there may be some correlation between the probabilities of winning the three prizes, they are assumed to be independent of each other for simplicity of the model, which helps with training and prediction via ARIMA and Random Forest models.

***Assumption* 2:** The same team members will participate in the 2028 Olympics as in the 2024 Paris Olympics.

Explanation: Considering that the title does not give information about the members of the 2028 team, the assumption is made in order to simplify the model to facilitate the estimation of the number of medals in future competitions from the performance of the athletes and thus the assumption is made.

***Assumption* 3:** The number of medals is influenced by a combination of factors, and these factors affect medals independently of each other without multicollinearity.

Explanation: A higher number of participants may increase a country's chances of winning a medal, but the effect is not the same as the direct tactical contribution of a great coach and can logically be considered independently. In reality, nonlinear relationships may occur (e.g. very high participation numbers may lead to dilution of resources and thus affect performance), but linear regression is an effective tool for preliminary analysis and simplifies the interpretation of complex relationships.

***Assumption* 4:** The effect of great coaches can only directly affect the number of medals in the sport they coach and not indirectly in other sports through rotation of resources.

Explanation: Coaches' training methods and management are primarily specific to a particular group of sports (e.g., volleyball, gymnastics). Changes in medal counts for specific sports can be attributed to coaching ability rather than to the overall national sports strategy.

2.2 Notations

Some important mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description
y_t	The observed value at time t
ϕ_i	The parameters of the autoregressive part
θ_j	The parameters of the moving average part
ϵ_t	Error term at time t
α	A constant in ARIMA mathetical expression
$\hat{y}_{t+h}^{Gold}$	The gold medal probability prediction for year $t+h$
N	The number of decision trees
$T_i(x)$	Predicted value of the ith decision tree
θ	Regression coefficient indicating the weight of each base learner in the final prediction
ϵ	Residual term used to test model robustness
β_0	Intercept denoting the number of medals predicted when all independent variables are 0
$\bar{X}$	Average value of sample
μ_0	The supposed average value
s	Standard value of sample
n	The size of the sample

other symbols instructions will be given in the text.

2.3 Data Cleaning and Preprocessing

2.3.1 Data Cleaning

The attached "summerOly_programs.csv" file, which is partially garbled and does not facilitate subsequent data processing, is here for data cleaning. Use Python to process the data, firstly, ensure that the range of the year column is correct, that is, every four years of the Olympic Games (1906 is more special, manually marked), secondly, the Discipline, Code and Sport three columns will be converted to a string type, and finally, the null value (NaN) in the data is filled with a value of zero. The "data_dictionary.csv" file is also garbled and cleaned accordingly, as shown in Table 2.

Table 2: Garbled Value Cleaning of summerOly_programs.csv

Status	Column	Sport	1900	1908
Before	10	Baseballnd Softball	?0	?0
After	10	Baseball and Softball	0	0

Other parts of the processing are not shown.

At the same time, data for 1940 and 1944 have been excluded from consideration because of the impact of the World War II.

2.3.2 Sports and Events Preprocessing

For the processed summerOly_programs.csv file, each sport category corresponds to several disciplines, and in order to simplify the data, the number of programs in each discipline was merged to get the number of events corresponding to each sport. In addition, to better predict and analyze the Olympic events, the programs in the file that were not held in the 2024 Paris Olympics were excluded, and only the sports events that were held in the most recent Olympics were retained.

And then, we manually supplemented the document with information from the Internet (https://www.olympics.com/), obtaining the content and number of programs for the 2028 Los Angeles Olympics and the 2032 Brisbane Olympics, as shown in Figure 3 only for 2028.

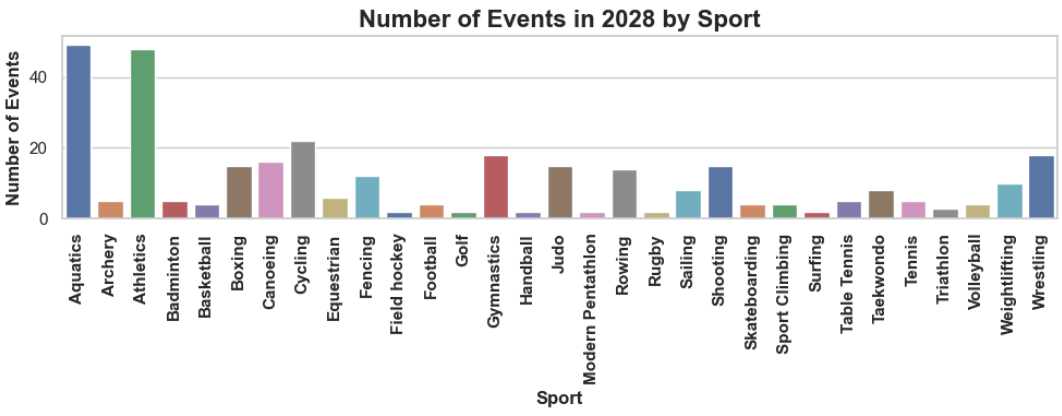


Figure 3: The number of events in 2028 Olympic game

3 Task One: Integrated Learning Model for Medal Prediction

3.1 Data Preprocessing

Athlete data, medal data, and organizer data were first read, and the athlete data was filtered to include only the awards and the total number of participants per year for each event was calculated. For each country, the number of gold, silver and bronze medals per year, per sport and the total number of medals were calculated. After this, the probability values for each medal won were added to the data separately and the indexes were reset for better machine learning afterwards.

The results of the visualization of the number of medals of each type and the probability of the total number of medals as a function of the transformation of the country (NOC) are shown in the Figure 4, where we have also re-labelled and coded the NOCs.

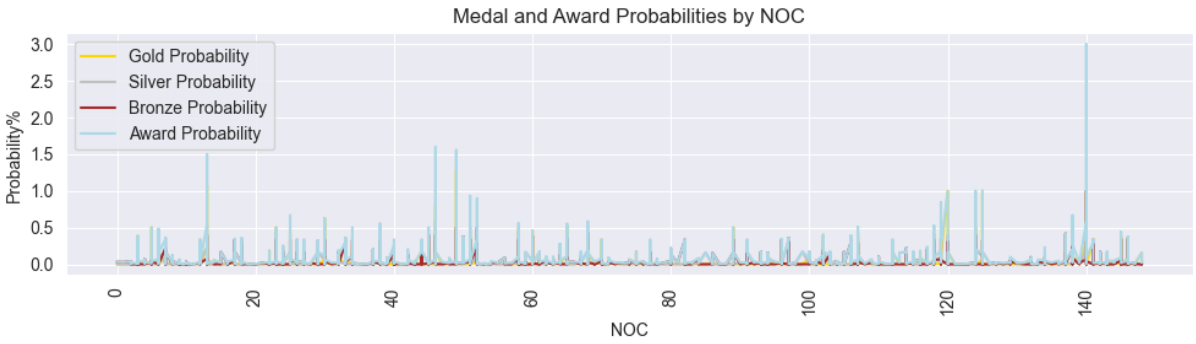


Figure 4: The medal and award probabilities visualization

For this problem and the assumptions set, we use data from two consecutive sessions to make predictions for accuracy.

3.2 Model Selection: ARIMA & Random Forest

(1) ARIMA Model

The ARIMA (AutoRegressive Integrated Moving Average) model is a widely used statistical method for time series forecasting. It combines three components:

1. Autoregressive (AR) Component: The AR part models the relationship between the current value and previous values in the series. The order of the autoregressive model is denoted as p, which represents the number of lagged observations included in the model.

2. Differencing (I) Component: The I component is used to make the time series stationary by subtracting the previous observation from the current observation. The degree of differencing is denoted as d.

3. Moving Average (MA) Component: The MA part models the relationship between the current value and past forecast errors. The order of the moving average model is denoted as q, which represents the number of lagged forecast errors [2].

The ARIMA Mathematical Expression is

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{1}$$

Algorithm 1: ARIMA Model Pseudocode

Data: Time series data y_t
Result: Forecasted values $\hat{y}_{t+h}$
Function `ARIMA Model`**:**
  Data: Time series data y_t
  Result: Forecasted values $\hat{y}_{t+h}$
  Data Preprocessing:;
  Split the data into training set S and test set T;
  Apply differencing d times to achieve stationarity;
  Model Identification:;
  Plot ACF and PACF to determine p and q;
  Model Fitting:;
  Fit ARIMA model with parameters p, d, and q;
  Estimate parameters ϕ_i, θ_j, and α using MLE;
  Model Evaluation:;
  Evaluate the model performance using the test set T;
  Forecasting:;
  Use the ARIMA model to forecast future values;
  if *Forecasting* **then**
    return The forecast formula is:

$$\hat{y}_{t+h} = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

  end

(2) Random Forest

Random Forest is an ensemble learning method primarily used for classification and regression tasks. It combines multiple decision trees to create a robust model that performs well on various types of data, especially when the relationship between the features and target variable is complex.

The main idea behind Random Forest is to build a large number of decision trees using a technique called bootstrap aggregating (or bagging) and to combine the predictions of these trees to make the final decision. This method greatly improves the accuracy and generalization ability of a single decision tree by reducing variance and overfitting [4].

The main steps are shown in the Figure 5 below:

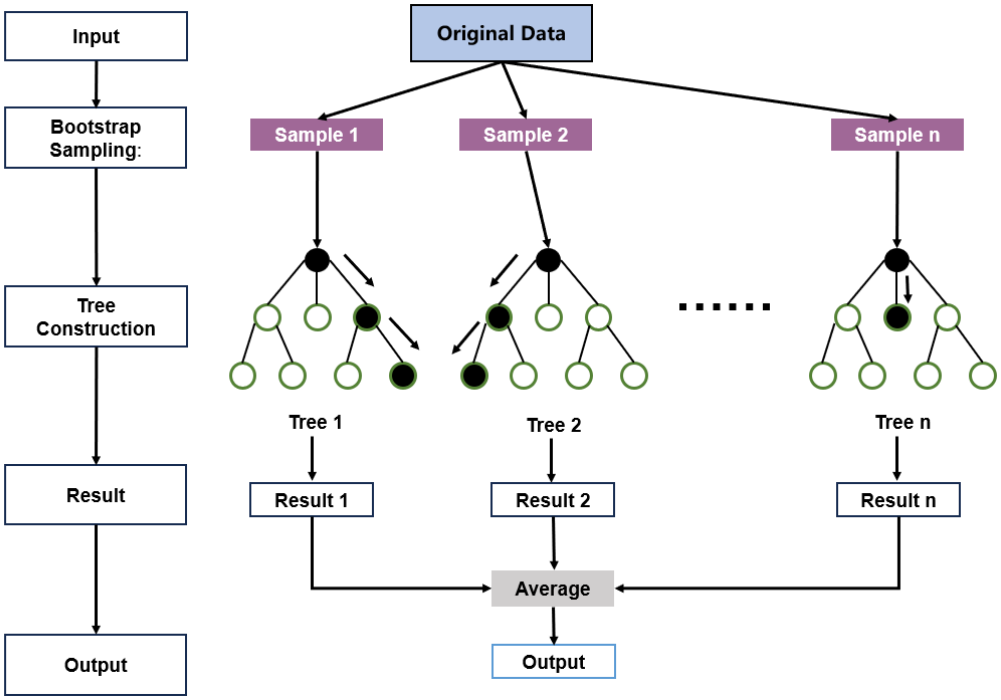


Figure 5: The structure and steps of random forest

3.3 Construction of Stacked Regression Model

Stacked regression is an integrated learning method that makes predictions by combining multiple base learners and a meta-learner. It typically improves the predictive power of the model and is more accurate than the base learners alone.

Here, we chose to extract features with ARIMA model, random forest regression for the base learner and linear regression for the meta learner to perform integrated prediction. The entire model is built in the following steps specifically:

(1) Step 1: Feature Extraction Based on ARIMA

Due to the distinct time-series nature of the medal data, time-series forecasting was performed using an ARIMA model. In order to predict the probability of gold, silver and bronze medals, a separate ARIMA model is developed here for each award. Based on the rationale, the $ARIMA(5,1,0)$ model was chosen, where p=5 for autoregression considering the last 5 years of data, d=1 for one differencing to make the series smooth, and q=0 for not using the sliding average component. Its main task is to use past medal data to predict future medal probabilities based on historical data. For example, the model to predict gold medal can be expressed as

$$\hat{y}_{t+h}^{Gold} = \alpha + \sum_{i=1}^{5} \phi_i \hat{y}_{t-i}^{Gold} + \epsilon_t \tag{2}$$

These predictions will be used as feature inputs for subsequent integrated learning models.

(2) Step 2: Base Learner Construction

Using Random Forest, an integrated learning method, as the base learner, the pre-

diction formula is as follows:

$$\hat{y}_{RF} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{3}$$

(3) Step 3: Meta Learner Construction

In stacked regression, the task of the meta learner is to make the final prediction based on the output of the base learner. Here, linear regression is chosen to be the meta learner, which can be expressed as

$$\hat{y}_{meta} = \theta_0 + \sum_{i=1}^{m} \beta_i \hat{y}_{base_i} \tag{4}$$

(4) Step 4: Overall Modeling

In order to take into account the strength of the country as well as the organizer, a special evaluation of the potential of the athlete and the country and a bonus point for the organizer were carried out. The overall modeling can be shown as Figure 6:

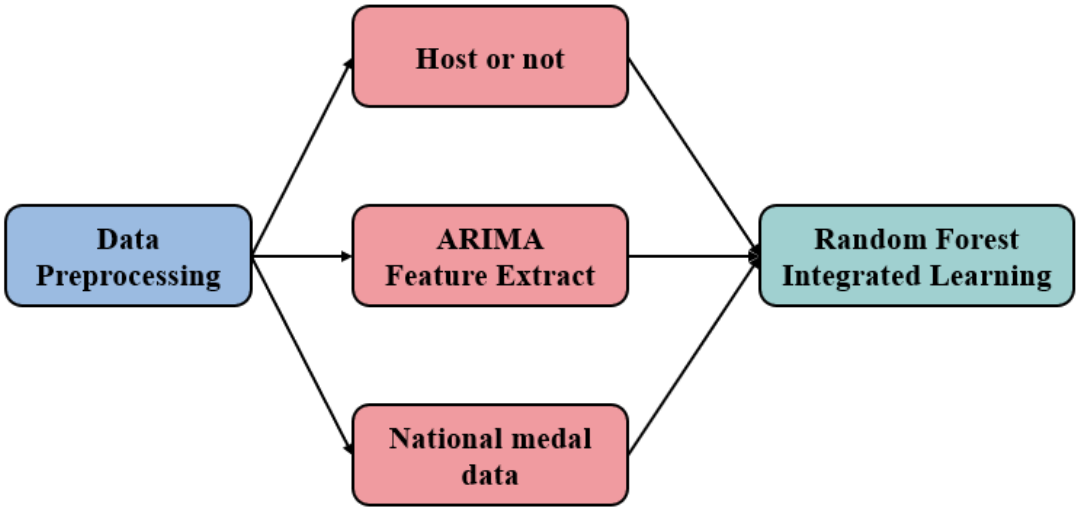


Figure 6: Overall model structure of the integrated learning

As mentioned above, the final prediction equation for the stacked regression model is

$$\hat{y} = \theta_0 + \theta_1 \hat{y}_{RF_{gold}} + \theta_2 \hat{y}_{RF_{silver}} + \theta_3 \hat{y}_{RF_{bronze}} \tag{5}$$

where, $\hat{y}_{RF_{gold}}$, $\hat{y}_{RF_{silver}}$ and $\hat{y}_{RF_{bronze}}$ are the results of predicting the probability of gold, silver and bronze medals based on the random forest regression model, respectively [3].

3.4 Result of the Stacked Regression Model

3.4.1 Results of the First Question

Using PYTHON modeling solution, the number of gold, silver and bronze medals of each country predicted for the next Olympic Games was obtained. And due to the excessive number of participating countries, 10 more representative large countries

were selected, their predictions of both the mean values and intervals are collected. The values are shown in the Table 3.

Table 3: Result of the prediction of the number of medals

	Australia	China	France	Germany	Great Britain
Gold	18.26	24.53	13.69	12.68	24.41
Silver	11.64	30.71	18.98	10.69	18.02
Bronze	16.56	42.05	15.60	12.12	15.73
	Italy	**Japan**	**Netherlands**	**South Korea**	**United States**
Gold	15.67	12.48	8.96	9.49	43.43
Silver	12.01	16.94	10.48	4.03	50.4
Bronze	12.6	17.56	11.45	6.76	47.9

To visualize the result, we have it in Figure 7:

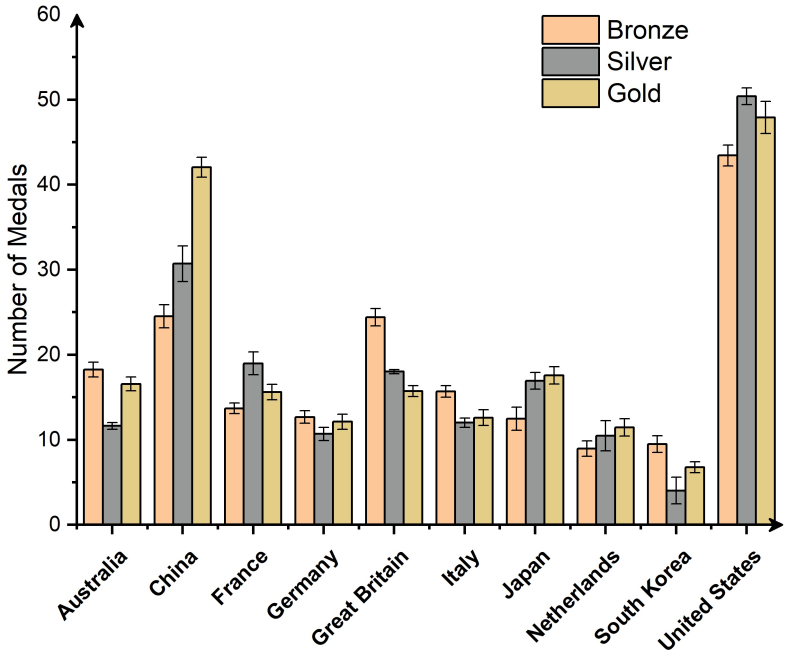


Figure 7: Prediction result of the medals

The number of gold medals was chosen as a reference, and the predicted number of gold medals of each country was compared with the actual number of gold medals won by each country in the 2024 Paris Olympics to analyze the progress and regression of each country's medals.

Similarly, for the first ten countries represented, the difference between the predicted number of gold medals and the number of gold medals at the Paris Olympics was visualized as in Figure 8:

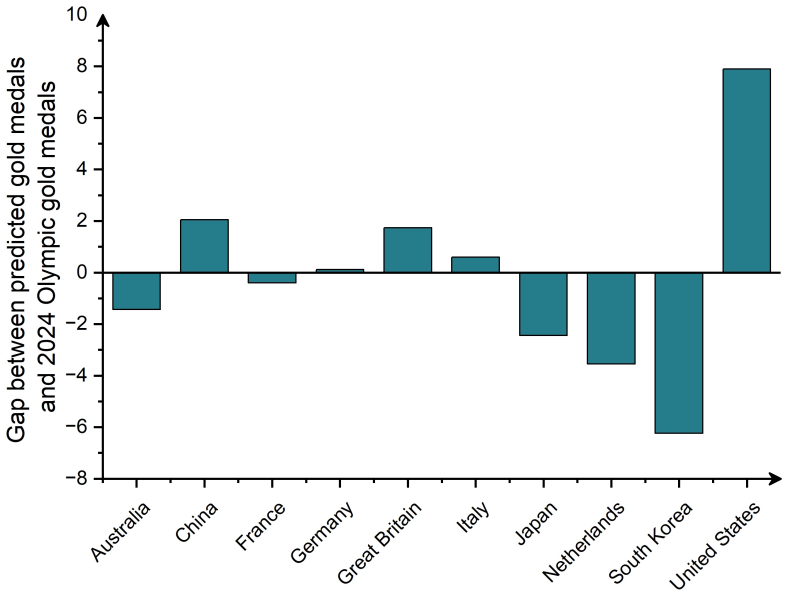


Figure 8: Gold medal counts for substracting previous data

As we can see, the United States, China, Germany, Great Britain and Italy will improve in the gold medal count, while the other five countries will regress in the gold medal count. Much of the U.S. gold medal count surge stems from its advantage as the host nation for the next Olympics.

3.4.2 Results of the Second Question

The assessment of ARIMA modeling was utilized to construct the athlete's Chisato system, and the countries that failed to get the award before were generally able to get it due to a handful of athletes in their country powering through. This leads to a system where the countries that failed to win the award get the award linked to the potential of the athlete.

By the model, we have the results as Table 4:

Table 4: Countries winning medals for the first time (indicated by code names)

HON	SLE	LAO	MAL	ROT	VNM	BRU	CRT

The accuracy of the prediction can be achieved:

$$acc = 0.77$$

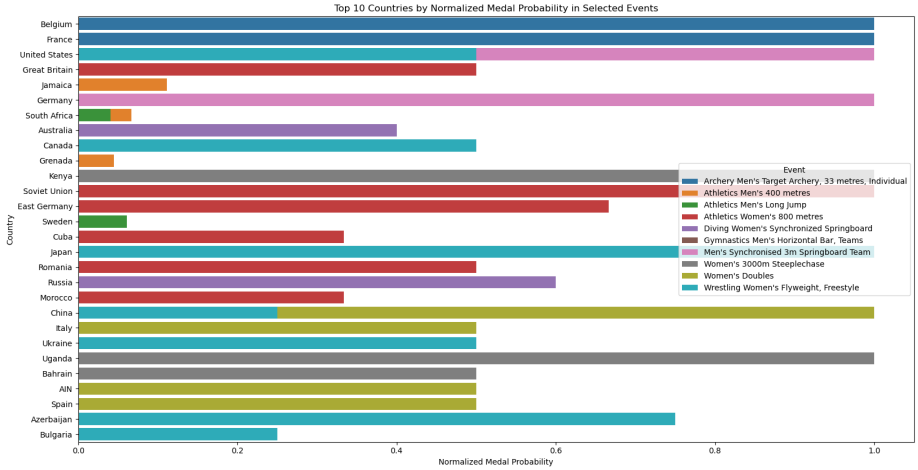
3.4.3 Results of the Third Question

Twenty countries were randomly selected to obtain the dominant sports for each country using the Pearson Correlation Coefficient, which is calculated as follows

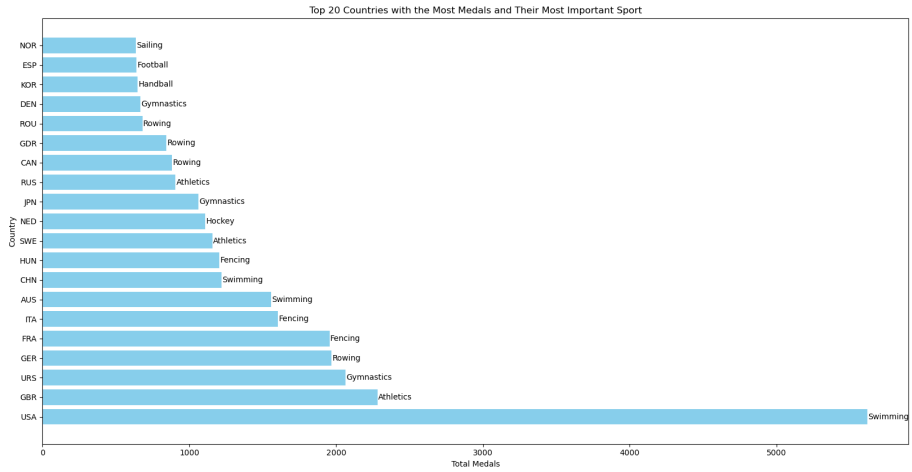
$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{6}$$

where X and Y are the two sample data, respectively. r's absolute value closer to 1 means more relevant, closer to 0 means less relevant.

Calculations were made to obtain the relevance rates for the 20 countries and the important campaigns for each country, displayed as in Figure 9.



(a) PCC value of each country



(b) Important sport of each country

Figure 9: Result of important sport event of each country

If the home country selects one of the following types of EVENT, there will be several corresponding results:

- **Selection of highly competitive events:** Some sports may be more competitive because of the larger number of countries involved. For example, sports such as track and field and swimming usually attract the best athletes in the world, so even a strong sports nation may face greater challenges in these sports. If a country chooses a sport that is too competitive, it may be limited in the number of medals it can win, especially if it does not have sufficient resources and top athletes.
- **Selection of resource-intensive events:** Some sports may require more infrastructure and resources to perform well. For example, sports such as skiing, sailing and equestrian sports usually require significant funding and special facilities, and certain countries may have limited resources in these areas, leading

them to underperform in these sports.

- **Selection of emerging or non-traditional events:** Some countries may choose to invest more resources in emerging or non-traditional Olympic sports. For example, surfing and skateboarding have been added as new disciplines in recent Olympic Games, which may provide more opportunities for certain countries. Certain countries may have excelled in these emerging sports, reflecting flexibility and adaptability in their choice of sports.

4 Task Two: Multiple Linear Regression Model of Great Coaches

4.1 Optimization Model Based on Multiple Linear Regression

In the field of Olympic competitive sports, the number of gold medals and the total number of medals won by an athlete are the key indicators of his/her performance. Coaches play a crucial role in the training and growth process of athletes, however, how the factors of coaches specifically affect the medal performance of athletes needs to be quantitatively analyzed through scientific methods. The multiple linear regression model can comprehensively consider the effects of multiple independent variables on the dependent variable, which provides a powerful tool for exploring the relationship between coaches and athletes' medal performance. Here, the multiple linear regression model is used to solve the task.

Figure 10: Two "Great Coaches"

4.1.1 Regression Objectives and Variable Determination

Given that each coach will only have a direct impact on a specific sport, and that total medal counts are also susceptible to interference from overall sport investment, policy and non-coaching factors, the objective of the regression analysis was set to be the medal dynamics of individual specific sports, i.e., the number of medals won by a country in a given sport. Thus, the dependent variable y is the number of medals won by a country in a given event.

After determining the dependent variables, the following five independent variables were designed:

- **"Great Coach" (x_1)**: Flag whether or not to bring in a great coach who will have a direct impact on the program ("1" means yes, "0" means no).

- **Historical performance (x_2)**: Weighted based on the number of medals won in the last three Olympic Games, calculated as follows

$$x_2 = \alpha_1 \cdot n_1 + \alpha_2 \cdot n_2 + \alpha_3 \cdot n_3 \tag{7}$$

  where n_1 indicates the number of medals in last term, n_2 indicates the number of medals two terms ago and n_3 indicates the number of medals three terms ago.

- **Number of athlete participants (x_3)**: Count the number of athletes from each country competing in the program, reflecting the level of national commitment to the program.
- **Intensity of medal competition (x_4)**: The intensity of medal competition can be measured by entropy. In the distribution of medals, the entropy value can reflect the uniformity of competition. The higher the entropy value, the more uniform the distribution of medals; the lower the entropy value, the medals are highly monopolized by a few countries.
- **Host effect (x_5)**: Record whether it is the host country or not, 1 for host, 0 for non-host

4.1.2 Modeling of the Multiple Linear Regression

The regression model is built as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon \tag{8}$$

where β_0 is the intercept denoting the number of medals predicted when all independent variables are 0 and ϵ is the residual term used to test model robustness [5] [6] [7].

4.2 Data Preprocessing

For the independent variable of intensity of medal competition, the process of using the entropy weight method is as follows:

1. Calculate the proportion of medals won by country i as a share of the total number of medals won:

$$p_i = \frac{N_i}{\bar{N}} \tag{9}$$

   where N_1 indicates the number of medals that country i won and $\bar{N}$ indicates the number of medals that all countries won.

2. Take the logarithm of p_i:

$$\log_2(p_i)$$

   This section measures the information or uncertainty associated with the proportion p_i.

3. Get entropy value H:

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) \tag{10}$$

   If a country has few medals, it contributes less to the entropy value because the uncertainty is lower. Whereas, when multiple countries have more medals, the

entropy value increases, indicating a more even distribution of medals and more competition.

After obtaining the medal competition intensity values, the other independent variables are obtained by definition and the final preprocessed data set is designed as in Table 5:

Table 5: Data pre-processing and presentation

Event	Country	Year	y	x_1	x_2	x_3	x_4	x_5
Volleyball	USA	2008	1	0	2	12	0.65	0
Volleyball	USA	2012	3	1	2	15	0.73	0
Volleyball	CHN	2012	0	0	1	14	0.70	0
Volleyball	CHN	2016	2	1	0.5	16	0.80	0
Volleyball	CHN	2020	1	1	2	15	0.78	0

4.3 Results of the Model

4.3.1 Calculation Results and Visualization

We selected data for volleyball athletics and swimming, respectively. The individual coefficients are obtained by MATLAB solution as shown in Table 6:

Table 6: Result of the multiple linear regression with volleyball

Sport	β_0	β_1	β_2	β_3	β_4	β_5	R^2
Volleyball	-0.1820	4.4222	0.0463	-0.0201	0.3098	0.7351	0.9331
Athletics	-0.0317	3.6853	0.0073	0.0420	-0.0503	-0.1775	0.8816
Swimming	0.3980	3.4794	1.0567	-0.0066	-0.1445	2.1990	0.9624

Their coefficient can be visualized in the Figure 11.

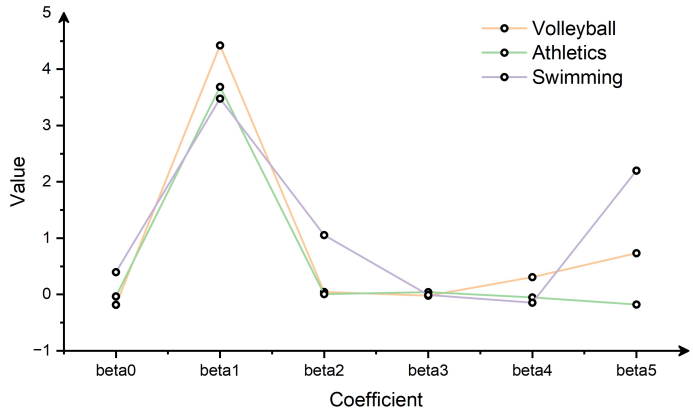
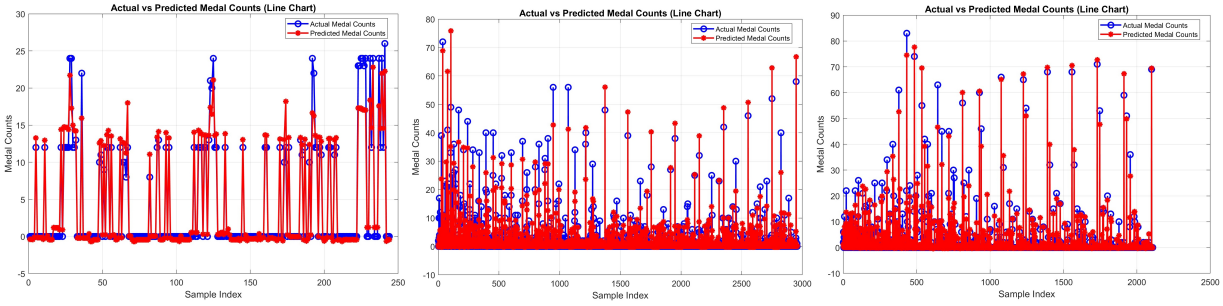


Figure 11: The value of different coefficients

As can be seen from the parameters of the three models, the great coach effect is one of the most significant factors in the model, and the difference in the number of medals won by countries with or without great coaches can reach 3-4 medals on average under the same conditions.

Based on the model we built, the prediction yielded the number of medals and the actual number of medals as shown in Figure 12:



(a) Volleyball data (b) Athletics data (c) Swimming data

Figure 12: Actual and predicted medal counts

4.3.2 Results Analysis

In the existing analysis, taking volleyball as an example, the regression coefficient β_1 (the great coach effect) is 4.4222, which indicates that the number of medals is expected to increase by an average of 4.42 medals after the introduction of a great coach. From the actual case, Lang Ping coached the Chinese women's volleyball team, the team's performance improved significantly, such as the 2016 Olympic Games, the number of medals won in line with or even exceeding expectations, which provides practical evidence to support the great coach effect. Similarly, Béla Károlyi's improvement in the performance of the U.S. gymnastics team further validates the existence of this effect.

According to the model, the contribution of the great coach effect varies across sports. In volleyball, as mentioned above, $\beta_1 = 4.4222$, which means that the number of medals is expected to increase by 4.42 medals on average for each introduction of great coaches; in athletics, $\beta_1 = 3.6853$, which means that the introduction of great coaches is expected to increase by 3.6853 medals on average; and in Swimming, $\beta_1 = 3.479376$, which is expected to increase the number of medals by 3.479376. These data are estimates based on model analysis and reflect the effect of great coaches on medal counts, controlling for other variables such as historical performance, number of participants, intensity of competition and home field effects.

4.3.3 Three Countries and Their Great Coach Investigation

Based on our work in this task, there are three countries that should focus on the project and influence estimation of the project of "Great Coach".

(1) Volleyball for Japan

The Japanese women's volleyball team has excellent defense, but the offensive ability needs to be improved. Italian coach Mazanti, who is good at offensive tactics.

Mazanti's coaching style is flexible, focusing on tactical cooperation and player characteristics, and led the Italian women's volleyball team to achieve many excellent results. His training strategy is to create a variety of offensive systems and tap the player's offensive potential. It is expected that in the next three years, the Japanese women's volleyball team is expected to win 3-4 medals in the World Championship, and the medal ranking is expected to increase by 2-3.

(2) Athletics for the US

There is a significant gap between American and African athletes in middle and long-distance events, such as the marathon, 5000m, and 10,000m, with American runners trailing behind athletes from Kenya and Ethiopia. To improve performance, it is recommended to hire coaches like Niels, who has helped athletes like Bolt and Blake achieve success. With his guidance, it is expected that the number of medals in the Olympics and World Championships will increase by 4, with 2-3 additional gold medals over the next five years.

(3) Swimming for China

The Chinese swimming team has shown impressive results, but still faces challenges in long-distance events compared to traditional powerhouses like Australia and the United States. To improve, hiring Australian coach Dennis Cottrel is a strong option. His scientific coaching approach, combining technology and physical training, has helped Australian swimmers win 5 gold, 4 silver, and 2 bronze medals. Cottrel also contributed to Sun Yang's gold medals. With his guidance, China is expected to win 3-4 medals in international competitions over the next four years.

4.4 Testing of the Model

In order to better measure the contribution of the respective variables in the model to the dependent variable, we chose to test the model with both t-value test and p-value test.

(1) t-value test

The basic principle of the t-test is to determine whether the original hypothesis can be rejected by calculating the difference between the sample means and comparing it to the variability between the samples. Its calculation formula is expressed as follows:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{11}$$

where $\bar{X}$ is the average value of sample, μ_0 is the supposed average value, s is the standard value of sample and n is the size of sample.

The decision to reject the original hypothesis is made by calculating the t-value and comparing it to the critical value.

(2) p-value test

The p-value is a method used to determine the significance of the results of a statistical hypothesis test. It indicates the probability of observing the current or more extreme outcome if the original hypothesis is true. The smaller the P-value, the less likely the observed outcome is to occur if the original hypothesis is true, thus increasing the evidence for rejecting the original hypothesis.

The test was conducted using the data from volleyball, athletics and swimming data. The results of the t-value test and p-value test were obtained as in Figure 13, respectively:

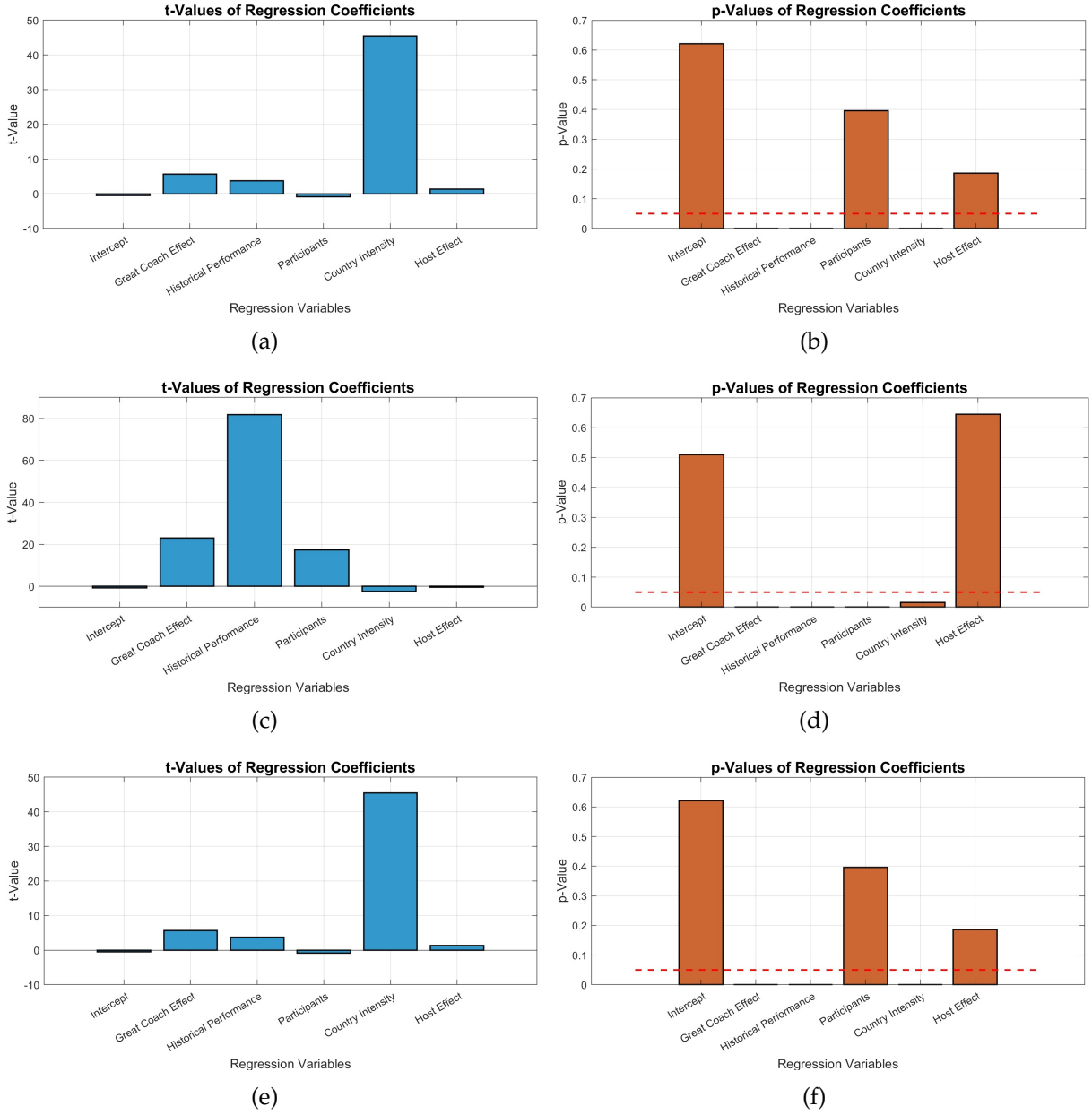


Figure 13: Testing of regression coefficients: (a) t-value test of volleyball data; (b) p-value test of volleyball data; (c) t-value test of athletics data; (d) p-value test of athletics data; (e) t-value test of swimming data; (f) p-value test of swimming data

5 Task Three: Insights and Implications from Model

5.1 Insights from the Model

An in-depth analysis of the Olympic medal count prediction model reveals some other unique insights about the Olympic medal count, which provide new perspectives on understanding the competitive landscape and sports development trends of

the Olympic Games from different angles. These insights can be summarized in the following box:

Insights

- Impact of the gender factor
- Impact of sports organizations
- Impact of the attributes of the sport
- Impact of country typology

(1) Impact of the gender factor

With the development of society and the advancement of the concept of gender equality, the participation and competitiveness of women in the Olympic Games have been increasing. Judging from historical data, more and more countries have begun to pay attention to the development of women's sports programs and increase their efforts to train and support female athletes, as can be seen from the following chart (Figure 14):

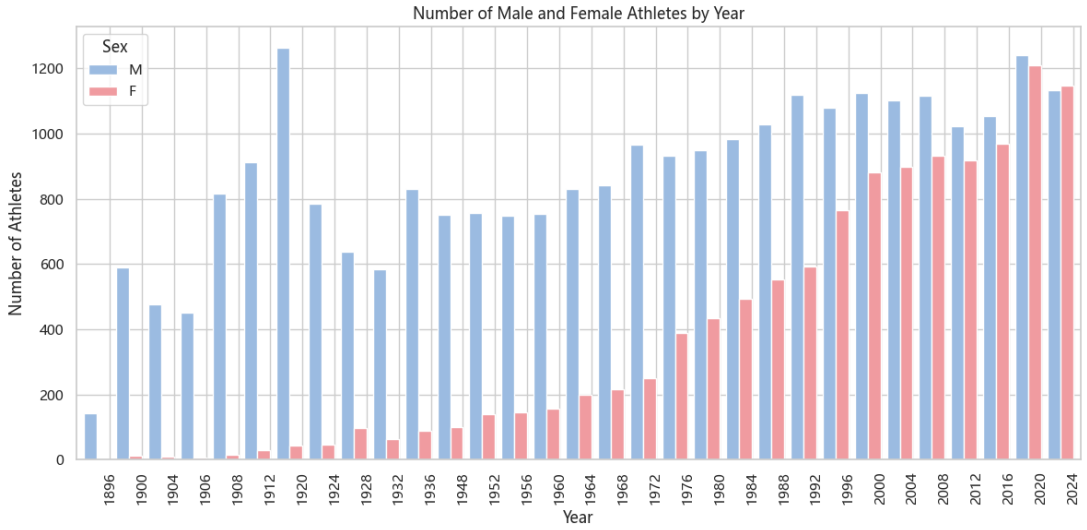


Figure 14: Number of male and female members of the Olympic team

In some traditionally male-dominated sports, female athletes are also gradually emerging. In track and field, the performance of women in sprinting and middle-distance running has been improving, and the number of medals won by women athletes in these sports has also been increasing. This demonstrates that the advancement of gender equality has not only helped to increase the status and participation of female athletes, but has also provided new opportunities for countries to win more medals at the Olympic Games.

(2) Impact of sports organization

An efficient sports management organization can optimize resource allocation, create scientific training programs, and develop top athletes, thereby improving a country's Olympic medal count. For instance, China's State General Administration of Sport has significantly enhanced the nation's Olympic performance through the National Fitness Program and the Olympic Competitiveness Program. At the 2008 Beijing Olympics, China topped the gold medal tally with 51 golds, 21 silvers, and 28

bronzes. This success was driven by strategic management, the focused investment in key sports, and the training of outstanding athletes.

(3) Impact of the attributes of the sport

Given the type of sport, individual programs often tend to have a higher intensity of medal competition, which affects the coefficient of the model. Selecting three sports from the data, basketball, cycling and weightlifting, for comparison, it can be seen that the entropy values used to calculate the intensity produced significant differences, and the visualization results are shown in Figure 15.

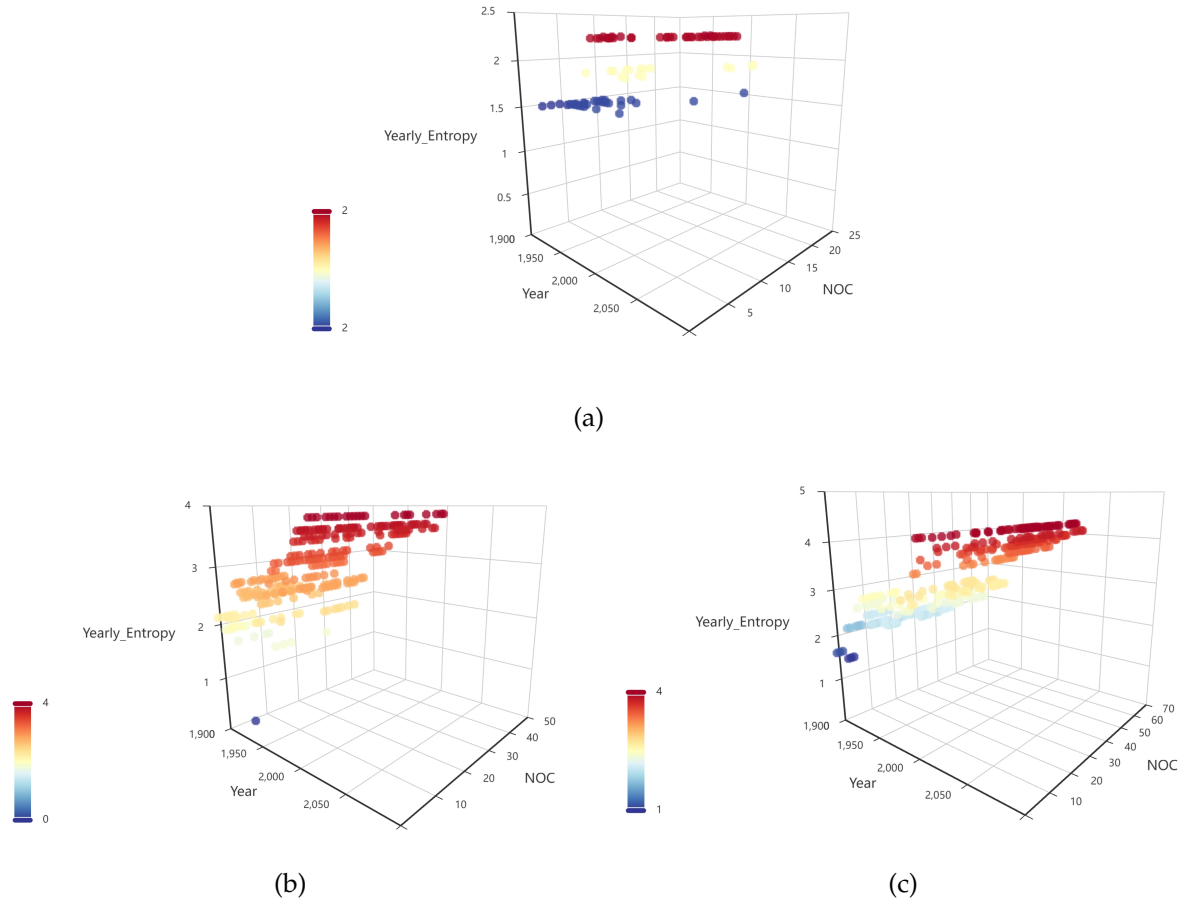


Figure 15: The yearly entropy values of three kinds of sports: (a) Basketball; (b) Cycling; (c) Weighing

(4) Impact of country typology

Developed countries typically have abundant sports resources, advanced training facilities, and robust talent development systems, giving them a competitive edge in the Olympics. However, some developing countries, like China and India, have also improved their Olympic performance by tailoring sports strategies to their national conditions. China has made significant progress by implementing a national sports system and focusing on key sports, while India has recently increased sports investment and achieved breakthroughs, such as winning multiple medals in shooting at the 2024 Paris Olympics.

5.2 Implications for NOCs

Based on the results of this study, we provide the following insights and recommendations for National Olympic Committees to help them make scientific decisions in sports strategic planning, resource allocation and talent development, and to improve Olympic medal counts and international competitiveness.

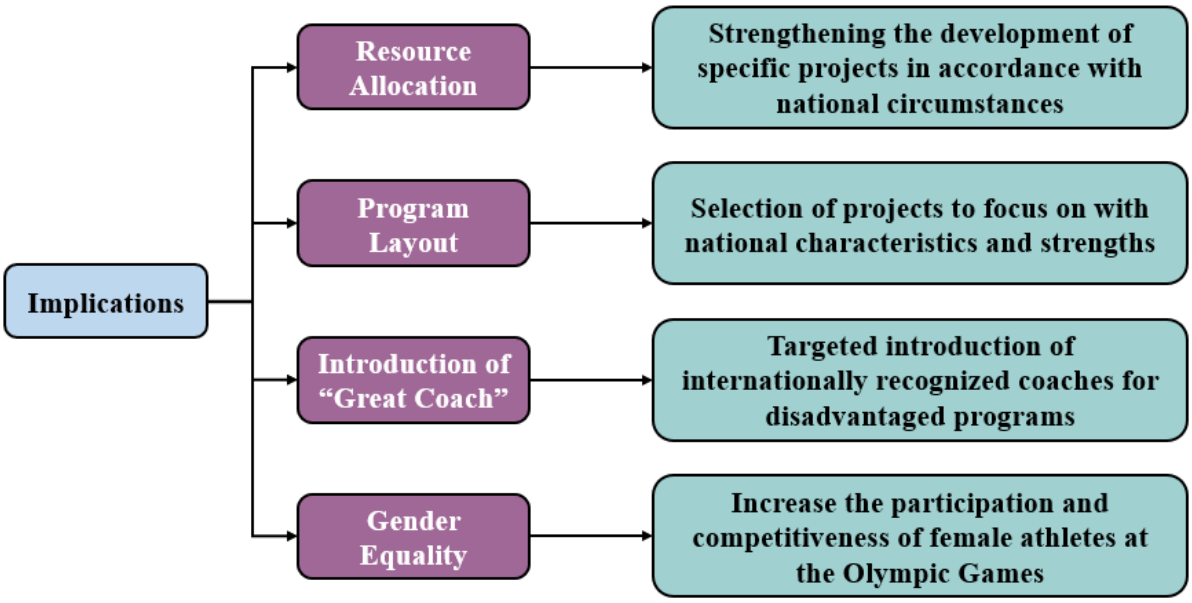


Figure 16: Implications for NOCs

National Olympic committees should allocate resources based on model predictions and program analysis. Economically weaker countries should focus on high-potential projects, like Kenya with long-distance running. Sports powerhouses should invest in emerging programs while maintaining support for traditional strengths. For example, China has increased its backing for e-sports and skateboarding while continuing to support table tennis and diving.

In terms of program layout, national Olympic committees should combine sports traditions, cultural backgrounds and geographical advantages to select programs with characteristics and potentials for development. Australia should further strengthen the layout of the swimming program and improve the talent training system. In addition, with the changes of the Olympic program, countries should adjust the layout of the program in time and layout the emerging programs in advance.

With regard to the introduction of coaches, national Olympic committees should pay attention to the introduction and training of excellent coaches, especially for programs with poor performance. For example, India can introduce excellent tennis coaches to improve the level. At the same time, countries should pay attention to the training of local coaches to improve their professional quality.

Finally, the Olympic Committee should focus on gender equality and the building of a sporting culture. It should increase its support for women's sports programs and promote the participation and competitiveness of female athletes. At the same time, it should enhance society's interest in sports and cultivate more sports talents.

6 Sensitivity Analysis

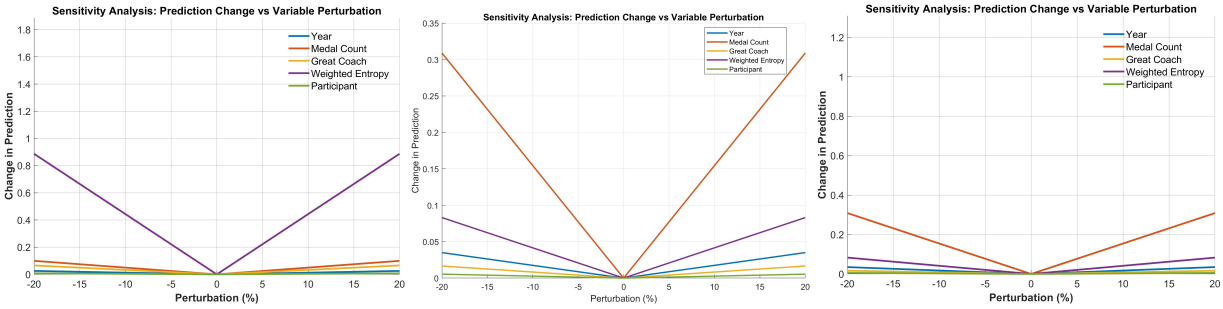
Perform sensitivity analysis on the multiple linear regression model in task 2. Set multiple perturbation ratios, and for each independent variable (except the bias term), adjust it according to different perturbation ratios to obtain a new independent variable matrix. Use the perturbed independent variable matrix to calculate the new predicted values, and compare them with the original predicted values to calculate the average absolute change as the sensitivity index.

The values for setting the perturbation ratio are shown in Table 7.

Table 7: Result of the multiple linear regression with volleyball

-20%	-10%	0%	10%	10%

The visual display of the results after the sensitivity calculation is shown in Figure 17. The trend of the change of the predicted value of each independent variable under different perturbation ratios can be obtained, which makes it easy to observe the change of the sensitivity of different independent variables.



(a) Volleyball data (b) Athletics data (c) Swimming data

Figure 17: Sensitivity Analysis of Multiple Linear Regression Model

The figure shows that the "Great Coach" variable has a steeper slope, indicating its greater sensitivity to changes, with small adjustments causing significant fluctuations in predicted medal counts. In contrast, the "Historical Performance" variable has a gentler slope, suggesting its weaker impact on the predictions. This comparison highlights the importance of investing in great coaches for countries aiming to boost medals in sports like volleyball, as it can lead to a more substantial increase in medal counts than improving historical performance alone.

7 Model Evaluation and Futher Discussion

7.1 Strengths

- The model combines several methods (ARIMA, Random Forest, Stacked Regression) to be able to predict medal probabilities from multiple perspectives, integrating time series characteristics with other social and economic factors (e.g., the

host country effect, the influence of coaches, etc.), which improves the predictive ability of the model and the accuracy of inter-series forecasting ability.

- Compared with other complex machine learning models (e.g. neural networks, XGBoost, etc.), the training process of multiple linear regression has less computational overhead and can quickly obtain training results, which is suitable for the case of a relatively small amount of data and a strong linear relationship between features.
- The multiple linear regression model allows us to quantify the impact of great coaches on medal counts and to distinguish them from other variables (e.g., historical performance, training inputs, number of athletes involved, etc.).The regression coefficient (β_1) for the Great Coach variable in the model directly represents the average incremental impact of this factor on medal counts, which facilitates comparisons across countries or sports.

7.2 Weaknesses

- ARIMA models and stacked regressions rely on large amounts of historical data to build valid new or incomplete historical data for countries and programs that may not have enough data to make accurate predictions.
- Integrated learning methods, especially random forest and stacked regression models, often lack strong interpretability. In practice, the models are able to provide a quasi-explanation of how each input feature affects the prediction results though, which may lead to less transparency in the model.

7.3 Futher Discussion

- **Improve the model to include more influences: Longitudinal Analysis**
  The current regression model assumes that the impact of great coaches on medal scores is relatively constant, but that this impact may change over time, including continuity and decline. This can help to predict the persistent impact of great coaches on medal performance, and thus guide long-term investments.
- **International Coaching Mobility and Systemic Impacts: Synergistic Effects of Non-Coaching Factors on Medal Counts**
  While coaching is important, the interaction between the effects of great coaching and other resource inputs could be investigated in the future. For example, in the absence of high-level athletes, can great coaches be fully effective or do they need to be complemented by pre-selected athletes with great potential; if there is a lack of infrastructure or training conditions, can great coaches overcome these obstacles, and so on.

8 Conclusion

This study provides crucial insights into the factors affecting Olympic medal distribution and offers effective models for predicting future performance. The Integrated Learning and Multiple Linear Regression models have enabled the prediction of medal counts, emphasizing the importance of strategic resource allocation and the significant

role of coaches in improving medal outcomes. National Olympic Committees (NOCs) can use these findings to focus on the development of key sports based on national strengths, allocate resources more efficiently, and invest in elite coaching to enhance athletic performance. Looking ahead, further refinement of these models, including considering the dynamic impacts of new sports and changing global competition patterns, will provide even more accurate predictions. These insights can help NOCs stay adaptable in a rapidly evolving Olympic landscape and make informed decisions that maximize their chances of success in future games.

References

[1] Bian, X. (2005). Predicting Olympic medal counts: The effects of economic development on Olympic performance. The park place economist, 13(1), 37-44.

[2] Newbold, P. (1983). ARIMA model building and the time series analysis approach to forecasting. Journal of forecasting, 2(1), 23-35.

[3] Ornstein, J. T. (2020). Stacked regression and poststratification.Political Analysis,28(2), 293-301.

[4] Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

[5] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

[6] Hansen, P. Ø., & Andersen, S. S. (2014). Coaching elite athletes: How coaches stimulate elite athletes' reflection. Sports Coaching Review,3(1), 17-32.

[7] De Bosscher, V., & van Bottenburg, M. (2010). Elite for all, all for elite?: An assessment of the impact of sports development on elite sport success. In Routledge handbook of sports development (pp. 579-598). Routledge.