Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2522127

One Model to Rule Them All: Olympic Medal Prediction and Strategy

Summary

The Olympic Games represent a globally celebrated sporting event that garners immense attention from nations worldwide. Countries, ranging from sporting powerhouses like the United States to developing nations such as Dominica, closely monitor the medal standings. Our team is dedicated to accurately predicting the medal tally for all nations ahead of the Summer Olympics, enabling each country to strategically plan training programs and adjust their sports development strategies in advance.

For question 1, we proposed a AHH model, which employs multiple linear regression model as a carrier to predict kinds of medal count. AHH combines the prediction results from the random forest model, the Prophet model, and the host effect estimation algorithm we proposed. Then, we apply the ADF test to examine the stationarity of the annual first-medal country count and use the ARIMA model to predict the count for 2028. Logistic regression is then employed to estimate the probability of non-medal countries achieving a breakthrough, with the highest-probability countries identified. Next, Spearman's rank correlation is used to analyze the relationship between national medal counts and the number of events. We propose an importance calculation model, where performance points are derived from the weighted sum of annual medals across events. These points are aggregated to assess the significance of each sport to a country's medal performance. Analysis of official data reveals that changes in events due to the organizer effect tend to favor the host country and impact nations differently.

For question 2, we employed the Difference - in - Differences (DID) method to quantify the contribution of great coaches to the number of medals. For women's volleyball, we chose the United States and China as the treatment group, and Serbian and the Netherlands as the control group. Then, by applying the parallel trends assumption, we calculated the DID estimate to be 0.5. This indicates that after controlling for other common factors, Lang Ping's coaching led to a relative increase in the medal count of the Chinese women's volleyball team at the 2016 Rio de Janeiro Olympics. By combining these examples and the medal tables of previous Olympic Games, we estimated that after hiring outstanding coaches, teams can achieve a breakthrough of one or two gold medals, and the total number of medals can increase by 5% to 8%.

For question 3, our model reveals the dynamic changes in dominant events, drives the decline of traditional events and the rise of emerging ones. It also highlights the need for tailored strategies: sports powerhouses rely on historical advantages, while less-developed nations benefit significantly from the host-country effect. Furthermore, we show that the number of first-time medal-winning countries reflects global sports balance, with increases indicating reduced dominance by traditional powers. Finally, it uncovers key factors for a country's first medal, such as event participation and Olympic appearances. Based on these insights, we recommend that country Olympic committees optimize athlete selection and training processes and increase investment in sports research and development.

Contents

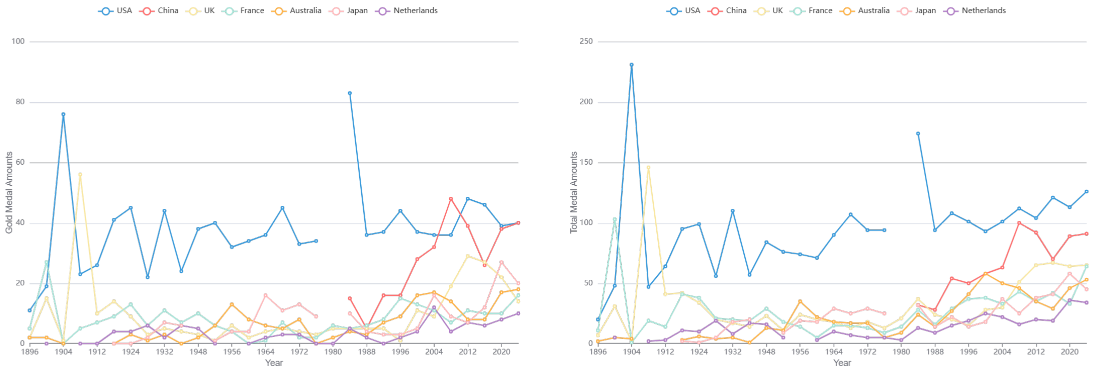
1 Introduction

1.1 Problem Background

As the most influential sporting event in the world, the Olympic Games have long been a topic of great interest, not only for global audiences but also as a crucial benchmark for governments to assess Olympic performance.

As shown in Figure 1, we selected the seven countries with the highest gold medal counts in the 2024 Paris Olympics, presenting the number of gold medals they won at the Olympics from 1896 to 2024 (Figure 1(a)) and their total medal counts (Figure 1(b)). If a country can accurately predict its medal outcomes before the Games, the government can adjust national sports policies based on these predictions, ensuring that funding and resources are concentrated in the most promising sports, thereby enhancing overall Olympic performance [2]. At the same time, accurate medal forecasts aid sports betting companies, media, and sponsors in allocating resources efficiently, improving their return on investment, and fostering the development of the sports industry [4].

Therefore, it is not only sports powerhouses, as highlighted in the figure, but also other nations that need a model to help predict their medal outcomes. This is also of significant importance for motivating more citizens to engage in sports activities, thus improving national health levels.



(a) Gold Medals by Country

(b) Total Medals by Country

Figure 1: Olympic Medal Performance of Top 7 Countries in the 2024 Paris Olympics

1.2 Problem Restatement

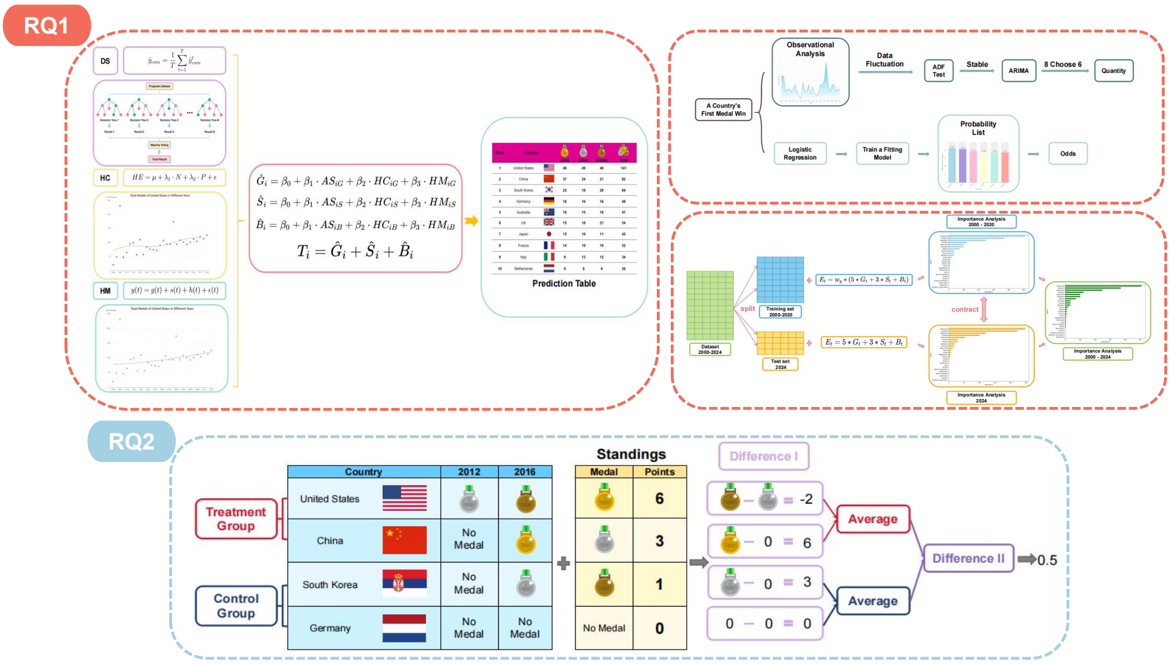
A country's medal tally may be influenced by factors such as coaches and the number of events. By developing a model that considers these factors, we can provide predictions for the future Olympic Games and put forward suggestions to enhance the national competitiveness. In light of this, our paper aims to address the following research questions:

- **RQ1:** How to develop a model to predict the number of gold, silver, and bronze medals for each country in the Olympics, including estimates of uncertainty and model performance measures?
  - The model should provide projections for the 2028 Los Angeles Olympics medal table, including prediction intervals. It should also compare the 2028

  predictions with the actual results from 2024 to identify countries likely to perform better or worse.
  - The model should estimate how many countries that have not earned medals in previous Olympics are likely to win their first medal in 2028, and provide the probability of this happening.
  - The model should analyze the impact of the number and types of Olympic events on medal counts, identify key events for each country and the reasons behind them, and explain how the host country's event selection influences the overall results.

- **RQ2:** How can the "great coach" effect, where coaches significantly influence medal performance, be quantified? What is the impact of this effect on medal counts for specific countries?
- **RQ3:** How can the insights gained from the model inform National Olympic Committees in their strategies for future Olympic Games?

1.3 Our Work


RQ1
DS
$\hat{y}_{csm} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}^t_{csm}$
HC
$HE = \mu + \lambda_1 \cdot N + \lambda_2 \cdot P + \epsilon$
HM
$y(t) = g(t) + s(t) + h(t) + \epsilon(t)$
$\hat{G}_i = \beta_0 + \beta_1 \cdot AS_{iG} + \beta_2 \cdot HC_{iG} + \beta_3 \cdot HM_{iG}$
$\hat{S}_i = \beta_0 + \beta_1 \cdot AS_{iS} + \beta_2 \cdot HC_{iS} + \beta_3 \cdot HM_{iS}$
$\hat{B}_i = \beta_0 + \beta_1 \cdot AS_{iB} + \beta_2 \cdot HC_{iB} + \beta_3 \cdot HM_{iB}$
$T_i = \hat{G}_i + \hat{S}_i + \hat{B}_i$
Prediction Table
A Country's First Medal Win
Observational Analysis
Data Fluctuation
ADF Test
Stable
ARIMA
8 Choose 6
Quantity
Logistic Regression
Train a Fitting Model
Probability List
Odds
Dataset 2000-2024
split
Training set 2000-2020
Test set 2024
$E_t = w_y * (5 * G_t + 3 * S_t + B_t)$
$E_t = 5 * G_t + 3 * S_t + B_t$
Importance Analysis 2000 - 2020
contract
Importance Analysis 2024
Importance Analysis 2000 - 2024
RQ2
Treatment Group
Control Group
Standings
Difference I
Average
Average
Difference II
0.5

Rank	Country	Gold	Silver	Bronze	Total
1	United States	46	49	46	141
2	China	37	24	21	82
3	South Korea	25	18	26	69
4	Germany	16	16	16	48
5	Australia	16	15	16	47
6	UK	15	18	21	54
7	Japan	15	16	11	42
8	France	14	19	19	52
9	Italy	9	13	12	34
10	Netherlands	9	8	9	26

Country	2012	2016
United States		
China	No Medal	
South Korea	No Medal	
Germany	No Medal	No Medal

Medal	Points
	6
	3
	1
No Medal	0

−	=	-2
− 0	=	6
− 0	=	3
0 − 0	=	0

Figure 2: Framework Architecture of Our Work

This problem requires us to predict the future medal tally, estimate the probability of countries winning medals for the first time, and explore the relationship between events and the number of medals. Our work mainly includes the following:

1. Construct a large model integrating the Random Forest model, Prohert model, and Linear Regression model to predict the medal tally.
2. Predict the number of countries winning medals for the first time using the ARIMA model, and then calculate the probability with a logistic model.
3. In this paper, the applicability and effectiveness of the model are significantly improved through sensitivity analysis and accuracy estimation.

The flow chart is shown in Figure 2.

2 Assumptions and Justifications

- **Assumption 1: Countries that have performed outstandingly in past Olympics and have long been at the top of the medal table will still rank high in future Olympics.**
  Justification: With their profound sports heritage, well - established sports talent cultivation systems and sufficient resource investment, they will continue to maintain strong competitiveness.
- **Assumption 2: Athletes from all countries can maintain good competitive conditions during the Olympics, without widespread injuries or large - scale disqualifications due to doping and other reasons.**
  Justification: Since athletes' physical and competitive states directly determine their performance in competitions, which is crucial for winning medals.
- **Assumption 3: Host country often has certain home-field advantages in future Olympics.**
  Justification: Host country's athletes have advantages in terms of audience support, adaptation to the competition venue, daily routines, and eating habits.
- **Assumption 4: There will be no major and disruptive adjustments to the sports development policies of countries before the future Olympics.**
  Justification: As sports development policies are important guidelines for a country's sports undertakings, stable policies can ensure the continuous progress of sports project development plans and guarantee the coherence and stability of the athlete cultivation system.
- **Assumption 5: There will be no large - scale wars, political conflicts, global public health crises and other major international events before and during the Olympics.**
  Justification: Major international events often have a serious impact on the schedule of the Olympics, athletes' participation eligibility and psychological states.

3 Notations

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

4 RQ1.1: Predict 2028 Medals with Intervals

For most prediction model, it is usually quite easy to make rough guesses about the results of the Olympic events based on various objective conditions. Despite this,it is still difficult to quantify these certain factors and evaluate the degree of each cause to specific country, some models, such as LSTM [5], may only consider the historical medal counts of countries as a reflection of their overall national strength. When the advantages of such countries are concentrated in certain major categories, and the number of these events is significantly reduced in a given year, the performance of that country in that year will be greatly diminished. In contrast, countries with a wide distribution of advantageous events will not be affected. Therefore, such models, due to their singular evaluation criteria, find it difficult to provide convincing results. Therefore, a model that can customize predictive solutions by incorporating both the di-

Table 1: Notations used in this paper

Symbol	Description
AS	Advantage Score
HC	Host-Country effect
HM	Historical gold, silver, and bronze medal counts
$\hat{G}_i$	Gold medal count
$\hat{S}_i$	Silver medal count
$\hat{B}_i$	Bronze medal count
$\vec{\beta}_G$	Coefficient vector
$\vec{X}_i$	Feature vector
T_i	Total medal count
N	the number of event changes in each Olympic Games
P	the host - country's event participation rat
E	the number of events they participated per Olympics
T	denotes the number of times a country has participated
H	indicates whether it is the host nation

versity of conditions of different countries and the outer objective conditions is quite important.

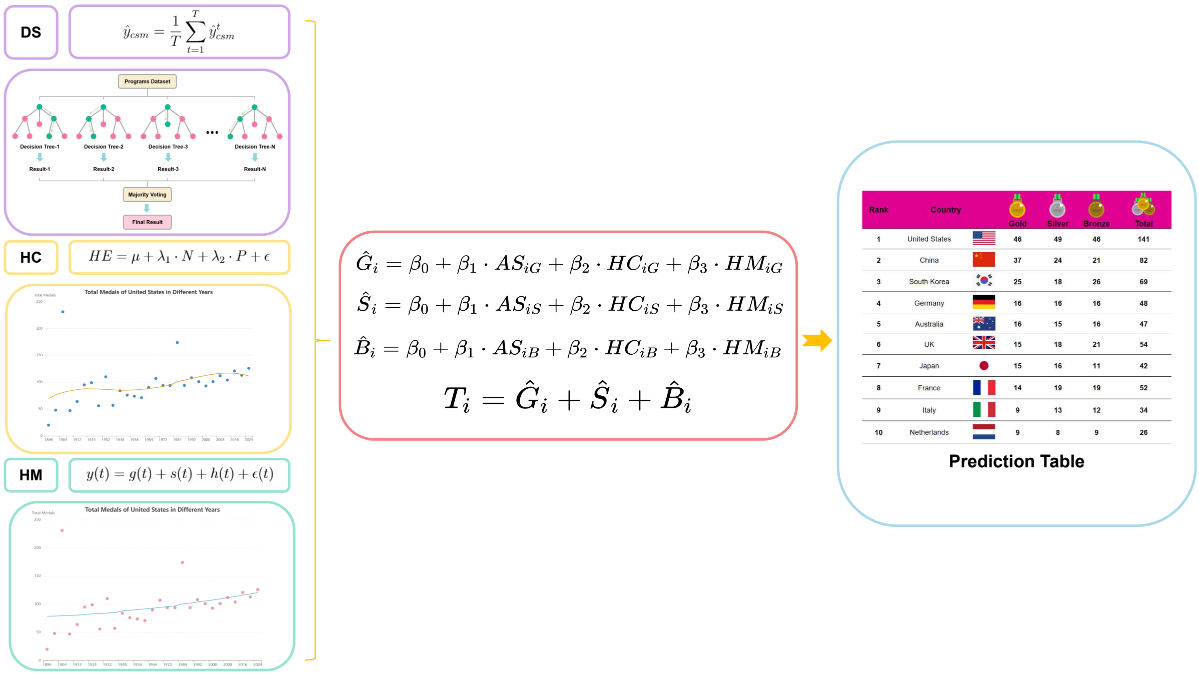
Inspired by the methodology of James et al.[3], we adopted the multiple linear regression method as a carrier to construct our AHH model, which is capable to consider the three features estimated by different aspects and algorithms: **Advantage Score(AS), Host - Country Effect (HC), Historical Medal Counts (HM)**. These three features have an impact on the results from different dimensions and together constitutes a comprehensive impact on the number of medals. The multiple linear regression is commonly used to predict the result by the estimated regression coefficient. The regression coefficient is calculated as a vector where each value from each dimension represents a customized effectiveness weight for a specific country, to adapt different national conditions.

$$\hat{G}_i = \beta_0 + \beta_1 \cdot AS_{iG} + \beta_2 \cdot HC_{iG} + \beta_3 \cdot HM_{iG} \tag{1}$$

The above - mentioned equations can be simplified as $\hat{G}_i = \vec{\beta}_G \cdot \vec{X}_i$ .Where coefficient vector $\vec{\beta} = [\vec{\beta_0}, \vec{\beta_1}, \vec{\beta_2}, \vec{\beta_3}]$ where $\vec{\beta_{1i}}$ represents the coefficient number for i-th country, and feature vector $\vec{X}_i = [1, AS_i, HC_i, HM_i]^T$. β_0 is the intercept term, and $\vec{\beta_1}\vec{\beta_2}, \vec{\beta_3}$ are the regression coefficients corresponding to the features. AS_i, HC_i, HM_i are the values of AS, HC, HM of the i-th country respectively. Similarly, the formulas for predicting the number of silver and bronze medals are: $\hat{S}_i = \vec{\beta} \cdot \vec{X}_{iS}$, $\hat{B}_i = \vec{\beta} \cdot \vec{X}_{iB}$.

For the total medal count, we sum up our predicted numbers of gold, silver, and bronze medals. The formula is as follows: $T_i = \hat{G}_i + \hat{S}_i + \hat{B}_i$.

The whole construction of AHH model is demonstrated in Figure 3 and pseudocode 0.


DS
$\hat{y}_{csm} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_{csm}^{t}$
Programs Dataset
Decision Tree-1
Decision Tree-2
Decision Tree-3
...
Decision Tree-N
Result-1
Result-2
Result-3
Result-N
Majority Voting
Final Result
HC
$HE = \mu + \lambda_1 \cdot N + \lambda_2 \cdot P + \epsilon$
Total Medals of United States in Different Years
Total Medals
250
200
150
100
50
0
HM
$y(t) = g(t) + s(t) + h(t) + \epsilon(t)$
Total Medals of United States in Different Years
Total Medals
$\hat{G}_i = \beta_0 + \beta_1 \cdot AS_{iG} + \beta_2 \cdot HC_{iG} + \beta_3 \cdot HM_{iG}$
$\hat{S}_i = \beta_0 + \beta_1 \cdot AS_{iS} + \beta_2 \cdot HC_{iS} + \beta_3 \cdot HM_{iS}$
$\hat{B}_i = \beta_0 + \beta_1 \cdot AS_{iB} + \beta_2 \cdot HC_{iB} + \beta_3 \cdot HM_{iB}$
$T_i = \hat{G}_i + \hat{S}_i + \hat{B}_i$

Rank	Country	Gold	Silver	Bronze	Total
1	United States	46	49	46	141
2	China	37	24	21	82
3	South Korea	25	18	26	69
4	Germany	16	16	16	48
5	Australia	16	15	16	47
6	UK	15	18	21	54
7	Japan	15	16	11	42
8	France	14	19	19	52
9	Italy	9	13	12	34
10	Netherlands	9	8	9	26

Prediction Table

Figure 3: Framework Architecture of AHH model

4.1 Data Preprocessing

Handling Missing Values: For example, the United States and other countries missed the number of medals in 1980, which was the result of their resistance against the Soviet invasion of Afghanistan that year. We employed time - series filling to handle these missing values.

Handling Outlier Values: We plotted box - plots for each country's medal rankings in Figure 4. Q1, Q2, and Q3 are the first quartile, median, and third quartile respectively, which divide the data into four equal parts. Values outside the range defined by upper = Q3 + 1.5 * IQR and lower = Q1 - 1.5 * IQR (where IQR is the inter - quartile range, IQR = Q3 - Q1) are considered outliers. Taking the top 7 countries in the 2024 gold medal ranking as an example, background research shows that their outliers often result from hosting the Olympics, indicating the host - country effect is vital for accurately predicting medal rankings as host countries usually exceed their long - term medal - winning averages.

Handling German Medal Data (1968 - 1988): We noticed that in the medal_counts table, Germany was divided into East Germany and West Germany from 1968 to 1988. To obtain the medal count for Germany during this period, we summed up the medal data of East Germany and West Germany for each corresponding year.

4.2 Advantage Score(AS)

One task of our AHH model is to consider the influence to the final result from the event arrangement every Olympic games. Every country has its own dominant sports. In these advantageous sports, athletes often have a higher level and stronger competitiveness, and the probability of winning medals is relatively high. For example, swimming and basketball in the United States, table tennis and diving in China. These countries usually win more medals in their respective advantageous events. Ac-

Algorithm 1 AHH Model

```
Input: Feature vector $X_s$: Number of events in each sport for the current Olympics.$AS$,$HM$,$N$
Step 1: Predict Advantage Score ($AS$):
Use a Random Forest model to predict $AS$ based on input features $X_s$.
Step 2: Predict Historical Medal Counts ($HM$):
Use Prophet model to predict $HM$ for the current Olympics.
Step 3: Calculate Host-Country Effect ($HC$):
Compute $HC$ using the formula:
    $$HC = \mu + \lambda_1 \cdot N + \lambda_2 \cdot P + \epsilon$$
Step 4: Fit Linear Regression Model for Host Medal Model ($AHH$):
Use data from the past 19 Olympic Games to fit the following linear regression model:
    $$\hat{G}_i = \beta_0 + \beta_1 \cdot AS_{iG} + \beta_2 \cdot HC_{iG} + \beta_3 \cdot HM_{iG}$$
    $$\hat{S}_i = \beta_0 + \beta_1 \cdot AS_{iS} + \beta_2 \cdot HC_{iS} + \beta_3 \cdot HM_{iS}$$
    $$\hat{B}_i = \beta_0 + \beta_1 \cdot AS_{iB} + \beta_2 \cdot HC_{iB} + \beta_3 \cdot HM_{iB}$$
Final Output: Predicted medal number and linear regression model for $AHH$.
```

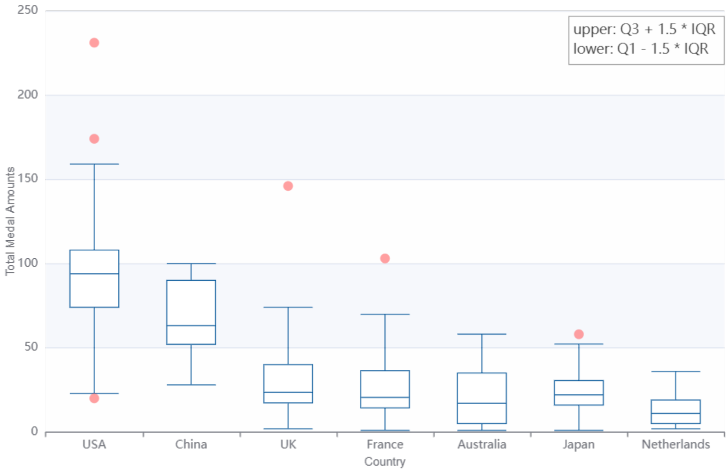


Figure 4: Boxplot of Total Medal Amounts for Multiple Countries

curately grasping the advantages of major events in various countries can more accurately predict their performance in the medal table. However, it is often challenging to exactly learn how much one country is strong in specific programs,it is abstract yet vital for predicting outcomes. Therefore, there is an urgent need for a model that can account for the varying degrees of advantage that different programs confer to a country.

We propose to employ the Random Forest model to solve this problem, due to its capability to deeply understand the effect of the arrangement of the events through the large data training, and its great mechanism brought by ensemble learning ensures the robust result.

4.2.1 Data Preparation and Feature Engineering

We first extract relevant data from a given dataset containing the counts of events by sport and discipline in all summer Olympics from 1896 to 2024, and combine it with historical country - specific medal data.

We also have historical medal data for each country c in each sport s and medal type $m \in \{G, S, B\}$ (where G represents gold, S represents silver, and B represents bronze). The ultimate task of our AS model presented below is to predict the Advantage Score (AS), this out come is crucial to our model, as the effect of the dominant program can be considered. AS can be utilized to reflect the ability for gaining the medals. Let y_{cm} be the number of medals of type m that country c won in previous Olympics, and our sub-model in this section is used to predict y_{cm} as a through the events arrangement as a form of Advantage Score (AS).

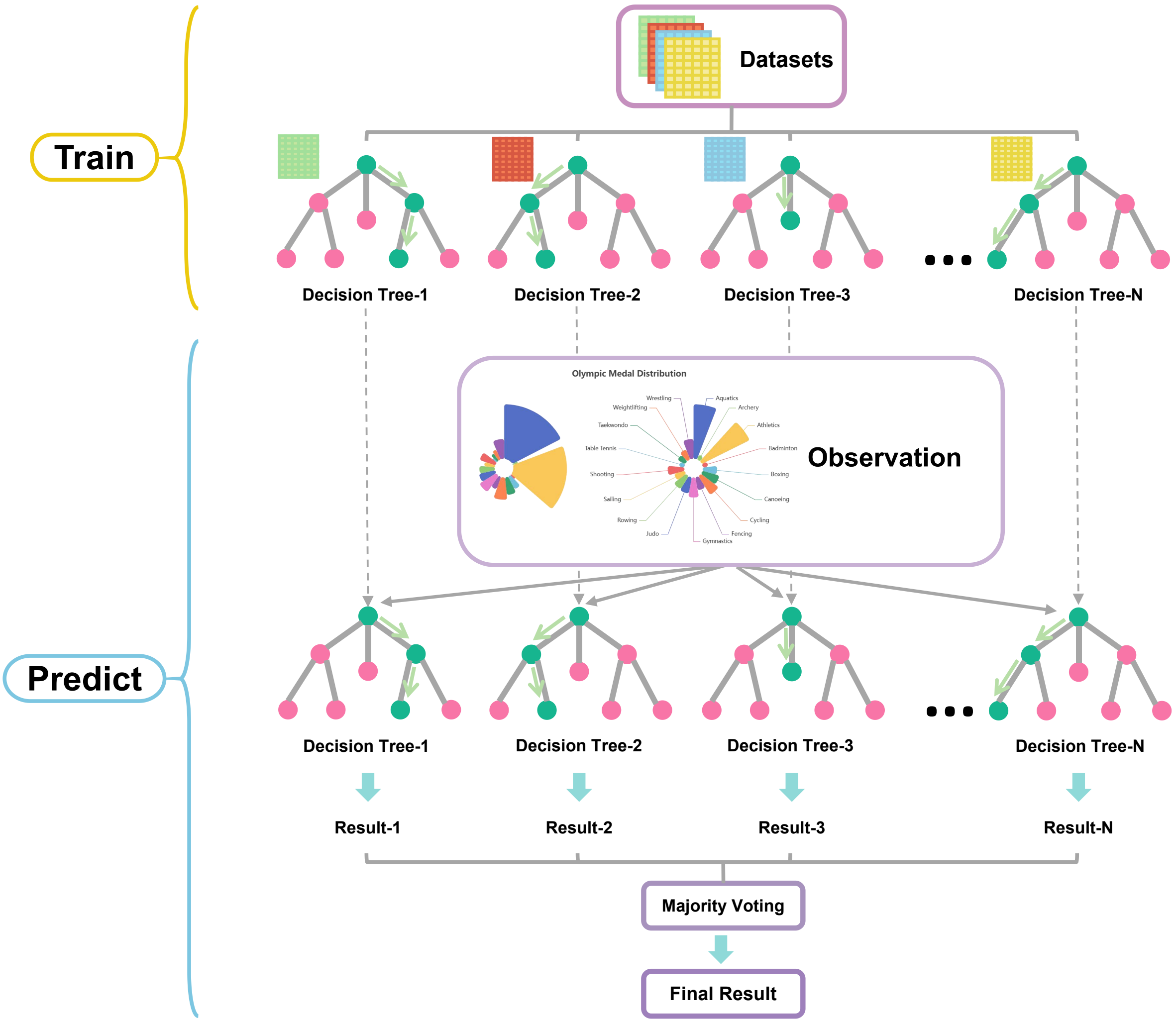


Figure 5: Principle Diagram of Random Forest

4.2.2 Random Forest Model

The Random Forest (RF) model is an ensemble of T decision trees,the basic principle is shown in Figure 5. When training, the whole datasets can be divided to be different sub-datasets which are distributed to different decision trees. Therefore decision trees in the forest learn the various situations. From different sub-datasets , different decision trees can derive different insights —— a decision tree can learn that a absence of a specific event can reduce the chance of a specific to gain the medals, otherwise another decision tree may reveal that the inclusion of a certain program could enhance the probability of achieving a favorable ranking. Therefore, each decision tree t in the forest makes a unique prediction $\hat{y}_{cm}^{t}$ for the number of medals of type m that country c will win. In the end the result of all the trees are Integrated together and all the influences of events to a specific country are learned to predict the final result.

The final prediction $\hat{y}_{cm}$ of the Random Forest model is the average of the predictions of all T decision trees:

$$\hat{y}_{cm} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_{cm}^{t} \tag{2}$$

Here, $\hat{y}_{cm}$ is the RF predicted number of medals of type m that country c will win. The averaging process helps to reduce the variance of the predictions and improve the model's generalization ability.

4.2.3 Overall Medal Count Prediction for a Country

We define

$$AS = \hat{y}_{cm} \tag{3}$$

where $\hat{y}_{cm}$ can fully consider effect to countries the project arrangement brought during training and prediction, which is significant but not enough to be considered as the final result. It is just an aspect of our model, we wrote the predicted $\hat{y}_{cm}$ as AS which we mentioned above because such a result from RF can reflect the advantages of the project schedule for this year in a concrete way.

4.3 Host-Country Effect (HC)

In addition to utilizing event-specific features applicable to all countries, incorporating country-specific features is crucial for accurately predicting the medal standings of a particular nation. Therefore, we introduce the Host-Country Effect. Generally, host nations invest substantial resources in developing sports facilities and training athletes, while also having the privilege to select or add events that favor their strengths. Moreover, host athletes benefit from familiarity with the competition environment and strong support from local audiences.

To estimate the impact of hosting the Olympics, we construct the following linear regression model:

$$HC = \mu + \lambda_1 \cdot N + \lambda_2 \cdot P + \epsilon \tag{4}$$

where N represents the number of event changes in each Olympic Games. P denotes the host - country's event participation rate, which is the proportion of all Olympic events that the host country participates in. μ is the intercept term, representing a basic host - country effect that exists without considering the influence of the variables N and P. λ_1 and λ_2 are regression coefficients, which measure the influence of N and

P on the host - country effect respectively. ϵ is the error term, including random factors that affect the host - country effect and are not considered in the model other than N and P.

We plan to use historical data from previous Olympic Games hosted by various countries to estimate the values of λ_1 and λ_2. By applying linear regression techniques to these historical data, our goal is to determine the values of λ_1 and λ_2 that minimize the total squared error between the calculated HE values and the observed impact of host countries on medal counts.

4.4 Historical Medal Counts (HM)

Our work AHH model also consider the effect brought by comprehensive national power. As it undoubtedly plays an important role in the Olympic result, nonetheless it is challenging to directly qualify this essential to make it into use for any prediction algorithm, so a proper replacement is necessary. We propose that historical medal counts can be viewed as a symbol of Comprehensive national power, the steady high score and progress in result are able to indirectly map a steady development of a country.

The medal counts of many countries exhibit long-term trends: some emerging nations with great potential may rise to prominence, while certain traditional powerhouses may gradually decline —— the Olympics often deliver unexpected outcomes. That is also an importance factor that made people always filled with anticipation for the Olympic games beyond the chance to enjoy a spectacular visual feast. Since the Olympics are held every four years, medal counts may also be influenced by cyclical factors such as preparation cycles, generational shifts among athletes, and the host country effect. Additionally, unforeseen events, such as the postponement of the Tokyo Olympics due to the pandemic, add further uncertainty to medal count variations. The interplay of these time-dependent factors makes the yearly medal rankings highly unpredictable.

Given these complexities, we propose using the Prophet model to capture the characteristics of historical medal counts and predict future outcomes. The Prophet model decomposes time series data into three fundamental components: trend, seasonality, and special event effects, making it particularly suitable for modeling medal counts influenced by long-term trends, cyclical patterns, and unexpected disruptions.

To further validate the appropriateness of the Prophet model, we performed the Augmented Dickey-Fuller (ADF) test on the medal count data. In ADF, the p-value represents the probability of the null hypothesis (the data is non-stationary) being true. Comparing it with a 0.05 significance level, we obtained a p-value of 0.1589, which is greater than 0.05. This result suggests that we cannot reject the null hypothesis, indicating that the medal count data may indeed be non-stationary. Since the Prophet model is well-suited for handling non-stationary data without requiring transformation to stationarity, this further reinforces its rationality and applicability in our study.

4.4.1 Prophet Model

We propose

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \tag{5}$$

where $y(t)$ is the number of medals a country has won at time t. $g(t)$ is the linear trend function, which describes the long - term change of the time series as follows:

$$g(t) = (k + \boldsymbol{\delta}^T \mathbf{a}(t))t + (m + \boldsymbol{\gamma}^T \mathbf{a}(t)) \tag{6}$$

where k is the basic growth rate, δ is the adjustment vector of the growth rate at the change point, $\mathbf{a}(t)$ is a vector of indicator functions used to mark whether a change point occurs, m is the intercept, and γ is the adjustment vector of the intercept at the change point.

$s(t)$ is the seasonal function, used to capture the periodic changes in the time series. For annual seasonality, it is usually expanded using a Fourier series:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{365.25}\right) + b_n \sin\left(\frac{2\pi nt}{365.25}\right)\right) \tag{7}$$

where a_n and b_n are Fourier coefficients, and N is the order of the Fourier series.

$h(t)$ is the holiday function, used to consider the impact of specific holidays or special events on the time series. Suppose we have L holidays, each with a corresponding time window, then:

$$h(t) = \sum_{l=1}^{L} \kappa_l \mathbf{1}(t \in D_l) \tag{8}$$

where κ_l is the impact coefficient of the l - th holiday, and $\mathbf{1}(t \in D_l)$ is an indicator function that takes the value 1 when time belongs to the time window D_l of the l - th holiday, and 0 otherwise.

$\epsilon(t)$ is the error term, usually assumed to follow a normal distribution:

$$\epsilon(t) \sim N(0, \sigma^2) \tag{9}$$

4.5 Validation and Parameter Estimation

To validate the performance of the AHH model in medal prediction, we integrated the three core features mentioned earlier: AS, HC, and HM. Based on these features, we conducted parameterized adjustments and constructed a model to predict the gold medal ranking for the 2028 Olympic Games. Table 2 presents the top ten countries in the predicted 2028 gold medal ranking output by the model and their corresponding parameters.

Using the data from the past 19 Summer Olympics, we employed a multiple linear regression model to fit the weights of each feature. Specifically, the model estimated the parameters through the Ordinary Least Squares (OLS) method, aiming to minimize the Mean Squared Error (MSE) between the actual medal data and the predicted values. Moreover, to mitigate the influence of multicollinearity among features, we adopted Ridge Regression for regularization, thereby enhancing the stability of parameter estimation.

From the result we can visually learn the importance weight of as to different countries. Taking the United States as an examplethe weights of it tends to balance, indicating that the influence of each factor is equally important, combining the outstanding

Table 2: Regression Coefficients and Bias Terms by Country

Country	β_0	β_1	β_2	β_3
United States	0.1036	1.2270665	1.56078422	1.2270665
Australia	-0.6895	0.4986528	17.18567948	0.4986528
China	-2.1163	0.50450247	2.85592263	0.50450247
South Korea	1.2194	0.58531698	-0.2555053	0.58531698
UK	-0.3791	0.44473451	22.36507219	0.44473451
Germany	0.5133	0.47363262	4.29180925	0.47363262
France	7.7425	0.35283314	0.65779497	0.35283314
Japan	1.3190	0.43642739	0.0	0.43642739
Italy	-0.6549	1.48089976	1.57248883	1.48089976
Netherlands	-3.1870	0.55046423	3.98255792	0.55046423
...	...	...	...	...

performance of the United States in the Olympics, it is inferred that the strong overall national strength and the wide distribution of advantageous projects, along with the opportunity to host the Olympics multiple times, have enabled the United States to consistently achieve impressive results in the Olympic arena.In comparison, the β_2 from the dimension of the UK is significantly larger, indicating that its strong overall national strength is the key to its success. On the other hand, β_1 at South Korea's dimension is relatively large, suggesting that South Korea relies more on its advantageous events in the Olympics.

4.6 Uncertainty Estimation and Accuracy

We Use the 0.95 prediction interval for uncertainty estimation, the specific medal table and prediction interval are shown in Figure 6.

$$\hat{y}_0 \pm t_{1-\alpha/2,n-k-1}\sqrt{MSE}\sqrt{1 + X_0^T(X^TX)^{-1}X_0} \tag{10}$$

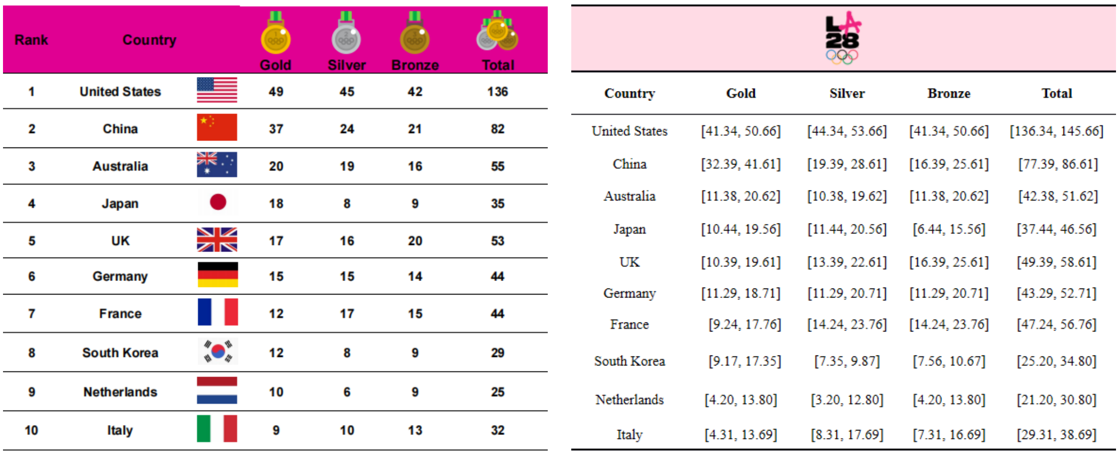
We reflect the accuracy of the AHH model by calculating the coefficient of determination (R^2).$\bar{y}$ represents the mean of true values, and $\hat{y}$ represents predicted values. This provides a more comprehensive evaluation of the model's accuracy in predicting the medal table.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

5 RQ1.2: Estimate First - time Medal Winners

5.1 Data Analysis and Prediction Methods

The following Figure 7 demonstrates the process in which we took non - medal - winning countries into account during the modeling.



Rank	Country	Gold	Silver	Bronze	Total
1	United States	49	45	42	136
2	China	37	24	21	82
3	Australia	20	19	16	55
4	Japan	18	8	9	35
5	UK	17	16	20	53
6	Germany	15	15	14	44
7	France	12	17	15	44
8	South Korea	12	8	9	29
9	Netherlands	10	6	9	25
10	Italy	9	10	13	32

(a) Medal Table

Country	Gold	Silver	Bronze	Total
United States	[41.34, 50.66]	[44.34, 53.66]	[41.34, 50.66]	[136.34, 145.66]
China	[32.39, 41.61]	[19.39, 28.61]	[16.39, 25.61]	[77.39, 86.61]
Australia	[11.38, 20.62]	[10.38, 19.62]	[11.38, 20.62]	[42.38, 51.62]
Japan	[10.44, 19.56]	[11.44, 20.56]	[6.44, 15.56]	[37.44, 46.56]
UK	[10.39, 19.61]	[13.39, 22.61]	[16.39, 25.61]	[49.39, 58.61]
Germany	[11.29, 18.71]	[11.29, 20.71]	[11.29, 20.71]	[43.29, 52.71]
France	[9.24, 17.76]	[14.24, 23.76]	[14.24, 23.76]	[47.24, 56.76]
South Korea	[9.17, 17.35]	[7.35, 9.87]	[7.56, 10.67]	[25.20, 34.80]
Netherlands	[4.20, 13.80]	[3.20, 12.80]	[4.20, 13.80]	[21.20, 30.80]
Italy	[4.31, 13.69]	[8.31, 17.69]	[7.31, 16.69]	[29.31, 38.69]

(b) Prediction Intervals Table

Figure 6: 2028 Los Angeles Olympics Medal Prediction

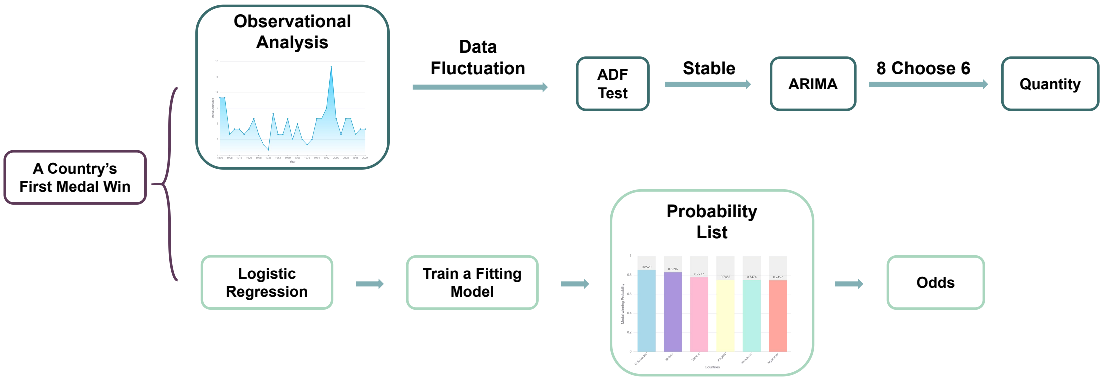


Figure 7: Flow Chart of Prediction Model

5.2 Quantity Prediction

As shown in Figure 8, we have compiled the number of countries that won their first Olympic medals in each Olympic Games from 1896 to 2024. We found that there always exist the countries that obtain their historical first medal every year. The number of them tend to be different.The number of countries that win medals for the first time usually has a certain time dependency in relation to historical data. For example, trends in global sports development and the promotion of regional sporting events may cause the number of first-time medal-winning countries in a specific year to be related to data from previous years. We are inspired to predict the number of the first medal countries by the time series prediction model.

To choose the suitable time series prediction model, we still firstly use the ADF test to estimate the stationarity of this data. Statistical value of ADF produced by the test is -4.0455686, with a p-value of 0.00686. Since the p-value is far below the typical significance level, we reject the null hypothesis and the data is stationary.

As a result, we propose to use the ARIMA model to predict the number of countries that will win their first-ever medals in 2028. ARIMA is a classic time series analysis method, suitable for handling historical data-driven problems such as the number of countries that first won medals, and for the aforementioned naturally stationary data like that, the ARIMA model can effectively capture its autocorrelation and random fluctuation characteristics, thereby achieving efficient modeling and accurate forecast-

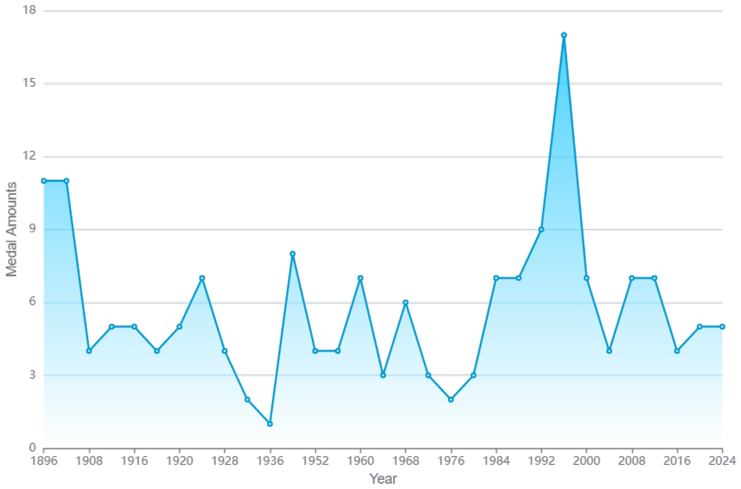


Figure 8: The Number of First - Medal - Winning Countries in Successive Olympics

ing.

We predict that there will be 6 first-medal countries by using ARIMA. Nevertheless, we found it difficult to predict the detailed list of the countries, because they all have not won any medals.

5.3 Probability Analysis

In order to exactly master the situations of the non-medal countries, we introduced two factors: the number of the events a country participate and the number of times a country has participated, as the more times a country participates in the Olympics and the more events it competes in, even if it does not win any medals, it can still reflect a strong overall national strength. Additionally, it accumulates rich competition experience, which gives it great potential for winning awards in the future.

Next, we construct a logistic model to calculate the probability of each country winning its first-ever medal as it is a suitable way to take into account the factors. The details are displayed in Figure 9.

$$P = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 \cdot E + \gamma_2 \cdot T + \gamma_3 \cdot H)}} \tag{12}$$

In our Logistic Model, E represents the number of events they participated per Olympics, T denotes the number of times a country has participated, and H indicates whether it is the host nation,γ_0 represents the distance, and γ_1, γ_2, γ_3 represent the coefficients.

Although most host countries have already won many medals, we considered the host factor for rigor by adding a binary variable H. $H = 1$ indicates host countries, while $H = 0$ indicates non-host countries.

We employ the data including E,T and H data of the ground truth first-medal country in historical Olympic games from 1952-2010 as the training data, and use the ones from 2012-2024 as test data. The RMSE-loss of our model is 0.489, achieving the high precision. The result that γ_1 = 0.328 and γ_2 = 0.542 indicates that the number of times a country participates in the Olympics may be more important for achieving a breakthrough in the medal number, because the comprehensive national strength and experience of a country capable of participating in competitions multiple times can indeed be guaranteed.

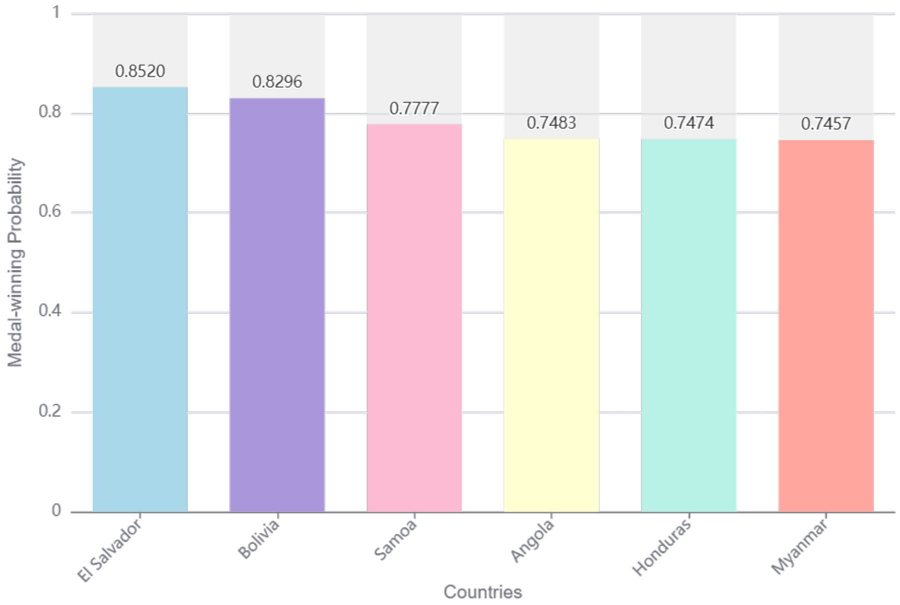


Figure 9: Predicted Probability List of Countries Winning Medals for the First Time in 2028

6 RQ1.3: Analyze Events' Impact on Medals

6.1 Correlation Analysis

First, we chose to analyze the correlation between the overall performance of various countries and the setup of sports events. Due to the interaction of multiple complex factors, this correlation is typically nonlinear. Therefore Spearman correlation coefficients is employed to capture the complex relationship [1].

We used all Olympic events and countries as inputs to compute the Spearman correlation coefficients between them. The equation of it is shown:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

- ρ: Spearman correlation coefficient.
- n: The number of data samples.
- d_i: The rank difference for each observation, $d_i = \text{rank}(x_i) - \text{rank}(y_i)$.
- $\sum_{i=1}^{n} d_i^2$: The sum of squared rank differences.

Then, we drew a heat map (Figure 10), in which a color closer to red indicates that the current sport is more relevant to the country, while a color closer to blue implies the opposite. By analyzing the map, we identified the most relevant sports for different countries. For example, the most relevant sports for China are swimming and diving, while for Great Britain, they are athletics and triathlon.

In spite of this, we just initially learn the correlation between the overall performance of various countries and the arrangement of events and we still cannot directly determine which sports are the most important for specific countries. Nonetheless, such analysis results can strongly assist us in conducting importance analysis for the models we establish in the following sections.

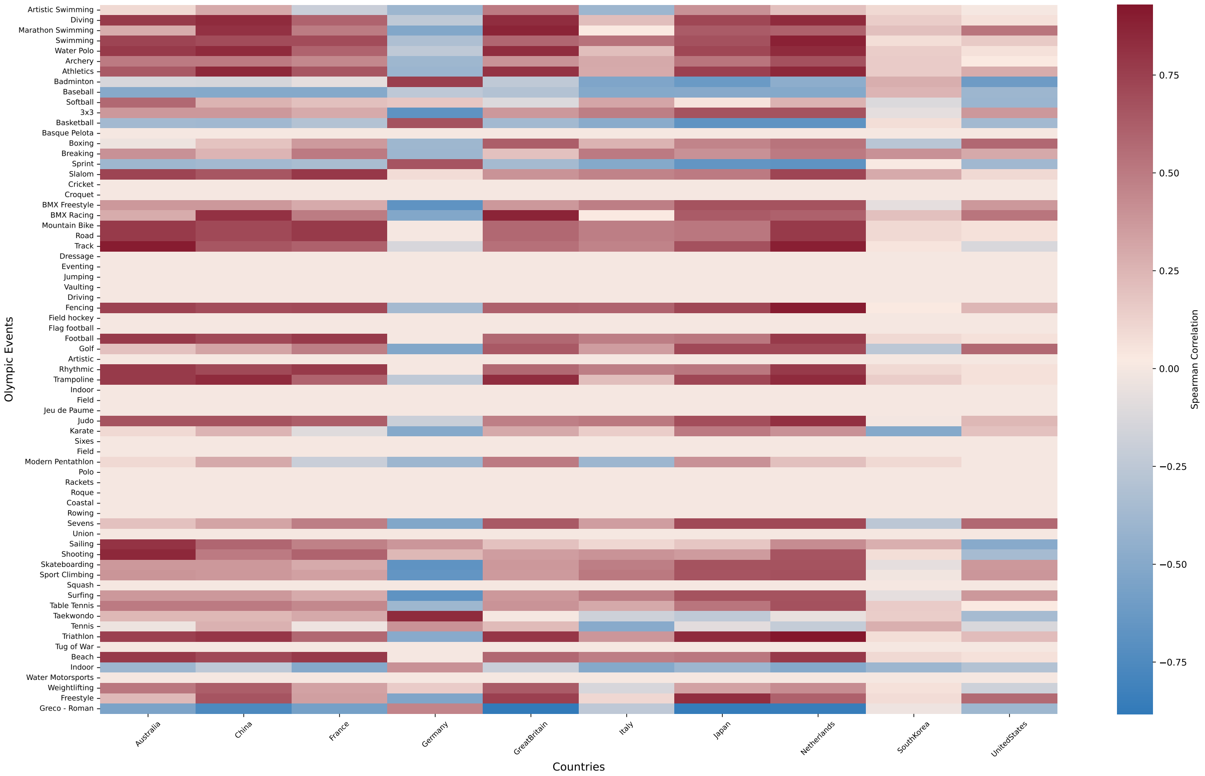


Figure 10: Heatmap of Olympic Events and Countries Correlation

6.2 Importance Estimation Algorithm

We have just explored the Correlation between event arrangement and performance of countries, which can be a important reference to our result. The main task of our model is to capture the importance of different sports to specific countries.

The framework architecture is shown in Figure 11. To begin with, let the influence coefficient of the year be w_y. The further back in time, the smaller the influence factor, because objective conditions such as competition rules and event settings change over time. Therefore, earlier data has less impact on the model's predictions. We selected the events which exists from 2000 to 2024 to analyze, starting from 2000 with $w_{2000} = 0.1$, w_y increases by 0.1 for each subsequent Olympics, etc. The model formula is:

$$\mathrm{E}_t = \sum_{i=1}^{n} w_i \times (5 * G_i + 3 * S_i + B_i) \tag{13}$$

where G_i represents the number of gold medals with a weight of 5, S_i represents the number of silver medals with a weight of 3, and B_i represents the number of bronze medals with a weight of 1. w_i ensures that more recent years contribute more to E_i, while earlier years contribute less.

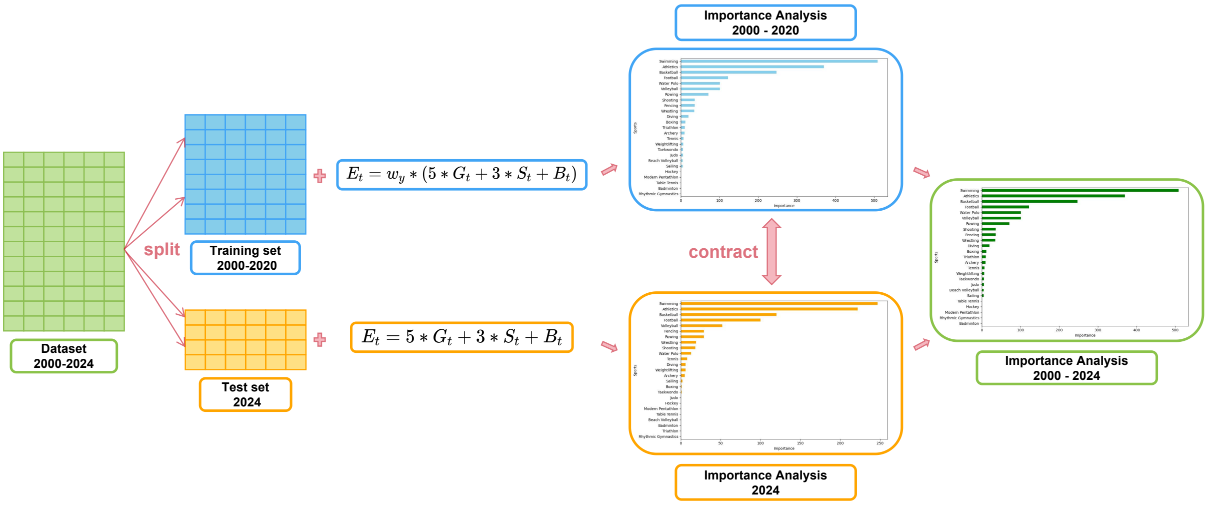


Figure 11: Framework Architecture of Model

To begin with, we used the sports event data from the 2000 - 2020 Olympics as a training set. In our model the equation $E_t = 5 * G_t + 3 * S_t + B_t$ is utilized to consider the different importance of different kind of medal (without considering the year factor) and this equation is also used to Simulating the importance of various projects in the year 2024 alone. Based on the estimated annual performance, we further introduced the model $E_t = w_y \times (5 * G_t + 3 * S_t + B_t)$ that considers the year factor. To sum up our importance prediction model is to emphasize the influence of the advantageous events and the development of Olympic competition. If a record of event to country is relatively good at one competition, the event must be important as it can indeed help a country reach the good rank.On the other hand, if a country does not win a medal, but it ever performed well at earlier competition in recent year, our model can also indicate its significance, because it means that the country reveals the potential to make an achievement at that event.

After that, by comparing the predicted results with the actual data of 2024, we found our model could accurately reflect the importance of different sports to specific countries.

Finally we incorporated all the data from 2000 - 2024 into the analysis. Figure 12 shows the results of the United States.

6.3 Host Effect

The host country holds a certain degree of influence in the selection of Olympic events. It can choose to include or add events in which it has an advantage and remove those where its edge is not prominent. Such event - selection decisions may exert an impact on the competition results. For example, it can increase the opportunities for the host country's athletes to win medals and may also reshape the competitive landscape of other countries across different events. When the year factor is taken into account, the influence of the host country's event selection in recent Olympic Games may be more conspicuously reflected in E_t.

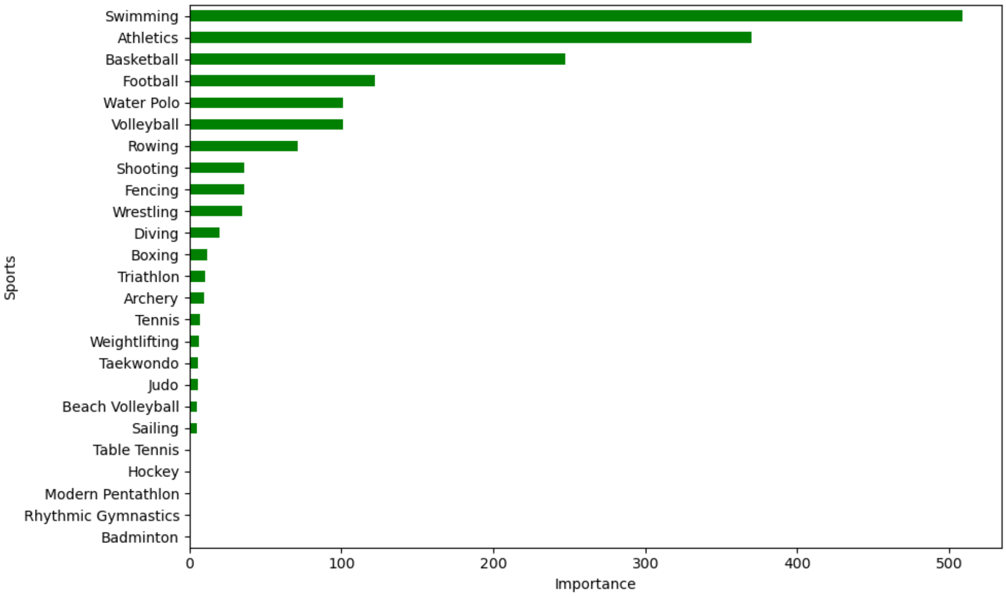


Figure 12: Importance of Sports for US Olympic Medals

7 RQ2: Quantify "great coach" Effect

7.1 Real World Example

- In 2012, **Karch Kiraly** became the head coach of the US women's volleyball team. He led them to win bronze in 2016 Rio Olympics, claim the first Olympic gold in 57 years in 2021 Tokyo Olympics, and secure silver in 2024 Paris Olympics.
- After hiring **Béla Károlyi**, the US women's gymnastics team saw remarkable improvement. In 1984 Los Angeles Olympics, his disciple Mary Lou Retton won the individual all - around gold, a major Olympic breakthrough. In 1992 Barcelona Olympics, as head coach, he helped the team claim the team bronze, their first in eight years. In 1996 Atlanta Olympics, the Károlyi couple led them to team gold.
- In 2022, the Chinese rhythmic gymnastics team hired Russian coach **BLIZNYUK Anastasia**, who led them to win the team all - around gold at the 2024 Paris Olympics, a historic zero - gold breakthrough. The Chinese women's boxing team hired a Cuban coaching team led by **Fernandez Liranza Raul Angel**, achieving 3 golds and 2 silvers, also a zero - gold breakthrough and doubling the medal count. In 2023, the Chinese synchronized swimming team hired **Anna Tarrés** and won their first gold in 2024.
- In 2003, the British cycling team hired **Dave Brailsford** as performance director, ending a 110 - year championship drought. At the 2008 Beijing Olympics, they achieved remarkable success, winning 8 gold medals, which accounted for 60% of the golds in cycling events. n 2012 London Olympics, on home turf, they broke 9 Olympic and 7 world records and won 8 golds again, continuing their success.
- After **Misbun Sidek** coached the Indian badminton team, Indian athletes achieved a zero - medal breakthrough in the Olympics. PV Sindhu won the silver medal in the women's singles badminton at the 2016 Rio Olympics and the bronze medal at the 2021 Tokyo Olympics.

By observing the experiences of these four countries in hiring renowned coaches, it is evident that they have made remarkable achievements in the Olympics . The contributions of outstanding coaches to a team are immense. They have helped teams reverse their mediocre performance that has persisted for nearly a century, achieved zero - breakthroughs in gold medals or medals in the team's history, and even enabled teams to win more medals, consolidating the dominant position of the sports. Meanwhile, the advanced training concepts and management methods of outstanding coaches have had a profound impact on the teams.

We adopt the Difference-in-Differences (DID) method to quantify the contribution of outstanding coaches to medal counts.

7.2 Data Collection

1. **Treatment group (USA women's volleyball team and Chinese Women's volleyball team)**

   Before Lang Ping's coaching: At the 2012 Olympic Games, the Chinese women's volleyball team did not win any medals, which is recorded as 0 medal points (assuming 6 for gold medal, 3 for silver medal, 1 for bronze medal, and 0 for no medal).

   After Lang Ping's coaching: At the 2016 Olympic Games, the Chinese women's volleyball team won the gold medal, which is recorded as 6 points.

   Before Karch Kiraly's coaching:At the 2012 Olympic Games, the US women's volleyball team won the silver medal, which was recorded as 3 points.

   After Karch Kiraly's coaching:At the 2016 Olympic Games, the US women's volleyball team won the bronze medal, which was recorded as 1 points.

2. **Control group (Serbian Women's volleyball team and Dutch Women's volleyball team)**

   At the 2012 Olympic Games, neither the Serbian women's volleyball team nor the Dutch women's volleyball team won any medals, and both were recorded as 0 points. At the 2016 Olympic Games, the Serbian women's volleyball team won the silver medal and was recorded as 3 points, while the Dutch women's volleyball team didn't win any medals and was recorded as 0 points.

7.3 Parallel Trends Test

During the period from 2012 to 2016, the world women's volleyball landscape was relatively stable. The strength improvement or fluctuation of top teams mainly relied on internal factors such as their own training systems and player development. There were no significant external shocks (such as major rule changes or the sudden emergence of new strong competitors) that had an unbalanced impact on all teams. Therefore, the change trend of the number of medals of the women's volleyball teams of the above four countries may conform to the parallel trend assumption.

7.4 DID Estimation

Calculate the medal count change of two groups by formula: Medal score in 2016 minus medal score in 2012.

For the Chinese women's volleyball team, it is 6 - 0 = 6. For the USA women's volleyball team, it is 1 - 3 = -2. Take the average of the two values, which is 2, as the data for the treatment group.

For the Serbian Women's volleyball team, it is 3 - 0 = 3. For the Dutch Women's volleyball team, it is 0.Take the average of the two values, which is 1.5, as the data for the control group.

The DID estimate = (Medal count change of the treatment group) - (Medal count change of the control group) = 2 - 1.5 = 0.5.

For ease of understanding, the above process is shown in Figure 13.

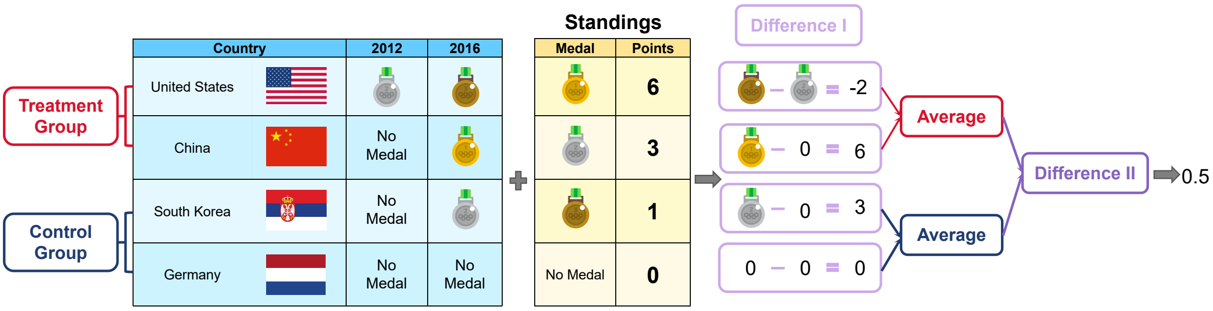


Figure 13: Visualization of an Instance within the DID Method

7.5 Result Analysis

The DID estimate is 0.5, which means that after controlling for the influence of other common factors (such as changes in the Olympic competition environment and the overall development trend of volleyball), the factor of Lang Ping coaching the Chinese women's volleyball team led to a relative increase of 1 levels in the medal count of the Chinese women's volleyball team at the 2016 Rio de Janeiro Olympics, Lang Ping's coaching had a significant positive contribution to the performance improvement of the Chinese women's volleyball team in those Olympic Games.

Specifically quantifying the contribution of outstanding coaches to the number of medals, combined with the examples cited and the medal tables of previous Olympic Games, after hiring outstanding coaches, teams have achieved a breakthrough of one or two gold medals, and the total number of medals has increased by **5%** to **8%**.

Based on the first question, we have predicted the six countries most likely to win their first medals in 2028 through our model. Now, we will select the three countries most likely to achieve awards from this group for analysis, as they have the greatest potential for breakthroughs in certain events. By analyzing the official data , we found that Bolivia has participated the most in the Athletics Men's Marathon event, and in addition, the country is also quite active in other athletics events. El Salvador dominates in the number of participations in swimming events. Lastly, Samoa has a unique presence in the Athletics Men's Discus Throw, with significantly fewer participations in other events. Therefore, Bolivia could introduce an excellent head coach for the athletics category or a coach specifically for the Athletics Men's Marathon event. El Salvador could bring in a top coach for swimming events, and finally, if Samoa recruits an outstanding coach for the Athletics Men's Discus Throw, they are likely to achieve good results.

8 RQ3: Guide NOCs' Future Strategies

Our model also reveals the impacts of the dynamic changes in dominant events, personalized development strategies, alterations in the global sports landscape, and the rise of emerging sports forces on the number of Olympic medals.

Changes in dominant programs: The random forest model can identify the dynamic changes in dominant sports events. The development of sports technology, the innovation of training methods, and the evolution of the international sports landscape can lead to the decline of traditional dominant events and the rise of emerging events. Analyzing these changes can help predict the focus of medal contention in the future and provide a reference for sports strategic planning and resource allocation.

Differences in model applicability: Multiple regression and various characteristic models have different applicability to different types of countries. For sports powers, the total number of medals is mainly influenced by dominant events and the number of medals won over the years, while the host - country effect has a relatively small impact. For countries with lagging sports development, when hosting the Olympic Games, the host - country effect may become the key to a breakthrough in the number of medals. Understanding the differences in model applicability can help countries formulate personalized sports development strategies.

Evolution of the Global Sports Landscape: The prediction of the number of countries winning medals for the first time by the ARIMA model reflects the trend of the global sports landscape. An increase in the number indicates that global sports development is becoming more balanced, breaking the monopoly of traditional sports powers. A decrease, on the other hand, means that the sports landscape is becoming more solidified, making it more difficult for emerging countries to achieve a medal breakthrough. Analyzing this trend can help evaluate the effectiveness of international sports organizations' policies and the sports development strategies of different countries.

Emerging Sports Forces: The logistic model calculates the probability of winning medals for the first time and reveals the rising path of emerging sports forces. By analyzing the weights of variables in the model, we can identify the key factors affecting the probability of winning medals for the first time. If the number of events participated in and the number of Olympic Games participated in are key factors, it points out a clear direction for countries seeking a breakthrough. They can increase support and improve the sports talent cultivation system to increase the probability of winning medals for the first time.

It can provide the following suggestions for the Olympic committees of various countries:

Optimize athlete selection and training: Based on the impact of the age structure on the medal count, when selecting and training athletes, the Olympic committees should focus on building a reasonable age echelon. They should not only pay attention to potential young players and provide them with more opportunities to participate in competitions to accumulate competition experience, but also arrange the competition events and time of veterans reasonably to give full play to their experience advantages.

Increase investment in scientific research: Recognizing the importance of scientific research investment in increasing the medal count, the Olympic committees of various countries should increase investment in sports scientific research, cooperate with scientific research institutions, and use advanced technologies and methods to improve the training quality and competition performance of athletes. This can help them gain an advantage in the fierce Olympic competition.

9 Sensitivity Analysis

We conduct a sensitivity analysis to evaluate the robustness of our proposed model, focusing on two key aspects: (1) adjusting the model's prediction interval, and (2) modifying the training dataset by randomly sampling 80% and 60% of the original data.

The altered information gain is shown below. In the left graph, extending the prediction interval shifts the model's focus toward long-term factors, such as "historical medal counts," while reducing the weight of short-term effects like the "host country effect." The right graph reveals notable changes when the model is trained on 60% of the data, reflecting its sensitivity to data quantity, whereas training with 80% ensures greater stability, indicating a threshold for data sufficiency.

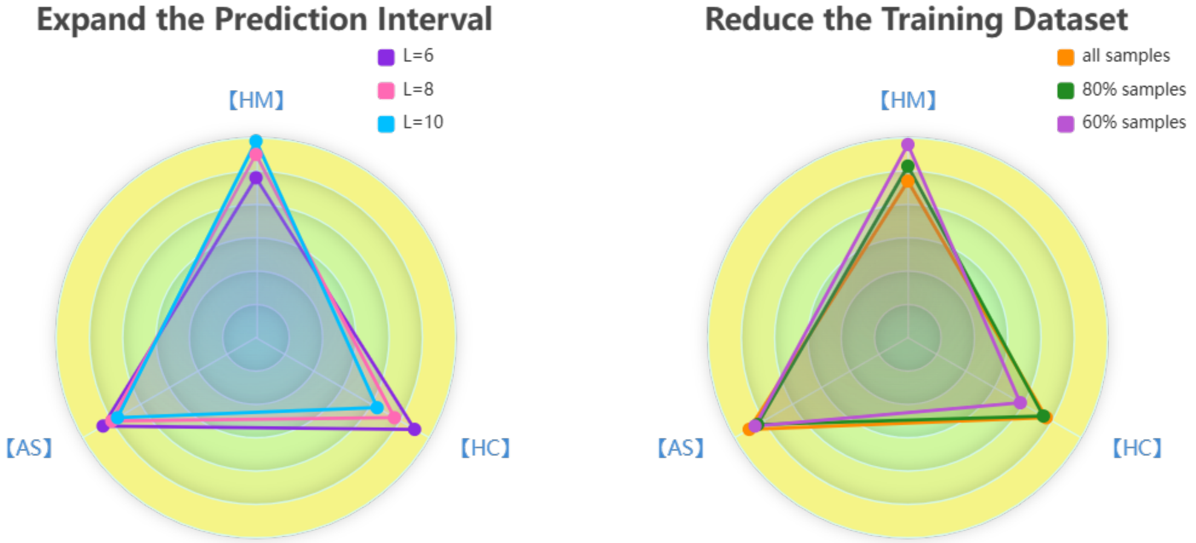


Figure 14: Radar Charts for Sensitivity Analysis

10 Strengths and Improvements

10.1 Strengths

- The multi-model fusion method analyzes the correlations between data features and medal counts from multiple dimensions. It significantly improves the comprehensiveness and reliability of feature predictions, laying a solid foundation for the final medal count prediction.
- By finely optimizing the parameters of the multiple regression model, the accuracy and credibility of the final medal count prediction results are greatly enhanced, increasing the model's practical application value.
- Innovatively combining ARIMA and Logical models for prediction analysis, the former analyzes trends in time - series data, and the latter calculates medal -

winning probabilities. Their complementary functions enhance the integrity and reliability of predictions.

10.2 Improvements

- Combining multiple models could pose challenges in model integration and calibration, potentially affecting overall performance. And the models may have limited adaptability to unforeseen external factors.
- The models might not fully capture the impact of emerging factors on future outcomes.

11 Conclusion

1. We construct AHH model to predict the medals that each country can win. Then we estimate the first-time medal country count by prophet model and use logistic regression model to calculate the breakthrough probability for the potential countries. Then Spearman correlation coefficient is employed to capture the country-event relationship and we construct a model to predict the event importance to specific country.Through the previous data, we capture the influence of effectiveness.
2. We utilized the DID method to measure the impact of exceptional coaches, demonstrating that hiring great coaches can lead to a breakthrough of one or two gold medals and an overall medal increase of 5% to 8%.
3. Our models reveal the evolving dynamics of dominant events, the significance of tailored strategies for different countries, and the implications of global sports balance. By leveraging methods such as Random Forest, ARIMA, and logistic regression, we identify key drivers of medal performance, including advancements in sports technology, the host-country effect, and first-time medal opportunities. These findings underscore the importance of optimizing athlete selection, refining training approaches, and investing in sports research and development to achieve competitive success.

References

[1] Ali Abd Al-Hameed, K. (2022). Spearman's correlation coefficient in statistical analysis. International Journal of Nonlinear Analysis and Applications, 13(1), 3249-3255.

[2] Humphreys, B. R., Johnson, B. K., Mason, D. S., & Whitehead, J. C. (2018). Estimating the value of medal success in the Olympic Games. Journal of Sports Economics, 19(3), 398-416.

[3] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer International Publishing.

[4] Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution–a socioeconomic machine learning model. Technological Forecasting and Social Change, 175, 121314.

[5] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.