International Business & Economics Studies
ISSN 2640-9852 (Print) ISSN 2640-9860 (Online)
Vol. 8, No. 1, 2026
www.scholink.org/ojs/index.php/ibes

Original Paper

Medal Prediction and Evaluation Model Based on Random Forests and a Difference-in-Differences with Multiple Time Periods Method

Xinyi Zhang[1,a]

[1] Xi'an International Studies University, Xi'an, China

[a] 13685291580@163.com

Received: December 24, 2025 Accepted: February 28, 2026 Online Published: March 10, 2026

doi:10.22158/ibes.v8n1p266 URL: http://dx.doi.org/10.22158/ibes.v8n1p266

Abstract

As people's attention to sports events increases, while watching individual events in the Summer Olympics, fans are also paying more and more attention to each country's overall Olympic medal standings. Based on this, this article is dedicated to deeply analyzing the historical data of the Summer Olympic medal standings and the factors affecting the counting, as well as predicting future development trends, to help countries formulate reasonable strategies for the development of sports events and provide potential information for the International Olympic Committee.

For question 1, first, pre-process the data by removing 1,484 duplicate values, etc. Second, construct a random forest regression prediction model, select a total of 6 feature values including historical medal count, host or not, dominant sport, number of Olympic participation sessions, number of participating athletes, and number of participating events, and predict the gold, silver, and bronze medals and the total number of medals in the 2028 Summer Olympics in Los Angeles, USA. Third, using the K-means clustering algorithm, the countries that will win their first medal at the next Olympics are Angola, Antigua and Barbuda, Bolivia, El Salvador, Guam, Honduras, Liechtenstein, Madagascar, Mali, Malta, Myanmar, Nicaragua, Nepal, Papua New Guinea, Samoa and Seychelles, with an SSE value of 604.70 and a silhouette coefficient of 0.69 , indicating that the model clustering effect is good and stable, and the odds are only 31.00 %. Finally, by screening the sports with the largest number of medals in each country and a significantly larger number of medals than other events, it is determined that these are the most important sports in that country. In addition, it can be seen from the host country feature value of 0.06 that the host country effect has a positive and promoting effect on the number of medals.

For question 2, first, using the multi-period double difference method, the weighted average modular *processing behavior results were used to explore the contribution values of the “great coach” effect to the number of medals in the volleyball and gymnastics events. The results showed that on average, each competition added approximately 3.74 and 5.15 medals, respectively, and the p-value was close to 0, indicating that the model as a whole was statistically significant. Next, we screened out three countries that had participated in a relatively large number of Olympic Games, with a large number of participating athletes and events, but had not won any medals. We suggested that they consider investing in the “great coach” effect. We believe that Canada and Tunisia should invest in the “great coach” effect of the volleyball project, so that their volleyball project medals will increase by 3.74 medals each year, while Australia should consider investing in the “great coach” effect of the gymnastics project, so that their gymnastics project medals will increase by 5.15 medals each year.*

For question 3, according to the model solution results of the random forest regression prediction model, the K-means clustering algorithm, and the multi-period double difference method, it can be seen that the number of participating athletes, the number of events, and the number of Olympic Games attended, etc., will have a significant impact on the medal count. Therefore, this can provide reference and reference for the IOC in terms of competitive talent selection, sports event development planning, long-term participation planning, and experience inheritance and innovation, etc..

Keywords

the fuzzy comprehensive evaluation method,the gray prediction model, a ridge regression model, the entropy weight-TOPSIS method

1. Introduction

1.1 Problem Background

As a key indicator of a country’s athletic strength, the medal standings in the Olympic Games, a major sporting event that attracts worldwide attention, carry great significance and are one of the core focuses of public attention.From a statistical point of view, the medal table visualizes the results of athletes from various countries in the event in a quantitative form, which not only reflects the athletes’ individual competitive level, but also comprehensively reflects the scientificity and effectiveness of a country’s sports talent cultivation system, as well as the degree of perfection of the sports infrastructure and other elements.

1.2 Restatement of the Problem

We need to analyze the Olympic medal table using the official data provided by COMAP and answer the following questions：

1.Develop a model to predict the total number of medals and the number of gold, silver, and bronze medals that countries will win in future Olympic Games and estimate the uncertainty and accuracy of the prediction.

(1) Predict the medal table for the 2028 Los Angeles Olympics based on the model, including trends in the performance of each country and which countries are expected to advance or regress.

(2) For countries that have not yet won a medal, project how many countries will win a medal for the first time in 2028 and assess the reliability of the projections.

(3) Explore the relationship between the number and type of events and the number of national medals, and analyze the performance of different countries in specific events.

2. Evaluate the impact of the "great coach effect" on medal tables by selecting three countries and analyzing their potential medal gains from investing in great coaches in specific sports.

3. Provide other insights related to Olympic medal counts to help the IOC make decisions.

1.3 Our Work

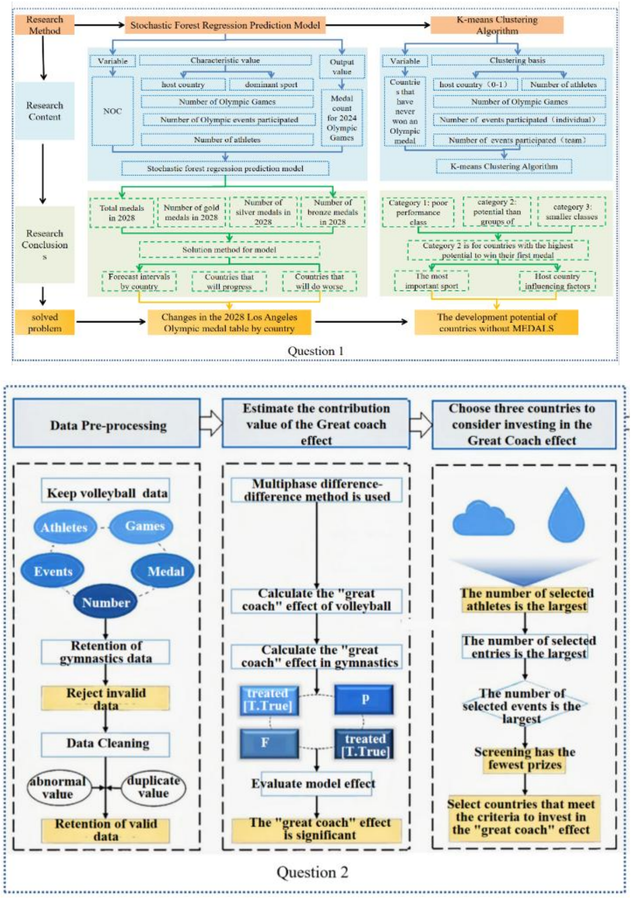
Research Method
Stochastic Forest Regression Prediction Model
K-means Clustering Algorithm
Research Content
Variable
Characteristic value
Output value
NOC
host country
dominant sport
Number of Olympic Games
Number of Olympic events participated
Number of athletes
Medal count for 2024 Olympic Games
Stochastic forest regression prediction model
Variable
Clustering basis
Countries that have never won an Olympic medal
host country（0-1）
Number of athletes
Number of Olympic Games
Number of events participated（individual）
Number of events participated（team）
K-means Clustering Algorithm
Research Conclusions
Total medals in 2028
Number of gold medals in 2028
Number of silver medals in 2028
Number of bronze medals in 2028
Solution method for model
Forecast intervals by country
Countries that will progress
Countries that will do worse
Category 1: poor performance class
category 2: potential than groups of
category 3: smaller classes
Category 2 is for countries with the highest potential to win their first medal
The most important sport
Host country influencing factors
solved problem
Changes in the 2028 Los Angeles Olympic medal table by country
The development potential of countries without MEDALS
Question 1
Data Pre-processing
Estimate the contribution value of the Great coach effect
Choose three countries to consider investing in the Great Coach effect
Keep volleyball data
Athletes
Games
Events
Medal
Number
Retention of gymnastics data
Reject invalid data
Data Cleaning
abnormal value
duplicate value
Retention of valid data
Multiphase difference-difference method is used
Calculate the "great coach" effect of volleyball
Calculate the "great coach" effect in gymnastics
treated [T.True]
P
F
treated [T.True]
Evaluate model effect
The "great coach" effect is significant
The number of selected athletes is the largest
The number of selected entries is the largest
The number of selected events is the largest
Screening has the fewest prizes
Select countries that meet the criteria to invest in the "great coach" effect
Question 2

Figure 1. Our Work

2. Assumptions and Justification

In order to rationally simplify the modeling process, the following assumptions are made.

Assumption1： The assumption of characterizing the variables in the 2028 Los Angeles Summer Olympics medal prediction evaluation model with the country code NOC as recorded by the National Olympic Committee eliminates the effect of non-national region information included by the team on the model solution.

Assumption2： It is assumed that Olympic-specific data prior to 1988 is invalid historical data and will have no impact on medal predictions, etc.

Assumption3： Assuming the official data provided by COMAP is true and reliable.

3 Notations

The key mathematical notation used in this paper is illustrated in Table 1.

Table 1. Symbol Description Table

Symbol	Description
NOC	IOC country codes
X_1	Number of medals 1988-2020
X_2	Whether or not it has been a host country 1988-2024
X_3	Dominant Programs 2024 Medals as a Percentage of Total Medals
X_4	Number of Olympic participation
X_5	Number of athletes participating in the Olympic Games
X_6	Number of Olympic events
A_1	Total number of medals
A_2	Total number of gold medals
A_3	Total number of silver medals
A_4	Total number of bronze medals

4. Predictive Model Based on the Random Forest Algorithm and the K-means Algorithm

4.1 Data Pre-processing

Firstly, considering that the attachment gives 250,000 data entries containing all participants, their sports, years and results, the amount of data is quite large. Therefore, the data of the athletes is sorted separately and presented in a table form for data cleaning. A total of 1,484 duplicate values are deleted, leaving 250,197 unique values. Secondly, given that in the athlete's dataset, "team" not only includes country or region, but also includes more detailed information, in order to simplify the model calculation, the National Olympic Committee (NOC) recorded by the International Olympic Committee at the time of a specific Olympic Games is selected as the research object to represent the characteristics of a country. Due to various reasons such as politics, the ambiguity of the "team" affiliation has a minimal impact on the model solution and can be ignored.

4.2 Establish a Medal Table Prediction Model

4.2.1 Random Forest Algorithm

Random forests use Bagging-type algorithms to combine multiple weak models trained into a single strong model, effectively improving the accuracy of prediction results. In addition, the random forest regression model can handle multiple feature variables, avoid overfitting, and has strong generalization capabilities. It also takes into account the possible nonlinear relationships between feature variables, fully captures the complex interactions between features, and provides robust prediction results [1]. Given that this question requires a prediction of the 2028 Summer Olympics medal standings in Los

Angeles, USA, based on a large amount of existing data, a random forest regression prediction model was considered for construction, and countries that are expected to improve or perform worse were determined based on the prediction results. The specific modeling process is as follows:

1) Modelling preparations

Model data includes three parts: variables, features, and output values. "Variables" are the abbreviations of the countries registered with the International Olympic Committee (IOC) as the country identifier (NOC). "Features" are the feature values that affect the number of medals won by each country. These include: the cumulative number of medals won by each country over the past nine years, whether the country has hosted the Olympics in the past ten years, the country's performance in dominant events (expressed as the proportion of 2024 medals to total medals), the country's historical participation since participating in the Olympics, the number of athletes the country has sent to the Olympics, and the number of different events the country has participated in at the Olympics. The "output value" is the total number of medals, gold medals, silver medals and bronze medals won by each country in 2024. Based on this benchmark, a random forest regression prediction model was constructed to predict the medal situation in 2028. Due to space constraints, the results of some data collation are shown in Table 2, using "total number of medals" as an example.

Table 2. Summary of Medal Count Characteristics and Output Values for each Country

NOC	X_1	X_2	X_3	X_4	X_5	X_6	A_1	A_2	A_3	A_4
AFG	2	0	0.00%	16	108	43	0	0	0	0
AHO	1	0	0.00%	13	52	39	0	0	0	0
AIN	0	0	66.67%	1	32	30	6	1	4	1
ALB	0	0	100.00%	10	54	51	2	0	0	2
ALG	15	0	33.33%	15	432	198	3	2	0	1
AND	0	0	0.00%	13	36	33	0	0	0	0
ANG	0	0	0.00%	11	188	68	0	0	0	0

2) Characteristic weight calculation

Random forests determine the weight of feature attributes by measuring the importance of eigenvalues. Here, the importance of each eigenvalue is determined based on the degree of Gini impurity in the Gini index. For each eigenvalue, the contribution of the eigenvalue to reducing Gini impurity in all decision trees is calculated to determine the weight of the feature[2]. The specific calculation formula is as follows:

(1) Calculate Gini impurity:

$$Gini(t)=1-\sum_{i=1}^{C}p_{i,t}^{2} \tag{1}$$

Where indicates that there are categories in the decision tree and is the probability that a random sample of belongs to category *i*.

(2) Calculate the reduction in Gini impurity for each eigenvalue for a random sample:

$$\Delta Gini(t,j)=Gini(t)-\frac{|t_1|}{|t|}Gini(t_1)-\frac{|t_2|}{|t|}Gini(t_2) \tag{2}$$

Where represents each eigenvalue and and represent the two sub-samples of the sample, respectively.

(3) Calculate the normalized weights of each eigenvalue:

$$W_j=\frac{\sum_{M=1}^{D}\Delta Gini_m(j)}{\sum_{l=1}^{n}\sum_{M=1}^{D}\Delta Gini_m(K)} \tag{3}$$

Where represents the number of eigenvalues and represents the number of trees.

3) Dataset prediction[3]

$$\hat{y_l}=\sum_{k=1}^{K}D_d\left(x_j\right) \tag{4}$$

Where represents each input feature value and represents the *d*th decision tree.

4) Confidence intervals calculated using the Bootstrap method

$$C_t=(\mu_t\text{-}1.96\sigma_t,\mu_t+1.96\sigma_t) \tag{5}$$

5) Model evaluation

(1) Calculate the mean square error (MSE):

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i\text{-}\hat{y}_i)^2 \tag{6}$$

(2) Calculate the mean absolute error (MAE):

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i\text{-}\hat{y}_i| \tag{7}$$

(3) Calculate the coefficient of determination (R^2):

$$R^2=1\text{-}\frac{\sum_{i=1}^{n}(y_i\text{-}\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i\text{-}\bar{y})^2} \tag{8}$$

4.2.2 K-means Clustering Algorithm

K-means clustering algorithm can not only handle complex data with a large number of data points and categories that are difficult to classify, but also process large amounts of data efficiently. Therefore, it was decided to use the K-means clustering algorithm in this paper. Based on iterative calculations, the data is divided into K clusters. The data from various countries over multiple years is processed to present the characteristics of each country's Olympic performance. At the same time, the relevant indicators are divided and the clustering results are intuitively displayed in the form of clusters to assist in evaluating the potential of countries that have not won a medal to win their first medal at the next Olympic Games. The specific modeling process is as follows:

1) Data Pre-processing

This article here is a prediction of whether unaward-winning countries can win prizes in the next

Olympics. However, the official COMAP dataset contains data on award-winning countries. Therefore, the dataset must first be cleaned to remove a series of data on award-winning countries. Second, for the retained data of countries that have not won awards, the first step is to check for missing values and duplicate values. Specifically, the Python isnull() function is used to check for missing values, and the data is filled in or deleted as appropriate. The duplicated() function is used to check for duplicate data in the cleaned data set and delete it. Second, organize by five key variables. First, the number of times each country has participated in the Olympics, which reflects its historical activity in participating in the Olympics. Second, the total number of participating athletes in each country, which indirectly reflects the scale of the country's participation in the Olympics. Third, the number of individual events participated in the Olympic Games, to understand the degree of participation of athletes from that country in individual competitions. Fourth, the number of team events participated in the Olympic Games, which shows the country's participation in team competitions. Fifth, the proximity of these countries to the host country of the 2028 Los Angeles Olympics. The proximity factor may affect the country's participation strategy, athlete status adjustment, etc. The third step is to convert discrete variables into dummy variables, converting discrete variables in the data, such as the distance of these countries from the host country of the 2028 Los Angeles Olympics, into dummy variables to facilitate the subsequent K-means clustering process. Due to space constraints, some of the results of data pre-processing are shown in Table 3.

Table 3. K-means Clustering Data Preprocessing Results Table

	Number of Olympic Games attended	**Total number of athletes participating**	**Number of individual events at the Olympics**	**Number of Olympic team events**	**Proximity to host country**
Andorra	13	36	32	1	0
Angola	11	188	67	1	0
Antigua and Barbuda	12	77	45	3	0
Aruba	10	38	36	1	0
American Samoa	10	34	29	1	0
Bangladesh	11	47	32	1	0
Benin	13	62	44	0	0
Bhutan	11	25	11	0	0
Bosnia and Herzegovina	9	52	49	0	0
Belize	14	55	31	2	0
Bolivia	16	85	61	6	0

Brunei	7	14	13	0	0
Central African Republic	12	56	36	0	0
GyoshuII	11	51	37	3	0

2) Establish a K-means clustering model

Step 1: Determine the initial cluster centers. From the sorted data of the 77 unmedaled countries, 3 sample points are randomly selected as the 3 initial cluster centers required for initial clustering. The selection of these 3 initial cluster centers will affect the final clustering results. Although they are randomly selected, the stability of the clustering results under different initial values can be found in multiple experiments, so as to evaluate the reliability of the clustering.

Step 2: Assign samples to clusters. Calculate the Euclidean distance between each unmedaled country's data and the three initial cluster centers using the Euclidean distance formula, and assign each country to the cluster with the initial cluster center at the shortest distance to complete the initial data classification. The specific calculation formula is:

$$D_{mn}=\sqrt{\sum_{k=1}^{5}(X_{mk}\text{-}K_{nk})^2} \tag{1}$$

Where (representing 3 cluster centers) and (representing 77 unmedaled country samples). Here represents the kth eigenvalue of the mth country (e.g., 5 indicators such as the number of Olympic Games participated in and the total number of participating athletes), and represents the kth eigenvalue of the nth cluster center.

Step 3: Update the cluster center. According to the clusters divided above, redefine the new cluster center, that is, adjust it to the center position of all sample points in each category in the previous step. The specific calculation formula is:

$$G=\sum_{m=1}^{k}\sum_{x_q\in A_m}\left(x_q\text{-}a_m\right)^2 \tag{2}$$

Here, is the number of clusters (k=3 in this study), represents the mth cluster, is a sample point in cluster A_m, and is the current cluster center of the mth cluster. This formula can be used to obtain a new cluster center that better represents the characteristics of each cluster.

Step 4: Iterative optimization: Steps 2, 3, and 4 are executed in a loop to analyze whether the cluster center has changed. During each iteration, the cluster center is continuously adjusted to make the data in each cluster more similar and the data in different clusters more different. The loop is stopped until the cluster center no longer changes or the preset number of iterations is reached (the running code is the third time), and a stable clustering result is finally obtained[4]. The formula for stopping iteration is as follows:

$$\sum_{j=1}^{k} \left|\mu_j^{'}\text{-}\mu_j^{''}\right| < \epsilon \tag{3}$$

Step 5: Evaluation of the adjustment. The within-cluster sum of squares (SSE) and silhouette coefficient of this K-means model are calculated to estimate the model's degree of fit and the odds of the prediction results.

$$SSE=\sum_{i=1}^{K}\sum_{x_j\in C_i} d\left(x_j,\mu_i\right)^2 \tag{4}$$

$$S=\frac{1}{n}\sum_{j=1}^{n} \mathrm{s}\left(x_j\right) \tag{5}$$

Where is the total number of samples.

4.3 Solve the Medal Table Prediction Model

4.3.1 Medal Table Prediction Results

First, the data in Table 1 is brought into Equations (1)–(3) to calculate the weight of each eigenvalue, and the results are shown in Table 4.

Table 4. Result Table for Eigenvalue Weights

	X_1	X_2	X_3	X_4	X_5	X_6
Weight	0.41	0.06	0.02	0.01	0.23	0.27

Second, the data in Table 1 is brought into Equations (4) and (5) to calculate the predicted value and predicted interval of the medals in 2028. Taking the "total number of medals" as an example, the results are shown in Table 5. The results of the prediction of the remaining gold, silver and bronze medals are shown in Appendices 1-3, and the codes are in Appendix 4.

Table 5. Results of the Prediction of the Total Number of Medals for the 2028 Summer Olympics in Los Angeles, USA

NOC	interval	NOC	interval	NOC	interval	NOC	interval	NOC	interval
AFG	0~0	COL	1~29	INA	3~20	MNE	0~1	SOM	0~0
AHO	0~0	COM	0~0	IND	5~22	MON	0~0	SRB	0~29
AIN	2~19	CPV	0~1	IOA	0~1	MOZ	0~0	SRI	0~1
ALB	1~4	CRC	0~0	IRI	0~29	MRI	0~0	SSD	0~0
ALG	1~7	CRO	0~29	IRL	1~11	MTN	0~1	STP	0~0
AND	0~0	CUB	0~40	IRQ	0~0	MYA	0~1	SUD	0~0
ANG	0~0	CYP	1~5	ISL	0~1	NAM	0~0	SUI	6~40

ANT	0~0	CZE	0~11	ISR	1~10	NCA	0~0	SUR	0~0
ARG	7~51	DEN	7~40	ISV	0~0	NED	16~139	SVK	0~8
ARM	1~7	DJI	0~0	ITA	57~168	NEP	0~0	SWE	16~156
ARU	0~0	DMA	0~0	IVB	0~0	NGR	0~10	SWZ	0~0
ASA	0~0	DOM	0~8	JAM	3~29	NIG	0~0	SYR	0~0
AUS	82~248	ECU	1~6	JOR	1~3	NOR	8~51	TAN	0~0
AUT	2~22	EGY	3~22	JPN	57~156	NRU	0~0	TCH	0~9
AZE	1~29	EOR	0~1	KAZ	5~22	NZL	16~51	TGA	0~0
BAH	0~8	ERI	0~1	KEN	1~29	OMA	0~0	THA	2~11
BAN	0~0	ESA	0~0	KGZ	3~7	PAK	0~8	TJK	2~6
BAR	0~0	ESP	18~83	KIR	0~0	PAN	0~1	TKM	0~1
BDI	0~1	EST	0~3	KOR	4~83	PAR	0~1	TLS	0~0
BEL	7~40	ETH	1~11	KOS	1~19	PER	0~4	TOG	0~0
BEN	0~0	EUN	0~29	KSA	0~0	PHI	2~7	TPE	0~9
BER	0~1	FIJ	1~22	KUW	0~0	PLE	0~1	TTO	0~1
BHU	0~1	FIN	0~23	LAO	0~0	PLW	0~1	TUN	0~4
BIH	0~0	FRA	57~321	LAT	0~4	PNG	0~0	TUR	6~22
BIZ	0~0	FRG	0~29	LBA	0~0	POL	12~83	TUV	0~0
BLR	0~29	FSM	0~0	LBN	0~1	POR	3~22	UAE	0~0
BOL	0~0	GAB	0~0	LBR	0~0	PRK	0~8	UGA	1~7
BOT	1~7	GAM	0~0	LCA	1~7	PUR	0~4	UKR	9~51
BRA	12~83	GBR	57~156	LES	0~0	QAT	1~7	URS	0~35
BRN	1~6	GBS	0~1	LIB	0~0	ROC	0~26	URU	0~4
BRU	0~0	GDR	0~35	LIE	0~0	ROT	0~1	USA	82~321
BUL	6~23	GEO	2~8	LTU	0~8	ROU	6~29	UZB	1~29
BUR	0~0	GEQ	0~0	LUX	0~6	RSA	4~22	VAN	0~0
CAF	0~0	GER	82~168	MAD	0~0	RUS	0~168	VEN	0~8
CAM	0~0	GHA	0~1	MAR	1~11	RWA	0~0	VIE	0~1
CAN	57~168	GRE	6~40	MAS	1~7	SAM	0~0	VIN	0~0
CAY	0~0	GRN	1~7	MAW	0~0	SCG	0~9	YAR	0~1
CGO	0~0	GUA	1~7	MDA	1~7	SEN	0~0	YEM	0~0
CHA	0~0	GUI	0~0	MDV	0~0	SEY	0~0	YMD	0~1
CHI	0~5	GUM	0~0	MEX	7~23	SGP	0~1	YUG	0~22
CHN	16~168	GUY	0~0	MGL	1~7	SKN	0~0	ZAM	1~2
CIV	1~2	HAI	0~0	MHL	0~0	SLE	0~0	ZIM	0~0
CMR	0~0	HKG	0~4	MKD	0~1	SLO	0~8		

COD	0~0	HON	0~0	MLI	0~0	SMR	0~0
COK	0~0	HUN	16~67	MLT	0~0	SOL	0~1

Third, the data in Table 1 are brought into Equations (6)–(8) to calculate the mean square error, mean absolute error and coefficient of determination in turn. The results are shown in Table 6 and the code is provided in Appendix 4.

Table 6. Evaluation Results of the Random Forest Regression Prediction Model

	MSE	MAE	R
Prediction of the number of medals	132.79	4.02	0.87
Gold Medal Prediction	51.82	1.95	0.67
Silver Medal Prediction	74.83	2.38	0.64
Bronze Medal Prediction	15.41	1.33	0.71

Finally, in order to determine which countries are most likely to make progress or perform worse than in 2024, the predicted results based on the random forest regression prediction model are compared with the number of Olympic medals won by each country in 2024 to obtain the medal growth rate data. The results are shown in Figure 2.

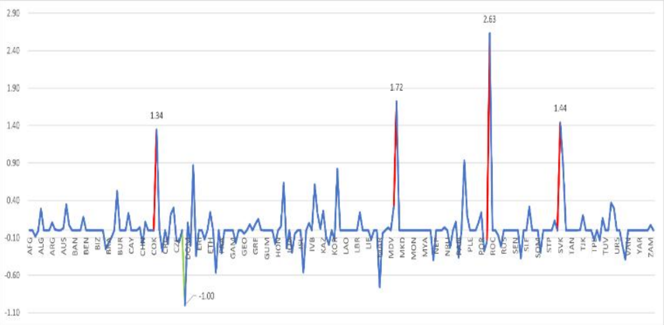

Figure 2. Projected Growth Rates of Olympic Medals by Country in 2028

As can be seen in Figure 2, Colombia, Slovakia, Mongolia and Qatar have the highest growth rates, at 1.34, 1.44, 1.72 and 2.63 respectively. Therefore, Colombia, Slovakia, Mongolia and Qatar are most likely to make progress. Albania, Botswana, Brazil, Bahrain, China, Cape Verde, Croatia, Czech Republic, Denmark, Dominica, Ecuador, Spain, Fiji, France, United Kingdom, Georgia, Hong Kong, India, Iran, Ireland, Israel, Kenya, Kyrgyzstan, South Korea, Lithuania, Morocco, Malaysia, Netherlands, New Zealand, Panama, North Korea, Puerto Rico, Romania, South Africa, Singapore, Serbia, Chinese Taipei, Tunisia, the United States, Uzbekistan and other countries will perform even worse than in 2024.

4.3.2 Results of the Analysis of the Potential of Countries that Have not won a Medal

First, the data in Table 3 are brought into Equations (9)–(11), and the optimal number of aggregations (k value) in the model is determined through iterative optimization, and the index (SSE) that evaluates the clustering effect is obtained as a function of the number of clusters or iterations and the final clustering result. The specific results are shown in Figures 3 and Table 7, and the code is in Appendix 5.

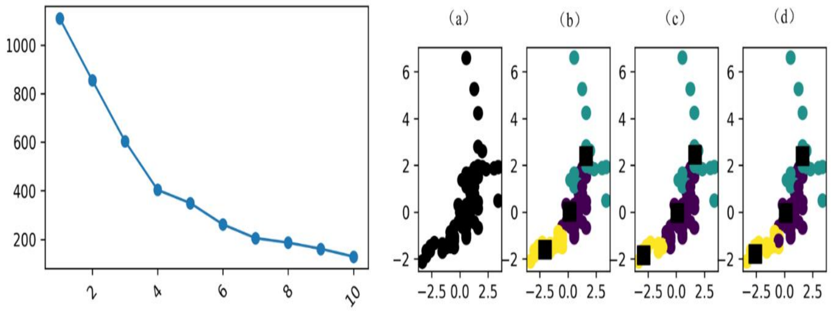

Figure 3. Graph Showing the Change in the Index for Assessing the Effectiveness of Clustering (left)and Visualization of the Results of the Cluster Analysis (right)

Table 7. Clustering Results Table

Country	Category	Country	Category
Andorra	1	El Salvador	2
Aruba	1	Guam	2
American Samoa	1	Honduras	2
Bangladesh	1	Liechtenstein	2
Benin	1	Madagascar	2
Bhutan	1	Mali	2
Bosnia and Herzegovina	1	Malta	2
Belize	1	Myanmar	2
Central African Republic	1	Nicaragua	2
GyoshuII	1	Nepal	2
Cayman Islands	1	Papua New Guinea	2
Congo (Brazzaville)	1	Samoa	2
Chad	1	Seychelles	2
Congo (Kinshasa)	1	Brunei	3
Cook Islands	1	Comoros	3
Gambia	1	Crete	3
Equatorial Guinea	1	Federated States of Micronesia	3
Guinea	1	Guinea Bissau	3
British Virgin Islands	1	Kiribati	3
Laos	1	Lebanon	3
Libya	1	Malaya	3

Liberia	1	Marshall Islands	3
Lesotho	1	North Borneo	3
Malawi	1	Newfoundland	3
Maldives	1	Nauru	3
Mauritania	1	Palau	3
Oman	1	Rhodesia	3
Palestine	1	Refugee Olympic Athletes	3
Rwanda	1	Saar	3
Sierra Leone	1	Saint Kitts and Nevis	3
Solomon Islands	1	South Sudan	3
Somalia	1	Sao Tome and Principe	3
Swaziland	1	Timor Leste	3
Vanuatu	1	Tuvalu	3
Saint Vincent and the Grenadines	1	Unknown	3
Yemen	1	South Vietnam	3
Angola	2	North Yemen	3
Antigua and Barbuda	2	South Yemen	3
Bolivia	2		

Table 7 shows that countries that did not win a medal were classified according to the optimal number of clusters. There were 36 countries in cluster 1, the underachievers, 16 in cluster 2, the high potentials, and 25 in cluster 3, the small countries. Among them, a total of 16 countries in category 2 are most likely to win prizes in the next Olympic Games, namely Angola, Antigua and Barbuda, Bolivia, El Salvador, Guam, Honduras, Liechtenstein, Madagascar, Mali, Malta, Myanmar, Nicaragua, Nepal, Papua New Guinea, Samoa and Seychelles.

Next, the data in Table 3 is brought into Equations (12) and (13), and the within-cluster sum of squares (SSE) and silhouette coefficient of the model are calculated based on the clustering results of the K-means model and the optimal number of clusters to evaluate the clustering effect of the model. The SSE value is approximately 604.7040, and the silhouette coefficient is approximately 0.6861. This shows that the clustering effect of the model is good and the stability is strong, indicating that there is a 68.61% probability that the prediction result is accurate, which is approximately a 31.39% probability.

4.3.3 Results of the investigation into the relationship between the number of events and the number of medals

First, as can be seen from Table 4, the random forest regression prediction model shows that the host country characteristic value is 0.06, which shows that the host country tends to add the number of Olympic events with its own advantages, etc. Therefore, the number of Olympic medals in the host country in that year is higher than in other Olympic Games. Therefore, the host country effect has a positive impact on the number of medals.

Next, the number of medals won by each country in each year was summed up according to different project categories. At the same time, in order to simplify the data, a total of 2,780 project data that had never won a medal in history and 292 project data that had only won one medal were eliminated, and they were not considered as corresponding to the dominant projects of the country. The sports with the most medals and significantly more medals than other sports in each country were selected as the country's dominant sports, that is, the most important sports for them. The specific selection results are shown in Table 8.

Table 8. The Most Important Sports in the Corresponding Countries

NOC	Most important item	NOC	Most important item	NOC	Most important item
USA	Swimming	AUS	Rowing	PAK	Hockey
GBR	Athletics	FRA	Cycling	ARG	Hockey
ITA	Fencing	BRA	Volleyball	CUB	Baseball
NED	Hockey	FIN	Athletics	NZL	Rowing
URS	Gymnastics	SWE	Athletics	ROU	Rowing
GER	Rowing	NOR	Sailing	GDR	Rowing
HUN	Fencing	KEN	Athletics	JAM	Athletics
CAN	Rowing	CHN	Diving	JPN	Gymnastics

5. Causal Inference Model Based on the Multistage Double Difference Method

5.1 Establish a Contribution Value Model for the "great coach" Effect

The multi-period double difference method can deal with the implementation of policies at multiple time points, comprehensively consider differences in temporal trends, make full use of multi-period data information, increase the amount and information content of data, and effectively alleviate the problem of sample loss. In addition, the multi-period double difference method can accurately predict dynamic effects, observe the change of treatment effects over time, and the robustness of the estimation results is more significant than that of the traditional double difference method. Therefore, this paper decides to consider using the multi-period double difference method to estimate the contribution of the "great coach" effect to the number of medals won by each country.

5.1.1 Data Pre-processing

First, it is known that there is a "great coach" effect in volleyball and gymnastics. Therefore, the attached data is screened to eliminate invalid data other than volleyball and gymnastics, and all the data of each specific competition event in each year of volleyball and gymnastics in all countries is retained.

Second, the data required for constructing the multiple-period double difference method for volleyball and gymnastics were sorted and selected separately. Taking the volleyball project as an example, first, the data for each year from 1984 to 2024 on the combined number of medals won by each country in all volleyball events were retained. Second, construct the "Treatment Group" data column, introduce a 0-1 variable to define the experimental group and the control group, i.e., countries that have received volleyball coaching are recorded as 1, and countries that have not received volleyball coaching are recorded as 0. Third, construct the "Treatment Time" data column to represent the relevant data of each country's coaching year and beyond, in order to measure the effect of coaching behavior on the number of medals. The same applies to the gymnastics project. Taking the volleyball project as an example, due to limited space, some of the specific pre-processing results are shown in Table 9.

Table 9. Results of Some Data Preprocessing Using the Difference-in-Differences with Multiple Time Periods Method

NOC	Treatment Group	Year	Medal	Treatment Time	NOC	Treatment Group	Year	Medal	Treatment Time
BRA	0	1984	12	0	CHN	0	1984	10	1996
BRA	0	1988	0	0	CHN	0	1988	11	1996
BRA	0	1992	12	0	CHN	0	1992	0	1996
BRA	0	1996	12	0	CHN	1	1996	9	1996
BRA	0	2000	12	0	CHN	1	2000	0	1996
BRA	0	2004	12	0	CHN	1	2004	12	1996
BRA	0	2008	24	0	CHN	1	2008	12	1996
BRA	0	2012	24	0	CHN	1	2012	0	1996
BRA	0	2016	12	0	CHN	1	2016	12	1996
BRA	0	2020	12	0	CHN	1	2020	0	1996
BRA	0	2024	13	0	CHN	1	2024	0	1996

5.1.2 Model Establishment

In order to eliminate the possible impact of unobservable factors and time effectson the accuracy of the estimated results, this paper simultaneously controls the individual effects of changes in the number of Olympic medals in each country and the time effects to conduct a double difference estimation[5]. At the same time, in order to compensate for the limitations of the multi-period double difference method, the overalldata is split into small sub-modules, and the traditional double difference method is performed within each module, and the effects of the treatment behaviors obtained within each module are reasonably weighted and averaged. The benchmark model is set as follows:

$$Y_{it}=\beta_0+\beta_1 D_i T_t+\beta_2 D_i+\beta_3 T_t+\beta_4 X_{it}+\epsilon_{it} \tag{1}$$

Among them, is the dependent variable, the number of medals, which indicates the total number of medals won by country in year in this event. is the treatment variable, whether or not the country has received coaching. is the time variable, where the year before coaching is recorded as 0 and the year after coaching is recorded as 1.

5.2 Resolution of the Contribution-value Model for the "great coach" Effect

5.2.1 Determination of Contribution Value

Substitute the data in Table 9 into Equation (14) to calculate the contribution of the “great coach” effect to a country’s Olympic medal count in volleyball and gymnastics. See Tables 10-11 for the specific solution results and Appendix 5-6 for the code.

Table 10. Results of the Volleyball “great coach” Effect Contribution Value Table

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	43.6467	78.605	0.555	0.579	-111.436	198.729
treated[T.True]	3.7442	0.976	3.838	0	1.819	5.669
Treatment_Group	3.7442	0.976	3.838	0	1.819	5.669
Year	-0.0199	0.039	-0.508	0.612	-0.097	0.057

Table 11. Results of the Contribution Values for the “Great Coach” Effect in Gymnastics

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	35.0645	40.106	0.874	0.383	-43.809	113.938
treated[T.True]	5.1511	0.606	8.496	0	3.959	6.343
Treatment_Group	5.1511	0.606	8.496	0	3.959	6.343
Year	-0.0167	0.02	-0.828	0.408	-0.056	0.023

As can be seen from Tables 10-11, first, in terms of volleyball, the coefficient of treated[T.True] is 3.7442, indicating that the medal count of a volleyball team that has accepted a great coach and started coaching increases by an average of 3.74 medals. In other words, in the sport of volleyball, the “great coach effect” has a positive contribution to the medal count, with an average of about 3.74 medals per competition. Second, the p-value of treated[T.True] is close to 0, and a p-value less than 0.05 means that the variable’s impact on the number of medals is statistically significant, i.e., the “great coach effect” significantly affects the number of medals. Third, the F value is 7.380 and the p value is 0.000826, which also indicates that the model as a whole is statistically significant. Secondly, similarly, in gymnastics, the treated[T.True] coefficient is 5.1511 and the p value is 0, which indicates that after accepting the coaching of a great coach, the number of medals increases by an average of 5.15. It can be seen that the coaching of a great coach has a significant positive contribution to the number of medals.

5.2.2 “Great Coach” Investment Projection

Based on the conclusions in Tables 10-11, it was decided to select the country with the most Olympic appearances in gymnastics or volleyball, the largest number of athletes, and the largest total number of events, but the fewest medals won in that event, to consider investing in the “great coach” effect of the corresponding sport.

First, for the volleyball project, invalid data other than volleyball data was eliminated, and then the data was cleaned to remove duplicate values. The number of Olympic Games each country has participated in, the total number of athletes participating in volleyball, the total number of volleyball events, and the total number of volleyball medals were calculated, and countries with a zero medal count were retained. Second, for the gymnastics project, invalid data other than gymnastics data was eliminated, and the number of Olympic Games participated in, the number of athletes participating, the number of events participated in, and the total number of medals won in this project were also calculated. The specific data calculation results are shown in Tables 12-13 and Figure 5.

Table 12. Results Table for Volleyball Project “Great Coach” Investment Priority

NOC	Olympic Games	Number of athletes	Number of events	NOC	Olympic Games	Number of athletes	Number of events	NOC	Olympic Games	Number of athletes	Number of events
CAN	7	96	4	HUN	4	34	2	SWE	1	12	1
TUN	7	71	2	FRG	2	34	2	YUG	1	12	1
EGY	6	65	2	ESP	2	33	2	BEL	1	12	1
GER	6	57	3	TUR	3	32	3	UKR	1	12	1
AUS	3	44	2	ALG	3	26	2	PUR	1	12	1
KEN	4	39	3	GRE	1	24	2	KAZ	1	12	1
DOM	4	37	3	GBR	1	24	2	CRO	1	11	1
MEX	2	36	2	IRI	2	22	2	CMR	1	10	1
VEN	2	34	3	SLO	1	13	1	LBA	1	9	1

Table 13. Results Table for Investment Priority of Gymnastics Event “Great Coach”

NOC	Olympic Games	Number of athletes	Number of events	NOC	Olympic Games	Number of athletes	Number of events	NOC	Olympic Games	Number of athletes	Number of events
AUS	16	71	14	SAA	1	6	8	HKG	1	2	9
LUX	10	56	16	SVK	6	6	7	CHI	2	2	6
CUB	12	43	14	LTU	4	5	12	GEO	4	1	7
MEX	15	36	15	COL	4	5	11	TUN	2	1	7

EGY	8	22	12	PHI	2	5	10	LIE	1	1	7
POR	10	21	13	ALG	3	4	11	BAR	1	1	7
ARG	10	20	13	ARM	3	4	11	SMR	1	1	7
ISR	9	14	12	VEN	4	3	11	CYP	1	1	6
TPE	5	12	13	IRL	3	3	11	BOL	1	1	5
IND	4	11	13	TUR	2	3	10	NAM	1	1	5
NZL	4	9	12	VIE	3	3	7	YEM	1	1	4
MAR	2	7	13	ECU	1	3	7	PAN	1	1	4
PUR	6	7	12	MON	2	3	6	TTO	1	1	4
CZE	6	7	12	AZE	2	3	6	SGP	1	1	4
KAZ	4	7	12	GUA	4	3	5	PER	1	1	4
SLO	5	6	12	BOH	4	3	3	JAM	1	1	4
RSA	5	6	12	IRI	1	2	12	BAN	1	1	3
MGL	3	6	12	ISL	4	2	11	DOM	1	1	2
UAR	1	6	8	MAS	2	2	9				

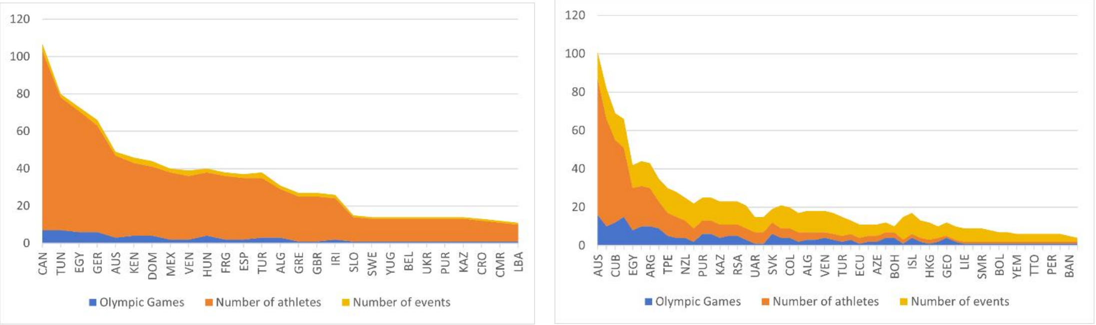

Figure 4. Results of the Investment Priority Ranking for the “Great Coach” Volleyball (left) and Gymnastics (right) Projects

As can be seen in Table 12-13 and Figure 4, the countries with the largest number of Olympic volleyball sessions, the largest number of participating athletes, and the largest total number of participating events, but have not yet won a volleyball medal, are Canada and Tunisia, while Australia is in the gymnastics category. Therefore, Canada and Tunisia should consider investing in the “great coach” effect in the volleyball project, so that their volleyball project medals will increase by 3.74 medals each year. Australia, on the other hand, should consider investing in the “great coach” effect in the gymnastics project, so that its gymnastics project medals will increase by 5.15 medals each year.

6. Other Original Insights into Olympic Medal Counting

6.1 The Effect of the Number of Athletes Competing on Medal Counts

Based on the data from the random forest model as well as the data used in the K-means model, it was found that there is a strong link between the number of participating athletes and the medal count.For example, the total number of medals won by the United States grew from 121 to 137 when its number of participating athletes increased from 550 in 2016 to 620 in 2020, while the number of participating athletes in Greece decreased from 110 in 2016 to 85 in 2020, with the number of medals dropping from 14 to 9.

6.1.1 Original Insights

1. Project Coverage Expansion： The increase in the number of participating athletes will help the country achieve broader program coverage at the Olympics.This trend is exemplified by Russia’s performance at the Olympics, as shown in Figure 6.At the 2012 Olympics, Russia sent 436 athletes to participate in 19 of the 22 major sports and collected 82 medals. By the 2016 Olympics, the number of athletes increased to 478, participation expanded to 21 of the 22 major sports, and the total number of medals jumped to 161 (56 gold, 53 silver, 52 bronze).When there are a large number of Olympic sports and a limited number of athletes, countries tend to be able to focus on only a few dominant sports. With an increase in the number of athletes, however, it is able to lay claim to a wider range of smaller events. In track and field, for example, it is possible to expand from participation in sprints only to multiple smaller events such as long-distance running and long jump, creating more opportunities to win medals.

2. Team Battle Enhancement： In team sports, an increase in the number of athletes provides coaches with a richer selection of personnel, which in turn helps to form stronger teams.This is exemplified by the increasing team play of the U.S. basketball team at the Olympics.As shown in Figure 5, at the 1936 Berlin Olympics, the U.S. basketball team competed with a limited number of athletes and a relatively small pool of talent, ultimately winning 14 medals. By the 1976 Montreal Olympics, the U.S. basketball talent pool had increased dramatically, with the number of medals rising to 24 from the previous 12-14. This was due to a significant increase in the number of athletes involved in basketball training, allowing coaches to have more high-level players to choose from when assembling their Olympic teams. For example, when selecting the starting lineup, they are able to choose a combination of more skillful players who work better together, and when preparing substitutes, there are players of comparable strength ready to step in.

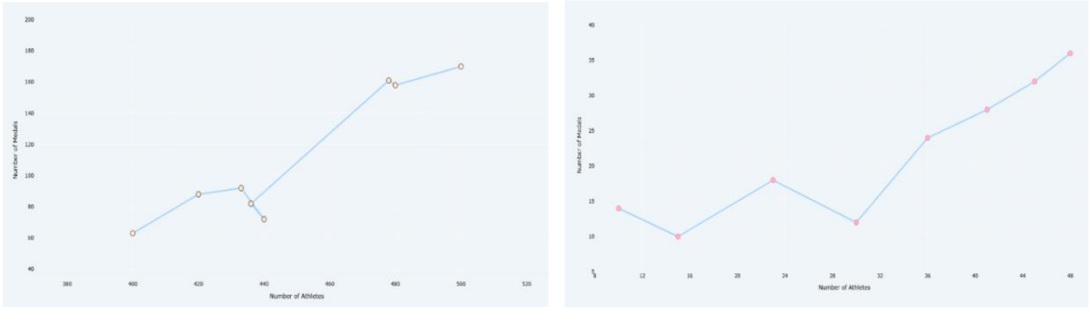

Figure 5. Number of Russian athletes (left) vs. U.S. basketball players (right) vs. medal counts

6.1.2 References to the Olympic Committee

1. Increase talent selection：The Olympic Committee should broaden the channels for athlete selection, tap potential athletes from more regions and age groups, and expand the talent pool.

2. Rational allocation of resources： While increasing the number of athletes, it is important to rationally allocate training resources to ensure that each athlete receives adequate training and enhancement.

6.2 Effect of the Number of Entries on Medal Counts

According to the analysis of related data by the random forest model, the number of participating events is closely related to the medal count. Taking the performance of China and Japan in the 2016 Olympic Games as an example, China participated in 24 of the 26 major events and finally won 70 medals; Japan participated in 22 events and won 41 medals. China dominated the medal count with more entries.

6.2.1 Original Insights

1.Risk-sharing stabilization： Increased participation in projects spreads the risk of losing medals due to poor performance in individual projects.Taking the UK as an example, Figure 7 presents the distribution of medals for each of the UK's events at the London 2012 Olympic Games.Britain focused on cycling and rowing, with 10 medals in cycling and 8 in rowing, while archery only managed 2 medals, a poor performance. However, Britain's extensive participation in 23 of the 26 major sports, such as track and field and swimming, which contributed 6 and 7 medals respectively, made up for archery's loss.In the end, Great Britain's medal tally settled at 29 gold, 17 silver and 19 bronze, for a total of 65.In the Olympic Games, the results of the events are uncertain, and even the dominant events may not perform well due to various factors. However, when the participating events are rich and diversified, the medals of other events can make up for the losses caused by the lost events, thus ensuring that the total number of medals remains relatively stable.

2. Uncovering Potential Strengths： More participation in sports provides an opportunity to explore potential strengths in the country. There are sports that may not have been emphasized or participated in before, but through active experimentation, athletes' talents in these sports can be discovered. Korea is a prime example.In the early Olympic Games, Korea's dominant programs were mainly focused on archery and short track speed skating, which won 8 and 9 medals respectively. With the continuous expansion of the participating events, in the 2016 Olympic Games, Korea made a breakthrough in the golf event, with a Korean player winning the gold medal in the women's golf individual event, and the results of this breakthrough can also be seen from the number of 3 medals in the golf event in the chart. This is undoubtedly the culmination of Korea's continued experimentation with new events and the deep exploration of its athletes' potential, adding a whole new dimension to Korea's medal tally.

6.2.2 References to the Olympic Committee

1. Project Expansion Planning： The NOC should develop a comprehensive program expansion plan to encourage athletes to participate in more different types of sports, not just limited to traditional strengths.In-depth study and evaluation of the Olympic program can be carried out to understand the development trend, competition pattern and the potential of national athletes, so as to select the program with potential and focus on cultivation and investment.

2. Rationalization of resources： It is important to deploy resources wisely when expanding your entries.Different proportions of training and coaching resources and financial support will be allocated according to the potential of the project and the possibility of winning awards. Newly participating programs may first invest certain resources for basic training and talent development.

6.3 The Effect of the Number of Olympic Games on Medal Counts

By analyzing the data from the K-means model, it is found that there is a certain cumulative effect between the number of Olympic sessions attended and the medal count. In the case of Australia, for example, which has participated in a number of consecutive Olympic Games since 1984, its total medal count has grown from an average of about 20 medals in the first few sessions to an average of about 40 medals in recent sessions. On the other hand, some newly participating countries, such as Timor-Leste, had relatively small and unstable medal counts in the previous Olympic Games in which they participated.

6.3.1 Original Insights

1. Tournaments familiarize with efficiency gains: The trend of Germany's cumulative medals across the Olympic Games presented in Figure 6 shows that as the number of Olympic Games attended increases, the familiarity of athletes and related staff with the event does increase, which in turn has a positive impact.As can be seen in Figure 8, Germany's cumulative medal count at the Olympics continued to grow as the number of participation sessions continued to accumulate, and by the time of the 2016 Rio Olympics, this was backed by the fact that the German athletes and staff were already well versed in the Olympics process, judging scales, and so on.German athletes were able to better adjust their status and play level on the field, and finally won 17 gold, 10 silver, 15 bronze, a total of 42 medals; the staff also became more competent in organizing the participation and logistic support, which created good conditions for the athletes and effectively improved the efficiency of the awards.

2. Training refinement and quality improvement: The change in China's cumulative medal count at the Olympics, shown in Figure 6, is a strong testament to the remarkable results brought about by the continuous refinement of the country's sports training system over the course of many Olympic Games.Since China returned to the Olympic Games in 1984, it has continued to optimize its training system by summarizing and analyzing the results of each competition. In the early years of the Olympic Games, China's performance in swimming and some other sports was average. However, with the increase in the number of competitions, the Chinese swimming team has continuously adjusted its training methods and programs and introduced advanced training concepts and techniques based on the problems exposed during the competitions. By the 2021 Olympics (2020 in the figure), China's swimming team won 3 gold, 5 silver, 8 bronze, a total of 16 medals, a qualitative leap compared with the early days, which makes China's cumulative medal count in the Olympic Games show a more obvious upward trend in Figure 6.This continuous improvement of training based on feedback from competitions has actually improved the quality of training and significantly strengthened the athletes' award-winning performance.

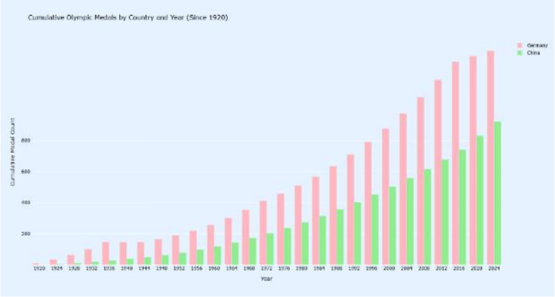

Figure 6. Cumulative Olympic Medals by Country and Year (since 1920)

6.3.2 References to the Olympic Committee

1. Long-term participation planning： NOCs should formulate long-term and stable participation plans to ensure continuous participation in the Olympic Games in order to fully utilize the advantages brought about by the accumulation of experience and the improvement of the system. Multi-year participation plans can be formulated on the basis of national sports development strategies and goals, with clear objectives for key programs and talent development at different stages.

2. Passing on experience and innovation： Focus on passing on the experience of previous Olympic Games to the new generation of athletes and staff, and encourage innovation in training methods, event organization and other aspects to adapt to the development of the Olympic competitive environment. Experience-sharing sessions and seminars can be organized to allow veteran athletes and senior staff to pass on their experience, while focusing on international sports dynamics and introducing advanced concepts and technological innovations.

7. Model Evaluation and Further Discussion

7.1 Strengths

1. In this paper, the random forest regression model is chosen to predict the total number of medals and the number of gold, silver and bronze medals of each country in the future Olympic Games, which is able to deal with highly nonlinear data and complex feature interactions, avoid the problem of overfitting, deal with the missing values and noisy data well, and improve the stability and accuracy of the prediction, as compared with the ordinary regression model.

2. The K-means algorithm is simple, efficient, capable of handling large-scale datasets, and scalable.

3. A multi-period double-difference approach is used to assess the long-term impact of the "Great Coach Effect" on medal counts, making full use of data from multiple points in time and avoiding errors due to short-term fluctuations or single-time idiosyncrasies.The dynamic analysis over multiple periods enables a more accurate estimation of the cumulative contribution of the great coaching effect

over multiple Olympic cycles, providing a more reliable basis for the development of long-term coach introduction and development strategies.

7.2 Weakness and Further Discussion

1. Since Random Forest is an integrated model consisting of multiple decision trees, it is difficult to intuitively understand the specific impact of each feature on the prediction results, and the interpretability is poor.

2. Random forests may over-rely on certain features when dealing with highly correlated features, leading to model preference for specific variables and the efficiency of the model.

2. K-means clustering results are highly influenced by the initial clustering centers, which may affect the analytical judgment of country characteristics.

3. Although the multi-period double-difference approach is able to capture long-run policy effects or intervention impacts, it has its own limitations.First, the double-difference method assumes parallel trends between different countries, i.e., the trend in the number of medals over time would be the same for all countries if there were no coaching changes, but this assumption is difficult to fully substantiate in practice.In addition, multi-period double differencing may not be able to effectively handle temporal heterogeneity or unobserved variable disturbances, which can affect the accuracy of the results. To address these issues, we used **CSDID (Double Difference with Common Trend)** to test and correct the model.The CSDID allows for adjustments to be made for time trends in different countries, better addressing heterogeneity and time-related issues and ensuring that assessment results are more robust and credible.

References

Lu, J. Y., & Guo, J. H. (2025). Impacts of Three-Industry Integration on Farmers' Income-Empirical Evidence from National Rural Industrial Integration Pilot Policies. *East China Economic Management*, 1-11[2025-01-27].

SONG, Y. p., ZHU, J. M., YANG, Q. et al. (2021). Analysis of first week box office prediction of domestic movies based on random forest regression. *Journal of Higher Education Science*, *41*(01), 21-26.

YANG, J. B., & ZHAO, C. (2019). A review of research on K-Means clustering algorithm. *Computer Engineering and Applications*, *55*(23), 7-14+63.

ZHANG, M. X., ZHAO, T. M., & WANG, S. Y. (2023). Research on the application of random forest method in the selection of mother ship. *Applied Science and Technology*, *50*(05), 126-132+174.

ZUO, Y. (2025). Decision Tree Based Flood Prediction Model. *Water Resources and Hydropower Technology (Chinese and English)*, 1-15.