ISSN:2790-1688 Volume-14-(2025)

Uncovering the Secret of Olympics Medals

Yihan Chai[a], Ruifeng Du[b]

School of Marxism, Xi'an Jiaotong University, Xian, China.

[a]cyh2225120091@163.com, [b]18568032589@163.com

Abstract. This study aims to predict Olympic medal counts and analyze factors influencing national competitiveness in the Olympics. Firstly, a Two-stage Random Forest Model was developed using multiple random forest regressions to forecast the number of medals and gold medals for each country in the 2028 Olympics. The model achieved an average R^2 of 0.985 on the test set, demonstrating high prediction accuracy. Additionally, an XG-BOOST Classifier was employed to identify countries most likely to achieve a medal breakthrough from zero to one. Secondly, a Cluster-OLS Prediction Model was established to explore the relationship between changes in events and medal counts. The concept of "market share" was applied to represent a country's project advantages, and clustering algorithms were used to group countries. The regression coefficients revealed that events with greater suspense had higher coefficients, indicating their importance for national competitiveness. The Host Effect was also included as a positive indicator in the model. Thirdly, a PSM-DID Model was used to quantify the impact of "great" coaches. The DID value was calculated as 7.9619, confirming the significant effect of coaching. Based on this, China, France, and the United States were recommended to invest in ice hockey, basketball, and volleyball, respectively, to maximize the benefits of effective coaching. Finally, the study suggests that countries should balance training resources for athletes of different genders based on regional culture and project advantages to enhance their Olympic preparations.

Keywords: medals; Random Forest; XG-BOOST; PSM-DID; Cluster-OLS; Olympic.

1. Introduction

The modern Olympic Games began in 1896 and are held every four years. Due to its consistent inheritance of the ancient Greek Olympic spirit of "peace, unity, and excellence" and the wide participation of countries around the world, it has become the world's most influential comprehensive sports event. Olympic medals represent the highest honor in sports competitions. They are not only a high recognition of an individual's outstanding sports skills, but also an important indicator for demonstrating a country's sports strength and even measuring a country's soft power and status. In each Olympic Games, the changes in the Olympic medal table affect the hearts of the country as a whole and the audience as individuals - all countries strive to move up a level in the medal table; while watching the wonderful performances of athletes in the competition, people also pay close attention to the overall medal table rankings of various countries.

Therefore, the demand for the development of models that can accurately predict the number of national medals is growing. This will help countries to recognize their own advantages and disadvantages in sports development in advance based on the forecast situation, reasonably allocate funds, manpower, technology and other resources, formulate more targeted development strategies, and enhance the country's overall sports competitiveness and international influence. To this end, we plan to answer the following questions and we have drawn a flow chart (see Fig.1).

Q1: Develop a model to predict and analyze the number of medals. Based on the model, predict the medal standings for the 2028 Summer Olympics in Los Angeles, CA and include prediction intervals showing all results. Determine which countries are most likely to improve and which countries are most likely to decrease their results compared to the 2024 Olympics.

Q2: First medal predictions. For countries that have not yet won a medal, estimate the number of people who will win their first medal at the next Olympic Games.

Q3: Analyzing the relationship between Olympic events and the number of medals. Analyze the relationship between Olympic sports (both number and type) and the number of medals won by a

country, exploring the most important sports for each country and the reasons for this, and explaining how the choice of sports made by the host country affects the final result.

Q4: Explore the impact of the 'great coach' effect on the number of medals won and estimate the contribution of this effect to the number of medals won.

Q5: Explain the model's unique insights into Olympic medal counts and explain how it can inform National Olympic Committees.

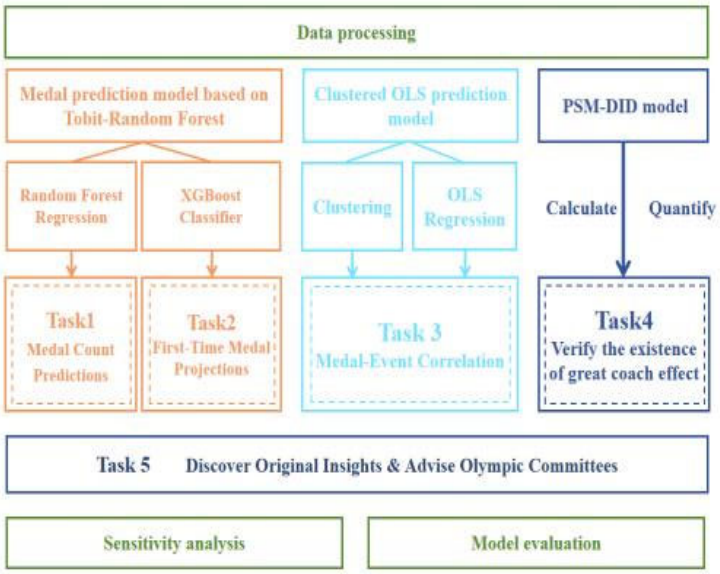

Fig.1 Workflow of this study.

In order to simplify the problem, we make the following basic assumptions, each of which is accompanied by a reasonable explanation:

Assumption 1: Gold, silver and bronze medals have different impacts on the number of medals a country can win in the future and the time of impact. The three medals represent different levels and have different degrees of motivation on the mentality of athletes and the country. Although this is not a simple linear relationship, we can still quantify it through certain models.

Assumption 2: The main basis for predicting medals is the current state of the country's overall athletes, not historical data. As long as high-level athletes are obtained, the country is likely to win medals regardless of past historical results, and vice versa. The state of athletes includes their level and the number of events per capita (reflecting their fatigue level)

Assumption 3: The composition and characteristics of athletes sent by various countries to the 2028 US Olympics will remain basically unchanged. We have not yet learned the composition and characteristics of athletes sent by various countries in 2028, but considering that there is only one time difference between this and 2024, there is a high probability that there will be no large-scale changes. Even if there are small-scale changes, the impact on the results is within an acceptable range.

Assumption 4: We assume that different types of sports can reflect some of the same sports indicators. Although each project has its own differences, its research objectives can be abstracted into several fixed categories, such as static, dynamic, and mixed. Countries in different regions may have the same project advantages in similar sports due to their ethnic composition and cultural factors.

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Definition	Symbol	Definition
nB	The number of bronze medals attained	ty	Year when athletes participated in Olympics Games
nmale	The number of male athletes	nG	The number of gold medals attained

nfemale The number of female athletes ns The number of silver medals attained

2. Q1: Medal prediction model based on Tobit-Random Forest

2.1 Introduction to the Tobit-Random Forest Model

The random forest regression algorithm randomly extracts samples and features to build multiple unrelated decision trees. Each decision tree can obtain prediction results in parallel. Finally, the results of all trees are combined and averaged to obtain the regression prediction result of the entire forest. XGBoost classifier is an ensemble learning algorithm based on gradient boosting decision trees. It achieves the classification task by combining multiple weak classifiers (decision trees) into a strong classifier. Each decision tree is trained based on the residual of the previous tree, and it gradually optimizes the loss function by iteratively reduce the residuals and introduce regularization terms to reduce the risk of overfitting.

Noticing that there are a large number of countries have not yet earn medals in Olympics, we established by combining the random forest regression algorithm with the XGBoost classifier not only has the advantages of both, but also has a significant advantage: after classifying the data, it can simplify the complexity of the data during regression degree, reducing the impact of noise and greatly improving the accuracy of our predictions.

2.2 Predict the number of medals

Based on the topic, we know that the athlete's competition plan and personal characteristics are important references for predicting the number of medals. Therefore, we find the following features from the original dataset and put them into the XGBoost and random forest:

The year is selected to reflect the continuity of the athlete's competitive status.

The number of medals of different types is selected to reflect the encouraging impact of last year's awards on this year's athletes.

The average number of events each athlete participates in is calculated to measure the athlete's fatigue level.

The variance of number of entries for each player is calculated to measure the change in the player's competitive status.

Next, we constructed an indicator based on the data that can reflect the project's strengths.

(1) Constructing events advantage indicators

First, to describe a country's competitiveness in a certain event, we consider the weighted sum of the number of three types of medals to obtain the corresponding competitiveness index, constructing a variable *score* to estimate countries' competitiveness. In order to finally obtain an index that describes a country's event advantages in all events, we should also weigh the advantages of a country in different events and the weighted sum is *sscore* .

$$score = \sum_{i=1}^{3} w_i M_i$$

$$sscore = \sum_{i=1}^{events} W_i score_i$$

(2) Determining Medal Weights

The Kaplan-Meier estimator of the survival function was used to estimate how long each type of medal could last, thereby measuring the value of each type of medal. Where Si represents the probability of survival at a point in time ti , here means the probability of athletes keep his medal, di represents the number of athletes who lost their medals at that point in time, and ni represents the number of athletes who still had medals before that point.

$$S_i = S_{i-1}(1 - \frac{d_i}{n_i})$$

The final medal duration T is:

$$T = \lim_{n\to\infty} \sum_{i=1}^{n} S_i(t_i - t_0)$$

We use the proportion of T to determine the weight of each medal.

Finally, we figured out that the weight of the gold medal is about 0.37, the silver medal is 0.28, and the bronze medal is 0.35.

Table 2: the proportion of medals to the entire Olympic Games, used as weight.

Sport	2008	2012	2016	2020	2024
Aquatics	0.152318	0.152318	0.150327	0.144543	0.148936
Archery	0.013245	0.013245	0.013072	0.014749	0.015198
Athletics	0.155629	0.155629	0.153595	0.141593	0.145897
Badminton	0.016556	0.016556	0.01634	0.014749	0.015198

After constructing the above variables, we calculated the correlation coefficients between them (see Fig.2). It can be seen that the correlation between these indicators is relatively high. In order to reduce multi-collinearity, we consider using factor analysis to reduce the data dimension.

Fig.2

(3) Determining Number of Factors

Suppose we have factors $f_1, f_2, \ldots, f_n$; and several indicators $x_1, x_2, \ldots, x_p$

$$\begin{cases} x_1 = u_1 + \sum_{i=1}^{n} \alpha_{1i} f_i + \epsilon_1 \\ x_2 = u_2 + \sum_{i=1}^{n} \alpha_{2i} f_i + \epsilon_2 \\ \vdots \\ x_p = u_p + \sum_{i=1}^{n} \alpha_{pi} f_i + \epsilon_p \end{cases} \qquad \begin{cases} E(f) = 0 \\ E(\epsilon) = 0 \\ Var(f) = I \\ Var(\epsilon) = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2) \\ \mathrm{cov}(f, \epsilon) = 0 \end{cases}$$

Among them, $f_i (i = 1,2,3...n)$ is a common factor, $\epsilon_i (i = 1,2,3...p)$ is a special factor, α_{ij} is called the loading matrix, and p is the number of original indicators .

Before conducting factor analysis, we will standardize the data to reduce the impact of different dimensions. Then we will conduct KMO test and Bartley sphericity test to determine the rationality of our factor analysis. The KMO value is 0.835, the significance level here is 0.01, much less than 0.05 Based on KMO and Bartlett's Test, it is believed that the original data is suitable for factor analysis.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy:

$$\mathrm{KMO} = \frac{\sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} r_{ij,\mathrm{partial}}^2}$$

where r_{ij} represents the sample correlation coefficient between the original samples x_i, x_j, *and* $r_{ij,partial}$ represents the partial sample correlation coefficient between the original samples x_i, x_j.

To determine the number of factors to select, we drew a scree plot (see Fig.3), based on which the appropriate number of factors was selected. It can be seen from the figure that when the factor number is 4, the image begins to become flat, so we set the factor number to 4 for subsequent discussion. we can see that the first factor mainly extracts information related to the total number of medals and athletes, reflecting the winning rate information; the second factor mainly extracts abstract statistics such as mean and variance, focusing on the fatigue status of athletes; the third factor mainly extracts time information; the fourth factor mainly extracts factors with larger weights such as the number of gold medals, reflecting the number of elite athletes in the country. Therefore, we named these four factors: *medals*, *fatigue*, *years*, and *elite*.

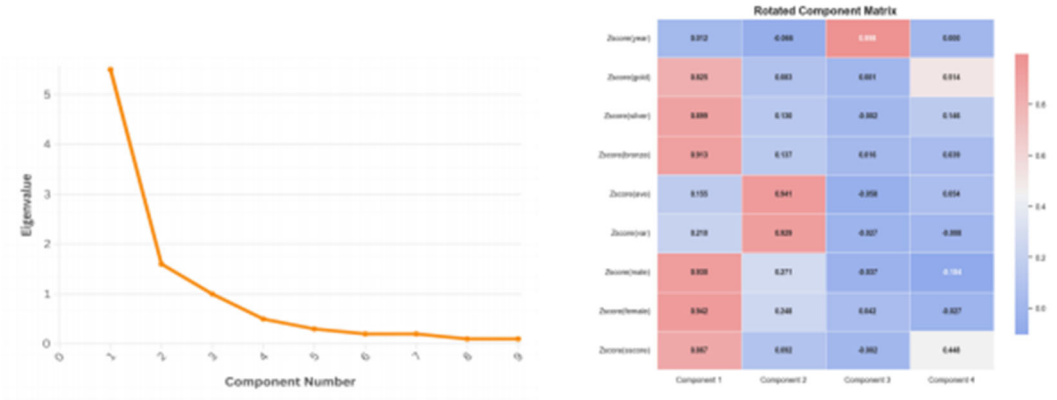

Fig. 3

Considering that there maybe a complex nonlinear relationship between the factors separated above and the predicted results, we choose to use machine learning methods to discover the potential relationship between independent variables and dependent variables. Given the advantages of random forests such as **strong anti-overfitting ability** and **high computational efficiency, we use this method to predict the number of medals for each country.**

(4) Data Process anf Model Building

Before building the model, in order to maintain the consistency of data distribution, we used the Bootstrap method to generate training subsets for the Olympic Games data from 1980 to 2020. Among them, the unselected samples are suitable for model evaluation and are called out-of-bag data (OOB). We conducted a generalization ability test based on this part of the data and took the average values of the obtained R^2 and MSE as indicators of the generalization ability of the random forest model.

When building model, we first input the sample x (including The four factors medals, fatigue, years, elite), each decision tree t predicts a value yt (the prediction of the total number of medals or gold medals of each country in the 2028 Olympics), and the final prediction result is the average of the prediction results of all decision trees. Next, to further find out the prediction interval of the number of medals, we divided 10trees in the random forest to a group, for a total of 100 groups. In this way, 100 training data can be obtained. After calculating the mean, the 10 data points with the

largest deviation from the mean are eliminated through **the MSE comparison method**, and the remaining 90 values are used to determine **the 90% confidence interval**.

After sorting, we took the top ten in the final values predicted by the model as the top ten countries expected to top the medal table in 2028. The results are shown in the figure below:

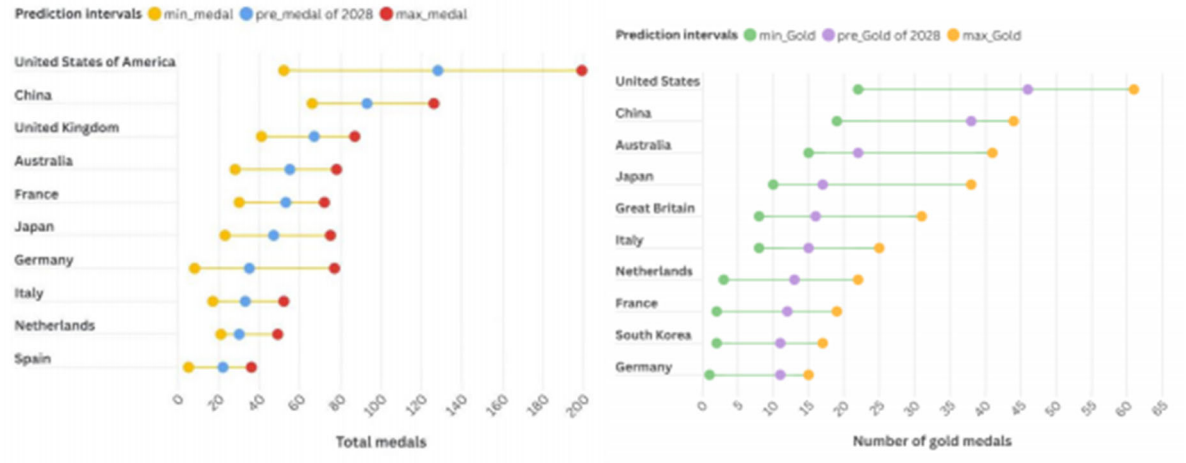

Fig.4: 2028 Olympic (gold) Medal Prediction and Prediction Range

The number of medals can reflect a country's strong comprehensive strength in the field of sports. As can be seen, the United States, China, and the United Kingdom rank in the top three in the 2028 predicted medal list. Gold medals are the symbol of the highest honor in the Olympics, and the ranking of gold medals can directly and clearly reflect the sports strength of a country or region. Although the ranking prediction results of the gold medal list have changed to some extent compared with the medal list, major sports powers (such as the United States, China, etc.) ranked at the top in both predictions, which shows that they have both strong comprehensive strength and top level in events. In addition, we also predicted based on the model that the countries that will make progress or decline in the number of Olympic medals in 2028 compared to 2024. Overall, countries in Oceania, North and South America, and Eastern Europe are showing a trend of winning more medals, while countries in Asia and Africa are mostly declining. Specifically, sports powerhouses such as the United States, Canada, and Australia are likely to continue to demonstrate their advantage in the total number of medals at the 2028 Olympics.

Fig. 5: Progress / deterioration of medal counts by country in 2028 Olympics

3. Q2: First Medal Breakthrough Prediction based on XG-BOOST

In addition to the top-ranked countries those countries that may achieve a breakthrough in the number of medals from scratch should also not be ignored. To date, there are more than 60 countries that have never won an Olympic medal. It is also our goal to analyze whether they can achieve this

historic breakthrough in 2028 and the possibility of it. Considered the fact that the XG-BOOST algorithm has regularization terms, supports to custom loss functions and parallel computing to the more commonly used GBDT algorithm, we chose to use the XG-BOOST classifier to improve optimization efficiency and enhance model interpretability and robustness.

Before building the model, in order to maintain the consistency of data distribution, we preliminarily divided the original data set into a training set Train (80%) that can be used to train the model and a test set Test (20%) for testing the generalization ability of the model. On this basis, $k-fold$ **cross-validation is used to** randomly divide the training set to k equal-sized subsets. In each iteration, $k-1$subsets are selected for training, and the remaining 1subsets are used for validation. In order to reduce the error caused by different sample divisions, the process is repeated p times, and finally the average value of all validation results is used as the basis for model evaluation. Setting $p = 10, k = 10$, which indicates that 10 times 10- fold cross-validation is used to ensure the reliability of model evaluation.

When answering whether we can achieve a breakthrough in predicting the number of medals from scratch, since the types of data and content available for countries that have never won medals are relatively limited, we mainly consider the influencing factor of the total number of athletes participating in the Olympics when making predictions. This can be reflected in the rotated factors *medals* and *fatigue*. Therefore, we used themas a feature for training. Then, the features required to predict the number of medals in 2028 are input, and a new probability array is obtained, representing the probability of attaining a medal. Finally, all countries predicted as "likely to win a medal" are extracted, and countries that have won medals before are removed, and then the six countries that are most likely to achieve a breakthrough in medals are selected. We calculate the odds of each country, and found that Cooperative Republic of Guyana, Republic of Iraq, and The Independent State of Samoa are likely to achieve their medal breakthroughs.

4. Q3: Clustered OLS Prediction Model

Note that in considering tournament dominance, we have considered the number of tournaments themselves and focused more on the change in the number of tournaments over time. We then used least squares to obtain standardized regression coefficients to reveal the relationship between the number and type of tournaments and the number of medals. Before regression, we use systematic clustering to reduce data complexity and better extract the information in the data.

In order to quantify the advantages of events, we need to extract the data from summerOly_ athletes.csv and summerOly_programs.csv. Note that the program data given in the question and the sport column in the athlete data do not correspond well, so we unify the data of the two. For example, " baseball " and " softball " are all marked as " baseball and softball " to make the data of the two tables easier to interact. Due to the diversity of sports, we divide sports into **three types: Dynamic, Static, and Combination of dynamic and static,** abbreviated as **dyn**, **sta**, and **com.**

4.1 Select Suitable Indicators

In order to figure out how well a county preform in a certain event, we borrowed **the concept of market share in economics**. This concept can reflect the monopoly power of an enterprise. As in our model, it can represent a country's dominance in an event p_i. This dominance is defined as:

$$p_i = \frac{s_i}{\sum_i s_i}$$

Where si is the medal score, and si = 3Goldi + 2Silveri + Bronzei.

we designed 6 explanatory variables $x_i (i = 1, 2...6)$ to describe the increase and decrease of each sport type. The dummy variable D indicates whether it is the home country, and constructed a linear relationship between the country events advantage score:

$$d_{ss} = \sum_{i=1}^{6} \beta_i x_i + \delta D + \epsilon$$

Where x_1, x_2, x_3 respectively means the increase of events in three categories; x_4, x_5, x_6 respectively means the decrease of events; and D is a dummy variable (1 when host country, else 0).

4.2 Model Establishment

We use the Euclidean square distanced to describe the distance between two samples. For two clusters, we use the Between Groups Linkage method to calculate the distance D between them.

$$D = \frac{1}{|A||B|} \sum_{a_i \in A, b_j \in B} d(a_i, b_j)$$

We choose the Within-Cluster Sum of Squares (WCSS) to choose the right number of clusters, which can well describe the effect of clustering. It is definite as follows:

$$WCSS = \sum_{k=1}^{K} \sum_{x \in C_k} (x - \mu_k)^2$$

Where μ_k is the barycentric coordinates of this cluster. According to the elbow rule and the result shows in Fig.7, we choose the number of clusters K to be 4. Fig.8 shows the clustering results.

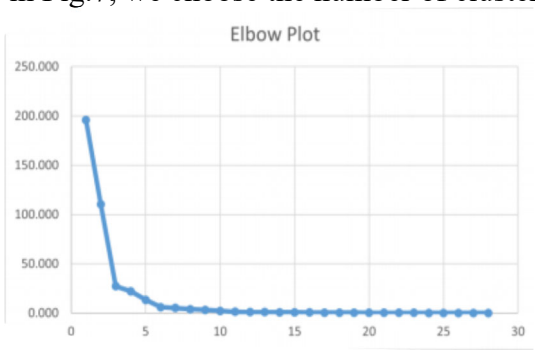

Fig. 7

Fig. 8

4.3 Standard Least Squares Regression

Based on the original algorithm, standard least squares regression standardizes the data, removing the impact of different dimensions between the indicators, and finally obtains the importance of each indicator by directly comparing the absolute values of regression coefficients.

(1) Determination of regression equation

In order to obtain the relative importance of each indicator, we considered using the standard least squares regression method to calculate the standard regression coefficients of the four categories of countries. For the above four categories of countries, our standard regression coefficients are shown in Fig. 9.

(2) Interpretation of Standardized Regression Coefficients

Observing the regression coefficients of the third group (traditional sports powers), we can find that combine-type sports have a greater impact on these countries. And the coefficients of x_1, x_5, are negative, the coefficients of x_2, x_6, are positive correspondingly, indicating that these countries do not dowell in dyn -type and sta -type sport events. For the fourth group of countries the values of these coefficients are very small,indicating that the change of events has little effect

on the country's awards, In addition, we observed the indicator $D$are always positive, indicating that the home country could enhance their events advantage.

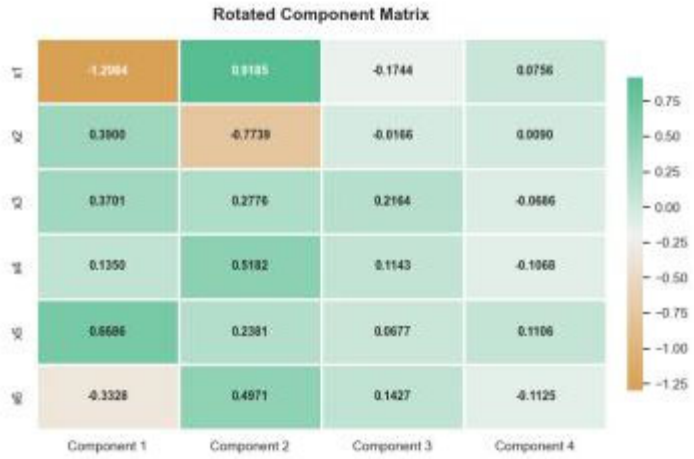

Fig. 9

(3) BP test for disturbance

For the disturbance term i, we hope it is a spherical disturbance term, which means the disturbance term and the explanatory variables are homoscedastic and have no autocorrelation.

H0: There is no heteroskedasticity in the disturbance term

H1 : The disturbance term is heteroskedastic

Under the premise of the original hypothesis, the BP statistic follows the chi-square distribution, where $BP = nR^2$, the calculated p is 0.873 >> 0.05, indicating that there is no heteroskedasticity in the disturbance term.

4.4 Model and results

When the number of Olympic sports events, most standard regression coefficients are positive, reflecting an objective law that the more events there are, the more medals are awarded, and the greater the advantages of each country in the events, thus the higher the medal scores they obtain. We substituted the data back into the model and found that for every additional event, the dominant country (whose share in an event exceeds 0.2) can get 1.17 more medal points, with an average increase of 0.007 points. We noticed that the countries with the highest percentage in the pie chart do not necessarily have the highest standard regression coefficients for the sport categories, which may indicate that these countries are more specialized in these sports and invest less in new sports. Indicators with larger relative regression coefficients are obviously more important to a country. For example, China and Russia have alarge weight indyn- type sports, such as badminton, tennis and swimming, but these are not the best sports for these two countries, they only share parts of proportion. These countries do not have an overwhelming advantage in this sport. Instead, they are evenly matched with other countries. Therefore, we conclude that the most important events for a country are those with suspense, not those who win gold medals without difficulties. From the previous analysis, we know that the presence of the host country has a positive incentive effect on the advantages of the event. We analyzed the specific cases of event changes and found that the host country will expand its event advantages by increasing the events that are beneficial to it and reducing the events that are not beneficial to it.

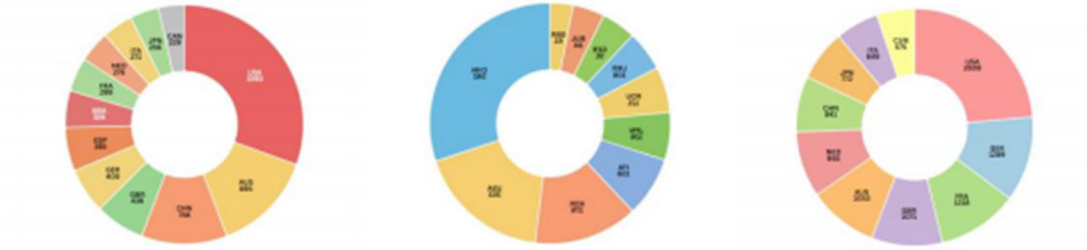

Fig. 10

5. Q4: 'Great Coach' Effect based on PSM - DID

When analyzing the data of each Olympic Games in each country, we found that in addition to the host effect, there are also some years with a large increase in medal data. Some countries have achieved a breakthrough in medals in non-medal events, and some have a large increase in the original ratio, which reveals that the country has made a leap in the level of a certain sports field in a short period of time. Considering that when great coaches change the country they coach, they can bring advanced training models and teaching ideas to the country and can quickly achieve better results in the recent Olympic Games, we analyzed the abnormal data to verify this conjecture.

5.1 DID

In order to quantify the impact of great coach, we first introduced the double difference method (DID) for analysis. It is a common method for estimating intervention and event treatment effects in social science research. It estimates the net effect of the policy by comparing the differences before and after the intervention and the differences between the intervention group and the control group. Its core assumption is that if there is no intervention, the change trends of the treatment group and the control group are the same (parallel trend hypothesis). In DID, $Y_{1,1}$ denotes the mean value of the treatment group after the intervention; $Y_{1,0}$ denotes the mean value of the treatment group before the intervention; $Y_{0,1}$ denotes the mean value of the control group after the intervention; and $Y_{0,0}$ denotes the mean value of the control group before the intervention:

$$DID = (Y_{1,1} - Y_{0,1}) - (Y_{1,0} - Y_{0,0})$$

In the analysis of the "great coach" effect, the time dimension is used before and after the coaching, and the countries or projects that received and did not receive coaching are used as the grouping dimension. The effect of the "excellent coach" is evaluated by comparing the changes in the number of medals of the two groups at different times. However, since the intervention measures in reality are essentially a non-randomized experiment (or quasi-natural experiment), the DID method used in the evaluation of the intervention effect is inevitably subject to self-selection bias. To solve this problem, we next introduced the score propensity matching (PSM).

5.2 PSM

Propensity score matching calculates the probability of each sample receiving intervention (propensity score) and matches samples with similar propensity scores in the treatment group and the control group, thereby reducing the impact of confounding variables. The PSM method can match each sample that has undergone coaching training to a specific control group sample, making the quasi-natural experiment approximately random. This just fills the gap in the DID method mentioned above. We use logistic regression or other models to estimate the propensity score, where T is the treatment variable (whether to receive intervention) and X is the covariate:

$$P(T = 1 \mid X) = \frac{}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

In this case, since there maybe differences in many aspects (such as the number of athletes) between countries that receive "excellent coaches" and those that do not, these differences may interfere with the accurate assessment of the coaching effect, that is, there is selection bias. The PSM algorithm constructs a propensity score function to calculate the probability (i.e., propensity score) of each country receiving guidance from an "excellent coach" based on the characteristics of the country. Then, the countries that receive guidance and those that do not receive guidance are matched according to the propensity scores, so that the two groups of countries after matching are as similar as possible in these characteristics, thereby eliminating selection bias. In Fig. 11, by filtering out the similar parts of the treatment and control groups, we can more clearly recognise the role of imposed variables.

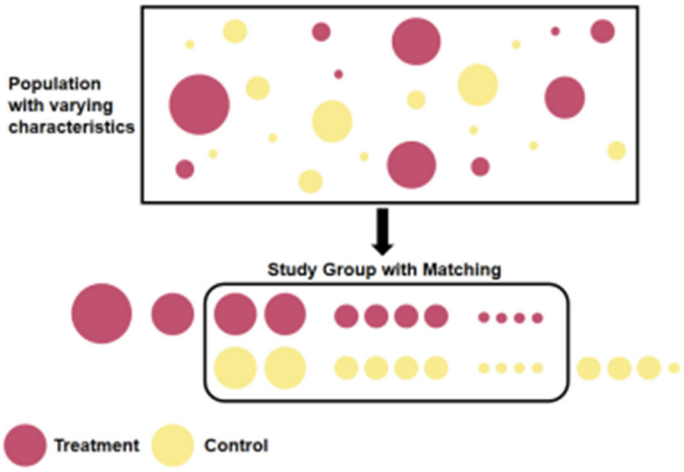

Fig. 11

We grouped the existing data by country, event, and year, calculated the number of medals and the number of athletes, and calculated the annual growth rate by subtracting the mean of the data from previous years. Considering that some countries may achieve a breakthrough from zero to some medals in some events, the growth rate will reach infinity. We will set a growth rate threshold and an absolute medal increase threshold, and screen two categories: breakthroughs in non-medal events and enhancements in non-traditional advantage events, and then screen out special event cases (host effect) to obtain preliminary results. We then conducted 10,000 permutation tests for each sudden growth case and calculated the p-value. If the p-value is <0.05, the medal growth is considered significant, supporting the existence of the "great coach" effect. In the end, we obtained the following four figures, which show the historical data we selected to prove the "great" coach effect. Taking the average of the selected data, we get: the average contribution of the coach effect DID is 7.9619 medals.

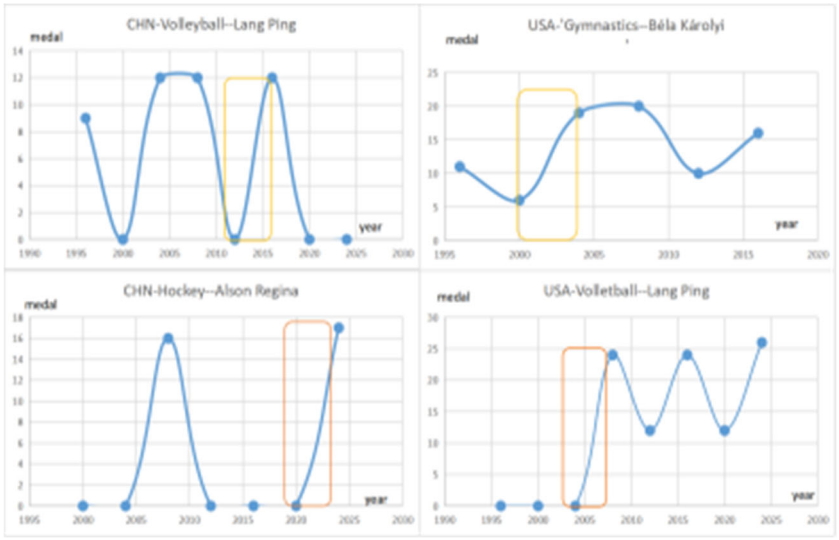

Fig. 12

5.2 Results

It should be noted that if a country already has a considerable advantage in a certain field (traditional advantage projects, such as basketball in the United States), the "great" coach effect is difficult to manifest. To maximize the "great coach" effect,we should focus on those projects that are underperforming or are steadily increasing and have room for improvement. Therefore, we recommend that countries that have high domestic attention but are non-traditional advantage projects (such as Chinese football); and countries that are eager to achieve a breakthrough in the number of

medals in a certain field from scratch invest in "great" coaches. Here, we list three countries and their reasons in Table 2. Based on the average impact of the coaching effect, we estimate that Chinese football can achieve a breakthrough of 3 to 4 medals at most (here we also take into account that one sport can win multiple medals), and the United States and France can achieve an increase of 41.64% to 65.29% in volleyball and basketball.

Table 2

country	Event	Reasons
China	football	No medal since 2000, eager for change
United States	volleyball	Lang Ping coached from 2005 to 2008, reaching the highest level in history, but there is no obvious advantage in recent years.
France	basketball	In recent years, the development momentum of basketball has been stable and positive. We can invest in the "great coach" effect to seek better development.

6. Q5: Further Insights

6.1 The rising status of women

In our factor analysis, we found that the amount of information extracted from the number of female athletes is generally greater than that of male athletes. For this reason, we specifically examined the impact of changes in the number of female athletes on the distribution of national medals. We found that the increase in female sports events has made a huge contribution to the rapid growth of national medals. By screening the data provided, we found that female athletes only began to participate in the Olympics in 1900, and the number was small, accounting for only 2.2% of the total number of athletes. As the concept of gender equality became more popular, by 2000, the proportion of female athletes in the Olympics reached 39%. Countries such as China and the United States have risen rapidly in new events due to systematic training of female athletes. Since 1992, the number of female athletes participating and the medal contribution rate have continued to exceed that of men, which has become an important reason for their long-term top rankings in the Olympics. Finally, in 2012, the number of male and female events was basically equal. This adjustment caused a differentiation among traditional sports powers. Some countries that attach more importance to female events (China, the United States, and the United Kingdom) maintained their advantages by balancing the development of male and female events, while other countries (Germany and Russia) did not make timely adjustments, resulting in fluctuations in rankings.

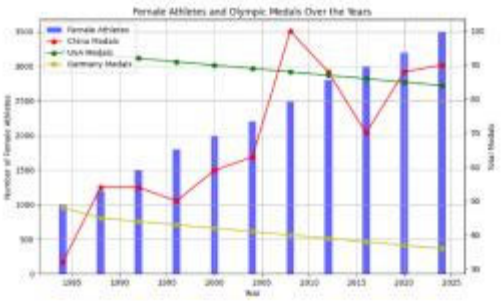

Fig. 13

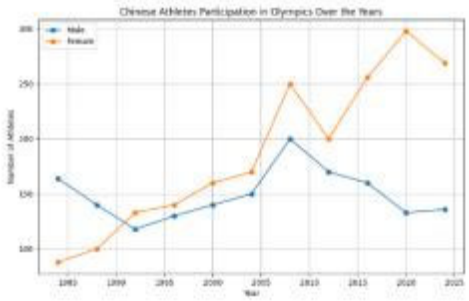

Fig. 14

6.2 Valuing women's abilities

Countries with an unbalanced ratio of male and female medal winners should pay attention to the training of female athletes. The increase in female events has not only had a profound impact on the

Olympic rankings of countries around the world but also reflects the popularity of the concept of gender equality in the world. Paying attention to the training of female athletes is not only a hard requirement for increasing medals and national sports capabilities, but also amoral act to conform to social development, promote gender equality, and create a good social atmosphere. The National Olympic Committee can encourage more female athletes to participate and grow by setting up special women's competitions and appropriately increasing the commercialization and sponsorship drive of women's events.

7. Conclusions

Our algorithm can effectively capture the non-linear relationship between the independent and dependent variables, and is suitable for complex sports data analysis scenarios. Using Clustering-OLS model, countries with multiple repeated features are classified into the same category, which reduces the difficulty of data processing and the impact of multi-collinearity, and also increases the regression data sample, which is more conducive to our analysis. The PSM-DID model is a commonly used and effective causal evaluation method. Compared with the randomized experiment (ABTest) method, it does not require random assignment of treatment groups and control groups, but uses existing observational data to estimate causal effects. This can avoid some ethical, cost, and feasibility issues. Moreover, the combination of PSM and DID methods solves the selection bias problem caused by observable and unobservable variables, and improves the accuracy and credibility of the estimation of the intervention policy effect. However, there are still areas that need improvement. First, if we have more complete data, the prediction of the number of medals in 2028 will be more accurate. For example, we can further improve the practicality of our model by analyzing major competitions such as the World Championships and the world rankings of athletes. Second, Our PSM-DID algorithm assumes that if there is no intervention, the trend of change in the treatment group and the control group is the same (parallel trend assumption). But this is not always true in real life. Countries may be affected by external unobservable factors, which may lead to changes in the number of medals, or even the opposite of our predictions.

References

[1] Csurilla, G., & Fertő, I. (2022). How long does a medal win last? Survival analysis of the duration of Olympic success. Applied Economics, 54(43), 5006-5020.

[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] 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.

[4] Pfau, W. D. (2006). Predicting the medal wins by country at the 2006 winter Olympic Games: An econometrics Approach. *Korean Economic Review*, *22*(2), 233-247.

[5] Becker, A. J. (2009). It's not what they do, it's how they do it: Athlete experiences of great coaching. *International Journal of Sports Science & Coaching*, *4*(1), 93-119.

[6] Simko, I. (2022). We are the champions. The index for evaluating concentration of championships using a sliding window approach. *Heliyon*, *8*(12).

[7] Mitchell, J. H., Haskell, W., Snell, P., & Van Camp, S. P. (2005). Task Force 8: classification of sports. *Journal of the American College of Cardiology*, *45*(8), 1364-1367.

[8] Balmer, N. J., Nevill, A. M., & Williams, A. M. (2001). Home advantage in the Winter Olympics (1908-1998). *Journal of sports sciences*, *19*(2), 129-139.