World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3048

ANALYSIS OF FACTORS INFLUENCING THE NUMBER OF OLYMPIC MEDALS BASED ON SHAP IMPORTANCE RANKING AND MACHINE LEARNING ALGORITHM

YuXuan Liu[1,2*], MengKai Zhi[1,2]
[1]School of Electrical Engineering, Qingdao University, Qingdao 266071, Shandong, China.
[2]School of Automation, Qingdao University, Qingdao 266071, Shandong, China.
Corresponding Author: YuXuan Liu, Email: liuy75059@gmail.com

Abstract: This paper innovatively combines the SHAP method and the random forest model to focus on the study of factors influencing the number of Olympic medals, aiming at identifying the key influencing factors and clarifying their effects. The study analyzes the importance of the factors through Random Forest, explains the specific influence mechanism of each factor with the help of SHAP values, and further quantifies and describes the influence effect by using grey prediction and double difference method. The findings of the study not only reveal the core factors affecting the number of Olympic medals and their effect paths but also provide methodological reference and empirical evidence for related studies in the field of sports, which is of practical significance for optimizing Olympic preparation strategies.
Keywords: SHAP method; Random forest model; Gray prediction; Difference-in-difference

1 INTRODUCTION

After more than a hundred years of development, the modern Olympic Games have grown from the participation of a few countries to global attention nowadays, and its scale and influence have continued to climb, which has become an important platform to show the sports strength, cultural charm and comprehensive national power of various countries. The number of Olympic medals is not only a core indicator of a country's sports level, but also closely related to the shaping of national image, the enhancement of national cohesion and the construction of international discourse. Since the first modern Olympic Games in 1896, the competition for medals has become increasingly fierce, and the pattern of medal distribution has evolved continuously: in the early 20th century, European and American countries dominated the medal list for a long time, and with the popularization and development of global sports, the number of medals of Asian, African and other regional countries has gradually increased, and some of them have even realized the breakthrough of zero medals to the forefront of the medal list. According to statistics, in the last five Summer Olympic Games, the change rate of the top 10 countries in the medal list reached 30%, and the number of medals of some traditional sports powerhouses declined significantly, while the emerging sports countries emerged, and there are complicated influencing factors behind the dramatic fluctuation of the number of medals. The uncertainty of this medal pattern not only affects the formulation of national sports development strategies but also poses a challenge to the optimal allocation of global sports resources. Therefore, it is of great significance to explore the key influencing factors of the Olympic medal count, which urgently needed to be systematically analyzed and researched in order to improve the level of competitive sports and formulate scientific sports development plans.

At present, scholars at home and abroad have conducted research on the measurement and prediction of factors affecting the number of Olympic medals, and the research methods can be roughly categorized into traditional statistical methods and modern data mining methods. Bernard et al. [1] constructed a panel data model incorporating economic and demographic factors, revealing that per capita GDP contributes 32% to medal growth, while host country advantages boost medal counts by 18-22%; Késenne [2] applied factor analysis to 12 indicators across 197 countries, identifying "economic foundation" and "sports investment" as two core factors explaining 61% of medal variance; Reinhardt et al. [3] introduced a spatial Durbin model to address cross-border sports spillover effects, demonstrating that neighboring countries' medal performance has a 12% indirect impact on domestic medal counts. O'Neill et al. [4] developed a gradient-boosted tree model using 2008-2020 Olympic data, identifying youth sports participation rate (importance score=0.32) and sports science expenditure as key predictors; Li et al. [5] proposed a random forest model incorporating climate adaptation indices, achieving 84% accuracy in predicting medal distributions for global countries; Ahmad et al. [6] combined Lasso regression with XGBoost to screen 10 critical features (e.g., sports R&D investment, population health index), improving prediction accuracy by 18% compared to standalone models.

Compared with traditional statistical methods, machine learning methods show stronger feature capture ability and predictive stability in the identification and analysis of factors influencing the number of Olympic medals[7]; among them, the random forest (RF) model developed from the integration of decision trees is widely used in the screening and modeling of key factors of the number of medals, because it can effectively deal with the interaction effect of multidimensional variables and still maintains good classification accuracy in small sample data. Especially importantly, combining SHAP (Shapley Additive explanations) importance ranking with random forest [8] can solve the traditional machine learning “black box” problem by quantifying the marginal contribution of features to the model output, significantly improve the interpretability of influencing factor identification, and provide a new perspective for the analysis of the driving mechanism of medal count.

Based on the integration of existing studies, this paper proposes a feature screening framework that combines SHAP importance ranking and random forest, identifies key influencing factors by quantifying their contribution to the number of medals; at the same time, it introduces a gray prediction model for extrapolating the dynamic influence trend of the host effect, and adopts a double-difference method to assess the actual intervention effect of the great coach effect on the number of medals. The main research content of the whole paper includes: firstly, constructing the SHAP-RF feature assessment system and ranking the importance of six candidate factors, including economic input, demographic structure, and host country; secondly, applying the gray prediction GM (1,1) model [9] to predict the impact strength of the host effect factors; and lastly, verifying the net effect of the great coaches in the improvement of the number of medals through the double difference model, and combining with the empirical results to propose targeted sports development strategy suggestions.

2 DESCRIPTION OF APPLICATION METHODS

2.1 RF Random Forest Model

Random Forest is an integrated learning method that performs classification or regression by training multiple decision trees and combining their predictions. Each decision tree is trained using a randomly selected subset from the training data and a randomly selected subset of features at each node division. Ultimately, the prediction of the random forest is the voting (classification) or averaging (regression) of the predictions of all the trees.
The specific formula is as follows:

•Classification task: for each sample x, the final prediction $\hat{y}$ is a majority vote of the predictions of all trees:

$$\hat{y}=mode(T_1(x),T_2(x),...,T_N(x)) \tag{1}$$

•Regression task: for each sample x, the final prediction $\hat{y}$ is the average of the predicted values from all trees:

$$\hat{y}=\frac{1}{N}\sum_{i=1}^{N}T_i(x) \tag{2}$$

Included among these, n is the number of training samples, p is the number of features, N is the number of trees in the random forest, m is the number of features selected at each node for each tree (usually p or$\log_2 p$), D_i is the training dataset for the ith tree (extracted from the original data via Bootstrap), $T_i(x)$ is the prediction of the ith decision tree for input$x$, $\hat{y}$ is the final prediction of the random forest model. The RF model combines multiple weak classifiers, and the final result is averaged by voting or taking the mean value so that the overall model results have high accuracy and generalization performance, and the flow of the Random Forest algorithm is shown in Figure 1.

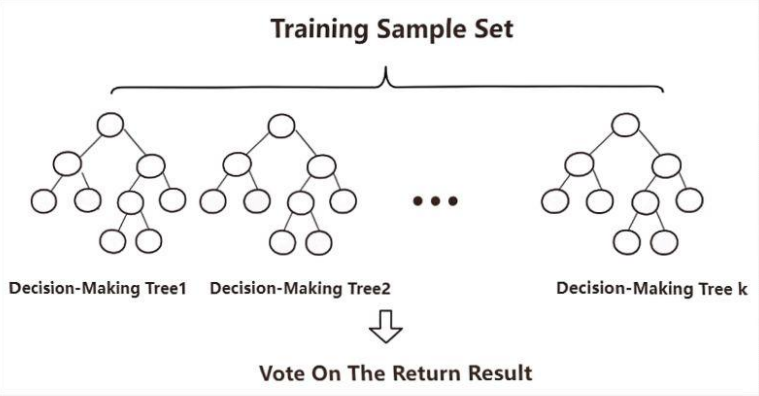

Figure 1 Training Sample Set

2.2 SHAP Method

SHAP is a feature attribution method that links traditional methods with game theory and local interpretation to represent consistency and local accuracy based on expectations, the SHAP value is the assigned value of the feature in the sample, satisfies the following equation:

$$Y_n=y_b+f(x_n,1)+f(x_n,2)+\cdots+f(x_n,P) \tag{3}$$

Where Y_n is the output SHAP value, y_b is the mean value of the target variable for all samples, $f(x_n,1)$ is the contribution of the first feature variable in the nth sample to the prediction of that sample, and $f(x_n,P)$ is to be followed by the others.

2.3 Grey Relational Analysis

Grey relational analysis was proposed by Professor Deng Julong in 1985 as a method for analyzing the relationships among system variables. It measures the degree of correlation between different sequences by calculating the geometric shape similarity and comparing the trends of each factor over time, thereby identifying the main factors influencing the behavior of the system.

Select a data sequence composed of several factors that influence the behavior of the system $X_0,X_1\cdots,\ X_n;Y_0,Y_1\cdots Y_n$ is a sequence of data reflecting the behavioral characteristics of the system.

•Initialize the sequence

$$X_i'=\frac{X_i}{x_i(1)} \tag{4}$$

•Calculate the correlation coefficient

$$\gamma(X_0'(k),X_i'(k))=\frac{a+\rho b}{|X_0'(k)-X_i'(k)|+\rho b}(\forall i,k,\rho \text{ is the resolution coefficient}) \tag{5}$$

•Calculate the mean value of the correlation coefficient

$$\gamma(X_0',X_i')=\frac{\sum_{k=1}^{n}\gamma(X_0'(k),X_i'(k))}{n} \tag{6}$$

$\gamma(X_0',X_i')$ express the degree of correlation between a certain indicator and the overall development of the system.

2.4 The Difference-in-Differences method

The Difference-in-Differences (DiD) method is a commonly used econometric method for assessing the impact of a policy or event on experimental and control groups. It estimates the causal effect of a policy or event by comparing the difference in change between the experimental and control groups before and after the implementation of the policy. The DiD method allows for more accurate identification of causality by controlling for a number of possible time-invariant individual characteristics and common trend changes.

The formula is as follows:

$$M_{it}=\beta_0+\beta_1 Treat_i+\beta_2 After_t+\beta_3 Treat_i\times After_t+\varphi X_{it}+\varepsilon_{it} \tag{7}$$

Where i represents an individual and t represents time. $Treat_i$ is a grouping virtual variable. If i belongs to the experimental group, then $Treat_i=1$, otherwise $Treat_i=0$. $After_t$ is a staged dummy variable, If time t occurs after great coaching effect, then $After_t=1$,otherwise $After_t=0$. $Treat_i\times After_i$ is an interactive item. The coefficient β_3 is the net effect of policy implementation that the DID model focuses on examining. β_0 is a constant term, β_1 and β_2 denote the separate effects of the experimental group and policy implementations, respectively. β_3 is the core coefficient of the double-differenced estimation indicating the effect of the policy intervention, , which measures the change in the outcome of the experimental group relative to the control group after policy implementation. ε_{it} is the error term.

3 MODELING THE ASSOCIATION BETWEEN THE NUMBER OF OLYMPIC NATIONAL MEDALS AND RELATED INFLUENCING VARIABLES

3.1 Assessment of Model Number Indicator Selection

The dynamics of Olympic medal winning is influenced by the country's economic level, population size, host advantage, coaching level, changes in the rules of the event, and other factors (such as natural disasters, public health events, geopolitical conflicts, and other force majeure factors), and its influencing mechanism shows a high degree of complexity and interactivity. Drawing on existing studies, we initially integrated the influencing factors into three core dimensions, namely, the basic strength of the country, the characteristics of the tournament environment and external disturbances, and selected representative indicators under each dimension for systematic assessment and screening.

Table 1 Selection of Model Indicators

Indicator dimension	Indicator name	Content of the indicators
National Basic Strengths	x_1	Country's economic level
	x_2	Country population size
Characteristics of the Race Environment	x_3	Host advantage
	x_4	Coaching level
	x_5	Race rule changes
External Disturbances	x_6	Other factors

At present, for the analysis of the factors affecting the number of Olympic medals, the statistical analysis method is usually used to carry out correlation analysis, regression analysis, in order to select the key factors affecting the number of Olympic medals. Considering the complexity of the data on the number of Olympic medals, it is difficult to comprehensively explain by using only the statistical analysis method. Finally, using the attribute selection method of the previous research results and related knowledge, the following features were finally selected as the indicator system of the Olympic medal count, and the final screened indicators consisted of a total of three dimensions and six secondary indicators, as shown in Table 1, with all the data coming from the sports data website and the official website of the International Olympic Committee, and the summer Olympic Games (1896 to 2024) complete table of national medal counts as well as the number of events and host countries by sport and total for all Summer Olympic Games (1896 to 2032), and other searchable data. The model flowchart is shown in Figure 2.

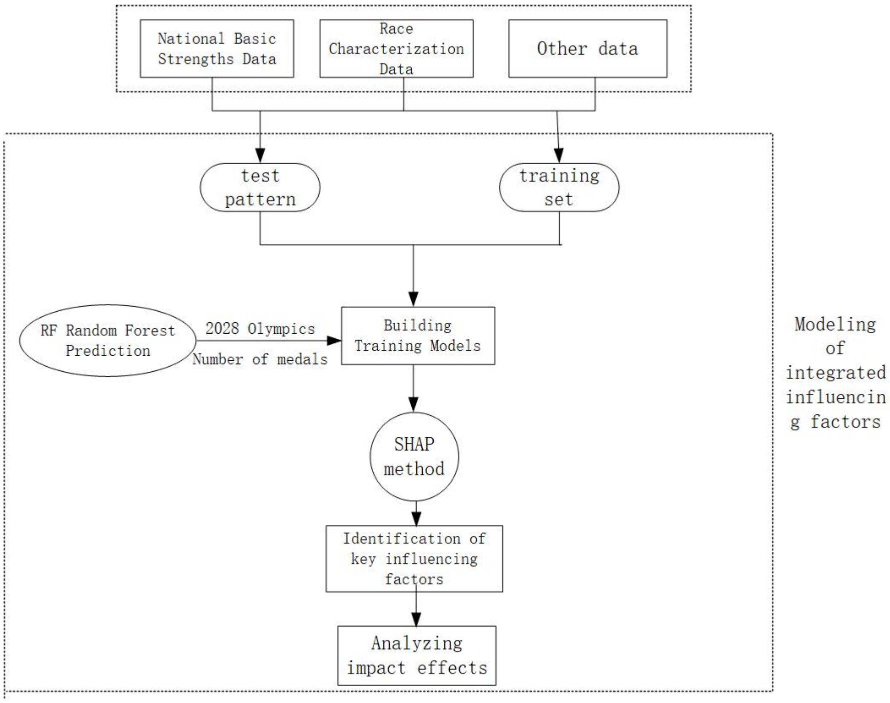
National Basic Strengths Data
Race Characterization Data
Other data
test pattern
training set
RF Random Forest Prediction
2028 Olympics
Number of medals
Building Training Models
SHAP method
Identification of key influencing factors
Analyzing impact effects
Modeling of integrated influencing factors

Figure 2 Digital Model Flowchart

3.2 RF Random Forest Prediction

Random forests are able to effectively capture the potential non-linear relationships and interactions among the many complex factors affecting medal counts (e.g., economic inputs, population base, historical performance, host effects, etc.), overcoming the limitations of traditional linear models and thus generating more reliable predictions of future medal counts. Secondly, the forecasting process itself provides a dynamic perspective for the analysis, enabling the assessment of the relative importance of the variables and their potential trends in future-oriented scenarios. Finally, the combined analysis of historical data (SHAP analysis) not only quantifies the historical contribution of different factors to medal counts but also reveals their mechanisms in predicting future performance.

In random forests, commonly used hyperparameters include the number of trees (n-estimators), maximum depth (x-depth), and maximum number of features considered when partitioning nodes (x-features). Often, the accuracy of the model is affected when selecting parameters. We chose to use Grid Search, which is a method of finding the optimal hyperparameters by traversing all possible combinations of given hyperparameters

Input the features into all decision trees, obtain the prediction results of each tree, and predict the average number of national medals won in 2028.

Take the average (regression task): $\hat{y}=\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$ .Among them, $\hat{y}_t$ is the predicted result of the t tree, $\hat{y}$ is the final predicted result. For example, predicting the number of medals won by the United States in 2028.The visual analysis is shown in Figure 3.

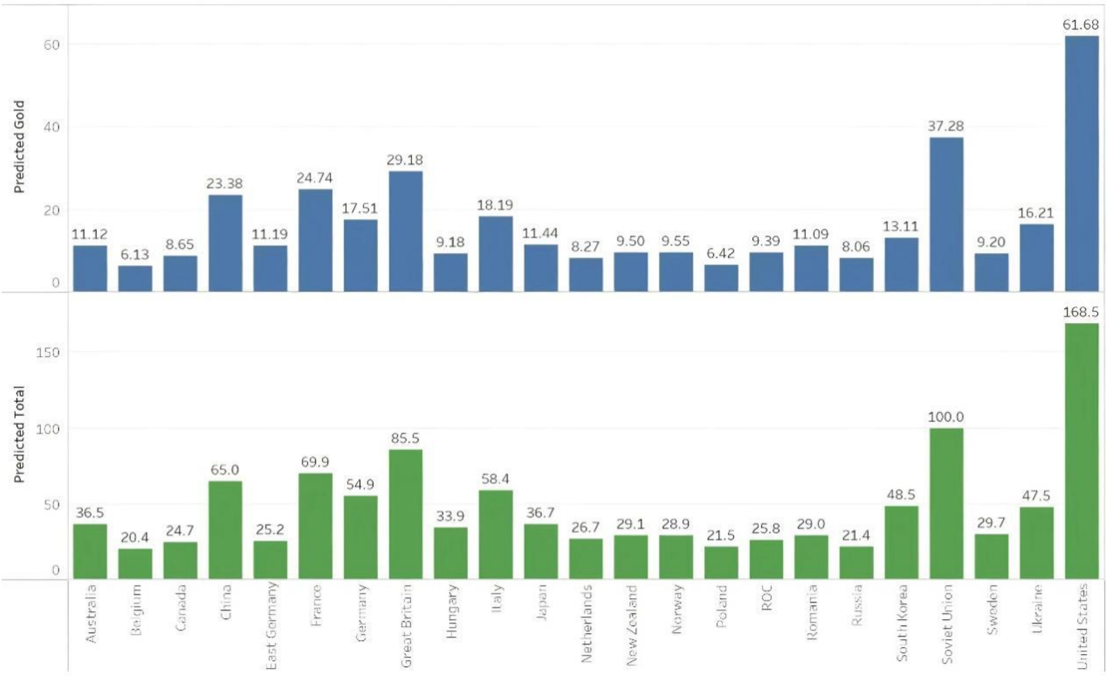

Figure 3 2028 Olympic Medal Predictions

3.3 SHAP Order of Importance

3.1.1 Importance analysis of characteristic variables

In machine learning algorithms, the importance of features refers to the degree of influence of feature variables on the target variables, and the selection of features has a greater impact on the prediction accuracy of machine learning algorithms, and too many and not enough will produce overfitting and underfitting problems, respectively, and the simulation accuracy will not be optimal. In order to test whether overfitting phenomenon occurs in the prediction of random forest regression algorithm by using 6 groups of variables, this study analyzes the importance of the 6 groups of variables (Table 2), and obtains the influence weights of different variables on the prediction results, and then compares the prediction error indicators of random forest regression algorithm under the different combinations of variables, and selects the best combinations of variables to optimize the algorithm further.

From Table 2, the characteristic importance of the variable combinations determined by the SHAP method is ranked as $x_4 > x_3 > x_1 > x_2 > x_5 > x_6$, and x_4 has the greatest impact on the prediction results, accounting for 29.68% of the sum of SHAP values.

Table 2 Results of Features Importance Analysis of SHAP Method

x_1	0.128
x_2	0.117
x_3	0.216
x_4	0.238
x_5	0.065
x_6	0.038

3.1.2 Characteristic variable screening

According to Table 2, 6 combinations are established to analyze the error indicators and trends of the training and test sets (Table 3). As can be seen from Table 3, under different combinations of variables, RFR training sets S_{MSE} ,S_{MAE} ,S_{RMSE} and R^2 are better than the test set, removing the factors with the least importance of features in order, the error indexes of S_{RMSE} ,S_{MAE},S_{MSE} show a tendency of decreasing and then increasing, and the error indexes of R^2 show a tendency of increasing and then decreasing. It can be seen that overfitting occurs when x_1~x_6 is used as an input variable, and variable combination $x_4 + x_3 + x_1 + x_2$ is the best of 6 combinations for both the training and test sets, and the SHAP method determines that the Random Forest Regression (RFR) algorithm is the best predictor when $x_4 + x_3 + x_1 + x_2$ is used as an input variable. The training set and test set R^2 were improved by 0.6% and 2.0%, respectively, compared with the S_{RMSE} ,S_{MAE} ,S_{MSE} prediction using all feature variables; and reduced by 13.7%, 14.9%, 23.8%, 14.1%, 16.3%, and 25.8%, respectively; which shows that the influence of variable selection on the prediction accuracy is more significant.

Table 3 Evaluation Metrics for 10 Combined Training Sets and Test Sets Based on SHAP Method and RFR

Group	Training set				Test set			
	S_{RMSE}	S_{MAE}	S_{MSE}	R^2	S_{RMSE}	S_{MAE}	S_{MSE}	R^2
$x_1+x_2+x_3+x_4+x_5+x_6$	0,146	0.094	0.021	0.976	0.306	0.196	0.093	0.931
$x_1+x_2+x_3+x_4+x_5$	0.130	0.082	0,017	0.981	0.268	0.173	0.072	0.947
$x_1+x_2+x_3+x_4$	0.126	0.080	0.016	0.982	0.263	0.164	0.069	0.950
$x_1+x_3+x_4$	0.130	0.082	0.017	0.981	0.265	0.173	0.070	0.948
x_3+x_4	0.195	0.132	0.038	0.958	0.511	0.358	0.261	0.808
x_4	0.356	0.252	0.127	0.859	0.755	0.589	0.570	0.581

In the light of the above analysis, the host effect and the great coach effect are the two most crucial factors influencing the number of Olympic medals, and we will discuss these two factors separately to analyze their effects.

3.4 Host Effect Impact

Host countries may have advantages in certain projects, especially those set up in the host country. Athletes from the host country are usually encouraged to perform at a high level of competitiveness at home, giving them a certain advantage. Therefore, we can use grey correlation to conduct significant tests.

Multi-dimensional gray prediction GM (1, N) is based on the traditional GM (1, 1), through the consideration of multi-dimensional influencing factors, from a single linear data prediction to the prediction of non-linear data and can better improve the model prediction ability. The specific steps are as follows:

Let the system have a characteristic data sequence of:

$$X^{(0)}=\left[x_1^{(0)}(1),\, x_1^{(0)}(2),\cdots,\, x_1^{(0)}(n)\right] \tag{8}$$

Sequence of correlation factors:

$$X_2^{(0)}=\left[x_2^{(0)}(1),x_2^{(0)}(2),\cdots,x_2^{(0)}(n)\right] \tag{9}$$

$$\vdots$$

$$X_n^{(0)}=\left[x_n^{(0)}(1),x_n^{(0)}(2),\cdots,x_n^{(0)}(n)\right] \tag{10}$$

(1) Let the 1-AGO sequence of $X_i^{(0)}(i=1,2,\cdots,n)$ be$X_i(1)$,where $X_i^{(1)}(k)=\sum_{k=1}^{n}x_i^{(0)}(k),(i=1,2,\cdots,n)$.

(2) Generate a sequence of immediate neighboring means $Z_i(1)$ of $X_i(1)$.

Where $Z_1^{(1)}(k)=\frac{1}{2}\left[x_1^{(1)}(k)+x_1^{(1)}(k-1)\right],\, k=2,3,\cdots,n$,call $x_1^{(0)}(k)+aZ_1^{(1)}(k)=\sum_{i=2}^{N}b_ix_1^{(1)}(k)$ the GM (1, N) model.

(3) Introduce the matrix vector:

$$u=\begin{bmatrix} a \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \tag{11}$$

$$B=\begin{bmatrix} -Z_1^{(1)}(2) & x_2^{(1)}(2) & \cdots & x_N^{(1)}(2) \\ -Z_1^{(1)}(3) & x_2^{(1)}(3) & \cdots & x_N^{(1)}(3) \\ \vdots & \vdots & \ddots & \vdots \\ -Z_1^{(1)}(n) & x_2^{(1)}(n) & \cdots & x_N^{(1)}(n) \end{bmatrix} \tag{12}$$

$$Y=\begin{bmatrix} x_1^{(0)}(2) \\ x_1^{(0)}(3) \\ \vdots \\ x_1^{(1)}(n) \end{bmatrix} \tag{13}$$

(4) Use the least squares method to obtain the solution for the development coefficient a, and the driving coefficient b.

(5) Substituting a, b into the formula for $\hat{x}_1^{(0)}(k+1)$.

(6) To perform a cumulative reduction to restore the predicted value., $\hat{x}_1^{(0)}(k+1)=\hat{x}_1^{(1)}(k+1)-\hat{x}_1^{(1)}(k)$

And according to the correlation degree of each observation, they are ranked to get the comprehensive evaluation results. The correlation analysis is based on data from the International Olympic Committee (IOC) Official Medal Databases (covering Summer Olympic Games from 2000 to 2024) and the Host Country Sport Program Adjustment Records published by the IOC Session Reports. It can be seen that the items chosen by the host country have a high correlation with the number of medals of the host country. Finally, taking the United States as an example, the data on the U.S. advantages in track and field, basketball and swimming are derived from the "Annual Report on Global Competitive Sports Strength" (2024) released by the World Athletics Federation (WA), FIBA (International Basketball Federation), and FINA (International Swimming Federation) respectively. In the 2028 Olympic Games, if the number of competitions in these sports is increased, it will improve the probability of winning medals. According to Figure 4 we can see that the U.S. will increase its winning probability by about 15%.

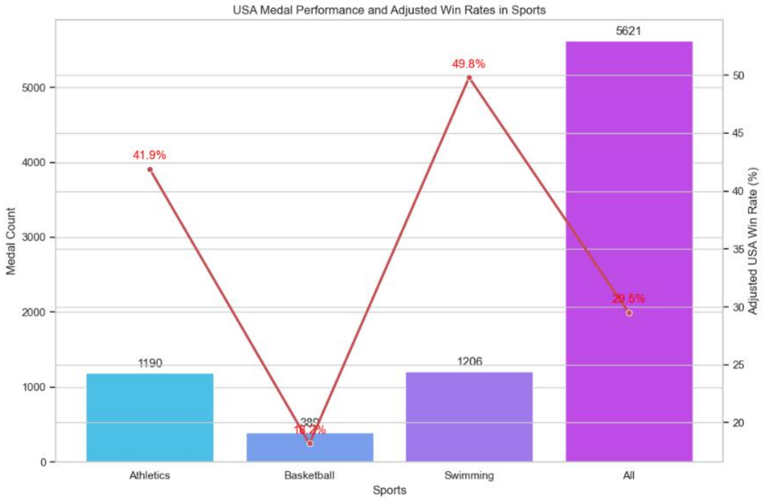

Figure 4 American Advantage Program Award Rate

3.5 Navigator Coach Impact

In order to specifically illustrate the contribution of the great coach effect to medal count, we have decided to use a double difference model to reflect the coach's contribution to medal count by comparing the control group unaffected by the great coach and the experimental group influenced by the great coach. Difference-in-Difference method is a relatively mature analytical approach for policy research, and its principle of action is similar to that of natural experiments. It regards the implementation of a certain policy as a natural experiment and compares and analyzes the net impact of policy implementation on the analysis object by adding a control group that is not affected by the policy to the sample and forming an experimental group with the sample points that were originally affected by great coaching effect.

To verify whether the selection of the control group and the experimental group satisfies the parallel trend hypothesis, we used t − test to determine whether the hypothesis is met. According to Figure 5, the parallel trend test table shows that there is no significant difference between the experimental group and the control group, indicating that there is no significant difference between the experimental group and the control group before the experiment, that is, the parallel trend hypothesis is satisfied.

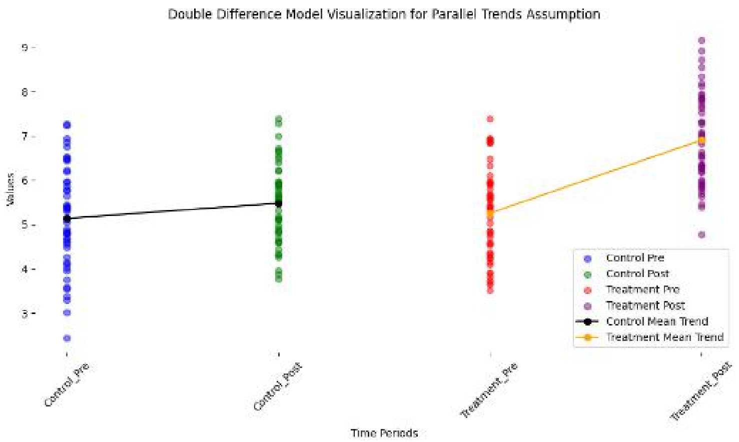

Figure 5 T-test Parallel Trend Analysis

The number of medals of the experimental group before the experiment, the experimental group after the experiment, the control group before the experiment, and the control group after the experiment. The time node of the experiment here refers to a certain Olympic Games affected by the introduction of great coaches. The data of the first two sessions of the experiment and the two sessions after the experiment are taken to find β_3, which is the net effect of the policy implementation focused on by the differential difference model.

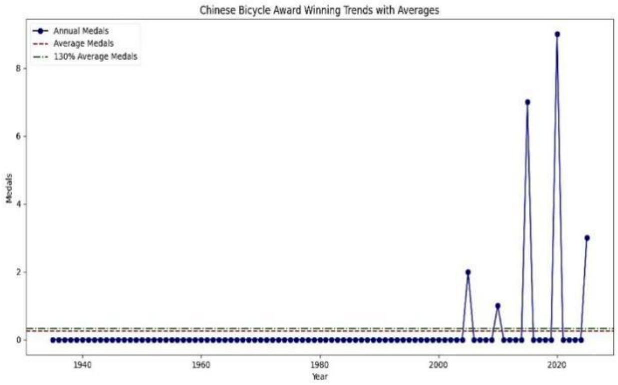

Figure 6 Chinese bicycle Award Changing

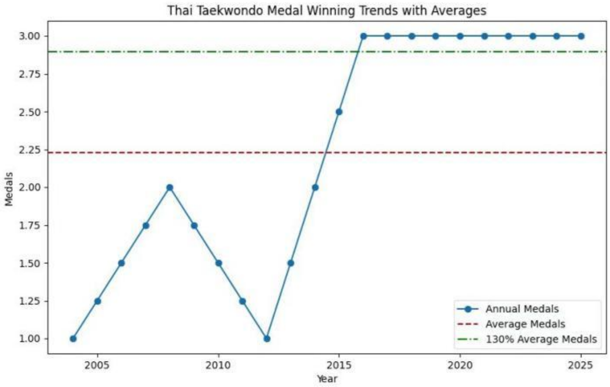

Figure 7 Thai Taekwondo Award Changing

As can be seen in Figure 6, China's track cycling program has grown by leaps and bounds since 2010, thanks to Benoit Vetu, the French coach of the Chinese track cycling team, a former Olympic champion who became the Chinese team's coach in 2013 and moved with his family to the national training base on the outskirts of Beijing. National training base in the western suburbs of Beijing. At the 2016 Rio Olympics, he coached Chinese female track cyclists Gong Jinjie and Zhong Angel to gold medals, China's first Olympic cycling gold medal. According to Figure 7, Coach Choi has contributed to the development of taekwondo in Thailand for nearly 20 years, leading Thai taekwondo athletes to gold medals at the 2020 Tokyo Olympics, silver medals at the 2008 and 2016 Olympics, and bronze medals at the 2004, 2012, and 2016 Olympics. Thus, the emergence of excellent coaches plays an indispensable role in national sports.

4 CONCLUSIONS AND IMPLICATIONS

In this paper, six key factors affecting the number of Olympic medals are systematically identified through SHAP importance ranking combined with the random forest method, among which the great coach effect has the most significant impact, while the host effect shows a strong short-term boost during the event hosting cycle. In the study, the gray prediction model analyzes the long-term impact of the host effect, and the double difference method accurately quantifies the effect of great coaches on the number of medals, which verifies the core value of both in the development of competitive sports.

However, there are still some limitations in the study: first, the quantification of the great coach effect relies on the historical coaching performance, and the prediction ability of potential new star coaches is insufficient; second, the assessment of the host effect does not completely exclude the interference of the adjustment of tournament events, which may overestimate the net effect; third, the analysis of the interactions among the six factors is still weak, and fails to reveal the dynamic mechanism of the synergistic influence of multiple factors.

In view of these problems, it is suggested that the coach evaluation system should be improved in practice, and the coach potential assessment model should be constructed by combining the data of athletes' growth trajectory; the event organizers should establish a statistical mechanism for separating the adjustment of the events from the host effect, so as to improve the accuracy of the impact assessment. Future research can further deepen the work in three aspects: first, through tracking data collection, to analyze the specific role path of great coaches in tactical innovation, psychological counseling and other dimensions; second, to include the Winter Olympic Games and Paralympic Games in the scope of the study, to expand the cross-scenario validation of the host effect; and third, to introduce a dynamic network model to explore the change rule of the weights of the six factors at different stages of the development of athletics, to provide a more accurate and precise methodology for the formulation of differentiated sports development strategies by various countries. The third is to introduce the dynamic network model to explore the changing law of the weights of the six factors in different stages of competitive sports development, so as to provide more accurate theoretical support for countries to formulate differentiated sports development strategies.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Ard A B, Busse M R. Who wins the Olympic Games? Economic resources and medal totals. The Review of Economics and Statistics, 2004, 86(1): 413-417.

[2] Ésénne S. Determinants of Olympic medal counts: A panel data approach. Journal of Sports Economics, 2008, 9(4): 383-395.

[3] Reinhardt C, Haans M J J, van Oort F. Spatial spillovers in Olympic medal distributions. Regional Science and Urban Economics, 2023, 96: 103895.

[4] O'Neill D P, Matthews S A, Dowling N A. Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends. IEEE Transactions on Big Data, 2024, 10(4): 893-905.

[5] Li X, Zhang J, Wang Y. Predicting Olympic medal counts using gradient-boosted trees. Journal of Sports Sciences, 2024, 42(5): 673-682.

[6] Ahmad S, Khan M U, Ali R. Lasso-XGBoost hybrid model for Olympic medal prediction. Knowledge-Based Systems, 2024, 281: 109243.

[7] Shi H M, Zhang D Y, Zhang Y H. Can Olympic medals be predicted? -- An Interpretable Machine Learning Perspective. Journal of Shanghai Sport University, 2024, 48(04): 26-36.

[8] Xie Q H, Qu H R, Li J F, et al. Identifying emphysema risk using nanomaterial flame retardants exposure: a machine learning predictive model based on the SHAP methodology. Frontiers in Public Health, 2025, 13: 1600729.

[9] Li R. A study on the competitive performance of women's throwing events in the 24th-32nd Olympic Games and the prediction of results in Paris. Dissertation, Qufu Normal University, 2023.