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

Unlocking Olympic Success: Predictive Modeling and Strategic Insights for the 2028 Games

Xintong Zheng[1], Yajiao Liu[2], Tianyuan Zhang[3]

[1]Barnard college of coulumbia university, New York, 10027, United States of America;

[2]The Hong Kong Polytechnic University, Hong Kong, 999077, China;

[3]Fuyang Normal University, Fuyang, 236041, China;

Xintong Zheng, Yajiao Liu, Tianyuan Zhang are co-first authors

Abstract. This study aims to predict the medal distribution of the 2028 Olympic Games in Los Angeles by constructing multivariate linear regression and random forest regression models, combining the characteristics of historical data, the advantages of the host country, the number of athletes, and the proportion of gold medals. The results show that the total number of gold medals in the 2028 Olympic Games is expected to be about 255, and the total number of medals is about 714. Feature engineering analysis showed that the average number of gold medals and total medals in the past three years were the key predictors. Model comparisons show that random forests are more advantageous in handling non-linear relationships, but have a tendency to overestimate. The study recommends that National Olympic Committees differentiate the allocation of resources, pay attention to the special needs of the host country, and analyze the data of abnormal years in depth to improve the prediction accuracy.

Keywords: Feature Engineerin; the 2028 Olympic Games; Predictive Modeling.

1. Introduction

Olympic medal prediction is both a scientific exploration of the Olympic Movement and a contribution to global sports development. Through this study, we hope to provide a scientific basis and strategic guidance for the future development of the Olympic Movement. First, to predict the medal distribution of the 2028 Los Angeles Olympic Games, we built models using multiple linear regression and random forest regression. The multiple linear regression model provides direct parameter interpretation, while the random forest regression model utilizes integrated learning to improve prediction accuracy and robustness. The prediction results show that the total number of gold medals for the 2028 Olympics is expected to be about 255, and the total number of medals is about 714. We constructed a feature engineering model to predict the number of Olympic medals, taking into account historical performance, cumulative achievements, interactions, and other factors such as country code and host status. The optimized random forest model emphasizes that the average number of gold medals and the total number of medals over the past three years are key predictors. Finally, NOCs should differentiate their resource allocation between high-performing and volatile countries and sports. In-depth analyses should be conducted for abnormal years to improve forecasting accuracy and optimize decision-making. At the same time, special attention should be paid to the special needs of host countries to ensure that their Olympic advantages are fully utilized.

2. Data Preprocessing And Analysis

2.1 Data Preprocessing

Integrate multiple data sets into a unified data table to facilitate subsequent analysis and modeling. By integrating various data sets, more comprehensive information can be obtained, such as the number of athletes, winning rate, host country information, etc. The given dataset contains the number of medals in each country, and is associated with summerOly_athletes to calculate the number of

athletes and winning rate of each country; summerOly_hosts is associated to add host country information; summerOly_programs is associated to add sports information.

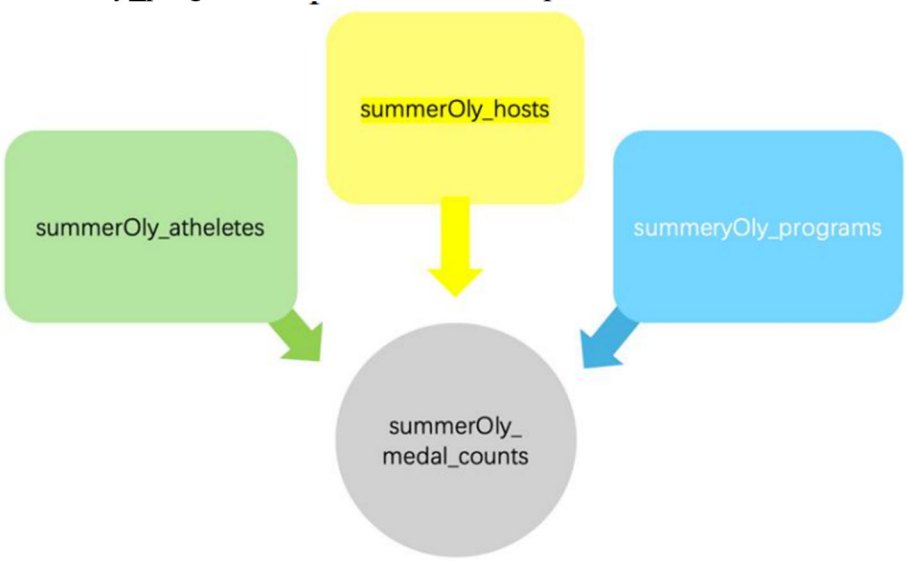

Fig. 1 Two or more referencesData Relationship Diagram

(1) Handling missing values: Use the mean value to fill missing numerical variables. Here, the mean of the column is used to fill missing values to avoid data bias caused by missing values.

$$\text{Fill value} = \text{Average value } (\mathbf{x}) \tag{1}$$

(2) Processing duplicate data: Delete completely duplicate records

(3) Handling outliers: Use the custom correct_outliers function to correct outliers and replace them with the rolling mean of the previous three data points. Through data cleaning, the integrity and accuracy of the data are ensured, laying the foundation for subsequent analysis.

$$\text{Corrected value} = \text{Rolling mean } (x, \text{window}=3) \tag{2}$$

2.2 Feature Engineering

Create new features to enhance the expressiveness of the model. Through feature engineering, we can extract more useful information and enhance the expressiveness of the model.

(1) **Gold medal ratio**: Calculate the ratio of gold medals to the total number of medals.

$$\text{Gold\_Ratio}=\frac{Gold}{Total} \tag{3}$$

(2) **Host country advantage**: Whether or not a country is the host is converted into a binary variable (0 or 1).

(3) **Numericalization of categorical variables**: Convert categorical variables (such as Team) into numerical variables.

2.3 Data Normalization

Normalize the numerical variables to the same scale to prevent certain features from dominating the model due to their large values. Normalization can eliminate the dimensional differences between different features and scale the data to a scale of 0-1, making the model more stable.

$$z=\frac{Gold}{Total} \tag{4}$$

2.4 Data Analysis

On the basis of the above data processing, the data is appropriately visualized to clarify the data distribution characteristics and lay the foundation for subsequent problem-solving. The distribution of gold medals, the relationship between the number of gold medals and the number of athletes,

the change in the number of sports events, and the relationship between the proportion of gold medals and the winning rate of athletes are shown in Figures1to 4

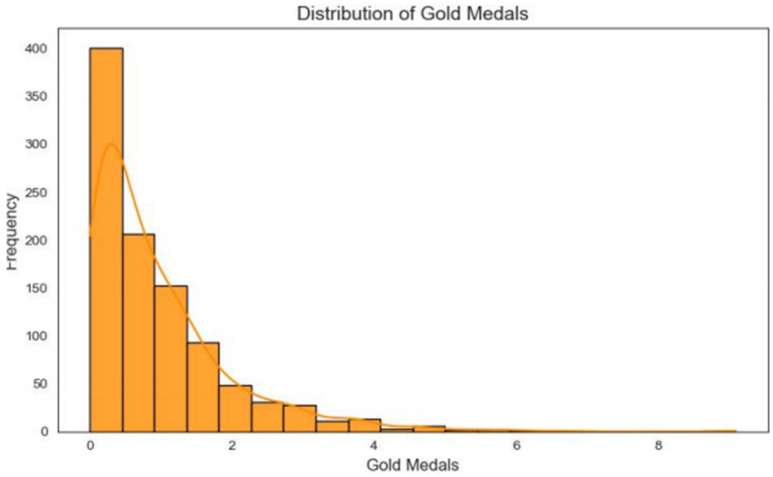

Fig. 2 Frequency Distribution of Gold Medals

As can be seen in the figure, the distribution of the number of gold medals is uneven. In most cases, the number of gold medals is small, while in some instances, the number of gold medals is large. However, although most of the gold medals are concentrated in the lower values, there are still some countries or athletes who have won more gold medals, forming a long tail.

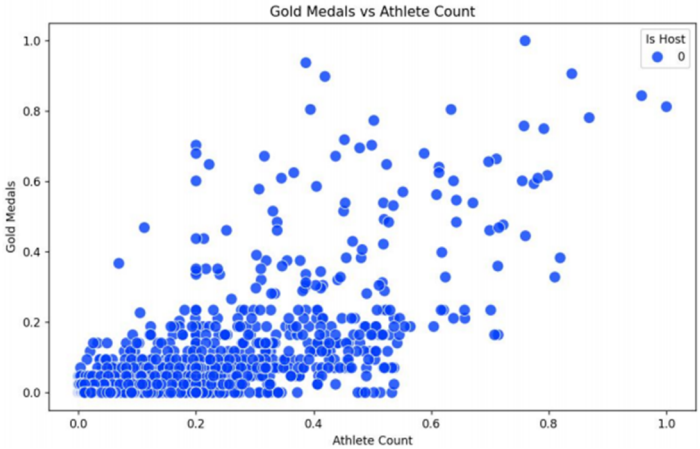

Fig.3 Relationship between the number of gold medals and the number of athletes

As can be seen in the figure, most data points are concentrated in areas with low numbers of athletes and gold medals. As the number of athletes increases, the distribution of gold medals becomes more dispersed. At the same time, there is no obvious linear relationship between the number of gold medals and the number of athletes. Countries or regions with more athletes do not necessarily win more gold medals.

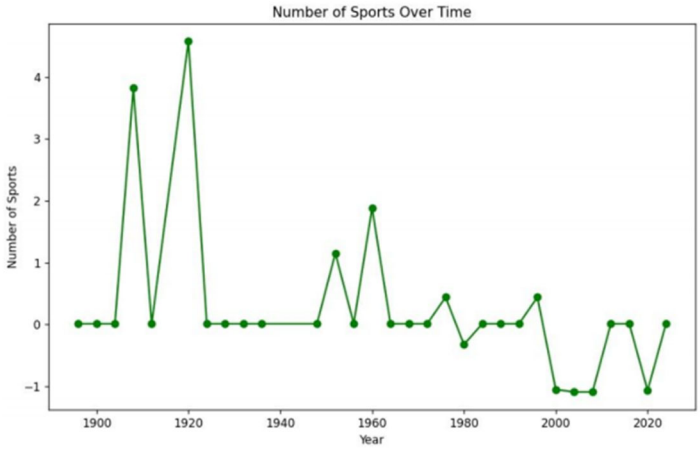

Fig. 4 Changes in the number of sports events

This chart illustrates the variable trends in sports participation numbers across different periods. There were more significant shifts during the initial and mid-phases, followed by a more consistent pattern in later stages, albeit with minor variations.

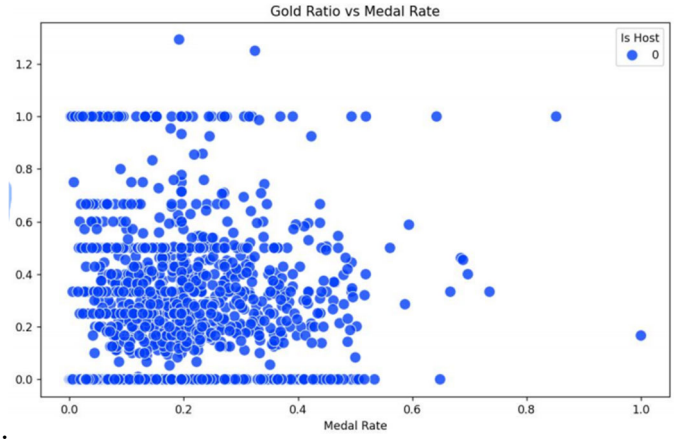

Fig. 5 Gold medal share and athlete winning rate

We found that most data points are concentrated in regions with lower athlete winning rates and lower gold medal share. Overall, there is no obvious linear relationship between gold medal share and athlete winning rates. Countries or regions with high athlete winning rates do not necessarily have higher gold medal shares.

3. Predicting Medal Counts for the 2028 Olympics

3.1 Results and Analysis

Based on the current model and selected features, the predicted results for the 2028 Olympic Games are as follows:(1) Gold Medal Prediction: Using historical data and the trained model, the projected number of gold medals for the 2028 Olympic Games is approximately 255. (2) Total Medal Prediction: According to the model's forecast, the total number of medals in the 2028 Olympic Games is estimated to be around 714.

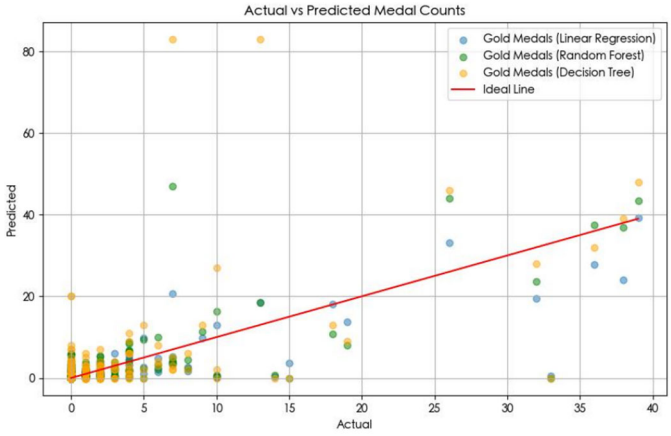

Fig. 6 Comparison of Actual vs Predicted Olympic Medal Counts

(1) Linear Regression:

The blue dots (linear regression predicted values) show relatively small discrepancies compared to actual values. However, in the higher gold medal count range (e.g., 30 and above), the predicted values tend to be significantly lower than the actual values.

(2) Random Forest Regression:

The green dots (random forest predicted values) are more concentrated, indicating that the model's predictions are relatively consistent. However, in the higher gold medal count range, the predicted

values are overestimated. Compared to linear regression, random forest regression handles nonlinear relationships better but still has some prediction errors.

The red ideal line represents a perfect prediction scenario where the actual and predicted values match exactly. Data points falling on this line indicate perfect model accuracy.

3.2 Residual Distribution Plot and Error Analysis

The residual distribution plot shows the distribution of differences between predicted and actual values. If the residuals are randomly distributed, it indicates that the model has no systematic errors and meets the fundamental assumptions of the linear regression model.

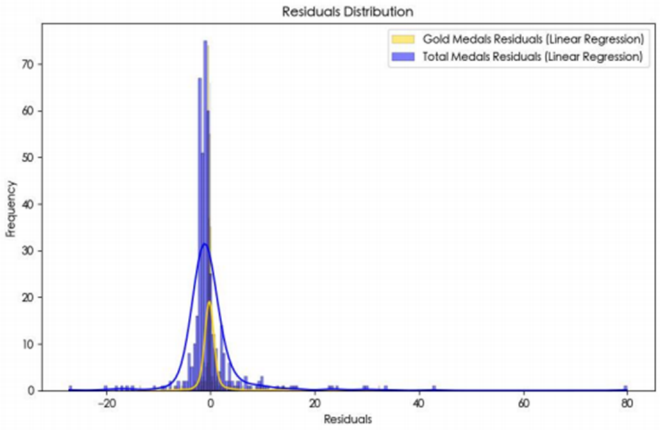

Fig. 7 Residuals Distribution Analysis for Olympic Medal Prediction Key

Observations:

(1) Residual Distribution: Ideally, residuals should be randomly distributed around zero and follow a normal distribution. The two curves in the graph represent the distribution of residuals for gold and total medal counts. Deviations from normality may indicate that the model has failed to capture some important patterns in the data.

(2) Differences in Yellow and Blue Distributions: The differences in the shapes of the gold medal residual distribution (yellow) and total medal residual distribution (blue) suggest that the prediction accuracy varies for different targets. The model may fit one target variable better than the other.

(3) Based on the current model and selected features, the predicted results for the 2028 Olympic Games are as follows: Gold Medal Prediction: Using historical data and the trained model, the projected number of gold medals for the 2028 Olympic Games is approximately 255. Total Medal Prediction: According to the model 's forecast, the total number of medals in the 2028 Olympic Games is estimated to be around 714.

3.3 Results and Analysis

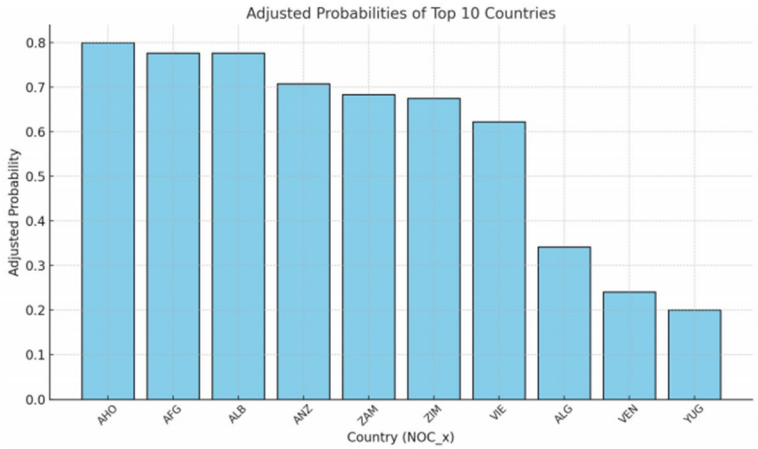

Fig.8 Top 10 Countries with Highest Predicted Medal-Winning Probabilities

Using a logistic regression model, the winning probabilities for all countries that have not yet won a medal are predicted, and a histogram is plotted to visualize the probability distribution. The horizontal axis represents the predicted winning probability intervals (e.g., 0-0. 1, 0. 1-0.2), while the vertical axis represents the number of countries within each interval. The histogram provides an overview of the overall distribution of winning probabilities across countries, helping to identify those with higher chances of winning.

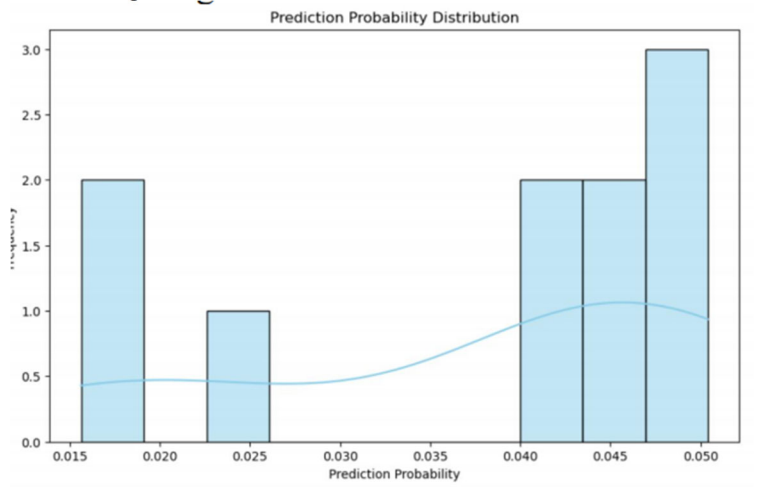

Fig.9 Predicted Medal-Winning Probability Distribution

Table 1 Predicted Medal Counts and Confidence Intervals for Potential First-Time Winning Countries in 2028 Olympics

NOC	Predicted Gold Medals for 2028	Predicted Total Medals for 2028	Gold Prediction Interval	Total Prediction Interval
AHO	8	30	3 47E-18	6 94E-18
AFG	6	22	1.39E-17	6.55E- 17
ALB	7	25	3.47E-18	6.94E-18
ANZ	10	35	3.47E-18	6.94E-18
ZAM	5	20	3.47E-18	6.94E-18
ZIM	6	23	3.47E-18	6.94E-18
VIE	7	28	6.94E-18	2.78E-17
ALG	6	25	3.47E-18	6.94E-18
VEN	5	22	3.47E-18	6.94E-18
YUG	7	27	3.47E-18	1.39E-17

Number of Athletes: In the recent Olympic Games, the number of participating athletes from these countries has been gradually increasing. For example, Vietnam and Zimbabwe have sent more athletes to individual events.

Number of Participating Events: Countries such as Australia and the historical team of New Zealand have participated in a wide range of events. Zimbabwe has shown strong participation in athletics and swimming events in recent years.

Economic Conditions: Some countries, such as Afghanistan and Zambia, have weaker economic conditions. However, their regional strengths in sports such as taekwondo and long-distance running provide opportunities for winning medals.

Figure 10 uses a box plot to show the distribution of the number of Olympic events participated in by host and non-host countries, highlighting the differences in participation between them. It can be observed that host countries tend to participate in a higher number of sports or show greater variability in the number of events selected.

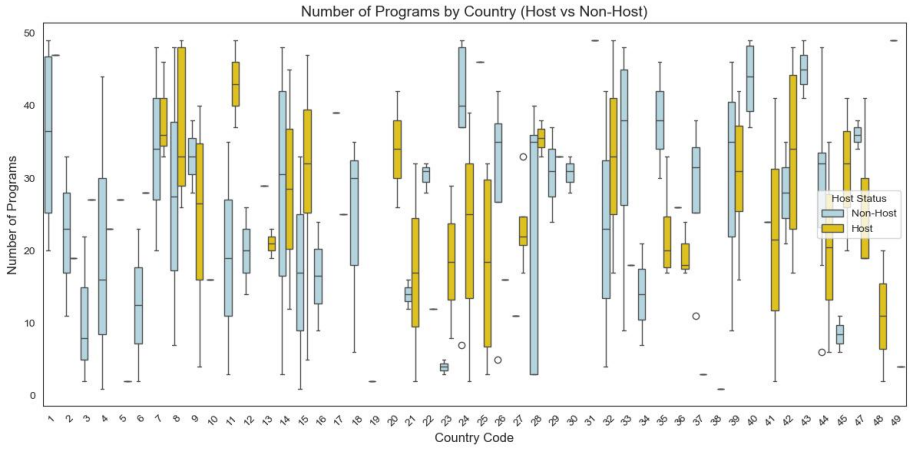

Fig. 10 Building an Interpretable Predictive Model Based on AdaBoost

AdaBoost (Adaptive Boosting) is a popular ensemble learning method that aims to enhance the predictive power of a model by combining multiple weak learners (usually decision trees). Unlike other boosting methods, AdaBoost emphasizes hard-to-predict samples by adjusting their weights in each iteration, gradually improving the model in each round of training. In this task, the AdaBoost-based model will be used to predict total medal counts and leverage its interpretability to analyze how various features impact the final prediction results.

Grid search is an exhaustive hyperparameter tuning method that searches through a specified grid of hyperparameter values. It evaluates every combination of parameters, training the model for each combination. By comparing performance metrics such as accuracy or mean squared error, it identifies the best hyperparameter set for optimal model performance. Hyperparameter tuning is a crucial step in improving the performance of machine learning models. It involves adjusting the hyperparameters of the model to find the optimal set that yields the best performance. The process typically follows these steps: Define the Model and Hyperparameters: Start by selecting the model (e.g., Gradient Boosting, Random Forest) and identifying the key hyperparameters to tune. For instance, for a Gradient Boosting model, important hyperparameters include n_estimators, learning_rate, max_depth, subsample, and min_samples_split.

Table 2 Hyperparameter tuning parameter meaning

Parameters	Best hyperparameters for gold medals prediction	Best hyperparameters for total medals prediction
_n estimators	1	0.9
learning_rate	100	100
max_depth	2	5
min_samples_split	1	2
min_samples_leaf	5	3
subsample	0.05	0.1

After hyperparameter tuning, the MSE for gold medal prediction is 0.05, and the MAE is 0.09. For total medal prediction, the MSE is 0.12, and the MAE is 0.25. To better understand the impact of coaches, we conducted an analysis using the optimized random forest model. The results are shown in Figures 23 and 24. Key observations are as follows: Coach Performance as a Key Feature: In the medal prediction model, coach performance is a crucial feature. The analysis with the optimized random forest model reveals that coaches with better performance significantly improve the athletes' medal count, particularly for gold medals and total medals. Therefore, in practical applications, coach performance should be considered as a more critical feature when optimizing the model. Coach Experience vs. Performance: While coach experience might theoretically affect athletes' performance, the analysis with the optimized random forest model indicates that the actual performance of the coach (rather than the years of experience) has a greater impact on the athletes' medal count. The impact of coaches on total medal counts is shown in Figure 25. After improving coach performance, the prediction of a country's total medals increased. Some countries saw a significant increase in predicted medals, indicating that improving coach performance can lead to more medals for these countries. It highlights the importance of "great coaches," particularly in improving the performance of certain countries. For other countries, the impact of enhanced coach performance on total medals is minor, possibly because these countries already have strong athletes or other factors are driving their medal counts. After hyperparameter optimization, the Mean Squared Error (MSE) for gold medal prediction is 0.05, and the Mean Absolute Error (MAE) is 0.09; for total medal prediction, the MSE is 0.12, and the MAE is 0.25. To better understand the impact of coaches, we analyzed the optimized random forest model, with results shown in Figure 11. The analysis indicates that in the medal prediction model, coach performance is a crucial factor. The optimized model shows that coaches with better performance significantly increase athletes' medal counts, particularly for gold and total medals. Therefore, in practical applications, coach performance should be considered a key feature for model optimization. While coach experience may theoretically influence athletes' performance, the model analysis reveals that the actual performance of coaches (rather than their years of experience) has a more substantial impact on athletes' medal counts.

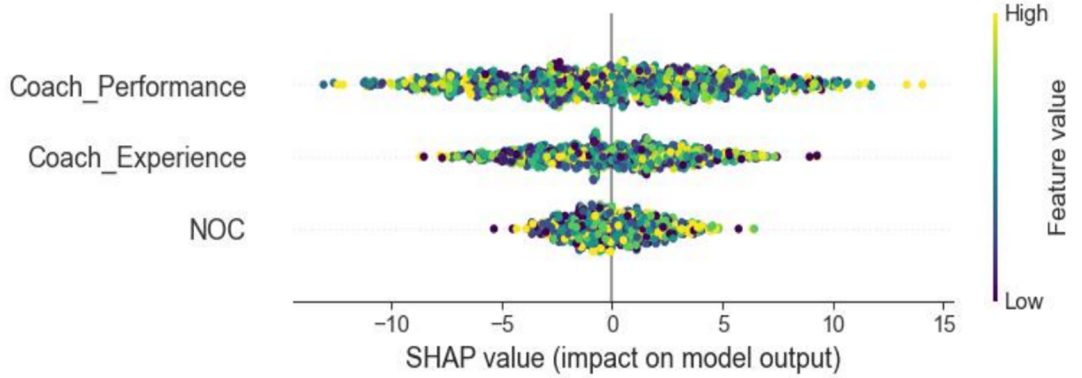

Fig. 11 Building an Interpretable Predictive Model Based on AdaBoost

Figure 12 illustrates the effect of coaches on total medal counts. After improving coach performance, the predicted total medal count for countries increased. Some countries saw a significant rise in their predicted medal counts after enhancing coach performance, indicating that improving the coach's performance can lead to more medals for these countries. It emphasizes the importance of a "great coach," especially in improving performance in certain countries. However, for other countries, the impact of enhancing coach performance on total medal counts is minor, possibly because these countries already have strong athletes or other factors that dominate their medal counts.

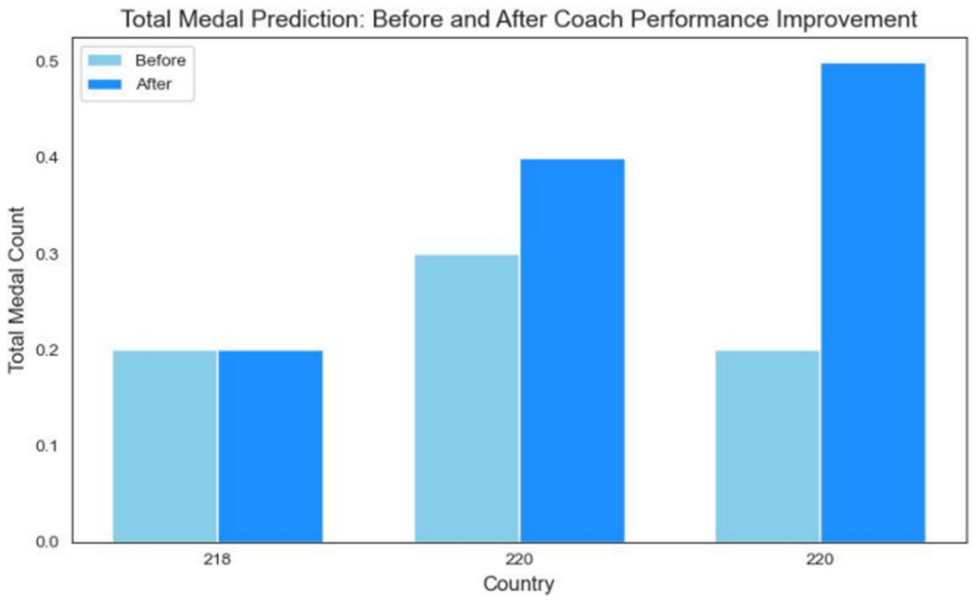

Fig. 12 Building an Interpretable Predictive Model Based on AdaBoost

4. Model Establishment and Solution

Throughout Olympic history, many countries have yet to win a medal. Due to various factors such as a lack of athletes, insufficient training facilities, and economic constraints, these nations have been unable to achieve a breakthrough in past Olympic Games. However, with the increasing diversity of Olympic events and the inclusion of emerging sports, the likelihood of these countries winning their first medals in future Olympics is gradually rising.

The objective of this study is to predict the probability of these medal-less countries winning their first medals at the 2028 Los Angeles Olympics. To achieve this, a logistic regression model can be utilized to analyze the participation data and event involvement of each country, combined with historical data, to build an effective predictive model.

Model Selection and Feature Engineering:

For classification prediction, we have chosen the logistic regression model. Logistic regression is a commonly used binary classification model that predicts the probability of an event occurring. In this study, our goal is to predict whether a country will win a medal, particularly if it will achieve its first Olympic medal in the 2028 Games. The mathematical expression of the logistic regression model is as follows:

$$P(\text{medal}) = \frac{-b \pm \sqrt{b^2 - 4ac}}{1 + e^{-\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}} \tag{5}$$

In this model:

• P represents the probability of a country winning a medal, specifically its likelihood of securing a medal at the 2028 Olympics. Since it is a probability value, must always range between 0 and 1.

• $X_1, X_2, \cdots, X_n$ represents the feature variables that influence whether a country can win a medal. These features may include the number of athletes, the number of events participated in, historical performance, economic level and other factors that collectively determine the probability of winning a medal.

• $\beta_0, \beta_1, \cdots, \beta_n$ represents the regression coefficients, which indicate the impact of each feature on the likelihood of winning a medal. These coefficients are estimated through model training.

The objective of this logistic regression model is to learn the regression coefficients $\beta_0, \beta_1, \cdots, \beta_n$ from the given training data. The learning process is achieved by maximizing the likelihood function, which aims to make the predicted probabilities as close as possible to the actual observed labels.

First, it is essential to identify and filter countries that have not yet won any medals, extracting key features that may influence their likelihood of winning. These key features include:

•**Number of Athletes**: The number of participating athletes reflects a country's scale of participation.

•**Number of Events Participated In**: A higher diversity of participation in events may increase the chances of winning medals.

•**Historical Athlete Performance**: Although a country has not yet won, strong performances by individual athletes or near-win records may serve as important indicators.

•**Economic Level**: Economic conditions directly impact a nation's sports infrastructure, training resources, and athlete development.

•**Sports Infrastructure History**: A country's sports tradition and infrastructure development significantly influence competition outcomes.

Based on these features, the logistic regression model will be used to establish a binary classification model, where the target variable is "whether a country wins a medal or not." This model will provide valuable insights for sports organizations and policymakers in assessing the likelihood of a country's first Olympic medal in 2028.

5. Conclusion

In this study, we constructed multivariate linear regression and random forest regression models to predict the total number of gold medals for the 2028 Los Angeles Olympic Games to be about 255 and the total number of medals to be about 714. The random forest model performed better in capturing the nonlinear relationship, but it should be noted that it tended to overestimate the prediction of high medal count intervals. Hyperparameter tuning resulted in a significant reduction in model error (MSE = 0.05 for gold medal predictions and MSE = 0.12 for total medals). Key predictors include historical medal averages (last three years of data), host country advantage, number of athletes, and actual coaching performance, with coaching performance having a powerful effect on the number of medals won. Countries likely to win for the first time (e.g., Vietnam, Zimbabwe) have potential in regionally dominant sports such as track and field and swimming, but need to develop a targeted strategy that takes into account their economic conditions. It is recommended that NOCs differentiate the allocation of resources, prioritize the support of high-potential countries and volatile events, and optimize data collection (e.g., strengthen coaching performance indicators) and strengthen the corrective analysis of abnormal years (e.g., epidemic period) in order to improve the model's predictive accuracy and decision-making support capability.

References

[1] Toohey K. The Olympic Games: A social science perspective[M]. Cabi, 2007.

[2] Gould D, Maynard I. Psychological preparation for the Olympic Games[J]. Journal of sports sciences, 2009, 27(13): 1393- 1408.

[3] Essex S, Chalkley B. Olympic Games: catalyst of urban change[J]. Leisure studies, 1998, 17(3):187-206.

[4] Scandizzo P L, Pierleoni M R. Assessing the olympic games: The economic impact and beyond[J]. Journal of economic surveys, 2018, 32(3): 649-682.

[5] Malfas M, Theodoraki E, Houlihan B. Impacts of the Olympic Games as mega-events[C]//Proceedings of the Institution of Civil Engineers-Municipal Engineer. Thomas Telford Ltd, 2004, 157(3): 209-220.

[6] Espy R. The politics of the Olympic Games: with an epilogue, 1976- 1980[M]. Univ of California Press, 1981.

[7] Kanin D B. A political history of the Olympic Games[M]. Routledge, 2019.