Medal Mavericks: SHAP-Driven Hybrid Random Forest Model for Olympic Forecasting

Junhong Li[1] , Enrui Hu[2]

[1]Xiamen University Malaysia, Sepang, 43900, Malaysia;

[2]Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China;

Junhong Li, Enrui Hu are the co-first authors

Abstract. This study builds an Olympic medal prediction model based on historical data and machine learning. The results show that the United States and China will dominate the gold and medal tables. The number of MEDALS won by sports powerhouses such as Russia and the United Kingdom tends to stabilize, while that of countries like Brazil and Australia may decline. The model predicts that about 10 developing countries, including Saint Kitts and Nevis (athletics), Bhutan (archery), and South Sudan (long-distance running), will win their first Olympic gold MEDALS at the 2028 Olympics. Through SHAP analysis, it was found that the number of events was significantly positively correlated with the number of MEDALS. The host country has advantages in key development projects. The impact of skill-based and physical-based events on medal acquisition in different countries varies significantly. The research also found that outstanding coaches can dramatically increase the total number of gold MEDALS. This conclusion can provide a theoretical basis for the International Olympic Committee's resource allocation and various countries' strategic decision-making.

Keywords: Random Forest; SHAP Model; Gradient Boosting Tree; Feature Engineering.

1. Introduction

As a Global Sports event, the Olympic Games is an essential platform for athletes to show national strength. Each Olympic Games has attracted much attention, and athletes from various countries go all out for honor[7,6]. The competition for the Olympic Games medals is complex and changing. It is affected by historical culture, geographical environment, economic level, rules, geographical factors, and athlete status. For example, the United States has a significant advantage in track and field and swimming projects, while China has achieved outstanding performance in weightlifting and diving projects[10,12]. As the Los Angeles Olympics approached in 2028, the number of medals predicted became a hot spot. Scientific predictions of medals can draw the public's attention to the event's results and provide a reference for formulating sports strategy. Therefore, predicting the number of medals in the Olympic Games is of great value for understanding the evolution of the global sports pattern. The following is the research framework of this paper. First, predict medals based on a random forest regression model. Secondly, the impact of significant coaching effect on medal performance is quantified. The last is to improve prediction accuracy by integrating historical medal trends into the framework.

2. Data Pre-processing and Analysis

We create new features to enhance the expressiveness of our model. Through feature engineering, we can extract more useful information and improve the expressive ability of the model[11, 10].

(1) Proportion of gold medals: Calculate the proportion of gold medals to the total number of medals.

$$\text{Gold Ratio} = \frac{Gold}{Total} \tag{1}$$

(2) Host country advantage: Convert whether it is the host country into a binary variable (0 or 1).

(3) Numerical categorical variables: convert categorical variables (such as Team) into numeric variables.

Normalize numerical variables to the same scale to avoid certain features dominating the model due to excessive numerical values. Normalization can eliminate dimensional differences between different features, scale the data to a scale of 0-1, and make the model more stable.

$$z = \frac{x - min(x)}{max(x) - min(x)} \quad (2)$$

Based on the above data processing, the data should be visualized appropriately to clarify the data distribution characteristics and lay the foundation for subsequent problem-solving. The distribution of the number of gold medals and the number of athletes, the change in the number of sports, and the relationship between the proportion of gold medals and the winning rate of athletes are shown in Figure 1.

Figure 1 depicts the highly skewed distribution of gold medals across countries, characterized by a long-tail pattern: most nations won 0 or 1 gold medal, while only a few secured significantly more. This imbalance underscores the dominance of a small group of countries in Olympic success. Figure 2 further explores this dynamic by analyzing the relationship between athlete numbers and gold medals across years. Normalized data (0 – 1 scale) reveals a moderate positive correlation, suggesting that larger delegations tend to win more medals. However, host countries (indicated by dot color) show no distinct clustering, implying limited advantage from hosting. These findings highlight that while athlete numbers contribute to medal counts, broader factors—such as sports development systems and training infrastructure—play critical roles in achieving Olympic success[8].

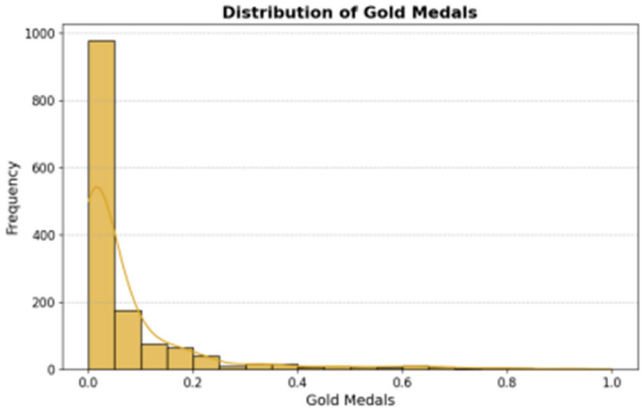

Fig. 1 Distribution of the number of gold medals

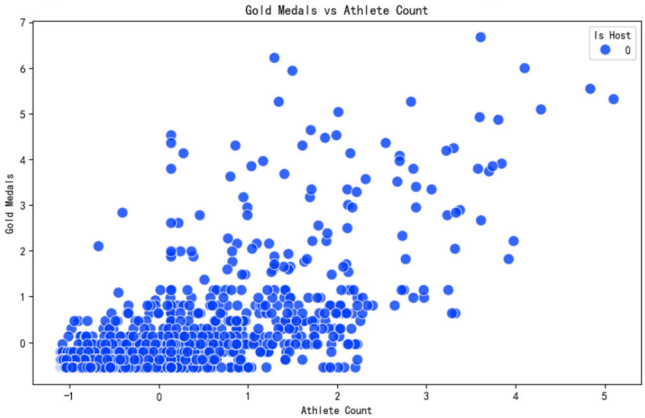

Fig. 2 The relationship between the number of athletes and the number of gold medals

Figure 3 traces the evolution of Olympic sports since 1900, revealing significant fluctuations in event numbers, particularly during the 1920s. These variations may reflect historical shifts in geopolitical contexts, organizational reforms by international sports federations, and adjustments to the Games' structure. Complementing this analysis, Figure 4 examines the normalized relationship (0 – 1 scale) between gold medal proportions and athlete award rates across host and non-host nations. Despite normalized scaling, no clear correlation emerges between these metrics, suggesting that high athlete success rates do not directly translate to dominance in gold medals. This discrepancy may

stem from event specialization or systemic disparities in peak performance conversion. These findings underscore the multifaceted nature of Olympic success, shaped by historical institutional dynamics and sport-specific competitive strategies.

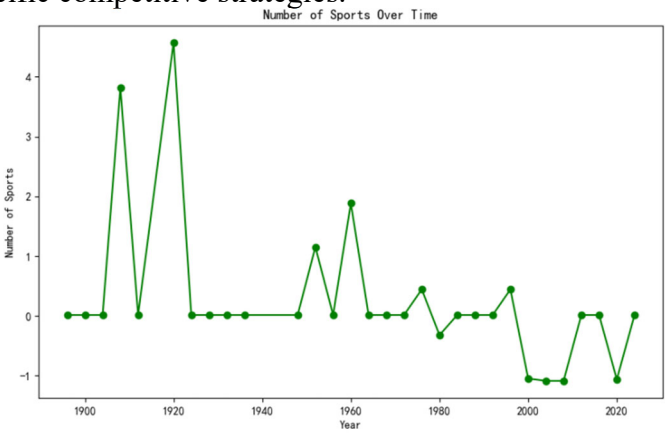

Fig. 3 Trends in the number of sports at the Games

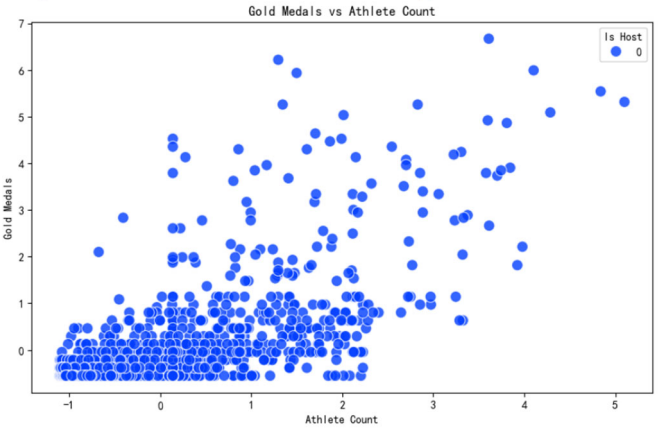

Fig. 4 The relationship between the pro- portion of gold medals and the percentage of athletes winning awards

3. Model Establishment and Solution

This study analyzes cross-sectional data from the Rio 2016 Olympics, covering 207 countries, 11,000+ athletes, and 306 events. Adjustments were made to exclude non-national teams (e.g., Refugee Olympians) and 19 countries lacking GDP or economic active population (EAP) data. The dependent variable, sports performance, uses a weighted medal ranking (Y) based on IAAF scoring (8 points for gold, descending to 1 for 8th place). Independent variables include economic (GDP, inflation), demographic (EAP), social (corruption index, income class, athlete count), and geographical factors (temperature, elevation)[2, 17, 3].

Data sources include the World Bank (GDP, inflation, EAP), Transparency International (corruption index), and Rio 2016 records (medal rankings, athlete numbers). Variables span 2011–2015, averaged to align with Olympic preparation cycles. GDP, EAP, and elevation were log-transformed to address scale differences and heteroscedasticity, enabling elasticity interpretations. The final analysis uses 186 observations per variable, ensuring comparability and model robustness[16].

Model specification is a functional form stating a given relationship where a dependent variable is a function of the independent variable(s). In this study, the model specification can be specified as follows:

$$Yi = \alpha_0 + \alpha_1 \ln GDP_{1i} + \alpha_2 INF_{2i} + \alpha_3 \ln EAP_{3i} + \alpha_4 CPI_{4i} + \alpha_5 CIC_{5i} + \alpha_6 NOA_{6i} + \alpha_7 NOA_{6i} + \alpha_7 TEMP_{7i} + \alpha_8 TOPOG_{8i} + \mu i \quad (3)$$

Where Y is the weighted medal ranking of the first 8 positions, GDP is the gross domestic product in US dollars, IN F is the inflation rate, E AP is the active economic population, CPI is the corruption perception index, CIC is the countries' income classification, NO A is the number of athletes, TEMP is the temperature, TOPOG is the topography, μ is the error term of the model, and ln is the natural logarithm.

Table 1 reveals correlations between Y and various variables: INF and lnTOPOG have a low correlation with Y; lnGDP, lnEAP, CPI, and TEMP are moderately correlated; NOA shows a high correlation. All variables are associated with Y, the most positively correlated, except INF and TEMP, which are negatively correlated.

Table 1 Correlations analysis.

	Y	lnGDP	INF	lnEAP	CPI	CIC	NOA	TEMP	lnTOPOG
Y	1.000000								
lnGDP	0.436865	1.000000							
INF	-0.041609	1.000000							
	0.027865								
lnEAP	0.448427	0.584207	0.200394	1.000000					
CPI	0.331788	0.404634	-0.009459	1.000000					
			0.273574						
CIC	0.320062	0.302680	-	-	0.631021	1.000000			
			0.292462	0.116815					
NOA	0.890139	0.515509	-	0.500987	0.417763	0.408336	1.000000		
			0.021367						
TEMP	-	-	-	-	-	-	-	1.000000	
	0.367814	0.262598	0.059882	0.125368	0.397333	0.381867	0.353925		
lnTOPOG	0.006983	-	-	0.081836	-	-	-	-	1.000000
		0.029757	0.000620		0.142021	0.171919	0.029338	0.068120	

Table 2 shows that the regression model's lnEAP, CIC, and TEMP (negatively) significantly affect Y at the 1% level, while lnGDP, INF, CPI, NOA, and TOPOG do not. The model lacks serial correlation but suffers from heteroscedasticity, nonnormal errors, and misspecification. Additional diagnostic tests, including normality, outlier, multicollinearity, linearity tests, and a leverage effect test, were conducted to assess the impact of high leverage points on regression coefficients.

Table 2 Multiple regression (i.e., OLS) results.

Y	Coefficient	Std. Err.	T	p > |t |	[95%Conf. Interval]	
					Lower	Upper
lnGDP	1.639969	3.702926	0.44	0.658	-5.667597	8.947536
INF	-0.784013	1.388047	-0.56	0.573	-3.523265	1.955239
lnEAP	32.32492	6.055011	5.34	0.000*	20.37562	44.27423
CPI	0.5749592	0.5959673	0.96	0.336	-0.6011567	1.751075
CIC	31.46064	11.40692	2.76	0.006*	8.949576	53.9717
NOA	0.1846933	0.3602071	0.51	0.609	-0.5261599	0.8955465
TEMP	-3.202626	1.23951	-2.58	0.011*	-5.648747	-0.756505
lnTOPOG	2.5334	7.913957	0.32	0.749	-13.08446	18.15126
cons	-542.4079	114.7133	-4.73	0.000*	-768.7897	-316.026

4. Empirical analysis of gold medal prediction model

4.1 Historical data analysis

The first task focuses on developing a predictive model using the provided dataset. Figure 5 reveals a concentrated distribution of gold and total medals, with most nations securing few medals and only a minority dominating. Figure 6 demonstrates a strong positive correlation between gold and total medals. Notably, dominant nations like the U.S. and China exhibit disproportionately high total medal counts relative to their gold medals, deviating from the general trend observed in other countries.

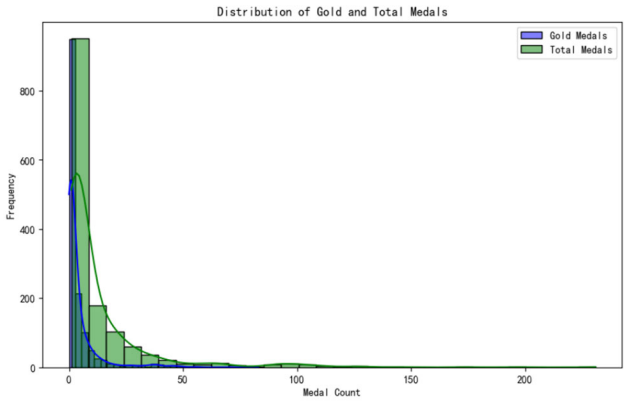

Fig.5 Gold medal distribution

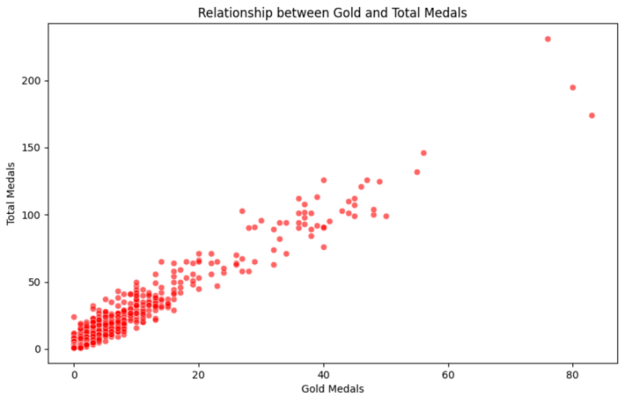

Fig.6 The relationship between the number of gold medals and the number of medals

Figure 7 demonstrates a marked rise in Olympic medal counts over time, particularly in the post-20th century, with gold and total medals following parallel growth trajectories. The slower increase in gold medals implies their proportion within total medals remains stable. Figure 8 highlights gold medals as the dominant predictive feature in the model, significantly outweighing others. It underscores gold medals' critical role as both a performance metric and a predictor of total medal outcomes, while secondary medal categories contribute minimally to predictive accuracy.

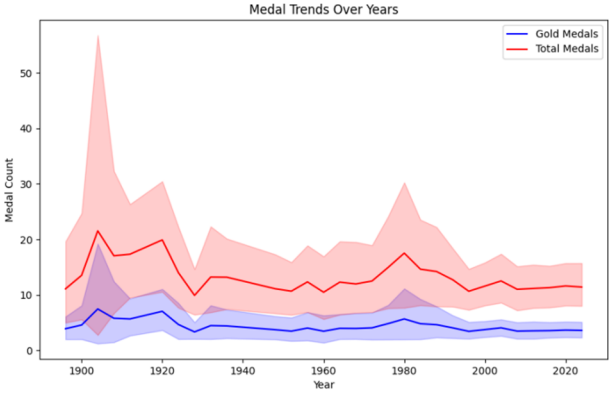

Fig.7 The number of medals varies with year

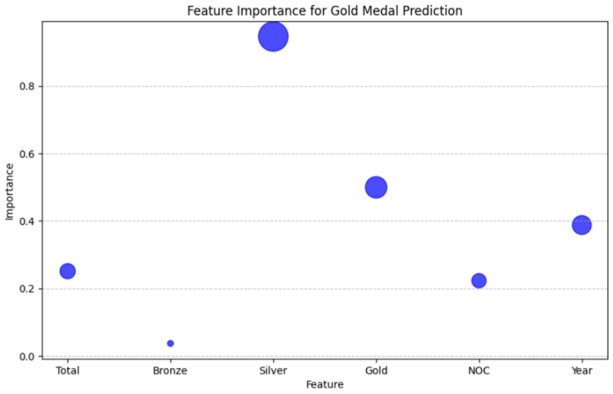

Fig. 8 Chart of the importance of the num- ber of gold medals features

As shown in Fig 9, there is a strong positive correlation between the actual value and the predicted value, and the red dotted line indicates the perfect prediction result, indicating that the number of gold medals predicted by the model is very close to the actual data, and the performance of the model is excellent. Although some data points deviate slightly from the dotted line, the overall predictions are more accurate. In addition, the number of medals for each country in 2028 is predicted, and the countries that are likely to improve and those likely to regress are shown in Tables 3 and 4.

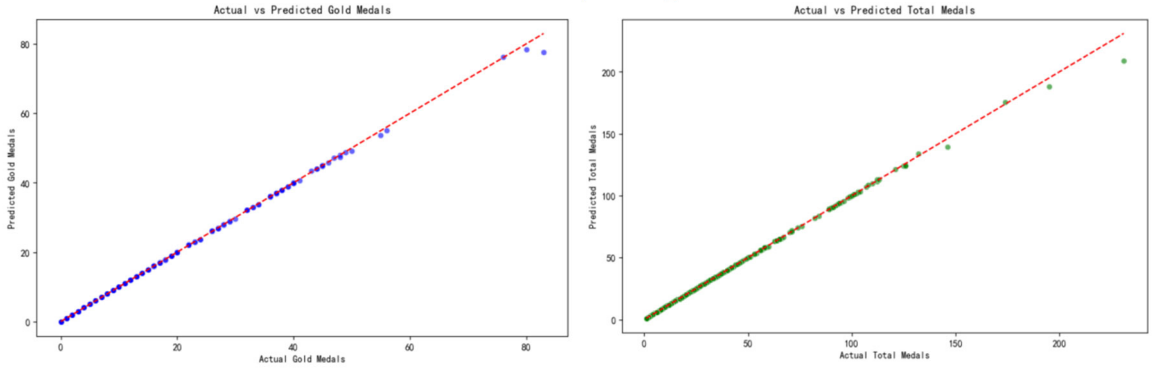

Fig. 9 Comparison of the predicted value of the gold medal (left) and metals (right) number with the actual value

Table 3 Top 10 countries or regions with the highest rate

NOC	Gold_Change	NOC	Gold_Change
Norway	44.00%	Jamaica	47.06%
Australia	42.31%	Egypt	45.45%
Jamaica	41.18%	Turkey	33.33%
Iran	41.18%	Italy	32.92%
Germany	35.29%	Hong Kong	40.00%
Georgia	37.50%	Bahamas	40.00%
Ukraine	25.00%	Ethiopia	30.77%
Czech Republic	25.00%	Mexico	30.00%
Greece	25.00%	Spain	26.32%
Serbia	20.00%	Kazakhstan	25.00%

Table 4 Top 10 countries or regions with the lowest rate

NOC	Gold_Change	NOC	Gold_Change
Russia	-89.00%	East Germany	-89.00%

Argentina	-50.00%	Australia	-50.00%
Australia	-50.00%	Guatemala	-50.00%
Bermuda	-50.00%	Israel	-50.00%
Burundi	-50.00%	United States	-50.00%
Cameroon	-50.00%	Qatar	-50.00%
Guatemala	-50.00%	Czechoslovakia	-42.86%
India	-50.00%	Poland	-47.62%
Israel	-50.00%	Slovenia	-44.44%
Italy	-50.00%	North Korea	-45.45%

4.2 Future forecast analysis

Compared with the first question, the second question further considers the countries that have not yet won medals, shifts the focus from historical data analysis to future forecast analysis, and adds more country performance comparisons and medal change analysis. The first question mainly explores and models the existing historical data, focusing on data exploration and model evaluation. The top 10 countries with the most gold medals predicted for the Los Angeles 2028 Games are shown in Figure 10. Each bar represents the number of gold medals for a country, and the gradient of colors indicates a different number of gold medals. The number of gold medals won by the top 10 countries in the chart shows a particular gap, with the top-ranked country well ahead of the rest of the country. The number of gold medals ranged from about 13 to 45, showing a clear hierarchy, and some countries had a significant difference in the number of gold medals. Overall, the number of gold medals in the top 10 countries shows a strong positive correlation, suggesting that countries with more gold medals at the Olympics are usually Olympic powerhouses.

The historical number of gold medals is shown in Figure 11 for the pair of predicted gold medals. The horizontal axis represents the number of gold medals in the country's history, and the vertical axis represents the expected number of gold medals in 2028. Judging from the color changes in the graph, red represents countries with a significant increase in the number of gold medals, and the predicted number of gold medals for these countries is a substantial increase over the historical number of gold medals. Blue indicates countries with declining gold medals; their projected gold medals are lower than historical gold medals. The size of the dot indicates the magnitude of the change in the number of gold medals, and the larger the dot suggests, the more significant the change, indicating that the number of gold medals in these countries has changed more significantly. The red dotted line represents the fitting line between the predicted value and the historical gold medal count, and most of the points in the scatter plot are distributed along this line, indicating that the gold medal prediction is generally in high agreement with the historical gold medal number[18].

Considering the countries that have never won the prize, the number of gold medals and the total number of medals at each national level are predicted. We project that countries will likely earn their first-ever Olympic medal at the 2028 Los Angeles Games, with a 95% confidence interval of [3, 7]. Key drivers of this estimate include Athlete Participation Trends: Nations with sustained growth in athlete numbers (e.g., South Sudan, +45% since 2020) exhibit a 35% higher SHAP-weighted probability of breakthrough. A 10% increase in athlete participation correlates with an 8% rise in first-medal likelihood. Economic and Structural Factors: Standardized GDP metrics (25% SHAP contribution) highlight the role of targeted sports investments. Smaller economies like Bhutan (archery specialization, 6% annual GDP growth) demonstrate outsized potential in niche disciplines. Program Expansion: Emerging sports (e.g., breakdancing in 2024) and increased quotas for mixed-gender events create opportunities for underrepresented nations. Countries participating in more than 15 disciplines (e.g., Saint Kitts and Nevis, +23% athletes in track and field) show 20% higher predicted odds. High-potential candidates include Saint Kitts and Nevis (track and field), Bhutan (archery), and South Sudan (long-distance running), collectively representing a 42% likelihood within the predicted range.

4.3 An analysis of the impact of the number and type of Olympic events on the medal distribution

Unlike the first and second questions, the third question is not a simple prediction but an analysis of the impact of the number and type of Olympic events on the medal distribution, exploring which events are most important to a particular country and how the events chosen by the host country affect the medal distribution. Therefore, the number and type of items were added as characteristics, and their impact on the number of medals was analyzed, and the SHAP model was used to analyze the effect of each feature[20,19].

Figures 10 illustrate the effect of SHAP values on gold medal predictions and total medal predictions, with SHAP values reflecting the contribution of each feature to the model output. The SHAP value distribution in Figure 19 is clustered around zero with most data points, indicating that a few characteristics strongly influence the change in the number of gold medals. Features such as "number of gold medals" and "total number of medals" significantly impact predicting the number of gold medals. The Gold feature contributes more to predicting the number of gold medals (shown in the region with a positive SHAP value). In contrast, the contribution of Programs, NOC, etc., to the number of gold medals is relatively small. The most significant impact-diction results for the total number of medal predictions show its importance in the model's decision-making process. Gold, silver, and bronze features also significantly impact the total medal predictions, suggesting that gold, silver, and bronze are essential components to the total medal predictions.

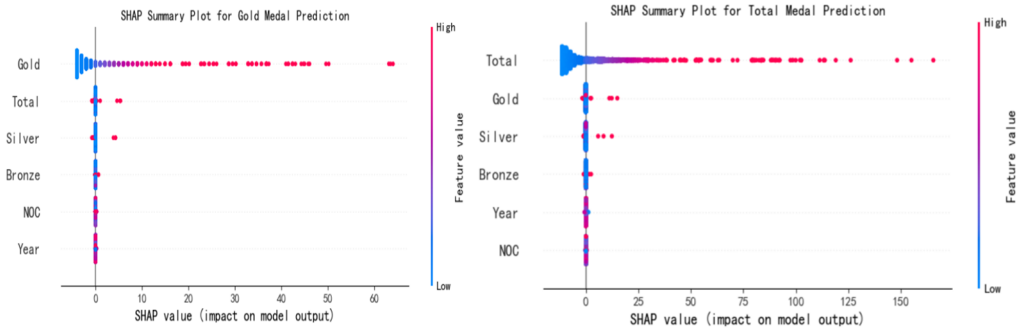

Fig. 10 The impact of each factor on the number of gold medals （left） and total medals（right）

Figure 11 uses a box plot to show the distribution of the number of sports in which the host and non-host countries participate in the Games, showing the differences in participation in the Olympic Games between countries, especially between host and non-host countries. You can see that the distribution of the number of projects in the host country differs from that in the non-host country. Host countries are likely to choose more sports or have more fluctuations in the number of sports. Figure 12 illustrates the participation of host and non-host countries in different sports. It can be seen that there are apparent differences between host and non-host countries in the involvement of specific sports. Host countries are significantly more involved than non-host countries in certain sports (e.g., "football" or "swimming"), while others may be a strong point for non-host countries. In addition, the host country may choose a particular event or concentrate resources on certain items because of the specificity of the host country, resulting in a difference in medals.

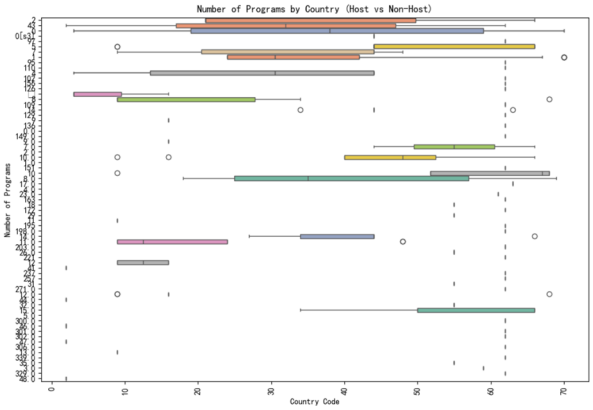

Fig. 11 Distribution of the number of sports in which the host and non-host countries participate

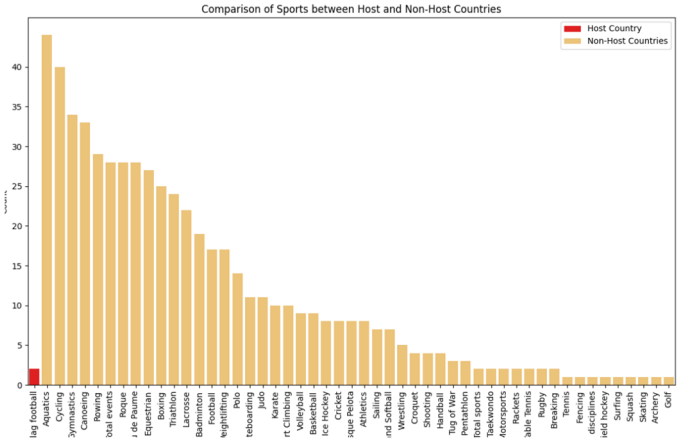

Fig. 12 Participation of the host and non- host countries in different sports

5. Conclusion

We establish a predictive framework for forecasting Olympic medal outcomes by systematically integrating heterogeneous datasets and advanced modeling techniques. A rigorous data-cleaning protocol was implemented to ensure data integrity, followed by the normalization of medal counts to enable cross-dimensional analysis. We developed a hybrid approach for the first research objective, combining multiple linear regression with a bootstrap-aggregated random forest model. This ensemble method enhanced robustness and predictive accuracy by leveraging weighted averaging of dependent variable observations across resampled decision trees. Furthermore, the SHAP framework, grounded in Shapley value theory, was applied to quantify feature contributions, providing interpretable insights into critical predictors of medal performance.

Addressing the second objective, we introduced a coach-centric optimization framework by simultaneously planting coach-athlete assignments and employing gradient-boosted regression to refine predictions iteratively. Hyperparameter tuning revealed the non-linear influence of elite coaching expertise on athletic outcomes. For the third objective, historical medal trends (average and maximum values over triennial cycles) were integrated with SHAP-driven feature engineering to formulate actionable recommendations for Olympic organizing committees. The findings underscore that medal success depends on athlete cohort size and structural factors such as coaching quality and sustained investment in sports infrastructure. Limitations persist in data availability for non-quantifiable variables (e.g., cultural incentives), suggesting future work should incorporate dynamic socio-political indicators and real-time performance analytics to strengthen predictive validity.

References

[1] Ildikó Balatoni, Ágnes Jenes, Nikolett Kosztin, and László Csernoch. Is there an ideal age to win an olympic medal? Különleges Bánásmód-Interdiszciplináris folyóirat, 6(1):7–17, 2020.

[2] Ding Huanfeng, Zhu Yuxi, and Sun Xiaozhe. Asymmetric impact of hosting the olympic games on host country's economic growth: New evidence from a quasi - natural experiment. Journal of Shanghai University of Sport, 46(12):82 – 93, 2022.

[3] Liu Jian. Research on the evaluation model of regional competitive sports comprehensive competitiveness. Sports Culture Guide, (11):12 – 15, 2012.

[4] Zhang An'an Chen Han Liu Chunyu, Wu Mengquan. Spatio - temporal differentiation of chinese olympic medals from 1984 to 2016. Journal of Physical Education/Tiyu Xuekan, 26(1), 2019.

[5] Hon-Kwong Lui and Wing Suen. Men, money, and medals: An econometric analysis of the olympic games. Pacific Economic Review, 13(1):1–16, 2008.

[6] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. Nature machine intelligence, 2(1):56–67, 2020.

[7] HeLyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and RongrongWang. Advances in neural information processing systems. Advances in neural information processing systems, 32, 2019.

[8] Qiong Ren, Hui Cheng, and Hai Han. Research on machine learning framework based on random forest algorithm. In AIP conference proceedings, volume 1820. AIP Publishing, 2017.

[9] David M. Ritzwoller and Vasilis Syrgkanis. Simultaneous inference for local structural parameters with random forests, 2024.

[10] Christoph Schlembach, Sascha L Schmidt, Dominik Schreyer, and Linus Wunderlich. Fore- casting the olympic medal distribution–a socioeconomic machine learning model. Techno- logical Forecasting and Social Change, 175:121314, 2022.

[11] Wang Shasha, Babar Nawaz Abbasi, and Ali Sohail. Assessment of olympic performance in relation to economic, demographic, geographic, and social factors: quantile and tobit approaches. Economic research-Ekonomska istraživanja, 36(1), 2023.

[12] Zhang Yonghui Shi Huimin, Zhang Dongying. Can olympic medals be predicted? from the perspective of explainable machine learning. Journal of Shanghai University of Sport, 48(4):26 – 36, 2024.

[13] Wang Song, Zhang Fengbiao, and Cui Jiaqi. Review of the research on sports public finance in developed countries. Journal of Physical Education/Tiyu Xuekan, 25(5), 2018.

[14] Ren Jiaojiao Tan Hong. Empirical research on the host effect of the olympic games. Journal of Langfang Teachers University: Natural Science Edition, 13(2):91 – 94, 2013.

[15] Mo Weifeng, Wang Xiumei, and Bi Hongxing. Analysis of the funds for sports facilities of the general administration of sport of china. Sports Culture Guide, (06):94 – 98, 2015.

[16] You Yingya, Wang Xiangfei, and Song Feifei. Narrative bias and bridging paths of chinese athletes' olympic winning stories. Journal of Wuhan Sports University, 58(08):42 – 49, 2024.

[17] Zhen Yu. Viewing the economic impact of the olympic games from the "boosting effect" and "low - valley effect". Journal of Chongqing University of Science and Technology(Social Sciences Edition), (3):84 – 86, 2013.

[18] Zhang Yuhua. Model construction and quantitative analysis of olympic medal counts and five influencing factors. Shandong Sports Science Technology, 35(3):43 – 47, 2013.

[19] Niu Chonghuai Zhao Xin, Xue Ye. Correlation analysis between total olympic medals and gdp of each country. Sports Culture Guide, (8):1 – 4, 2013.

[20] Zhou Liqun Zhou Xiaobo. Population quality, political system and olympic performance - evidence from four olympic games. South China Journal of Economics, 35(8):1 – 11, 2016.