Olympic Medal Prediction Based on Random Forest Modeling

Hao Zhou *, Zehao Li, Zhikuan Wang

School of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhuang, China, 050043

* Corresponding Author Email: m18832123161@163.com

Abstract. The Olympic Games are the most influential sporting event in the world, conveying the idea of higher, faster, and stronger. It promotes the pursuit of excellence in sportsmanship and the dedication to sportsmanship. At the same time, the results of the Olympic Games also reflect the level of sports of each country, so the number of medals is the focus of attention, and the factors that affect the number of medals predicted are very complex, usually based on the historical data of each country, the strength of the players, etc. This paper proposes a random forest prediction model based on multiple linear regression and regression to predict the interval of each country's medal count at the Los Angeles 2028 Olympic Games, as well as the ranking of each country. At the same time, this study provides reference opinions for the prediction of the medal table of the next Olympic Games.

Keywords: Random Forest, Interval Estimation, Hypothesis Testing, Factor Analysis.

1. Introduction

The Olympic Games, the world's largest comprehensive games organized by the International Olympic Committee and held every four years, are the most influential sports event in the world. The Olympic Games have a far-reaching impact and contribution. It is not only a sports event, but also an important platform for promoting international exchanges, safeguarding world peace, promoting economic development and realizing personal and social values.

Regarding the prediction of the medal table, scholars in China have proposed different methods, Zheng Yuxin et al [1] proposed a prediction model based on random forest regression; Shi Huimin et al [2] predicted based on the interpretable machine learning perspective; Li Zhen et al [3] predicted the number of Chinese medals in the Olympic Games as well as the development of the trend by using the cluster analysis method; Yi Jiandong [4] believed that the 2028 Olympic Games in Los Angeles traditional Olympic major events to join the e-sports sub-items or minor events have a certain possibility; Liu Jun et al [5] based on two-way long and short-term memory neural network of reservoir porosity prediction method for research; Fei Luo [6] for the special period of the Tokyo Olympic Games gold medal list of the new pattern of analysis; Peng Zhaofang et al [7] summed up China's development of competitive gymnastics historical experience and the outlook for the new era.

Compared with other models, Random Forest has become a “universal model” in the field of data science by virtue of its advantages of high accuracy, robustness, ease of use and interpretability [8]; among them, Random Forest naturally mitigates the risk of covariance caused by feature correlation through the double randomization mechanism and the integrated learning feature, which is especially suitable for dealing with moderately correlated data sets [9].

During the 2024 Summer Olympics in Paris, spectators not only follow the events, but also pay close attention to the “medal standings” of each country. Prior to each Olympics, predictions are made about the final medal count. The prediction of the “medal table” takes into account the historical performance of each country, the information of the participants, the information of the organizers, and the events of the competitions. Despite the complexity of the prediction method, the predicted “medal count” is still an important reference in the development of national sports, and a key tool to facilitate business decisions and the development of the sports industry.

In this paper, the collected data is first preprocessed. screen out missing values and outliers, select feature engineering and data visualization, and then use multiple models for prediction, and select the best prediction model by comparing the relevant indicators of different models.

In terms of revealing insights about the Olympics, this paper presents five aspects of the sports talent pool, racial disparities, investment in new programs, average age of competitors, and breadth of programs, and provides feasible recommendations on these aspects.

2. Methods

2.1. Core Mechanisms of Random Forest Prediction Models

Random forest model is an algorithm that integrates multiple uncorrelated decision trees and is suitable for high dimensional datasets [10]. Due to the high dimensionality of the given dataset, it is suitable to use Random Forest for prediction as shown in Figure1.

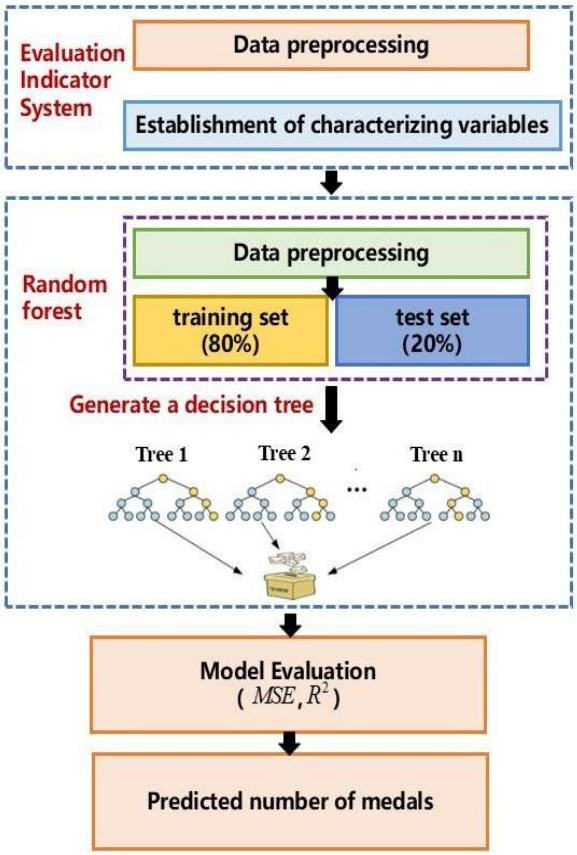

Figure 1. Random forest flowchart

2.2. Core Architecture of Bidirectional Long and Short-Term Memory Networks

$BILSTM$ is an improved recurrent neural network that excels in processing sequence data.$BILSTM$ is able to capture both past and future directions in a sequence by combining the outputs of both forward and backward $LSTM$ networks.

In this paper, when predicting the number of gold medals and the total number of medals in the 2028 Olympic Games in Los Angeles, U.S.A., the $BILSTM$ can take into account the medal performances of previous Olympic Games of various countries, as well as the possible future development trend. Thus, it can provide a more comprehensive and accurate basis for the medal prediction.

2.3. Theoretical Architecture of Multiple Linear Regression Models

Multiple linear regression modeling is a method of fitting regression coefficients based on linear regression by minimizing the sum of squared residuals and thus fitting the regression coefficients [11].

In this paper, the intercept as well as each of the regression coefficients are obtained by fitting the eigenvalues, then the regression equation is:

$$y = -0.32 + 0.01x_1 + 0.01x_2 + 2.22x_3 + 0.01x_4 + 1.24x_5 - 0.02x_6 - 0.64x_7 \quad (1)$$

where $x_1, x_2, \cdots x_7$ are the seven eigenvalues.

2.4. Theoretical Architecture of $LASSO$ Regression

$LASSO$ has L_1 regularization, which adds a regularization term to the objective function of multiple linear regression to constrain the magnitude of the regression coefficients and thus achieve feature selection.

$$F(a) = \sum_{j=1}^{n}\left(y_i - \sum_{k=1}^{7} a_k \cdot x_{jk}\right)^2 + \gamma \sum_{k=1}^{7}|a_k| \quad (2)$$

where y_i is the observation, x_{jk} is the k feature of the j observation, and γ is the L_1 regularization parameter.L_1 regularization compresses some of the regression coefficients a_j to 0, and as γ increases, more regression coefficients will be compressed to 0, resulting in feature selection and sparsification of the model.

3. Results

3.1. Evaluation of the model

This article uses Mean Squared Error (MSE) as the metric, the $RMSE$ metric, and the coefficient of determination R^2, to evaluate the models.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

Specific calculations are shown in Table 1. where y_i is the true value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the true value.

Table 1. Predictive Modeling Results and Errors

Model	MSE	$RMSE$	R^2
Random Forest	2.38	1.54	0.85
BILSTM	2.96	1.72	0.83
Multiple linear regression	5.40	2.32	0.79
Lasso regression	4.55	2.13	0.81

As shown in Table 1, the random forest model has the best fitting effect and the highest accuracy, so this model is chosen to solve the problem.

3.2. Medal table predictions for the 2028 Olympics

(1) Data collation

As there are no data in the dataset about the number of male and female participants in each country and the number of events in each country in 2028. Therefore, this paper will obtain these data on its own and analyze them in combination with the existing data.

(2) Interval estimate

In this paper, the quantities are found to follow a normal distribution by conducting 1000 random forest simulation trials such that the confidence level is $1 - \alpha$. This paper sets the significance level

α to 0.05, which means that the confidence level is 95%. The overall distribution is normal and the variance σ^2 is known, then the interval is:

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}\right] \tag{6}$$

where α is the significance level and n is the sample size.

(3) Medal prediction range

This article not only resolves the prediction interval of the number of gold medals and the number of medals for each country in the 2028 Summer Olympics in Los Angeles, USA, but also the prediction intervals for the number of silver medals and the number of bronze medals for each country.

Therefore, the dependent variables of the model are the number of gold medals, the number of silver medals, the number of bronze medals and the number of medals for each country in the 2028 U.S. Summer Olympics in Los Angeles, as shown in Figure 2. This paper finds that the data variables as a whole are normally distributed.

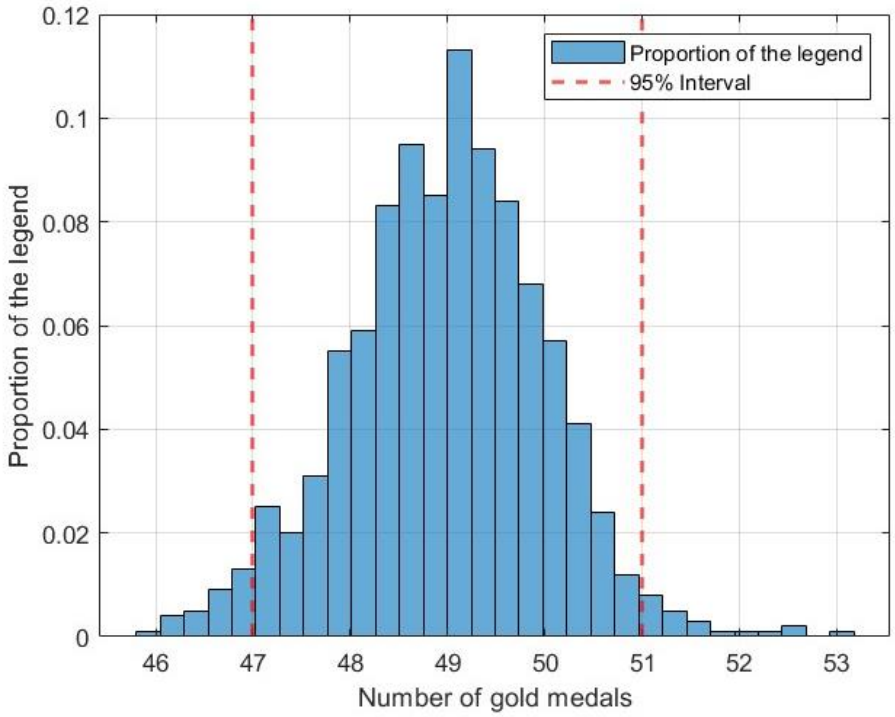

Figure 2. Medal Forecast Interval Estimates

The table below shows the medal prediction ranges for the major participating countries in the 2028 Los Angeles Olympics.

Table 2. Range of medal projections

Country name	Gold medal	Silver medal	Bronze medal	Total
United States	49±2	42±2	32±1	123±4
China	38±2	26±1	26±1	91±4
France	24±1	23±1	16±1	63±3
Germany	20±1	16±1	19±1	55±3
Australia	19±1	17±1	20±1	55±3
Japan	18±1	15±1	16±1	49±3
Great Britain	17±1	17±1	20±1	55±3
Italy	15±1	13±1	21±1	50±3
Netherlands	11±1	9±1	12	32±2
Canada	9	7	14±1	30±1

Based on Table 2, the U.S. ranks first in gold medals with 49±2 and medals with 123±4. China follows with 38±2 gold medals and 91±4 medals.

(4) Projections for improved and worse-performing countries

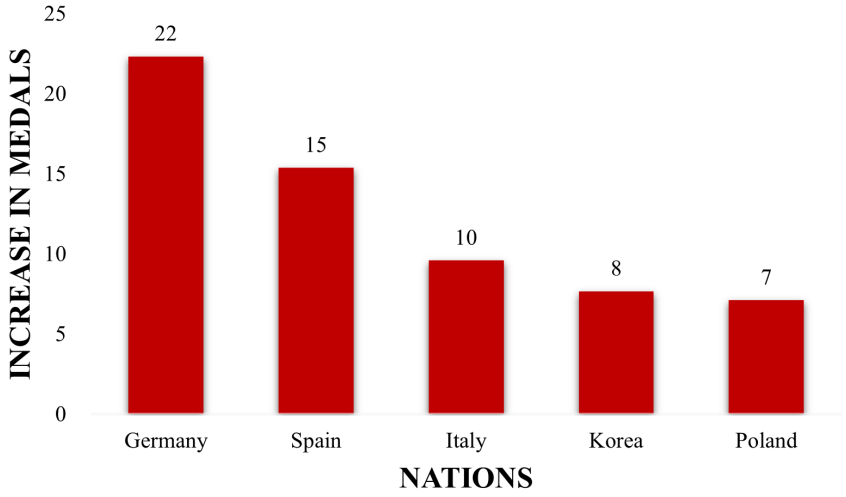

Figure 3. Better performing countries

According to the results of the model solving, the countries most likely to improve are Germany, Spain, Italy, Korea and Poland, with Germany having the most potential, with a chance to increase the number of medals by 22. As shown in Figure 3.

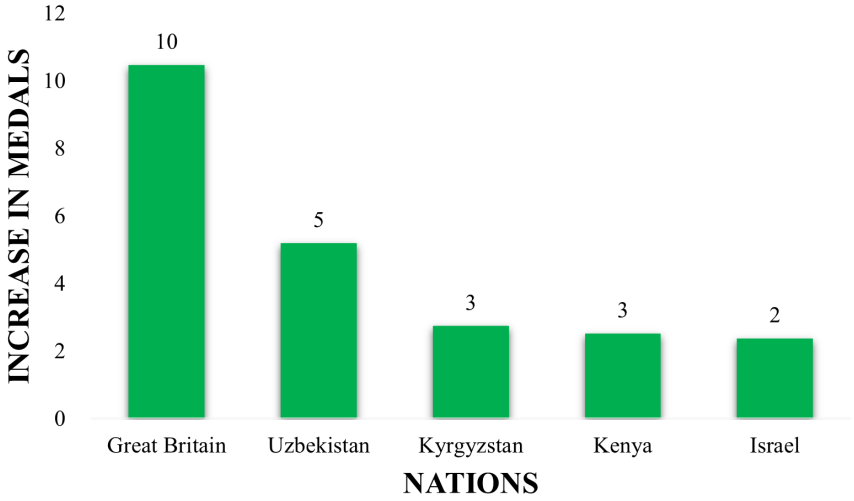

Figure 4. Worse performing countries

According to the results of the model solving, the worse performing countries are Great Britain, Uzbekistan, Kyrgyzstan, Kenya and Israel, with Great Britain likely to perform the worst, with a potential reduction in the number of medals by 10. As shown in Figure 4.

(5) Forecast of the number of countries that will win their first medal

The list of all countries that have not won a medal is obtained by looking up the countries that have won Olympic medals for each Olympics in the data, which facilitates the subsequent solution.

Regression analysis using the random forest model established above to predict the number of countries that will win a medal for the first time in each Olympic Games, and combined with the actual number of countries that will win a medal for the first time in each Olympic Games to determine the judgment criteria. The regression analysis can be concluded that when the predicted number of medals of a non-winning country in that Olympics is greater than 0.7, the country is considered to have a great probability of winning the award, as shown in Table 3.

Table 3. Distribution of missing values

Name of the Olympics	Number of first-time recipients
1988 Seoul Olympic Games	1
1992 Barcelona Olympic Games	3
1996 Atlanta Olympic Games	2
2000 Sydney Olympic Games	2
2004 Athens Olympic Games	3
2008 Beijing Olympic Games	2
2012 London Olympic Games	2
2016 Rio Olympic Games	2
2020 Tokyo Olympic Games	2
2024 Paris Olympic Games	5

The regression analysis leads to the conclusion that when the predicted number of medals for a non-winning country in that Olympics is greater than 0.7, the country is considered to have a very high probability of winning.

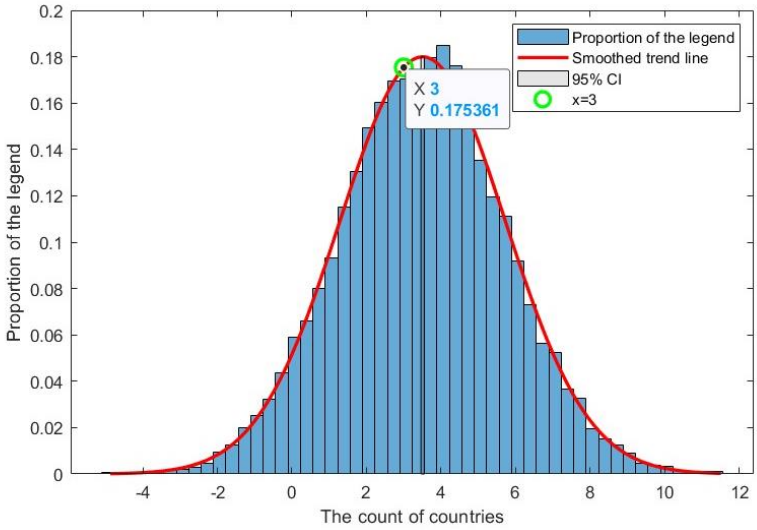

Figure 5. Normal distribution chart

(6) Projected results

The distribution of the sample data matches well with the theoretical curve of the normal distribution, so it can be assumed that these data are approximately normally distributed. As can be seen from the graph, the probability density value (0.175361) at x =3 is higher, indicating that the number of countries winning the first Olympic Games in 2028 has a higher probability of occurrence around 3, and that most of the data will be clustered within the 95% confidence interval. As shown in Figure 5

(7) Hypothesis testing

This paper use confidence intervals to test the number of countries winning medals for the first time in the 2028 Olympics obtained by solving for the number of countries winning medals for the first time in the 2028 Olympics, and since it obeys the distribution of χ^2 , we can get the confidence interval for the variance σ^2 is:

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right] \tag{7}$$

where μ is the average of the predicted number of countries that will win the Olympic Games for the first time in 2028, $\alpha = 0.05, \chi^2_{\alpha/2}(n)$ is the quantile of $\alpha/2$.

Hypothesis testing shows that the number of countries winning medals for the first time in the 2028 Olympics obtained from the solution is within the 95% confidence interval. Therefore, there is a 95% certainty that the mean of the predicted number of first-time medal-winning countries for the 2028 Olympics is three, and this estimate has a high level of confidence.

(8) Project changes

Based on the analysis of the data, this paper concludes that for a country, the type and number of changes in the dominant items have a great impact on the total number of medals, so the dominant items of each country should be evaluated first, and then the characteristic values should be set according to the changes in the dominant items of each country in the past years.

(9) Assessing Strengths Programs

Factor analysis is a multivariate statistical method that converts a large number of variables that may be correlated with each other into a smaller number of composite indicators that are not correlated with each other. Because of the correlation of the factors influencing the assessment of the strengths program, in order to simplify the calculations, this paper adopts the factor analysis method to assess the strengths program of each country.

In this paper, the indicators of the influencing factors are set as follows: the medal contribution of the country in a particular event, the international ranking of the country in a particular event, the historical participation experience of the country in a particular event, and the proportion of participants from the country in a particular event.

For p indicators $Y_1, \cdots Y_p$, the observation matrix Y is obtained, which p indicator variables may be influenced by $m(m < p)$ factors $f_1, \cdots f_m$, plus other influences, denoted as:

$$\begin{cases} Y_1 = b_{11}f_1 + b_{12}f_2 + \cdots + b_{1m}f_m + e_1 \\ Y_2 = b_{21}f_1 + b_{22}f_2 + \cdots + b_{2m}f_m + e_2 \\ Y_p = b_{p1}f_1 + b_{p2}f_2 + \cdots + b_{pm}f_m + e_p \end{cases} \tag{8}$$

where the influences $f_1, \cdots, f_m$ are random variables with mean 0 variance 1; b_{ji} is the loading of the jth variable on the ith common factor; e_i is an indicator-specific factor. Its mean is 0 variance is σ_i^2,The special factors are independent of each other and of the special factors and the common factors.

(10) Factor analysis score

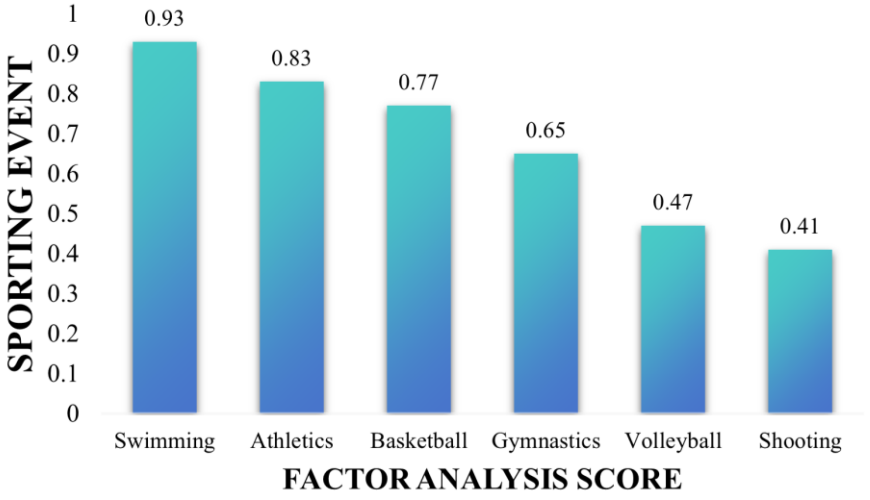

Figure 6. Factor Analysis Score Chart

Results are derived from this Random Forest Prediction Model, using the top 6 scoring programs in the United States as an example. As shown in Figure 6.

As you can see, Swimming is the most dominant sport in the U.S., while Athletics, Basketball and Gymnastics also have a big advantage. Based on comprehensive considerations, it was decided in this paper that the top four ranked programs would be the dominant programs.

Based on the framework of the previous research, some of the countries' strengths in specific sports areas are now systematically presented.

Table 4. Distribution of missing values

Nations	Project1	Project2	Project3	Project4
United States	Swimming	Athletics	Basketball	Gymnastics
Germany	Athletics	Cycling	Equestrian	Sailboard
Japan	Judo	Wrestling	Gymnastics	Baseball
France	Athletics	Equestrian	Fencing	Cycling
Brazil	Football	Judo	Volleyball	Sailboard

(11) Construct new eigenvalues and solve

Based on the statistical analysis of the dominant programs of each country completed in the previous study, this paper add feature 8 - the dynamic evolution dimension of the dominant programs of each country to the feature system of the existing Random Forest Model. This feature will focus on the migration pattern of dominant items in the time series, the structural adjustment of the competition pattern and the quantitative change of the intensity of the dominance of the items, etc. By constructing multi-dimensional feature vectors to achieve the dynamic portrayal of the evolution of the competition situation of each country, so as to optimize the model's explanatory power and prediction accuracy of the complex sports competition pattern. The introduction of this feature will break through the limitations of the traditional static feature modeling, and provide a new quantitative analysis perspective to reveal the internal mechanism of the dominance change of sports.

$$Q = \sum_{i=2}^{l} \left(A \cdot \frac{V_i - V_{i-1}}{V_{i-1}} \right) \tag{9}$$

where Q is the quantitative value of project changes in each country, V_{i-1} is the number of sports competing in the country's ith Olympics, l is the total number of Olympic events, A for whether the country is a dominant player in a given project, If yes, take 1, otherwise take 0.

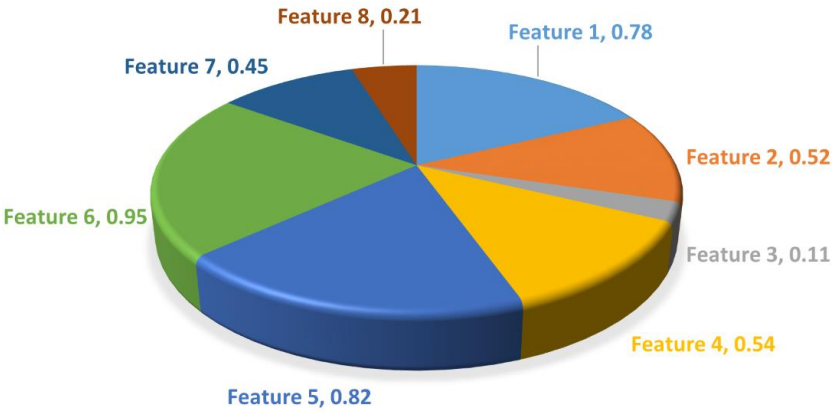

Figure 7. Importance map

As can be seen from Figure 7, the coefficient of importance of feature 8 - dominant program changes in each country is about 0.21, which strongly suggests how and to what extent program changes affect the number of medals. It suggests that when countries make program changes, they may affect resource allocation, training priorities and strategic planning, which in turn may have an impact on overall competitive performance.

3.3. Strategies of the organizing country

Since, for a given country, the number of medals for that country will increase somewhat if the items added are items of strength for that country, or decrease somewhat if the items removed are items of strength for that country, when the host country chooses the items.

For the U.S., as the host of the 2028 Olympics, it can create more medal opportunities by breaking down the program, adding more competition formats and expanding the number of dominant events.

4. Conclusions

This paper mainly adopts a random forest model based on Olympic medal count prediction, aiming at predicting the future Olympic medal list while analyzing the dominant events of different countries. The model is applicable to finance, health care, network security, marketing and other aspects, through the deep fusion with other models, parameter optimization, adjusting the selection of feature engineering, Random Forest can be used in a variety of application scenarios to predict the satisfactory results, and the credibility as well as accuracy is high.

Through the random forest model, this paper predicts the medal count interval and the ranking of each country in the 2028 Los Angeles Olympic Games, in which the United States has 49 ± 2 gold medals and 123 ± 4 total medals, both of which are ranked first in the world. The number of male participants and the number of female participants in each country have high importance coefficients on the prediction results, the more participants, the number of medals will also increase, and the number of participants indirectly reflects the talent pool of each country in sports.

The research limitations of this paper lie in the limitations of research methods, the limitations of data, as well as the insufficiency in theoretical depth and innovation. To overcome the limitations of research methods, future studies can actively introduce advanced research techniques and methods.

References

[1] Zheng Yuxin, Zhao Shengbo, Zhong Weny an, et al. Olympic Medal Prediction Model Based on Random Forest Regression [J]. Statistics and Application, 2025, 14: 47.

[2] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic Medals be Predicted? From the Perspective of Explainable Machine Learning [J]. Journal of Shanghai University of Sport, 2024, 48 (4): 26-36.

[3] Chen Rui, Huang Gaoduan, Liu Yongfeng. Research on the Evolution and Influencing Factors of the Provincial Competitive Sports Pattern in China since the 21st Century [J]. Journal of Chengdu Sport University, 2024, 50 (1): 122-131.

[4] Yi Jiandong. On the Possibility of "E-sports Entering the Olympics" from the Selection Criteria and Procedures, Concepts and Trend Evolution of Olympic Events [J]. Journal of Chengdu Sport University, 2022, 48 (3): 10-17.

[5] Liu Jun, Cao Junxing, Ding Weinan, et al. Research on reservoir porosity prediction method based on bidirectional long short-term memory neural network [J]. Progress in Geophysics, 2022, 37 (5): 1993-2000.

[6] Fei Luo. Analysis of the New Pattern of the Tokyo Olympics Medal Table in a Special Period [J]. Frontiers of Modern Education, 2022, 3 (1): 418-421.

[7] Peng Haojie, Zhou Yang, Hu Xiaofei, et al. PM 2.5 Concentration Prediction Model Based on Deep Learning and Random Forest [J]. Journal of Remote Sensing, 2023, 27 (2).

[8] Liu Shuai, Wang Tao, Cao Jiawen, et al. Evaluation of rainfall - induced landslide susceptibility based on an optimized random forest model [J]. Geological Bulletin of China, 2024, 43 (6): 958 - 970.

[9] Yu Wenmeng, Zhang Tingting, Shen Dajun. Analysis of the Pattern and Evolution of Influencing Factors of County - level Carbon Emission Intensity in China Based on the Random Forest Model [J]. China Environmental Science, 2022, 42 (6): 2788 - 2798.

[10] Liu Shuai, Wang Tao, Cao Jiawen, et al. Evaluation of rainfall-induced landslide susceptibility based on optimized random forest model [J]. Geological Bulletin, 2024, 43 (6): 958-970.

[11] Chen Jun, Su Chunyang, Jiang Yaqing. Calculation Model of Lime Dosage in Converter Based on Multiple Linear Regression [J]. Special Steel, 2024, 45 (1): 42.