Research on the prediction of the number of 2028 Olympic Medals based on K-Means and random forest

Cheng Zhao *

Northwestern Polytechnical University, Xi'an, China

* Corresponding Author Email: 13963860230@163.com

Abstract. In order to better predict the number of MEDALS in the next Olympic Games and provide a theoretical basis for developing resource optimization strategies, this study first classified the sports intensity of each country into four categories through K-Means clustering method, and combined the time series ARIMA model and random forest method. At the same time, the "host effect" is introduced, quantify the influence on the medal number, and introduce into the medal prediction model based on the historical award, GDP, total population, 2028, and get the medal list. Finally, the results show that the top three gold MEDALS in 2028 are: the United States, China and the United Kingdom. This study has demonstrated significant innovation in the field of sports by making a detailed division of national sports intensity, successfully constructing a targeted prediction model. This model more accurately reflects the intensity levels of sports activities and effectively utilizes the hierarchical characteristics of data, thereby making the results more precise.

Keywords: K-Means, ARIMA, random forest, host effect.

1. Introduction

Previous studies have focused on the distribution of medals at the overall team level. Schlembach predicted the performance of each team in the Olympics using a random forest model, assessing the contribution of different characteristic variables to the prediction [1]. After a review of previous studies, Coumeya established a theoretical framework for home-field advantage [2]. Wilson D compared the performance of the host countries at the 1960-2016 Summer Paralympics and the 1976-2014 Winter Paralympics, and their performance on the road [3].

Based on the above analysis, this study applied the K-Means clustering algorithm to subdivide the study subjects into four clusters with significant differences according to the activity of sports activity and medal status in each country. Subsequently, this study combined the ARIMA model in the field of time series prediction with random forest techniques in machine learning to make customized future trend predictions for the countries in the four clusters. On this basis, this study will be " host effect&quot, and through the quantitative method to evaluate the promotion effect of the total number of MEDALS, and the historical MEDALS data, national GDP, population size and athletes win multiple information integration into the medal prediction system, finally in 2028 countries in the performance of the medal table.

2. Clustering with the prediction algorithm

2.1. Data preprocessing

Upon reviewing the data from the Olympic website, it was found that some events in the data table summer Oly_programs have data gaps for different years, and some of the data are from a long time ago, making it difficult to find data with high reliability. To address this, we adopted the method of filling in zeros; in the data table summer Oly_medal_counts, there are spaces before and after some data in the NOC field, which affect program processing, so we performed the operation of deleting spaces; data from the periods of World War I and World War II were removed.

Conduct data integration verification on the matches, athletes, and events in the given multiple data tables; check if the medal counts in the data tables summer Oly_athletes and summer Oly_medal_counts are consistent.

Temporarily filter out countries that did not participate in the 2024 games, and among the remaining countries, select those that have participated consecutively up to 2024, calculate the mean and standard deviation for both gold medals and total medals, and also calculate the coefficient of variation:

$$k = \frac{\sigma_i}{\mu_i} \tag{1}$$

Among them, $\sigma_i$is the standard deviation of gold medals and total medals, $\mu_i$is the mean of gold medals and total medals to eliminate the impact of the volume of medals on the stability of standard deviation evaluation, reflecting relative fluctuations. From the data table summer Oly_athletes, filter out countries that have never won a medal.

2.2. K-means Clustering Method

In the k-means algorithm [4], the distance to the center is generally calculated using Euclidean distance:

$$d(x, c_i) = \sqrt{\sum_{j=1}^{d}(x_j - c_{ij})^2} \tag{2}$$

Wherein: x represents the data points, that is, the distribution of medals; d represents the number of indicators used to classify the strength of national sports;x_j, c_{ij}represent the values of x and c_i on the j-th dimension, respectively;

The objective function of K-means is:

$$J = \min \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x_i - c_k\|^2 \tag{3}$$

Where: $w_{ik}$is the prompt function. If the data point is within the cluster, w_{ik}= 1; otherwise, w_{ik}= 0. Differentiate the target function with respect to, you can get:

$$c_k = \frac{\sum_i^m w_{ik} x_i}{\sum_i^m w_{ik}} \tag{4}$$

The new cluster centroid is the weighted average of all data points x_i within that cluster, and the weight is w_{ik}.

2.3. Time Series ARIMA

The AR autoregression model is used to predict the [5], and the formula of the p-order autoregression process is as follows:

$$y_t = c + \sum_{i=1}^{p} \gamma_i y_{t-1} + \varepsilon_t \tag{5}$$

It is a constant $c\varepsilon_t$term; it is an error.

After the first-order difference of the original data, the prediction is made again, and then the prediction model formula is expanded as follows:

$$X_i = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \tag{6}$$

In the ARIMA model, the future value of the sequence is expressed as a linear function of the current and lag periods of the lag terms and the random interference terms. The general form of the model is as follows:

$$Y_t = c + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{7}$$

Where: $\alpha_i$is the autoregressive coefficient; q is the order of the moving average part; $\beta_i$is the moving average coefficient; using past observations of the series, future values of the series can be forecasted.

2.4. The Random Forest Approach

Random Forest is an ensemble learning method, the general process of which is shown in Figure 1. The basic architecture is decision trees, and by introducing randomness mechanisms, the model's resistance to overfitting and noise performance is enhanced [6]. RF exhibits stochastic characteristics in the process of sample and feature selection: (1) In terms of sample selection, the training samples for each decision tree are subsets obtained from the original dataset through a resampling technique with replacement; (2) In terms of feature selection, Random Forest randomly selects feature variables when constructing each base learner and further selects the optimal features for splitting. This stochastic mechanism significantly improves the model's generalization performance and learning ability.

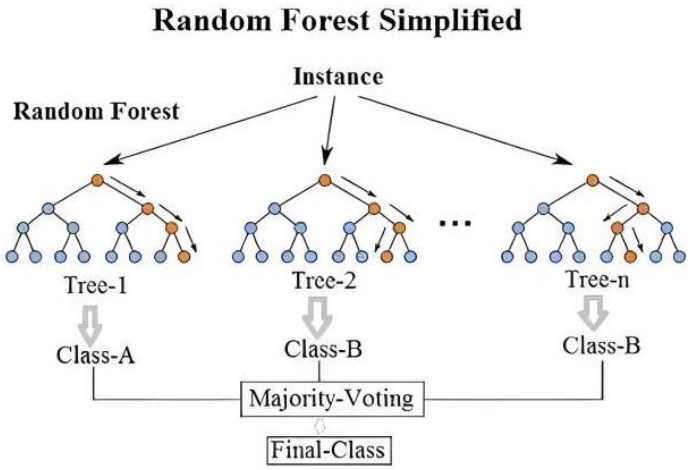

Figure 1. Random Forest Process Diagram

We interpret the contribution value of each feature variable to the model's prediction outcome as "the contribution of the variable (x) to the final prediction result (y) when participating in the model prediction" [7]. The "total prediction contribution" of a prediction model can be expressed as:

$$g(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i(x), x = x_{1,}, x_2, \cdots, x_m \tag{8}$$

x_m represents the explanatory or feature variable for the m-th dimension, such as a country's GDP, population count, etc.

By calculating, one can determine the impact of changes in on the number of medals or gold medals awarded, with its mathematical expression being:

$$\varphi_i(x) = \sum_{S \in F \backslash \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_s(x_s) \right] \tag{9}$$

F is the set of feature variables used by the model; S is a subset of F\{i}; represents all feature variables included in S; $f_{S \cup \{i\}}(x_{S \cup \{i\}})$and are the prediction results obtained from models trained based on the feature variable sets, respectively; $|F|$and $|S|$are the number of elements in the sets F and S, respectively [8].

3. Results analysis

3.1. National classification results

Through data organization, we have charted the distribution of medals since the year 2000, as shown in Figure 2. The distribution of medals is not uniform, with most being awarded to sports powerhouses such as the United States and China. The characteristics of medal distribution vary significantly among different countries; the number of medals for sports powerhouses is influenced by overall strength and the breadth of events, whereas other countries rely on the outstanding performance of individual athletes. Using a uniform model to predict medals for all countries could result in inaccuracies. To

enhance prediction accuracy and reflect the actual patterns of sports competition, we have employed the k-means clustering method to classify the sports strength of each country, thereby optimizing the medal prediction model.

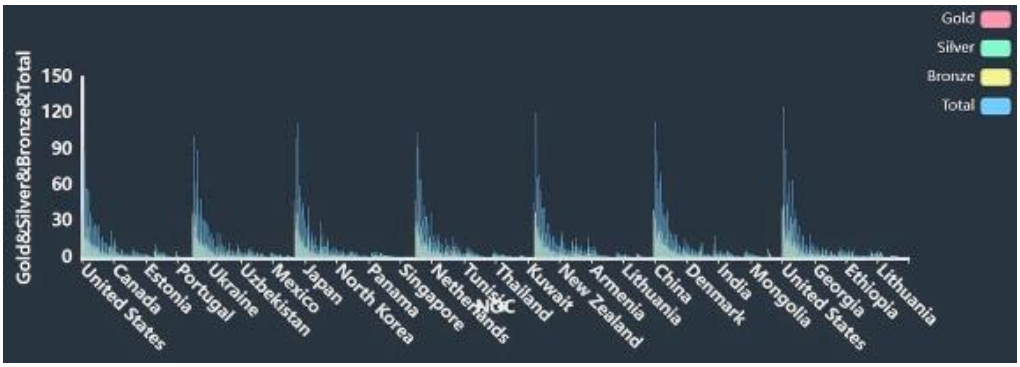

Figure 2. Medal Distribution Chart

We performed a comparative analysis of hierarchical clustering, k-means clustering and second-order clustering [9], and drew the distribution map of national sports intensity under the hierarchical clustering and K-means clustering methods, as shown in Figure 3 and Figure 4.

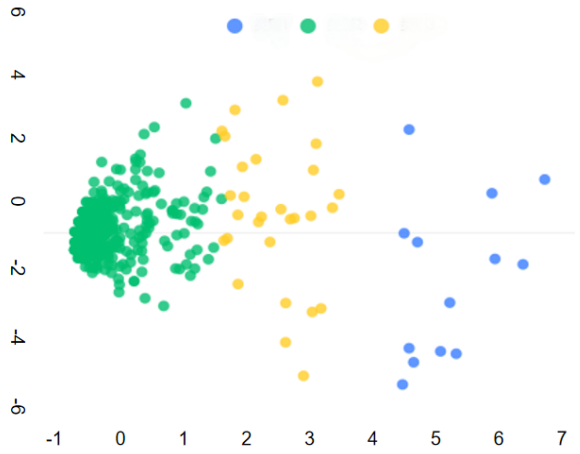

Figure 3. Hierarchical Clustering

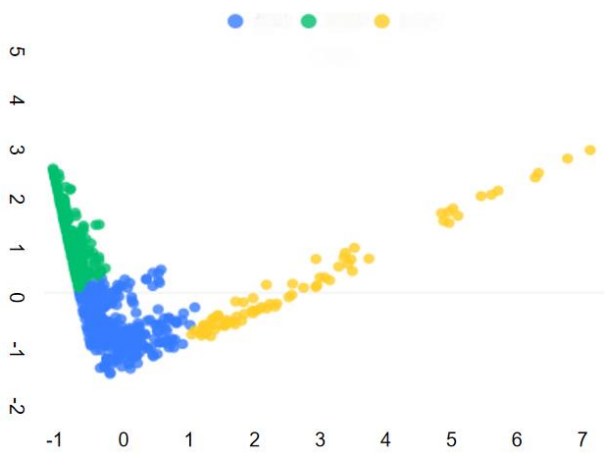

Figure 4. K-means Result Diagram

Among them, when the second-order clustering is performed according to the previous number of classifications, the silhouette score of the clustering model is calculated to be 0.007, indicating an unsatisfactory clustering result. Through analysis, it was found that after forcibly dividing the countries into three categories, the cluster center coordinates of the second and third categories are very close. Therefore, we changed the number of classifications to 2, and finally obtained a silhouette score of 0. 693. By calculating and comparing the model profile, DBI and CH, we finally used k-means clustering to initially divide countries into three categories according to sports intensity, and continued to divide the third category more extreme.

In the K-means clustering image, category 3 represents high sports intensity, category 1 represents medium sports intensity, and category 2 represents low sports intensity. Analysis of the image reveals that for the same principal component "1", its growth is accompanied by a trend of change in the dispersion of points, which is different for sports powerhouse countries and sports weaker countries. Under our hypothesis that "the distribution of medals is limited by the ceiling effect of intensified global competition, " for sports powerhouse countries, the number of medals (gold, silver, bronze) increases with the increase in gold medals. This is a macro reflection of the enhancement of their sports intensity due to the combined effects of various factors; whereas for countries with weaker sports intensity, due to limitations in resources and talent, their medal count maintains a dynamic balance. That is, an increase in the number of gold medals leads to a corresponding decrease in the number of silver and bronze medals. When the number of gold medals reaches a certain level, it confirms that the

increase in gold medals is due to the enhancement of sports intensity, rather than by chance factors. At this point, the corresponding number of silver and bronze medals will also increase accordingly, which is the turning point category 1 in the image.

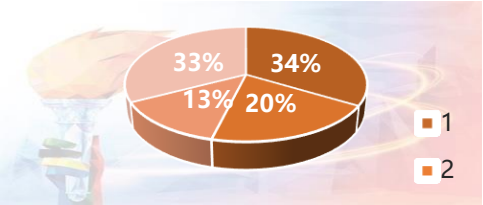

Figure 5. National Sports Intensity Classification

In the end, we divided all the participating countries into four categories: (1) sports powers; (2) many records, but low ranking and unstable; (3) fewer records; (4) never won a medal. The proportion of each category is shown in Fig. 5.

3.2. Forecast results of national medal number

We conducted a robust regression analysis specifically on the number of medals won by countries that have previously qualified to host the Olympics. Figure 6 is a robust regression of the United States medal count. By comparing the medal data from when these countries hosted the event with the predicted data from the robust regression, we obtained the impact of the host country effect on the number of medals.

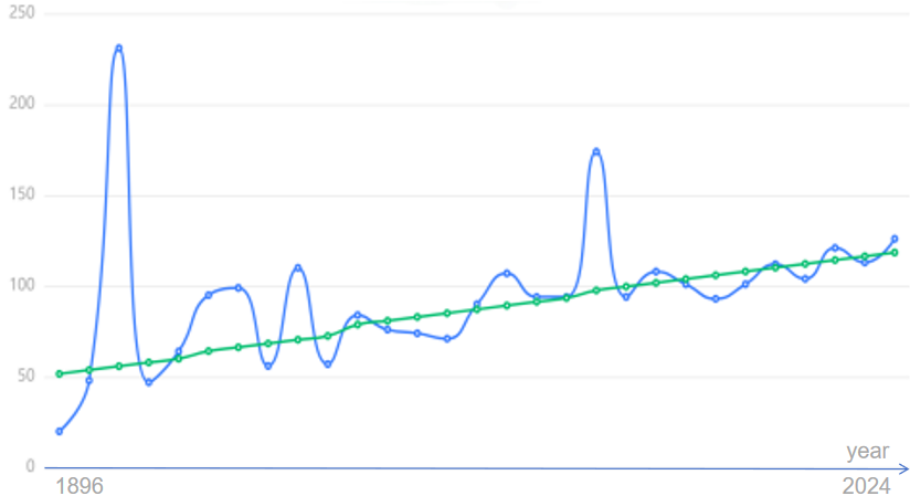

Figure 6. Stable Return of Medal Counts of the United States

Define the impact degree of the host effect:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - y_i'}{y_i} \tag{10}$$

N represents the number of times a country has participated in the Olympics as a host; is the predicted number of medals for the i-th time participating as a host through robust regression; is the actual number of medals for the i-th time participating as a host. Calculate the impact of the host countries in previous Olympics and, through multiple optimization fittings, obtain the image of the impact and time.

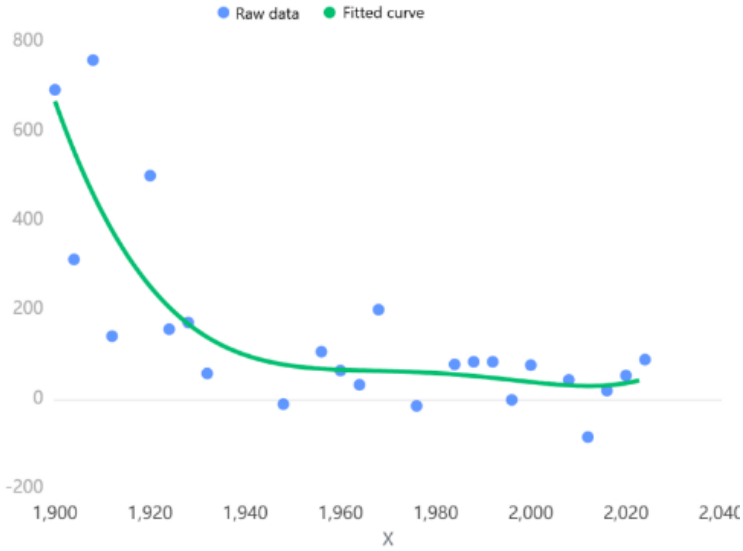

Figure 7. Host Country Effect Impact Degree Fitting Results

By comparing the division of countries with sports intensity, the sports power countries accounted for 93.7%, so the fitting results can be regarded as the fitting of the host effect of the specialization of sports power. The fitting results are shown in Figure 7. By analyzing and fitting images, before 1950, the host effect had a great impact on the medal number of a country, up to 758% (1908, Britain). After 1950, the influence area of the host effect was stable, finally converging to 44.78%, floating about 15%, and fluctuating between [29%, 59%].

3.2.1. Category 1 national medal number forecast

Under the premise of a small sample size, medal predictions cannot be based on historical medal counts; the number of Olympic medals does not exhibit a stable pattern, with a significant degree of randomness and volatility.

However, a sports powerhouse typically exhibits strong competitiveness across most events, and the fluctuations in individual athletes' performances are insufficient to affect the overall competitive trend, demonstrating the large-sample effect and long-term stability of the sports powerhouse. Therefore, we employ the Autoregressive Integrated Moving Average (ARIMA) model to forecast the number of gold and medals for the sports powerhouse.

The prerequisite for the applicability of the ARIMA model is the stationarity of time series data [10]. Through the Augmented Dickey-Fuller (ADF) test results, based on the analysis of the t-statistic, we determine whether the null hypothesis of non-stationarity can be statistically significantly rejected ($P<0.05$). By comparing the data graphs before and after differencing, we assess the stationarity of the series (fluctuations should be relatively stable) [11]. We perform partial autocorrelation analysis on the time series and infer its p, q parameter values based on the truncation phenomenon. Another requirement of the ARIMA model is the pure randomness of the residual series, meaning that residuals should manifest as white noise. We test the white noise characteristics of the residuals using the P-value of the Q statistic from the model test table ($P>0.05$). Additionally, we evaluate the model using information criteria such as AIC and BIC values (the smaller the value, the better the model), and we conduct auxiliary analysis through the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) graphs of the model residuals [12, 13].

Here are the examples of gold MEDALS and MEDALS:

Organize the historical data of American medals, and replace the medal data when the US was the host with the data of robust regression fitting for ADF test and partial autocorrelation analysis.

According to the ADF test, the ARIMA model (2, 1, 0), and the fitting formula is as follows:

$$y(t) = 6.802 - 0.926 * y_{(t-1)} - 0.608 * y_{(t-2)} \quad (11)$$

Among them, indicates the number of MEDALS won in t year.

Considering that the United States is the host country of the 2028 Olympic Games, combined with the host effect, the final number of MEDALS in the 2028 Olympic Games is predicted to be 181, and the change range is within the [161, 198] range.

Figure 8. Time Series Fitting of the Number of Medals of the United States

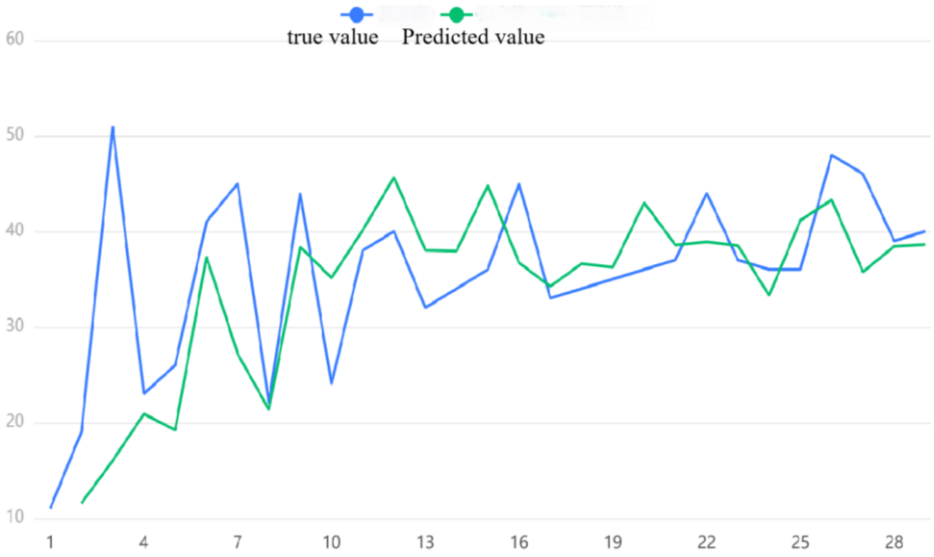

Figure 9. Time Series Fitting of the Amount of Gold Medals of the United States

Figure 8 and Figure 9 are the prediction result graphs during the time series training sample process. And similarly, the US gold medal time series fitted image is obtained. Finally, combined with the host effect, it is predicted that the number of gold MEDALS in the 2028 Olympic Games will be 65, and the change range is within the [58, 71] range.

3.2.2. Category 2 national MEDALS forecast

For category 2 countries (with many records but low ranking and unstable), a random forest model was used to predict. Organize the GDP, population and project participation in the Olympic Games, and establish the project advantage of each country [14]; integrate the GDP of the previous year and the current year into the unified sports input index according to the weight of 0.75 and 0.25 respectively. Combined with the genetic algorithm [15], the dominance of other items is randomly expanded according to the sports investment index each time, and the expansion coefficient is determined through machine learning. A random forest regression model was calculated from the training set data. The established random forest regression model was applied to the training and test data, and the model evaluation results were obtained.

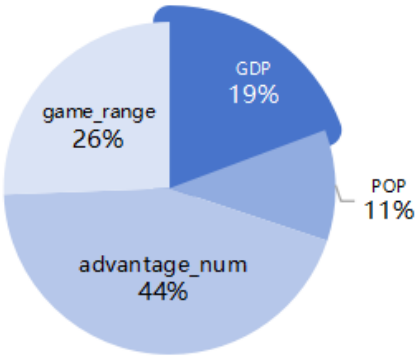

Figure 10. Feature Vector Interpretability

The figure 10 shows the prediction of the training set of some national random forest models and the importance of each feature vector in the model. It can be found that the national GDP, the number of population and the breadth of participating projects all influence the prediction of the number of MEDALS, while the influence degree of dominance is the largest, accounting for 44% of the total proportion.

3.2.3. Category 3 and 4 national medal prediction

For category 3 and 4 countries, targeted modeling is conducted because they have few entry records or have never won a medal. We take the Morocco teams in category 3 countries as an example to predict the number of their 2028 Olympic gold MEDALS:

Table 1. Morocco Participation in the previous Olympic Games

Name	Sex	Team	Sport	Event	Medal
Soufiane El	M	Morocco	Athletics	Men's 1500m	No medal
Soufiane El	M	Morocco	Athletics	Men's 3000m Steeplechase	Gold
Soufiane El	M	Morocco	Athletics	Men's 3000m Steeplechase	Gold

Identify competitors with competitiveness in the men's 3000-meter steeplechase and their participation records, The specific situation can be seen in Table 1. It is observed that the breadth of medal-winning events in this country is narrow, with only the men's 3000-meter steeplechase being a medal event; the number of athletes is small, with only one athlete's participation record. Therefore, for the prediction of medals for this country, it is necessary to analyze specifically from certain events and athletes. Identify competitors with competitiveness in the men's 3000-meter steeplechase and their participation records, the results are shown in Table 2, calculate the probability of each athlete winning a gold medal, the mathematical expression is:

$$p_k = \frac{\sum_{i=1}^{3} \varepsilon_i n_{k,i}}{\sum_{i=1}^{3} n_{k,i}} \tag{12}$$

p_k is the probability of athlete k winning a gold medal; is the evaluation weight coefficient; is the number of medals of type i that athlete k has won.

Table 2. Gold medal prediction for the men's 3000 m steeplechase in the 2028 Olympic Games

name	win_pre
Lamecha Girma	0.15
Benjamin Kigen	0.1
Soufiane El	0.5
Abraham Kibiwot	0.05
Kenneth Rooks	0.3

That is, the Morocco team has a probability of winning a gold medal in the men's 3000 m obstacle event, and because the only team has won a medal in the event, the number of gold MEDALS in the 2028 Olympic Games is 1, with a probability of 0.5.

3.2.4. Summary

Combining the establishment and solution of the above four types of national MEDALS prediction models, we get the final prediction results as shown in Figure 11:

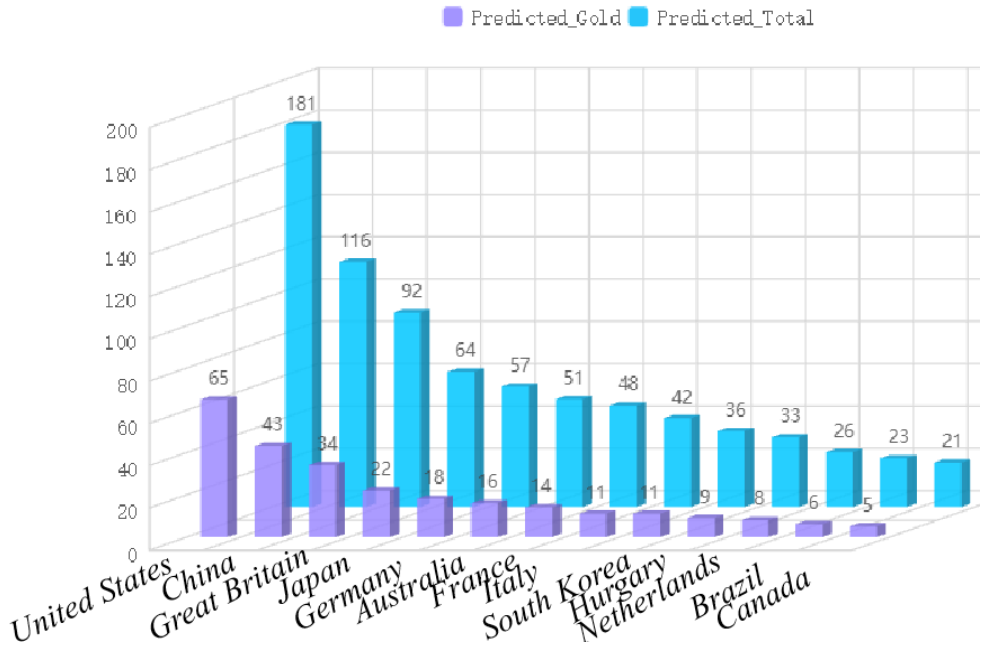

Figure 11. 2028 Olympic Games Gold Medal and Medal Predictions for some Countries

Countries of Category 1 and Category 2 are specific forecast data, while countries of Category 3 and Category 4 are the forecast fluctuation range. By comparing the predicted number of golds in 2028 with the actual number of golds in 2024, we can identify which countries have a significant increase or decline in 2028. Combined with the prediction interval, the significance of these changes can be judged.

4. Summary

This study used the K-Means algorithm to divide countries into four clusters according to physical activity and medal count, revealing different performance patterns and laying the foundation for customized predictions. Cluster analysis revealed significant differences between clusters in sport participation and ability to medal, demonstrating the validity of the analysis. Studies combine ARIMA model and random forest technology for customized prediction, considering cluster differences to improve accuracy. The study also assessed the impact of the host country effect on the total number of medals and found that it significantly increased the number of host countries, providing a new perspective for understanding home field advantage. Integrate the information such as historical MEDALS, GDP, population and win rate to make the prediction models closer to the reality and improve the practicability and accuracy of the prediction.

(1) Data analysis shows that the distribution of Olympic MEDALS is not uniform. If the same model fits are used, it may cause a prediction bias. Through cluster analysis, the division of sports intensity in each country can make full use of the hierarchical characteristics of the data, improve the accuracy of prediction, and better reflect the realistic law of sports competition.

(2) ARIMA is good at trend capture, and random forests are better at handling complex relationships and interactions. Combining both enhances the comprehensiveness and reliability of the prediction.

In future research, we can also attempt to combine machine learning methods with other statistical methods, such as Bayesian networks and deep learning, to explore more efficient and accurate prediction models. At the same time, for the characteristics of different sports and countries, we can conduct more detailed and in-depth customized research to provide more specific and targeted advice

and guidance for sports managers, coaches, and athletes worldwide. With these efforts, we hope to bring more innovation and breakthroughs to the field of Olympic medal prediction and sports competition analysis.

References

[1] SCHLEMBACH C, SCHMIDT S L, SCHREYER D, et al. Forecasting the Olympic medal distribution: A socioecono-mic machine learning model [J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[2] Carron A V, Loughhead T M, Bray S R. The Home Ad-vantage in Sport Nompetitions: Courneya and Carron's (1992) Conceptual Framework a Decade Later [J]. J SportsSci, 2005, 23 (4): 395 - 407.

[3] Wilson D, Ramchandani G. An Investigation of Home Advantage in the Summer Paralympic Games [J]. Sport Sciences for Health, 2017, 13 (3): 625 - 633.

[4] Mimenbayeva A, Artykbayev S, Suleymanov R, et al. Determination of the number of clusters of normalized vegetation indices using the k-means algorithm [J]. Eastern-European Journal of Enterprise Technologies, 2023, 5 (2): 42 - 55.

[5] Feng J, Huang B. Forecasting Carbon Emission Using ETS Exponential Smoothing, ARIMA and Regression with ARIMA errors Techniques [J]. International Journal of Engineering and Technology, 2024, 16 (3).

[6] Johansson U, Boström H, Löfström T, et al. Regression conformal prediction with random forests. [J]. Machine Learning, 2014, 97 (1-2): 155 - 176.

[7] Indrayudh G, Giles H. Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and Its Variance Estimate [J]. Journal of Computational and Graphical Statistics, 2020, 30 (2): 493 - 502.

[8] Mantas J C, Castellano G J, Moral-García S, et al. A comparison of random forest-based algorithms: random credal random forest versus oblique random forest [J]. Soft Computing, 2019, 23 (21):10739 - 10754.

[9] Xue L, Jiameng Z, Yong Z, et al. Clustering and Bellerophon state in Kuramoto model with second-order coupling. [J]. Chaos (Woodbury, N. Y.), 2019, 29 (4): 043102.

[10] Basariya N M, Murugesan P. An approach to arrive at stationarity in time series data [J]. International Journal of Applied Management Science, 2022, 14 (3): 221 - 245.

[11] Belashov Y V, Belashova S E, Asadullin I A. Testing the time series for stationarity in systems for processing of experimental data [J]. Radio physics and quantum electronics, 2013, 55 (9): 587 - 592.

[12] Davies N, Petruccelli D J. On the Use of the General Partial Autocorrelation Function for Order Determination in ARMA (p, q) Processes [J]. Journal of the American Statistical Association, 2012, 79 (386): 374 - 377.

[13] A New Algorithm for Automated Box-Jenkins ARMA Time Series Modeling Using Residual Autocorrelation/Partial Autocorrelation Functions [J]. Industrial Engineering & Management Systems, 2006, 5 (2): 116 - 125.

[14] BERNARD A B, BUSSE M R. Who wins the Olympic Games: Economic resources and medal totals [J]. Reviewof Economics and Statistics, 2004, 86 (1): 413 - 417.

[15] Genetic Algorithms; Findings on Genetic Algorithms Detailed by Investigators at Otsuma Women's University (Optimization technique by genetic algorithms for international logistics) [J]. Journal of Technology & Science, 2014.