Research on Prediction Methods Based on Random Forests and Quantile Regression

Ming Liu *

Department of Communication engineering, University of Science and Technology Beijing, Beijing, China

* Corresponding Author Email: liumingcam@163.com

Abstract. This study constructs a hybrid prediction framework that integrates random forest and quantile regression. The study first integrates multi-dimensional variables reflecting trends and attribute characteristics to construct the input feature system for the prediction model. By leveraging the ensemble learning mechanism of the random forest regression model to integrate the output results of multiple decision trees, the framework captures the distribution patterns of the data to achieve predictions for continuous variables. Additionally, the quantile regression model is introduced to select feature variables, quantify model uncertainty by setting quantiles, and generate prediction intervals. Furthermore, the study extends the application scope by incorporating a random forest classifier and conducts correlation analysis to uncover positive associations between feature variables and target variables, verifying the significant impact of multi-dimensional inputs on prediction outcomes and enhancing the model's interpretability. This framework enhances data analysis and prediction capabilities through multi-model collaboration and multi-method comprehensive analysis, demonstrating its applicability in related data processing tasks and providing a reference framework for similar studies.

Keywords: Random forest regression model, quantile regression model, correlation analysis, multi-model collaboration.

1. Introduction

In the field of data-driven predictive research, effectively integrating multi-dimensional features and quantifying predictive uncertainty have become key challenges. Existing methods often face the dilemma of balancing model complexity and predictive accuracy when dealing with high-dimensional non-linear data, and they also lack adequate assessment of uncertainty in the results [1]. This study focuses on constructing a hybrid predictive framework that integrates random forests and quantile regression, aiming to capture the distribution patterns of data through ensemble learning mechanisms while introducing quantile analysis to enhance the robustness of predictive results [2].

The study first constructs an input feature system for the predictive model by integrating multi-dimensional variables that reflect trend changes and attribute characteristics. Based on this, the random forest regression model effectively handles non-linear relationships in the data by integrating the output results of multiple decision trees, achieving high-precision predictions for continuous variables [3]. Quantile regression provides interval estimates for prediction results by setting different quantiles, enhancing the model's ability to characterise data uncertainty [4]. Additionally, the study introduces a random forest classifier and combines it with correlation analysis to reveal the positive association mechanism between feature variables and target variables, confirming the significant impact of multi-dimensional inputs on prediction results and further enhancing the model's interpretability and reliability [5].

The core objective of this study is to construct an analytical framework that combines prediction accuracy with uncertainty quantification capabilities through multi-model collaboration and multi-method fusion, providing a reusable methodological reference for similar data processing tasks and promoting the cross-application of ensemble learning and quantile analysis in the field of prediction.

2. Random Forest Regression Model

2.1. Model Building

The core idea of this model is to use random forest regression to predict the distribution of Olympic medal counts. The random forest is an ensemble learning method that combines the outputs of multiple decision trees to improve the accuracy and stability of predictions [6]. The model can be expressed as:

$$y_{\text{pred}} = f(X) = \frac{1}{T}\sum_{t=1}^{T} T_t(X) \tag{1}$$

Where $T_t(X)$ is the prediction of the t-th decision tree for the input feature X, T is the total number of decision trees. X is the vector containing all input features.

To predict medals, this paper needs to build a mathematical model based on a series of historical data features. The goal is to predict the future medal counts for each country using a random forest regression model. Let the target variable be y (the predicted medal count), which is expressed as:

$$y = f\left(x_1, x_2, x_3, x_4, x_5, x_6, x_7\right) \tag{2}$$

Where f is the Random Forest regression model, which predicts by integrating the outputs of multiple decision trees. Each decision tree T_t makes a prediction based on the input features x, and the final predicted result is the average of these predictions:

$$y_{RF} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{3}$$

Where T is the number of decision trees and $\hat{y_t} = T_t(X)$ is the prediction of the t-th tree.

2.2. Model Evaluation

To evaluate the performance of this prediction model, this paper compared it with the XGBoost regression model. XGBoost (Extreme Gradient Boosting) optimizes the objective function by constructing a series of weak learners (typically regression trees) to make the final prediction. For each training sample, XGBoost aims to predict the medal count by learning a weighted sum of decision trees. The final prediction of XGBoost is the weighted sum of all base learners, represented as:

$$y_{\text{pred}} = f(X) = \frac{1}{T}\sum_{t=1}^{T} T_t(X) \tag{4}$$

Where y_i is the predicted medal count for sample i, T is the number of decision trees (number of iterations), $f_t(x_i)$ is the prediction of the t-th tree for input feature x_i.

To compare the performance of both models, this paper evaluated the random forest regression model and XGBoost using R^2 to measures how well the model fits the data. The closer the value of R^2 is to 1, the better the model explains the data.

Compared with the two, the R^2 of random forest is closer to 1. It can be seen that the fitting effect of random forest on the actual situation is better.

2.3. Data Feature Extraction

Considering that the prediction of the number of medals is usually not based on the historical medal data, but on the list of known contestants, in terms of data feature processing, this paper only selected the data of the past 2016, 2020 and 2024 Olympic Games, so as to prevent dependence on all the historical Medal data. And this paper selected the following features to establish the model:

(1) Gold: the number of gold medals in 2016, 2020 and 2024 Olympic Games of the country.

(2) Total: total medals of 2016, 2020 and 2024 Olympic Games of the country.

(3) Athlete_count: the number of athletes participating in the Olympic Games in each session of the country.

(4) Athlete_growth_rate: it reflects the changing trend of athletes in the country over time.

(5) Gold_growth_rate: reflects the change trend of the number of gold medals over time.

(6) Total_growth_rate: reflects the trend of the total number of medals over time.

(7) Is_Host: it reflects whether the country is the host country. The host country usually performs better due to its home advantage, which affects the distribution of gold medals and total medals.

(8) Medal: predict the target variable of the classification model in the number of countries winning the first prize. The award is expressed as 1, otherwise it is expressed as 0.

2.4. Medal Forecast Results

In order to quantify the uncertainty of the model, this paper used quantile regression,

$$y_\tau = Q_\tau(y|\ X) \tag{5}$$

Where $Q_\tau(y|X)$ denotes τ-quantile under a given characteristic x, $\tau \in (0,1)$ is quantile (for example, $\tau = 0.5$ corresponds to the median regression, $\tau = 0.1$ corresponds to the 10th percentile regression, and $\tau = 0.9$ corresponds to the 90th percentile regression). X is the input eigenvector (such as the number of historical medals, the number of participants, etc.). $\beta(\tau)$ is the parameter vector of quantile regression model, which varies with τ.

The output of quantile regression is the predicted number of medals under each quantile. Then the gold medal number prediction interval and the total medal prediction interval are as follows:

$$Gold\ Prediction\ Interval = \left[\ Gold_{0.1}, Gold_{0.9}\right] \tag{6}$$

$$Total\ Prediction\ Interval = \left[\ Total_{0.1}, Total_{0.9}\right] \tag{7}$$

The prediction results are shown in Fig. 1and Fig. 2.

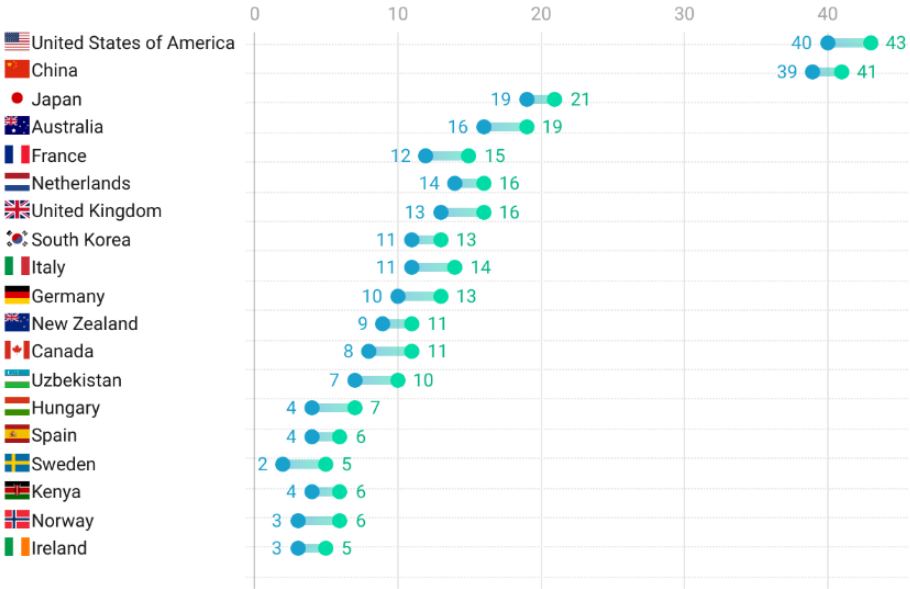

Fig. 1 Prediction range for the gold medals number

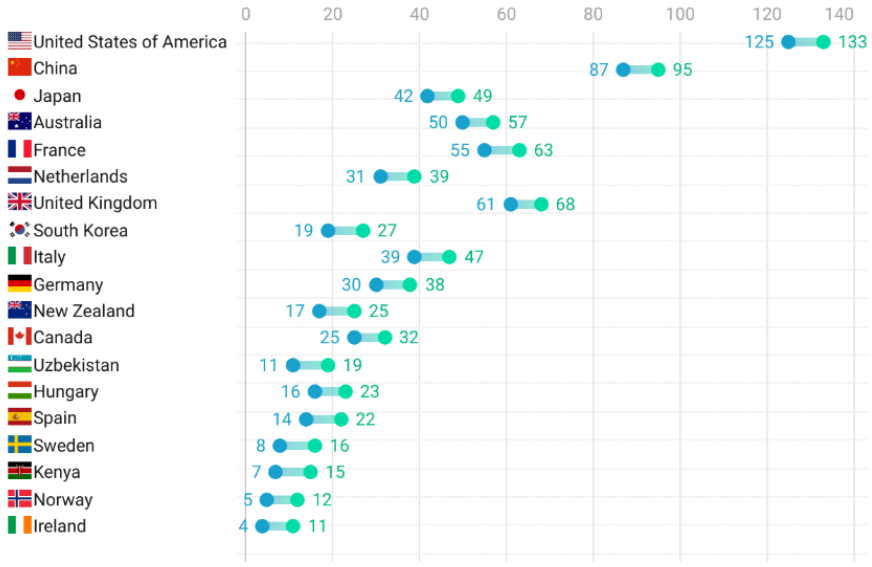

Fig. 2 Prediction range of total medal count for some countries

3. Analysis of National Medal Performance Trends

3.1. Country Performance Analysis

After predicting the number of gold medals and the total number of medals in the 2028 Olympic Games, this paper can use the following formula to calculate the changes in the number of gold medals and the total number of medals:

$\Delta Gold$ and $\Delta Total$ are the changes in the number of gold medals and total medals of the 2028 Olympic Games compared with the 2024 Olympic Games. If $\Delta Gold$ >0, it means that the number of gold medals in this country has increased and played better; if $\Delta Gold$ <0, it means that the number of gold medals in this country has decreased compared with the previous one and played worse. If $\Delta Total > 0$, it means that the total number of medals increases and the national level rises; $\Delta Total < 0$, indicating that the total number of medals has increased compared with the previous session, and the national level has declined.

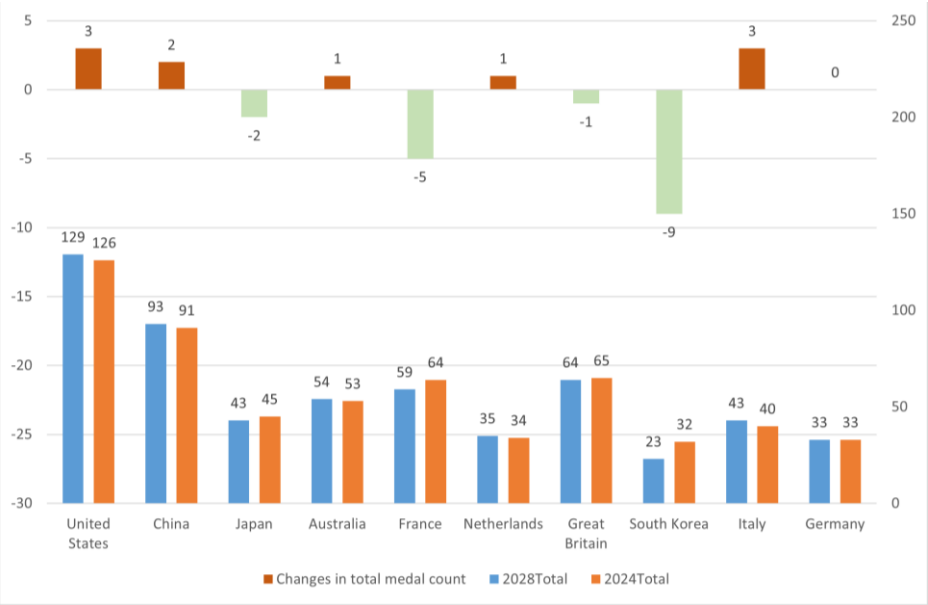

Fig. 3 Progress and regression in some countries

As shown in Fig. 3, the number of medals of the United States, China, Australia, Italy, the Netherlands and other countries is expected to increase, indicating that their competitiveness in the

future Olympic Games will be enhanced. The number of medals in Germany and other countries is expected to remain unchanged and the performance is relatively stable. The number of medals in Britain, France, South Korea and Japan is expected to decline, and their performance is expected to decline.

3.2. Predict the countries that win medals for the first time

3.2.1. Random forest classification model

This paper still used the random forest classifier to predict whether each country will win a medal in the next Olympic Games. In this task, this paper paid special attention to those countries that have not won medals in history, predict whether they will win the first prize in the next Olympic Games, and estimate the probability of winning. For a country with zero gold medals and total medals (i.e. a country that has not won a medal in History), if the model predicts that it will win the prize, it will be considered as the first country to win the prize.

$$\text{FirstTime Medal} = \begin{cases} 1 & , \quad \text{Predicted\_Medal}=1 \cap \text{Gold}=0 \cap \text{Total}=0 \\ 0 & , \text{otherwise} \end{cases} \tag{8}$$

3.2.2. Result analysis

The model predicts that several countries may win medals for the first time in the next Olympic Games, which provides each country with the prediction probability of winning medals. Fig. 4 shows some of the results of the first prize prediction.

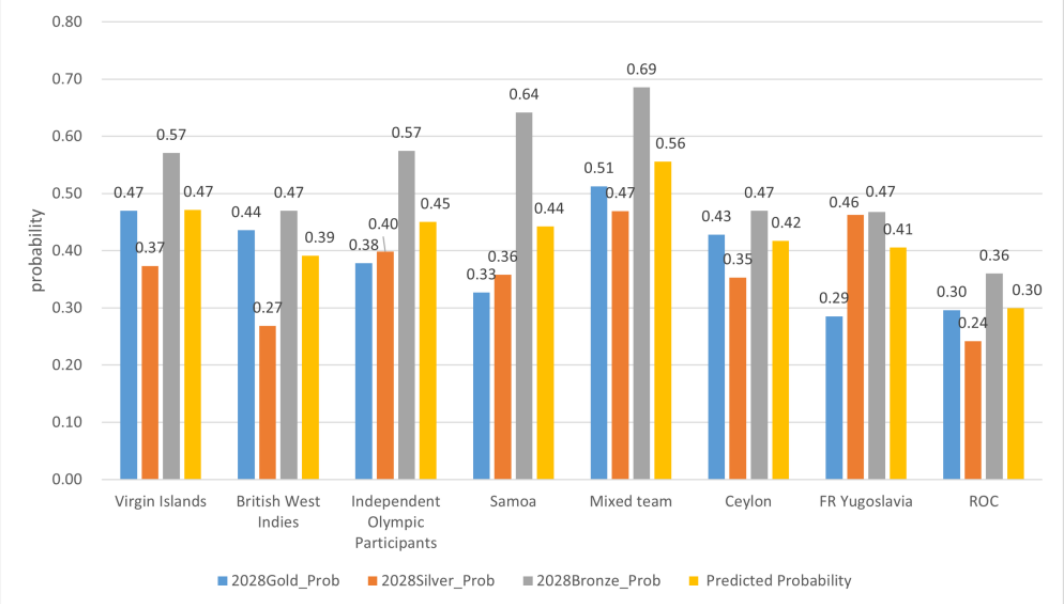

Fig. 4 Prediction first-time winning countries

For those countries that have not won medals, countries with a higher probability of pre-diction are more likely to achieve breakthroughs in the future Olympic Games.

4. A Study on Factors Affecting Medal Distribution

4.1. The Impact of the Project on the Distribution of Medals

After data preprocessing, this paper can analyze the correlation between the type and number of events and the number of medals, and then get the impact of the type and number of events on the number of medals. In order to do the correlation analysis, this paper only needs to calculate the Pearson correlation coefficient between the variety and number of events and the number of medals. The calculation formula is as follows:

$$\text{Correlation}(a,b) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{9}$$

Where a_i is the quantity change of item i, b_i is the number of medals for the project, $\bar{a}$ and $\bar{b}$ are the average value of the number change and the number of medals.

By calculating the Pearson correlation coefficient between the change of events and the number of medals, this paper can get the correlation between the two. It can be seen that the correlation coefficient between the change of 2016-2024 projects and the number of medals is 0.3, and the correlation coefficient between the change of 2020-2024 projects and the number of medals is 0.2. The data shows that the number of events has an impact on the number of medals, and is positively correlated.

Analyze which projects are the most important to the country. This paper drew the following chart of the types of projects and the number of medals in each country.

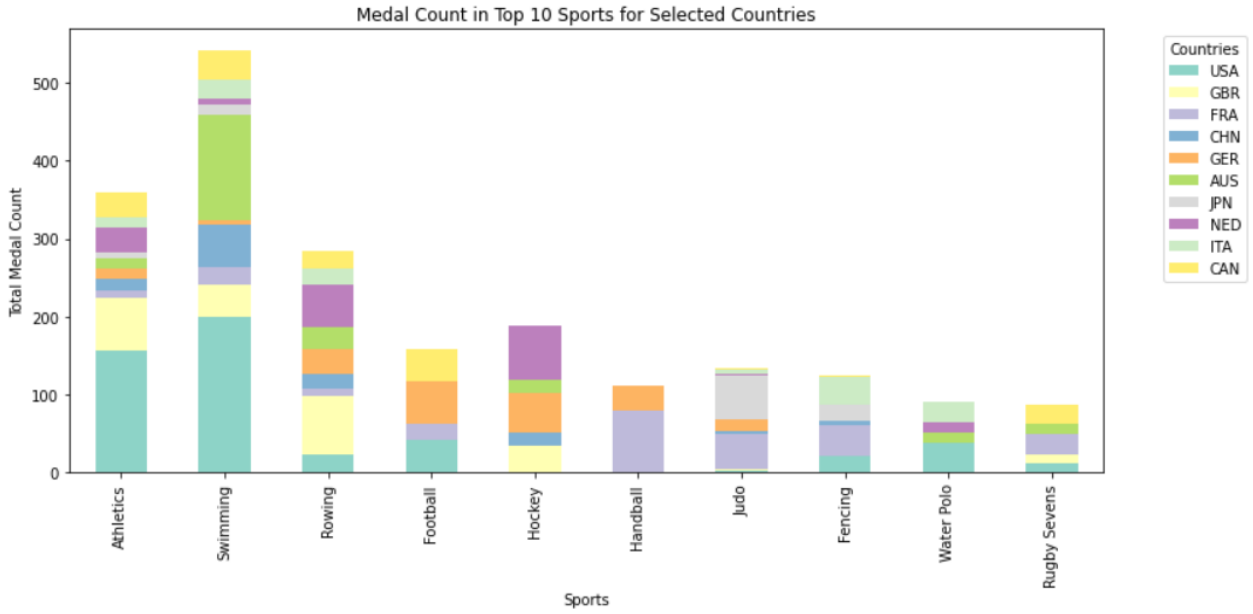

Fig. 5 Types of projects and number of medals awarded by each country

It can be seen from the data in Fig. 5 that athletes and swimming are important to the United States, while hockey and rowing are important to the number of medals.

4.2. The Impact of Host Country's Event Selection

When analyzing the relationship between the host country's event selection and national medal counts, the number of medals a country wins in different sports events may be influenced by both the quantity and type of events. For each host country, this paper calculated the correlation be-tween the changes in medal counts across various events and the types of events chosen. This will help to reveal which host country event selections have a significant impact on their medal counts.

$$\Delta X_{\text{Host}} = X_i - \frac{1}{Y_{\text{NonHHost}}} \sum_{i \in \text{NonHost}} X_i \tag{10}$$

X_i means the medal counts of different events for the host country, and $Y_{NonHost}$ means the number of events in which the host country participates. By calculating the difference between the changes in the number of medals for each event, this paper can determine the contribution of each event to the total medal count.

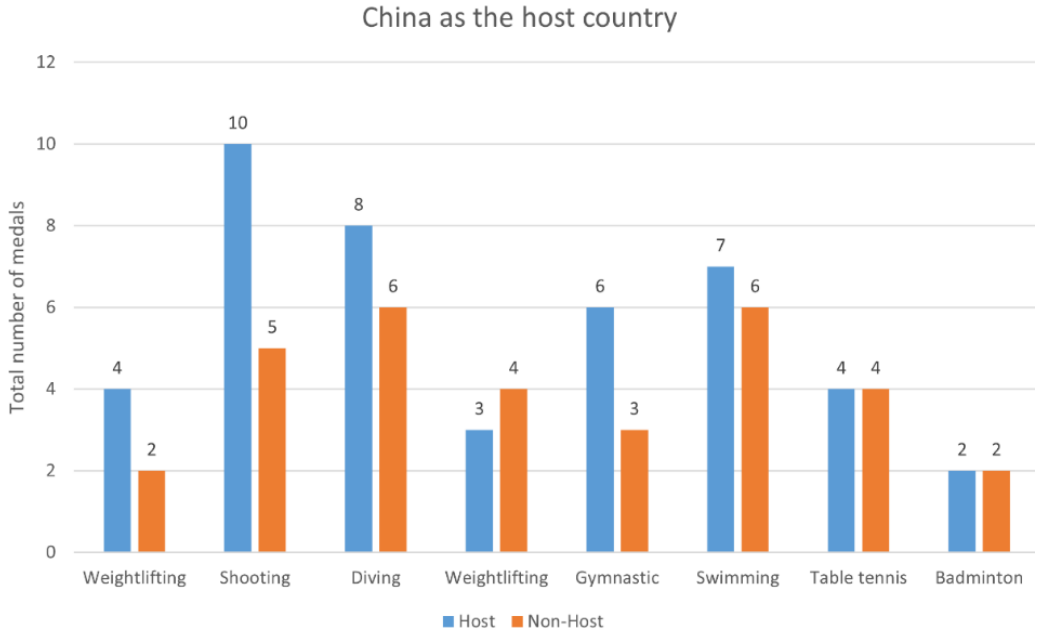

Fig. 6 When China as the host country

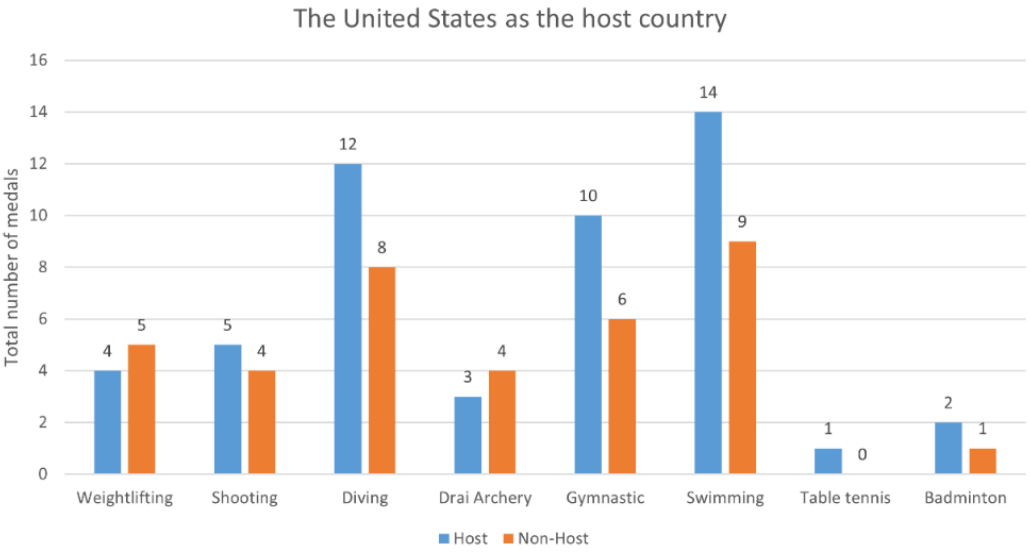

Fig. 7 When United States as the host country

As shown in Fig. 6 and Fig. 7, in some events, the host country performs better than when it competes as a non-host. The choice of events by the host country also has an impact on its medal distribution. Host countries tend to choose events where they have a competitive ad-vantage and allocate more resources and support to these events. This "host country effect" is particularly evident in the medal distribution. For example, in the 2008 data, China won 20% to 30% more medals in its dominant events (such as table tennis, diving, and badminton) compared to other years.

The reason lies in the fact that the significance of these events is not only reflected in the Olympic medal counts but also influenced by cultural and historical traditions. For instance, the United States has a strong advantage in swimming, athletics, and basketball, which is closely linked to its large sports industry, athlete selection system, and national resource allocation.

The changes in the selection of events significantly affect the medal counts of various countries, and there are clear differences in performance across countries in different events. The choice of events by the host country has a potential impact on its medal distribution, espe-cially when the host country adds more events where it holds an advantage, which may lead to significant changes in the medal count.

5. Summary

The study developed a hybrid prediction framework combining random forest and quantile regression, achieving effective prediction of the target variable and quantification of uncertainty. First, the random forest regression model integrates the output results of multiple decision trees to capture the non-linear distribution patterns of the data, enabling high-precision prediction of continuous variables. Quantile regression addresses the issue of traditional point prediction lacking uncertainty metrics by setting a quantile τ to generate prediction intervals. Additionally, the study used a random forest classifier and correlation analysis to reveal the positive association mechanism between feature variables and target variables, indicating that multi-dimensional inputs reflecting trend changes and attribute characteristics significantly influence prediction results, thereby enhancing the model's interpretability and reliability. In summary, this hybrid framework demonstrates efficient analytical capabilities and prediction robustness for high-dimensional data through multi-model collaboration and multi-method comprehensive analysis, providing a reference analysis paradigm for similar studies. Future research can further explore the dynamic update mechanism of feature variables to improve the timeliness and generalisation capabilities of the model in time series data prediction.

References

[1] Yang Yanping, Li Rong. Feature selection for high-dimensional data classification based on machine learning [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2025, 37 (01): 23-31.

[2] Wang Lei, Yang Yaya, Zhao Yini, et al. Analysis of Factors Influencing Hospitalisation Costs for Elderly Injury Patients Based on Quantile Regression and Random Forest [J]. Chinese Journal of Chronic Disease Prevention and Control, 2023, 31 (04): 246-250.

[3] Gan Ruiping, Ren Xinmin, Jiang Jun, et al. Daily energy consumption prediction for ship spe-cial coating maintenance based on random forest regression [J]. Big Data, 2024, 10 (01): 170-184.

[4] Zhao Shuran, Song Ningjing, Ren Peimin, et al. Systemic Risk in the Energy Industry from a Group Analysis Perspective: Based on a Network Dynamic Quantile Regression Model [J]. Systems Engineering, 2024, 42 (01): 15-26.

[5] Li Xinhai. Application of Random Forest Models in Classification and Regression Analysis [J]. Journal of Applied Entomology, 2013, 50 (04): 1190-1197.

[6] Zhang Kunbin, Chen Yuming, Wu Kesheng, et al. Research on a Random Forest Classification Algorithm Driven by Grain Vectors [J]. Computer Engineering and Applications, 2024, 60 (03): 148-156.