Is it feasible to achieve a precise and scientific prediction regarding the tally of medals awarded in the Olympic Games?

Summary

In this paper,we want to predict the number of Olympic medals using scientific methods, with the aim of addressing the following issues: first, to analyze the factors influencing the number of Olympic medals and to build an accurate prediction model; second, to provide prediction intervals for the total number of gold medals and medals won by each country at the 2028 Los Angeles Olympic Games; third, to recommend sports events that should be prioritized for investment for three countries.

Firstly,we used the **independent samples t-test to identify** the host. Subsequently, feature extraction was conducted, including the number of participating athletes, the number of events entered, whether the country is the host, and the overall sports level of the year as feature variables, and the **zero-inflated Poisson regression model** was employed for prediction. A preliminary prediction of the medal outcomes for the 2028 Olympics was made. To identify the events on which each country relies, the **Spearman correlation coefficient** between the number of events hosted and the number of medals won was calculated.

Besides, **correlation analysis** led to the introduction of the number of athletes participating in specially advantageous events as a new feature variable.

Additionally,the **random forest quantile regression method** was used to re-predict the medal outcomes for each country at the 2028 Olympics, and a **sensitivity analysis** of the model's **hyperparameters** was conducted, achieving high fitting accuracy and good **robustness.** This model was applied to predict the prediction intervals and confidence levels for the number of Olympic medals for each country. For the probability of countries that have never won a medal to win their first medal, the **random forest quantile regression model** of countries that won their first medal in 2024 was used for **model fusion**, resulting in the probability of winning at the 2028 Olympics.

Furthermore, we analyzed the impact of the "great coach" effect on the number of medals through an **independent samples t-test** and found it to be highly significant. Secondly, to quantify the change in the number of medals after the addition of a "great coach," we selected the national competition medal data coached by Lang Ping and Bella Karolyi as training samples and constructed a **random forest regression model**, which was then fused. Using the optimized random forest model, we predicted the number of medals won by China, South Korea, and Spain in specific dependent events with and without the guidance of a "great coach," and recommended the three projects with the greatest potential for improvement as key investment areas.

In summary, this study could help the IOC to effectively allocate resources and optimize competition preparation strategies.

Keywords: medal prediction;zero-inflated poisson regression;quantile regression forests;independent samples t-test;correlation analysis

Contents

1 Introduction

1.1 Problem Background

During the Paris Summer Olympics, the United States topped the medal table with a total of 126 medals, tying with China for the lead in gold medals, each winning 40 golds. The host nation, France, ranked fifth in the gold medal count but fourth in the total medal tally. Britain ranked seventh with 14 gold medals and third in the total medal count.

In addition to the aforementioned top-ranked countries, Albania, Cape Verde, Dominica, and Saint Lucia, among others, won their first Olympic medals at these Games, with Dominica and Saint Lucia each securing a gold medal. There are still over 60 countries that have not yet won an Olympic medal.

The prediction of Olympic medal tallies is typically based on the current athletes' competition plans rather than historical medal counts. Olympic medal tallies are often seen as an important indicator of a country's sports strength and competitive level. Predicting medal counts can assess a country's status and strength on the international sports stage, which is of great significance.

The 2028 Los Angeles Olympics(Olympic Logo shown in **Figure 1.1.1**) in the United States is becoming the focus of worldwide attention! More and more people are concerned about how to scientifically predict the exciting changes that might occur in the medal table of the next Olympic Games.

Figure 1.1.1:The logo for the 2028 Los Angeles Olympics in the United States.

1.2 Statement of the Problem

To address this issue, I has been tasked with developing a model to predict the number of medals (including gold medals and the total number of medals) that countries will win in

future Olympic Games. The model and its analytical methods must meet the following requirements:

• **problem 1：** The model should be capable of estimating the accuracy of its predictions and measuring its performance. The predictions should include the following:

1.Predict the medal table for the 2028 Summer Olympics in Los Angeles, USA, and provide a prediction interval. Analyze which countries are likely to improve their performance and which may fare worse than in 2024.

2.Predict how many countries will win their first medal at the next Olympics and provide the probability of this prediction.

3.Investigate the relationship between the number and type of Olympic events and the number of medals won by countries. Analyze which events are most important for different countries and how the selection of events by a country affects medal outcomes.

• **problem 2：** Seek evidence of changes that may be caused by the "great coach" effect and estimate the impact of this effect on the number of medals. Additionally, select three countries and determine which sports they should consider investing in "excellent" coaches for, and estimate the impact of such investments.

• **problem 3：** Reveal unique insights regarding Olympic medal counts and explain how these insights provide valuable information for national Olympic committees.

1.3 Literature Review

In recent years, the prediction of Olympic medals has become a significant topic in the field of sports science. This paper systematically reviews the progress of relevant research, including Wang Fang's (2019) use of the Cobb-Douglas production function and neural network methods to predict the medal standings of seven major sports powers at the 2020 Tokyo Olympics, as well as Shi Huimin et al.'s (2024) application of the random forest model and SHAP method to assess the predictability of gold and other medals in different events, along with an in-depth analysis of the impact of socio-economic factors on medal performance.

However, predicting Olympic medals involves the interaction of multi-dimensional and complex factors, making it a highly challenging research task. Yang Jinghan (2018) combined time series models with related factor regression models to forecast the medal table for the 2020 Olympics, effectively integrating the strengths of different models to enhance the reliability of predictions. The research reveals significant differences in the predictability of various Olympic events, with sports like table tennis, badminton, and swimming showing

higher predictability, while events such as water polo, modern pentathlon, and volleyball exhibit lower predictability. Additionally, the traditional strengths of teams, gender differences, and economic development levels all have a significant impact on medal predictions.

Given the deficiencies in accuracy and reliability of existing research methods, future research needs to further explore more scientific, accurate, and reliable prediction methods. To this end, our team is dedicated to constructing a more rigorous and precise mathematical model, aimed at providing strong support for the preparation for the Olympics and the sustainable development of competitive sports.

1.4 MY Work

My work for this problem is shown in **Figure 1.4.1**.

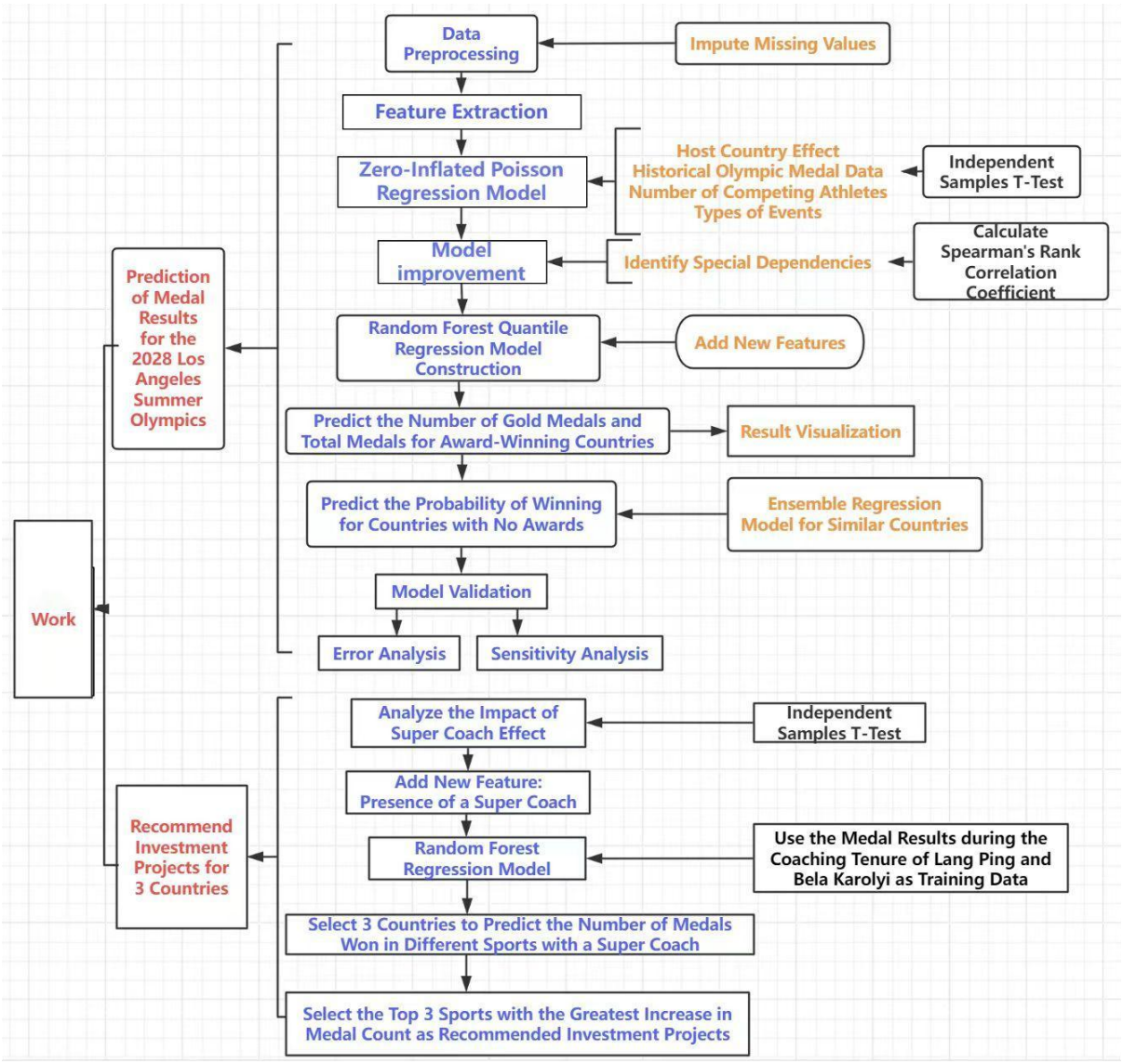

Figure 1.4.1: Workflow

2 Assumptions and Justifications

To simplify the problem, we make the following assumptions, each of which is thoroughly justified.

• Hypothesis 1: The number of events and participants from various countries in the new sports added to the next Olympics is highly correlated with the performance of active countries in the World Championships for these new sports.

Explanation: Due to the fact that new sports often require specific training facilities and professional talent, countries that participate more in the new sports at the World Championships are more likely to have these conditions in place, and therefore are more likely to participate in the new sports added to the Olympics.

• Hypothesis 2: When assessing a country’ s comprehensive strength at the Olympics, the average number of medals won in the past three Olympic Games is an effective indicator.

Explanation: An athlete’ s career typically does not span more than three Olympic Games, therefore the average number of medals over the past three Olympics can reflect the stable performance and competitive level of the country at the Olympic Games.

• Hypothesis 3: For different events within the same sport category, a country’ s competitive level is roughly the same.

Explanation: Different events within the same sport category typically share similar training methods, technical requirements, and competitive experience, hence a country’ s performance in these events tends to be consistent.

• Hypothesis 4: For countries that have never won a medal, if they manage to secure a medal in the next Olympics, their performance can be referenced against the predictive model trained on the data of countries that won their first medal in 2024.

Explanation: Countries that win their first Olympic medal often have similar development trajectories and backgrounds in competitive sports. Therefore, a model trained on the data of these countries can be used to predict the future performance of other non-medal-winning countries.

3 Notations

The key mathematical notations used in this paper are listed in **Table 1.**

Table 1: Notations used in this paper

Symbol	Description
$host_{ij}$	0 indicates that country i is not the host in year j. 1 indicates that country i is the host in year j.
$athletes_{ij}$	Represents the total number of participants from country i in the Olympics in year j.
$Sports_{ij}$	Represents the number of sports events participated in by country i at the Olympics in year j.
$Gold_{ij}$	Represents the total number of gold medals won by country i in the three previous Games by the year j.
$Medal_{ij}$	Represents the total number of gold, silver, and bronze medals won by country i in the three previous Games by the year j.

4 Model establishment and development

4.1 Data Description

4.1.1 Data Collection

Based on the provided data, it can be observed from the summerOly_athletes table that a total of 234 countries have had athletes participate in Olympic competitions. According to the data from the summerOly_medal_counts table, it is found that 164 countries have won medals at the Olympics. Furthermore, from the summerOly_programs table, we can learn that a total of 48 sports have been featured in the Olympics from 1896 to 2024.

Before conducting data analysis, we performed preprocessing on the data: First, we filled in the missing values in the summerOly_programs table with 0s; secondly, we converted numeric data from character type to numeric type and removed unnecessary spaces. Upon reviewing the summerOly_athletes table, we noticed that all sports events were not held in the years 1916, 1940, and 1944. After analysis, we speculate that this may be due to war reasons causing the Olympics to not be held as scheduled. Therefore, we decided to delete the data from these three years from the dataset and not include them in the subsequent analysis considerations.

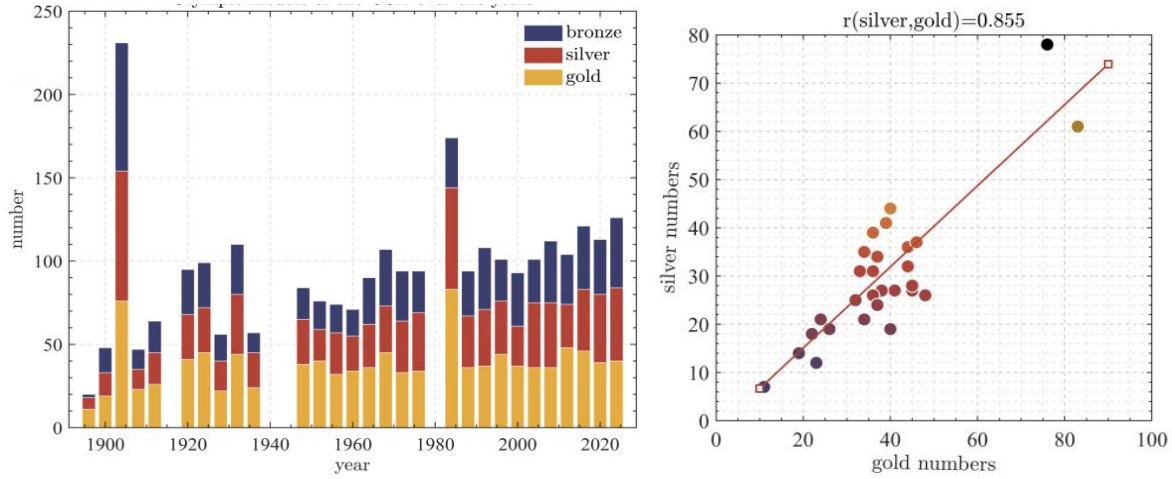

Figure 4.1.1.1 Correlation between the United States' medal count and the number of gold and silver medals.

Taking the United States as a case study(data shown in **Figure 4.1.1.1** and **Figure 4.1.1.2**), the following is a preliminary analysis of the distribution characteristics of Olympic medals:

1. The proportion of gold, silver, and bronze medals shows little difference.
2. 2.In the years when the United States serves as the host country, there is a significant increase in the total number of medals won.
3. There is a high correlation between the number of gold and silver medals won, indicating that the stronger a country's comprehensive strength, the greater the likelihood of winning more medals at the Olympic Games.

4.1.2 Data preprocessing

The following processing will be carried out on the collected data:

1. For the missing values, outliers, and special symbols in the provided table, conduct differentiated interpolation processing based on their distribution characteristics.
2. Correct classification errors such as "France-1", "France-2" to a single country name, for example, "France".
3. Remove data for countries that historically participated in the Olympics but no longer exist or do not have independent sovereignty.
4. Delete the relevant data records for the year 1906.

4.2 Feature extraction

4.2.1 Determining the impact of being the host country on the number of medals won.

To conduct statistical analysis on six representative countries (China, Japan, Australia, France, the United Kingdom, and the United States), an independent samples t-test method was used to determine the impact of being a host country on the number of medals won. **The specific steps are as follows:**

Firstly, the data were grouped by whether they had a super coach for different years, and a dummy variable (Dummy Variable) $host_{(i,j)}$ was defined. If the country was a host, $host_{(i,j)} = 1$; if not, $host_{(i,j)} = 0$. An independent t-test was conducted on $host_{(i,j)}$ and $mental_{(i,j)}$ to reflect whether there is a significant difference in the number of medals won by the host country compared to non-host countries[5].

The calculation results are as follows:

Table 2:result of Independence t-test

Variable value	Average value	Standard deviation	Welch's T-test
No	85.92	26.005	T=-2.209 P=0.109
Yes	154	60.756	T=-2.209 P=0.109

We examined and visualized the results of the independent t-test using violin scatter plots, which combine density distribution plots, box plots, and scatter plots. The green triangles represent the mean, the white horizontal lines represent the median, and the box plot

represents the interquartile range. It clearly shows from multiple dimensions that the upper and lower quartiles, mean, and median of the host's medals are significantly higher than those of the non-hosts. We tested and visualized the results of the independent t-test through violin plots.

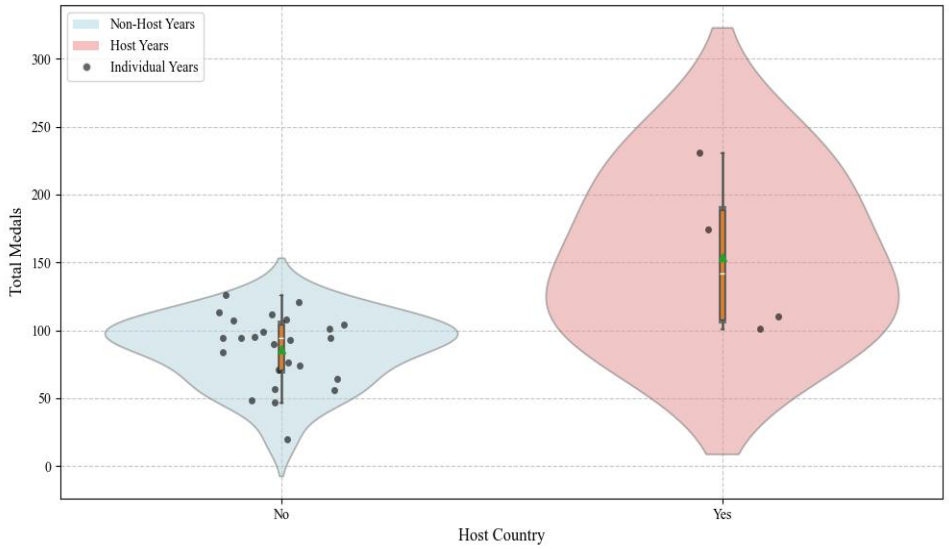

Figure 4.2.1.1:Scatter plot of the violin influenced by the American event organizers

4.2.2 Detailed feature extraction

To predict the medal counts for each country at the 2028 Los Angeles Olympics, we will extract predictive features for each country. These features include historical gold medal counts for measurement, whether the country is the host, and the number of event entries for each country in 2028. Specifically:

1. The total number of participants from each country in the Olympics;
2. The number of Olympic events each country participates in;
3. Whether the country is the host nation;
4. The average number of gold medals won by each country in the past three Olympic Games (if a country has participated in less than three sessions, the average is calculated based on the actual number of sessions in the three previous Games leading up to the current year);

In addition, we will count the number of gold medals and the total number of medals won by each country in various sports as the target values for training samples.

When we are extracting features from the 2028 data, it is easy to know the average number of medals from previous games and the host country, but it is difficult to predict the number of participants and events for each country. Through research, we learned that five new events will be added[6], and one event will be removed. Therefore, we reasonably speculate that the number of events and participants for the new events in the next Olympics will be consistent with the countries that are active in the new events at the World Championships. Assuming that the number of participants in the other events remains unchanged, by looking up the number of participants in each new activity, we can roughly determine the change in the five characteristics for each country.

4.3 Predicting with Zero-Inflated Poisson Regression

4.3.1 Model Building

Using Poisson regression for prediction, we can obtain the distribution of predicted values for the target variable and determine the prediction range under different confidence intervals. However, since some countries fail to win medals in certain years, the value of the target variable becomes zero, which clearly does not conform to the characteristics of the Poisson distribution. Therefore, we employ a zero-inflated Poisson regression for prediction, which adds a mechanism for generating zeros to the traditional Poisson regression[1].

Next, we will combine the four aforementioned characteristics with the target variable of the training samples to apply the Poisson regression model to predict the medal-winning situations of countries that have won medals at the Olympics in the 2028 Olympic Games. The structure of the model is as follows:

For cases where the target variable is zero, we will use the zero-generation part of the model to make predictions:

$$\ln\frac{\pi_i}{1-\pi_i} = y_0+y_1Z_{i1}+y_2Z_{i2}+y_3Z_{i3}+y_4Z_{i4} \quad (1)$$

Where π_i represents the probability predicted/observed as 0 by the logistic regression model，

$Z_{i1},Z_{i2},\cdots\cdots,Z_{i4}$represent the probabilities of the zero-inflation part,

$y_0,y_1,\cdots\cdots,y_4$ represent the regression coefficients of the zero-inflation part.

For the case where the target variable is non-zero, predictions are made using the count part:

$$\ln\lambda_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\beta_4X_{i4} \quad (2)$$

Where λ_i is the expected value of the target variable for the ith country,

β_0, β_1，……，β_4 represent the regression coefficients of the count part.

The final combined model:

$$P(G_{ij}=y_i)=\begin{cases}\pi_i+(1-\pi_i)poisson(0|\lambda_i) & if\ y_i=0\\ (1-\pi_i)poisson(y_i|\lambda_i) & if\ y_i>0\end{cases} \quad (3)$$

*The tr*aining process of the model is as follows:

Construct the likelihood function, which represents the joint probability that the predicted number of medals equals the actual number of medals given the regression parameters of the model. The regression coefficients are solved using the maximum likelihood estimation method.

The expression of the likelihood function is as follows：

$$L(\gamma,\beta|Y,X,Z)=\prod_{i=1}^{n}P(G_i=y_i|X_1,\ldots\ldots,X_4,Z_1,\ldots\ldots,Z_4) \quad (4)$$

Utilizing the gradient descent method to obtain the model parameters, we subsequently apply the same technique to train the prediction model for the total number of medals.

4.3.2 Model Evaluation

Taking the United States as an example, the following is the presentation of results: (where the red dashed line represents the scenario when the United States is the host country).

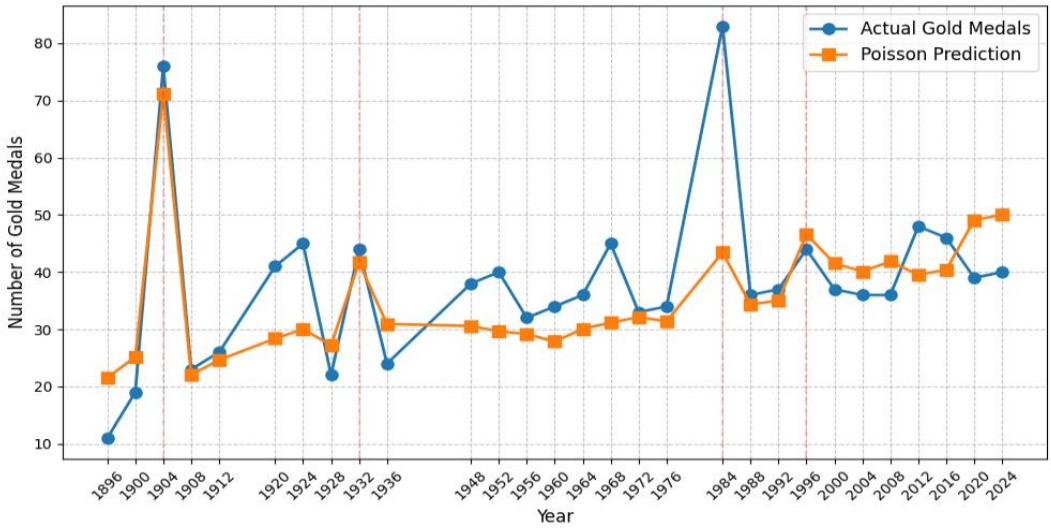

Figure 4.3.2.1:Comparison between the actual number of gold medals won by the United States and the Poisson predicted number of gold medals.

The observation results（data shown in **Figure 5**） indicate that the model's fitting effect is relatively satisfactory, with a correlation coefficient R value of 0.718. However, in certain years, the model's predictions are contrary to the actual trends, and the degree of fit could be improved. In future work, we will draw on the experience of this model and make corresponding optimizations and improvements to address its shortcomings.

Model Improvement

When using Poisson regression for prediction, it is necessary to meet the condition where the variance of the target variable is similar to its mean, but this may not be the case for the number of medals won at the Olympics. Poisson regression tends to compress the advantages in the prediction results, showing a conservative trend of raising the lower limit, which makes it more suitable for predicting low-probability events and deviates from the patterns of sports competitions. These factors may have led to the previous predictions being less accurate, and therefore we plan to use other methods for re-prediction.

In fact, some countries may have a high dependence on certain specific events, which could significantly impact their medal tallies, and this was not taken into account in our earlier models. This oversight may be one of the reasons for the inaccuracy of the prediction results. **To address this, we will take the following steps to re-predict:**

Firstly, we will identify the specific events on which each country heavily relies. In this process, we will use the Spearman's rank correlation coefficient for analysis. The Spearman's rank correlation coefficient is suitable for measuring the monotonic relationship between two variables, with a value range from -1 to 1. The closer the coefficient value is to 1, the stronger the monotonic increasing relationship between the two variables[4]. We will calculate the Spearman's rank correlation coefficient between the number of medals won by each country in past Olympic Games and the number of competitions held for various sports. A higher correlation coefficient indicates a stronger positive correlation between the number of competitions for an event and the number of medals won by the country.

Use a 164x30 matrix named Country to represent the medal information of each country in every Olympic Games:

$$\text{Country}=[C_1, C_2, \ldots\ldots, C_i, \ldots\ldots C_{164}]^T \quad (5)$$

$$\text{Among } C_i=[c_{i,1}, c_{i,2}, \ldots\ldots, c_{i,j}, \ldots\ldots c_{i,30}]$$

Where $c_{i,j}$ represents the number of medals won by the i-th country in the j-th Olympic Games

Use a 48x30 matrix named Sport to represent the number of competitions held for various sports at each Olympic Games:

$$\text{Sport}=[S_1, S_2, \ldots\ldots, S_k, \ldots\ldots S_{48}]^T \quad (6)$$

$$\text{Among } S_k=[s_{k,1}, s_{k,2}, \ldots\ldots, s_{k,j}, \ldots\ldots s_{k,30}]$$

Where $s_{k,j}$ represents the number of events for the kth sport held in the jth Olympic Games

Calculate the Spearman correlation coefficients between X1 to X164 and Y1 to Y48, and store the results in a 48x164 matrix named Spearman：

$$\text{Spearman}=[P_1, P_2, \ldots\ldots, P_i, \ldots\ldots P_{164}] \quad (7)$$

$$\text{Among } P_i=[p_{i,1}, p_{i,2}, \ldots\ldots, p_{i,k}, \ldots\ldots p_{i,48}]^T$$

Where $p_{i,k}$ represents the Spearman correlation coefficient between the number of medals won by the ith country and the number of events held for the kth sport.

The following is a visualization of some of the information in the Spearman matrix, presented as a heatmap. The data in each cell represents the Spearman correlation coefficient between the number of medals won by the country on the horizontal axis and the number of events held for the sport on the vertical axis. The closer the color is to yellow, the stronger the positive correlation, and the closer to purple, the weaker the correlation.

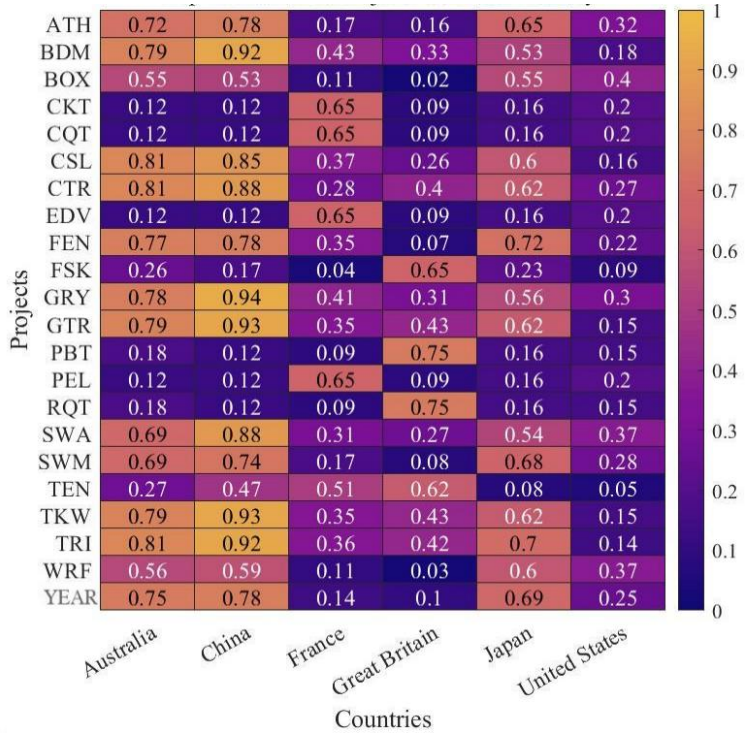

Figure 4.4.1 Heatmap of countries' proficiency in a particular sport

We will set the threshold to 0.8. If ρi,k is greater than 0.8, it is considered that the k-th sport is a specific strength event for the i-th country. After identifying the strength events for the 164 countries that have won medals, we will count the number of participants from these countries in their specific strength events at each Olympic Games. This count will be used as a new feature for subsequent model training.

4.4 Predicting with Random Forest Quantile Regression

4.4.1 Model establishment

After introducing the new feature——the number of participants in specific advantageous events, we employ the Random Forest Quantile Regression method to predict the medal outcomes for countries that have won medals in previous Olympics at the 2028 Olympic Games. **The specific steps are as follows:**

Firstly, we select five features and the corresponding actual medal counts for each country in each Olympic Games to construct the sample set, and then divide this sample set into a test set and a training set.

Random Forest is an ensemble model composed of multiple decision trees, where each decision tree is built by training on a randomly sampled subset of the training data. After training, each decision tree will make medal predictions based on the five features of the test set according to the rules it has learned.

Subsequently, we determine the final prediction result by calculating the mode of all the decision trees' prediction results. Moreover, based on the prediction results of multiple decision trees, we can calculate the predictive distribution[2] at a given confidence level and compare this distribution with the actual values to obtain error parameters. Therefore, once we have the five feature data for 2028, we will be able to predict the medal count more accurately.

This method not only provides the conditional distribution of the target variable, rather than just mean prediction, which meets the requirements of implementing multivariate regression analysis for this problem. It also offers prediction intervals at different levels of uncertainty, which is why we have decided to use this model for prediction.

The specific steps for this problem are as follows:

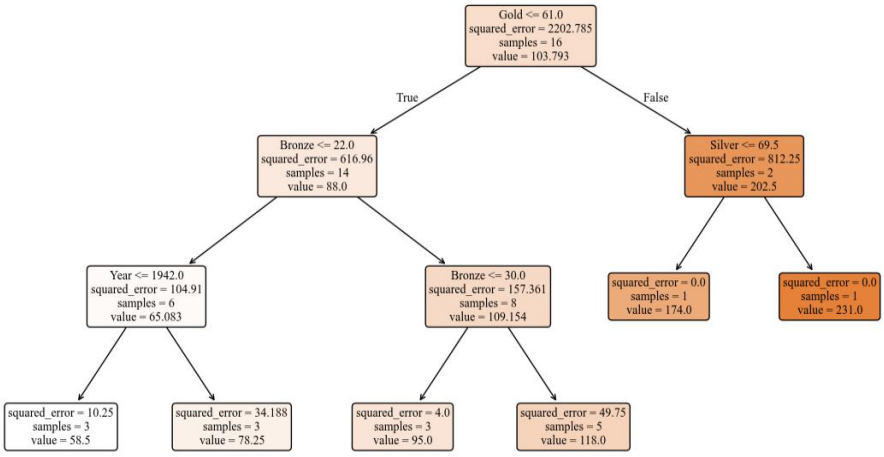

Figure 4.5.1.1 Diagram of the Random Forest Structure

1.Organize training data, use a 6-dimensional vector to store the 5 characteristics of each country along with the number of gold medals and total medals won.

$$x_{ij}=[athletes_{ij}, Sports_{ij}, host_{ij}, Gold_{ij}, advantage_{ij}] \quad (8)$$

2. Building a Decision Tree: First, perform bootstrap sampling (Bootstrap), randomly select several samples from the training data to form a sub-sample set. Second, feature random selection: at each node, randomly select two features from all features for splitting. Finally, recursively split until the tree reaches the preset maximum depth.

3. Quantile Estimation: Traverse the leaf nodes of each tree, collect the target values of the training samples in the leaf nodes, and form a collection $G_b(x_{ij})$

4. Integrate Quantiles: **Combine all the values of $G_b(x_{ij})$ to form a collection:**

$$G(x_{ij})=\cup_{b=1}^{B} G_b(x_{ij}) \quad (9)$$

5. Then sort the collection for each target variable and find the quantiles of τ-th quantile **as the final prediction:**

$$Q_T(G|X_{ij}=x_{ij})=\text{quantile}(G_{(x_{ij})}, T) \text{ 其中 } G(x_{ij})=\cup_{b=1}^{B} G_b(x_{ij}) \quad (10)$$

4.4.2 Model Evaluation

To ensure that the fluctuation range of the final gold medal prediction result is within 5 medals, we define the confidence level as 38% after observing the output results through multiple attempts. Taking the number of gold medals won by the United States and Spain as an example, we present the results:

the red solid line represents the actual number of gold medals won by the United States in each Olympic Games, the blue dashed line represents the model's median estimate of the gold medals won, and the three different shades of color represent different confidence levels. The darkest color represents a confidence level of 38%, the medium color represents a confidence level of 60%, and the lightest color represents a confidence level of 90%.

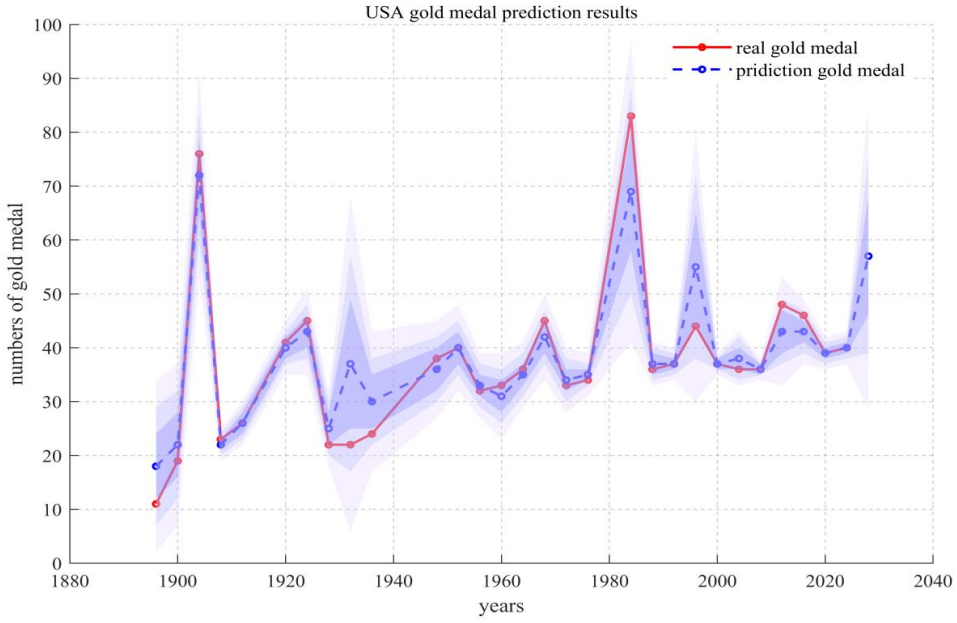

Figure 4.5.2.1 Prediction results of the United States' gold medal count

Evaluation results of the US regression model:

1. Mean Squared Error (MSE): 0.80; Root Mean Squared Error (RMSE): 0.89; Coefficient of Determination (R^2): 0.91

2. Feature importance: Number of participants: 0.2996, Number of events: 0.0923, Host status: 0.0992, Strength rating: 0.1922 Number of participants in dominant events: 0.3167

3. Mean Absolute Error: 0.32, Mean Relative Error (excluding years with zero gold medals): 14.17%

Combining the analysis results, we believe that the true curve falls within the 38% confidence interval, indicating a good fit of the model.

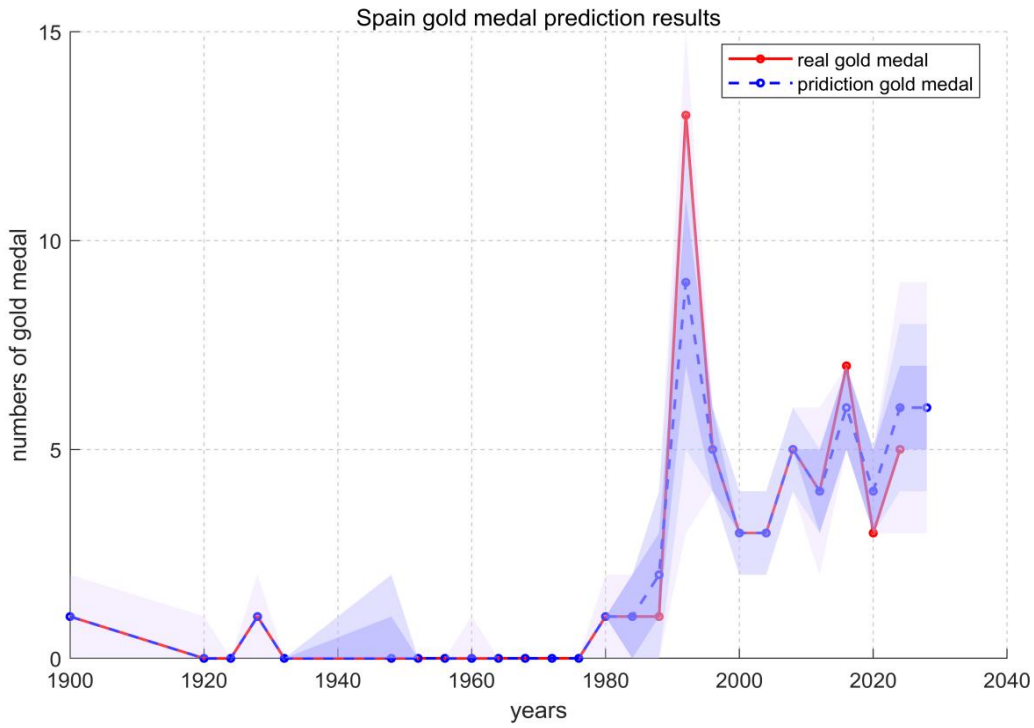

Figure 4.5.2.2 Prediction results of Spain’s gold medal count

Evaluation results of the Spanish regression model:

1. Mean Squared Error (MSE): 134.76; Root Mean Squared Error (RMSE): 11.61; Coefficient of Determination (R^2): 0.91

2. Feature importance: Number of participants: 0.6080; Number of events: 0.1400; Host status: 0.1441; Strength rating: 0.1079; Number of participants in dominant events: 0.0

3. Mean Absolute Error: 7.10; Mean Relative Error (excluding years with zero gold medals): 10.38%

Combining the analysis results, we believe that the true curve falls within the 38% confidence interval, indicating a good fit of the model.

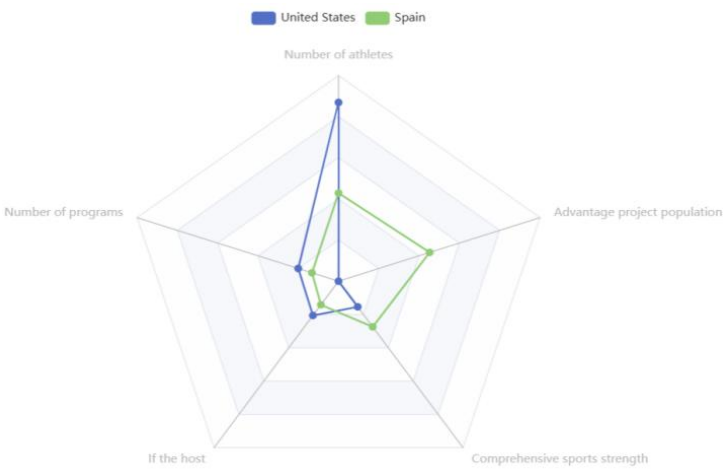

Figure 4.5.2.4 prize dimension comparision

To visually compare the differences in the analysis of medal influencing factors between two countries, we can use radar charts to clearly show that the United States, as a sports powerhouse, has a significant advantage in medal acquisition due to its large number of athletes, and the U.S. has a balanced development across various sports without a severe dependence on any single event. In contrast, countries like Spain, which have relative strengths in specific events, heavily rely on the performance of these advantageous events to win more medals.

It can be observed that the model’s prediction accuracy is very high and the analysis is comprehensive. By applying it to all countries, we can obtain the predicted results for the 2028 gold and total medals. Here, we present the top 10 countries with the highest predictions:

Table 3：prediction of 2028 gold medals

country	Gold medal range	Total medal range
United States	[46, 68]	[101.0, 174.0]
China	[39, 41]	[89.0, 91.0]
Great Britain	[20, 24]	[89.0, 91.0]
Japan	[15, 19]	[40.0, 45.0]
Australia	[16, 18]	[50.0, 53.0]
Germany	[12, 20]	[33.0, 68.0]
France	[11, 15]	[33.0, 64.0]
Netherlands	[11, 13]	[33.0, 36.0]
Italy	[10, 12]	[27.0, 40.0]
Canada	[8, 9]	[24.0, 27.0]

4.5 Predicting for Non-Winning Countries

Due to the lack of training data for countries that have not won any medals, we cannot directly establish a medal prediction model using the random forest regression method previously employed. To address this issue, we assume that the situations of the four countries that won medals for the first time in 2024 (Albania, Cabo Verde, Dominica, Saint Lucia) are similar. Referring to their random forest prediction models, we incorporate the

characteristics of each country that has not won a medal in 2028 into the data from 1896 to 2024 of these four countries to train a comprehensive prediction model.

The stacked method combines the quantile models of the four countries:

To fully utilize the model information of Albania, Cabo Verde, Dominica, and Saint Lucia, and to reduce the impact of factors of inter-country variability on the model's general applicability, we separately use the random forest quantile regression models of the four countries for prediction, and then take the average to obtain the final results.

By statistically analyzing the prediction distribution of each decision tree, we calculate the probability of these countries winning one or more medals in 2028, which is the probability of winning medals in the next competition. We present the results of the top ten countries with the highest probability of winning new medals:

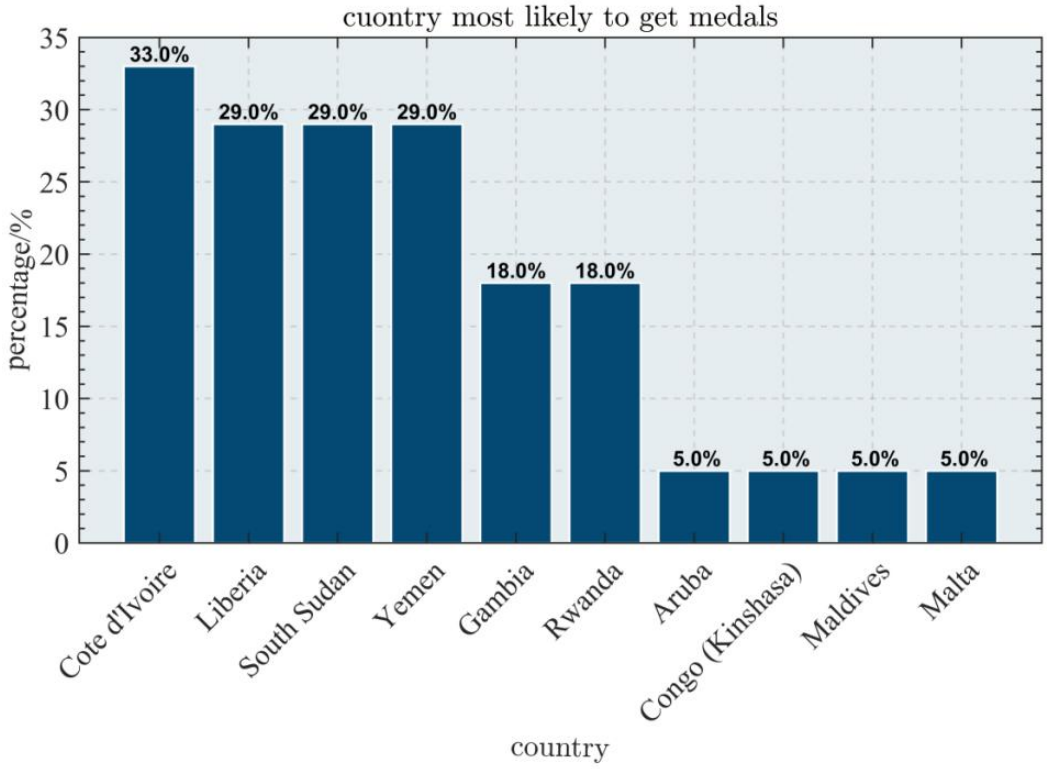

Figure 4.6.1 The country most likely to win its first medal

4.6 Analysis of Results

4.6.1 Analysis of changes in the number of awards won

Based on the previously established random forest quantile regression model, the predicted medal counts for each country in 2028 can be obtained. After comparing with the medal counts from 2024, it is easy to identify the changes in the number of medals for each country. **Here, the six countries with the most significant changes in medal counts are selected for display:**

Table 4：Increase Notations

country	number	increase
Germany	44.0	11.0
Jamaica	9.0	3.0
United States	129.0	3.0
Ethiopia	6.0	2.0
Kazakhstan	9.0	2.0
Cuba	11.0	2.0

Table 5：Decrease Notations

country	number	decrease
France	54.0	-10.0
Italy	37.0	-3.0
United States	3.0	-2.0
Ethiopia	63.0	-2.0
Kazakhstan	4.0	-2.0
Cuba	43.0	-2.0

Most obviously, it is easy to see that because the next host is the United States, and this time the host is France, due to the host factor, France will decline, while the United States will make progress.

4.6.2 Model Sensitivity Test

To verify the stability and parameter dependency of the model, we took the gold medal count prediction model for the United States as an example to conduct sensitivity analysis. We selected the two most important hyperparameters for the random forest, namely, the depth of each decision tree and the number of decision trees in the forest, and **analyzed them as shown in the following figure:**

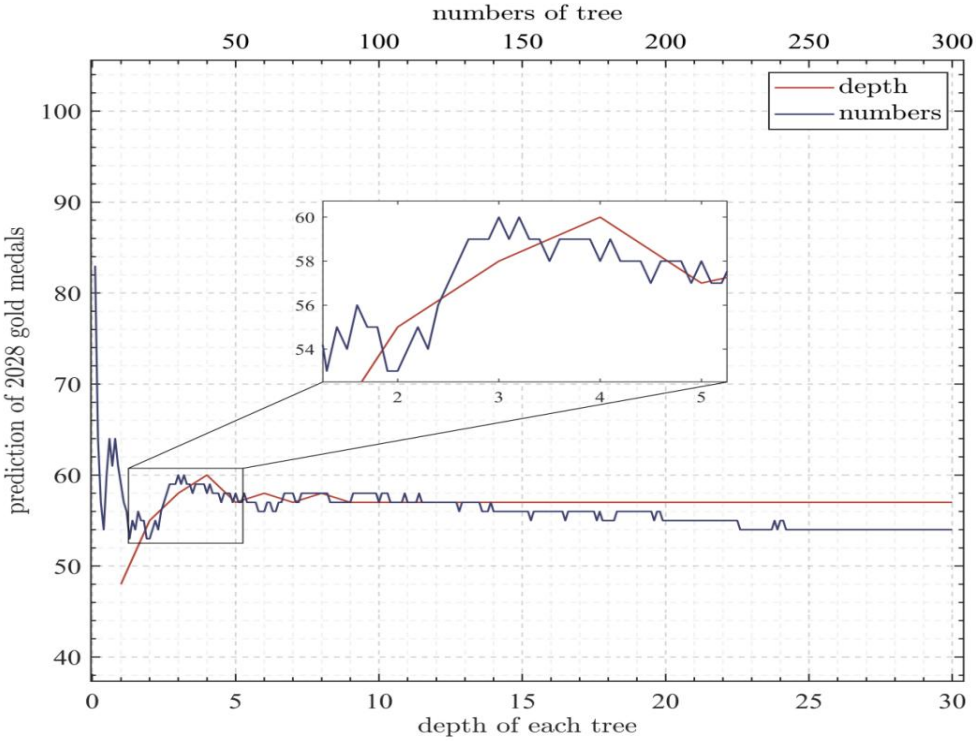

Figure 4.7.2.1 Sensitivity Analysis of Predicted US Gold Medal Counts

From the observation of the image, it can be seen that when the depth of the trees is low, the wisdom learned by each tree is not sufficient to make correct decisions; when there are too few decision trees in the forest, it is difficult to conduct comprehensive analysis, leading to one-sided decisions. However, the model converges very quickly, and the prediction results of the model no longer change and converge around 57 when the depth of the decision trees is greater than 5 or the number of decision trees is greater than 50. This indicates that the model is not sensitive to hyperparameters and has very good robustness.

5 Analyzing the Impact of the great coach Effect

5.1 Data Description

Based on the conditions provided in the question, volleyball coach Lang Ping and gymnastics coach Lá Károly are two possible examples. We will use the Chinese women's volleyball team coached by Lang Ping as a case to analyze the impact of Lang Ping's coaching on the team's medal achievements. We will collect the annual medal records of the Chinese women's volleyball team from the summerOly_athletes table. The scoring will be set to 3 points for gold medals, 2 points for silver medals, 1 point for bronze medals, and 0 points for no medal[7].

$$\text{score}_j=\begin{cases}0,\ Indicate\ th\ at\ th\ e\ j-th\ \ Olympic\ Games\ did\ not\ win\ medals\\ 1,\ Indicate\ th\ at\ th\ e\ j-th\ \ Olympic\ Games\ won\ a\ bronze\ medal\\ 2,\ Indicate\ th\ at\ th\ e\ j-th\ \ \ Olympic\ Games\ won\ a\ silver\ medal\\ 3,\ Indicate\ th\ at\ th\ e\ j-th\ \ Olympic\ Games\ won\ a\ golden\ medal\end{cases} \quad (11)$$

Define a dummy variable: Lang Ping coached the Chinese women's volleyball team at the Olympics in 1996, 2016, and 2020, respectively.

$$\text{Coach}_{ij}=\begin{cases}0,\ the\ i-th\ country\ did\ not\ have\ a\ \ coach\ in\ the\ j-th\ Olympic\ Games\\ 1,\ the\ i-th\ \ country\ had\ a\ \ coach\ \ in\ the\ j-th\ \ Olympic\ Games\end{cases} \quad (12)$$

Conduct an independent t-test on score_j and Coach_{ij}.

5.2 The Establishment of Model for Problem Two

5.2.1 Using t-test to determine the impact degree of the super coach.

The mean number of medals with and without a great coach are: 1.099/2.526: Due to the failure to meet the homogeneity of variances, Welch's T-test was used, with a significance level of P = 0.0001. Therefore, the statistical result is significant, indicating a significant difference in the number of medals with and without a great coach; the magnitude of the difference is Cohen's d = 1.249, which is very large (values of 0.20, 0.50, and 0.80 correspond to small, medium, and large effect sizes, respectively).

The following is a display of the calculation results:

Table 6:result of Independence t-test

Variable value	Average value	Standard deviation	Welch's T-test
0.0	1.099	1.241	T=-7.875 P=0.0001
1.0	2.526	0.513	T=-7.875 P=0.0001

The results of the independent t-test indicate that the Great Coach Effect has a significant impact on the number of medals won. Next, we will further validate the independent t-test and provide a visual representation. In the box plot above, the red line represents the median number of victories for China with and without a great coach. It can also be seen that the contribution of the great coach to the improvement of the Chinese women's volleyball team is very significant.

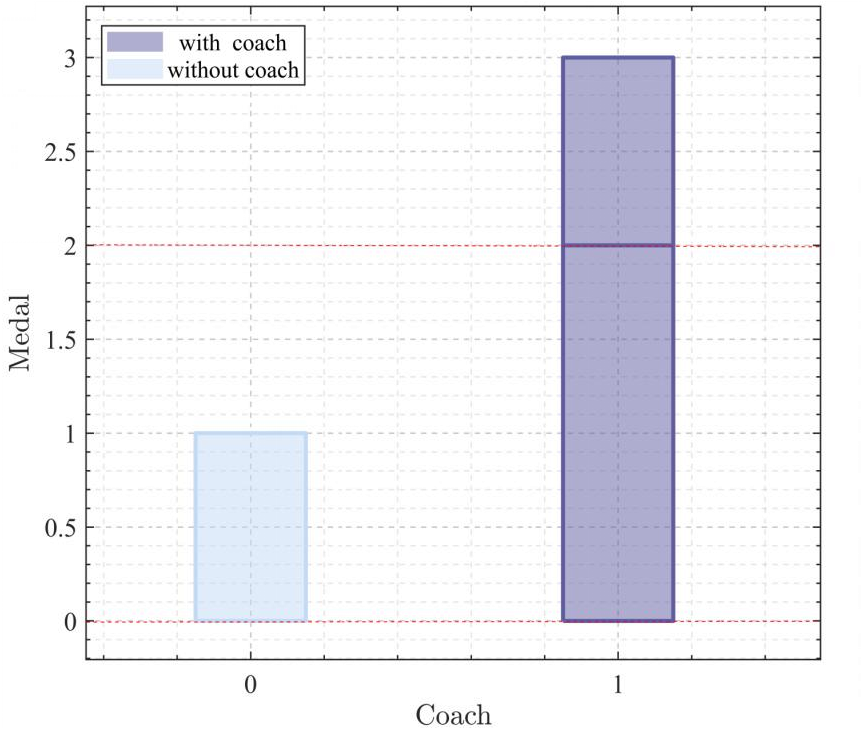

Figure 5.2.1 The impact of a great coach on the medal count of the Chinese women's volleyball team.

Next, we will use the medal counts of the countries coached by the great coaches Lang Ping and Bella Karolyi in the respective events as training data[8] to establish a random forest regression model. This will quantify the impact of having a super coach on the number of medals won.

5.2.2 Feature extraction

First, prepare the training data, taking the Chinese women's volleyball team as an example. There are a total of 4 features: the number of women's volleyball team participants, whether China is the host, the comprehensive strength of Chinese volleyball, and whether Lang Ping is coaching. The specifics are as follows:

1. $athletes_{(china,\ j)}$: represents the total number of participants in women's volleyball at the jth Olympic Games from China.

2. $host_{(china,\ j)}$: indicates whether China is the host of the jth Olympic Games, where 1 means yes and 0 means no.

3. $medal_{(china, j)}$: represents the average number of medals won in the women's volleyball event in the first three participations at the jth Olympic Games (if less than 3 years, calculate based on the actual number of medals won).

4. $coach_{(china, j)}$: indicates whether there is a super coach for China at the jth Olympic Games, where 1 means yes and 0 means no. The target variable is: M(china, j): represents the total number of medals won by the Chinese women's volleyball team in the jth year.

5.2.3 Random Forest Regression Model Establishment

The training process of the Random Forest regression model is similar to that of the first question, simply requiring adjustments to the input features and target variables. No further explanation is needed here.

Another super coach, women's gymnastics coach Bella Karolyi, coached in the United States during the 1984, 1988, and 1992 Olympic Games. We have collected the medal records of the U.S. women's gymnastics team at the Olympics up to and including 1992. The feature extraction method is the same as that used for the Chinese women's volleyball team, and a Random Forest regression model is trained in the same manner.

Referencing the previous method of integrating the Random Forest regression models of Albania, Cabo Verde, Dominica, and Saint Lucia, the above two models are combined to obtain a prediction model that considers the impact of a super coach on a country's medal count in a specific sport.

5.2.4 Select three countries for prediction.

We have obtained a prediction model for the number of medals won in individual sports events with or without a super coach. Now, we will analyze China, South Korea, and Spain. In Section 4.4, we identified the sports events that have a significant impact on the total number of medals won by each country. Next, we will use the refined random forest model to predict the number of medals won in these countries' particularly dependent events with and without a super coach, and calculate the difference. This difference represents the predicted increase in the number of medals for that event after hiring a super coach, reflecting the size of the return on investment in that event. For these three countries, we will select the three events with the largest differences as the recommended events for them to invest in.

5.3 Result Analysis

For China, investments should be considered in Gymnastics, Triathlon, and Handball. For South Korea, investments should be considered in Basketball, Gymnastics, and Canoeing. For Spain, investments should be considered in Gymnastics, Aquatics, and Field Hockey. The predicted increase in the number of medals is presented in the table below.

Table 7：Decrease Notations

country	sports	increase number of medals
China	Gymnastics	6
China	Triathlon	5
China	Handball	5
South Korea	Basketball	6
South Korea	Gymnastics	6
South Korea	Canoeing	5

Spain	Gymnastics	6
Spain	Aquatics	5
Spain	Field hockey	5

6 The model's unique insights into Olympic medal counts

6.1 Olympic medal counts reflect the modernization differences among countries.

Many non-Western countries were unable to participate independently in the early Olympics due to historical reasons. Based on this, we innovatively incorporated the year into the correlation analysis of event prediction and conducted a comparative analysis of the correlation over time between the People's Republic of China, Australia, Japan, and the United States of America, the French Republic, and the United Kingdom of Great Britain and Northern Ireland. It is easy to conclude that the Olympic year has a relatively small impact on Western countries, while it has a significant impact on emerging countries. Through the trend of Olympic medals, we can observe the differences in the modernization process of different countries.

6.2 Olympic performance is related to a country's economic level.

We found that countries that rely solely on a single event as their strength are generally located in economically underdeveloped regions, and their past medal tallies are mostly around zero. Based on this, we can roughly conclude that there is a correlation between a country's Olympic medal count and its Gross Domestic Product (GDP). Economically underdeveloped countries need to focus on their own development issues first, rather than participation in international Olympic events. Some underdeveloped regions eager to participate in the Olympics often end up relying on a single event for a long time due to a lack of funds. This phenomenon reminds us that while we pay attention to the countries at the top of the Olympic medal table, we should not neglect the sports competitive situation in underdeveloped areas. For major countries, they should actively provide peaceful humanitarian assistance.

6.3 Politics influences Olympic participation and the medal table.

When statistical data was compiled, we found that the United States typically participates in the Olympics, but there have been absences in certain years, such as 1916, 1940, 1944, and 1980. After consulting historical records, we learned that the first three absences were due to the cancellation of the Olympics because of World War I and World War II. As for 1980, it was because the United States, along with 64 other countries, boycotted the Moscow Olympics in protest against the Soviet Union's invasion of Afghanistan, which was considered an unjust war. Additionally, there are some countries that no longer participate in the Olympics after certain years because they have ceased to exist or

have disintegrated. This shows that the composition of the Olympic medal table is also influenced by political factors. This phenomenon reminds countries that they should always remember the peaceful spirit of the Olympics and jointly maintain the fairness and justice of international sports competition.

7 Model Evaluation and Further Discussion

7.1 Strengths

7.1.1 Using the random forest quantile regression prediction method.

The notable advantage of this model is that the entire dataset in the sample collection is often not necessary, as these data may obscure the significance of features to a certain degree. The random forest algorithm, by randomly sampling, enables each decision tree to make independent judgments based on the features it has learned, thereby achieving a more profound and accurate multi-level analysis. Compared to the traditional random forest model, the model we employ is capable of producing a distribution interval. By analyzing the distribution patterns at different confidence levels, the prediction results are made more reliable and rigorous. After gaining a deeper understanding of the model, we discovered a similarity to the American jury system in this approach: it involves randomly selecting different individuals to examine issues from their respective perspectives and ultimately synthesizing various opinions, thereby revealing the essential truth of matters more objectively and comprehensively. This reflects the cutting-edge nature and broad applicability of the model's conceptual framework.

7.1.2 Model fusion was performed when predicting non-award-winning countries.

For countries that have not won any medals in history, the prediction task is challenging due to the zero medal count. Innovatively, we used the four countries that won medals for the first time in 2024 as a reference for prediction. The model training results for these countries are closer to the actual situation of countries that have not won medals. Considering that each country has its uniqueness, we fused the models of these four countries to eliminate the influence of specific factors, thus achieving a better fitting effect. Moreover, we also calculated the probability of each country winning more than one medal, which allows us to predict which countries are likely to achieve a breakthrough in their medal tally.

7.1.3 This model has good comprehensive performance.

Based on the chart analysis, it can be observed that the medal prediction model has a high degree of fit and good smoothness, with a high determination coefficient R value of 0.91. Additionally, the model demonstrates strong robustness and rapid convergence speed. Furthermore, we can also utilize this model to analyze the proportion of importance of the five characteristics for different countries, thereby indirectly revealing the comprehensive situation of each nation.

7.2 Weaknesses

1. When discussing dominant events, a more refined classification of events can be

made, such as distinguishing between field events and track events in athletics.

2. When conducting medal predictions and data processing, it is necessary to differentiate between team events and individual events to ensure the accuracy of predictions and the appropriateness of processing methods.

7.3 Further Discussion

1. When constructing the features of our model, more factors can be included, such as: the year the event is held, the geographical distance from the host country, the gender ratio of athletes in the team, the distinction between individual and team events, and the overall quality of substitute players, etc.

2. In the prediction process, a variety of models can be used for regression analysis, and a broader strategy for model fusion can be implemented.

3. In predicting outstanding coaches, we only considered the two coaches mentioned in the question. However, the number of high-level coaches in reality should be more numerous. If all relevant samples are taken into consideration, the prediction accuracy of the model is expected to be further improved.

References

[1] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1–14.

[2] Meinshausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research, 7(6), 983–999.

[3] https://www.olympics.com/zh/sports/

[4] Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13(1), 25 – 45. https://doi.org/10.1093/biomet/13.1.25

[5] Fagerland, M. W., & Sandvik, L. (2009). A comparison of the independent samples t-test and the Mann-Whitney U test for non-normal data. BMC Medical Research Methodology, 9(1), 1 – 10. https://doi.org/10.1186/1471-2288-9-42

[6] https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/).

[7] https://olympics.com/en/athletes/ping-lang

[8] https://usagym.org/halloffame/inductee/coachingteam-bela-martha-karolyi/