Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2512002

2028 Olympics Medal Prediction Model and the Secrets Behind It

In addition to watching individual events during the Games, spectators also pay attention to the overall “medal standings” of each country. While the top positions in the standings are undeniably watched closely, the number of medals won by certain countries is also emphasized. The aim of this report is to develop predictive models to forecast the number of medals won by each country at the 2028 Olympic Games and whether there is a “great coach” effect. The following four models were developed: Model 1: A model for predicting the number of medals based on random forests；Model 2: Logistic regression model to predict first medal win；Model 3: A regression-based model exploring the relationship between the number of Olympic events and the number of medals won by each country； Model 4: The “Great Coach” problem model.

For model 1, we first collected data of all types of countries from 2000 to 2024. Then we analyzed the data based on the intrinsic connection to determine the relevant variables that affect the number of national medals, and then we used the Random Forest plus bootstrap integrated prediction model, and finally predicted the medal list and prediction intervals of the 2028 Olympic Games, the top 30 results are shown in Figure 7.

For model 2, we are building a probabilistic prediction model based on a logistic regression model. We train the prediction model with the help of the data of countries that have won awards in the past, from the three characteristics of the number of athletes, the number of participating programs, the number of programs opened, and finally we get that there will be 11 countries that will win the first medal.

For Model 3, we first conducted an analysis of the collated data to find out which events each country is better at, and then carried out various regression analyses and comparisons to conclude that cubic polynomial regression analysis is more effective, as shown in Figure 10. Next, the number of medals won by each country in each sport was statistically analyzed to find out which sport was the most important for each country, and the effect of the addition of new sports by the host country on the number of medals was examined, as shown in Figure 13.

For Model 4, we determine the existence of the “great coach” effect based on the data, and then take Lang Ping as an example to prove that the existence of the great coach has an effect on the number of national medals through a comprehensive evaluation model based on OLS regression and so on. Then, we screen out the items that have basically no fluctuation on the medal counts of the selected three countries through CUSUM, and substitute them into the evaluation model to get the effect of increasing about 1~2 medals.

In addition, we analyzed the results predicted by our model and found that it revealed some mysteries behind the Olympic Games, such as the decreasing probability of winning prizes for large sports countries and the increasing probability of winning prizes for small ones. We have rationalized the reasons for these mysteries, and based on them, we have made two recommendations to the IOC.

Finally, we conducted a model performance analysis of the model, and we found the possible overfitting or underfitting of the model under some specific conditions. We then optimized the parameter settings of the model to improve the accuracy of the prediction, and the standard deviation of the prediction results is shown in Figure 15.

Keywords: Olympic medal; Random Forest Prediction; logistic regression; host country

Contents

1 Introduction

1.1 Problem Background

As the most influential sports event in the world, the Olympic Games brings together the best athletes from all over the world. The Olympic medal table is not only a symbol of a country's sports strength, but also reflects, to a certain extent, the comprehensive strength of a country's economy, science and technology, culture and so on. At the same time, in addition to the ranking of the medal table, the number of medals won by other countries is also crucial, so the correct prediction of the number of medals won at the Olympic Games is of great significance to countries in formulating their sports development strategies. Since the probability of winning a medal is affected by the athletes, coaches, hosts and other factors, we want to build a model to predict the total number of medals and the number of gold medals for each country in the 2028 Olympic Games in Los Angeles, USA.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Develop a mathematical model to predict the total number of medals and the number of gold medals for each country at the 2028 Olympic Games in Los Angeles, USA, and estimate the accuracy of the model. Based on your model, which countries do you think are most likely to win more medals? Which countries will do worse than they did in 2024?
- For countries that haven't won a medal so far, how many do you expect to win their first medal at the next Olympics?
- Explore the relationship between the number and type of events at the Olympics and how many medals a country wins.
- Coaches can easily change nationality, so there is the possibility of a “great coach” effect, i.e., a coach's skills and experience will have an impact on the athletes' level of performance. What is your estimate of the contribution of this effect to the number of medals won? Choose three countries to analyze as examples.
- What other behind-the-scenes secrets or features does your model reveal about the number of Olympic medals? Explain how these characteristics would help the International Olympic Committee.

1.3 Our work

First, through data preprocessing, we obtained several larger datasets with high-dimensional feature spaces. In order to predict the medal table for the 2028 Summer Olympics in Los Angeles, USA, and to include prediction intervals for all outcomes, we decided to use the random forest model Bootstrap [1]. The features include the number of gold medals and medals won by each country in previous editions, the number of events per country participating in each edition, the number of events offered in each event, whether they were hosted or not, and the number of athletes from each country. By comparing the 2028 medal table obtained from the projections with the 2024 data, it was obtained which country improved and which did worse.

Next, for predicting countries that have not yet won a medal, we collated the data from the Athlete Level Confusion to get the data for countries that have not yet won a medal including the number of sports involved, the number of athletes, and the number of sports offered at each session. We used a logistic regression model to train the model with the collated dataset to predict who will win the first medal at the next Olympics in 2028 and how likely they are to do so.

We then explored the relationship between the number and type of events in the Olympics and how many medals a country wins separately. We first made scatter plots of the number of events versus how many medals countries won, and then analysed them using linear regression fits, quadratic and cubic polynomial regression fits, and exponential logarithmic fits, respectively, and got the best results from the cubic polynomial regression fit. Exploratory analysis of the data was then used to get a relationship between the type of sport and how many medals a country won, and data visualisation was used to derive what sports were most important to each country. For the question of how the sport chosen by the host country affects the results, we processed the data to get the number of medals won by host and non-host countries in new and traditional sports and visualised them.

Finally, for the great coaching effect, we explored the evidence for the existence of the great coaching effect using Lang Ping in the USA and Bela Karolyi in Romania as examples and demonstrated that the presence of a great coach has a significant impact on the number of medals won by the country through a variety of methods including linear regression analyses, independent samples t-tests, and analyses of variance (ANOVA). We believe that programmes with consistently low or unchanged medal counts need to consider investing in ‘great’ coaches. We used the CUSUM model to process the data we had compiled on the medals won by the three selected countries in each programme to obtain the programmes that need to be invested in, and estimated the effect of investing in ‘great’ coaches by calculating The impact of investing in a ‘great’ coach on the project was estimated by calculating the impact of investing in a ‘great’ coach on the project.

In summary, the whole modeling process can be shown as follows:

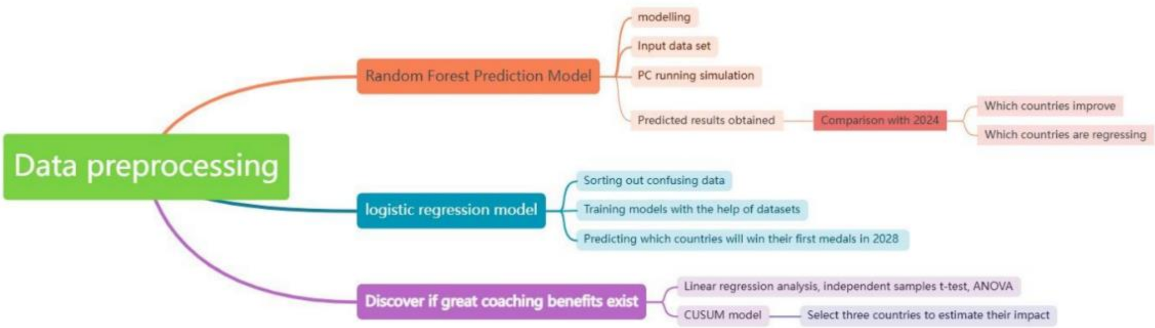


Figure 1: Mind map of our work

2 Assumptions and Justifications

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

- **Assumption 1: Athletes' performance reflects the ability of the country to win gold medals at the Olympic Games.**

Justification: Athletes‘ performance can directly reflect their ability to perform in a specific sport, we ignore their injuries, withdrawals and other factors, the athletes’ award records have a high predictive value for their future performance, so the athletes' data is used as an important input feature of the model, which directly affects the probability of winning the award.

- **Assumption 2: The factor of the host country of the Olympic Games has a significant contributing effect on the number of medals won by that country.**

Justification: The host country usually has a home field advantage, and the host country weakens the effect of historical medal counts by gaining advantages in a variety of ways, such as the number of athletes, the type of event, and the athletes' adaptation to the competition environment. Therefore, it is reasonable to assume that the host country has an additional medal-winning advantage in the projections.

- **Assumption 3: Historical medal count data is only taken from the seven Olympic Games since 2000, assuming no major changes in the sporting prowess of the countries, and assuming that the number of athletes, types of events, and participating countries in 2028 will be the same as in 2024.**

Justification: Due to the evolution of sports and the world, it is difficult to reliably reflect the future performance of Olympic data that are too old, such as disappeared countries (USSR, Yugoslavia, etc.), obsolete events (Polo, Roque, etc.), and athletes that have long since been retired, etc. Therefore, to avoid the negative impact of invalid data on the model, we adopt this assumption for subsequent modelling.

3 Model 1: A model for predicting the number of medals based on random forests

The number of gold medals and the total number of medals won by each country at the 2028 Los Angeles Olympics in the United States of America is determined by a variety of factors, i.e., the performance of athletes, the number of participants in the Olympics, the history of awards, and so on, all of which can cause changes in the number of awards won. Therefore, for such a complex system, we decided to use the predictive forest model to solve it. By using historical data, athletes' information and various parameters of the current Olympic Games, we built a regression model based on integrated learning. This model not only has high prediction accuracy, but also can handle complex nonlinear relationships.

3.1 Data Description

We removed all data before 2000, removed invalid sports, corrected and standardized wrong country names (German-1, etc.), we used NOC instead of country names, and collated to get a high-dimensional dataset containing multiple variables such as country, year, number of events competed in, number of events hosted, number of athletes, whether or not it was a host country, etc., and we also extracted the historical medal totals for each country at each event and the change of medals won by a single country at a single event over time in the summerOly_athletes file to extract the total number of historical medals won by each country in each event and the change in medals won by individual countries in individual events over time.

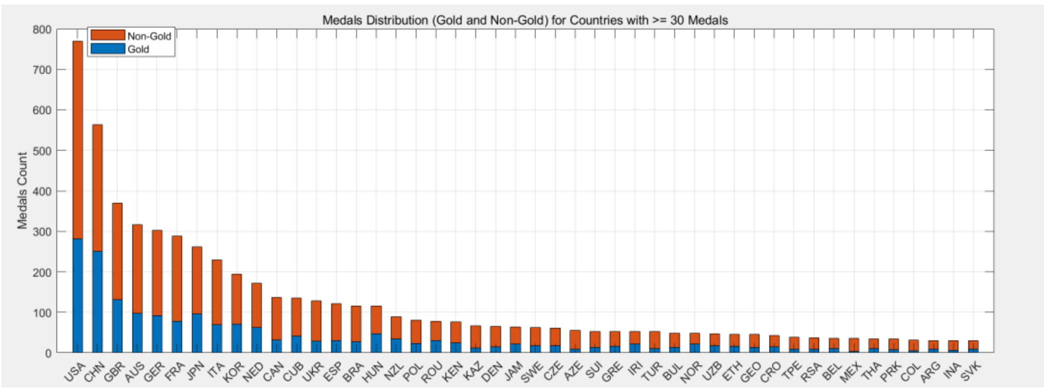


Figure 2: National Historical Gold and Medal Totals Statistical Chart

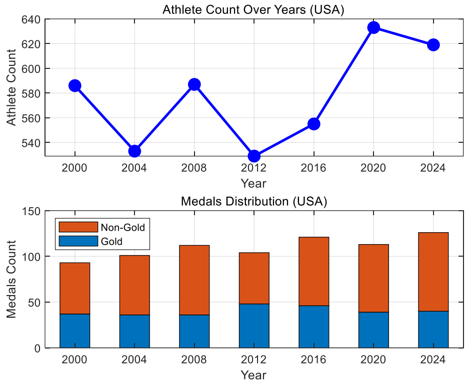


Figure 3: U.S. athletes and medal counts over time

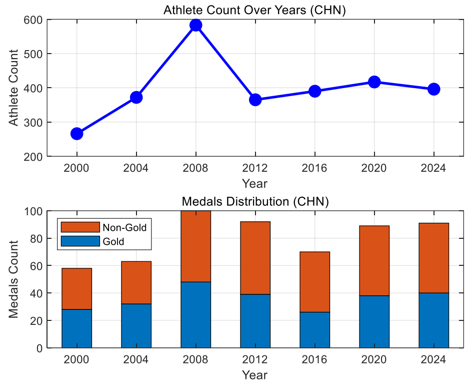


Figure 4: Chinese athletes and medal counts over time

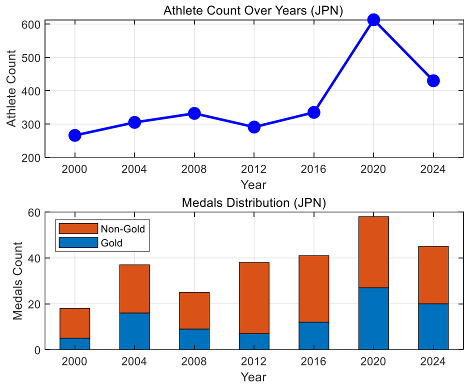


Figure 5: Japanese athletes and number of medals over time

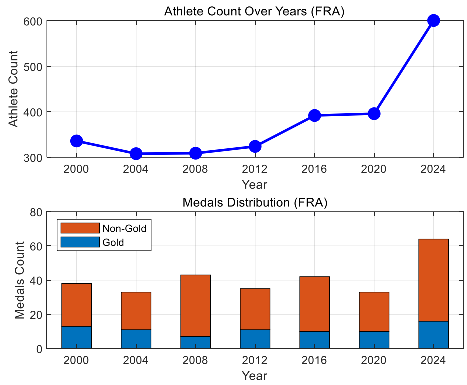


Figure 6: French athletes and number of medals over time

3.2 Notations For Model 1

Table 1: Notations used in Model 1

Symbol	Description	Unit
$Y_{i,gold}$	Country i Number of gold medals won in the Olympics	°
$Y_{i,total}$	Country i Total number of medals won in the Olympics	°
$X_{i,j}$	Indicates the characteristics of country i in sport j, including the number of athletes, historical results, etc.	°
$T_{olympics}$	Total number of Olympic events	°
D	Training dataset	°
X_i	eigenvector	°

Y_i	target variable	。
$f_t(X)$	predictive output	。
T	Number of decision trees	。
Y	The final prediction, which is the average of all tree predictions	。

3.3 The Establishment of Model 1

The training set data is first randomly sampled to obtain multiple subsets, each subset D_t containing n samples. For the tth tree, the training set

$$D_t=\{(Xi',Yi')\}^{n}_{i=1} \tag{1}$$

Then train the decision trees, the random forest model is an integrated model consisting of T decision trees. Each tree t is trained on a subset Dt and a decision tree $f_t(X)$ is generated. For each training sample X_i, a prediction is made by each tree $f_t(X_i)$ and the predicted values of all trees are obtained:

$$Y_t(X_i)=f_t(X_i) \tag{2}$$

Then each tree uses the input feature X in its prediction and outputs the corresponding number of medals (e.g.number of gold medals, total number of medals).

Finally the predictions of all trees are averaged, which is a regression problem, to get the final predicted output.

$$Y(X_i)=\frac{1}{T}\sum_{t=1}^{T} f_t(Xi) \tag{3}$$

3.4 The Solution of Model 1

Here are our predictions

Table 2: Predicted ranking of national gold medals and medals (top5)

Country	Predicted_Medal	Confidence_Interval_1	Confidence_Interval_2	Predicted_Medal	Confidence_Interval_1	Confidence_Interval_2
USA	109	104	118	41	37	44
CHN	80	70	92	37	30	40
GBR	56	41	65	19	13	25
AUS	45	36	54	14	10	17
FRA	42	33	48	11	9	14

We visualized the above table using MATLAB and got the result as follows:

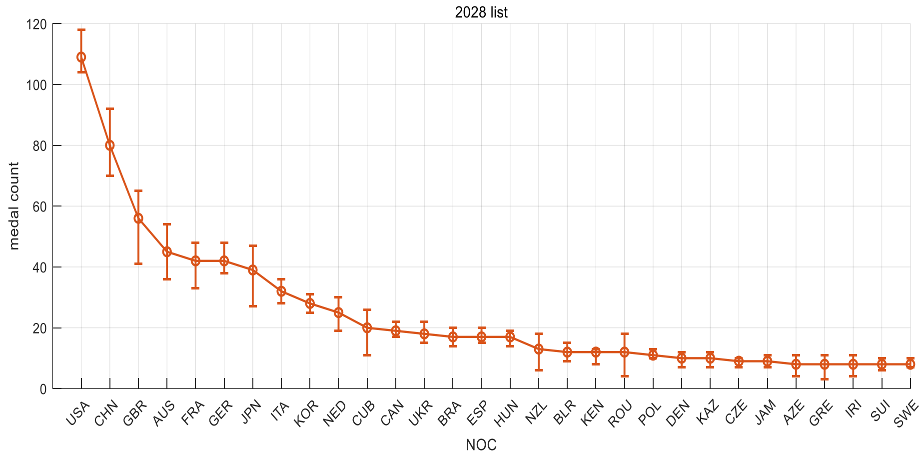


Figure 7： Top 30 predicted medal standings for the 2028 Olympics

4 Model 2: Logistic regression model to predict first medal win

The likelihood that a country that has not yet won a medal will win its first medal at the Olympics is affected by a number of factors, such as the number of events involved, the number of athletes, and the performance in previous competitions. Based on this, this paper proposes a logistic regression model [2] for predicting the probability of these countries winning their first medal at the 2028 Olympics.

4.1 Notations For Model 2

Table 3： Notations used in Model 2

Symbol	Description	Unit
X_{proj}	Number of events in which the country has participated in previous Olympic Games	°
$X_{totproj}$	Total number of all Olympic sports	°
X_{ath}	Total number of athletes from countries participating in previous Olympic Games	°
f_1	Percentage of participation in projects	°
f_2	Average number of programs in which athletes participate	°
f_3	Athletes' historical performance averages	°
f_4	target variable	°
X_{perf}	Athlete Competition Performance	°

Y	Number of medals won by the country

4.2 The Establishment of Model 2

First calculate the participation program ratio based on our assumptions:

$$f_1=\frac{\mathrm{Xproj}}{\mathrm{Xtotproj}} \tag{4}$$

Calculate Average number of programs in which athletes participate

$$f_2=\frac{\mathrm{Xproj}}{\mathrm{Xath}} \tag{5}$$

Calculate Athletes' historical performance averages and target variable

$$f_3=\mathrm{mean}(X_{perf}) \tag{6}$$

$$f_4=\mathrm{var}(X_{perf}) \tag{7}$$

Whether the country wins a medal state $Y \in \{0,1\}$

A logistic regression probabilistic model is then developed for estimating the probability that a state wins a medal $p(y=1 \mid X)$. The model is of the form:

$$p(y=1 \mid X)=\frac{1}{1+e^{-(b+w^T X)}} \tag{8}$$

where $X=[f_1, f_2, f_3, f_4]^T$ is the feature vector, w is the weight vector, and b is the bias term.

4.3 The Solution of Model 2

Table 4: Countries predicted to win their first medal

Country	Year	CountryEv	YearEvent(	AthleteCo	medal_in_h	in_host	Predicted_	Predicted_Probability
AIN	2028	10	50	32	0	0	1	0.763005
SEN	2028	8	50	11	0	0	1	0.68676
ESA	2028	7	50	8	0	0	1	0.623196
LBN	2028	7	50	9	0	0	1	0.620381
LUX	2028	7	50	13	0	0	1	0.609043
ANG	2028	7	50	25	0	0	1	0.574359
MON	2028	6	50	6	0	0	1	0.552133
GUM	2028	6	50	7	0	0	1	0.549171
NCA	2028	6	50	7	0	0	1	0.549171
NEP	2028	6	50	7	0	0	1	0.549171
PLE	2028	6	50	8	0	0	1	0.546206

We analyzed the accuracy of the model's prediction results and obtained the following ROC plot results:

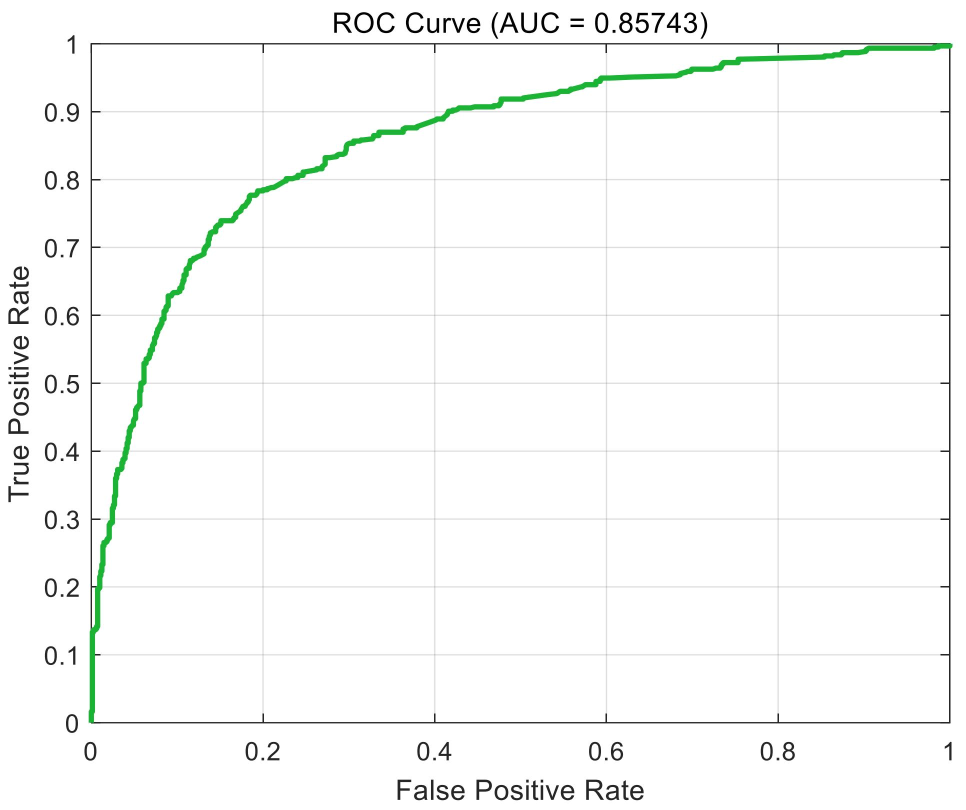


Figure 8: ROC curve predicted by the model

5 Model 3: A regression-based model exploring the relationship between the number of Olympic events and the number of medals won by each country

5.1 Notations For Model 3

Table 5: Notations used in Model 3

Symbol	Description	Unit
X	Number of Olympic events	○
Y	Total number of medals won by the country	○

5.2 The Establishment of Model 3

Firstly, we draw a scatterplot with X on the horizontal axis and Y on the vertical axis, and determine whether there is a nonlinear relationship between X and Y by observing the scatter distribution, and determine the relationship between X and Y by using cubic polynomial regression.

Then the model is built.

$$Y = w_0 + w_1 X + w_2 X^2 + w_3 X^3 + \epsilon \tag{9}$$

where:w0,w1,w2,w3 are the regression coefficients to be fitted;

ϵ is the error term, which is usually assumed to follow a normal distribution.

The model parameters are estimated by the least squares method by

$$\text{Minimize} \sum_{i=1}^{N} \left(Y_i - (w_0 + w_1 X_i + w_2 X_i^2 + w_3 X_i^3) \right)^2 \tag{10}$$

Finally, with the help of MATLAB tool, the operations were solved using the preprocessed data.

5.3 The Solution of Model 3

We wanted to plot pie charts to determine what programs each country is best at. We chose four countries, the United States, China, Japan, and France, to draw the following graph, and we can find that the United States is better at swimming, Japan is better at handball, etc.

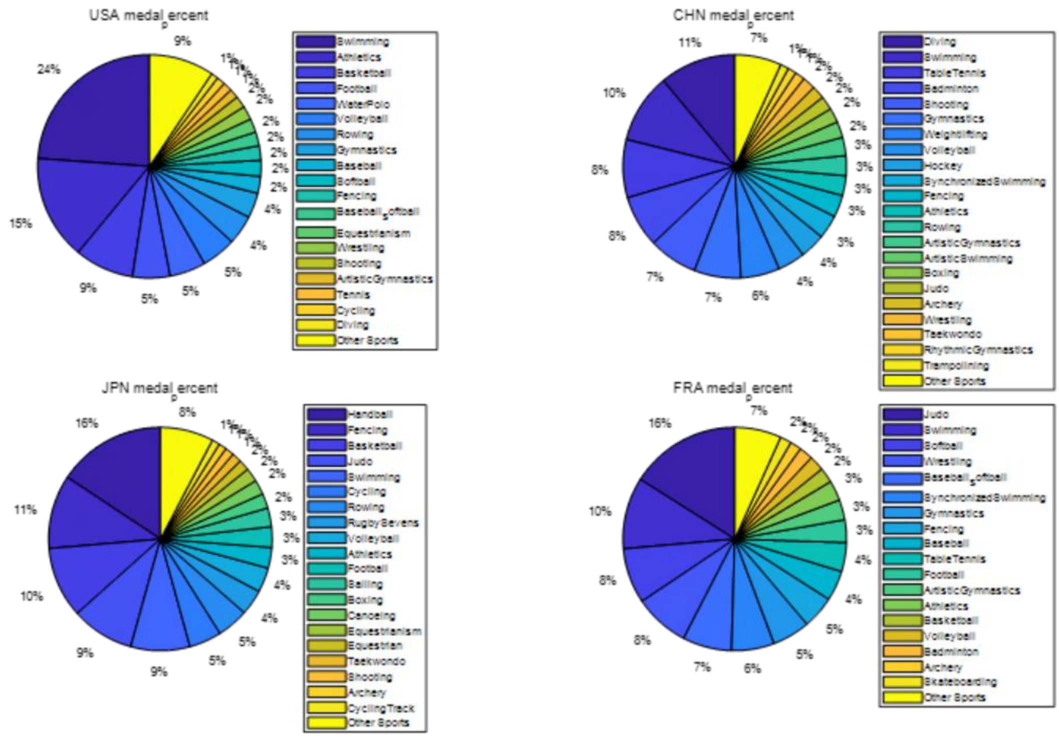


Figure 9: Awards for 4 countries on different projects (pie charts)

We performed multiple regression analyses to synthesize the relationship between the number of programs and the number of medals, and obtained the following graphs.

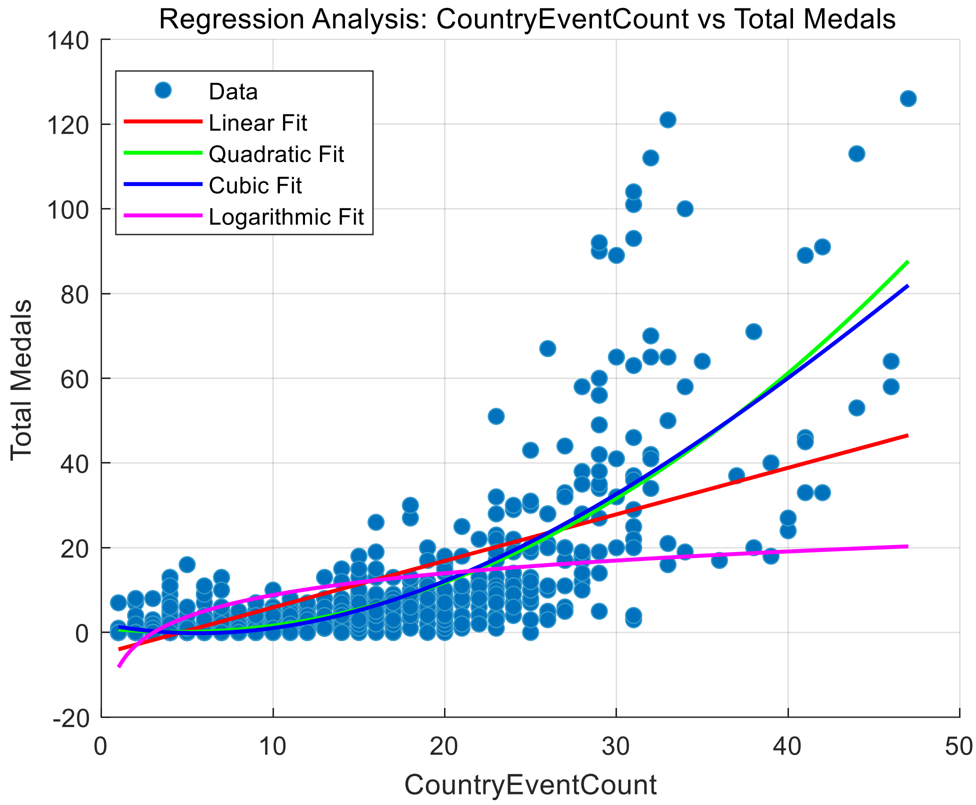


Figure 10: Multiple regression analysis plots for number of projects and number of awards won

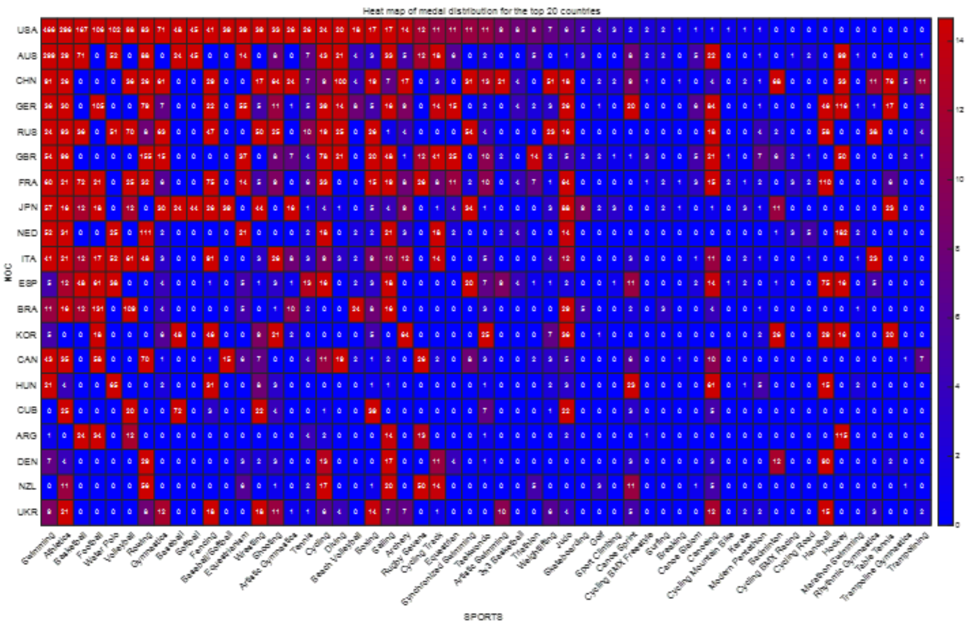


Figure 11: Heat map of awards for different projects in 30 countries

Our statistical analysis of the impact of the events chosen by host and non-host countries on the number of medals yields the following two graphs.

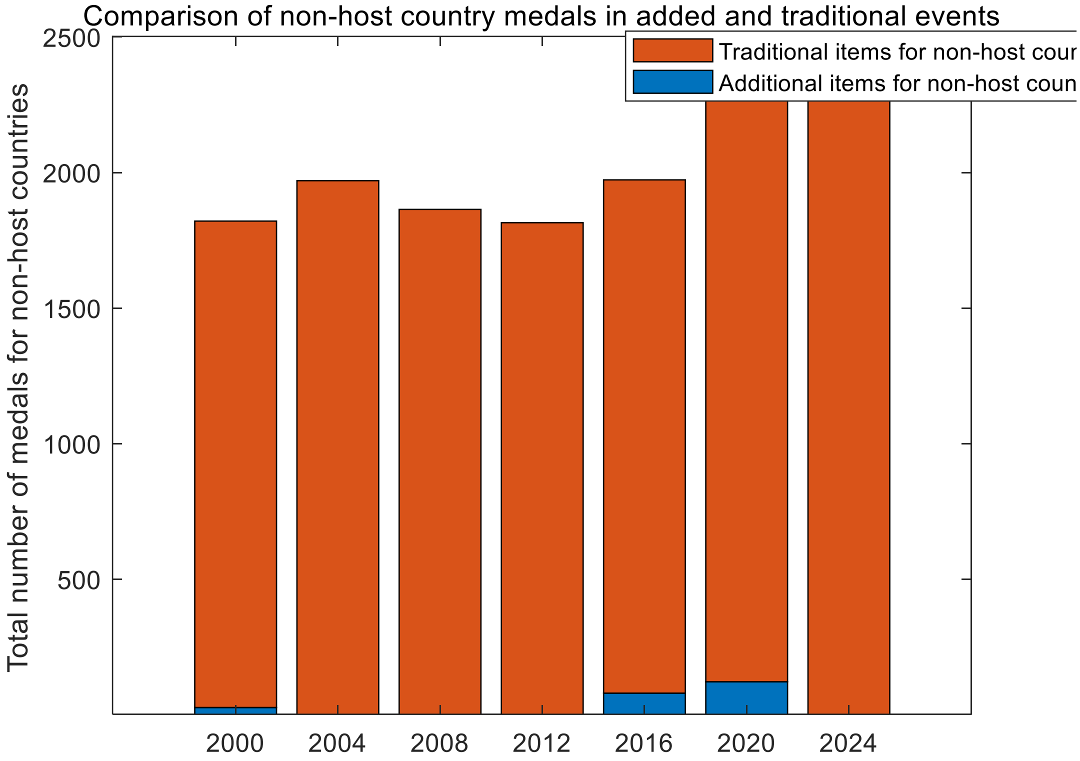


Figure 12: Comparison of non-host country medals in added and traditional events

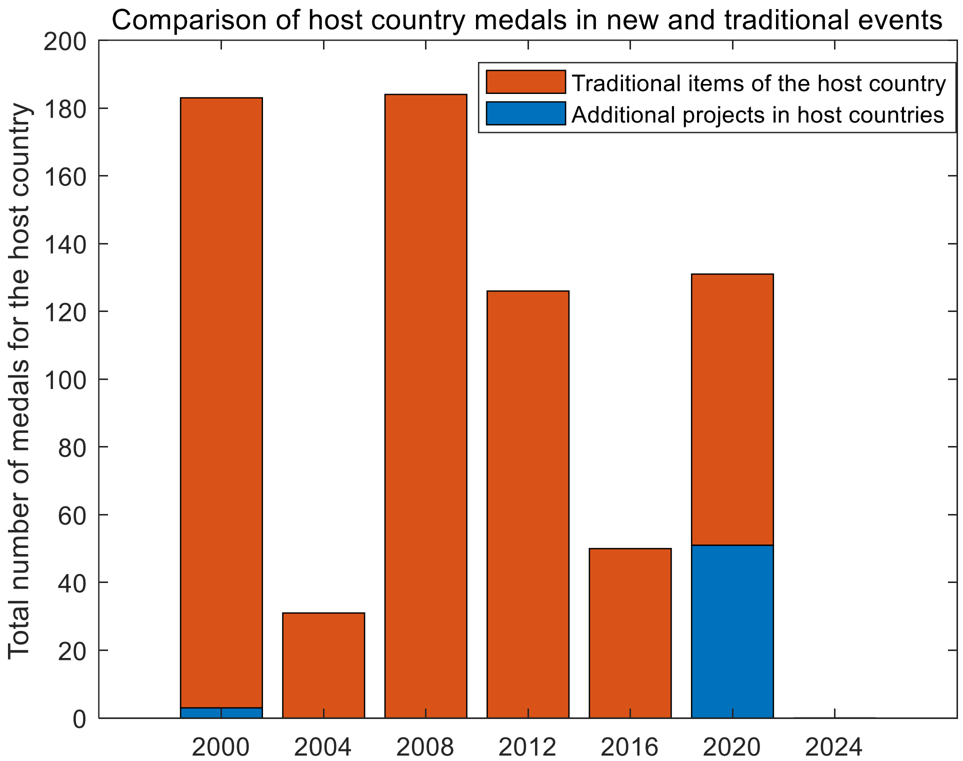


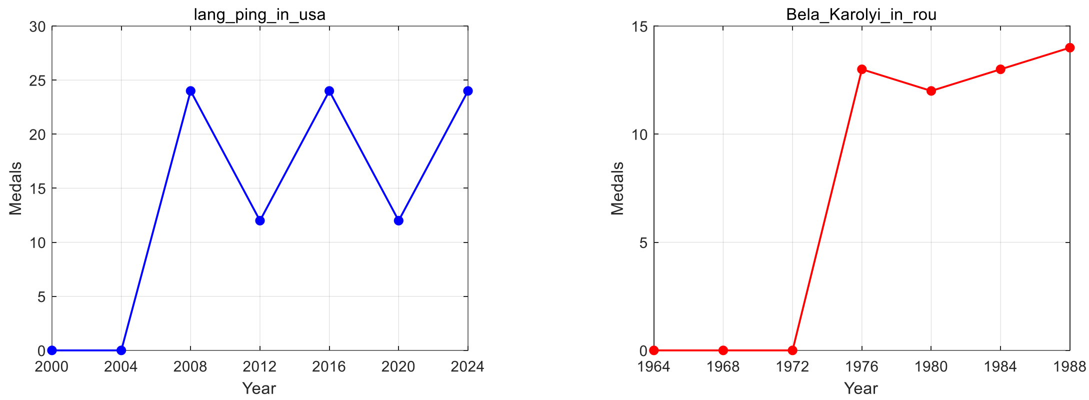
Figure 13: Comparison of host country medals in new and traditional events

6 Model 4: The "Great Coach" problem model

6.1 Proof of the "Great Coach" Effect Based on Multiple Tests

6.1.1 Data preprocessing

We looked up the data of Lang Ping and Bela Karolyi to determine the period of their coaching, determined the time interval of the data and plotted the following two graphs, it is not difficult to see that the performance of the corresponding country has been significantly improved after the "great coaches" coaching, so we take this as the evidence for the existence of the "great coach" effect and carry out the following analysis. Therefore, we take this as evidence of the existence of the "great coach" effect and analyze it as follows.



(a) lang ping in usa (b) Bela Karolyi in rou

Figure 14: Proof of the existence of the great coach effect

6.1.2 model building

We used OLS regression, independent samples t-test, and ANOVA to determine whether the "great coach" effect exists.

We solved the data for Lang Ping's time in the U.S. After solving all three models, it was determined that Lang Ping had a significant effect on the U.S.'s medal improvement, with the OLS regression model showing that the presence of a great coach wins 1.6 more medals for that country in that event, proving the existence of the great coach effect.

6.2 The Solution of Model 4

We believe that programs with little change in medal figures should invest in "great coaches". We used the CUSUM model to identify programs in the 3 countries we selected (China, USA, Japan) that have had little change in their previous medal counts. In the Appendix, we present the programs that we think are most likely to invest in "great coaches" through our calculations. We bring these items into the previous model and get their impact as an increase of 1 or 2 medals.

7 Sensitivity Analysis and Error Analysis

We conducted a detailed error analysis of the random forest model, aiming to identify the bias and uncertainty in the model prediction. By analyzing the error between the training data and the prediction results, we can identify the possible overfitting or underfitting phenomenon of the model under some specific conditions, and then optimize the parameter settings of the model to improve the accuracy of the prediction, and the standard deviation plot of the prediction results is shown in the following figure.

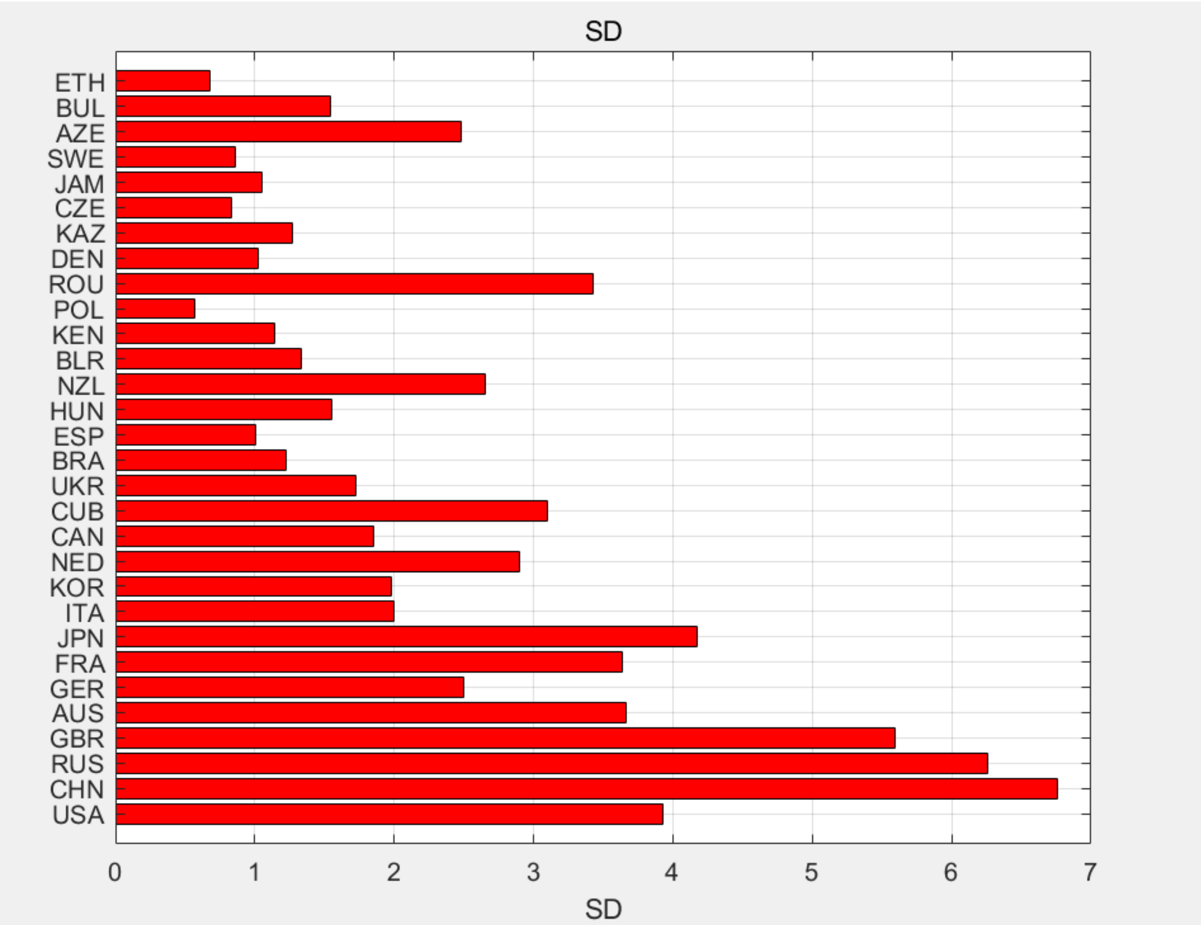


Figure 15: Standard deviation of projected medal counts for the top 30 countries

Meanwhile, we conducted a comprehensive model performance analysis of the logistic regression model, mainly evaluating its performance under different feature combinations and dataset divisions. By calculating the model's precision, recall, F1 value and other indicators, we were able to get a clearer understanding of the applicability and limitations of logistic regression in Olympic medal prediction, and the residual analysis and the comparison of measured and true values are shown in the following figure.

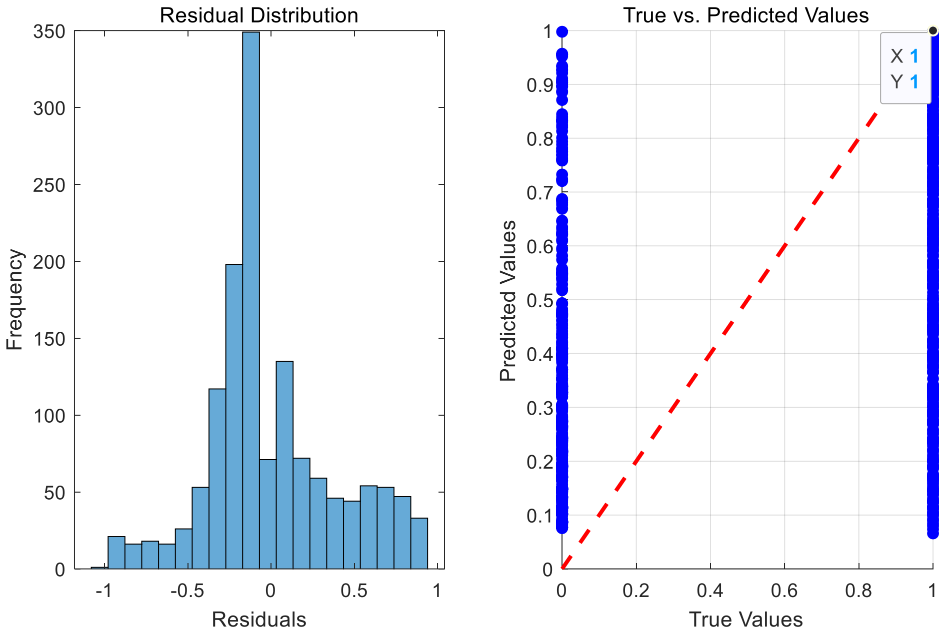


Figure 16: Residual analysis and comparison of true and measured values

8 Model Evaluation

8.1 Strengths

1.Powerful non-linear prediction ability: in Olympic medal prediction, the number of medals is affected by a variety of complex factors, such as the number of athletes, the number of competing events, the impact of the host country, etc. Random Forest can capture these complex non-linear relationships and provide high-precision prediction.

2. Olympic medal prediction involves a variety of feature types, and can directly handle these data types without normalization or special conversion operations. As Random Forest predicts through the integration of multiple decision trees, it reduces the risk of overfitting of a single tree.

3. Combined with the Bootstrap method, Random Forest can provide confidence intervals for each country's medal prediction results, clarifying the uncertainty range of the prediction. This is especially important for problems such as Olympic medal prediction where the results need to be interpreted. Powerful non-linear prediction ability: in Olympic medal prediction, the number of medals is affected by a variety of complex factors, such as the number of athletes, the number of competing events, the impact of the host country, etc. Random Forest can capture these complex non-linear relationships and provide high-precision prediction.

8.2 Weaknesses

1.Poor model interpretability, although Random Forest can provide more accurate predictions, it is a “black box” model, it is difficult to explain the specific contribution of each feature to the final prediction, which may be a problem in scenarios that require a high level of model interpretability.

9 Our insights and information

After our modeling, solving, and computing above, we found the following 4 unique insights by visualizing the data.

9.1 Declining probability of winning prizes for large sports countries and increasing probability of winning prizes for small countries

The results predicted by our random forest model show that the medal advantage of traditionally large sports countries (e.g.the United States, China) is diminishing and the probability of winning a medal is decreasing. Smaller countries with limited resources (e.g. Belarus, Cuba) are more likely to win. We analyze that the reason for this may be that the globalization of the world has promoted the sharing of sports technology and resources, and small countries have begun to cultivate world-class athletes. In addition, due to different geographical locations, dietary habits and other conditions, each country specializes in different Olympic sports, which makes it easier for some countries to win medals or even gold medals. While the traditional big sports countries will still rank high in terms of the number of medals won, the probability of winning medals will decrease compared to the previous period, and will be affected by the impact of other countries.

9.2 The Critical Role of Great Coaches in the Probability of Winning Awards

Our logistic regression model shows that countries with great coaches have a significantly higher probability of winning the prize than countries without great coaches. We believe that a great coach not only improves the individual athlete's skills, but also develops a scientific preparation strategy. They will provide athletes with their own competition experience, which can motivate athletes on a psychological level and increase their confidence in competition.

9.3 The centrality of high medal rate programs

We find that countries' medal totals are significantly correlated with their performance in high-medal-rate events. Model analysis suggests that countries that focus on high medal rate events are more likely to top the medal table. With limited funds and manpower, it is more cost-effective to focus on a small number of sports in the Olympic medal distribution.

9.4 The impact of countries' GDP on Olympic prizes

We find that a country's ability to win medals is correlated with its national GDP, as shown in the chart below, where the top fifteen countries in terms of GDP are almost all in the top fifteen of the predicted medal table (Russia will not be participating in the 2028 Olympics), so we can speculate that the higher the country's GDP, the higher the number of medals it will win. We believe that countries with higher GDPs usually invest more in sports and pay more attention to sports, so they will perform better; countries with higher GDPs have been developing steadily for a longer period of time and have accumulated more experience in sports, making it easier for them to achieve excellent results.

Table 6: GDP vs projected medal standings

	NOC	GDP	Country	Predicted_I
1	USA	141,273	USA	109
2	CHN	86,816	CHN	80
3	GER	23,053	GBR	56
4	JPN	19,612	AUS	45
5	IND	18,696	FRA	42
6	GBR	17,433	GER	42
7	FRA	15,678	JPN	39
8	ITA	11,288	ITA	32
9	BRA	11,018	KOR	28
10	CAN	10,619	NED	25
11	RUS	9,876	CUB	20
12	MEX	9,625	CAN	19
13	KOR	9,414	UKR	18
14	AUS	8,741	BRA	17
15	ESP	8,183	ESP	17

9.5 Information to the National Olympic Committee

9.5.1 Supporting the development of small States and strengthening international cooperation

- Share training methods and resources in international cooperation to promote global sports development.
- Strengthen exchanges and cooperation with small countries and explore win-win mechanisms.

9.5.2 Optimize coach selection and training system

- Establish a global coaching database to evaluate coaches' historical performance and potential.
- Provide systematic training programs for young coaches to nurture the next generation of great coaches.
- Encourage international coaching exchanges to introduce advanced concepts to national athletes and improve their competitive level.

10 Conclusion

In this analytical study on the number of medals in the Olympic Games, we have used the Random Forest model, logistic regression model and polynomial regression model, etc., to make detailed predictions and trend analyses of the number of medals won by various countries in the previous Olympic Games. The core objective of our study is to predict the distribution of the number of medals in future Olympic Games through these models, especially to analyze the impact of various factors, such as the host country effect, the number of athletes, the number of participating events, and the historical performance on the number of medals.

We collected medal data from previous Olympic Games and performed data cleaning and preprocessing to ensure the accuracy and usability of the data. In order to capture complex medal distribution patterns, we selected three commonly used machine learning models:

1. random forest model: this model is capable of handling high-dimensional data and capturing non-linear relationships by integrating multiple decision trees. With this model, we identify the key factors affecting the number of medals, including national GDP, population size, and athlete participation.

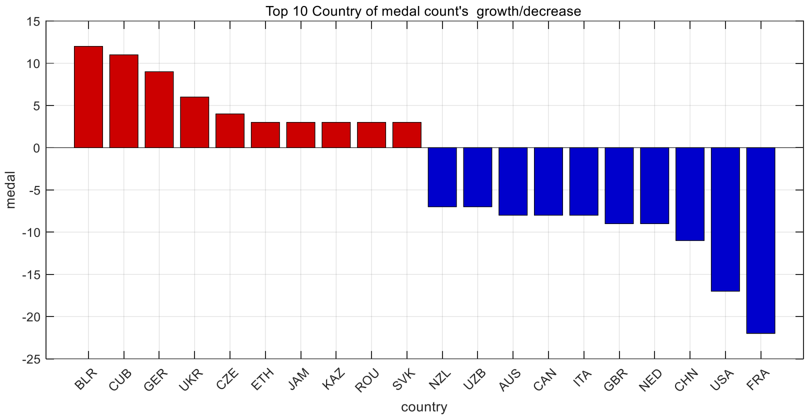


Figure 17: Top 10 Country medal count's growth/decrease

2. Logistic regression model: It is used to analyze the probability of countries winning medals. By building regression equations, we reveal the key factors affecting the distribution of medals, such as economic level and policy support.

3. Polynomial regression model: By analyzing the trend of the number of medals over time, we predicted the number of medals in the next few Olympics, especially in the medium and long term.

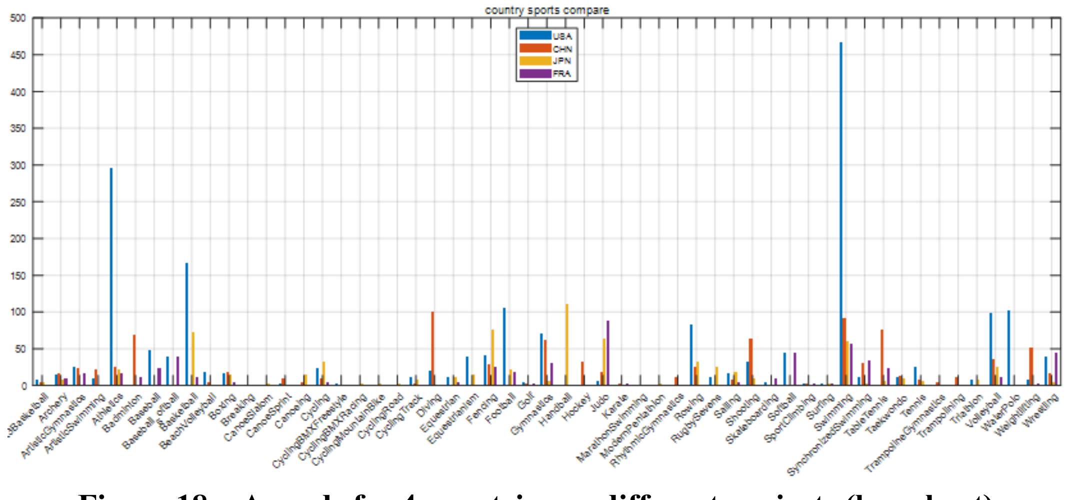


Figure 18: Awards for 4 countries on different projects (bar chart)

After model selection, we evaluated the models through methods such as cross-validation and used a large amount of simulated data to test the accuracy of the models. We ensured the stability and predictive ability of all models, which provided strong support for the subsequent analysis.

We used three statistical tools: OLS regression, independent samples t-test, and ANOVA to determine the existence of the “great coach” effect. We also used CUSUM to obtain the programs in which three countries need to invest in “great coaches”.

This study reveals the impact of different factors on the number of Olympic medals by analyzing the number of medals from multiple perspectives. Through the application of machine learning models, we not only made predictions on the future distribution trend of Olympic medals, but also provided valuable references for countries to formulate sports policies. Further analysis can help governments and sports administrations make more targeted decisions in future Olympic cycles.

References

[1] Yu H. Optimization research on stock index forecasting model based on random forest and LSTM network[D]. Harbin University of Commerce,2023.
[2] Long Yang. Research on financial risk early warning of Chinese medicine industry based on logistic regression model[D]. Northeast University of Finance and Economics, 2022.

Appendices

Appendix 1
Introduce: **Predicting whether the first medal will be won with a logistic regression model and evaluating model performance**

```
data_train = readtable('jibenchuli.csv');
data_new = readtable('nomedal（1）.csv');
data_train.FirstMedal = (data_train.Total_medal > 0);
features = {'CountryEventCount', 'YearEventCount', 'AthleteCount'};
X_train = data_train{:, features};
y_train = data_train.FirstMedal;
logit_model = fitglm(X_train, y_train, 'Distribution', 'binomial', 'Link', 'logit');
X_new = data_new{:, features};
predicted_probabilities = predict(logit_model, X_new);
predicted_labels = predicted_probabilities > 0.5;
data_new.Predicted_FirstMedal = predicted_labels;
data_new.Predicted_Probability = predicted_probabilities;
data_new_sorted = sortrows(data_new, 'Predicted_Probability', 'descend');
writetable(data_new_sorted, 'predicted_nominal_first_medal_with_probability.csv');
disp(data_new_sorted(1:10, {'Country', 'Predicted_FirstMedal', 'Predicted_Probabil-
ity'}));
[predicted_values, ci] = predict(logit_model, X_train, 'Alpha', 0.05);
lower_ci = ci(:, 1);
upper_ci = ci(:, 2);
residuals = y_train - predicted_values;
mse = mean(residuals.^2);
rmse = sqrt(mse);
sst = sum((y_train - mean(y_train)).^2);
ssr = sum((predicted_values - mean(y_train)).^2);
r_squared = 1 - sum(residuals.^2) / sst;
fprintf('Mean Squared Error (MSE): %.4f\n', mse);
fprintf('Root Mean Squared Error (RMSE): %.4f\n', rmse);
fprintf('R-squared (R^2): %.4f\n', r_squared);
figure;
subplot(1, 2, 1);
histogram(residuals, 20);
title('Residual Distribution');
xlabel('Residuals');
ylabel('Frequency');
```

Appendix 2

Introduce: **Random forest and bootstrap estimation of medals and stability analysis**

```
data = readtable('medalallnoc.csv');
countries = unique(data.Country);
predictions = cell(length(countries), 3);
nTrees = 5;
nBootstrap = 5;
for i = 1:length(countries)
country_id = countries{i};
country_data = data(strcmp(data.Country, country_id), :);
X_train = country_data{1:end-1, {'CountryEventCount', 'YearEventCount', 'AthleteCount', 'medal_in_history', 'in_host'}};
y_train = country_data{1:end-1, 'Total'};
X_2028 = country_data{end, {'CountryEventCount', 'YearEventCount', 'AthleteCount', 'medal_in_history', 'in_host'}};
rf_model = TreeBagger(nTrees, X_train, y_train, 'Method', 'regression');
predicted_medal = predict(rf_model, X_2028);
predicted_medal_int = round(predicted_medal);
bootstrap_preds = zeros(nBootstrap, 1);
    for j = 1:nBootstrap
idx = randi([1, size(X_train, 1)], size(X_train, 1), 1);
X_bootstrap = X_train(idx, :);
y_bootstrap = y_train(idx);
rf_model_bootstrap = TreeBagger(nTrees, X_bootstrap, y_bootstrap, 'Method', 'regression');
bootstrap_preds(j) = predict(rf_model_bootstrap, X_2028);
    end
predicted_medal_mean = mean(bootstrap_preds);
predicted_medal_std = std(bootstrap_preds);
predicted_medal_lower = round(predicted_medal_mean - 1.96 * predicted_medal_std);
predicted_medal_upper = round(predicted_medal_mean + 1.96 * predicted_medal_std);
predictions{i, 1} = country_id;
predictions{i, 2} = predicted_medal_int;
predictions{i, 3} = [predicted_medal_lower, predicted_medal_upper];
end
predictions_table = cell2table(predictions, 'VariableNames', {'Country', 'Predicted_Medal', 'Confidence_Interval'});
writetable(predictions_table, 'predictions_results2.csv');
figure;
```

```
mean_predictions = cell2mat(stability_results(:, 2));
variances = cell2mat(stability_results(:, 3));
countries = stability_results(:, 1);
bar(mean_predictions);
hold on;
errorbar(1:length(countries), mean_predictions, sqrt(variances), '.');
hold off;
xticks(1:length(countries));
xticklabels(countries);
xtickangle(45);
ylabel('predict medal');
title('stable');
```

Appendix 3

Introduce: **Three ways to test the existence of the "great coach" effect**

```
disp('--- OLS ---');
X = [ones(size(LP)), LP];
beta = (X' * X) \ (X' * MEDAL3);
disp('OLS :');
disp([' ( β 0): ', num2str(beta(1))]);
disp(['( β 1): ', num2str(beta(2))]);
y_pred = X * beta;
disp('--- T ---');
MEDAL_GM1 = MEDAL3(LP == 1);
MEDAL_GM0 = MEDAL3(LP == 0);
[h, p, ci, stats] = ttest2(MEDAL_LP1, MEDAL_LP0);
disp('--- (ANOVA) ---');
[p_anova, tbl, stats_anova] = anova1(MEDAL3, LP, 'off');
```

Table 7: Three countries predicted to invest in the "Great Coach" program

NOC	Sport	NOC	Sport	NOC	Sport
CHN	3x3 Basket-ball	JPN	Boxing	USA	Badminton
CHN	Cycling Road	JPN	Canoe-ing	USA	Canoeing
CHN	Cycling Track	JPN	Cycling	USA	Golf
CHN	Equestrian	JPN	Diving	USA	Hockey
CHN	Football	JPN	Golf	USA	Judo

CHN	Golf	JPN	Hand-ball	USA	Surfing
CHN	Handball	JPN	Hockey	USA	Table Ten-nis
CHN	Rugby Sev-ens	JPN	Rowing	USA	Taekwondo

Artificial Intelligence Usage Report

1.Which of the American contests can you help me with?

The American Collegiate Mathematical Modeling Competition (MCM/ICM, USA) is a competition involving mathematical modeling, computation, and analysis. If you participate in it, the areas I can help with include:

1. Model construction and design

- Problem understanding: help analyze the requirements of the topic, define the research objectives and key questions, and organize them into a task list.

- Model selection and design: according to the characteristics of the problem, assist in selecting suitable mathematical models (e.g. linear programming, regression analysis, differential equations, stochastic models, etc.) and designing specific calculation steps.

2. Data Analysis and Processing

- Data preprocessing: including missing value processing, outlier detection, normalization, etc.

- Data analysis tools: guide the use of Python, MATLAB, R and other tools for data visualization, statistical analysis or simulation.

3. Algorithm implementation and optimization

- Algorithm design: help implement specific algorithms (e.g. dynamic programming, Monte Carlo simulation, genetic algorithm, etc.) according to model requirements.

- Algorithm design: help to implement specific algorithms (e.g. dynamic programming, Monte Carlo simulation, genetic algorithm, etc.) according to model requirements.

4. Document Writing and Layout

- Thesis framework: provide the standard structure of a thesis (abstract, introduction, model assumption, symbol definition, model construction, solution process, result analysis, conclusion, etc.).

- LaTeX layout: help realize high-quality LaTeX layout, including formulas, charts and reference formats.

- Language embellishment: improve the logic and language expression level of the paper to avoid grammatical errors.

5. Result analysis and validation

- Result visualization: assist in drawing graphs and charts to help present model output more clearly.

- Discussion of results: extract key conclusions from the results, point out the strengths and weaknesses of the model and the direction of improvement.

6. Time management and strategy

- Suggestions on Time Allocation: Make a reasonable time plan according to the difficulty of the topic to ensure that each stage of the task can be completed on time.

- Teamwork suggestions: provide strategic suggestions for team division of labor to improve overall efficiency.

If you have chosen a certain topic or have specific needs, feel free to tell me, I can help you more precisely! 😊

2.When we talk about predictions we usually base them on the athletes currently scheduled to compete, now that we have data on all individual Olympians, including their sport and performance (medal type or no medal), can you provide me with an idea for my reference?

Combining the athlete data you mentioned, including sport and performance, with the available medal data, you can build a more sophisticated and comprehensive predictive model by following the steps below. These steps will combine athlete-specific performance, characteristics of the Olympic program, and historical data to predict the number of medals each country will win in the 2028 Olympics.

Summary of Ideas:

1. **Data Integration**:

- **Historical Medal Data**: includes the number of medals each country has won in past Olympics.

- **Athlete Data**: Includes each athlete's events and results (medal type or no medal). This part of the data will reflect the performance of each athlete and can help predict the number of medals for a particular country.

- **Program Data**: Includes the type of program, number of programs, and performance of each country for each Olympic Games.

2. **Data preprocessing**:

- **Athlete Characterization Construction**:

- Create features for each athlete's performance (e.g. gold, silver, number of bronze medals, no medals).

- A series of features can be generated for each athlete based on the events in which they have competed, combined with their historical performance (e.g., whether they have won medals in the past few years, the type of events in which they have won medals, etc.).

- **Country features constructed**:

- Historical medal data for each country: number of medals at previous Olympics, percentage of gold, silver and bronze medals.

- Number of athletes involved per country and their historical performance: consider the performance of athletes within the country to add a measure of overall national strength.

- **Program characteristics constructed**:

- Types of events and number of events: types of events in the Olympics (e.g., individual events, team events, etc.), and different weights can be set for each event.

- Historical medal distribution for each event: different types of events may affect the difference in medal distribution.

3. **Characteristic design**:
- **National level**:
- National historical medal data: the number of medals and the proportion of gold, silver and bronze medals for each country in the past Olympic Games.
- Number of national athletes: number of athletes sent by each country.
- **Athlete level**:
- Medal performance per athlete: number of gold, silver and bronze medals, no medals.
- Program performance per athlete: whether or not they won a medal in a program, type of medal.
- **Program level**:
- Number of gold, silver, and bronze medals in each Olympic sport: distribution of medals in different sports.

4. **Model Selection and Training**:
- A **regression model** is used to predict the number of medals for each country in the 2028 Olympics. Modeling can be done in the following ways:
- **Multi-task learning**: train separate regression models for gold, silver and bronze medal counts.
- **Integrated methods**: e.g. using **Random Forest** or **XGBoost** for these diverse data features.
- Three separate models can be built for each country's medal count (predicting gold, silver, and bronze medal counts, respectively), and these predictions can eventually be combined to form a complete medal prediction.

5. **Model Validation**:
- Use methods such as cross-validation to evaluate the performance of the models.
- For each prediction (e.g., gold, silver, bronze), its margin of error can be calculated and a prediction interval given.

By combining athlete data, program data, and historical country performance, you can build a prediction model for each country that accurately predicts its number of gold, silver, and bronze medals in the 2028 Olympics. By taking into account factors such as athlete performance, number and type of events, you will be able to create a more accurate and reliable model.

3.Disadvantages of the Random Forest Model？

Disadvantages of the random forest model include:

1. **Large computational overhead**: Since random forests make predictions by con-

structing multiple decision trees, the training and prediction process requires large computational resources and time, especially when the amount of data is large.

2. **Poor model interpretability**: Although Random Forest can provide more accurate predictions, it is a "black box" model that makes it difficult to explain the specific contribution of individual features to the final prediction, which can be a problem in scenarios that require model interpretability.

3. **May be sensitive to noise**: Although Random Forest reduces the risk of overfitting by integrating multiple trees, the model may still exhibit a certain amount of overfitting when the data contains a large amount of noise, affecting its ability to generalise.

*** Translated with www.DeepL.com/Translator (free version) ***