Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2523094

2028 Olympic Medal Prediction and Coach Effectiveness: A Historical Data-Based Random Forest and Regression Analysis

As a grand event in the sports world, the Olympic Games, is a dream destination for countless athletes around the globe. After more than a century of inheritance and development, the Olympics now carry many comprehensive social functions, including political, economic, and cultural exchanges. The distribution of competition medals is influenced by various factors. To forecast the medal acquisition of different countries at the 2028 Los Angeles Olympics, analyze the effects of project selection, host nation influence, coaches, and other factors on the results, and to provide suggestions for host countries in selecting new events and hiring coaches, we established two models. **Model I: Medal Prediction Model,Model II: Coaching Effectiveness Analysis Model.**

Model I first analyzes the participation and award results of various countries and events in previous Olympic Games, adding binary variables related to the host country. We establish a Random Forest Model and a Negative Binomial Regression Model. Historical multi-dimensional data is used as the training set, and the credibility of the Random Forest Model is evaluated using R and RMSE. If $R < 0.5$, a second training session is conducted. For projects where the results from both sessions are unsatisfactory, the Negative Binomial Regression Model will be used. The R obtained will be compared with the previous two sessions, and the optimal result will be selected to estimate the final medal situation for 2028.

In Model II, we use a Random Forest Model for prediction. All historical data prior to the coach's involvement in the sport are treated as the training set, and the predicted medal outcomes are compared with the actual outcomes when the coach was involved. The model's fitting degree is assessed using RMSE, and based on the optimized model, deviations are estimated to evaluate the coach effect.

Regarding the host nation effect, we additionally set a binary variable. We also analyze the correlation between the number of medals won by different countries and their selection of events. By combining multiple factors, we provide recommendations for event selection and coach hiring for the chosen host countries.

Keywords: Grid Search; Olympic; Negative Binomial Regression; Random Forest; One-hot encoding

Contents

1 Introduction

1.1 Problem Background

The four-year Olympic Games are an irreplaceable jewel in the world of sports, where numerous athletes pour their heart into competing for the ultimate symbol of glory—medals, which represent both national and personal achievement. With the increasing number of events, the total number of medals awarded at the Olympics has also gradually risen. In addition to the inherent appeal and competitiveness of the sports, now the changes in total medal counts and rankings by country attract the attention of more viewers.

Aside from objective factors that influence a nation's overall sporting level, such as economic and educational standards, individual sports or events with fewer participants are significantly affected by the skills and capabilities of the athletes themselves.

This intense reliance on the individual athlete's performance highlights the need to prioritize the current comprehensive strength of athletes being cultivated in each event when predicting a country's future gold medal count, rather than simply relying on past total medal counts.

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:

- Process the existing dataset to observe the historical medal acquisition and participation rates of various countries. Based on the processed dataset, select appropriate methods to establish gold medal and overall medal prediction models to forecast the medal acquisition situation for the 2028 Olympics.
- Analyze the trends of certain countries based on the medal prediction model. Which countries are likely to achieve their first-ever medal?
- Establish an appropriate mathematical model to analyze the coaching effect, comparing the prediction results with situations where no coaches were introduced, and provide recommendations regarding coach hiring.
- Process recent data on medal acquisition by countries, analyze the probability of obtaining non-gold medals, and, in combination with the models established above, provide suggestions for event selection based on the correlation between different countries and events.

1.3 Our work

The topic requires us to predict the medal outcomes for the 2028 Olympics and explore strategies regarding coach selection and host country event selection. Our work mainly includes the following aspects:

✧ Establishing a medal prediction model based on historical medal data from various countries to provide insights into the medal outcomes and fluctuations for each country in 2028.

✧ Developing a coach effect analysis model based on the data fluctuations brought by star coaches and past trends.

✧ Analyzing how event selection influences the ultimate medal acquisition through a medal benefit model, providing reasonable suggestions for both the host country and participating countries.

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

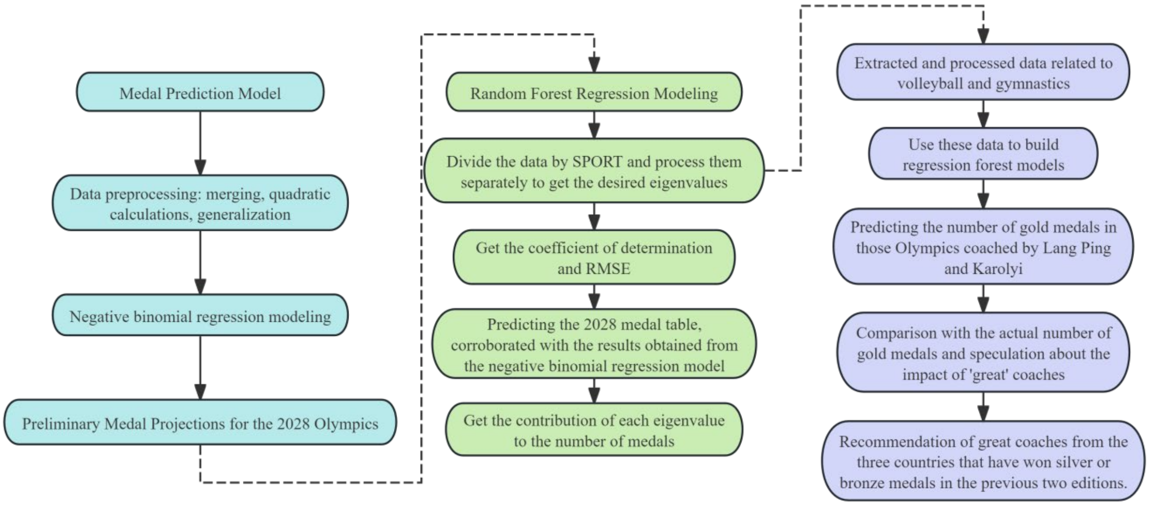

Medal Prediction Model
Data preprocessing: merging, quadratic calculations, generalization
Negative binomial regression modeling
Preliminary Medal Projections for the 2028 Olympics
Random Forest Regression Modeling
Divide the data by SPORT and process them separately to get the desired eigenvalues
Get the coefficient of determination and RMSE
Predicting the 2028 medal table, corroborated with the results obtained from the negative binomial regression model
Get the contribution of each eigenvalue to the number of medals
Extracted and processed data related to volleyball and gymnastics
Use these data to build regression forest models
Predicting the number of gold medals in those Olympics coached by Lang Ping and Karolyi
Comparison with the actual number of gold medals and speculation about the impact of 'great' coaches
Recommendation of great coaches from the three countries that have won silver or bronze medals in the previous two editions.

Figure 1: Model Overview

2 Assumptions and Justifications

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

➢ **Assumption 1: The international environment remains stable.**

Justification: Before the next Olympic Games（2028）, there will not be significant events such as large-scale wars or pandemics that could notably affect the hosting of the Olympics and the training of athletes.

➢ **Assumption 2: The project selections by athletes will not deviate from reality.**

Justification: Over a long practice span, there will not be a significant number of athletes switching sports.Also, countries invest effort in projects which are not drastically different from their past endeavors. Changes within a normal range can be considered negligible.

➢ **Assumption 3: The medal acquisition status among countries is predictable.**

Justification: Historical medal data indicates the competitive level of the country's sports and the capabilities of its athletes, which are highly correlated with future sports performance.

- **Assumption 4: Assume the research data is accurate.**

Justification: We assume that the acquisition of medals objectively reflects the level of the country's performance in the sport. This ensures the results obtained from the existing dataset are credible.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
$Y_{c,t}$	The gold count of Country c at Year t
$\mu_{c,t}$	The expected number of gold medals calculated by the model(1)
u_c	Random effects of country c by the modal(1)
$Y_{c,t}^{Total}$	The Medal count of Country c at Year t
$\mu_{c,t}^{Total}$	The expected number of gold medals calculated by the model(2)
v_c	Random effects of country c by the modal(2)

4 Model Preparation

4.1 Data Collection

Since the amount of data is large and not intuitive, we directly visualize some of the data for display.

The new data sources are summarized in Table 2.

Table 2: Data source collation

Database Names	Database Websites Data	Type
2028 Olympic Event	https://zh.wikipedia.org/wiki/	Net Book
2028 Olympic Event	https://www.olympics.com/en/sports/	Official Website
Coach News	https://www.chinanews.com/ty/2013/04-25/4763823.shtml	News Coverage

4.2 Data Cleaning

Based on preliminary observations, there are errors in the medal counts in summerOly_counts.csv. Therefore, we will recalculate the data in summerOly_athletes.csv to compile specific information for each country for each event by year, including the total number of events held, the number of participants in each event, the total number of medals, and the number of gold medals.

According to the dataset provided, we will merge summerOly_athletes.csv and summerOly_host.csv by year and NOC. A new column, "host," will be added to summerOly_athletes.csv. If the NOC in summerOly_athletes.csv matches the corresponding country in summerOly_host.csv, the "host" value for that row will be set to 1; otherwise, it will be set to 0.

The NOC country codes will be standardized according to the descriptions provided on Wikipedia. As nations like East Germany and the Soviet Union no longer exist and therefore cannot win medals at the 2028 Olympics, they will be removed or their historical information will be updated to reflect the names of the current countries.

The summerOly_programs dataset provides information on the number of types of events each year. There are two events that were only held at the Winter Olympics after 1920; these events will not aid in predicting future Summer Olympic events, and thus, we will handle the participation data associated with these two events as missing values.

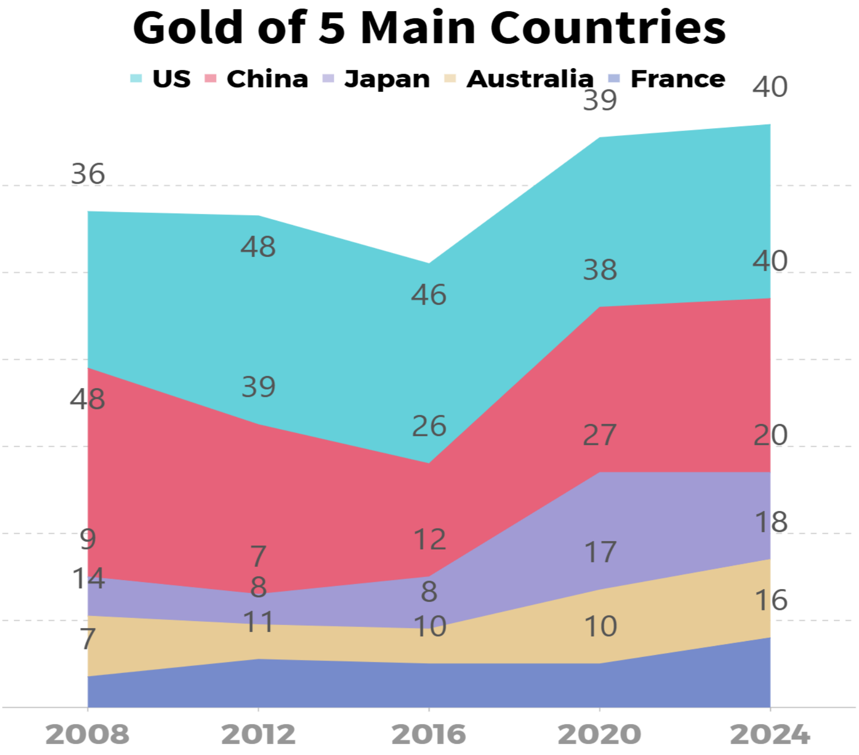


Figure 2:Gold_Count of 5 Countries

5 Model I: Medal Prediction Model

The acquisition of medals is determined by various factors, including the investment in athletes for each event, past performance in the events, the host country effect, the total number of athletes, and the selection of events, among others. Generally speaking, countries that have won medals in multiple previous editions of a particular event have an overall performance level that is significantly higher than that of countries that have not won medals, which also increases their likelihood of winning in the upcoming Olympic Games. Additionally, due to the discontinuous and discrete distribution of medal data over time, we employ a Negative Binomial Regression Model.

5.1 Data

Historical Gold:The total number of gold by this country before next Olympic.

$$\text{Historical Gold}_{c,t} = \sum_{t'<t} Y_{c,t'}$$

Historical Medal:The total number of medal by this country before next Olympic.

$$\text{Historical Total}_{c,t} = \sum_{t'<t} Y_{c,t'}^{\text{Total}}$$

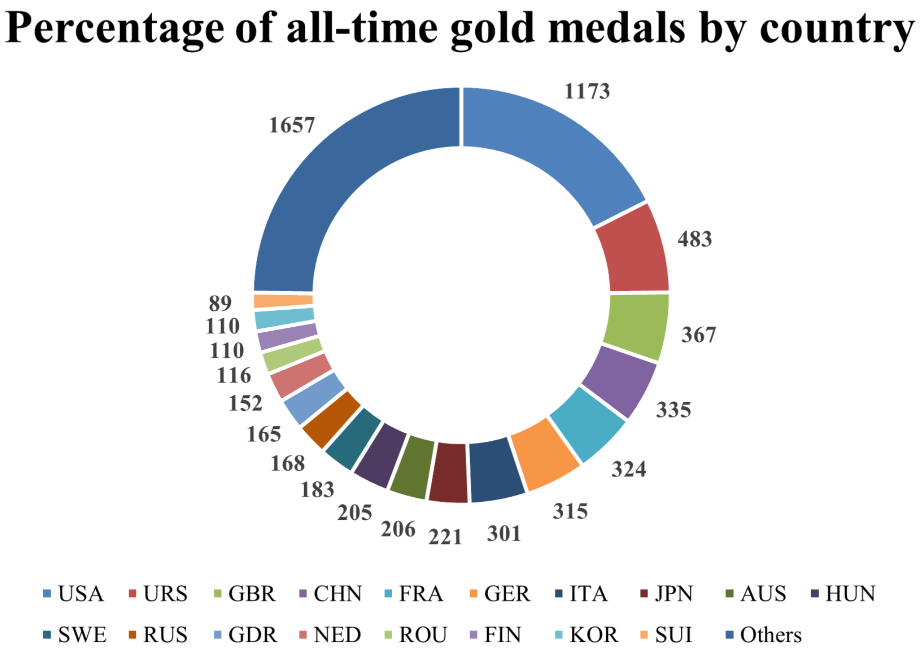


Figure 3:Percentage of gold by country

New dummy variable:host.Add this binary variable to the data as described when merging the data, where 1 represents the host country and 0 represents the non-host country.

$$\text{Host}_{c,t} = \begin{cases} 1, \text{if the country c is host in year t} \\ 0, \text{or not} \end{cases}$$

Merge the data of athletes.

Sport_Count:The number of sports that the country participated in at the current Olympic Games.

Athlete_Count:The total number of athletes from the country who competed in the current Olympic Games.

$$\text{Sport\_Count}_{c,t} = \text{Number of the distinct Sports}_{c,t}$$
$$\text{Athlete\_Count}_{c,t} = \text{Number of Athlete}_{c,t}$$

5.2 Negative Binomial Regression Model

5.2.1 Modal Discription

We will set the participation and award data of the projects before the 2028 as feature values for the training set of the random forest model, categorized by country and project, and evaluate it using R after the first training.

If the R^2 obtained from the first training is less than 0.5, we will optimize the parameters for a second training. If the difference between two R^2 is greater than 0.05, we will adopt the results from the second round of training.

But if it is still less than 0.5, to address the issue of overdispersion in the athletes' medal acquisition data, we will use a negative binomial regression model for a third training.

Gold Predictive Model Expressions:

$$\begin{aligned}\log(\mu_{c,t}) = {} & \beta_0 + \beta_1 * \text{Historical\_Gold}_{c,t} + \beta_2 * \text{Sport\_Count}_{c,t} + \beta_3 * \text{Athele\_Count}_{c,t} + \beta_4 \\ & * \text{Host\_Flag}_{c,t} + \beta_5 * \text{Total\_Events\_Sport\_Year}_{c,t} + \beta_6 * \text{Olympic\_Number}_{c,t} \\ & + \beta_7 * \text{GreatCoach\_Proxy}_{c,t} + u_c\end{aligned} \tag{1}$$

Negative binomial distribution:

$$Y_{c,t} \sim \text{NegativeBinomial}(\mu_{c,t}, \alpha)$$

Medal Predictive Model Expressions:

$$\begin{aligned}\log(\mu_{c,t}^{\text{Total}}) = {} & \gamma_0 + \gamma_1 * \text{Historical\_Total}_{c,t} + \gamma_2 * \text{Sport\_Count}_{c,t} + \gamma_3 * \text{Athele\_Count}_{c,t} \\ & + \gamma_4 * \text{Host\_Flag}_{c,t} + \gamma_5 * \text{Total\_Events\_Sport\_Year}_{c,t} + \gamma_6 \\ & * \text{Olympic\_Number}_{c,t} + \gamma_7 * \text{Historical\_Gold}_{c,t} + v_c\end{aligned} \tag{2}$$

5.2.2Model training and evaluation

- Treat all data before 2028 as the training set, with 2028 serving as the test set.
- In the training set, fit the aforementioned negative binomial regression model separately for the number of gold medals and the total number of medals.

Model Evaluation

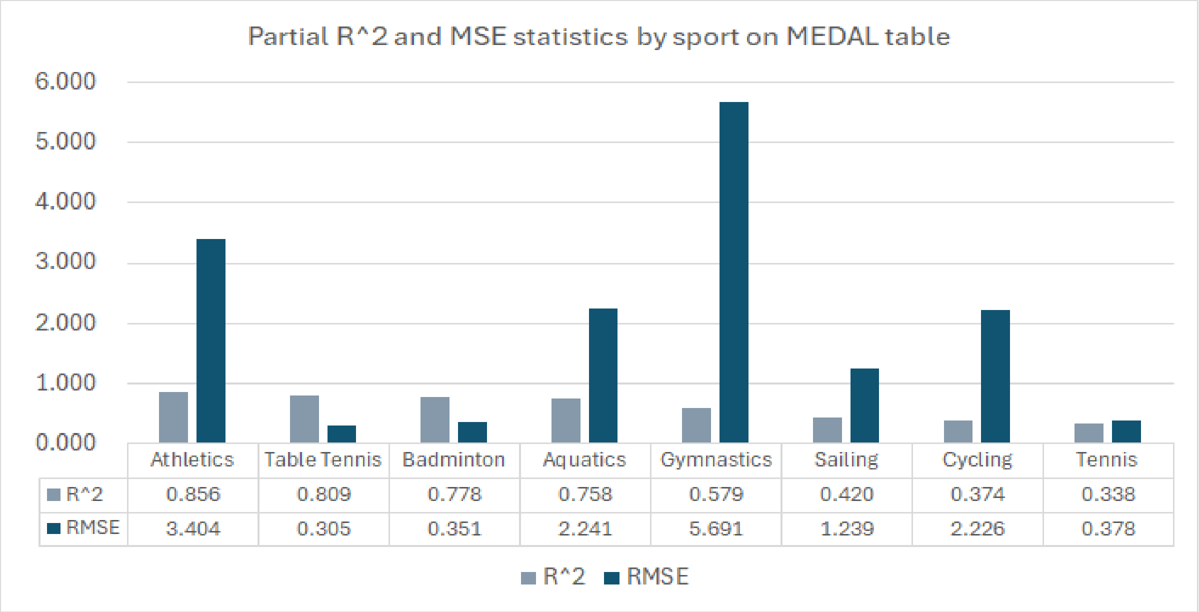
Use the training set data to calculate the Root Mean Square Error (RMSE) and the coefficient of determination (R^2) to evaluate the performance of the obtained models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2}$$

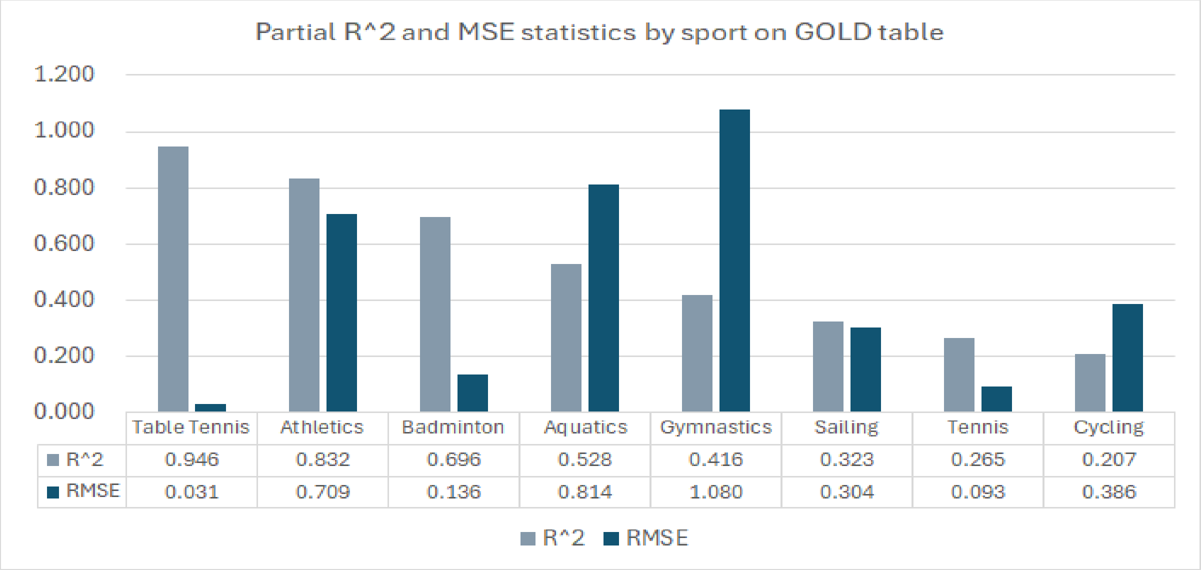
All GOLD model： R^2 = 0.458 RMSE = 0.524

All MEDAL model: R^2 = 0.610 RMSE = 2.082



Partial R^2 and MSE statistics by sport on MEDAL table

	Athletics	Table Tennis	Badminton	Aquatics	Gymnastics	Sailing	Cycling	Tennis
▨ R^2	0.856	0.809	0.778	0.758	0.579	0.420	0.374	0.338
■ RMSE	3.404	0.305	0.351	2.241	5.691	1.239	2.226	0.378



Partial R^2 and MSE statistics by sport on GOLD table

	Table Tennis	Athletics	Badminton	Aquatics	Gymnastics	Sailing	Tennis	Cycling
▨ R^2	0.946	0.832	0.696	0.528	0.416	0.323	0.265	0.207
■ RMSE	0.031	0.709	0.136	0.814	1.080	0.304	0.093	0.386

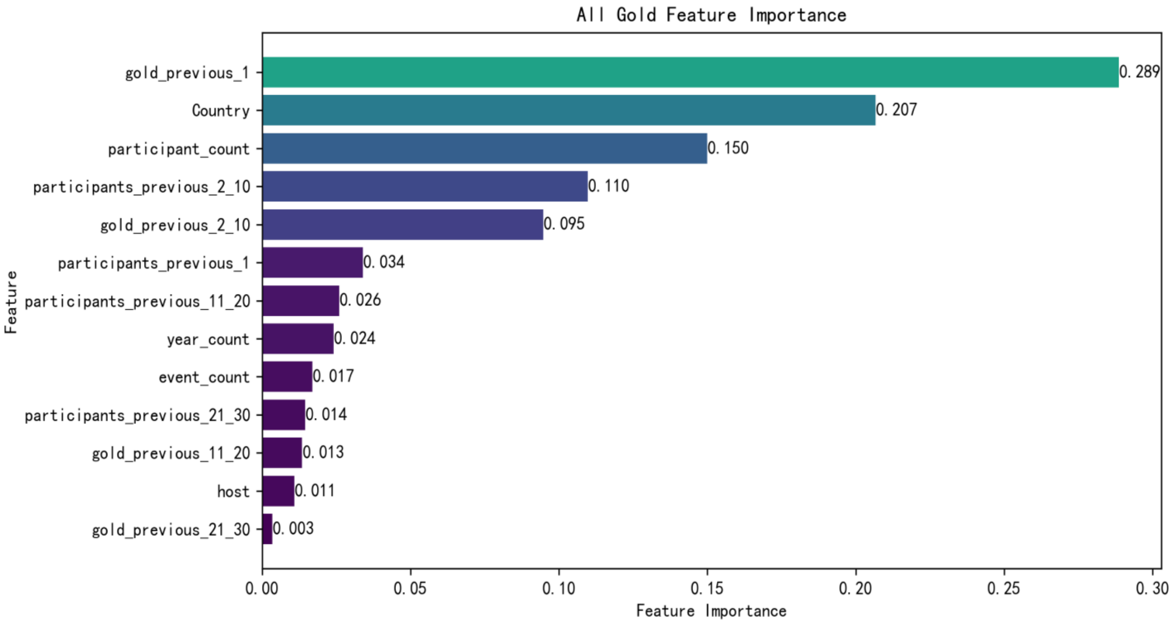
Figure 4 and 5 :The Particial Results of RMSE and R^2

5.2.3 Feature importance value assessment

Feature importance:

$$\textbf{Feature Importance}_{\mathbf{i}} = \frac{\text{Total Reduction in Impurity for Feature i}}{\text{Total Reduction in Impurity across all Features}}$$

The higher the importance value of the feature, the greater the contribution of the feature to the prediction of the predictor.



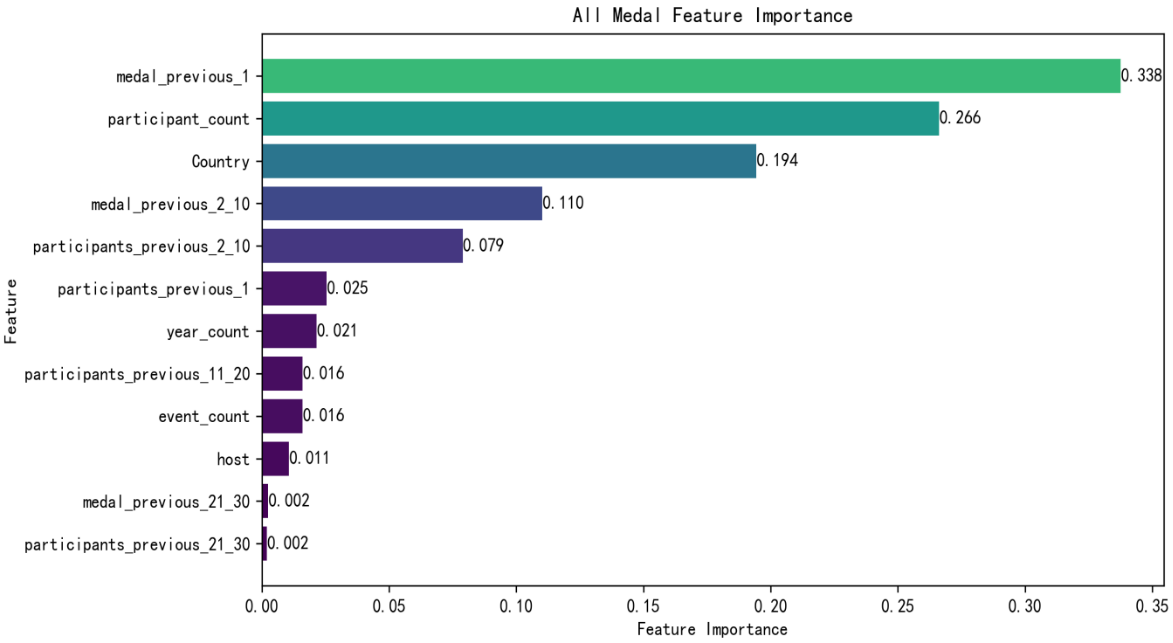


Figure 6 and 7:The results of feature importance

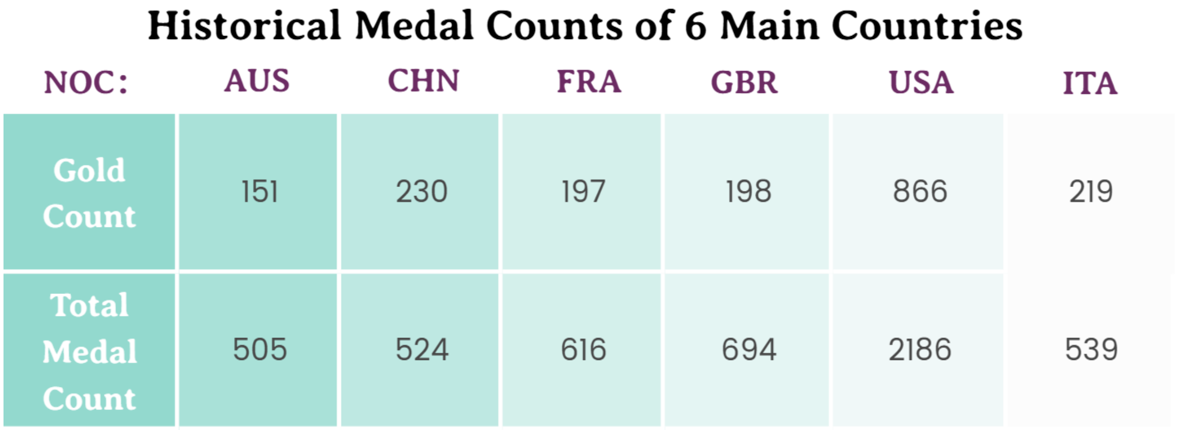
5.3 Results

5.3.1 Data preparation

Historical Gold and historical medal:

$$\text{History Gold}_{c,2028} = \sum_{t<2028} Y_{c,t}$$

$$\text{History Total}_{c,2028} = \sum_{t<2028} Y_{c,t}^{\text{Total}}$$



Historical Medal Counts of 6 Main Countries

NOC:	AUS	CHN	FRA	GBR	USA	ITA
Gold Count	151	230	197	198	866	219
Total Medal Count	505	524	616	694	2186	539

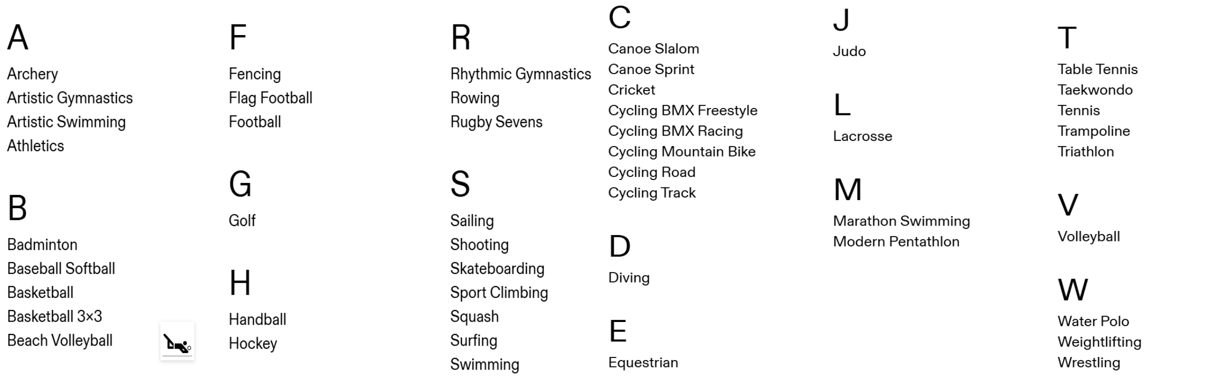
Figure8:Historical Medal Counts

The number of participating athletes and the number of events in each country are projected based on the number of participating athletes in the last two Olympic Games.

$$\text{Athlete\_Count}_{c,2028} = \frac{1}{2}\sum_{t=2020}^{2024} \text{Athlete\_Count}_{c,t}$$

$$\text{Sport\_Count}_{c,2028} = \frac{1}{2}\sum_{t=2020}^{2024} \text{Sport\_Count}_{c,t}$$

2028 Olympic Games are based on the projects currently announced by the Olympic officials.



A
Archery
Artistic Gymnastics
Artistic Swimming
Athletics

B
Badminton
Baseball Softball
Basketball
Basketball 3×3
Beach Volleyball

F
Fencing
Flag Football
Football

G
Golf

H
Handball
Hockey

R
Rhythmic Gymnastics
Rowing
Rugby Sevens

S
Sailing
Shooting
Skateboarding
Sport Climbing
Squash
Surfing
Swimming

C
Canoe Slalom
Canoe Sprint
Cricket
Cycling BMX Freestyle
Cycling BMX Racing
Cycling Mountain Bike
Cycling Road
Cycling Track

D
Diving

E
Equestrian

J
Judo

L
Lacrosse

M
Marathon Swimming
Modern Pentathlon

T
Table Tennis
Taekwondo
Tennis
Trampoline
Triathlon

V
Volleyball

W
Water Polo
Weightlifting
Wrestling

Figure 9:The event of 2028 Olympics

Host:Host of USA=1,Host of Others=0.

5.3.2 Model prediction

Using the trained negative binomial regression model, the predicted values and prediction intervals of the number of gold medals and total medals in 2028 were generated according to the variables given above.

Prediction Interval:

$$\left(\mu_{c,t} * e^{-z*\sqrt{\frac{1}{\mu_{c,t}}+\frac{\alpha}{\mu_{c,t}^2}}}, \mu_{c,t} * e^{z*\sqrt{\frac{1}{\mu_{c,t}}+\frac{\alpha}{\mu_{c,t}^2}}}\right)$$

where z is the critical value of the standard normal distribution.

5.3.3Calaculation Results

The prediction results are visible in the figures.

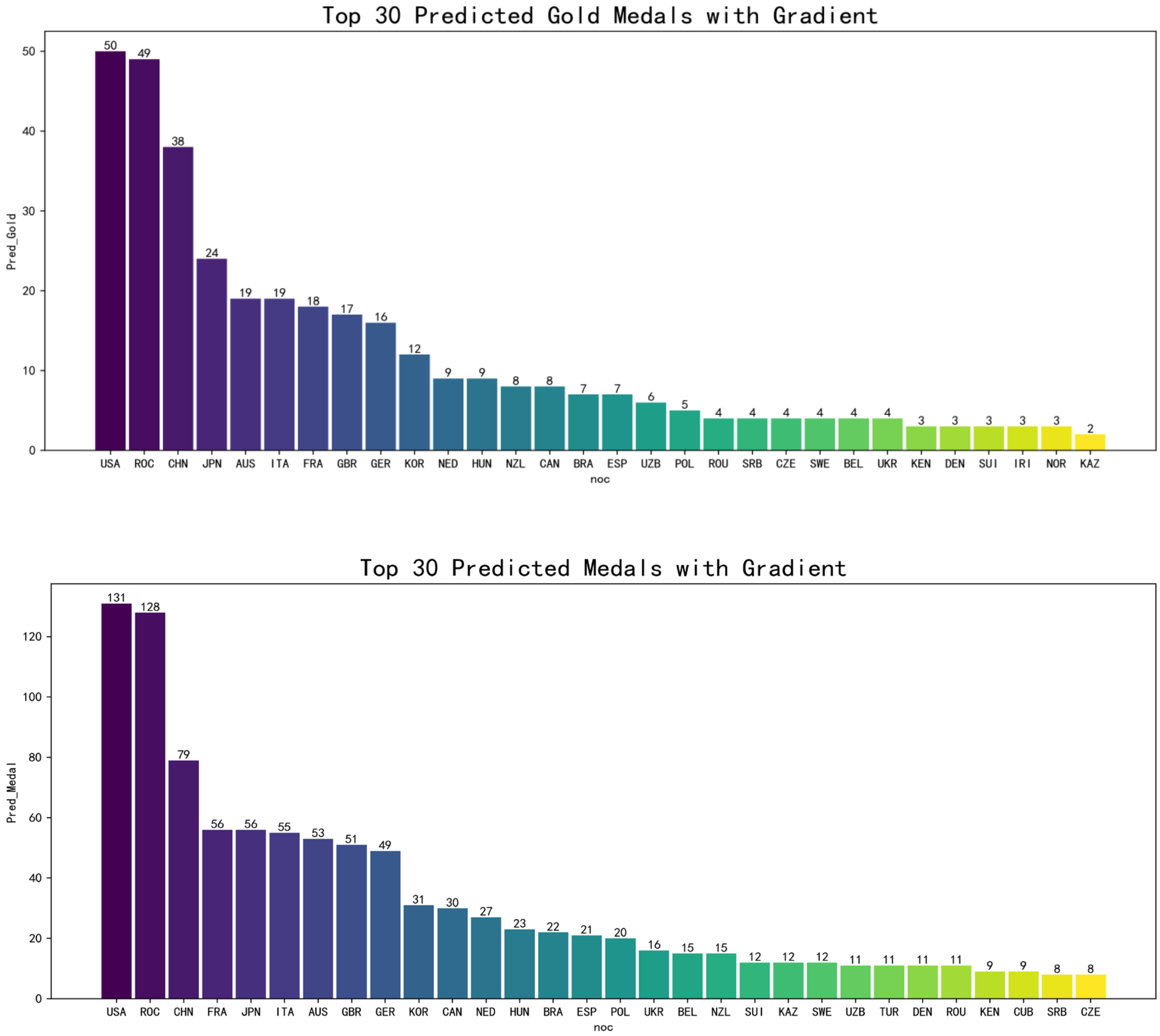


Figure 10 and 11: Top 30 Predicted Medals

6 Medal Breakthrough Prediction

Many countries have still never won an Olympic medal to date, but with the development of sports, an increasing number of countries have the potential to achieve their first medals. This possibility of a breakthrough can be based on analyses of historical participation in events and the number of athletes, as these data often reflect the emphasis that the country places on sports.

6.1 Identification of Countries Yet to Win Medals

After basic data processing, we can identify the countries shown in the figure below that have not yet won any medals. We will focus on analyzing the potential for these countries to achieve a medal breakthrough.

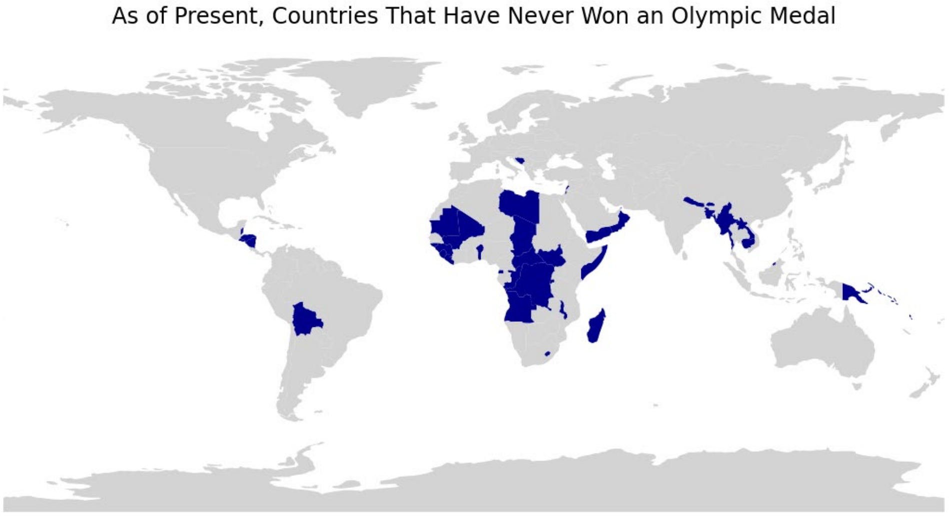

As of Present, Countries That Have Never Won an Olympic Medal

Figure 12 :World Map of Non-Medals Country

6.2 Data Observation and Results

Based on the results from Model I, we will find that the projected number of medals for these countries at the 2028 Los Angeles Olympics is not necessarily zero. We can obtain the expected number of medals for these countries; the greater the expected number of medals, the more likely they are to achieve a breakthrough.

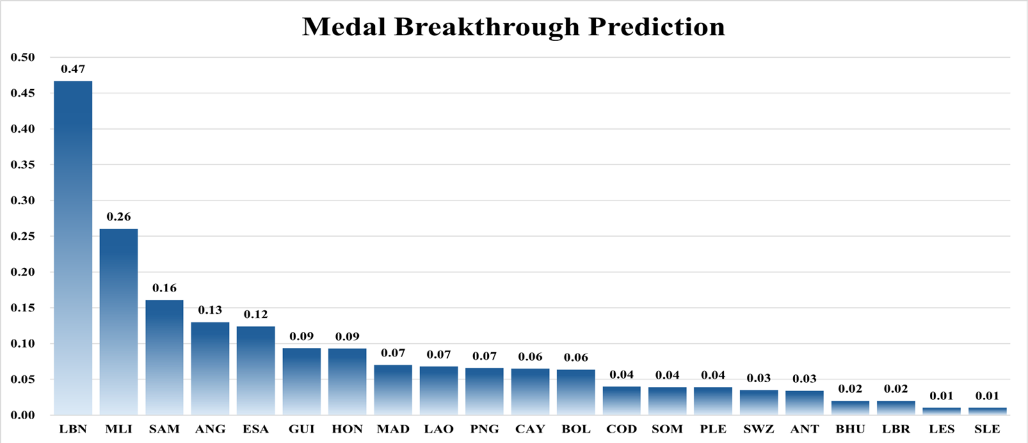


Figure 13:Medal Breakthrough Prediction

7 Model II: Coaching Effectiveness Analysis Model

7.1 Data Choose

Select two coaches: Lang Ping and Karch Kiraly. Both have coached in CHN (China), USA (United States), and ROU (Romania). Therefore, we will select data from these three countries.

Lang Ping: Coached the USA women's volleyball team in 2008 and the China women's volleyball team in 2013 (participated in the 2016 Olympics).

Bela Karolyi: Held a position in the USA in 1984, and coached in ROU in 1976 and 1980.

Thus, we will select data prior to their coaching tenures as the training set for the Random Forest model, which includes variables such as participant numbers, years of participation, award achievements, and more.

7.2Model Construction and Evaluation

Select the data from before the aforementioned star coaches began their tenures as the training set for the Random Forest model. The model will predict the performance of the project in the year if the coach had not been hired, based on the development trends prior to the coach's engagement.

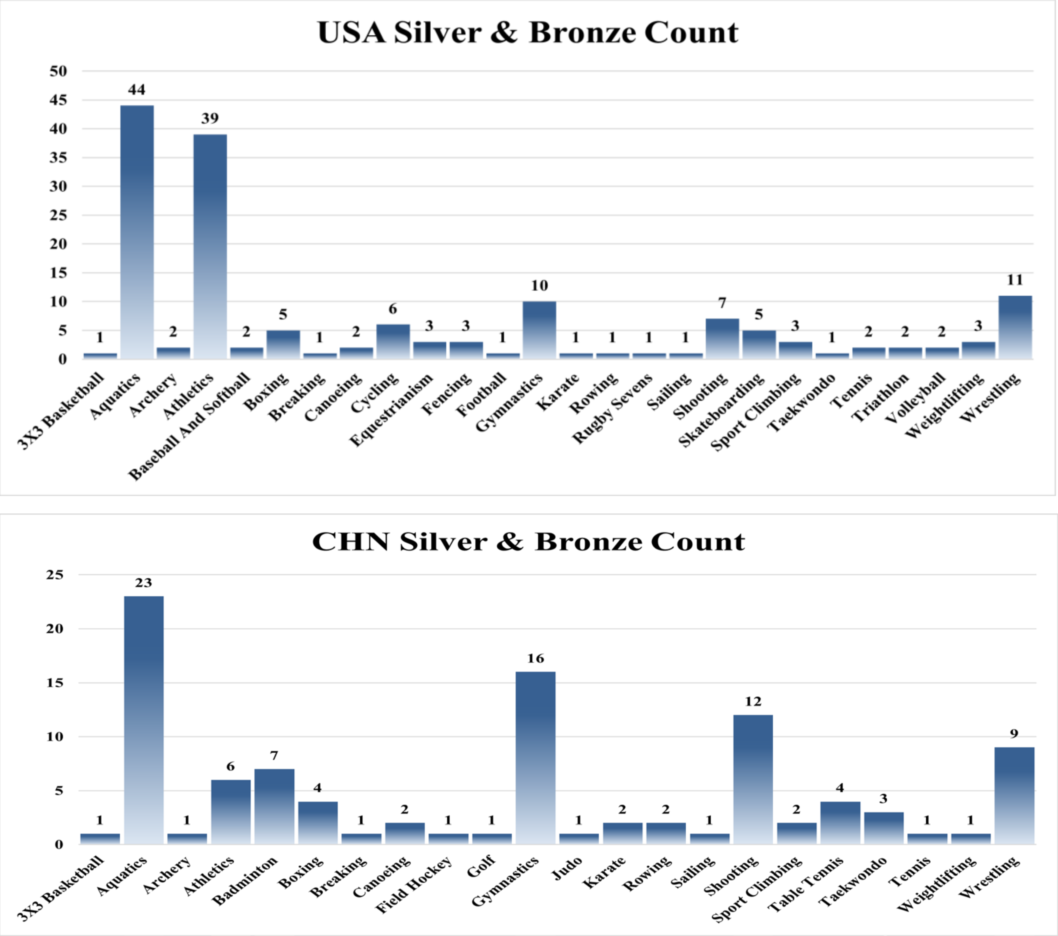


Figure 14 and 15:Silver and Bronze Count of 2 Countries

The difference between the predicted results and the actual outcomes is highly correlated with the coaching effect.

We will use $R^2$and RMSE to evaluate the model.

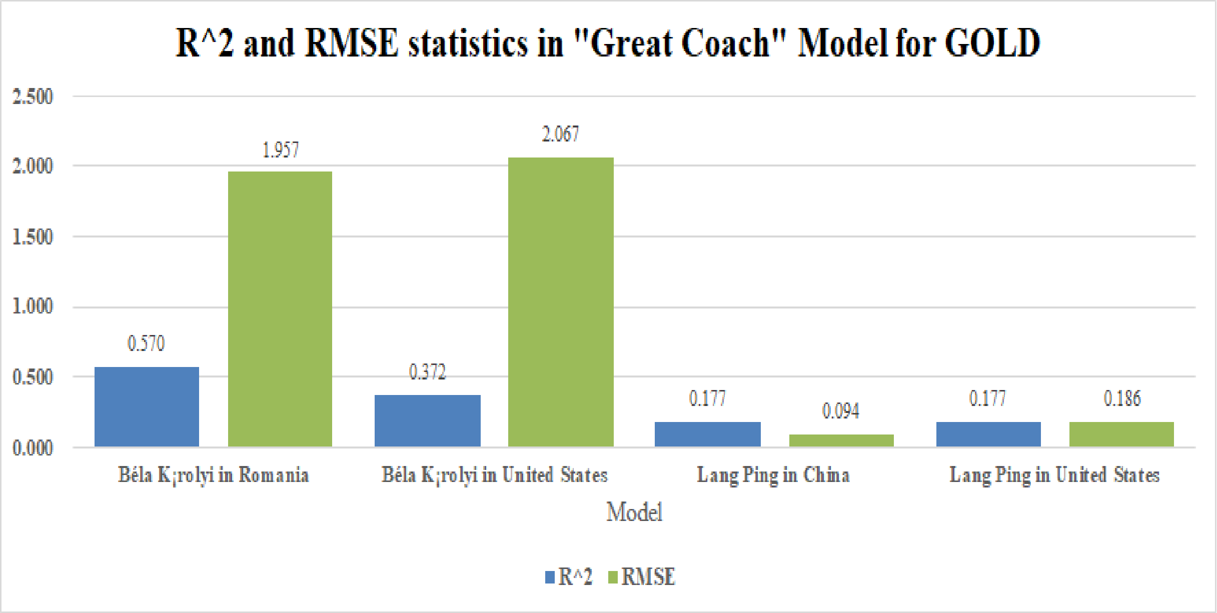
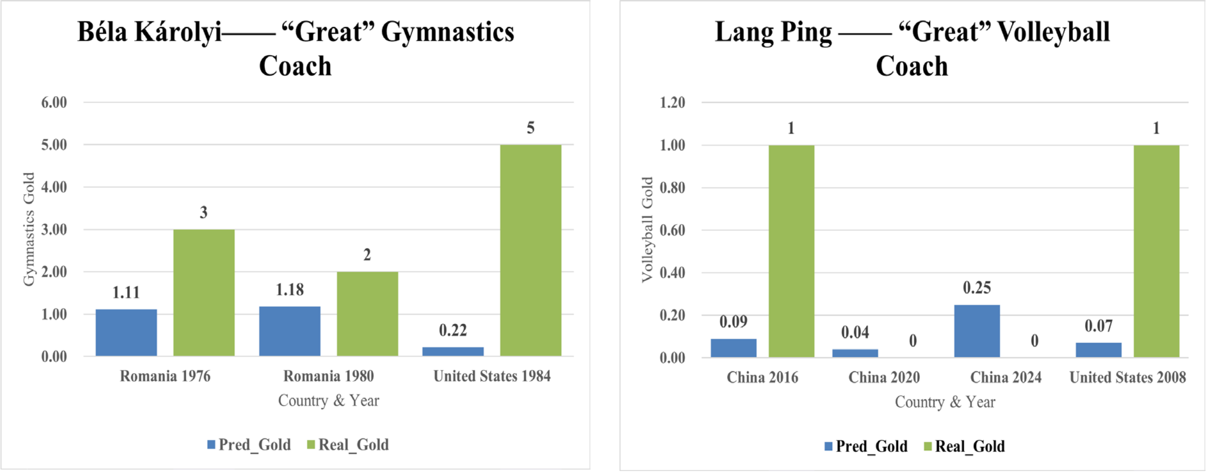


Figure 14:The result of RMSE and R^2

7.3 Results



(a) Karoly (b) Lang Ping

Figure 15: Great Coach

Blue represents the predicted number of medals, while green is the actual number of medals brought about by the star coaches, although there are many endogenous variables in this difference, it can still be seen that there is a very obvious coaching effect.

8 Conclusion

8.1 Summary of Results

8.1.1 Result of Problem 1

Based on the results of the model run, the predicted medal standings for the 2028 Los Angeles Olympics are as follows:

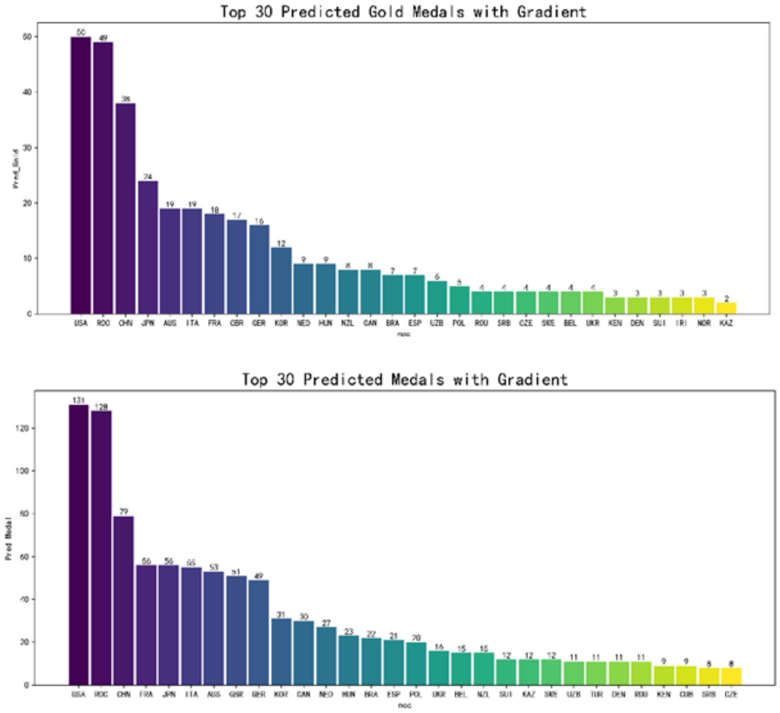


Figure 16: The Predicted Medal

The model performance metrics for some of the projects are as follows：

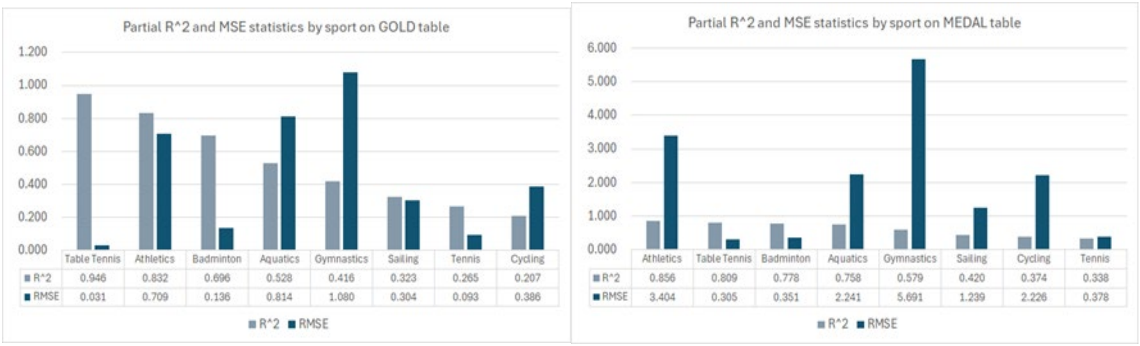
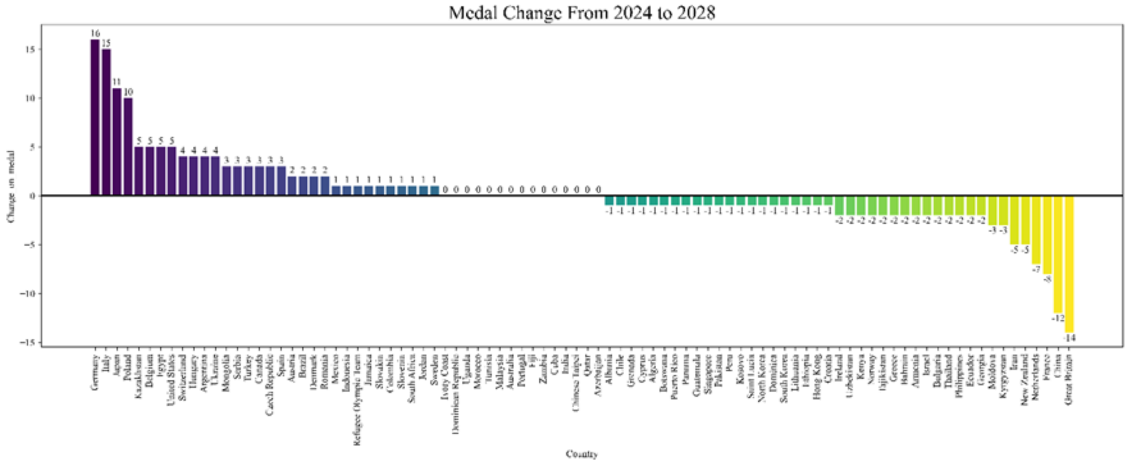


	Table Tennis	Athletics	Badminton	Aquatics	Gymnastics	Sailing	Tennis	Cycling
■ R^2	0.946	0.832	0.696	0.528	0.416	0.323	0.265	0.207
■ RMSE	0.031	0.709	0.136	0.814	1.080	0.304	0.093	0.386

	Athletics	Table Tennis	Badminton	Aquatics	Gymnastics	Sailing	Cycling	Tennis
■ R^2	0.856	0.809	0.778	0.758	0.579	0.420	0.374	0.338
■ RMSE	3.404	0.305	0.351	2.241	5.691	1.239	2.226	0.378

Figure 17:The results of R^2 and MSE

Based on the results of the model, we conclude that the number of medals for countries such as the United States and Germany is likely to rise, while the number of medals for countries such as China and the United Kingdom is likely to fall, as detailed in the chart below：



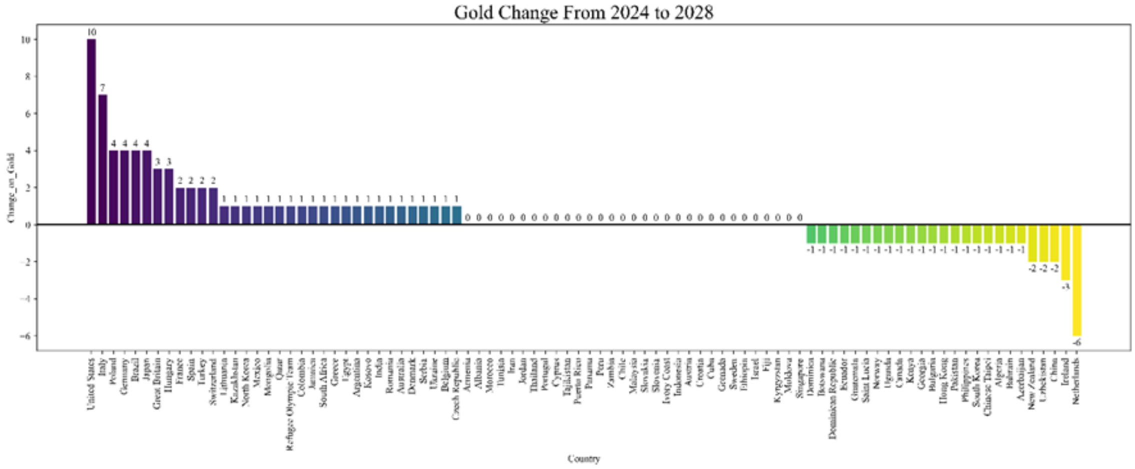


Figure 18:The predicted diffrence

According to the number of medals predicted by the model, LBN and MLI are more likely to achieve a breakthrough of 0 and get their first Olympic medal, with LBN having the highest likelihood of 47 percent, as detailed in the chart below:

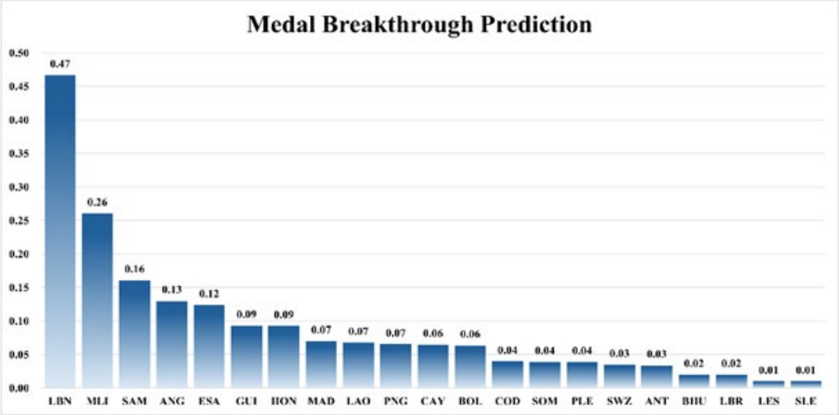


Figure 19:The Medal Breakthrough Predictions

Programs that are more important to most countries should have more participating countries and greater competition, so we filtered by the number of participating countries and then compared the variance of the winning countries in these programs, with the higher variance programs being more competitive, yielding programs such as Aquatics and Athletics as being more important, as detailed in the chart below:

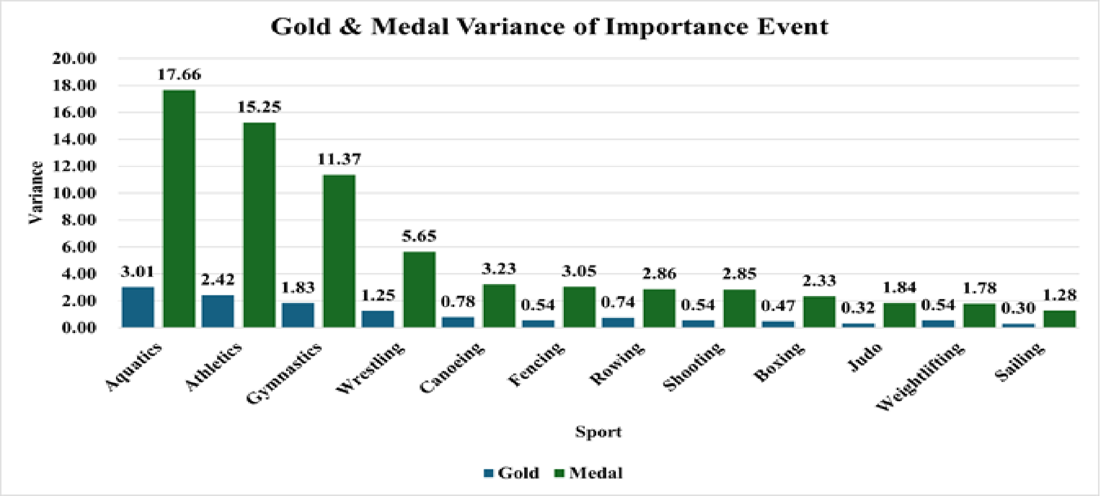
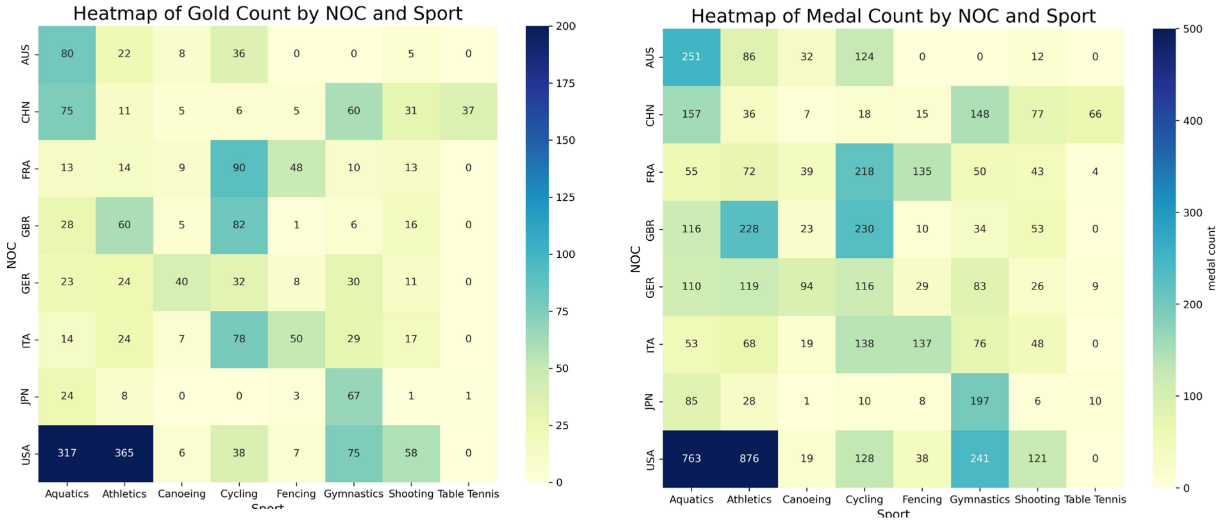


Figure 20:Gold and Medal Variance of Importance Event

The U.S. should select the programs in which it is more proficient and eliminate the program in which it is less proficient. By comparing the number of medals won by different countries in different sports, it is possible to identify the sports in which the U.S. excels and those in which it has obvious deficiencies. The following chart shows the number of medals won by some of the major countries in some of the sports, from which we can see that the U.S. could add some water sports and track and field and cut down on table tennis, among other sports. This method of choosing sports will allow the host country to participate in more strong events and more likely to win more medals.

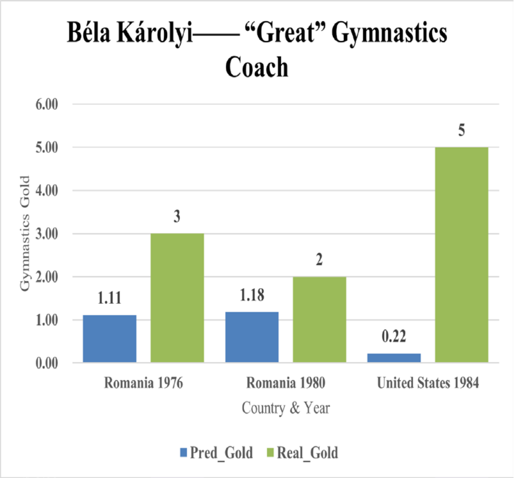


(a) Gold Count (b) Total Medal Count

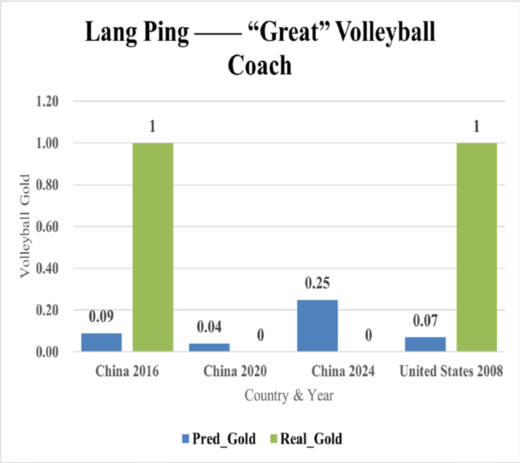
Figure 21:Heatmap of Medel Count/Gold Count by NOC and Sport

8.1.2 Result of Problem 2

By putting the relevant volleyball and gymnastics data from China, the United States, and Romania into a regression forest model to predict and validate the number of gold medals, we can predict the extent of Lang Ping's and Karolyi's influence on the outcome of the tournament, as shown in the figure below:



(a) Béla Károlyi



(b) Lang Ping

We then studied the silver and bronze medals won by the three countries mentioned above in the previous two Olympic Games (these can be seen as the more promising events in each country and the ones with the most significant coaching effects), and chose among them the team events that had won a higher number of medals. By filtering, the sport that needs the greatest coaching in the **United States is soccer, Romania's is rowing, and China's is field hockey.**

8.1.3 Result of Problem 3

In our previous study we summarized 12 sports with a high number of participating countries, most of which have a very small variance, suggesting that their medal winners are concentrated in a very small range, and that some of the countries that are weaker in this area can shift their focus to other sports that do not have as many participating countries. (See previous image for details)

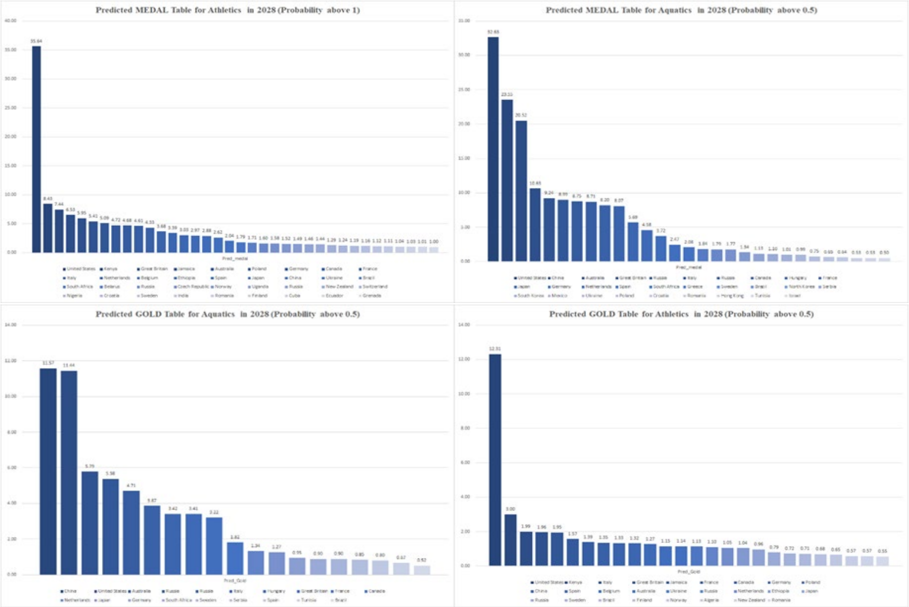


Figure 20: Simulation results of fishing company operations over the next 50 years

In these competitive and popular sports, some countries have strong competition for medal counts in some sports, and based on the projected medal counts, the U.S. can focus on countries with projected medal counts close to theirs and give them a targeted strategy.

8.2 Strengths

- **Our model has strong robustness and accuracy, using multi-model integration, selection and optimization mechanism.** The final prediction of the medal table R^2 is higher than 0.6, and a combination of random forest model, negative binomial regression model, grid search optimization is used. In the prediction of the medal table of each "sport", we select the model with the highest value of R^2 for prediction.
- In predicting the 2028 medal standings, we used an algorithm that predicts the medal standings for each sport **(using different models trained based on different sports)** and ultimately sums them by country. This is because we recognize that an algorithm that predicts the 2028 medal standings directly based on the historical trend of the overall medal standings is unscientific because the historical trend of the overall medal standings is not a time series and is not regular. However, it is feasible to predict the medal table for individual sport because the overall level of the country in a single sport is relatively stable and the historical medals and participation of the country in that sport can explain the level of the country's sport in that sport, or even serve as a proxy variable.
- **One-Hot Encoding o**f country names was trained as "features" for the model. This is because we know that the differences between countries due to GDP, population, etc. can have a significant effect on the distribution of the number of medals. However, since reference to additional data is not allowed, we try to take into account the influence of countries on the medal table by encoding the country names with One-Hot Encoding to improve the model interpretability.

8.3 Possible Improvements

- ◆ The analysis of medal predication can be more credible if we have more accueate data;
- ◆ Some approximate analysis methods are applied to model the predication of medals, which may lead to the situation contrary to the actual in extreme cases.

References

[1] Hu, Liangping, & Hu, Liangping (1955-). (2013). Nonlinear Regression Analysis and Intelligent Implementation with SAS. Beijing: Electronics Industry Press.J

[2] Han, Yong. (2008). An Overview of Olympic Culture: Ceremonies and Celebrations of the Olympic Games. Beijing: Beijing Sports University Press.

[3] Müller, & (De) Müller. (2018). Introduction to Machine Learning with Python. Beijing: People's Posts and Telecommunications Press.

[4] Layton, & (Aus) Layton. (2016). Introduction to Data Mining with Python: Theory and Practice. Beijing: Published by People's Posts and Telecommunications Press.

Appendices

Appendix 1

Introduce: Tools and software

Paper written and generated via Office 361.
Programming part and Graph generated using Spyder 6.0.1, Jupyter Notebook 7.2.2, Stata 18.0.

Appendix 2

Introduce:Model Code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
import os
from sklearn.model_selection import GridSearchCV

folder_path_A = r"C:\Users\Lucius\Desktop\data\train"
folder_path_B = r"C:\Users\Lucius\Desktop\data\pred"

xlsx_files_A = [f for f in os.listdir(folder_path_A) if f.endswith('.xlsx')]
xlsx_files_B = [f for f in os.listdir(folder_path_B) if f.endswith('.xlsx')]

pred_file_path = r"C:\Users\Lucius\Desktop\result\pred_2028_gold_sport_forests.xlsx"
performance_file_path =
r"C:\Users\Lucius\Desktop\result\performance_2028_gold_sport_forests.xlsx"

all_predictions = pd.DataFrame()
all_performance = pd.DataFrame()

non_zero_column_threshold = 1
threshold = 0.05
with open(os.path.join(folder_path_A, 'model_output.txt'), 'w') as output_file:
    for file in xlsx_files_A:

        if file in xlsx_files_B:
            file_path_A = os.path.join(folder_path_A, file)
            file_path_B = os.path.join(folder_path_B, file)
            data = pd.read_excel(file_path_A)
            if data.shape[0] < 10:
                print("The sample size is less than 10, skip using the Random Forest model.")
                output_file.write("The sample size is less than 10, skip using the Random Forest model.")
                continue

            features = ['participant_count','Country','event_count','host','year_count','gold_previous_1','participants_previous_1','gold_previous_2','participants_previous_2','gold_previous_3','participants_previous_3','gold_previous_4','participants_previous_4','gold_previous_5','participants_previous_5','gold_previous_6','participants_previous_6','gold_previous_7','participants_previous_7','gold_previous_8','participants_previous_8','gold_previous_9','participants_previous_9','gold_previous_10','participants_previous_10','gold_previous_11','participants_previous_11','gold_previous_12','participants_previous_12','gold_previous_13','participants_previous_13','gold_previous_14','participants_previous_14','gold_previous_15','participants_previous_15','gold_previous_16','participants_previous_16','gold_previous_17','participants_previous_17','gold_previous_18','participants_previous_18','gold_previous_19','participants_previous_19','gold_previous_20','participants_previous_20','gold_previous_21','participants_previous_21','gold_previous_22','participants_previous_22','gold_previous_23','participants_previous_23','gold_previous_24','participants_previous_24','gold_previous_25','participants_previous_25','gold_previous_26','participants_previous_26','gold_previous_27','participants_previous_27','gold_previous_28','participants_previous_28','gold_previous_29','participants_previous_29']
            target = 'sport_gold'
            X = data[features]
            y = data[target]

            X = X.fillna(0)
            X = X.loc[:, (X != 0).any(axis=0)]
            X = pd.get_dummies(X, columns=['Country'], drop_first=True)
            new_features = list(X.columns)
            non_zero_column_count = X.shape[1]

            if non_zero_column_count > non_zero_column_threshold:

                X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
                param_grid = {
                    'n_estimators': [50, 100, 200],
                    'max_depth': [None, 10, 20, 30],
                    'min_samples_split': [2, 5, 10],
                    'min_samples_leaf': [1, 2, 4],
                }

                rf_model = RandomForestRegressor(random_state=42)
                rf_model.fit(X_train, y_train)
                y_pred = rf_model.predict(X_test)
                mse = mean_squared_error(y_test, y_pred)
                r2 = r2_score(y_test, y_pred)
                grid_search = GridSearchCV(estimator=rf_model,
        param_grid=param_grid, cv=3, scoring='r2', verbose=0, n_jobs=-1)
                grid_search.fit(X_train, y_train)
                best_params = grid_search.best_params_

                best_rf_model = grid_search.best_estimator_
                y_pred = best_rf_model.predict(X_test)

                mse_2 = mean_squared_error(y_test, y_pred)
                r2_2 = r2_score(y_test, y_pred)

                if r2 < 0.5 and r2_2 < 0.5:
                    reg_model = LinearRegression()
                    reg_model.fit(X_train, y_train)

                    y_pred = reg_model.predict(X_test)

                    mse_3 = mean_squared_error(y_test, y_pred)
                    r2_3 = r2_score(y_test, y_pred)

                    if r2_3 >= max(r2,r2_2):
                        r2_use = r2_3
                        mse_use = mse_3
                        feature_importances = reg_model.coef_
                        model_to_use = reg_model

                    else:
                        if (r2_2 - r2) > threshold:
                            model_to_use = best_rf_model
                            r2_use = r2_2
                            mse_use = mse_2
                            feature_importances =
        best_rf_model.feature_importances_
                            sorted_indices = feature_importances.argsort()

                        else:
                            model_to_use = rf_model
                            r2_use = r2
                            mse_use = mse
                            feature_importances = rf_model.feature_importances_
                            sorted_indices = feature_importances.argsort()

                else:
                    if (r2_2 - r2) > threshold:
                        model_to_use = best_rf_model
                        r2_use = r2_2
                        mse_use = mse_2
                        feature_importances = best_rf_model.feature_importances_
                        sorted_indices = feature_importances.argsort()

                    else:
                        model_to_use = rf_model
                        r2_use = r2
                        mse_use = mse
                        feature_importances = rf_model.feature_importances_
                        sorted_indices = feature_importances.argsort()

                data_B = pd.read_excel(file_path_B)
                if data_B.shape[0] == 0:
                    continue

                X_B = pd.get_dummies(data_B, columns=['Country'], drop_first=True)
                X_B = X_B[[col for col in new_features if col in X_B.columns]]
                train_columns = X.columns
                pred_columns = X_B.columns
                missing_columns = set(train_columns) - set(pred_columns)
                for col in missing_columns:
                    X_B[col] = 0
                X_B = X_B[train_columns]

                y_pred_B = model_to_use.predict(X_B)

                predictions_df = pd.DataFrame({
                    'Country':data_B['Country'],
                    'noc':data_B['noc'],
                    'Year':2028,
                    'sport':data_B['sport'],
                    'Pred_Gold': y_pred_B,
                })
                all_predictions = pd.concat([all_predictions, predictions_df], ignore_index=True)
                performance = []
                sorted_indices = feature_importances.argsort()

                for idx in sorted_indices[::-1]:
                    feature_name = X.columns[idx]
                    importance = feature_importances[idx]
                    sport = data_B['sport'].iloc[0]
                    performance.append({
                        "sport": data_B['sport'].iloc[0],
                        "r2": r2_use,
                        "mse": mse_use,
                        "feature": feature_name,
                        "feature_importance": importance
                    })
                performance_df = pd.DataFrame(performance, columns=['sport', 'r2',
'mse', 'feature', 'feature_importance'])
                non_zero_importances = performance_df[performance_df['feature_importance'] != 0]
                all_performance =
pd.concat([all_performance,non_zero_importances],ignore_index=True)

        else:
            print('file not found in B, skip')

with pd.ExcelWriter(pred_file_path, mode='w', engine='openpyxl') as writer:
    all_predictions.to_excel(writer, index=False, sheet_name='Predictions')
with pd.ExcelWriter(performance_file_path, mode='w', engine='openpyxl') as writer:
    all_performance.to_excel(writer, index=False, sheet_name='Performance')
print('Done.')
```