Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2510185

Maximizing Medal Performance: Athlete and Event Potential Indices

Summary

As the highest-level global sporting event, the Olympic Games attract worldwide attention. Olympic achievements not only represent the limits of human athletic performance but also reflect comprehensive national strength.

First, after **data preprocessing**, we introduce the **Athlete Potential Index (API)** to quantify athlete performance. We define the API as an exponentially decaying function of **Medals won by an athlete/team (PWM)**. The PWM exhibits a linear relationship with the number of Gold, Silver, and Bronze medals. Following a comparison of various models, we select a **Random Forest (RF)** model optimized using **Grid SearchCV**. The optimal hyperparameters are **[max_depth=10, min_samples_split=4, n_estimators=180]**, achieving an R-squared (R^2) value of **0.8792**. Consequently, we employ the RF model with **bootstrapping** to generate 95% confidence intervals for medal predictions. Our analysis suggests that the **United States** is most likely to improve its medal count, while **South Korea** is most likely to experience a decrease.

Then, to predict medal acquisition by medal-less nations/regions, we employ a **BP neural network classifier** after comparing several classification models. The classifier achieves an Area Under the ROC Curve (AUC) of **0.94**. We apply this model to predict medal acquisition for 77 nations/regions. The nations/regions with the highest probability of winning their first medal are **Guam (GUM)** and **ESA**, with Guam having a probability of **0.153**.

Subsequently, we analyzed the six nations/regions' participation and medal counts across different events. Utilizing **the feature importance attribute** of the Random Forest model, we predict medal counts and derive the importance of various sports for each nation (Figure 11). **Swimming** emerges as a significant sport for most nations analyzed.

Next, to investigate the impact of "Great Coach" we compile a list of renowned coaches in sports such as gymnastics, athletics, and swimming. We use the Random Forest model to predict changes in medal counts attributed to these coaches. Our results indicate that "Great Coach" contribute up to **28.8%** to performance improvement. Specifically, at an average athlete API of 70, the presence of a "Great Coach" is associated with an increase of 0.2 gold medals and 1.8 total medals. We introduce the **Event Potential Index (EPI)** to quantify the potential improvement after hiring a "Great Coach". We recommend that **China** consider hiring a "Great Coach" for **women's gymnastics**, **Germany** for **men's shooting**, and **Jamaica** for **men's athletics**, which is projected to yield an increase of approximately **1.6 medals**.

Finally We made new discoveries and provided suggestions on (1)Number of participants; (2)Events with **negative importance**; (3)Projects suitable for hiring great coaches. Next, by varying the dataset size, we assessed the impact on prediction and information gain. we observed minimal changes in both. This demonstrates the model's strong robustness.

Keywords: Athlete Potential Index; Event Potential Index; Random Forest; BP Neural Network Classifier

Contents

1 Introduction

1.1 Background

The Olympic Games are the world's most prestigious sporting event, and in most of the time, becoming an Olympic champion is regarded as the highest honor in sports[1]. This pursuit of excellence naturally fosters intense competition between nations and regions, with the medal count serving as a key indicator of sporting prowess on the global stage.

At the recently concluded Paris 2024 Olympic Games, the United States and China led the medal count (tied for the most gold medals), followed by Japan, Australia, and host nation France[2]. Notably, several countries, including Albania, achieved historic first-time Olympic medals.

Some studies predict medal counts based on a country's political and economic context[3], while others focus on the recent performance of athletes at the Games[4]. These approaches have consistently achieved high accuracy. In this study, we aim to explore patterns in medal counts using historical Olympic data.

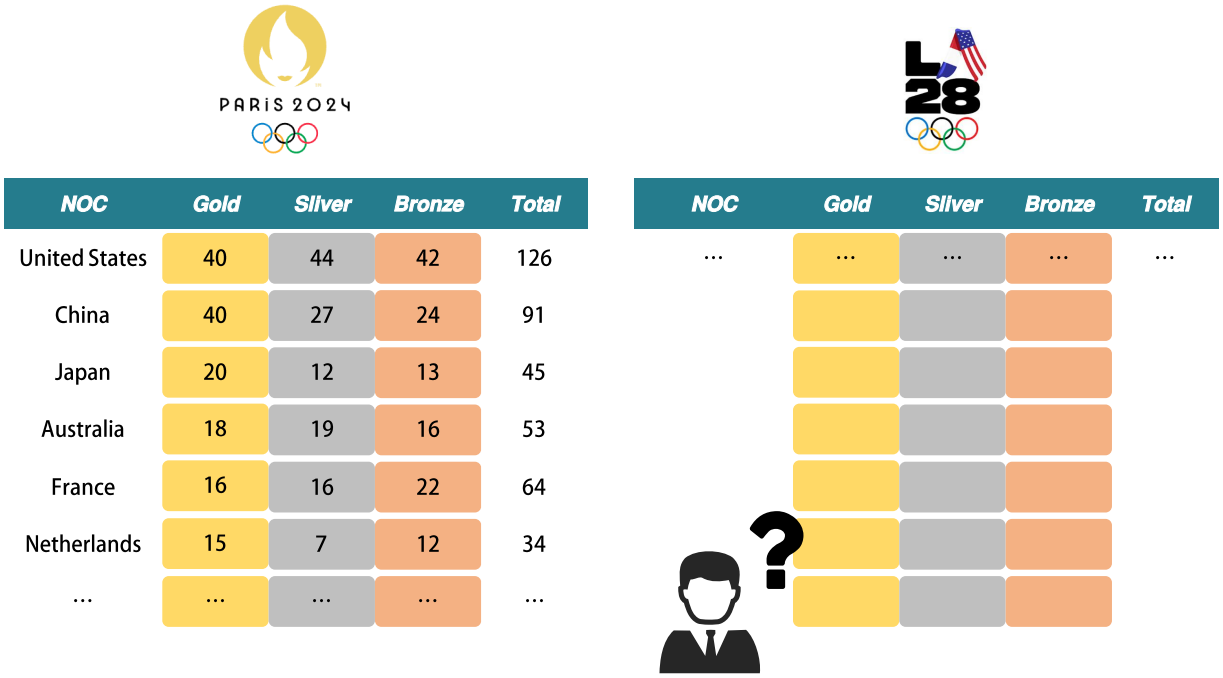

Figure 1: From Paris 2024 to LA 2028: Medal Table Prediction

1.2 Restatement of Problem

The challenge is to analyze and predict the medal table for the 2028 Los Angeles Olympics using publicly available data from all previous Summer Olympics. In addition, other patterns, including the "Great Coach" effect, should be explored. The specific questions are:

- **Build a Medal count prediction model:**
  The model should include at least gold medals and total medals. Measure the performance

of the model and estimate the uncertainty or accuracy of the prediction.

- **Specific Event type analysis:**
  Consider the number and types of events, explore the relationships between events, and the number of medals won by each country in each event. Examine the most dominant events in each country, and analyze the reasons and impact on the results.

- **Model prediction:**

  1. Predict the medal table for the 2028 Summer Olympics in Los Angeles and give a prediction interval for the result. Analyze countries that may improve or decline.
  2. Focus on countries/regions that have not yet won a medal and predict their probability of winning their first medal at the 2028 Olympics.

- **Explore the existence of the "Great Coach" effect:**
  Look for changes due to the "Great Coach" effect and estimate its impact on medal changes. Select three countries, determine the sports in which they should consider investing in a great coach, and estimate the impact.

- **Explore other patterns:**
  Analyze other patterns related to the number of medals displayed by the model. Use these findings to inform the Olympic committees of various countries.

1.3 Our work

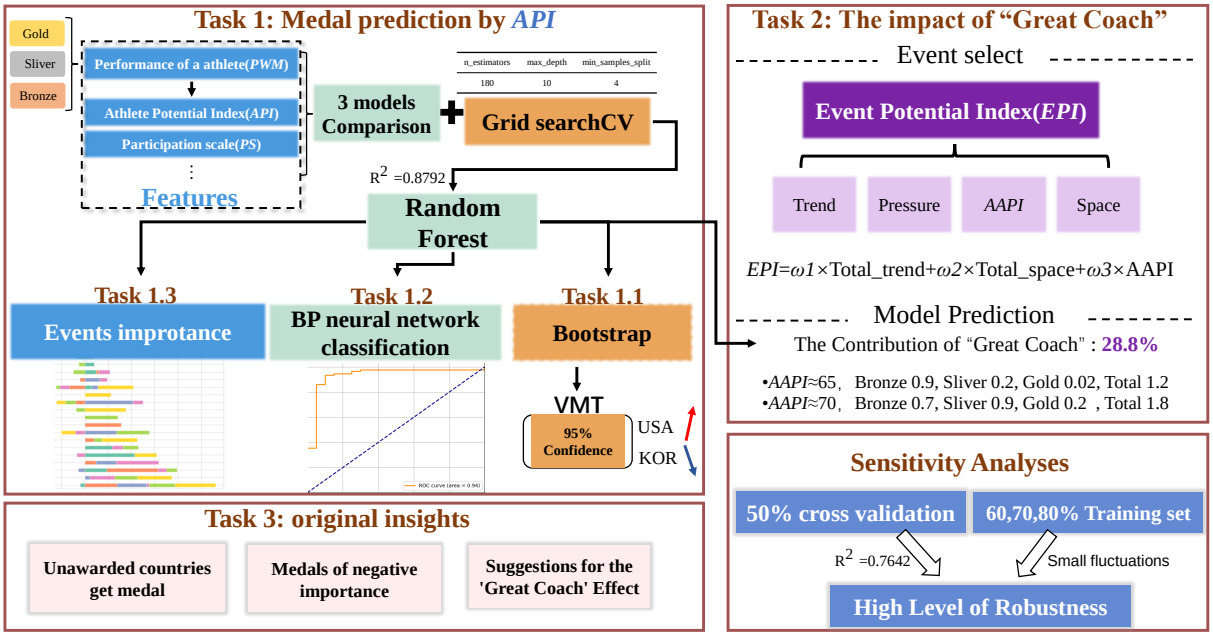

Figure 2: Our Work

2 Preparation for Modeling

2.1 Assumptions

We make the following assumptions to complete our model. In addition, we will refine these simplified assumptions later.

- We assume that **an athlete's competitive ability can be fully inferred from the historical data provided.** This is because the given dataset only contains information about a player's participation in previous Olympic Games and the medals he won (gold, silver, or bronze). We believe that a player's ability based on this information is sufficient.
- We assume that **the competitive ability of an athlete or team remains constant over the four years between games.** This is because we cannot obtain any other information about athletes between Games from the dataset provided.
- We assume that **the probability of a country that has never won a medal to win a medal is only related to the historical data of previous participation.** This is because there is no more realistic data on countries that have never won a medal in the data set provided. Although this may be a relatively important factor.

2.2 Notations

Table 1: Notations Table

Symbol	Description
PES	Performance score of an athlete/team
PWM	Medals won by an athlete/team
HA	Host advantage
N_i	The Olympic Games of number *i*
API	Athlete Potential Index
EPI	Event Potential Index
PS	Scale of Participation (by NOC)

2.3 Data Processing

After observing the datasets provided by the topic, *summerOly_athletes.csv*, *summerOly_hosts.csv*, *summerOly_medal_counts.csv* and *summerOly_programs.csv* contain some abnormal or missing values. The following describes the data processing of these abnormal and missing values. Further data processing will be done as needed in the later text. The processing procedures are shown in Figure 3.

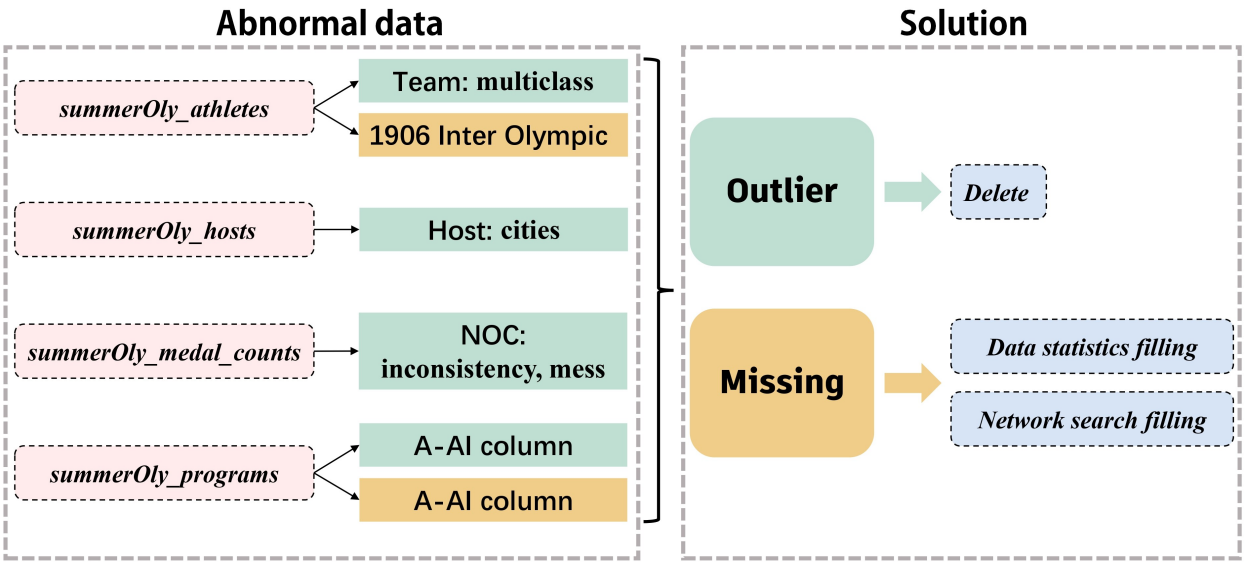

Figure 3: Data processing

SummerOly_athletes

The data in the ***Team*** column of the *SummerOly_athletes.csv* file is misaligned or garbled.

We noticed that the 1906 Olympics were an intercalated Olympics (unofficial Olympics), and the results were not officially recognized by the International Olympic Committee (IOC). We ultimately decided not to use any data from 1906.

summerOly_hosts

The ***Host*** column in the *summerOly_hosts.csv* dataset contains both countries and cities. To obtain precise information on the host countries, we performed a column-splitting operation to separate the data accordingly.

summerOly_medal_counts

In the *summerOly_medal_counts.csv* dataset, the ***NOC*** column contains 72 instances of garbled data marked as "??". These entries were removed directly.

Additionally, inconsistencies in country names, such as "Great Britain" in this dataset versus "United Kingdom" in other tables, caused mismatches in host information and athlete data. To address this, we standardized the country names across all datasets.

summerOly_programs

The *summerOly_programs.csv* file contains missing values and garbled data marked as "?". Missing or garbled entries in columns A–C were corrected using online resources. For columns beyond E, garbled data were removed, and missing values were confirmed to be zeros based on ***Total events*** statistics.

Significant fluctuations in medal rankings for 1904, 1908, and 1984, likely due to geopolitical factors, led to the exclusion of these years from the analysis.

3 Task 1: Medal Count Prediction Model

The flowchart for Task 1 is as follows:

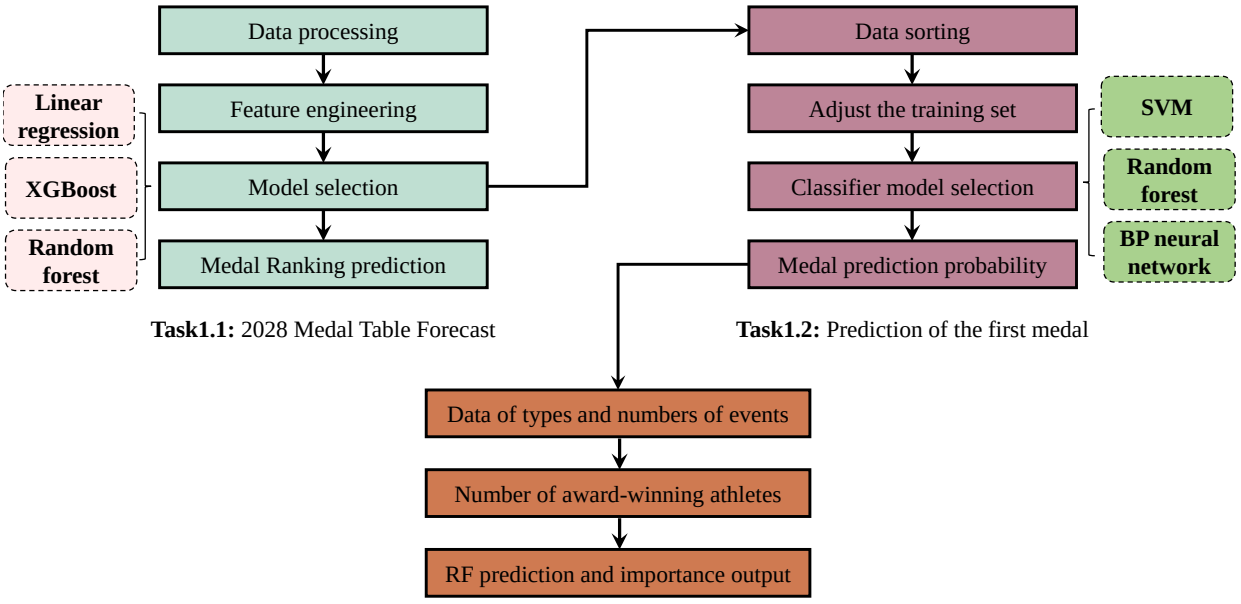

Figure 4: Flowchart for Task 1

3.1 Feature Engineering

3.1.1 Athlete Potential Index (API)

Inspired by Ruiz and Mario's (2024) research on the Olympic performance index of countries[5], we developed the Athlete Potential Index (API) based on athletes' previous Olympic medal performances to describe athletes' dynamic performance across Olympic Games. This potential index describes an athlete's past performance and the results they can achieve in the current Olympic event, dynamically adjusted for each Olympic Games.

Note that the dataset only provides information about an athlete/team for the last few Olympic Games. Combining relevant papers and parameter tuning practices, we define the API as follows:

$$API = \sum_{j=1}^{i} PWM_j \times \left(\frac{1}{2}\right)^{i-j} \tag{1}$$

where i represents the number of the next Olympic Games, and j represents the number of Olympic Games with recorded results.

$$PWM = 100 \times NG + 90 \times NS + 80 \times NB + 60 \times NM \tag{2}$$

where,

- NG represents the number of gold medals,

- NS represents the number of silver medals
- NB represents the number of bronze medals
- NM represents the number of events without medals
- If the athlete has never participated in a competition, their PWM is 0.

3.1.2 Participation Scale (PS)

The scale of a country's participation in the Olympic Games is a factor that is often taken into account[3]. According to the number of athletes representing the country in the Olympic Games, we divide the participation scale into five levels, denoted by 1-5, as defined below:

Table 2: Participation Scale (PS) based on Athlete Number

Athlete Number	Participation Scale (PS)
0-9	1
10-49	2
50-99	3
100-149	4
150 or more	5

3.2 Model Selection

After data processing and feature engineering, we obtain the following features:

Table 3: Feature Reference in Model 1

Symbol	Description
Year	The year of the Olympic Games
N_A	Number of Athletes
N_E	Number of Events
AAPI	Average of API
PS	Scale of Participation (by NOC)
HS	Host advantage
N_G	Number of Gold
N_S	Number of Silver
N_B	Number of Bronze
Total	Total medals

We compiled the above characteristics of the Olympics for five countries: Japan, the United States, France, China, and Great Britain, and formed a dataset to train the model.

The correlation matrix for the current feature mapping is shown in Figure 5.

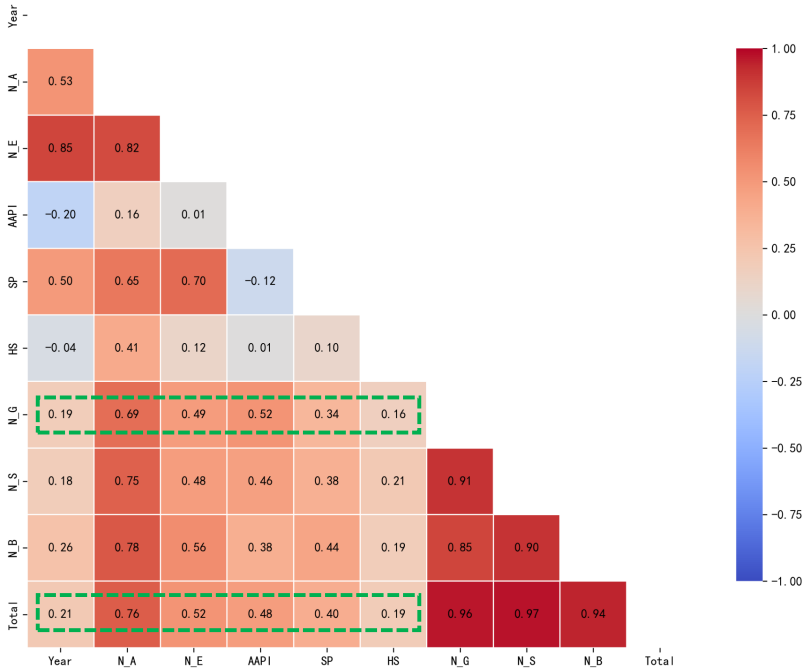

Figure 5: Correlation matrix of the characteristics of model 1

As can be seen from the image, for N_G, N_S, N_B, and Total, the six selected features are all correlated to some extent. Therefore, we select all of the above features to train the model.

We evaluated multiple models to select the best for prediction. Linear regression is chosen for its simplicity and interpretability. Random forests are selected for their strong predictive power and interpretability. XGBoost, known for its excellent performance, is also used for prediction.

For both random forests and XGBoost, we use a grid search to determine the best hyperparameters. A total of 27 combinations of three hyperparameters are set for random forests and 54 combinations for XGBoost. The optimal parameters are:

Table 4: Random Forests Hyperparameters

n_estimators	**max_depth**	**min_samples_split**
180	10	4

Table 5: XGBoost Hyperparameters

eta	**max_depth**	**subsample**	**colsample_bytree**
0.1	6	0.8	0.8

The fitting results of the three models' predictions for 70% of the training set and 30% of the test set are shown in Figure 6.

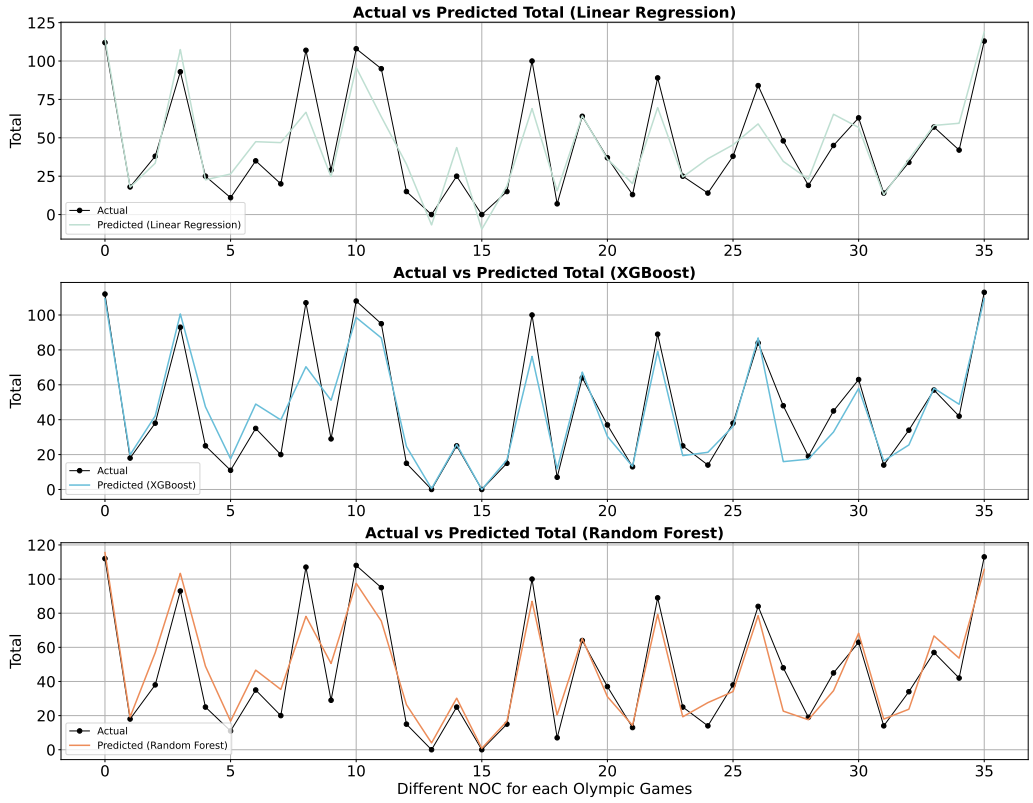

Figure 6: Fitted prediction curves for Linear Regression, XGBoost, and Random Forests (from top to bottom)

Additionally, we tested the stability of model predictions using 5-fold cross-validation.

Table 6: Performance Evaluation of Three Prediction Models

Model	**R^2**	**RMSE**	**5-fold Cross-Validation R^2**
Linear Regression	0.8063	15.44	0.7066
XGBoost	0.8764	12.33	0.7597
Random Forest	0.8792	12.19	0.7642

The R^2 and RMSE results for Random Forest and XGBoost are similar, with R^2 exceeding 0.85. Given the better interpretability of random forest regarding feature importance, we select the grid-searched Random Forest as the primary model.

In 5-fold cross-validation, both Random Forest and XGBoost achieve R^2 values above 0.75, demonstrating the robustness of these models. However, examining the R^2 results for each fold reveals occasional instances of poor predictive performance. We hypothesize that increasing the size of the training set improves this. Additionally, the size of the test and training sets may also influence the RMSE values.

3.3 Task 1.1: 2028 Olympic VMT Prediction

3.3.1 Grid SearchCV - Random Forest Model

Combining the data from the medal tables of recent Olympic Games, we select the top-ranked countries to make predictions for. For the medal table predictions, we will only predict medals for the major award-winning countries mentioned above. Note that Russia did not participate in the Olympics. The flowchart for Task 1.1 is as follows:

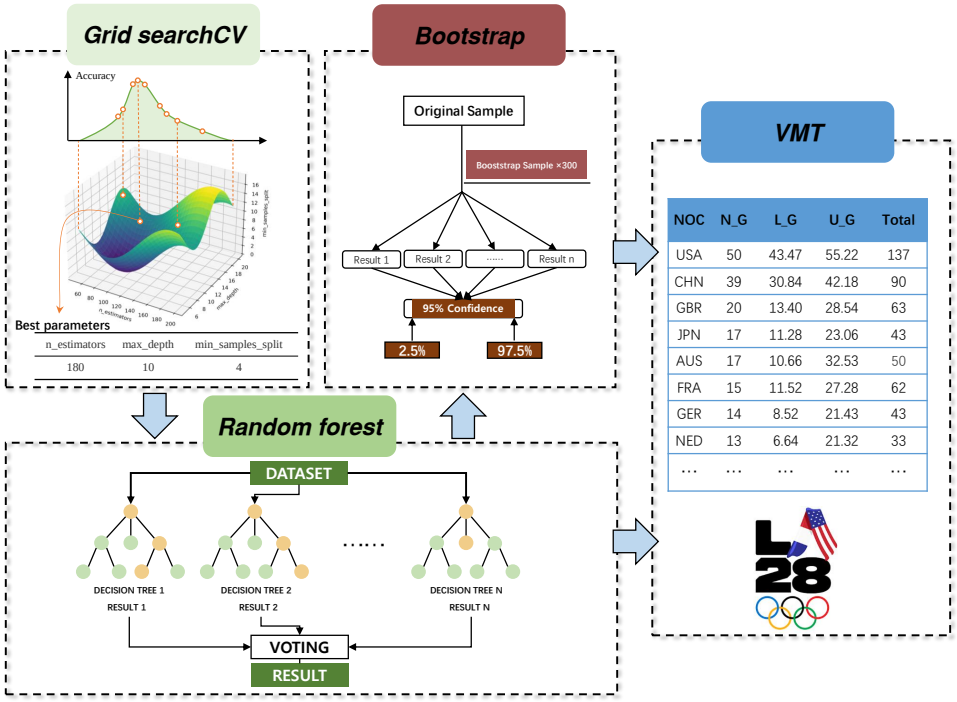

Figure 7: Flowchart for Task 1.1

3.3.2 Prediction Results

The random forest model is used to predict the medal rankings of each country in the 2028 Los Angeles Summer Olympics. We assume that the number of participants (N_A) and the number of events (N_E) are increased in proportion to the additional events in 2028. The prediction results are shown in Table 7:

The medal rankings are based on the predicted counts of gold, silver, and bronze medals in descending order. Integer values are rounded using the "round half to even" method. Due to space constraints, only a selection of countries is displayed in the medal table, while many others are not shown. However, the same methodology is consistently applied to all countries in the rankings.

Analyzing the forecast results, we can conclude:

- Countries most likely to increase the number of medals: the United States.
- Countries most likely to decrease the number of medals: South Korea.

Table 7: Los Angeles Olympics (2028) Virtual Medal Table

NOC	Gold	Silver	Bronze	Total	Total (2024)	△Total
USA	50	47	40	137	127	+11
CHN	39	29	22	90	91	-1
GBR	20	20	23	63	65	-2
JPN	17	15	15	47	45	+2
AUS	17	15	17	50	53	-3
FRA	15	25	22	62	64	-2
GER	14	8	9	31	33	-2
NED	13	8	12	33	34	-1
ITA	11	12	16	39	40	-1
KOR	10	7	9	26	32	-6
NZL	8	7	5	20	20	0
HUN	7	6	6	19	19	0
CAN	7	7	10	24	27	-3
ESP	6	5	8	19	18	+1
BRA	5	6	8	19	20	-1

3.3.3 95% Confidence Intervals by Bootstrap

For the Random Forest model, the 95% confidence intervals are estimated using the bootstrap method. The process includes the following steps:

1. Resample the training data 300 times and train a model for each resample.
2. Use the models trained on the resampled data to make predictions.
3. Determine the confidence intervals by calculating the 2.5th and 97.5th percentiles of the predictions.

The predicted intervals are presented in Table 8:

Our prediction intervals for Gold, Silver, Bronze, and Total are predicted separately, so the Total prediction interval is not determined by the intervals for the three medal categories.

3.4 Task 1.2: The Probability of First Medal

After observing and organizing the dataset, we find that there are currently 77 countries or regions that have participated in the Olympics but have never won a medal. Among them, there are many countries that have participated multiple times.

We also identify countries that have won medals for the first time in recent years, including Albania, Cabo Verde, Dominica, and Saint Lucia in 2024. Using historical data, we explore the commonalities of countries that have won medals for the first time. The following predictions are based on Hypothesis 3, which only considers the performance of athletes and not the political and economic situation of external countries.

Table 8: 95% Confidence Intervals for Medal Predictions Using Random Forest

NOC	L_Gold	U_Gold	L_Sil	U_Sil	L_Bro	U_Bro	L_Total	U_Total
USA	43.4673	55.2163	36.1520	48.5010	34.5945	42.3355	120.2138	146.0528
CHN	30.8390	42.1753	22.9743	34.2540	16.6925	26.8335	81.5058	103.2628
GBR	13.3978	28.5428	16.1528	27.2090	12.6220	24.3835	52.1725	76.1353
JPN	11.2848	23.0640	10.7348	18.5230	12.1013	19.8430	34.1208	56.4300
AUS	10.6555	32.5253	10.5470	22.6360	10.0188	31.1533	41.2213	76.3145
FRA	11.5233	27.2760	18.3410	29.6038	19.6150	30.0225	49.4793	86.9023
GER	8.5160	21.4323	3.3070	17.7790	7.1198	19.4945	22.9428	50.7058
NED	6.6443	21.3190	4.0565	13.9610	6.5465	15.9925	17.2473	51.2725
ITA	8.1100	15.2173	10.4003	14.1658	13.5638	19.4900	32.0740	48.8730
KOR	4.2660	13.1340	4.6835	10.3085	5.8170	11.1828	14.7665	34.6253
NZL	5.4268	11.3150	4.2243	9.9683	3.3828	7.8695	13.0338	29.1528
HUN	4.1548	7.1308	4.8633	8.1225	3.9523	7.5568	14.9703	22.8100
CAN	5.9835	12.6878	5.0590	10.6268	8.8628	13.2300	19.9053	36.5445
ESP	2.7575	10.2725	2.2690	8.8830	3.0180	11.2775	10.0445	30.4330
BRA	3.8375	8.6273	4.5943	8.8620	6.1783	11.9415	14.6100	29.4308

3.4.1 BP Neural Network Classifier

Unlike medal ranking predictions, this analysis cannot directly use previous models for training and prediction. This is because for countries that have never won a medal, the dependent variable, *Total*, is always 0 during training, leading to significant errors and making reliable results unattainable.

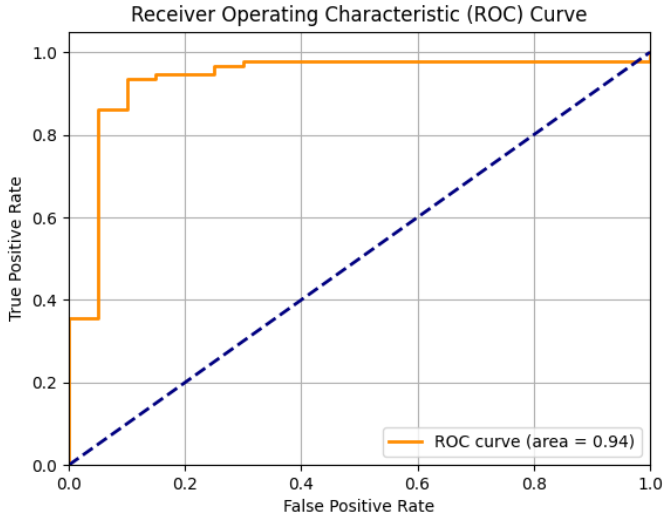

Figure 8: ROC Curve of BP Neural Network Classifier

We use countries that have won their first Olympic medals in recent years as the training set

and countries that have never won a medal as the test set. A classifier is used to map the likelihood of a country winning a medal (PWM = 0 or > 0) to a value between 0 and 1. After comparing SVM and Random Forest classifiers, we selected the BP neural network classifier, which has an ROC curve with an area of 0.94, as shown in Figure 8, indicating strong classification performance.

A BP neural network classifier is used to predict the 2028 Los Angeles Olympics for countries that did not win medals, and the resulting probabilities are shown in Table 9:

Table 9: Probability of Win first Medal in 2028 (Partial)

NOC	N_A	Prediction Probability
GUM	8	0.153210476
ESA	9	0.144216761
LBR	10	0.1420369
PLE	8	0.140068978
RWA	8	0.130942091
GAM	8	0.130942091
COD	6	0.12339206

As can be seen from the table, countries with a long history of participation and a large number of participants are more likely to win medals. The overall probability of winning is relatively low, and it cannot be ruled out that none of these countries will win a medal. Countries or regions with a relatively high probability of winning medals at the 2028 Olympics are: GUM, ESA, LBR, and PLE.

3.5 Task 1.3: The Impact of Events on Medals

In Task 1.1, we used only the number of projects and the number of participants to predict the number of medals. The correlation between the number of projects and the number of medals is 0.52. Below we examine the relationship between the number of events, the number of athletes, and the number of medals won in some countries.

By examining the table, we focus on exploring the relationship between medals and events for six countries: China, the United States, Great Britain, Germany, Japan, and France. We organize data on the number of participants in each event and the number of medalists across all Olympic Games in history for each country. Using random forests, we predict the total number of medals won, thereby determining the importance of each event in contributing to the medal count.

$$\text{PWMA} = \sum_{i=1}^{n} (N\_G_i + N\_S_i + N\_B_i) \tag{3}$$

where n is the total number of events and i is the i-th event in an Olympic Games.

Figure 9 shows the solution framework for this subtask.

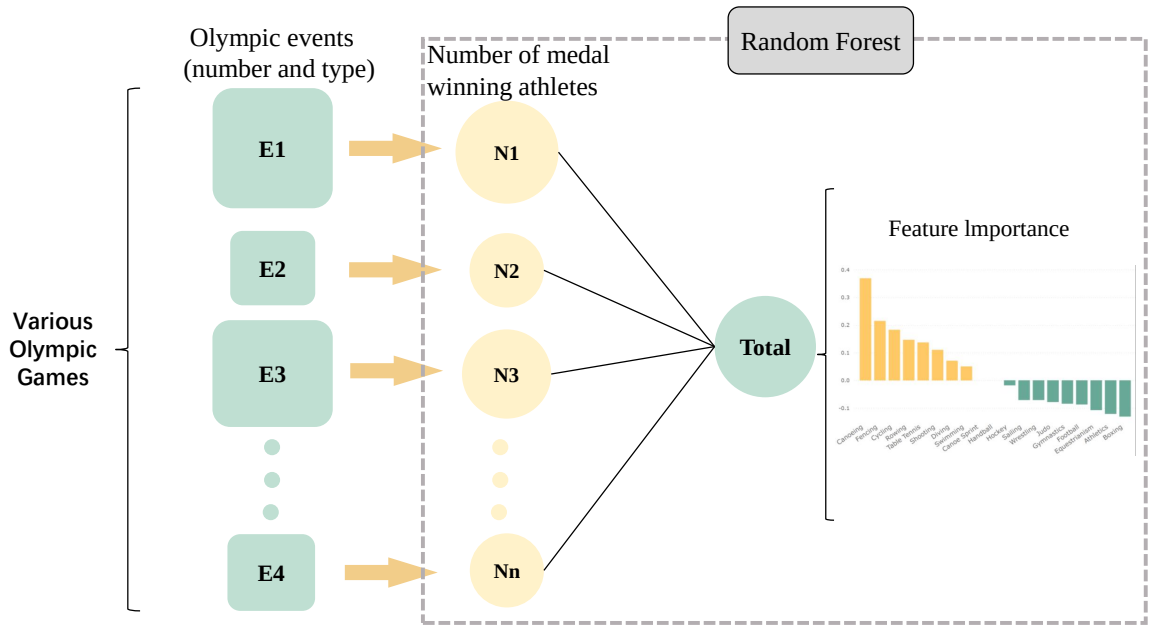

Figure 9: Flowchart for Task 1.3

3.5.1 Event Importance Based on Random Forests

We use random forests to predict the total number of medals based on the number of participants and medalists in each event. The data consists of the number of participants and medalists for each event across multiple Olympic Games for a given country. Random forests are then used to generate feature importance. For example, in predicting the medal count for Germany, the R^2 value reaches 0.91, and the feature distribution is shown in Figure 10.

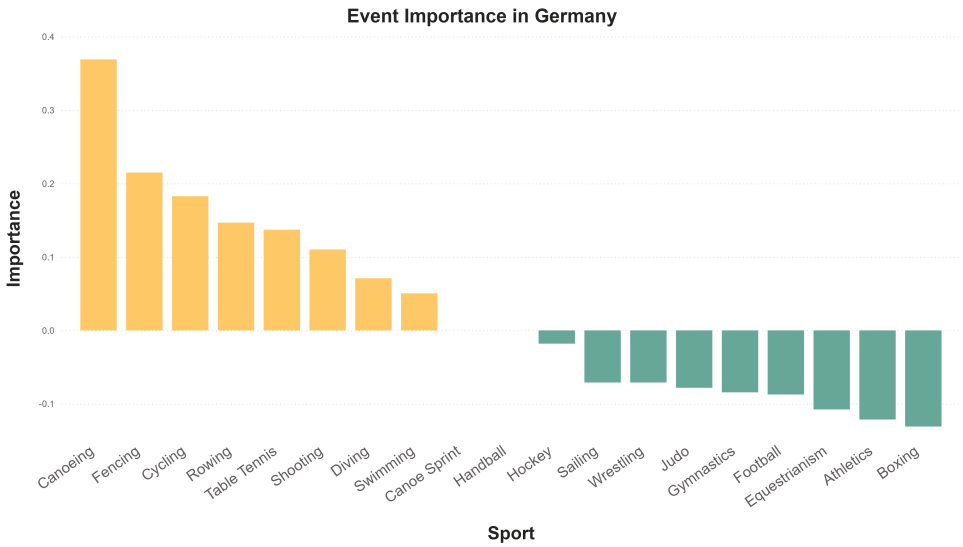

Figure 10: The event importance distribution of Germany

Using the same method, the results of the feature importance of the six countries are shown in Figure 11.

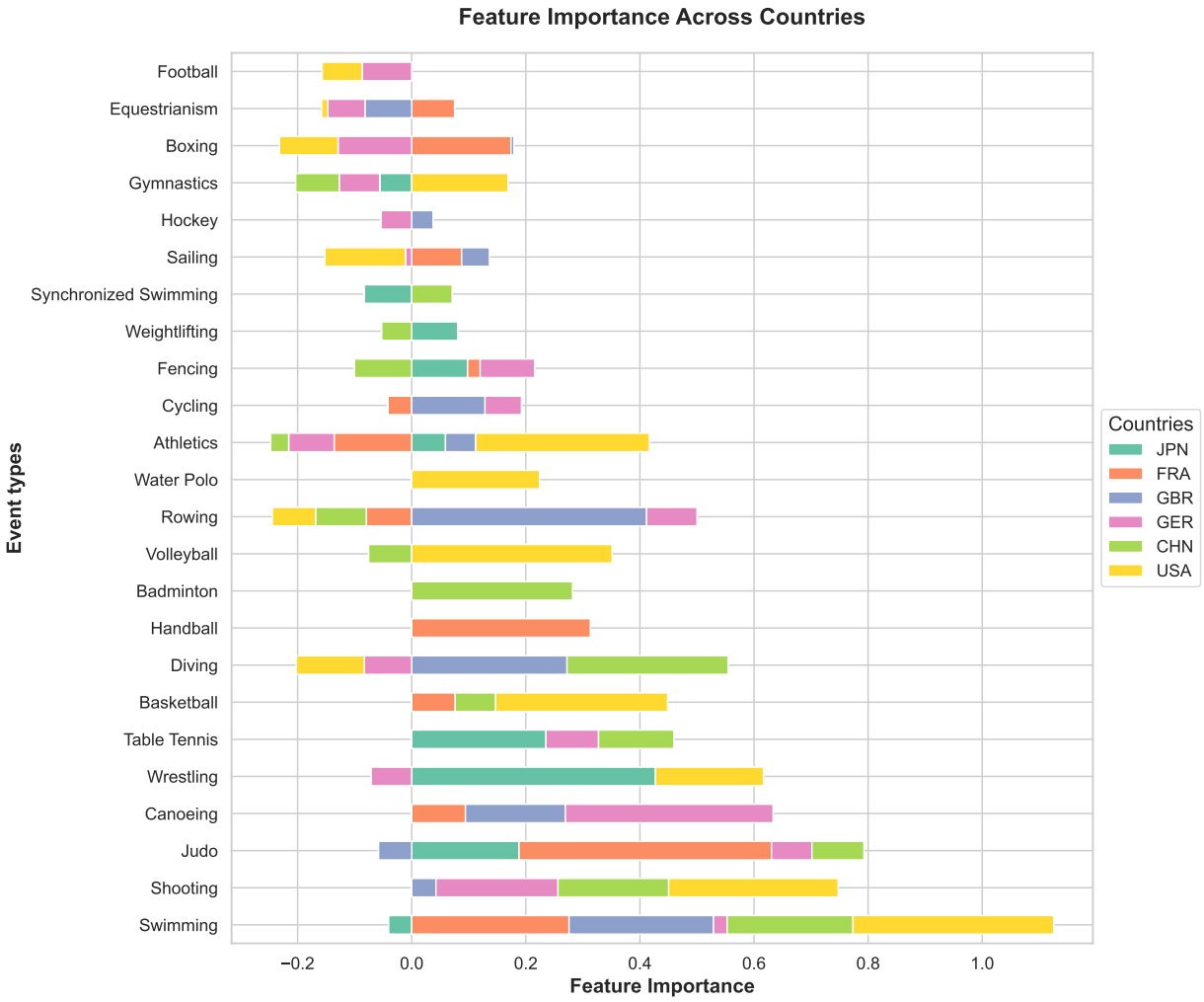

Figure 11: Importance of Sports Events by Medal Count for Different Countries

Some countries have too few participants in certain sports and no medals; therefore, these were not included in the training and are assumed to have a feature importance of 0. The feature importance distribution mainly focuses on sports with a good medal expectation. It is important to note that we treat the number of participants and medalists in a sport as the essence of that sport. Therefore, our predictions include the number of events as a type. However, our predictions may amplify the impact of team events.

Analyzing the chart, the key sports for medal count in six countries are:

- Japan: Wrestling, Judo, and Table Tennis.
- France: Judo, Handball, and Swimming.

- Great Britain: Rowing, Diving, and Swimming.
- Germany: Canoeing and Shooting.
- China: Badminton, Diving, Swimming, and Shooting.
- United States: Swimming, Basketball, Volleyball, and Athletics.

The model explains that these countries have more participants in key sports and are more likely to win medals in these areas. The identified important sports align closely with the countries' traditional strengths, which significantly influence their medal counts. This indicates that our model results are highly accurate.

3.5.2 The Impact of Home Country Events

The host country often adds new Olympic events. These events are related to the country in some way and are basically sports that the people of the country love. The newly added events of the host country can have a certain impact on the number of medals.

Examples include the 1996 Olympic Games in the United States, the 2020 Olympic Games in Japan, and the 2024 Olympic Games in France. We pay attention to whether the host country has certain advantages in the newly added events. According to the data provided, the main events added to these three Olympic Games were:

- 1996 Atlanta Olympics: Beach Volleyball, Mountain Biking, Softball, Lightweight Rowing, etc.
- 2020 Tokyo Olympics: Baseball, Softball, Rock Climbing, Karate, Surfing, and Skateboarding.
- 2024 Paris Olympics: Breakdancing, Skateboarding, Rock Climbing, and Surfing.

Here are the medal results for the three specific events:

Table 10: Medal Counts for New Events at the 1996 Atlanta Olympics

NOC	N_G	N_S	N_B	Total	P_G	P_T
USA	41	5	18	64	33.33%	17.63%
RUS	42	2	13	57	34.15%	15.70%
CHN	4	45	5	54	3.25%	14.88%
AUS	0	0	26	26	0	7.16%
BLR	0	12	6	18	0	4.96%
BUL	0	6	11	17	0	4.68%

Table 11: Medal Counts for New Events at the 2020 Tokyo Olympics

NOC	N_G	N_S	N_B	Total	P_G	P_T
JPN	43	4	4	51	78.18%	29.48%
USA	1	40	3	44	1.82%	25.43%
DOM	0	0	24	24	0	13.87%
CAN	0	0	15	15	0	8.67%
BRA	1	3	0	4	1.82%	2.31%
TUR	0	1	3	4	0	2.31%

Table 12: Medal Counts for New Events at the 2024 Paris Olympics

NOC	N_G	N_S	N_B	Total	P_G	P_T
USA	1	3	3	7	8.33%	19.44%
JPN	3	3	0	6	25%	16.67%
BRA	0	1	3	4	0	11.11%
AUS	2	1	0	3	16.66%	8.33%
FRA	1	1	1	3	8.33%	8.33%
CHN	0	2	1	3	0	8.33%

where,

- P_G is the proportion of the number of people who won gold medals.
- N_G is the number of people who won gold medals. When medals are won as a team, we treat it as if each person won a medal individually.

As can be seen in the analysis table, the United States and Japan were the top medal winners in the new events in 1996 and 2020, respectively. France also did well in the new events in 2024. After reviewing the relevant information[6][7], we believe that selecting events from the home country that are either advantageous or popular in the country can greatly improve the results of the competition. For example, the new events of karate and skateboarding in Japan, and the new event of softball in the United States. These sports were originally popular in that country.

4 Task 2: "Great Coach" Contribution to Medals

4.1 Existence of "Great Coach" Effect

We first looked up Lang Ping's coaching trajectory and considered her role as the head coach of the Chinese women's volleyball team at the 2016 Rio Olympics[8]. We examined the results achieved by the women's volleyball team at the Games during her tenure as coach. The results are shown in Table 13:

The Chinese women's volleyball team secured the gold medal in women's volleyball at the 2016 Olympics. However, in the eight years before and after this victory, their best Olympic result

Table 13: Performance of the Chinese Women's Volleyball Team in the Last Five Olympic Games

Team	Year	Event Participants	Great Coach	Gold	Silver	Bronze	Total
China	2024	13	0	0	0	0	2.41
China	2020	12	0	0	0	0	3.85
China	2016	12	1	1	0	0	1.41
China	2012	12	0	0	0	0	3.10
China	2008	12	0	0	0	1	2.41

was a single bronze medal. This leads us to conclude that the "Great Coach" effect does indeed play a role.

4.2 Contribution of "Great Coach"

We expanded our analysis by researching legendary Olympic coaches and introduced a binary feature, "Great Coach" (0 or 1), into the dataset. Additional features included "Year" and "AAPI." The dataset covered events such as U.S. women's volleyball, women's gymnastics, and men's sprinting; Chinese women's volleyball and men's swimming; German men's canoeing; Russian women's gymnastics; British men's hurdles; and Jamaican women's sprinting.

Using a random forest model to predict the "Total" medal count, the feature importance of "Great Coach" was 0.288. A BP neural network classifier was also applied to estimate the probability of winning a gold medal.

For Jamaican women's sprinting in 2024, the inclusion of the "Great Coach" feature increased the predicted medal count from 0.56 to 1.56.

- At an AAPI of approximately 65, the predicted increase was 0.9 bronze medals, 0.2 silver medals, 0.02 gold medals, and 1.2 total medals.
- At an AAPI of around 70, the predicted increase was 0.7 bronze medals, 0.9 silver medals, 0.2 gold medals, and 1.8 total medals.

These results highlight the significant impact of the "Great Coach" effect, contributing 28.8% to overall performance.

4.3 Event Investment and Impact

4.3.1 Event Potential Index (EPI)

We propose that sports with a history of multiple medals, athletes with higher AAPI, and events showing an upward trend are more likely to benefit from the introduction of a great coach. To estimate the improvement potential of an event after introducing a great coach, we define the Event Potential Index (EPI) using the following formula:

$$\text{EPI} = \omega_1 \times \text{Total\_trend} + \omega_2 \times \text{Total\_space} + \omega_3 \times \text{AAPI} \tag{4}$$

- Total_trend represents the trend in medal count over time.

- Total_space reflects the growth potential in medal count.

Calculation Details:

- **Total_trend:**

$$\text{Total\_trend} = \frac{\text{pre\_Total\_2028} - \text{Total\_2024}}{\text{Total\_2024} + 1} \tag{5}$$

where pre_Total_2028 is the predicted medal count for the 2028 Los Angeles Olympics, and Total_2024 is the medal count from the 2024 Paris Olympics.

- **Total_space:**

$$\text{Total\_space} = \left(\frac{\text{Total\_max}}{\text{Total}}\right)^{\frac{1}{j-1}} - 1 \tag{6}$$

where Total_max is the maximum medal count achieved in the event's history, and j is the corresponding Olympic session.

Additionally, the improvement potential is influenced by the presence of dominant nations in the event. Based on the historical Olympic data and factors influencing EPI, we recommend focusing on the following events:

- China: Men's and women's gymnastics, men's swimming.
- Jamaica: Men's and women's athletics.
- Germany: Men's shooting, men's and women's swimming.

4.3.2 Post-investment Results

Our prediction for the above-selected sport is as is shown in Table 14.

Table 14: The post-investment prediction results

NOC	Events	AAPI	G_C	N_B	N_S	N_G	Total
CHN	Men's gymnastics	67.8	0	0.50	1.33	0.50	2.41
CHN	Men's gymnastics	67.8	1	1.05	1.31	1.33	3.85
CHN	Women's gymnastics	62.7	0	0.57	0.45	0.72	1.41
CHN	Women's gymnastics	62.7	1	1.03	1.36	0.87	3.10
CHN	Men's swimming	67.8	0	0.50	1.33	0.50	2.41
CHN	Men's swimming	67.8	1	1.05	1.31	1.33	3.85
JAM	Men's athletics	68.38	0	0.71	1.61	0.907	3.32
JAM	Men's athletics	68.38	1	1.08	1.82	1.21	4.11
JAM	Women's athletics	65	0	0.66	1.26	0.72	2.68
JAM	Women's athletics	65	1	1.29	1.10	0.94	2.95
GER	Men's shooting	62.6	0	0.53	0.2	0.72	1.32
GER	Men's shooting	62.6	1	1.01	0.97	0.87	2.98
GER	Women's swimming	62.4	0	0.52	0.21	0.72	1.32
GER	Women's swimming	62.4	1	0.93	0.98	0.87	2.96
GER	Men's swimming	61.6	0	0.44	0.66	0	1.1
GER	Men's swimming	61.6	1	0.84	1.1	0.09	1.75

Based on our analysis, we recommend that the following three countries introduce world-class coaches in these specific events:

- China: Women's gymnastics
- Jamaica: Men's athletics
- Germany: Men's shooting

Our predictions indicate that these events are likely to see the greatest improvement in total medal counts after the introduction of great coaches, with an estimated increase of approximately 1.6 medals.

5 Task 3: Other Original Insights

5.1 Common Traits of No Medal Nations

We observed 77 nations that have never won an Olympic medal, including countries with populations of tens of millions, such as Myanmar. When predicting medal acquisition, we found that nations are more likely to win their first medal when they demonstrate:

1. Long-term participation in the Games.
2. A large number of participating athletes.
3. Participation in a diverse range of sports.

For these nations determined to win medals, we recommend using our model to project the increased probability of medal acquisition following changes in athlete count (N_A) and event participation (N_E). Additionally, these nations could consider focusing their investments on less-dominated sports.

5.2 Negative Impact Events

Through Figure 11 (Importance of Sports to National Medal Count), we identified sports that have a negative impact on overall medal count. Given a fixed number of participating athletes, increasing participation in sports with positive importance can lead to more medals. Conversely, increasing participation in sports with negative importance is likely to result in a disproportionate investment-to-return ratio.

We recommend using our derived sport importance predictions to increase athlete participation in sports with positive importance for each nation. For example, if Germany aims to increase its medal count, it could shift investment from Boxing towards Canoeing.

5.3 Recommendations of "Great Coach"

Our model predicts that a "Great Coach" can contribute nearly 30% to performance improvement. Furthermore, we found that this effect is more pronounced in technically demanding sports such as Gymnastics and Shooting.

National Olympic Committees can use the Event Potential Index (EPI) to assess the potential performance improvement of a sport after hiring a world-class coach. By weighing the actual investment against the expected return, committees can make informed decisions about whether to hire such coaches for specific sports.

6 Model Analysis

6.1 Sensitivity Analysis

We performed 5-fold cross-validation, and the results are shown in Table 6. To further analyze the sensitivity of our model, we varied the size of the training set by randomly sampling 60%, 70%, and 80% of the original dataset for training and prediction. The resulting fitting curves and changes in information gain are shown in Figure 12 and Figure 13.

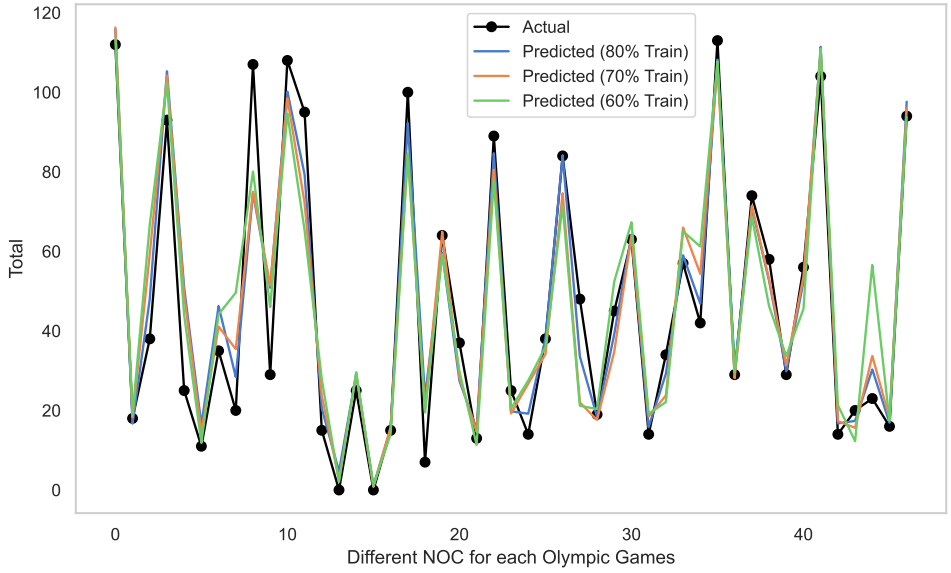

Figure 12: Fitting curves of different dataset

Changes in the training set size resulted in a slight decrease in predictive performance; however, the magnitude of this decrease was minimal. With training sets comprising 60%, 70%, 75%, and 80% of the original data, the model's information gain remained relatively stable. Between 70% and 80%, minor fluctuations in the importance of the AAPI and N_E features were observed. Overall, these two features exhibited a tendency towards increased importance compared to the 60% training set. In conclusion, our model demonstrates good robustness.

6.2 Strengths and Weaknesses

Strengths:

1. We introduced a novel metric, the Athlete Potential Index (API), to assess athlete performance. This index played a significant role in model training and improved predictive accuracy.

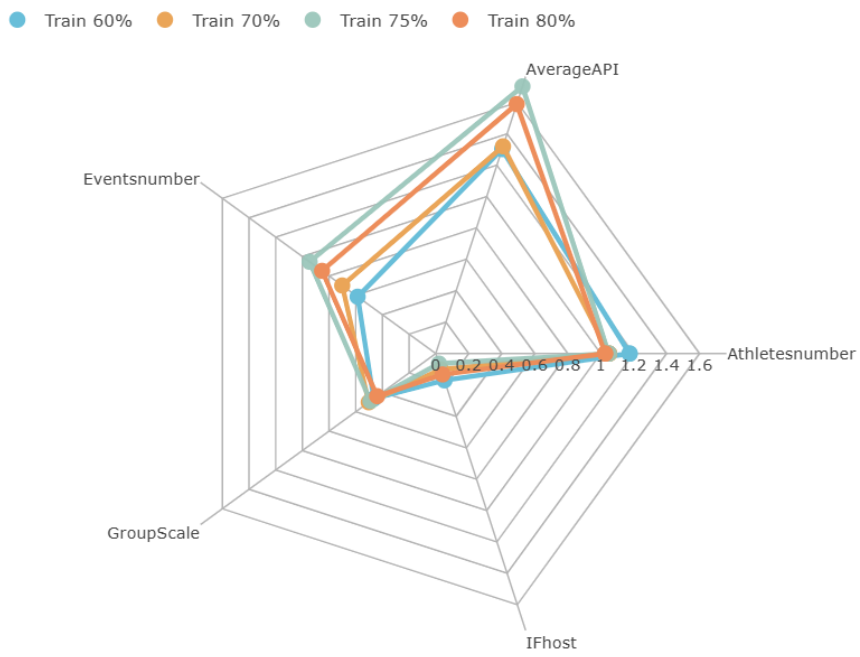

Figure 13: Changes in information gain

2. Through a comparative analysis of Linear Regression, XGBoost, and Random Forest models, we selected the optimal model for medal prediction. In the prediction set, our model achieved an R-squared (R^2) value exceeding 0.87, indicating a strong goodness of fit.

3. Rigorous validation through 5-fold cross-validation and sensitivity analysis demonstrated the robustness of our model to variations in training data.

4. For the binary classification task of predicting medal acquisition (yes/no), we employed a BP neural network classifier, achieving an Area Under the ROC Curve (AUC) of 0.94, demonstrating excellent discriminatory power.

5. We provided an Event Potential Index (EPI) to guide national Olympic committees in evaluating the potential return on investment of hiring world-class coaches for specific sports.

Weaknesses:

1. The performance of the binary classifier could potentially be further enhanced with a larger training dataset, which may lead to more reliable and generalizable results.

2. Predicting the "Great Coach" effect presents challenges in definitively identifying and quantifying the performance difference between existing coaches and world-class coaches. This inherent difficulty introduces uncertainty into the estimation of the coach's contribution.

3. While we introduced the Event Potential Index (EPI), further research is needed to develop a more precise and quantitative definition of this metric, enabling more accurate and nuanced analysis.

7 Conclusions

In this paper, we employed a BP neural network classifier and a random forest regression model to predict the medal count for the 2028 Los Angeles Olympic Games, providing prediction intervals. We also assessed the key sports contributing to medal acquisition for six specific nations, identifying the sports responsible for the majority of their medals. Our analysis suggests that including locally popular or dominant sports as new Olympic events can positively impact the host nation's medal count. We provided evidence supporting the "Great Coach" effect, estimating its contribution to be up to 28.8%, and identified potential coaching interventions for three nations (China, Jamaica, and Germany) projected to yield an increase of approximately 1.6 medals. Finally, we presented original insights regarding non-medal-winning nations, the importance of specific sports, and the impact of world-class coaches.

References

[1] https://en.wikipedia.org/wiki/Olympic_Games.

[2] https://olympics.com/en/paris-2024/medals.

[3] C. Schlembach, S. L. Schmidt, D. Schreyer, and L. Wunderlich, "Forecasting the olympic medal distribution – a socioeconomic machine learning model," *Technological Forecasting and Social Change*, vol. 175, p. 121314, 2022.

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

[5] M. A. Ruiz Estrada, "A new indicator to evaluate any country performance in the olympic games: The olympia-index," *Available at SSRN 4920383*, 2024.

[6] https://www.thepaper.cn/newsDetail_forward_16567025.

[7] J. Luo, "Research on the change of the project setting of the modern summer olympic games (1984-2020)," Ph.D. dissertation, Changsha: Hunan Normal University 2020, 2020.

[8] https://www.olympics.com/zh/athletes/ping-lang.