Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2527215

Predicting Olympic Medals: A Data-Driven Analysis and Strategic Insights

Summary

To predict the final medal counts of the Olympic Games, this paper proposes a data-driven model based on historical Olympic medal data.

First, we extract 14 features from the dataset: the number of participants from each country in 11 major categories for each Olympic Games, whether the country is the host, and the average number of gold medals and total medals won by each country in the past three Olympic Games. We then attempt to predict the medal table using four models—Linear Regression, Random Forest, LightGBM, and XGBoost. After comparing the errors (MSE, RMSE, R^2 Score), we select the Random Forest model for predicting the medal counts. The top three countries in our predicted medal table for the Los Angeles Olympics are United States with 48 GOlds and 129 in total, China with 39 Golds and 87 in total, United Kingdom with 24 Golds and 70 in total.

Next, to predict how many countries will earn their first medal in the next Olympic Games, we continue to use the same four models to obtain the probability of winning the first medal, we compare cross-entropy loss and ultimately select LGBM. We predict that there will be five countries earning their first medal. The top three countries or regions with the highest probability are Palestine, Guinea and EI Salvador.

We then identify the top two events in which the highest proportion of countries in the top ten medal rankings in the last four Olympic Games participated, as well as the three dominant events for China and the United States over the years. This analysis leads to the conclusion that there is a positive correlation between the dominant events of a country and its total medal count.

Furthermore, we explore the "great coach" effect. We label part of the dataset to train the XGBoost model and predict whether each country in every Olympic Games has a "great coach" for each event. After completing the predictions, we use the chi-squared test to examine the correlation between having a "great coach" and winning medals. Additionally, we analyze data from China, the UK, and Japan, applying Fisher's exact test to calculate p-values and identify which events would yield the greatest benefit from investing in a "great coach". Based on this analysis, we offer recommendations for these countries on where to invest in "great coaches".

Finally, during the modeling process, we discover new perspectives, such as the geographic clustering effect of medal distribution and the relationship between GDP and Olympic medal counts. Based on these findings, we provide recommendations for national Olympic committees.

Based on our model analysis and conclusions, we offer relatively objective predictions for future Olympic medal tables, investigate the influence of the "great coach" effect, and present practical recommendations for national Olympic committees.

Keywords: Olympic medal prediction, Random Forest, LGBM, "great coach" effect, XGBoost

Contents

1 Introduction

1.1 Background

In modern society, sports are not just physical activities but also powerful forces that carry significant spiritual and cultural values. The Olympic spirit, epitomized by the motto "Citius, Altius, Fortius" (Faster, Higher, Stronger), inspires individuals to push beyond their limits and strive for excellence. This spirit has become deeply ingrained in global sporting events and in the lives of people worldwide. As the largest and most influential sporting event on the planet, the Olympic Games capture the attention of audiences and media across the globe. Among all the competitions, the medal table stands out as one of the most prominent focal points, not only reflecting the athletic prowess of athletes from various countries but also symbolizing a nation's strength and competitiveness on the world stage.

After each Olympic Games, the shifts in the medal table become a hot topic of discussion for the global media and the public. In the lead-up to the Olympics, major media outlets attempt to predict the medal table based on historical data, athlete performance, and other influencing factors. This process goes beyond predicting the outcomes of sports events; it also explores the evolving position of each country in the global sports arena. With the rapid advancement of data science, predicting the Olympic medal table using scientific and systematic methods has become a crucial research area.

This paper aims to develop an accurate model for predicting the Olympic medal table by applying data mining and machine learning techniques to large-scale historical data from multiple sources. To achieve this, we will employ methods such as multiple linear regression and Random Forest to generate reliable predictions.

1.2 Restatement of the Problem

The challenge at hand involves the analysis of data from four datasets and models and data analysis must only use the provided datasets. Specifically, the primary tasks to be addressed include:

- Develop a model for medal counts for each country and predict the medal table in the Los Angeles, USA summer Olympics in 2028.
- Predict how many countries will earn their first medal in the next Olympics and give the likelihood of the estimate.
- Explore the relationship between the events and how many medals countries earn and identify the most important sports for each country.
- Provide a model to assess the impact of coaches on the performance of various national teams and explore how this factor contributes to the medal count of each country.
- Based on the models from the previous questions, we are required to reveal some insights for National Olympic Committees.

1.3 Overview of Our Work

First, we find a few key points in this question:

- The given dataset contains a lot of inconsistencies in data such as country names and sport names, which need to be processed.
- We need to model the sports events, but since there are many events, the feature space is high-dimensional, and we need to perform dimensionality reduction. How should dimensionality reduction be performed?
- The dataset lacks information on coaches. How can we analyze the impact of coaches?

On the basis of the above discussion, to determine the optimal medal prediction strategy, we may boil down the tasks to the following four steps:

- First, we clean the data. We standardize the sport names in *summerOly_programs* based on the sport names in *summerOly_athletes*. Additionally, we replace the country names in both *summerOly_athletes* and *summerOly_hosts* with the corresponding country abbreviations.
- Next, due to the lack of feature descriptions for the events, we estimate and categorize the sports into 11 major categories for dimensionality reduction.
- We train the model by labeling some well-known "great coaches" as the training set based on historical Olympic data. The model is then used to score and predict the coaches for each event in subsequent Olympics for other countries.
- Further analysis and discussion of the model.

2 Assumptions and Notations

2.1 Assumptions

- The number of events in the 2028 Olympic Games will remain almost consistent with the 2024 Games, excluding newly added events such as cricket, baseball/softball, rugby.
- The participation of countries and regions in the events of the 2028 Olympic Games will be almost identical to that of the 2024 Games.
- The historical performance of countries and athletes remains stable, meaning there will be no abnormal fluctuations in their performance during the competition.
- The influence of political and economic conditions on Olympic results is ignored.
- When considering the host country effect, we only take into account the impact on the host country itself.
- The influence of natural or political factors that may cause changes in the Olympic cycle is ignored.
- It is assumed that there will be no sudden changes in the sports level and sports investment of each country.
- The technological and equipment levels of each country change steadily.
- The impact of athletes' psychology and international competition experience is ignored.

2.2 Notations

Name	Definition	Denotation
XGBoost Prediction	Predicted value of sample	$\hat{y}_i$
Decision Tree Function	Function representing decision tree structure	$f_k(x_i)$
Decision Tree Structure Set	Set of decision tree structures	$F = \{f(x) = w_{q(x)}\}$
Feature Mapping Function	Maps sample to leaf node	$q : \mathbb{R}^m \to \{1, 2, \ldots, T\}$
Leaf Node Scores	Scores associated with each leaf node	$w \in \mathbb{R}^T$
Loss Function	Sum of squared errors	$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Regularization Term	Regularization term to prevent overfitting	$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$
Chi-Square Statistic	Statistical measure for independence	$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
Observed Frequency	Frequency of occurrences in contingency table	O_i
Expected Frequency	Expected frequency under null hypothesis	$E_i = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$
Contingency Table Value	Number of successes and failures in groups	a, b, c, d
Null Hypothesis	Hypothesis of no association between categorical variables	H_0
Binomial Coefficient	Number of ways to choose k successes from n trials	$\binom{n}{k}$
p-Value	Probability of obtaining the observed or more extreme table	P

3 Data Preprocess

Any rigorous statistical analysis begins with substantial data preprocessing. The raw data primarily consists of four datasets: *summerOly_medal_counts*, *summerOly_programs*, *summerOly_hosts*, and *summerOly_athletes*. During the analysis of these datasets, we identified several issues, including data inconsistencies, missing values, and redundancy. These issues needed to be addressed before proceeding with model fitting and analysis.

3.1 Data Cleaning

The initial data cleaning was performed in Python. We replaced certain variable values to simplify their representation, making them more conducive to model calculation and fitting.

Upon comparing the datasets *summerOly_athletes*, *summerOly_medal_counts*, and *summerOly_hosts*, we observed that country names in *summerOly_athletes* were represented using abbreviations (e.g., “Australia” as “AUS”), while *summerOly_medal_counts* used full country names (e.g., “Australia”), and *summerOly_hosts* used the full name of the host city and country (e.g., ‘’Athens, Greece”). To facilitate data integration and analysis, we replaced the full country names in the *NOC* column of *summerOly_medal_counts* and the *Host* column of *summerOly_hosts* with the corresponding country abbreviations.

At the same time, we made some exclusions for certain countries or regions. For example, due to the impact of the Russia-Ukraine war, Russia has been banned from participating in the Paris Olympics, and Russia did not exist before the dissolution of the Soviet Union. Considering the complexity of both the Soviet Union and Russia, we decided to exclude them entirely. We also treated both East Germany and West Germany as a unified Germany, they share the same country code "GER" for Germany.

Additionally, the “Medal” column in *summerOly_athletes* contained four values: “No medal”, “Gold”, “Silver”, and “Bronze”. We replaced these with numerical values 0, 1, 2, and 3, respectively, to simplify the representation without losing the meaning.

Additionally, the "Sport" column in the *summerOly_athletes* dataset has data consistency issues. Some discrepancies arise due to the mixing of major and minor sports, such as "3x3 Basketball" and "3x3 Basketball, Basketball," which both actually belong to the "Basketball" category. Other discrepancies occur because the same sport can be represented with different

terms; for example, "Equestrianism" and "Equestrian." In the 2016 Rio Olympics, this sport was referred to as "Equestrianism," while in the 2024 Paris Olympics, it is referred to as "Equestrian." To standardize for prediction, we have organized the "Sport" data by following the same structure as provided in the *summerOly_programs* dataset. For sports not listed in *summerOly_programs* (such as Art Competition, Tug_Of_War, etc.), we have grouped them under "Others."

3.2 Data Imputation

Next, while analyzing *summerOly_programs*, we found that the number of gold medals awarded for certain events in some Olympic Games was missing or represented by blank spaces, indicating that the event was either not held or did not produce any gold medals for some reason. For instance, Sailing and rowing were included in the 1896 Games program but were canceled due to bad weather. To facilitate calculations and modeling, we replaced these missing values and blank spaces with 0.

4 Task1: Model for Medal Counts Prediction

During each Olympic Games, the medal table is always a hot topic of discussion, with people frequently talking about which athlete has won another gold medal in a specific event. Many organizations and institutions also provide predictions for the medal table before the Games begin, aiming to attract attention and spark relevant discussions. Based on the provided dataset, we have also developed a model to predict the medal table for the 2028 Los Angeles Olympics. Additionally, we forecast how many countries will win their first Olympic medal and reveal the relationship between the number of medals and sports events for different countries in various Olympics.

4.1 Feature Extraction

The number of medals won at the Olympics is influenced by various factors. At the national level, it is affected by population size and economic development. It is also influenced by the country's sports development characteristics, such as historical Olympic performance, dominant events, the number of events participated in, and the number of athletes competing. Additionally, the number of medals is impacted by objective factors such as the host country. Given the numerous influencing factors and the limitations of the dataset for this topic, we have only selected historical Olympic performance (the average number of gold medals and total medals in the last three Olympics), whether the country is the host, the number of athletes, and the number of events as features.

Although we unified certain values in the *Sport* column of the *summerOly_athletes* dataset during the data preprocessing phase, the number of unique values in the *Sport* column remains excessively large. In the subsequent model fitting process, each unique value in the *Sport* column will be treated as a separate variable. An excessively large number of variables can significantly increase the dimensionality of the feature space, leading to several challenges. As the dimensionality increases, the data becomes increasingly sparse, and the possible combinations of subspaces grow exponentially. Moreover, a higher number of features results in more parameters to be learned, increasing the model's complexity. To prevent underfitting and simplify the model, it is therefore essential to reduce the dimensionality of the feature space.

Since the individual sports lack descriptive features, traditional dimensionality reduction techniques such as PCA or k-means clustering are not feasible. Consequently, we opted to

classify these sports based on empirical estimation, ensuring that the number of gold medals produced by each category is approximately balanced.

Sports involving balls, such as basketball, football, and volleyball, typically have a larger number of participating athletes but produce relatively fewer gold medals. Hence, we grouped these sports into the *Balls* category. Similarly, certain water sports, such as sailing, were grouped into the *WaterSport* category. For newer Olympic events or less popular sports that produce relatively few medals, such as breaking, we classified them into the *Others* category. The remaining sports were classified using a similar approach, which will not be elaborated upon here.

As a result, we categorized all sports into 11 groups: *Balls*, *Athletics*, *Swimming*, *Cycling*, *Combat*, *Weightlifting*, *Gymnastics*, *Shooting* (including archery), *WaterSport*, *Diving*, and *Others*. This approach significantly reduces the dimensionality of the feature space, facilitating subsequent modeling and training.

With the above special handling, we applied some special treatments for the participating athletes and sports. Not only did we consider the total number of participants and sports, but we also took into account the number of athletes from different countries and different Olympic Games in the 11 groups above.

Finally, the dataset we obtained covers 11 sports events, as well as fourteen features, including the average number of gold medals, average number of medals over the past three Olympics, and whether the country is the host.

4.2 Projections for the 2028 Los Angeles Olympics

4.2.1 Model Building

We attempted to fit the data using various models, including multiple linear regression, random forest regression, LightGBM and XGBoost. Among these, the random forest model performed the best. The evaluation metrics, such as mean squared error, root mean squared error and R^2 score for the four models are shown in Table1.

Table 1: Model Evaluation Metrics

Model	MSE	RMSE	R^2 Score
Linear Regression	5.16 / 25.87	2.27 / 5.08	0.78 / 0.83
Random Forest	3.76 / 17.24	1.93 / 4.15	0.84 / 0.89
LightGBM	5.83 / 23.83	2.41 / 4.88	0.75 / 0.85
XGBoost	4.32 / 21.92	2.07 / 4.68	0.81 / 0.86

The table1 shows the three types of errors for four models. The errors before "/" represent the error in predicting the number of gold medals, while the errors after "/" represent the error in predicting the total number of medals. From this, it is evident that the errors for the Random Forest model are smaller than those of the other three models, which is why we chose Random Forest as the model for medal prediction.

4.2.2 Projection Results

The prediction results are shown in the table2. The "Gold" and "Total" columns in the table represent the predicted number of gold medals and total medals, respectively. The

"Gold_95%_CI" and "Total_95%_CI" columns indicate the 95% confidence intervals for the predictions of gold and total medals. Combining the predicted data with the medal table from the 2024 Olympics, we find that the United States is expected to see a significant increase in its medal count. This is likely due to the 2028 Olympics being held in the U.S., which gives them a natural home advantage. Additionally, the U.S. is favored in new events like rugby, cricket, lacrosse, squash, and baseball/softball, among which baseball and rugby are key sports in America's four major professional leagues. At the same time, the United Kingdom, Italy, and Germany are also projected to improve their medal counts. Germany's medal count is particularly interesting; since the beginning of the 21st century, Germany's ranking in the Olympic medal table has been declining, dropping to 10th place in 2024. This is inconsistent with its status as the third-largest global economy by GDP, making it highly probable that Germany will improve its sports performance in this cycle to maintain a ranking in line with its economic status.

However, some countries may experience a decline in their medal counts. The most notable example is France, which has lost the advantage of competing at home. Moreover, breakdancing, which was added to the Olympic program in 2024, has faded and will be removed from the 2028 Olympics, reducing France's competitive edge. China, South Korea, and Japan are also expected to see a decrease in their medal counts, possibly due to regional differences. Asian countries will need more adjustments to adapt to different schedules and time zones, and the new events are not as popular in Asia, limiting their competitiveness. Additionally, after China broke the record for overseas gold medals in 2024, it will be challenging to continue breaking records in the subsequent years.

Table 2: Projections for the Medal Table in Los Angeles

Rank	NOC	Gold	Gold_95%_CI	Total	Total_95%_CI
1	United States	48	[38, 64]	129	[104, 153]
2	China	39	[26, 43]	87	[68, 100]
3	United Kingdom	24	[15, 32]	70	[54, 95]
4	Italy	22	[14, 28]	57	[44, 78]
5	France	15	[12, 23]	54	[40, 67]
6	Australia	17	[12, 25]	52	[38, 75]
7	Japan	18	[11, 29]	51	[37, 85]
8	Germany	17	[11, 21]	44	[35, 66]
9	Canada	15	[9, 18]	41	[29, 58]
10	Netherlands	12	[8, 13]	31	[24, 35]
11	South Korea	8	[6, 12]	25	[19, 33]
12	Spain	5	[4, 9]	25	[19, 30]
13	Hungary	6	[4, 8]	21	[16, 27]
14	Brazil	5	[4, 9]	21	[19, 28]
15	New Zealand	7	[5, 9]	18	[16, 23]
16	Ukraine	7	[3, 10]	17	[11, 25]
17	Uzbekistan	6	[2, 8]	15	[10, 20]
18	Belgium	4	[2, 6]	15	[11, 20]
19	Serbia	4	[2, 7]	14	[6, 19]
20	Poland	3	[2, 6]	14	[12, 18]
21	Denmark	4	[2, 6]	13	[10, 16]
22	Switzerland	3	[2, 4]	13	[9, 16]

Rank	NOC	Gold	Gold_95%_CI	Total	Total_95%_CI
23	Kazakhstan	4	[2, 8]	12	[8, 20]
24	Turkey	3	[2, 6]	12	[9, 18]
25	Egypt	5	[2, 7]	11	[6, 17]
26	Kenya	3	[2, 4]	9	[7, 14]
27	Nigeria	3	[0, 4]	9	[3, 13]
28	Romania	3	[2, 4]	9	[6, 13]
29	Jamaica	2	[1, 4]	9	[6, 15]
30	Sweden	2	[2, 4]	9	[7, 11]
31	Croatia	3	[1, 5]	8	[6, 13]
32	Greece	3	[1, 5]	8	[5, 13]
33	Ireland	2	[1, 3]	8	[5, 10]
34	Norway	2	[2, 3]	8	[6, 10]
35	Bulgaria	2	[1, 3]	7	[5, 9]
36	Czech Republic	2	[2, 3]	7	[6, 10]
37	South Africa	2	[1, 3]	7	[5, 9]
38	Azerbaijan	1	[0, 3]	7	[6, 11]
39	Cuba	2	[1, 2]	6	[5, 8]
40	Georgia	2	[1, 3]	6	[4, 9]
41	Chinese Taipei (Taiwan)	2	[1, 3]	6	[5, 8]
42	Colombia	1	[0, 2]	6	[5, 8]
43	Mexico	1	[0, 2]	6	[4, 9]
44	Philippines	1	[1, 2]	6	[4, 8]
45	India	2	[0, 5]	5	[4, 9]
46	Slovenia	2	[1, 3]	5	[4, 8]
47	Thailand	2	[1, 2]	5	[4, 6]
48	Austria	1	[0, 1]	5	[4, 6]
49	Iran	1	[1, 2]	5	[4, 7]
50	Israel	1	[0, 1]	5	[3, 7]

4.3 First Medal Projections for Unmedaled Countries

4.3.1 Model Building

We used similar features as in the previous question, retaining the participation of athletes in the eleven sports, but removed features like "host country," "average gold medals from the past three Olympics," and "average total medals," as these three features are not relevant for predicting whether a country will earn its first medal. Additionally, we included the total number of athletes as a feature, since a larger number of athletes generally increases the likelihood of winning medals. We also added a column, Get_Medal, as the target variable, where this column is 0 when the country has not earned a medal.

After selecting the features, we processed the dataset by retaining only the records of countries that earned their first medal in the Olympic Games, removing all subsequent Olympic participation records for those countries. For example, Japan earned its first Olympic medal in the 1912 Stockholm Olympics, so we retained only Japan's participation in the 1912 and earlier Olympics, and the Get_Medal value for 1912 would be 1. There would be no participation records for Japan in the subsequent Olympic Games after 1912.

A significant advantage of this approach is that it helps with regularization. As countries that have previously won medals participate in more Olympic Games over the years, their participation scale increases, causing feature values to grow, which could lead to model parameters becoming too large and introducing bias. By removing these records, the model focuses more on learning the characteristics of countries that are earning their first medal, leading to more accurate predictions.

This is a typical classification problem, but we treat it as a regression problem. The model will output a value between 0 and 1, with values closer to 1 indicating a higher likelihood of winning a medal. Similar to the previous question, we compared the performance of four models: Linear Regression, Random Forest, LightGBM, and XGBoost. Their cross-entropy losses are in table 3, and it is evident that LightGBM performs the best in terms of fitting, so we chose it for modeling.

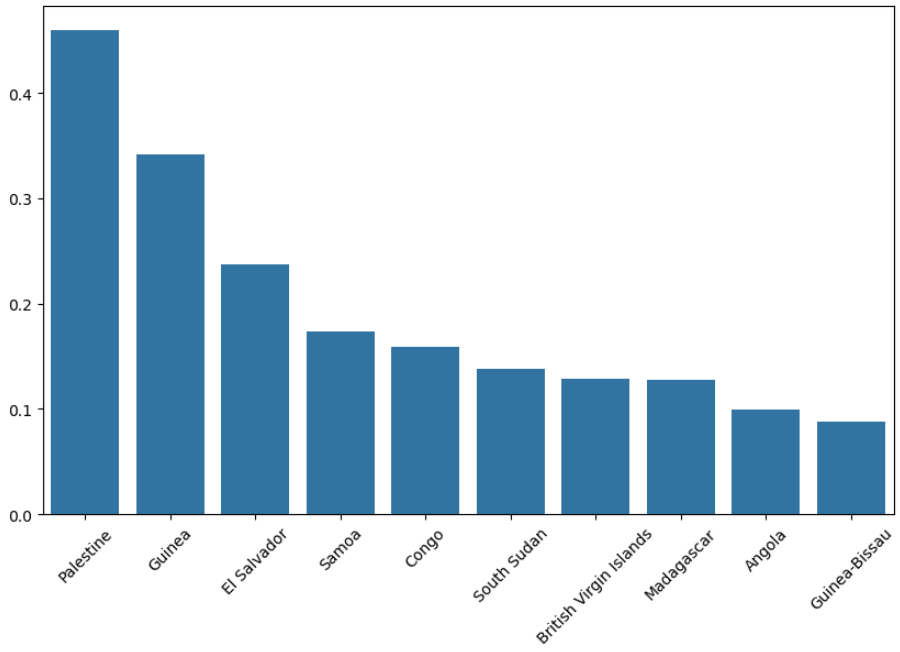
Table 3: Cross-Entropy Loss

	Linear Regression	Random Forest	LightGBM	XGBoost
Cross-Entropy Loss	0.40	0.39	0.34	0.65

4.3.2 Projection Results

Based on our model prediction, we expect 5 countries to win their first Olympic medal at the 2028 Los Angeles Olympics. The expected value is calculated by summing the medal-winning probabilities of all countries that have never won a medal, which is a simple probability model. The bar chart below lists the ten countries or regions with the highest probabilities of winning their first medal. The top three are Palestine, Guinea, and El Salvador, with probabilities of 45.9%, 34.1%, and 23.7%, respectively. These are the only countries or regions in the prediction with a probability exceeding 20%.

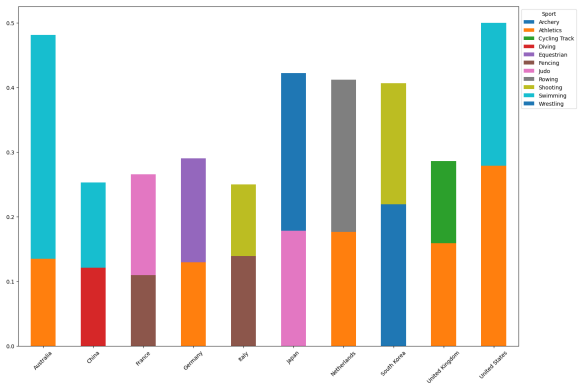
Figure 1: Probability of Winning a Medal



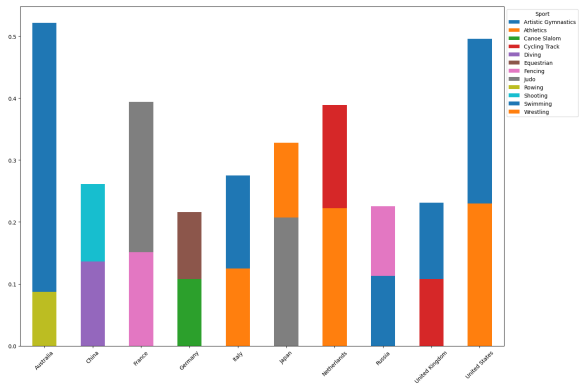
4.4 Impact of Events on Medal Counts

The dominant sports of each country vary, and the performance in these sports plays a crucial role in determining the total medal count for each nation. In this chapter, we explore the two sports with the highest medal share among the top ten countries on the medal table over the past four Olympic Games. We also focus on the performance of the three dominant sports of China and the United States in each Olympic Games. Finally, we examine the impact of new sports introduced in each Olympic Games on the host country's medal count.

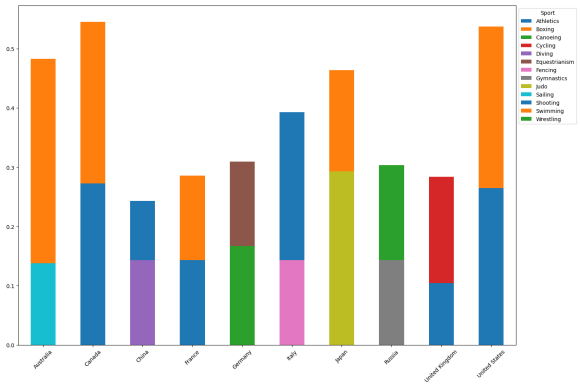
Figure 2: The Proportion of Events in Four Olympics



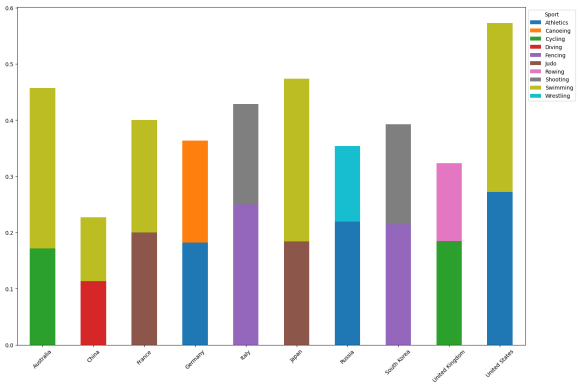
(a) Proportion in 2024



(b) Proportion in 2020



(c) Proportion in 2016



(d) Proportion in 2012

4.4.1 Top 2 Events for Countries

We first calculated the two sports with the highest medal share among the top ten countries on the medal table over the past four Olympic Games, as shown in the figure2. It is evident that different countries have different dominant sports. Athletics and Swimming are the most important for the United States, Diving, Swimming, and Shooting for China, Judo, Wrestling, and Boxing for Japan, Cycling and Rowing for the United Kingdom, Fencing and Judo for France, and Swimming and Sailing for Australia.

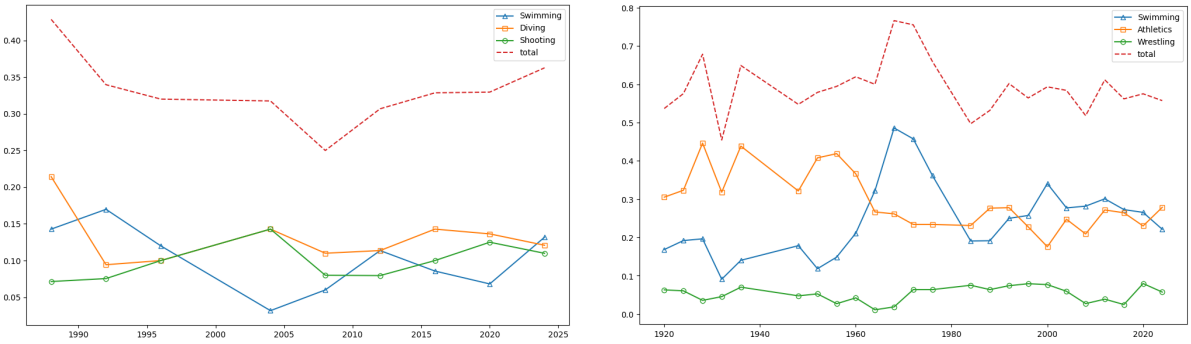
Of course, there is also overlap in dominant sports between countries. For instance, Swimming is a shared dominant sport for China, the United States, and Australia; Judo contributes significantly to the medal tally in both France and Japan; Athletics is a dominant sport for the United States, the United Kingdom, and the Netherlands. Therefore, the competition in these

shared dominant sports at the Olympics will be fiercer, and the events will become more exciting to watch.

4.4.2 Performance of the same event in different Olympic Games

In figure3, we have separately calculated the proportion of medals won in Swimming, Diving, and Shooting relative to the total medals for China, and the proportion of medals won in Swimming, Athletics, and Wrestling relative to the total medals for the United States in each Olympic Games. The red dashed line represents the cumulative proportion of these three sports. Undoubtedly, these sports play a very important role for each country. We also calculated the Pearson correlation coefficient and Spearman's rank correlation coefficient between each sport and the total number of medals, as shown in the table4 and table5. From this, we can conclude that there is a strong correlation between dominant sports and the total medal count.

Figure 3: Performance of strong events in various Olympics



(a) Events for China

(b) Events for United States

	Swimming	Diving	Shooting
PCCs	0.66	0.87	0.82
Spearman's ρ	0.69	0.82	0.67

Table 4: Correlation between events and medals(CHN)

	Swimming	Athletics	Wrestling
PCCs	0.64	0.74	0.66
Spearman's ρ	0.64	0.63	0.60

Table 5: Correlation between events and medals(USA)

5 Task2: Analysis of the "great coach" effect

It is widely acknowledged that coaches have a significant influence on the performance outcomes in sports. Unlike athletes, coaches are not required to hold citizenship of the countries they represent, allowing them to easily transfer between nations. Consequently, a "great coach" effect may arise, which is of considerable importance to investigate.

5.1 Model Development and Prediction

In the dataset provided, information on coaches is not explicitly available, and given the wide range of Olympic sports, identifying the "great coach" for each event is challenging.

Moreover, the definition of "great coach" itself lacks clear criteria. To facilitate the analysis, we first aim to develop a model capable of identifying a "great coach."

5.1.1 Labeled Dataset Construction

To train this model, a labeled dataset is required. As a representative case, the U.S. women's gymnastics team was selected due to its prominence and the relatively high number of medals it has achieved compared to sports such as volleyball. We extracted data related to the U.S. women's gymnastics team from the *summerOly_athletes* dataset and added a new column, "Legendary_Coach", which indicates whether the team was coached by a "great coach" during a particular Olympic Games (with a value of 1 for "great coach" and 0 otherwise).

Through research conducted via sources such as Wikipedia, we compiled information on the head coaches of the U.S. women's gymnastics team across various Olympic Games. We identified the head coaches of the 1976, 1984, 1996, 2000, 2004, 2008, 2012, 2016, and 2020 Olympics as "great coaches." After labeling the dataset accordingly, we randomly split it into a training set and a testing set in a 4:1 ratio, thereby preparing the dataset for model development.

5.1.2 XGBoost Model

Given the relatively small size of the labeled dataset, there is a risk of overfitting during model training. To address this issue, we chose to employ the XGBoost model for binary classification prediction.

XGBoost is an algorithm based on gradient-boosted decision trees. Unlike traditional gradient boosting decision trees (GBDT), XGBoost incorporates both L_1 and L_2 regularization terms (Lasso and Ridge) in its loss function. This feature enhances the model's ability to prevent overfitting and improves its generalization performance, making XGBoost particularly suitable for small datasets.

For a dataset containing n samples with m-dimensional features, the XGBoost model can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \ldots, n)$$

where $F = \left\{ f(x) = w_{q(x)} \right\} \quad \left(q : \mathbb{R}^m \to \{1, 2, \ldots, T\},\ w \in \mathbb{R}^T \right)$ is the set of CART decision tree structures, q denotes the structure mapping a sample to a leaf node, T is the number of leaf nodes, and w represents the scores associated with the leaf nodes. When constructing the XGBoost model, the optimal parameters are determined by minimizing the objective function to achieve the best model.

The objective function of the XGBoost model can be decomposed into the loss function L and the regularization term Ω:

$$\text{Obj} = L + \Omega$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\Omega = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

Here, γT represents the L_1 regularization term, and $\frac{1}{2}\sum_{j=1}^{T} w_j^2$ corresponds to the L_2 regularization term.

5.1.3 Prediction of "Great Coach"

After training the XGBoost model, we applied it to the remaining data to make predictions. For each Olympic Games, every country, and every event in which the country participated, the model generated a "great coach" score to quantify the quality of the coach for that event. To identify a threshold for the "great coach" score, which determines whether a coach can be classified as a "great coach," we needed to establish a "great coach" threshold.

To determine this threshold, we employed the elbow method. Specifically, we plotted a line chart with the "great coach" score on the x-axis and the number of coaches with a score exceeding that value on the y-axis (as shown in Figure 4). The x-coordinate of the inflection point, where the curve transitions from steep to flat, was selected as a candidate threshold. Since the number of coaches who qualify as "great coach" is expected to be relatively small, we further refined our selection by choosing a larger candidate threshold with a reasonable proportion on the y-axis.

Based on this approach, we determined the final threshold to be 0.71. Coaches with a "great coach" score exceeding 0.71 were classified as "great coaches."

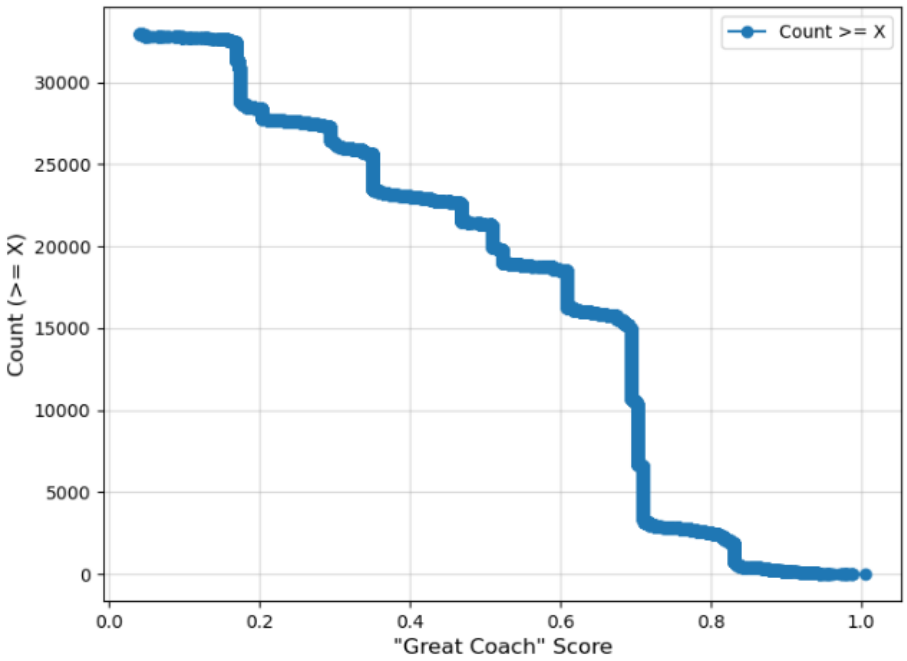


Figure 4: This figure is used to determine the "great coach" threshold.

5.1.4 Validation of "Great Coach" Prediction Results

Through the observation of the dataset and the search for background knowledge, we found some interesting trends and changes, especially in the performance of different countries and projects. These changes are closely related to the change of coaches.

Firstly, the Chinese badminton team made a remarkable improvement in the 2012 Olympics. In 2012, they won 5 gold medals, a significant achievement. This success can be credited not only to the athletes' hard work but also to the leadership of head coach Li Yongbo. Li's experience

and guidance were crucial in refining player skills, adjusting strategies, and building a strong, united team. His coaching methods contributed greatly to the team's excellent performance at the London 2012 Olympics.

Similarly, the Japanese gymnastics team saw a similar performance boost. In 2020, they improved significantly, winning 2 gold medals and 5 total medals. This success can be attributed to the leadership of head coach Hiroyuki Tomita. Under his coaching, the team improved their technical skills and tactical strategies, leading to better results at the Tokyo 2020 Olympics, surpassing past performances and making a lasting impact on the gymnastics world.

In the U.S., fencing also experienced significant improvement. In 2004, the U.S. fencing team won two medals, but by 2008, they earned six medals. This change can also be linked to a coaching shift. Greg Massialas was the coach in 2004, and in 2008, Yury Gelman took over. Gelman's training techniques and tactical changes brought fresh energy to the team, leading to improved performance and more medals at the 2008 Olympics.

Clearly, these coaches can be considered among the greatest. When our model predicts, it recognizes these coaches as the best, indicating that our predictions are accurate and convincing.

5.2 Correlation Between the "Great Coach" Effect and Medal Counts

To assess the contribution of the "great coach" effect to medal counts, we conducted a correlation analysis between the presence of a "great coach" and the medal outcomes. Using the data obtained in section 5.1, we grouped the events based on whether they were coached by a "great coach". Ignoring differences between sports and countries, we calculated the number of medals won and the number of non-medalists for both the "great coach" and "non-great coach" groups. This data was used to construct a contingency table, as shown in Table 6:

	Medals	**No Medal**	**Total**
Great Coach	11239	24852	36091
Non-Great Coach	26291	179991	206282
Total	37530	204843	242373

Table 6: Contingency Table for "Great Coach" and Medal Counts

5.2.1 Chi-Square Test Formula

The chi-square statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i represents the observed frequencies in the contingency table.
- E_i represents the expected frequencies under the null hypothesis, calculated as:

$$E_i = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

5.2.2 Results

Using the chi-square test formula, we computed the chi-square statistic for the contingency table. The calculated value was approximately:

$$\chi^2 \approx 7941.34$$

indicating a significant association between the presence of a "great coach" and the likelihood of winning medals.

The results suggest that whether a sport is coached by a "great coach" is significantly associated with the likelihood of winning a medal. Specifically, having a "great coach" significantly enhances a nation's probability of securing medals in a given event.

5.3 Investment Recommendations for "Great Coach"

In this section, we provide specific analyses based on the prediction results for "great coach," focusing on data from China, the United Kingdom, and Japan. We attempt to offer investment recommendations for these three countries concerning specific projects related to "great coach" and make rough estimates of its impact.

5.3.1 Data Processing and Analysis

Similar to the procedures in Section 5.2, we group the data by country, sport, gender, and the "great coach" indicator in sequence. For each group, we calculate the total number of medals won and the number of individuals who did not win medals.

Subsequently, we group the data by country, sport, and gender, and apply further data screening. Groups where the "great coach" value is either entirely 0 or 1 are excluded, as correlation analysis would be unfeasible. Additionally, groups with a small total number of participants are also removed, as this suggests that the sport is less prioritized in the country or that very few athletes have been sent, making the analysis of the "great coach" effect irrelevant.

After data screening, we proceed with analyses for China, the United Kingdom, and Japan, using Fisher's Exact Test to calculate the exact p-value for each 2×2 contingency table.

5.3.2 Fisher's Exact Test

Fisher's Exact Test is a statistical test used to determine if there are nonrandom associations between two categorical variables. It is particularly useful when sample sizes are small and the chi-squared test may not be applicable due to low expected frequencies.In our contingency table, some values may be 0, making it impossible to perform the chi-square test. Therefore, we opted to use Fisher's exact test to address this issue.

Given a 2×2 contingency table, we have the following general setup:

	Success	Failure
Group 1	a	b
Group 2	c	d

Where a represents the number of successes in Group 1, b represents the number of failures in Group 1, c represents the number of successes in Group 2, d represents the number of failures in Group 2.

The null hypothesis H_0 in Fisher's Exact Test is that there is no association between the two categorical variables (i.e., the distribution of successes and failures is independent of the groups).

To calculate the p-value, we compute the probability of obtaining the observed table (or one more extreme) under the assumption that H_0 is true. The exact p-value is given by:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Where $\binom{n}{k}$ is the binomial coefficient, which represents the number of ways to choose k successes from n trials.

The test calculates the probability of observing the given table's distribution, as well as all other possible distributions of the table under the assumption of independence. If the p-value is smaller than a significance level α (commonly 0.05), the null hypothesis is rejected, suggesting a significant association between the two categorical variables.

Thus, if the p-value is small, we conclude that the "great coach" effect is related to the outcome of winning medals, and it would be advisable to invest in "great coach" for that particular sport.

5.3.3 Recommendations for China's Investment in "Great Coach"

After calculating the p-values for each sport in which China participates (separately for men's and women's events), the p-values were rounded to four decimal places. We consider p-values less than 0.05 to reject the null hypothesis. Thus, all sports with p < 0.05 were selected as candidate recommendations. Additionally, the total number of Chinese participants in these sports across previous Olympic Games was included as a reference. The final recommendation table is presented as follows:

Index	Sport	Sex	p-value	total athletes
1	Gymnastics	F	0	392
2	Volleyball	F	0	149
3	Baseball and Softball	F	0	60
4	Athletics	M	0.0024	271
5	Table Tennis	M	0.0029	63
6	Gymnastics	M	0.0098	452
7	Taekwondo	F	0.0248	18
8	Judo	F	0.0296	58
9	Fencing	M	0.0315	155
10	Boxing	M	0.0358	49
11	Archery	F	0.0451	65

From the table, it is evident that both men's and women's Gymnastics have significantly low p-values, with the largest number of total participants. Therefore, we infer that investing in "great coach" for Gymnastics would yield the highest cost-effectiveness. Furthermore, based on a review of the medals won by the Chinese gymnastics team in previous Olympic Games, we estimate that the implementation of "great coach" could result in an increase of $5 \sim 12$ medals (combined for men's and women's teams).

In addition, China's women's Volleyball and women's Baseball and Softball also show significantly low p-values, making them strong candidates for investment in "great coach". Due to the medal allocation structure, it is estimated that "great coach" could potentially increase the Chinese women's Volleyball team's medal count by $0 \sim 1$ and the women's Baseball and Softball teams' medal count by $0 \sim 1$.

5.3.4 Recommendations for the United Kingdom's Investment in "Great Coach"

Using a methodology similar to that outlined in Section 5.3.3, the p-values for each sport in which the United Kingdom participates are shown in the table below:

Index	Sport	Sex	p-value	total athletes
1	Aquatics	M	0	1017
2	Canoeing	M	0	196
3	Field hockey	F	0	147
4	Cycling	M	0.0001	618
5	Gymnastics	M	0.0015	797
6	Modern Pentathlon	M	0.0069	100
7	Rowing	F	0.0076	209
8	Athletics	F	0.0083	725
9	Cycling	F	0.0247	123
10	Rugby	M	0.0494	55

Firstly, it can be observed that the p-value for the men's Aquatics events is significantly low, and the total number of participating athletes is high. Therefore, it is recommended to prioritize investment in "great coach" for this sport. Based on historical Olympic data, the effect of "great coach" is estimated to increase the medal count in men's Aquatics by $5 \sim 9$.

Additionally, the p-values for both men's and women's Cycling events are relatively low, with a considerable number of participants, making these events strong candidates for investment in "great coach". Historical Olympic data suggest that the "great coach" effect could result in an increase of approximately $3 \sim 4$ medals for the United Kingdom in Cycling, with men's and women's events combined.

Moreover, the p-values for the men's Canoeing and men's Field Hockey events are also close to zero. Although the total number of participants in these events is relatively small, investment in "great coach" for these sports could still be worthwhile. According to historical data, the "great coach" effect could lead to an increase of $1 \sim 3$ medals in men's Canoeing and approximately 1 medal in men's Field Hockey.

5.3.5 Recommendations for Japan's Investment in "Great Coach"

Using a similar methodology, the p-values for each sport in which Japan participates are presented in the table below:

Index	Sport	Sex	p-value	total athletes
1	Aquatics	F	0	622
2	Baseball and Softball	F	0	74
3	Baseball and Softball	M	0	135
4	Football	F	0	111
5	Volleyball	F	0	181
6	Table Tennis	F	0.003	59
7	Badminton	F	0.0218	59
8	Volleyball	M	0.0355	125

It is evident that the p-value for women's Aquatics is close to zero, and the total number of participating athletes is high. Thus, it is most recommended for Japan to invest in "great coach" for women's Aquatics. Based on historical Olympic data, the "great coach" effect is estimated to increase the medal count in this sport by $2 \sim 5$ medals.

Furthermore, the p-values for both men's and women's Baseball and Softball events are also close to zero, making them strong candidates for investment in "great coach". The "great coach" effect is estimated to increase the combined medal count in these events by $1 \sim 2$.

Lastly, the p-values for women's Football and women's Volleyball are close to zero, with a relatively large number of participants. Investing in "great coach" for these sports could also yield notable outcomes. Based on historical Olympic data, the "great coach" effect is estimated to increase the medal count in women's Football and women's Volleyball by $0 \sim 1$ medal each.

6 Task3: Other Original Insights about Olympic Medal Counts

In the process of building models to address the aforementioned tasks, we encountered some interesting findings and hypotheses that may provide new perspectives for studying the distribution of Olympic medals.

6.1 Geographic Clustering Effect

In the process of predicting countries likely to win their first Olympic medal, we observed an interesting phenomenon: countries that have never won an Olympic medal are often geographically surrounded by other countries with either no medals or very few medals. Conversely, the countries we predict to have the potential to win their first medal—namely Palestine, Guinea, and El Salvador—are located in regions where neighboring countries have won relatively more medals. This led us to a bold hypothesis: could the total distribution of Olympic medals across nations exhibit a geographical "clustering effect", where neighboring countries tend to have similar total medal counts?

To test this hypothesis, we first calculated the total number of Olympic medals won by each country based on historical Olympic medal data. Next, we categorized the total medal counts into eight intervals: 0, (0, 10), [10, 50), [50, 100), [100, 500), [500, 1000), [1000, 2000), and $\geq$2000. Using these intervals, we assigned a unique color to represent each category. Countries within each interval were shaded accordingly on a world map, while the three countries we predicted to potentially win their first medal were marked in black (also highlighted with a red circle in the map), as shown in Figure 5.

Analyzing this map, it is evident that adjacent medal intervals often correspond to geographically neighboring regions. For example, gray and purple, purple and blue, or orange and

yellow tend to appear in proximate locations. This observation suggests that the distribution of Olympic medals does indeed exhibit a geographic "clustering effect".

The geographic clustering of medal counts also implies that neighboring countries may excel in similar types of sports. This insight provides several strategic considerations for national Olympic committees (NOCs):

- **Regional Resource Integration and Cooperation**: National Olympic Committees (NOCs) can pursue cross-regional resource integration and international collaborations to achieve resource sharing and mutual advantages, thereby enhancing their national athletic performance.
- **Development of Advantageous Sports Based on Geographic Location**: NOCs can develop sports that align with their geographic features. For example, countries in high-altitude regions may focus on middle- and long-distance running or mountaineering, while tropical or coastal countries could emphasize water sports such as sailing or surfing.
- **Utilization of Geopolitical Advantages for Joint Training Programs**: NOCs can leverage geographical proximity to establish joint training programs and other collaborative initiatives to foster athletic development.

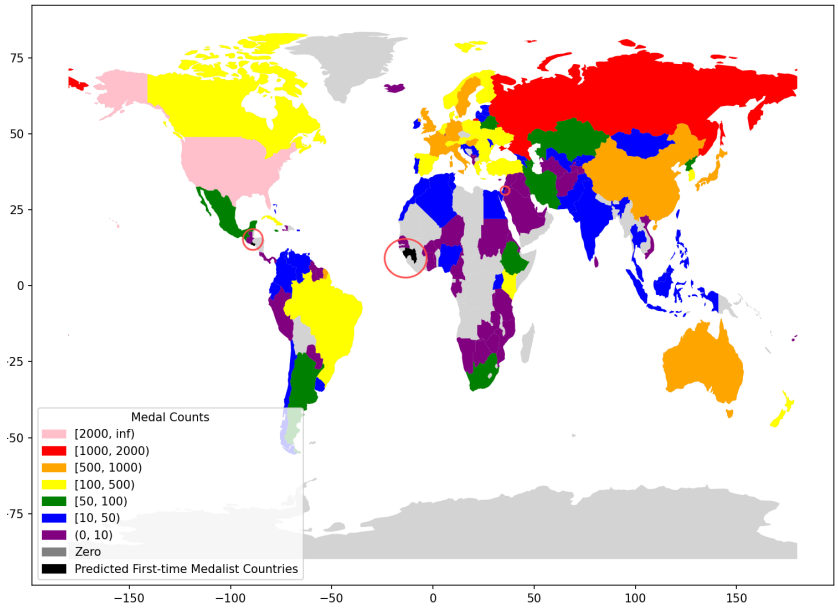


Figure 5: A world map displaying the distribution of total Olympic medals across all editions.

6.2 The Impact of GDP on Olympic Medal Counts

In our analysis of historical Olympic medal tables and the predictions generated by our model, we observed that countries ranking at the top of the medal table are typically those with the highest GDP, such as the United States, China, Japan, and the United Kingdom. Conversely, countries that have never won an Olympic medal often have relatively low economic levels. This raises an important question: is there a correlation between a country's GDP and its Olympic medal count?

The answer is affirmative. Generally speaking, nations with higher economic levels tend to have more abundant training facilities, more advanced equipment, and more scientific methods

for athlete development. They are also better equipped with comprehensive support systems, including dietary provisions, medical assistance, and psychological counseling. Furthermore, wealthier nations host a greater variety of competitions at different levels and scales, which significantly contribute to the improvement of athletes' technical skills, the enhancement of their physical and mental fitness, and the accumulation of practical experience. These factors—technical proficiency, physical and mental resilience, and practical experience—are critical indicators of an athlete's excellence in the Olympics.

In other words, countries with higher economic levels are more likely to produce a greater number of elite athletes, thereby achieving superior Olympic performance.

The positive impact of a country's GDP on its Olympic medal count offers the following insights for National Olympic Committees (NOCs):

- **Investing in Sports Development**: Higher GDP enables greater financial resources for sports infrastructure, training facilities, athlete development programs, and international exposure. NOCs can advocate for increased investment in these areas to enhance medal-winning potential.
- **Optimizing Resource Allocation**: Countries with lower GDPs can focus on strategically allocating resources to sports where they have a competitive advantage or higher chances of winning medals, thereby maximizing their returns on investment.
- **Promoting Public-Private Partnerships**: NOCs can collaborate with private enterprises to secure additional funding for sports development, particularly in nations with limited government budgets.
- **Fostering Economic Growth Through Sports**: Successful sports performance can stimulate economic benefits, such as increased tourism and brand visibility. NOCs can emphasize the broader economic value of investing in sports to secure support from governments and stakeholders.

6.3 Reasons for the medal differences between China and USA

As the two largest economies in the world, the sports competition between China and the United States has garnered widespread attention. The Paris Olympics have come to a close, and China won a total of 40 gold medals, a remarkable achievement worth celebrating. However, despite both China and the United States winning the same number of gold medals—40 each—America's total medal count reached 126, far surpassing China's 91. What is the reason behind such a large gap?

We believe this gap is largely due to the rules and regulations. For example, in the Paris Olympics, China's participation in weightlifting was limited to six athletes, while in the Tokyo Olympics, it was eight. This reduction led to two fewer gold medals for China. Without these restrictions, China would likely have led the medal table. However, the U.S. allows more athletes to compete in their strong events, thus winning more medals. Meanwhile, China faces strict limits on the number of participants in its strong events, such as diving, weightlifting, and table tennis, which directly affects its performance in these sports.

For instance, if China were allowed to send three athletes per weight category in weightlifting, its medal count would increase significantly, possibly surpassing the U.S. In table tennis

and diving, if China could increase its number of participants, it would certainly win more medals. Therefore, if China's participation in its strong events were not restricted, the U.S. would not only risk losing its position at the top of the gold medal count but also see its overall medal lead threatened.

The reasons behind the medal differences between China and the USA can provide the following recommendations to their respective National Olympic Committees (NOCs):

- **Steady progress with a broad vision**: While maintaining dominance in traditional strong events, it is also necessary to broaden the potential medal opportunities in other events.
- **Improve Athlete Support and Resources**: Both countries should continue to enhance athlete support systems, such as mental health resources, nutrition, training facilities, and access to top coaches. This ensures athletes perform at their peak in their respective events.
- **Advocating for Fairer Rules**: Both countries should advocate for fairer rules and regulations that better support the development of athletes from all sports. For instance, medal points for event-specific quotas could be adjusted to allow more equal representation for athletes from developing nations, thus ensuring a more equitable global competition.

7 Model Analysis

7.1 Strengths

- We compared the performance of various models on the same task and selected the best-performing model. We then fine-tuned and validated it to ensure its stability and reliability across different scenarios.
- Our model integrates multi-dimensional data, such as the number of participants in each event, whether a country is the host, and the average gold/medal counts from the past three Olympics to predict the medal table, achieving a low error rate.
- Our model can make predictions regarding the "great coach" through training although original dataset does not contain information about the "great coach". The use of correlation analysis to reflect the "great coach" effect is both clever and convincing.

7.2 Weaknesses

- The GDP of participating countries in different eras and other factors may also influence the predictions, but due to the lack of data, they were not included as features in the model.
- The lack of labeled data sets necessitates extensive manual annotation, leading to insufficient training data.
- The analysis of the "great coach" effect is based solely on historical Olympic data, and no precise method for calculating the effect has been proposed.

7.3 Conclusion

In this paper, we develope a medal prediction model based on random forests and make projections for the upcoming 2028 Los Angeles Olympics medal table. During this process, we analyze countries that might achieve their first-ever medals and examine the relationship between each country's dominant sports and its total medal count. Additionally, we build a model to explore the "great coach" effect. We use the XGBoost model to predict which events are coached by a "great coach" and employ correlation tests to verify whether the presence of a "great coach" is related to winning medals in those events. Additionally, we conduct Fisher's Exact Test to calculate the p-values for each event in three countries, offering recommendations on which sports these countries should invest in a "great coach". Finally, we highlight three other findings discovered during the data analysis and provide some recommendations for national Olympic committees based on these insights.