Predicting Olympic Glory:

Unveiling the Path to Victory through Data-Driven Insights

Yanqi Jiang[1], Yixuan Sun[2]

[1, 2]School of Computer Science and Engineering, Nanjing University of Science and Technology

Abstract

With the rapid advancement of artificial intelligence and machine learning, there is an increasing need for accurate prediction models in various fields, including sports. This study develops a prediction model for Olympic medal distribution using the XGBoost algorithm, leveraging historical athlete performance data to forecast medal outcomes for the 2028 Summer Olympics in Los Angeles. The model considers multiple factors such as historical medal counts, athlete experience, national background, and the influence of coaching on performance.

For the Model, we extensively reviewed existing literature and selected the XGBoost model due to its proven effectiveness in classification tasks. We incorporated GridSearchCV for hyperparameter tuning and Bootstrap resampling to account for un-certainty in predictions. Our preliminary analysis on historical data of 160 countries re-vealed a consistent trend of improvement in national performance over time. Specifically, the model predicts that the United States and China will continue to dominate, while coun-tries like France and Italy are expected to face a decline in medal counts. We used feature importance analysis to assess which factors most significantly influence medal outcomes. Our results show that national historical performance and the experience level of athletes are key determinants in predicting Olympic success. The model achieved an accuracy of 91.31%, with a high precision in predicting gold medal winners.

For the Mould, we conducted a Cost-benefit Analysis of the model’s implementation, focusing on the implications for National Olympic Committees (NOCs). The analysis suggests that the model can serve as a tool for strategic resource allocation and athlete development. We also explored the role of “great coaches” and found that their impact on medal outcomes is significant, with countries benefiting from strategic coaching invest-ments.

For the Movement, we applied the model to predict the medal distribution for the 2028 Olympics, specifically focusing on first-time medal winners. We found that 15 countries are likely to win their first Olympic medal, with a strong likelihood of success in certain disciplines like athletics and swimming. In addition, the analysis revealed that home-field advantages and the selection of event types have a substantial impact on medal tallies, which should inform future strategic decisions for NOCs.

Finally, we summarized our findings into a non-technical report aimed at providing insights to NOCs, helping them to optimize their training and resource allocation strate-gies. The sensitivity analysis highlighted that the model is adaptable and can improve as more data is integrated, making it an invaluable tool for future Olympic predictions.

Keywords

Olympic prediction model, XGBoost, machine learning, sports analytics, feature importance, coach effect, National Olympic Committees.

Introduction

Thomas Bach, President of the International Olympic Committee said: “The Olympic Games are the world’s greatest celebration of the human spirit.”

Question Background

The Olympic Games are not only a platform for athletes to showcase their talents and pursue excellence but also an important symbol of a country's sporting strength. The medal table, as a key indicator of a nation's performance at the Olympics, has long attracted global attention and analysis. Predicting a country's medal tally is not only a complex analytical task but also a highly meaningful endeavor. It not only reflects a nation's overall strength in the field of sports but also provides valuable insights for shaping future sports development strategies, while simultaneously sparking widespread public interest and discussion about competitive sports.

Figure 1(a): Glory at the Eiffel Tower.

Figure 1(b): Hurdler at the start.

Question Restatement

The ranking of the medal table is influenced by several factors, including national sports resources, training systems, athlete abilities, and coaching. Based on these factors, the 2028 Olympic medal table is predicted. The main tasks are as follows:

Task 1: Develop a model to predict the number of gold and total medals for each country. Evaluate the model's accuracy and uncertainty. Using this model, predict the medal tally for the 2028 Summer Olympics in Los Angeles and analyze which countries are likely to improve and which may perform worse than in 2024.

Task 2: Examine the impact of the "excellent coach effect" on medal distribution. Es-timate its contribution and identify three countries where investing in top coaches could significantly impact results.

Task 3: Investigate the relationship between Olympic events and medal counts. Ana-lyze which sports are most important to each country and how the host country's event se-lection affects medal distribution. Provide additional insights to help guide National Olym-pic Committees, based on findings about the importance of specific sports and host country event choices.

Our Work

In this study, we combine the XGBoost model with medal data to develop a prediction model for

Olympic medals. First, using the XGBoost model, we calculate the probability of each athlete winning a medal, which allows us to predict the expected medal table. We also estimate the number of countries that are likely to win their first Olympic medals. Next, we analyze the relationship between different sports and countries by examining the partial contributions of each sport to the overall medal tally for each country. Then we investigate the influence of "great coaches" on national medal counts. Using examples mentioned in the paper, we hypothesize that the presence of an great coach correlates with changes in a country's medal performance. Finally, we analyze the importance of different features that influence the medal distribution and draw some interesting conclusions.

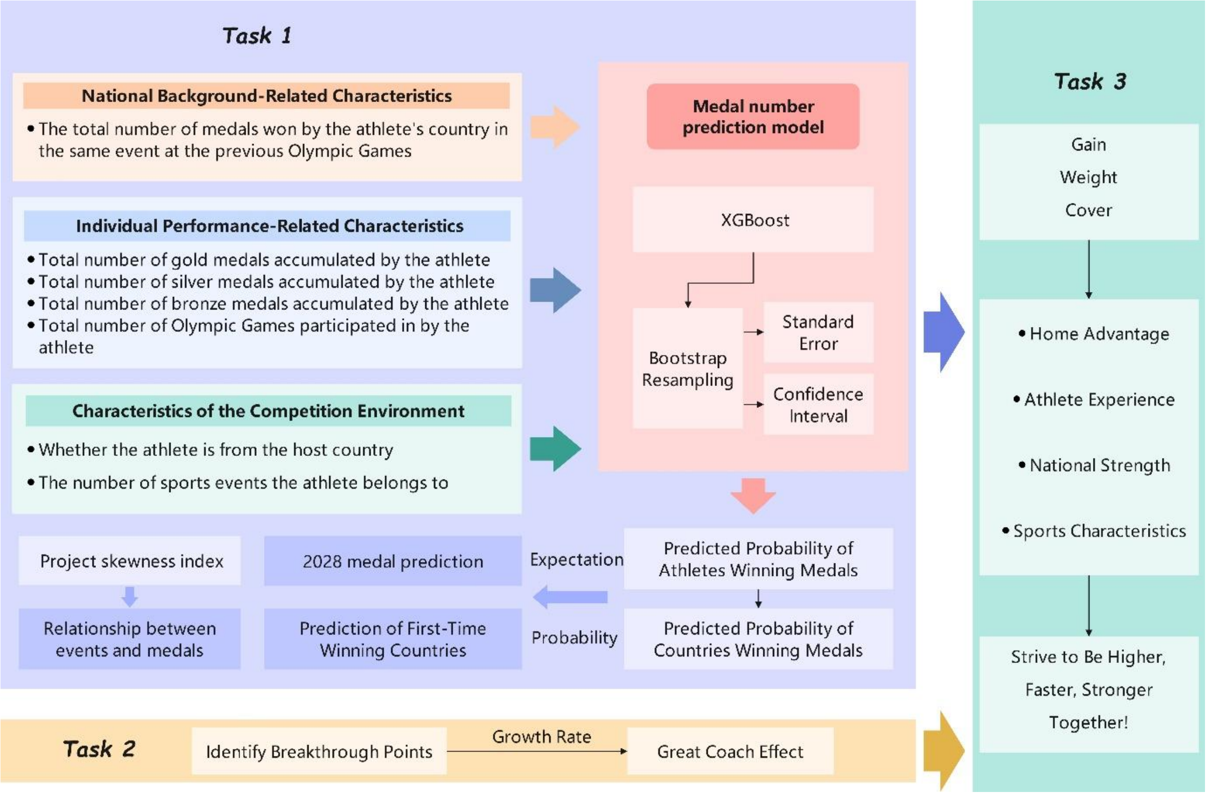

Task 1
National Background-Related Characteristics
• The total number of medals won by the athlete's country in the same event at the previous Olympic Games
Individual Performance-Related Characteristics
• Total number of gold medals accumulated by the athlete
• Total number of silver medals accumulated by the athlete
• Total number of bronze medals accumulated by the athlete
• Total number of Olympic Games participated in by the athlete
Characteristics of the Competition Environment
• Whether the athlete is from the host country
• The number of sports events the athlete belongs to
Medal number prediction model
XGBoost
Bootstrap Resampling
Standard Error
Confidence Interval
Project skewness index
Relationship between events and medals
2028 medal prediction
Prediction of First-Time Winning Countries
Expectation
Probability
Predicted Probability of Athletes Winning Medals
Predicted Probability of Countries Winning Medals
Task 2
Identify Breakthrough Points
Growth Rate
Great Coach Effect
Task 3
Gain
Weight
Cover
• Home Advantage
• Athlete Experience
• National Strength
• Sports Characteristics
Strive to Be Higher, Faster, Stronger Together!

Figure 2: Flow chart of our work

Assumption and Justification

To simplify our model and make it more manageable, we make the following well-justified assumptions based on a thorough analysis of the problem:

Assumption 1: Athletes who participated in the last Olympics will show no significant change in their physical condition for this Games.

Justifications: The majority of Olympic athletes are typically aged between 20 and 24, a range during which physical performance generally peaks. Athletes tend to maintain their physical capabilities from one Olympic cycle to the next, especially if they are in their peak years (between 20 and 30). This assumption allows us to treat their performance in the 2028 Olympics as comparable to their performance in 2024, assuming no major disruptions to their training or health.

Assumption2: The list of athletes participating in the 2028 Olympics will be the same as in

2024.

Justifications: It is challenging to predict the exact list of athletes who will compete in the 2028 Olympics based on the available data, especially since new athletes emerge and others may retire. Furthermore, the first-time Olympic athletes often face challenges in winning medals due to their lack of experience in high-stakes competition. For modeling purposes, we assume continuity in the athlete pool, as this simplifies the prediction process while still offering valuable insights into the potential outcomes for countries and athletes based on the current trends.

Notations

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

Table 1: Notations used in this paper

Symbol	Description
X	Feature set for each athlete
p_k	Predicted probability for medal category k
μ_k	Mean of predicted probabilities for medal category k
SE_k	Standard error for the predicted probabilities for medal category k
CI_k	Confidence interval for medal category k
w_k	Weight assigned to different medals in the model

Medal number prediction model

Model Selection

To predict the probability of each athlete winning a medal, we employ the XGBoost model. This model offers numerous advantages in predicting the Olympic medal table. It builds decision trees incrementally using the gradient boosting method, adjusting the model based on the errors from the previous round in each iteration, thereby optimizing the loss function for high prediction accuracy. Additionally, through parallel computing and optimi-zation strategies, the XGBoost model demonstrates a significant advantage in training speed, making it well-suited to handle the survey data provided in this problem.

Data Description

There are vacancy values for some fields in the data set, such as missing event or medal information in some years. To solve this problem, we use the following strategies: numerical fields are filled with the average, category fields are filled based on the context information or domain knowledge, and vacancy values that cannot be inferred are directly marked as "unknown" for exclusion or separate processing in subsequent analysis.

Outlier handling: outliers were found in some fields, such as abnormally high or low number of matches, unreasonable distribution in medal records, etc. For numerical fields, outliers were detected by box plot method and adjusted or removed according to the reasonable range. For outliers in the category field, this is corrected by alignment to the standardized name list.

Special cell analysis: In the event table, some cells not only contain numbers, but also the special symbol " • ", indicating that the event is an performance or informal event in an Olympic Games. We added a Boolean field Is_Artistic, extracting all cells containing " • " and marking them as "1".

Change rules for the Team field: using the list of standardized names on the International Olympic Committee (IOC) website, such as changing "USSR" to "Russia". For team numbers (e. g. "Germany-1"), extract only the country name section (e. g. "Germany") to maintain consistency with other fields. For team names that cannot be resolved, marked as "Unknown Team" and stored separately for subsequent manual verification.

Normalization of the Sport, Discipline, Event fields: The Sport, Discipline, and Event fields have spelling differences in the data, such as extra spaces, case differences, or special characters. To ensure the consistency in these fields, we performed the normalization.

Discipline Fuzzy match with Event: There is a correspondence between Discipline and Event, and the following three steps are used to match: preferentially check whether Sport and Event can directly match the standardized Discipline field. If there is a unique match, the direct record is a complete match; for the records that cannot be directly matched, the RapidFuzz library is used to calculate the similarity of Event and Discipline, and the threshold is set at 0.5 (50%). Only records with similarity greater than the threshold are considered to match successfully; **manual mapping**: For fuzzy matching records still unsuccessful, use manual mapping table alignment. For example, manually tag matching relations in some special cases.

After completing the data preprocessing, this study selected the characteristics closely related from the dataset to ensure that these characteristics can fully reflect the key factors such as historical performance, national background and competition environment. These characteristics mainly include the following categories:

Characteristics of the competition environment

Whether the athlete is from the host country (0 / 1)

Through the match of the competition year and the host country and the host country, the athletes of the host country may have home field advantage, such as more familiarity with the competition environment, the potential bias of referees, and national policy support. This home-court effect may significantly increase the medal probability of the athletes.

The number of sports events that the athletes belong to

According to the statistics of competition items, the number of projects in each Sport and the number of projects may affect the difficulty of competition. The more projects in the sports category may mean the more competition.

Individual performance-related characteristics of the athletes

Total number of gold MEDALS accumulated by the athletes

Total number of silver medals accumulated by the athletes

Total number of bronze medals accumulated by athletes

Athletes have participated in the Olympic Games accumulatively

The total number of athletes in the Olympic data. Participating experience may have an important impact on the performance of the athletes, and the experienced athletes may have an advantage in the competition.

National background-related characteristics

The total number of medals won by the athlete's country in the same event at the previous Olympic Games

According to the historical data of athletes' MEDALS, the total number of MEDALS in the corresponding events in the last Olympic Games is counted. National historical medal performance can reflect its overall strength in a particular event, and this national back-ground may have an important impact on the performance of athletes.

XGBoost model establishment

For a given dataset with n examples and the aforementioned features, a tree ensemble model utilizes a combination of additive functions to predict the output.

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

Where $F = \{f(x) = w_q(x)\}(q{:}R^m \to T, w \in R^T)$ is the space of regression trees

Based on the concept of gradient boosting, each round of training updates the model by focusing on the residuals from the previous round, and iteratively optimizes the model using an additive approach:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot h_t(x) \tag{2}$$

Where $\hat{y}^{(t)}$ is the current predicted value, η is the learning rate, and $h_t(x)$ is the decision tree in round t.

Next, a regularization term is incorporated into the model to control complexity and prevent overfitting. This regularization ensures that the model generalizes well to unseen data. The regularization function helps in balancing the model's bias-variance tradeoff by penalizing overly complex models, thus improving its performance and stability. In the case of XGBoost, both L1 (Lasso) and L2 (Ridge) regularization can be applied, which further enhances the model's ability to generalize by controlling the growth of individual trees and minimizing overfitting. The final model is then trained iteratively, using an additive approach to combine the results of each individual tree, progressively improving the accuracy of the predictions.

The probability of athlete

The dataset is divided into training and test sets, with the model trained on the former and evaluated on the latter. To optimize model performance, **GridSearchCV** is employed for hyperparameter tuning. This technique performs an exhaustive search over a specified parameter grid and selects the best combination of hyperparameters based on cross-validation performance, ensuring optimal model configuration. During the training process, particular attention is given to adjusting for any biases introduced by country-specific factors, such as differences in athlete participation, resources, or historical performance. These biases are mitigated to ensure the model's fairness and reliability across different countries.

The model achieved the following hyperparameters for optimal performance:

Learning rate: 0.2

Max depth: 10

Number of estimators: 300

By establishing the best model, we predicted the award situation of athletes participating in 2028.In Table 2,we only show the probability of some athletes:

Table 2:The probability of some athletes winning a medal

Name	2024 Medal	No medal	Bronze	Silver	Gold
Lena Grandvea-u	Silver	0.048	0.004	0.934	0.014
Alex Yee	Gold	0.353	0.097	0.191	0.359
Alexander Massialas	No medal	0.466	0.290	0.170	0.074
Maude Charron	Silver	0.875	0.050	0.038	0.037

Model evaluation

The final accuracy of the model was **91.31%**, demonstrating that the model could predict medal winners with high precision.

The table 3 summarizes the precision, recall, and F1-score for each category:

Table 3:Precision, Recall, and F1-Score by Medal Category

Category	Precision	Recall	F1-Score	Support
Gold	0.91	1.00	0.95	7456
Silver	0.92	0.34	0.49	422
Bronze	0.97	0.41	0.58	445
No Medal	0.95	0.52	0.67	418

Gold Medal: The model performed excellently in predicting Gold medal winners, with high precision (0.91) and recall (1.00), meaning nearly all Gold medalists were correctly identified by the model.

Silver Medal: The model achieved a high precision of 0.92 for predicting Silver medal winners, but the recall was relatively low (0.34), indicating that many Silver medalists were missed.

Bronze Medal: The precision for Bronze medal prediction was 0.97, but with a recall of 0.41, the model still missed a significant portion of actual Bronze winners.

No Medal: The model achieved high precision (0.95) for athletes not winning medals, but the recall was 0.52, suggesting room for improvement in accurately predicting non-medalists.

Resampling and the confidence intervals

We employed the **Bootstrap resampling method** to evaluate the uncertainty of the model's predictions. This technique involved generating multiple resampled datasets from the original data to assess the stability of the model's predictions. Additionally, we calculated the confidence intervals for each medal category, providing a range within which the true predictions are likely to fall. The following is a description of the process used to perform this analysis:

Resampling: We set the number of resampling iterations to 100. In each iteration, samples are randomly drawn with replacement from the original training dataset to generate new resampled datasets, denoted as **X_resampled** and **y_resampled**. This process allows for the creation of multiple variations of the training data, which helps in evaluating the stability and uncertainty of the model's predictions by testing it on these different subsets.

After generating the resampled datasets (**X_resampled** and **y_resampled**) in each iteration, the model is trained on these new datasets, and predictions are made for the test set. This process is repeated 100 times, providing a distribution of predictions for each medal category. From this distribution, we can calculate the **standard deviation** for the predicted medal counts.

$$S\tan dard\ Error=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^{2}} \tag{3}$$

Where x_i is the predicted probability of each resampling, $\overline{x}$ is the mean of these predicted probabilities, and n is the number of resampling.

To further quantify the uncertainty of the model predictions, we calculated the 90% confidence intervals for each category.

Mean calculation: For each medal category, the mean of the predicted probability is calculated.

$$\mu=\frac{1}{n}\sum_{i=1}^{n}P_{ij} \tag{4}$$

Where P_{ij} is the predicted probability of the j category of the $i-th$ resampling.

Confidence interval calculation: The confidence interval is calculated using the standard error and Z value (for 90% confidence interval, Z value is about 1.645).

$$Confidence\ Interval=\mu\pm Z\times SE_k \tag{5}$$

Where SE_k is the standard error of category k.

Table 4 presents the confidence intervals and standard deviations (with a confidence level of 0.9) for the probabilities of athletes winning various medals.

Table 4:The probability of some athletes winning a medal

Category	Standard error	Confidence interval
Gold	0.122	±0.201
Silver	0.109	±0. 179
Bronze	0.109	±0.179
No Medal	0.205	±0.337

National medal count forecast

First, we calculate their expectations of gold, silver, and bronze awards for each athlete j in each country C_i. For example, assuming that the probability of winning gold, silver, and bronze awards is $P_{Gold,ij}$, $P_{Silver,ij}$ and $P_{Bronze,ij}$, then the athlete's medal expectation is:

$$E_{medal,ij} = P_{Gold,ij} + P_{Silver,ij} + P_{Bronze,ij} \tag{6}$$

In other words, each athlete's medal is expected as the sum of their gold, silver and bronze awards.Number of countries winning the first medal

Next, we sum up the medal expectations of a national athlete j to obtain the total medal expectations of the countryC_i:

$$E_{medals,i} = \sum_{j=1}^{m_i} E_{medal,ij} \tag{7}$$

That is, the expectation of national medals is the sum of the expectation of all athletes

The results are as follow table 5 , showing only the top 10 countries in the medal table

Table 5: Predicted Range of National Medal Counts for the 2028 Olympic Games

Team	Bronze	Silver	Gold	Tota
United States	[35.784, 43.553]	[36.543, 44.312]	[42.193, 50.917]	[114.519, 138.782]]
China	[22.507, 29.431]	[25.295, 32.219]	[35.354, 43.129]	[83.156, 104.779]
Great Britain	[26.699, 32.928]	[17.682, 23.911]	[15.288, 22.283]	[59.668, 79.121]
France	[16.547, 23.631]	[16.781, 23.866]	[13.470, 21.425]	[46.799, 68.922]
Autralia	[16.966, 23.701]	[13.423, 20.157]	[13.728, 21.290]	[44.117, 65.149]

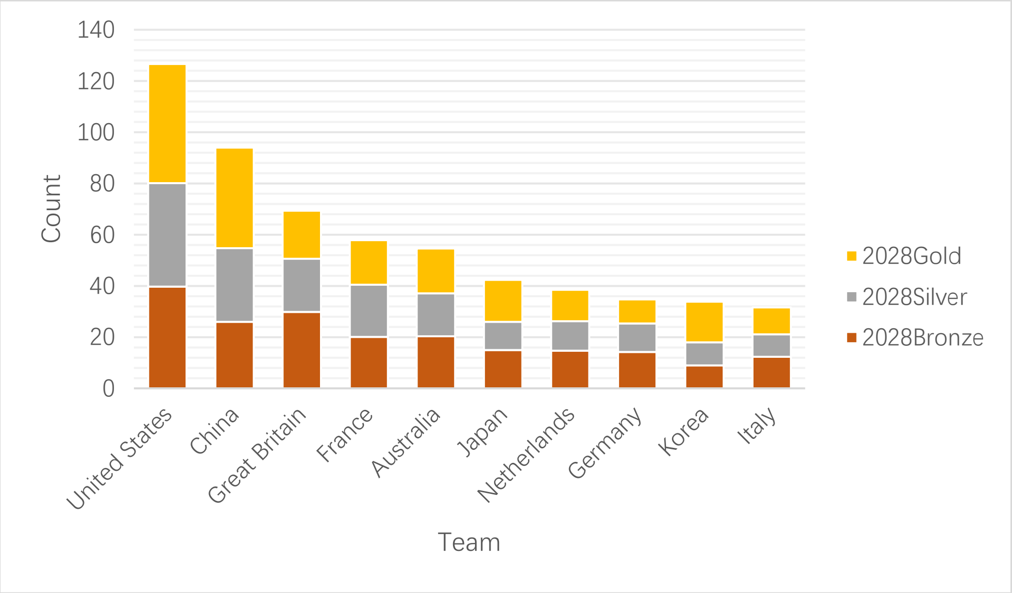


Figure 3: National Medals at the 2028 Olympic Games.

We know the change value of the 2028 medal after a simple calculation.

$$\varDelta N = N_{2028} - N_{2024} \tag{8}$$

The calculation results are shown in Table 6, which only displays the medal growth of the 10 countries in the medal table:

Table 6:medal number change table

Country	Bronze increase	Silver increase	Gold increase	Total increase
United States	-2.331	-3.573	6.555	0.650
China	1.969	1.757	-0.759	2.967
Great Britain	0.813	-1.204	4.785	4.395
France	-1.911	-5.677	1.447	-6.140
Australia	4.334	-2.210	-0.491	1.633
Japan	2.003	-1.009	-3.683	-2.689
Germany	6.696	-1.522	0.274	5.448
Italy	-0.788	-1.922	-2.588	-5.298
Netherlands	-3.019	1.968	0.866	-0.184
New Zealand	1.182	1.846	-0.952	2.077

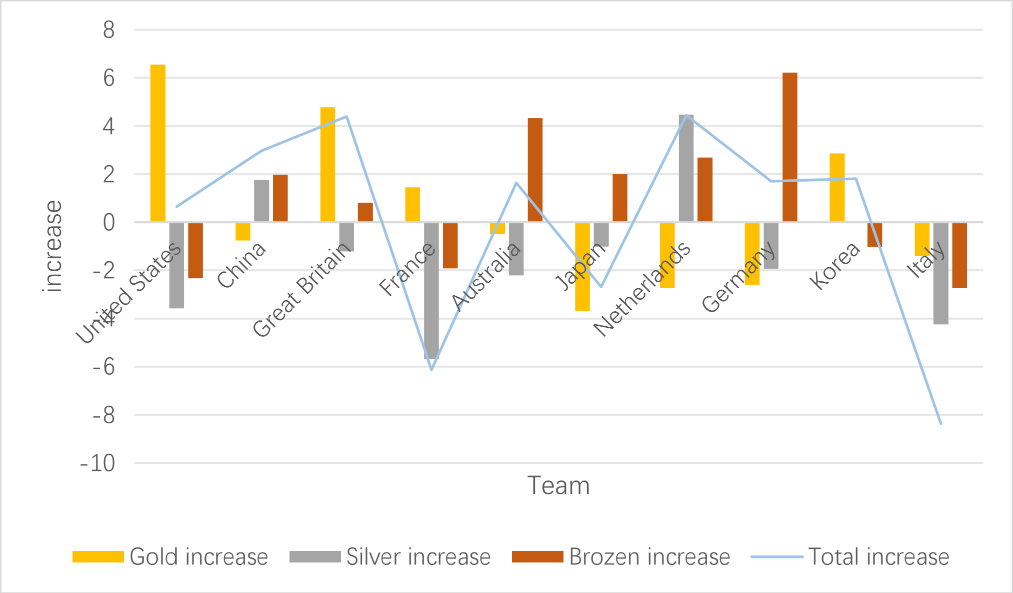


Figure 4 medal number change table

From the chart above, we can observe notable changes in the number of medals won in 2024. The United States maintains a medal count similar to that of the previous year, while countries such as Great Britain and the Netherlands are expected to show improvement. In contrast, France, Italy, and several other nations are likely to experience a decline in their medal counts compared to the last Olympic Games.

Countries with first-time medals

Consider n countries $C_i(i=1,2,...,n)$, with m_i athletes in each country C_i. Among them, each athlete The probability of winning the award is P_{ij} , where j indicates the j athlete. We define K_i as country C_i for obtaining at least one Probability of awards. We first find the probability that a certain athlete fails to get a medal, and multiply the probability of all of the country's athletes to obtain the probability of the country's first medal.This probability can be expressed as follows:

$$K_i = 1 - \prod_{j=1}^{m_i} (1 - P_{ij}) \tag{9}$$

Where$(1-P_{ij})$ indicates the probability that the athlete j does not receive the award.

After finding the probability of a country winning a medal for the first time, summing the expectations of all the probability countries expects the number of countries winning the first medal in the upcoming Olympic Games.The number of countries that won win for the first time in the upcoming Olympics is as follows:

$$E[N] = \sum_{i=1}^{n} K_i \tag{10}$$

The calculated expectations of new medal countries are:15.467.

The results are in the table 7, showing only the top 3 countries in the table:

Table 7:Expectations for some countries to win their first medal

Team	No medal Product	Have medal Product
Mali	0.336	0.664
South Sudan	0.414	0.586
Guinea	0.452	0.548

Project and the number of national MEDALS relationship

To measure the degree of importance of an event to a country, we define the value of a project as the sum of the value of the medal awarded in that event:

$$Total\ Value = 6 \cdot Gold + 3 \cdot Silver + 1 \cdot Bronze \tag{11}$$

From the value of the project, we can calculate the proportion of the total value of the medal of the countries participating in the project.

$$\begin{aligned} Gold\ Proportion &= \frac{Project\ Gold}{Country\ Total\ Gold} \\ Silver\ Proportion &= \frac{Project\ Silver}{Country\ Total\ Silver} \\ Bronze\ Proportion &= \frac{Project\ Bronze}{Country\ Total\ Bronze} \\ Total\ Value\ Proportion &= \frac{Project\ Total\ Valu}{Country\ Total\ Value} \end{aligned} \tag{12}$$

Using China and the United States as examples, the following illustrates the top ten sports contributing to their national medal counts:

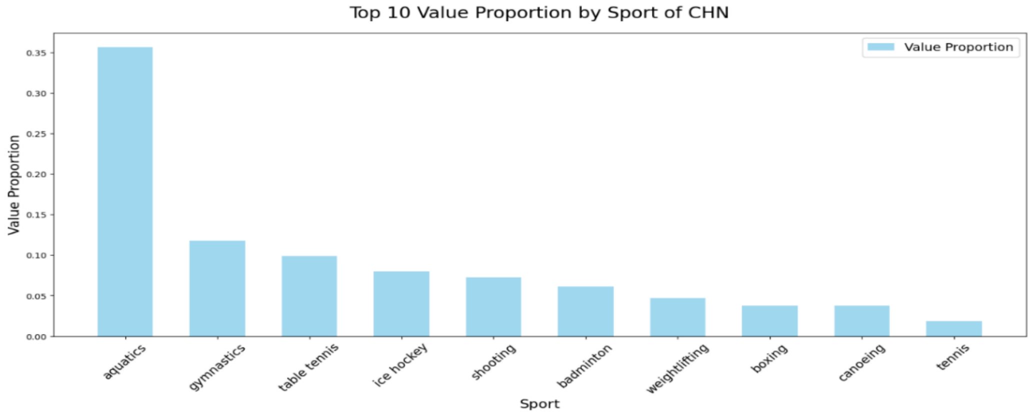


Figure 5:Top 10 Value Proportion By Sport of China

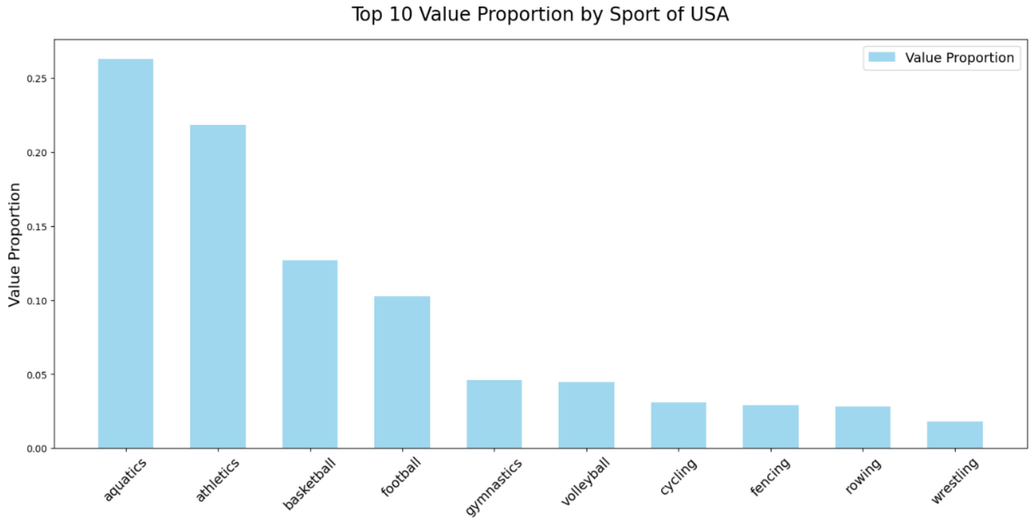


Figure 6:Top 10 Value Proportion By Sport of United Status

It can be seen that track and field events account for a large proportion of the United States and China, so we further analyze the impact of different Discipline.

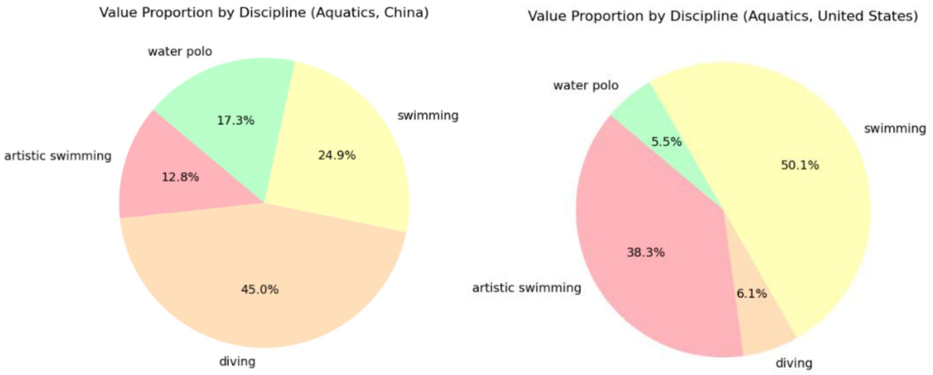
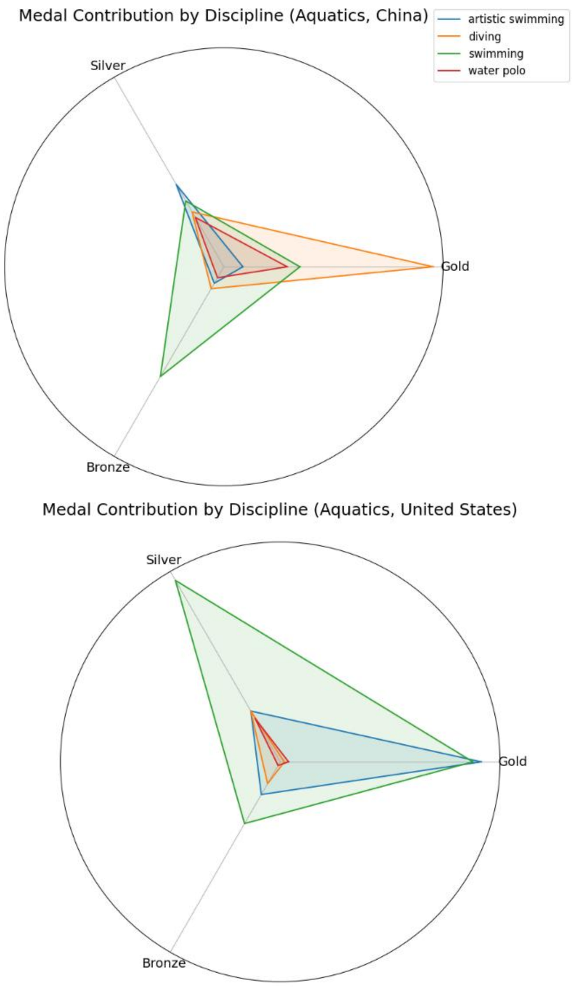


Figure 7: Value Proportion by Discipline (Aquatics, China vs. United States)

It can be seen that for China, diving accounts for the largest proportion, reaching 45%, which is the core advantage of Chinese water sports; followed by swimming (24.9%) and water polo (17.3%), the contribution of artistic swimming is relatively small (12.8%). For the United States, swimming contributed the most outstanding, accounting for 50.1%, becoming the absolute main force; art swimming accounted for 38.3%, ranking second, while diving (6.1%) and water polo (5.5%) are relatively low. It can be seen that the layout and focus of the two countries are obviously different under different Discipline.

Next, we will further analyze the specific effects of these events on the medal distribution.



Medal Contribution by Discipline (Aquatics, China)

Medal Contribution by Discipline (Aquatics, United States)

Figure 8: Medal Contribution by Discipline

(Aquatics, China vs. United States)

For China, diving has always been a core advantage, with the most prominent contribution of gold medals and a significantly broader coverage compared to other events. For the United States, swimming performances are particularly strong, with the coverage of gold, silver, and bronze medals far exceeding other events, showcasing the absolute dominance of the U.S. in swimming.

The distribution and value of medals across different sports and disciplines have a profound impact. This difference not only reflects the varied sports strategies of nations but also provides an important reference for understanding the patterns of Olympic medal distribution.

The skewness index is usually defined as a quantitative measure of the difference of students' differences in learning performance in different subject areas. This index can reflect the learning advantages and disadvantages of students in certain subjects. Obviously, the index is a better measure of national performance on different items, that is, the importance of different items to the country.

Here we directly give the calculation formula of the bias index

$$Skewness\ Index = \frac{|x_i - \mu|}{\sigma} \tag{13}$$

Where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of gold, silver, copper and total value of the selected projects in each country， and $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$ is standard deviation.

The Figure 9 below shows the bias index heat map of the top 10 countries in the 2024 Olympic medal table in different sports, reflecting the focus and performance differences of different countries in different events.

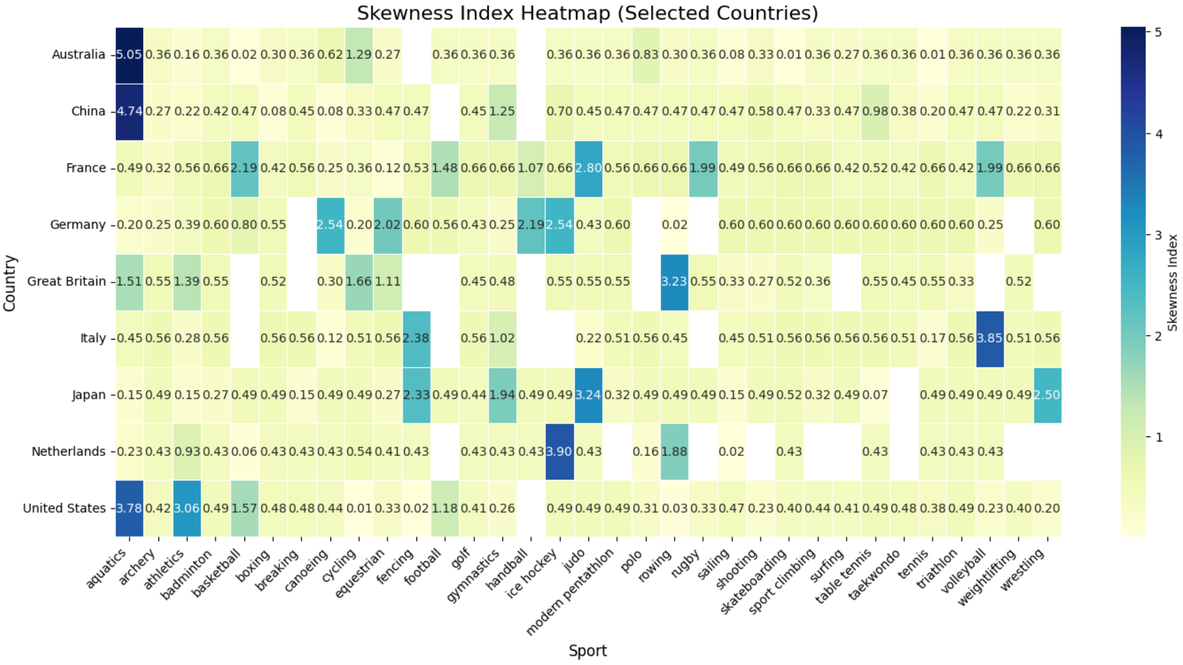


Figure 9: Skewness Index Heatmap

Almost every country has significant bias in some projects, and the bias index shows the traditional advantages of different countries in specific fields. In the heat map, the partial index of some projects is particularly prominent in some countries, while others are relatively balanced.

The United States has the highest bias index in swimming and track, consistent with its traditional dominance in these gold sports, and moderate focus in basketball, while others are more balanced. China also shows significant partial characteristics in diving, gymnastics, table tennis and weightlifting, showing China's focus on technical events. Australia has the highest bias index in swimming events, which is its core advantage in international events. Japan has the highest bias index in judo; Germany has the highest bias index in equestrian and rowing, the UK in cycling, and France has the highest bias index in fencing.。

The bias index provides an intuitive and quantitative tool to analyze the performance of countries in different sports. As can be seen from the partial index heat map, almost every country has its own core projects focused on development. Multiple gold events are crucial to the ranking of the medal table, and the cultural tradition and strategic layout have a profound impact on the phenomenon of biased subjects. By understanding the bias index, countries can better optimize the allocation of resources and achieve further breakthroughs in the future Olympic competition.

We used 75% quantiles as minor and 95% quantiles as significant bias threshold， all biased indices were merged into a single sequence and their quantiles were calculated.

$$\begin{aligned} threshord_1 &= Q3\,(75th\ Persentile) \\ threshord_2 &= Q9\,(90th\ Persentile) \end{aligned} \tag{14}$$

Determine whether the bias index of each item is greater than the set threshold, and if so, the project is considered more important for the country.

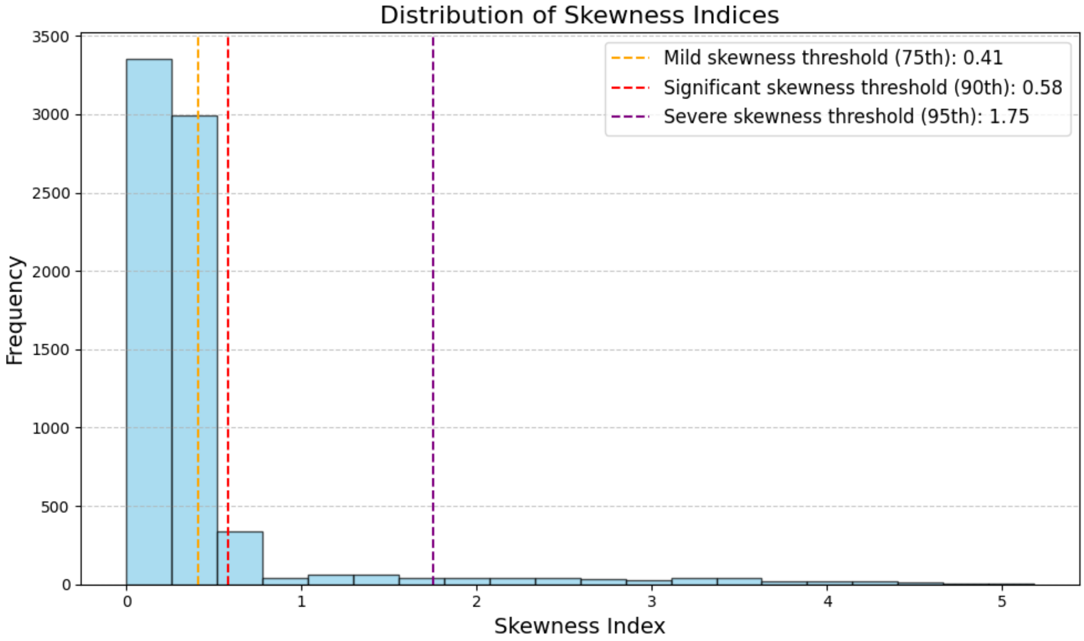


Figure 10:Distribution of Skewness Indices

Q3=0.41, Q9=1.75. If the bias index of all projects in a country is less than Q3, we believe that the country is not biased, that is, the change of the competition project has little impact on the project; if a country has a bias index greater than Q9, it is considered that the country is highly dependent on the project, it may have a great impact on the number of MEDALS. According to statistics, the proportion of balanced countries in the total countries:

At the same time, we counted which items caused the bias. The statistical graph is shown in Figure 11:

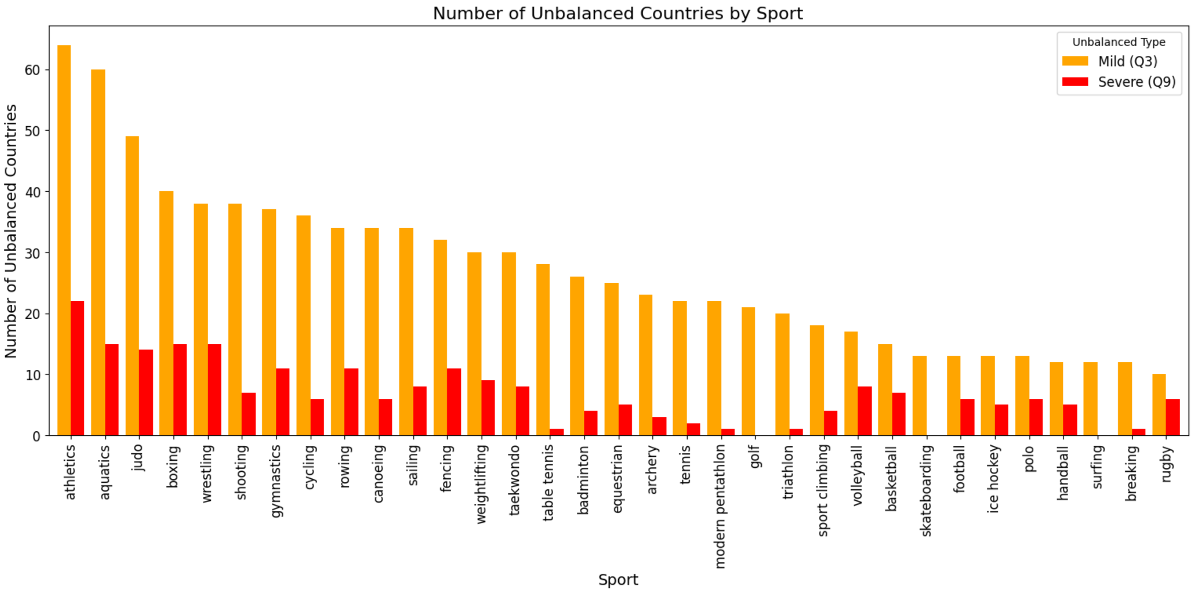


Figure 11:Number of Unbalanced Countries by Sport

Figure 11 shows the number of countries leading to bias in each project, with the number of countries with mild and significant bias families as the classification criteria. Track and field, swimming and other gold events are the main source of partial phenomenon, and the adjustment of these events has a great impact on the national medal table. The dependence of some countries on judo, shooting, boxing and other programs shows that the adjustment of these programs is equally important for their ranking. Some niche events have a good balance, and their adjustment has a relatively little impact on the overall medal distribution.

The “great coach” effect

Feature analysis

The impact of a great coach typically manifests in specific events within a country and may be observed over several Olympic Games within a particular timeframe. For instance, when a nation hires a new coach for a specific sport, there is often a noticeable improvement in the country's performance in that event compared to previous years. To identify the years, countries, and events that may exhibit significant coach influence across a wide range of annual medal charts, we categorize the data by national events and analyze the performance of each country in those events over the years (i.e., the number of medals).

An example in this context is Lang Ping, who coached volleyball teams from both the U.S. and China to championships. Analyzing the data, we observed that during her coaching tenures, the countries saw a notable increase in the number of medals. Therefore, we believe we can calculate the changes in medal counts for the countries involved and identify significant shifts, assuming that the presence of a great coach contributed to these changes.

The relationship between the medal growth and the great coach effect

Considering that different MEDALS represent different degrees of excellence, in order to facilitate the subsequent calculation and make the calculation results more reference value, we give different weights to different MEDALS in the table 7:

Table 8:Medal weight table

medal	Gold	Silver	Bronze	No medal
Weight	3	2	1	0

We can get the medal number formula as follows:

$$N_{ij}^{year} = N_{Gold,ij}^{year} \cdot W_{Gold} + N_{Sliver,ij}^{year} \cdot W_{Sliver} + N_{Bronze,ij}^{year} \cdot W_{Bronze} \tag{15}$$

Then a certain year in the number of MEDALS in a country in a certain event can be expressed as:

$$\varDelta N_{ij}^{year} = N_{ij}^{year,t} - N_{ij}^{year,t-1} \tag{16}$$

After conducting statistical analysis, we discovered that events displaying significant fluctuations in medal counts are predominantly team events. In these contexts, the influence of elite individual athletes tends to be relatively minimal regarding the overall team performance.

Additionally, during the same timeframe, we can reasonably assume that countries with similar levels of national development experience little variation in their advancements in sports technology. Therefore, we can assess the presence of a great coach in a given event by comparing the medal counts and growth rates of different teams participating in the same event.

This comparative approach enables us to evaluate the impact of coaching on team success and identify any notable coaching influences.

Based on the aforementioned methods, we have identified several countries that may be influenced by the "Great Coach Effect." The details of these findings are presented in Table 9.

Table 9:Teams and their medal data that may be influenced by great coaches

Team	Event	Sport	Year	Medal numeric	Medal increment
Argentina	Hockey Men's Hockey	Hockey	2016	54	54
Australia	Baseball Men's Baseball	Baseball	2004	48	48
Germany	Hockey Women's Hockey	Hockey	2004	48	48
Brazil	Volleyball Women's Volleyball	Volleyball	2008	36	36
China	Volleyball Women's Volleyball	Volleyball	2004	36	36

We have gathered data on the coaching circumstances of various national teams. For instance, Carlos Retegui has been the head coach of the Argentine men's field hockey team since 2016. Notably, the team experienced significant improvement in 2016, which can be attributed to the introduction of new coaching strategies and tactical adjustments.

We believe that this enhancement in performance was influenced by the "Great Coach Effect," aligning with our initial predictions regarding the impact of effective coaching on team success.

Characteristic importance analysis

The calculation principle of the feature importance of the XGBoost model is based on the construction process of the decision tree. In the calculation, the XGBoost will be evalu-ated based on the contribution of each feature in the model in the decision tree node.XGBoost.

Definition of feature importance

To measure the importance of the features, we give the following three valid indica-tors:

Feature importance based on split gain (Gain)

Split gain is the information gain brought when a feature is used to split when building a tree. Specifically, it is the degree to which the model's loss function (such as mean square error or log loss) decreases when using a certain feature.

Feature importance based on split number (Weight)

The number of divisions is the frequency at which a feature appears as a split node in all decision trees. Weight Reflects the frequency of features being selected as nodes dur-ing the splitting of all trees.

Coverage-based feature importance (Cover)

Coverage represents the amount of data divided by a feature during tree splitting. It measures the number of samples involved in the feature in the training set.

The calculation of importance

XGBoost In the training process, the tree structure of the model will be updated iteratively according to the objective function (such as mean square error, logarithmic loss, etc.), and each tree will reduce the loss by the optimized splitting features. For each feature, XGBoost calculates its contribution to all trees based on the different criteria (Gain, Weight, Cover).

We selected the more prominent features of these three assessment criteria (Gain, Weight, Cover), as shown in Table 10:X_1

Table 10:High importance characteristics

Feature	gain	weight	cover
X_1	1.087	2284	26.348
X_2	1.487	1305	20.077
X_7	1.374	547	16.843
X_3	2.913	337	1264.192
X_4	1.749	371	93.035
X_6	15.981	81	6586.331
X_5	2.633	311	1255.716

The model reveals that the status of being the host country plays a significant role in predicting medal counts, highlighting the substantial influence of home advantage. Home advantage is not only reflected in familiarity with the competition environment but also includes potential referee bias and strong policy support from the host nation. Athletes from the host country often receive more resources and opportunities, enabling them to perform better in competitions. National Olympic Committees (NOCs) can aim to secure host country status to capitalize on these advantages. Meanwhile, for non-host countries, it is essential to adapt to the competition environment in advance and implement psychological preparation strategies to minimize the disadvantages of competing away from home.

The number of times an athlete has participated in the Olympic Games has a significant impact on medal predictions, indicating that experienced athletes tend to perform more consistently. This may be due to their stronger psychological adaptability in high-pressure environments and better command of competition rhythm and tactics. NOCs should focus on the long-term development of young athletes by increasing their participation in international competitions to help them adapt to high-level events early. Additionally, experienced veterans should be provided with adequate support, and their competition frequency should be strategically planned to extend their athletic careers.

The number of medals a country won in specific events during the previous Olympics has a strong influence on the current predictions, reflecting the continuity of a country's dominance in certain traditional strengths. This highlights the importance of a nation's sports infrastructure, long-term investment, and technical expertise in achieving Olympic success. NOCs should continue to invest in traditional strong events to maintain competitiveness while introducing international coaches and leveraging advanced technology to explore breakthroughs in other potential events, thereby increasing their overall medal count.

The model shows that the influence of different sports on medal counts varies significantly. For instance, events like athletics and swimming, which have a large number of medal opportunities and intense competition, have a higher impact on predictions. In contrast, events like sailing or shooting, though less influential globally, may hold strategic importance for certain countries. NOCs can focus on developing specific sports based on their resources and competitive advantages. For example, they can prioritize medal-rich events like athletics or swimming while seeking breakthroughs in less popular but less competitive events.

These insights enable NOCs to make more informed decisions regarding resource allocation, athlete development, and strategic planning, ultimately helping them achieve greater success in future Olympic Games. These findings provide robust data support and theoretical guidance for the formulation of long-term development plans.

Model Evaluation and Further Discussion

Strengths

High prediction accuracy: XGBoost The decision tree is gradually constructed through the gradient lifting method, and the model is adjusted according to the error of the previous round in each iteration, so as to optimize the loss function, with high prediction accuracy.

Processing complex data: XGBoost Can handle large data sets and handle missing data and categorical variables.

Fast training speed: Using parallel computing and optimization strategies, XGBoost has a significant advantage in training speed, and can quickly process the survey data in this problem.

Prevent overfitting: L1 and L2 regularization are used to control the model complexity, and find the best balance.

Probability prediction: By calculating the probability of each athlete winning a medal, and then summarizing it to the national level, it can reflect the competitiveness of different countries in different events in more detail.

Quantifying project importance: By calculating the project knewness index, it can quantify the importance of different projects to the distribution of national MEDALS, and provide reference for the National Olympic Committee in project investment and resource allocation.

Weaknesses

Difficult to deal with emergencies: The model assumes that athletes' fitness and roster remain unchanged between the two Olympics, but in reality there may be emergencies such as injuries and retirement, which are difficult to predict.

It is difficult to consider the interaction between events: the model mainly focuses on the impact of individual events on national medals, and it is difficult to consider the interactions and synergies between different events.

High data requirements: a large amount of historical data is needed to accurately identify the impact of good coaches, which requires high data integrity and accurace.

Conclusion

We tried to solve the Olympic medal prediction problem by constructing a prediction model based on XGBoost. The advantage of the model is its high prediction accuracy and fast training speed, which can effectively handle large-scale datasets. We identified events that may have had the influence of good coaches by comparing the number of medals in different countries in the same sport. This analysis provides valuable insights for countries in coaching investments. Through feature importance analysis, we found that factors such as year, gender, country and sports items had a significant influence on the prediction results of the model. These findings provide an important reference for the national Olympic Committees in terms of project investment and resource allocation.

Future studies could further improve the model to better address the limitations. More real-time data and dynamic factors can be introduced to more accurately reflect the status of the athletes and the changing competition environment. Furthermore, more accurate methods to analyze interactions between different items can be explored.

In conclusion, this paper provides an effective tool for predicting Olympic medal distribution by developing a XGBoost based prediction model and a valuable reference for national Cs in resource allocation and coaching investment.

References

Chen, T. , & Guestrin, C. . (2016). Xgboost: a scalable tree boosting system. *ACM*.

Cook, G. M., Fletcher, D., & Peyrebrune, M. (2021). Olympic coaching excellence: A quantitative study of psychological aspects of Olympic swimming coaches. *Psychology of Sport and Exercise*, 53, 101876.

Forrest, D., Sanz, I., & Tena, J. D. (2010). Forecasting national team medal totals at the Summer Olympic Games. *International Journal of Forecasting*, 26(3), 576-588.

Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. *Technological Forecasting and Social Change*, 175, 121314.