

2028 LA Olympics: Who Will Dominate the Medal Count?

Summary

As the Olympic Games' importance grows, the focus on medal standings across countries has surged, making Olympic medal prediction an exciting area of research. This paper aims to develop a **comprehensive** model to **forecast** the medal counts (including predicted values and confidence intervals) for each country at the 2028 Olympics. We integrate linear regression, **ARIMA** (AutoRegressive Integrated Moving Average), and **random forest models** to predict both total and gold medal counts, providing a **90% confidence interval**. The **ARIMA** model captures trends and cycles, while the random forest model handles **complex nonlinear relationships**.

In Task 1, we **categorized** countries into three groups: host country, the U.S., medal-winning countries, and countries without medals. We used historical data from previous Olympics, linear regression, and average values to predict medal outcomes for both returning and new events. For new events, we used **Monte Carlo simulations** to estimate medal distributions. We forecasted the host country's medals based on patterns from past hosts, and for countries without medals, we applied random forest and ARIMA (with **parameters** $p=0, q=2, d=0$) to estimate gold medals. By combining **linear regression** with event participation data, we calculated expected medal counts and used a **Poisson distribution** to derive 90% confidence intervals. Our results showed that the U.S. is likely to see at least a 95% probability of a medal increase, while China, losing some dominant events, is predicted to regress.

In Task 2, we used **Fisher's exact test** to analyze the impact of "great coaches," exemplified by Lang Ping and Béla Károlyi, on medal outcomes. We recommended "great coaches" for Chinas women's volleyball team, Indias athletics, and Brazils gymnastics, providing the rationale for these suggestions.

In Task 3, we examined medal distribution across sports and advised smaller countries on improving their chances, considering factors like **GDP**. We found a **strong positive correlation** between a countrys economic strength and its medal count.

Overall, our model provides reliable predictions of medal counts and confidence intervals for the 2028 Olympics. While uncertainties remain for countries with fewer medals, the predictions for top medal winners are robust.

Keywords: ARIMA, Random Forest, Linear Regression, Olympic Medal Prediction, Great Coach Effect, Poisson Distribution.

Contents

1	Introduction	2
1.1	Background	2
1.2	Problem Restatement	2
2	Notations & Assumptions	4
3	Data Processing	5
3.1	Clustering Analysis	5
3.2	Country Classification	5
3.3	Event Addition and Removal	6
4	Task 1: Olympic Medal Prediction and Analysis	6
4.1	2028 Medal Table Prediction	6
4.1.1	Countries with Previous Medals (Non-Host Countries)	6
4.1.2	Host Country: United States	8
4.1.3	Prediction for Countries Winning Medals for the First Time	10
4.2	Analysis of the Relationship Between Events and Medals	16
5	Task 2: Great Coaches	17
5.1	Evidence of the Great Coach Effect	17
5.2	Selection of Three Countries and Specific Sports	19
6	Task 3: Recommendations	20
7	Model Evaluation	21
7.1	Sensitivity Analysis	21
8	Model Evaluation	22
9	Reference	24



1 Introduction

1.1 Background

The Olympic Games are a global spectacle where nations vie for medals. While **giants** like the USA and China lead the medal table, **smaller countries** such as Albania and Cape Verde have made history, with both winning their first gold medals at the 2024 Paris Olympics. This highlights not just their growing sports prowess, but their rising prominence on the world stage.

Predicting future Olympic medal outcomes is complex. Current models often focus solely on athlete performance and past results, missing key factors like **national investments, event changes, and shifting competition trends**. A more comprehensive model could greatly enhance predictions, particularly for the 2028 Olympics.

1.2 Problem Restatement

We are provided with historical data on past Olympics, including host countries, events, and detailed athlete performance, categorized by sport and medal type. The challenge is to build a model that can predict medal outcomes for the 2028 Olympics and provide insights into the underlying factors.

Task 1: Build a model to predict the total and gold medal counts for each country in 2028, with 90% confidence intervals.

- Compare the predictions with actual data from the 2024 Olympics to identify countries likely to see increases or decreases in their medal counts.
- Estimate the probability of each non-medal countries that will win their first medals in 2028.
- Explore the relationship between sports types, event numbers, and medal distributions across nations.
- Analyze the potential impact of new sports introduced by the host country on the overall medal standings.

Task 2: "The Great Coach Effect"

- Investigate the impact of great coaches on the medal performance of a nation, by examining changes before and after a coachs appointment.
- Quantify the effect of a coach's influence on medal counts.
- Identify three countries where hiring a coach could make a significant difference in their performance.

Task 3: Insights from the Model

- Explore the relationships between medal distributions, country-specific factors (e.g. sports events, economic development), and national strengths.
- Help smaller countries optimize their medal chances despite resource limitations.
- Explore the correlation between per capita medals and per capita GDP.

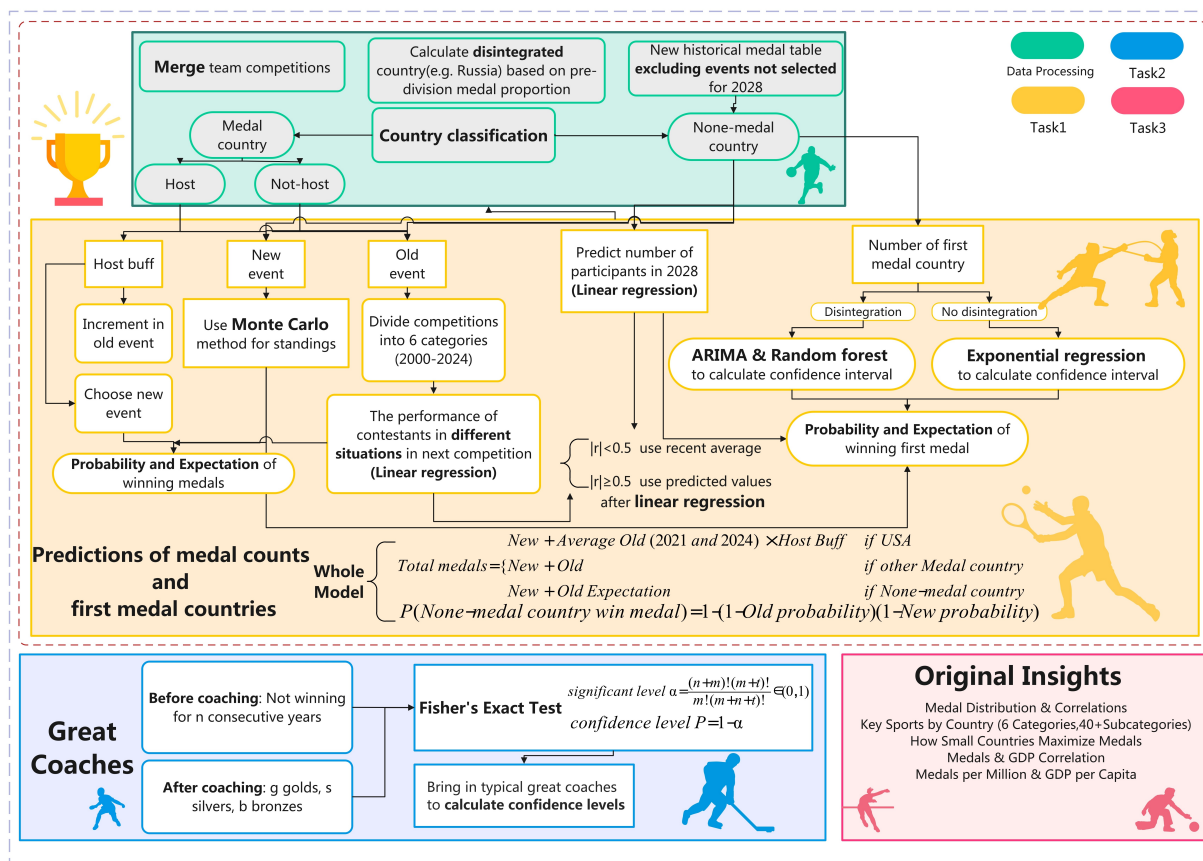


Figure 1: Our work



关注数学模型
获取更多资讯

2 Notations & Assumptions

Table 1: Notations

Symbols	Description
Class 1 ~3	1:Aquatic Sports 2:Ball Sports 3:Track & Field
Class 4 ~6	4:Gymnastics & Acrobatics 5:Martial Arts 6:Motorsport & Racing
$a_{jk}^i/b_{jk}^i/c_{jk}^i$	for one country, if Class i year $1996 + 4k$ won gold/silver/bronze, number of medals in year $2000 + 4k$; $j = 1, 2, 3$ refers gold,silver and bronze
d_{jk}^i/e_{jk}^i	for one country, if Class i year $1996 + 4k$ won no medal/didn't participate, number of medals in year $2000 + 4k$; $j = 1, 2, 3$ refers gold,silver and bronze
f_i	for one country, the expectation of all medals in all new events year 2028
$a_j^i/b_j^i/c_j^i$	for one country, Class i year 2024 won gold/silver/bronze, the predicted value of medals in year 2028; $j = 1, 2, 3$ refers gold,silver and bronze
d_j^i/e_j^i	for one country, Class i year 2024 won no medal/didn't participate, the predicted value of medals in year 2028; $j = 1, 2, 3$ refers gold,silver and bronze
G/S/B	for one country, the expectation of gold/silver/bronze in all events year 2028
T	for one country, the expectation of total medals year 2028
$\bar{G} / \bar{S} / \bar{B}$	for one country, the predicted value of gold/silver/bronze in all events year 2028
\bar{T}	for one country, the predicted value of total medals year 2028
$[G_{min}, G_{max}]$	for one country, the predicted value interval of gold in all events year 2028
$[T_{min}, T_{max}]$	for one country, the predicted value interval of total medals year 2028
g_i	the total medals of the ith host in ith Oly.(without new events)
m_i/n_i	the total medals of the ith host in (i-1)th/(i+1)th Oly. (consider special cases)
T_{boost}	the weighed increment of the host
$T_{old}/T_{new}/T$	the total medals (old/new/all events) of USA year 2028
$u_1 \sim u_{t-2}$	first medal country(no disintegration) in year 1904 ~2024
$v_1 \sim v_9$	first medal country(disintegration) in year 1992 ~2024
a_{ij}	the number of involved events of ith none-medal country year $1996 + 4j$
a_i	the predicted number of involved events of ith none-medal country year 2028
L_i	the expectation of medals of ith none-medal country year 2028
q_i	the probability of winning medals of ith none-medal country year 2028
x_{ij}/u_{ij}	the percentage/total number of the jth country Class i medals in year 1896 ~2024
r_i	the total number of Class i medals in year 1896 ~2024
d_{ij}	the difference of Class i and Class j medals
d'_{ij}	the correlation of Class i and Class j
x'_{ij}/u'_{ij}	the percentage/total number of the jth event i medals in year 1896 ~2024
r_i	the total number of event i medals in year 1896 ~2024
t	medals after changing great coaches

Considering the complexity of factors involved in the upcoming 2028 Los Angeles Olympics, we make reasonable assumptions to simplify the model. Each assumption is followed by its corresponding explanation:

- **Assumption 1:** All countries will be **unaffected** by factors like politics or doping and will participate with complete teams.
 - **Explanation:** This ensures that external factors do not impact the analysis of sports performance.
- **Assumption 2:** The United States will add five new sports for the 2028 Olympics: Baseball/Softball, Lacrosse, Cricket, Squash, and Flag Football, totaling ten events. (Men and Women) **And all are conducted as knockout tournaments.**
 - **Explanation:** This is based on IOC's preliminary discussions on adding new sports for the upcoming Olympics.
- **Assumption 3:** Four sports (Breakdancing, Boxing, Weightlifting, and Modern Pentathlon) will be removed from the Olympics.
 - **Explanation:** These sports are likely to be excluded based on IOC's review and changing interests.
- **Assumption 4:** The athletes' nationality will remain unchanged.
 - **Explanation:** This assumption maintains a consistent basis for comparing athlete performance by country.

3 Data Processing

3.1 Clustering Analysis

A country's Olympic performance depends on its proficiency in various sports, but for 70+ events, it's not practical to analyze all. To predict its medal count, we group events into broader categories due to the large number of Olympic events.

We assume each country excels in one or more major sports categories. Clustering analysis helps identify these strengths and groups similar events together.

We construct a matrix with countries as rows and events as columns, where the values represent medal-winning rates. **K-means clustering** uses these rates to group similar events. The optimal number of clusters, $K = 6$, is determined using the elbow method and literature insights. Misclassified points are **manually corrected** for accuracy.

The clustering visualization is shown below:

3.2 Country Classification

Countries are divided into **none-medal** and **medal-winning**. For **none-medal** countries, we predict the chance of them winning their first medal. Host countries are analyzed separately to account for their performance boost. We refine the model to handle missing data from countries that skipped certain Games.



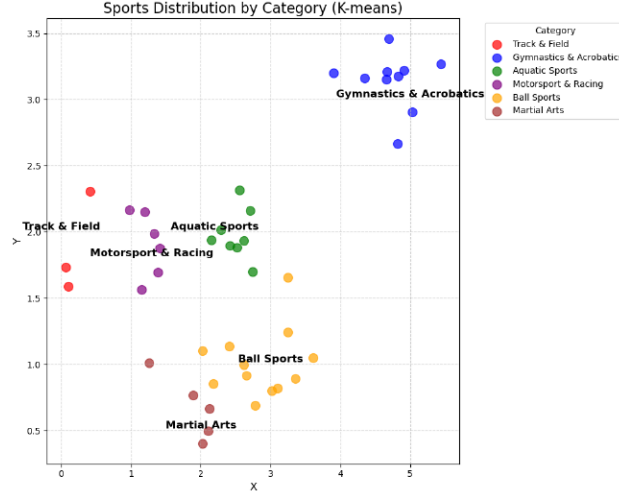


Figure 2: Sports Distribution Classification (K-means)

3.3 Event Addition and Removal

New Olympic events lack historical data, so we use current country rankings for prediction. Discontinued events are excluded from the dataset to prevent them from affecting future predictions.

4 Task 1: Olympic Medal Prediction and Analysis

4.1 2028 Medal Table Prediction

4.1.1 Countries with Previous Medals (Non-Host Countries)

Old Events We classify all events into six major categories with K-means method as mentioned before. Then, we analyze the data of athletes from each country within these categories.

Given that athletes often retain their medals in consecutive Olympic Games, our strategy is to use the results from the previous Olympic Games to predict the performances of athletes in the upcoming Games. We define the following variables and indices:

Let i represent the above categories of sports: $f_1 : \mathbb{R} \rightarrow \text{item}$ (where Water Sports, Ball Sports, Track and Field, Technical Events, Combat Sports, and Racing Events correspond to numbers 1 to 6).

Let j represent the outcomes in the next Olympic Games: $f_2 : 1 \rightarrow \text{gold}, 2 \rightarrow \text{silver}, 3 \rightarrow \text{bronze}$.

Let k represent the previous Olympic year: $f_3 : \mathbb{R} \rightarrow \mathbb{R}, f_3(k) = 4k + 1996$.

Symbols a , b , c , d , and e represent the medal types or absence thereof for the athletes in the previous Games: - a represents gold medals, - b represents silver medals, - c represents bronze medals, - d represents non-medal athletes, - e represents athletes who did not

participate in the previous Games.

We define a_{jk}^i as the total number of athletes who won gold in the previous Games in category i and who win medal j (gold, silver, or bronze) in the following Games. Similarly, we define b_{jk}^i , c_{jk}^i , and d_{jk}^i for silver, bronze, and non-medal outcomes, respectively.

To explore whether the number of medals in consecutive Olympic Games is related to the year, we propose different prediction algorithms. If the medal count shows a strong correlation with the year, we use linear regression to predict. If the correlation is weak, we use the average of the data as the prediction.

We use **0.5** as the threshold for determining strong or weak correlation (other thresholds will be considered in subsequent sensitivity analyses). The predicted number of medals is defined as:

$$a_j^i = \begin{cases} \frac{\sum_{k=1}^6 a_{jk}^i}{6} & , |r| < 0.5 \\ \max\{0, \hat{a}_j\} & , |r| \geq 0.5 \end{cases}$$

where a_j^i is the predicted number of gold medals in category i for the 2024 Games, and r is the correlation coefficient for a_{jk}^i , where $k = 1, 2, \dots, 6$. The linear regression value is:

$$\hat{a}_j = \frac{-10a_{j1}^i - 4a_{j2}^i + 2a_{j3}^i + 8a_{j4}^i + 14a_{j5}^i + 20a_{j6}^i}{30}$$

Since this value may be negative, we **take the maximum between the linear regression value and 0**.

Similarly, we apply the same methodology for b_{jk}^i , c_{jk}^i , and d_{jk}^i , as well as their total predicted values b_j^i , c_j^i , and d_j^i .

For athletes who did not participate in the previous Olympic Games but are expected to win medals in the upcoming Games, we define e_{jk}^i as the total number of athletes in category i who did not participate in the previous Games but win medal j (where $0 \leq j \leq 3$). The prediction for the number of such athletes is defined as:

$$e_j^i = \begin{cases} \frac{\sum_{k=1}^6 e_{jk}^i}{6} & , |r| < 0.5 \\ \max\{0, \hat{e}_j\} & , |r| \geq 0.5 \end{cases}$$

where r is the correlation coefficient for e_{jk}^i , and the linear regression value \hat{e}_j is:

$$\hat{e}_j = \frac{-10e_{j1}^i - 4e_{j2}^i + 2e_{j3}^i + 8e_{j4}^i + 14e_{j5}^i + 20e_{j6}^i}{30}$$

New Events For new events, we use **Monte Carlo** simulations. By gathering authoritative data on rankings and points for each country in these events, we compute the winning probability for each country:

$$P_{AB} = \frac{w_A}{w_A + w_B}$$

where w_A and w_B represent the points (weights) of countries A and B, respectively. We simulate the events according to the Olympic competition rules, where the top 16 athletes



are selected and paired for a **knockout tournament**. After two rounds, the top 4 athletes compete for the gold, silver, and bronze medals, as shown in the figure below:

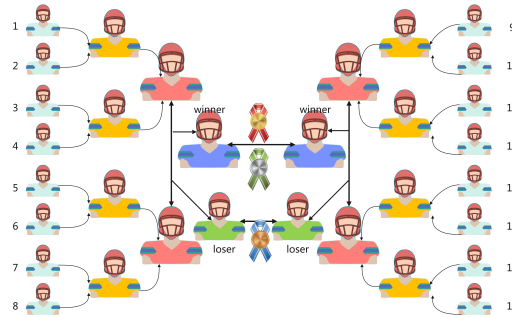


Figure 3: Monte Carlo Method for Simulating New Events

These 5 new events are split into male and female categories, making a total of 10 events. To obtain more accurate probabilities, we simulate each event 1 million times. The results are then used to estimate each country's expected medals for these new events, denoted as f_1 , f_2 , and f_3 .

For some events, like the Lacrosse tournament, where official data is sparse and only world rankings are available, we use **logistic regression** to assess the performance gaps between athletes:

$$P_{AB} = \frac{1}{1 + \exp(\alpha \cdot (R_A - R_B))}$$

where R_A and R_B are the rankings of countries A and B. For events like Squash, where top 20 player rankings and points are available, we calculate the probabilities of each athlete winning gold, silver, and bronze, and estimate the probability of each country earning medals.

The total medal count for a country is the sum of medals from both traditional and new events: Gold: $G = \sum_{i=1}^6 (a_1^i + b_1^i + c_1^i + d_1^i + e_1^i) + f_1$, Silver: $S = \sum_{i=1}^6 (a_2^i + b_2^i + c_2^i + d_2^i + e_2^i) + f_2$, Bronze: $B = \sum_{i=1}^6 (a_3^i + b_3^i + c_3^i + d_3^i + e_3^i) + f_3$.

The final predicted values for gold, silver, and bronze medals are rounded to the nearest integer:

$$\lfloor G + \frac{1}{2} \rfloor, \lfloor S + \frac{1}{2} \rfloor, \lfloor B + \frac{1}{2} \rfloor$$

$$\text{Total medals: } T = \lfloor G + \frac{1}{2} \rfloor + \lfloor S + \frac{1}{2} \rfloor + \lfloor B + \frac{1}{2} \rfloor$$

This gives the predicted total and gold medal count for non-host countries that have won medals in the past.

4.1.2 Host Country: United States

The **host country factor** in the Olympics cannot be ignored. The table below shows the medal counts for new events by host countries from 1984 onwards.

Table 2: Host's New Event Medal

	Year										
	1984	1988	1992	1996	2000	2004	2008	2012	2016	2021	2024
Gold	4	4	3	5	2	0	2	2	1	7	1
Silver	4	1	1	2	3	0	1	0	0	2	4
Bronze	1	1	0	1	2	0	0	1	0	1	2

By analyzing the data from 1896 to 2024, it is clear that the host country's medal count, both total and gold, has generally increased. The reasons for this increase can be attributed to two factors:

1. The host country may add events that are their strengths.
2. The undeniable "home advantage," including familiar venues, food and accommodation, and no time zone adjustment.

Therefore, we predict the United States' medals in the 2028 Los Angeles Olympics separately for old and new events. First, we calculate the increase in medals for traditional events based on the **host country's performance in previous** Olympics. This is done by comparing the number of medals won in the host country's traditional events with the average number of medals won in the two previous Olympics, then calculating the increment and rate of increase.

Old Events For countries with missing data from the previous two Olympic Games, we use the two closest Olympic Games in time (e.g., for the 1984 Los Angeles Olympics, since the U.S. did not participate in the 1980 Moscow Olympics, we use the data from 1976 and 1988 for the U.S.).

Year	Host Medals	Previous & Next Average	Increase Rate	Increment
1896	g_1	$\frac{m_1+n_1}{2}$	$\frac{2g_1}{m_1+n_1}$	$g_1 - \frac{m_1+n_1}{2}$
...
2024	g_t	$\frac{m_t+n_t}{2}$	$\frac{2g_t}{m_t+n_t}$	$g_t - \frac{m_t+n_t}{2}$
2028	to be predicted	115	to be predicted	to be predicted

Table 3: Calculation of Host Factors

Using the data from the table, we calculate the weighted increase rate caused by the host country factors as:

$$\text{Weighted Increase : } T_{\text{boost}} = \left(\frac{g_1 + g_2 + \dots + g_t}{\frac{m_1+m_2}{2} + \frac{m_2+m_3}{2} + \dots + \frac{m_t+n_t}{2}} - 1 \right) \times 100\%$$

Next, we take the average of the medals won by the United States in the 2020 Tokyo Olympics and the 2024 Paris Olympics for traditional events, multiply by the increase rate, and obtain the final prediction for the number of medals in traditional events. The predicted values for the expected gold, silver, bronze, and total medals in the 2028 Los Angeles Olympics are 53.64, 51.43, 31.56, and 136.63, respectively.



New Events Using the **Monte Carlo** simulation data from earlier for new events, if we use the expected values, the United States is predicted to win 0.937 gold, 0.735 silver, 0.706 bronze, and 2.378 total medals in new events at the 2028 Los Angeles Olympics.

By adding the expected values from both old and new events, we predict the United States will win **54.57 gold, 52.16 silver, and 32.37 bronze medals, totaling 139 medals** in the 2028 Olympics.

4.1.3 Prediction for Countries Winning Medals for the First Time

We separate the analysis into two parts: the old events before 2024 and the new events added in 2028. For the old events, we fully utilize the data provided in the task, while for the 5 new events added in 2028, we calculate predictions based on authoritative sources.

For countries that have never won medals before, we consider the number of countries that win their first medal in each Olympic Games for old events: starting from 1984, the data shows 4, 5, 8, 16, 6, 5, 6, 8, 3, 0, and 4. The average value is 5.91, and the standard deviation is 3.85. It is clear that the number of first-time medal winners in 1996 (16) is **significantly higher than twice the standard deviation**, so we need to understand why so many new countries won medals that year.

According to historical records, the Disintegration of the Soviet Union, Yugoslavia, and Czechoslovakia in 1991, 1992, and 1993, along with the reunification of Germany and Yemen in 1990, led to the emergence of new countries. This explains the unusually high number of first-time medal winners. Thus, in the subsequent model, we need to separate countries that emerged from Disintegration from those that were already existing.

Old Events First, we need to predict the number of countries without medals participating in the 2028 Los Angeles Olympics. Similarly, based on the number of participated events of each countries from 2000 to 2024, we select whether to use a specific module which contains the **average value or linear regression by examining the correlation**, and we obtain the predicted participation numbers for 2028.

$$a_j^i = \begin{cases} \frac{\sum_{k=1}^7 a_{ik}}{7} & , |r| < 0.5 \\ \max\{0, \hat{a}_j\} & , |r| \geq 0.5 \end{cases}$$

where a_{ij} represents the number of events country i participated in at the $1996 + 4j$ year Olympics. r is the correlation coefficient for a_{ij} , with $k = 1, 2 \dots 7$; the linear regression value is calculated as:

$$\hat{a}_j = \frac{-8a_{i1} - 5a_{i2} - 2a_{i3} + a_{i4} + 4a_{i5} + 7a_{i6} + 10a_{i7}}{7}$$

Next, we calculate the total number of events country i participated in over the years: $a = \sum_{k=1}^t a_{ik}$.

Considering the Disintegration of some countries between 1992-1996, which made the 1996 data abnormal, we classify the countries into two categories: countries that emerged from Disintegration and countries that already existed.

$$\begin{cases} \text{Non-disintegrated countries winning first medal} & u_1, u_2, \dots, u_{t-2} \\ \text{Disintegrated countries winning first medal} & v_1, v_2, \dots, v_9 \end{cases}$$

For Non-disintegrated countries, based on historical data, we apply a combination of the ARIMA model and Random Forest model, using time series data and feature engineering to generate prediction results and their confidence intervals.

Reason for the model choice: The number of countries winning their first medal in the Olympics is a typical time series data with an underlying trend (the number of participating countries increases, the number of old events increases, etc.), and a four-year cycle. The ARIMA model captures these characteristics well and is used to predict the data for 2028. The Random Forest model, on the other hand, can capture the complex nonlinear relationships between total countries participating, total medals, and the number of countries winning first medals. The ensemble of decision trees helps reduce overfitting, and when combined with the results from the ARIMA model, we can estimate the confidence interval. The ARIMA model captures linear dependencies in time series data, while the Random Forest model captures non-linear dependencies and additional features.

We split the data from 1896-2004 and 2008-2024 into training and testing sets, with each Olympic year as a time point. The target variable is the number of countries winning their first medal. After performing first and second-order differencing on the original target variable, we conduct an ADF test to verify the stationarity of the data. The ADF test results indicate that the original series is the most stationary, with an ADF statistic of ADF Statistic = -4.01 , and a p-value of 0.0014, as shown in the figure. After performing first and second-order differencing, there is no noticeable improvement in stationarity, confirming that the original data is optimal.

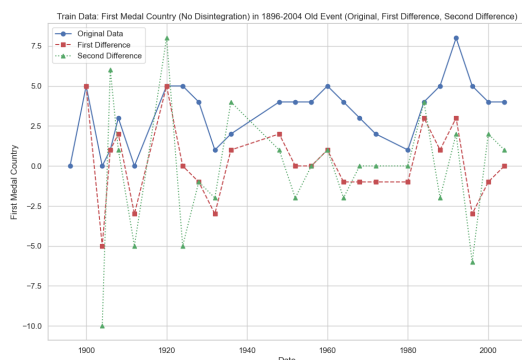


Figure 4: Trend chart of original data, first-order difference, and second-order difference. The original data exhibits the best stationarity.

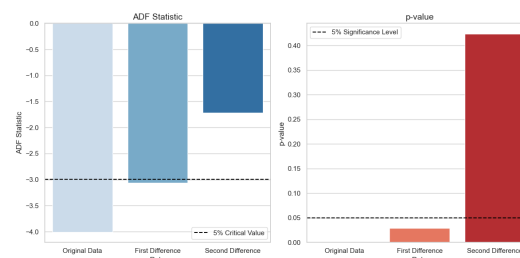


Figure 5: ADF Statistic and p-value. The ADF statistic is -4.01 , and the p-value is 0.0014, verifying the stationarity of the original series.

Next, we plot the **ACF and PACF of the original series**. The tailing patterns in the ACF and PACF indicate long-term dependencies, leading us to select the **ARIMA model for modeling**. To determine the optimal parameters for the ARIMA model, we use the **AIC criterion**. The AIC is calculated as $AIC = 2K - 2\ln(L)$, where K is the number of



关注数学模型
获取更多资讯

model parameters and L is the maximum likelihood value. By plotting the **AIC heatmap**, we find that when $p = 0$ and $q = 2$, the AIC value is minimized, so we select ARIMA(0,0,2) as the optimal model.

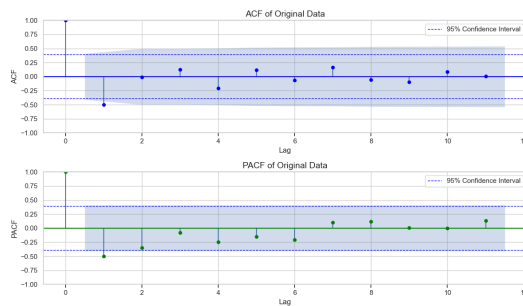


Figure 6: ACF and PACF plots of second-order differences. The plots show the long-term dependencies, supporting the choice of ARIMA modeling.

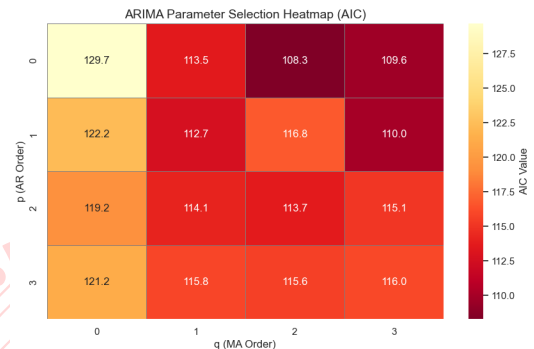


Figure 7: AIC heatmap. The figure shows that the AIC criterion selects the ARIMA(0,0,2) model, where $p = 0$ and $q = 2$ minimize the AIC value.

To validate the **ARIMA model's fitting performance**, we perform residual checks. First, we plot the **ACF and PACF of the residuals** and conduct the Ljung-Box test. The results show no **significant autocorrelation** in the residuals within the selected lag ($p > 0.05$). Additionally, we plot the **histogram and QQ plot** of the residuals and conduct the **Shapiro-Wilk test**, which confirms that the residuals are approximately normally distributed ($p > 0.05$). These results indicate that the **ARIMA model** fits well, with **no significant autocorrelation** in the residuals and a normal distribution, meeting the modeling assumptions.

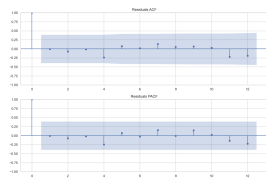


Figure 8: ACF and PACF of residuals. The residuals show no significant autocorrelation, indicating a good fit of the model.

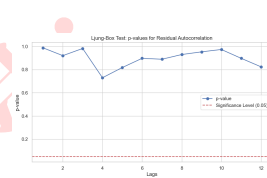


Figure 9: Ljung-Box test results. The Ljung-Box test shows no significant autocorrelation in the residuals ($p > 0.05$).

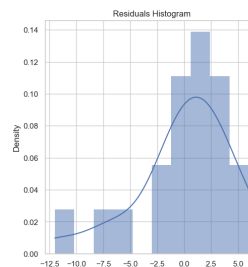


Figure 10: Shapiro-Wilk test results. The test indicates that the residuals are normally distributed ($p > 0.05$).

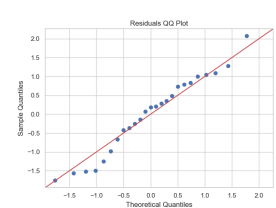


Figure 11: QQ plot of residuals. The residuals align with the fitted line, supporting their normality.

Based on the results from the ARIMA(0,0,2) model, we predict that the number of countries winning their first medal in 2028 will be 3.09. By inputting this prediction along with training set features into the Random Forest model, we obtain a 95% confidence interval

for the number of countries winning their first medal in 2028 of $[2.58, 4.87]$, as shown in the figure.

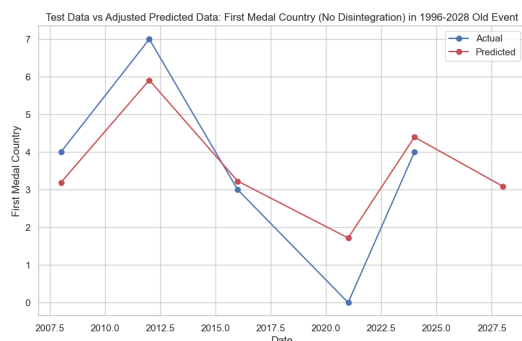


Figure 12: Comparison of actual vs predicted number of first-time medal-winning countries (2008-2028). The prediction aligns well with the actual results, indicating high accuracy of the model.

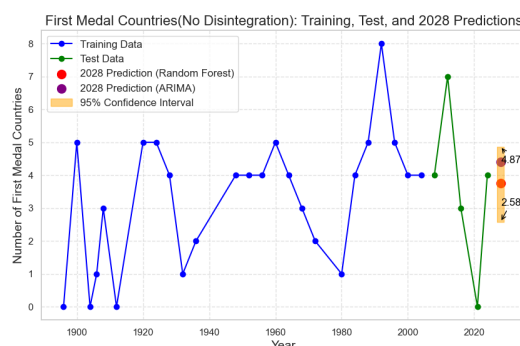


Figure 13: Training, testing, and 2028 prediction results for first-time medal-winning countries. The 95% confidence interval is $[2.58, 4.87]$, indicating stable and reliable predictions.

For countries emerging from Disintegration after 1996, the number of new countries winning their first medal is as follows: 11, 2, 1, 1, 1, 0, 0, 0, forming a **monotonically decreasing sequence approaching 0**. The original data, first-order difference, and second-order difference fail to pass the stationarity test. Considering that 18 of the 21 Disintegrated countries have already won medals, we use an **exponential regression model** to predict a 95% confidence interval of $[0, 0.01]$. Therefore, we conclude that the number of countries from Disintegrated nations winning their first medal in 2028 will be 0, i.e., $u = 0$.

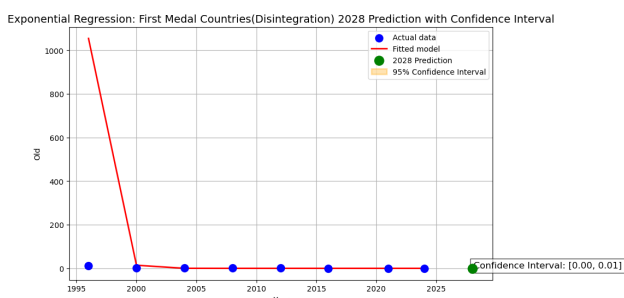


Figure 14: Exponential Regression Model: First Medal Country (Disintegration) 2028 95% Confidence Interval Prediction

Thus, we can calculate the probability of country i (which has not won a medal) winning a medal in the 2028 Los Angeles Olympics traditional events as:

$$L_i = \frac{a_i}{a} \cdot (u + v)$$



Since the number of events is large and the probability of winning in each event is low, the distribution of L_i can be approximated by a **Poisson distribution**: $Poisson(L_i)$. From the Poisson distribution formula, the winning probability q_i is the complement of the probability of zero medals:

$$q_i = 1 - P(x = 0) = 1 - e^{-L_i}$$

New Events Among the 5 new events, the countries with a chance of winning their first medal are Bangladesh and Nicaragua. Based on the previous **Monte Carlo simulation**, the following table is derived:

Country	Item	Gold	Silver	Bronze	Total
Nicaragua	baseball/softball (Men's)	0.058	0.062	0.062	0.182
Bangladesh	cricket (Men's)	0.066	0.067	0.068	0.201
Bangladesh	cricket (Women's)	0.075	0.077	0.077	0.229

Table 4: Medal Distribution by None-Medal Country and New Event

Based on this, we calculate the probability of the two countries winning medals in the new events:

$$\begin{cases} \text{Bangladesh: } P_{\text{Bangladesh}} = 1 - (1 - q_{\text{Bangladesh}})(1 - 0.229)(1 - 0.201) \\ \text{Nicaragua: } P_{\text{Nicaragua}} = 1 - (1 - q_{\text{Nicaragua}})(1 - 0.182) \end{cases}$$

For other countries, $P_{\text{country}} = q_{\text{country}}$.

For countries that have never won a medal, we combine the new and old events to derive the top ten countries most likely to win their first medal in 2024.

Next, we calculate the distribution function of medals based on the expected values of medals. Since for any given country, the distribution of medals in each event is a relatively **small probability binomial distribution**, by summing these binomial distributions, we

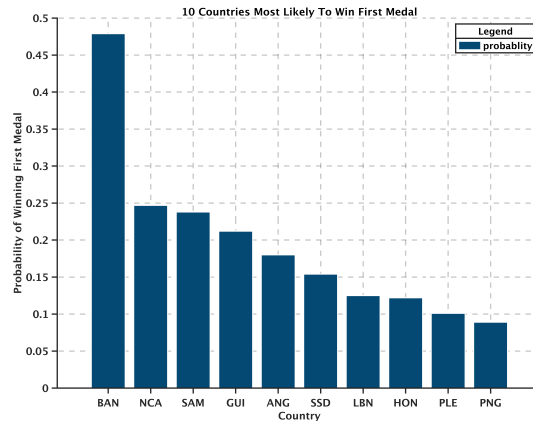


Figure 15: Top 10 Countries Most Likely to Win Their First Medal

approximate the total medal distribution using a Poisson distribution when the number of events is large. Therefore, when calculating the total medal and gold medal counts, we approximate these distributions as **Poisson distributions**, $Poisson(\lambda)$, with λ being T (total medals) and G (gold medals). Then, we simulate the Poisson probability distribution using a **Random Forest model**, calculating the 90% **confidence intervals** (with 90% confidence assumed for now; sensitivity analysis will consider other confidence levels): $[T_{\min}, T_{\max}]$ and $[G_{\min}, G_{\max}]$, which are then used as the **prediction intervals** for the total medals and gold medals for each country in Task 1.

Since Russia did not fully participate in the 2024 Olympic Games, the total medal count and gold medal count for Russia are zero. In the figures below, we can see that among the major Olympic medal-winning countries, the United States is most likely (95% confidence) to improve (in both total medals and gold medals), except for Russia. This is due to the home advantage for the host country, including, but not limited to, familiar venues, cheers from local audiences, accustomed accommodation and food, and no time zone adjustment. There is no significant decrease in the total medal count for the United States; however, China is expected to show the most notable decline in gold medals among these ten countries. This is mainly due to the cancellation of weightlifting, China's most successful event (where it won 5 gold medals at the 2024 Olympic Games).

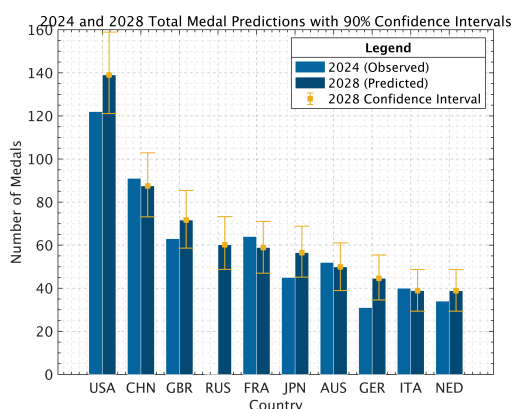


Figure 16: Top 10 Total Medal Predictions

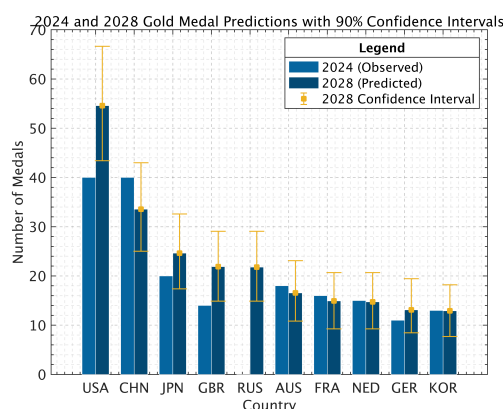


Figure 17: Top 10 Gold Medal Predictions

Analysis of model accuracy and **uncertainty**: For a given country, the uncertainty of the total medal count is calculated as:

$$T_{\text{uncertainty}} = \frac{T_{\max} - T_{\min}}{T}$$

We analyze the uncertainty for both countries with many medals and countries with fewer medals.



关注数学模型
获取更多资讯

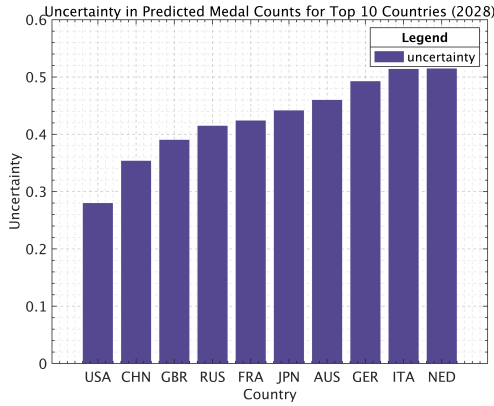


Figure 18: Uncertainty in Predicted Medal Counts for Top 10 Countries

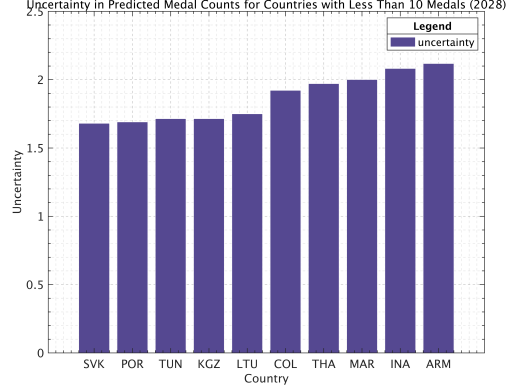


Figure 19: Uncertainty in Predicted Medal Counts for Countries with Less Than 10 Medals

Thus, we conclude that the uncertainty for the countries with more medals is relatively low, while the uncertainty for the countries with less medals is higher. This is because the countries with more medals have smaller fluctuation compared to the number of the medals, while the countries with less medals have greater fluctuation compared to the number of the medals.

4.2 Analysis of the Relationship Between Events and Medals

Correlation Between Six Major Categories of Events and Medals

Categorization Basis: Medal Distribution by Country

Let $i = 1, 2, \dots, 6$, corresponding to the following major categories of sports: Water Sports, Ball Sports, Athletics, Technical Sports, Combat Sports, and Racing Sports; j corresponds to the countries, which include all countries that have participated in the Olympics since 2000. Define x_{ij} as the proportion of medals for country j in the i -th major category from 1896 to 2024, where $x_{ij} = \frac{u_{ij}}{r_i}$, with u_{ij} being the total number of medals won by country j in the i -th category from 1896 to 2024, and r_i being the total number of medals won in the i -th category from 1896 to 2024.

Next, we calculate the difference in medal distribution between the i -th category and the j -th category:

$$d_{ij} = \sum_{k=1}^t (x_{ik} - x_{jk})^2 \quad (1 \leq i, j \leq 6)$$

We then create a 6×6 table. It is evident that in this table, $d_{ij} = d_{ji}$ and $d_{ii} = 0$. The larger the value of d_{ij} , the greater the difference between $\{x_{ik} - x_{jk}\}$, meaning that the correlation between these two categories of sports is weaker.

To normalize this table, we define:

$$d'_{ij} = 1 - \frac{d_{ij}}{\max\{d_{ij}\}}$$

This gives us a new 6×6 table that describes the correlation between every pair of sports categories. We then use this data to generate a **heatmap**. In the **heatmap**, the more **green** the cells are, the stronger the correlation, and the more red they are, the weaker the correlation.

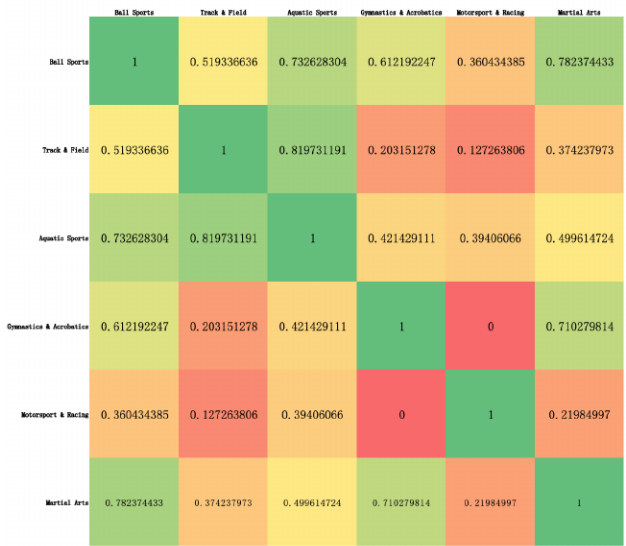


Figure 20: Heat Map

Analysis: The correlation between events is based on medal proportions, normalized to reflect the correlation between each pair. Apart from the self-correlation (which is 1), there is a strong correlation between Ball Sports and Combat Sports, as well as between Water Sports and Athletics. The former is because both Ball Sports and Combat Sports require physical confrontation, and in addition to strength, coordination, flexibility, and reaction time are also indispensable. For the latter, Water Sports and Athletics both fundamentally focus on speed, and some events require explosive power or endurance, which are shared characteristics. In contrast, the lower correlation between Technical Sports and Racing Sports is due to the fact that Technical Sports emphasize flexibility, balance, and coordination in the body to achieve high scores through complex movements, whereas Racing Sports focus on speed, reaction time, and explosiveness, with the goal of completing the race faster than the opponent.

5 Task 2: Great Coaches

5.1 Evidence of the Great Coach Effect

We examine the data for possible evidence of the "great coach" effect: a country has not won a medal in a specific event for several consecutive Olympics, but in the next one or subsequent ones, the number of medals significantly increases! Meanwhile, there was a change in coach in the previous or subsequent Olympics. Thus, we use **Fisher’s Exact Test** to explore whether the medals won are related to the change in coach.

Lets assume a country has participated in n consecutive Olympics for a particular event and has not won a medal, but after changing the coach, it won a medal. During the coach’s



tenure for k Olympics, the country won g gold medals, s silver medals, and b bronze medals, with m Olympics in which the country did not win any medals.

First, construct the **mapping** for gold, silver, and bronze medals:

$$\begin{cases} \text{gold} + \text{silver} + \text{bronze} = 3 \\ \text{gold} : \text{silver} : \text{bronze} = 3 : 2 : 1 \quad (\text{standard mapping of medals to points}) \end{cases}$$

The corresponding "medal points" for gold, silver, and bronze are 1.5, 1, and 0.5, respectively. Thus, the "medal points" for the medals won are given by:

$$t = \frac{3g + 2s + b}{2}$$

We use **Fisher's Exact Test** to calculate the confidence level of whether winning medals is related to the change in coach:

	Before Changing Coach	After Changing Coach
Medal Points	0	$\frac{3g + 2s + b}{2}$
No Medals (Count)	n	m

Table 5: Table Title

From the table, the **significant level** is given by:

$$\alpha = \frac{(n+m)!(m+t)!n!t!}{m!n!t!(m+n+t)!} = \frac{(n+m)!(m+t)!}{m!(m+n+t)!} \in (0, 1)$$

Thus, the **confidence level** is:

$$P = 1 - \alpha$$

We find examples of great coaches, satisfying the following conditions:

1. In the n Olympics before the coach's tenure, the country did not win any medals in the event.
2. After the coach took over, the country won medals, achieving a breakthrough from 0 to 1.

Using the examples from the task, we analyze the data for Lang Ping and Béla Károlyi. The following table shows their coaching records for different national teams, and the resulting **P -values** indicate how confident we are in attributing the success to the "Great Coach."

To measure the contribution of these "**Great Coaches**", we calculate the average medal points per Olympics before and after their coaching, and then take the difference. For example, Lang Ping's contributions to the Chinese and USA volleyball teams are respectively 0.75 medal points (between silver and bronze medal) and 0.5 medal points (equivalent to one bronze medal).

Coach	Country	n	m	g	s	b	t	P
Lang Ping	China	2	1	1	0	0	$\frac{3}{2}$	0.6190
Lang Ping	USA	3	0	0	0	1	$\frac{1}{2}$	0.5429
Béla Károlyi	USA	7	1	2	4	2	8	0.9993
Béla Károlyi	Romania	2	0	1	3	0	$\frac{9}{2}$	0.9441

Table 6: Calculation of P -values for Coaches in Different Coaching Periods

5.2 Selection of Three Countries and Specific Sports

To analyze the impact of great coaches on sports performance, we have selected the following three countries and their respective sports: China (Volleyball), India (Athletics), and Brazil (Gymnastics).

These selections are based on the following reasons: **China's volleyball** team, under the leadership of Lang Ping, once achieved great success, but in recent years, their performance has declined, and the team urgently needs a great coach to revive its glory. **India** has great potential in **sprinting**, but due to a lack of systematic training and scientific guidance, a great coach could significantly improve their performance. **Brazil's gymnastics** has made progress in recent years, but still lags behind top countries (such as the United States and China), and a great coach could help them break through this bottleneck. Therefore, bringing in great coaches for these three sports is seen as a key factor in enhancing their international competitiveness.

A great coach should address the current challenges faced by these countries. For China's volleyball, a great coach can introduce advanced tactical systems, help improve the team's overall competitiveness, especially through tactical innovation, with flexible and varied strategies to cope with intense international competition. For India's sprinting, after the introduction of a great coach, athletes will significantly improve their physical fitness and technical level through advanced training methods, particularly through strength training, technical analysis, and other specialized training to enhance explosiveness and speed. For Brazil's gymnastics, a great coach can help athletes master higher-difficulty routines, improve their technical skills, and increase the execution and stability of their routines through technical analysis and specialized training.

By introducing great coaches, these three countries are expected to make significant improvements in the respective sports. The expected impacts are as follows:

Country	Sport	Current Level	Expected Level	Main Impact Area
China	Volleyball	Top 8	Top 3	Tactical Innovation
India	Athletics	Top 5(Asia)	Top 8	Scientific Training
Brazil	Gymnastics	Top 20	Top 10	Technical Improvement

Table 7: Expected Impact of Hiring Great Coaches



关注数学模型
获取更多资讯

6 Task 3: Recommendations

From an **economic** perspective, a country's GDP is a reflection of its overall economic strength. A higher GDP typically means that more financial resources can be invested in sports infrastructure, athlete training, and scientific research support. For instance, developed countries like the United States and Japan can provide advanced training facilities, scientific nutrition plans, and professional coaching teams for their athletes. These factors directly enhance their competitiveness in the Olympics. Therefore, there should be a **significant positive correlation** between **GDP** and **Olympic medal count**. In fact, based on the GDP data of the top 20 countries in medal counts for the 2024 Olympics, the $\log(\text{GDP})$ has a strong correlation with the total medal count (the **correlation coefficient** r is approximately 0.83), as shown in the figure below:

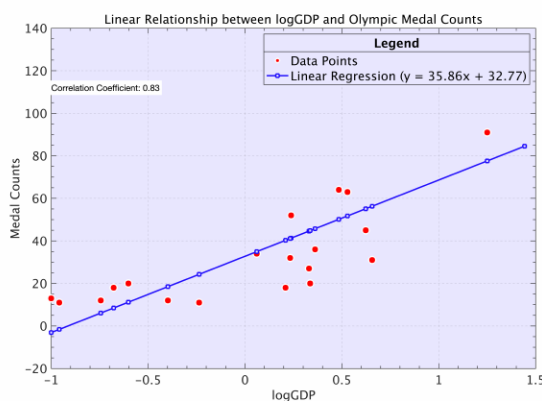


Figure 21: Relationship between GDP and Medal counts

For the i -th sport, the sum of the squares of the medal distribution across countries, $\sum_{k=1}^t x_{ik}^2$, is calculated to assess whether the medals in this sport are evenly distributed across countries. The smaller this value, the more evenly the medals are distributed, meaning that each country has a chance to win medals in this event. The ten events with the smallest values are: **Wrestling, Athletics, Football, Taekwondo, Judo, Shooting, Canoe, Rowing, Gymnastics, and Sailing**. As for countries with fewer medals or countries that have never won medals, they could improve their chances by bringing in a "great coach" to earn medals in these areas. Specifically, wrestling and taekwondo are low-cost, high-return medal investment events. Additionally, coastal countries could invest in Sailing.

7 Model Evaluation

7.1 Sensitivity Analysis

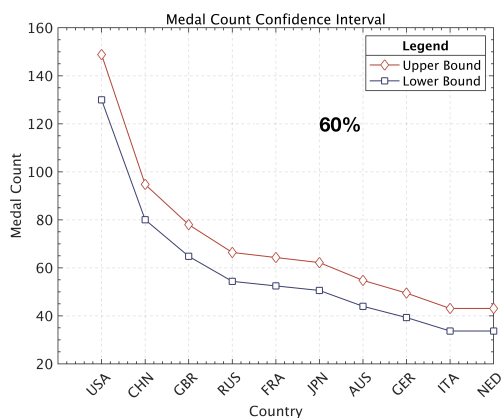


Figure 22: 60% confidence level

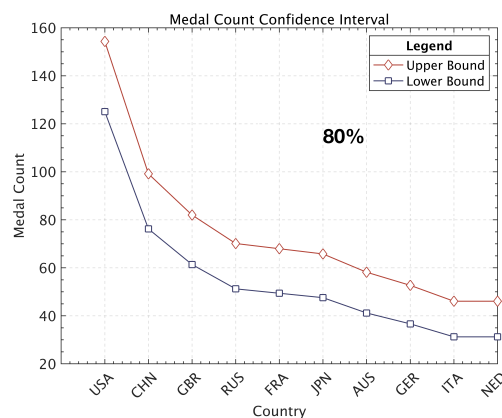


Figure 23: 80% confidence level

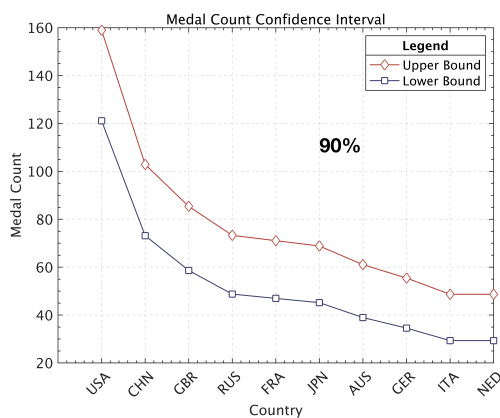


Figure 24: 90% confidence level

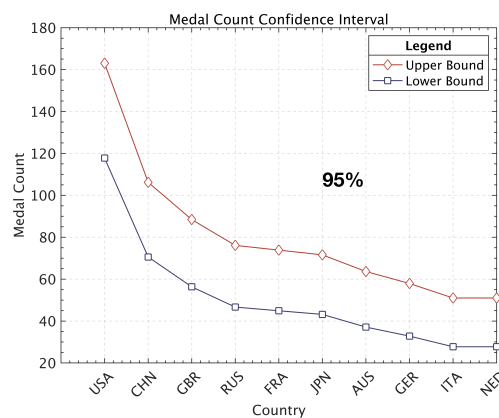


Figure 25: 95% confidence level

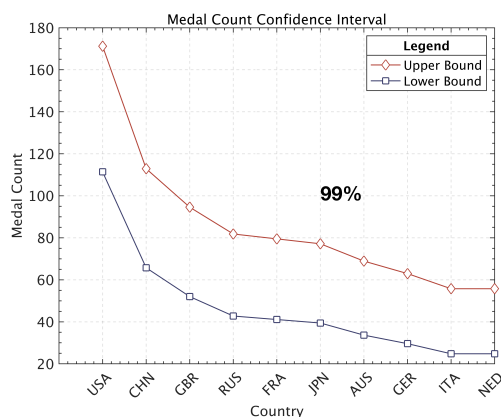


Figure 26: 99% confidence level



关注数学模型
获取更多资讯

Table 8: The total number of medals varies with the correlation coefficient

Country	0.3	0.4	0.5	0.6	0.7
USA	139	139	139	139	139
CHN	96	94	87	83	77
JPN	59	56	55	54	44
GBR	73	73	69	65	57
RUS	60	60	60	60	60
AUS	53	50	48	48	42
FRA	63	61	58	55	43
NED	44	41	38	36	28
GER	47	45	44	44	42
KOR	32	31	29	27	25

Table 9: The number of gold medals varies with the correlation coefficient

Country	0.3	0.4	0.5	0.6	0.7
USA	55	55	55	55	55
CHN	38	38	34	32	31
JPN	25	24	24	24	18
GBR	22	22	21	19	14
RUS	22	22	22	22	22
AUS	18	18	16	16	14
FRA	17	16	15	14	10
NED	17	15	15	15	11
GER	14	13	13	13	13
KOR	15	15	13	11	10

8 Model Evaluation

Sensitivity Analysis

In the prediction model of each country's medal counts and its interval, we initially assumed a confidence interval of 90% and a value of 0.5 for parameter r , which is the threshold for determining strong or weak correlation. Now we will take confidence intervals of 60%, 80%, 95%, and 99% respectively to explore the range of values of medal prediction intervals for different countries in different confidence intervals. Next, the values of the judgment parameter r will be taken as 0.3, 0.4, 0.6, and 0.7, respectively, to explore the changes in the expected total medal count and gold medal count of each country under different judgment parameters.

Gradually observe the medal table prediction interval at each confidence interval, and you will find that as the confidence value increases, the medal prediction interval for each country will also expand accordingly. Although the increase in confidence level can broaden the prediction interval of medals for each country, the confidence level does not affect the medal prediction value. Let r be the parameter of the threshold for determining strong or weak correlation. The medal predictions for the United States and Russia are very stable: when r varies between 0.3-0.7, the total medal and gold medal predictions for the United States in 2028 remain at 139 and 55 respectively, while the total medal and gold medal predictions for Russia in 2028 remain at 60 and 22 respectively, demonstrating strong stability. The trend of predicted Chinese medals: As r increases from 0.3 to 0.7, the predicted total number of Chinese medals in 2028 decreases from 96 to 77, showing a strong decreasing trend. Situation in other countries: Most countries such as Japan, the United Kingdom, Australia, France, the Netherlands, Germany, South Korea, etc. also show a trend of decreasing total medal predictions as r increases.

Advantages and Disadvantages Analysis

Advantages:

Table 10: The total number of medals varies with the correlation coefficient

Country	0.3	0.4	0.5	0.6	0.7
USA	139	139	139	139	139
CHN	96	94	87	83	77
JPN	59	56	55	54	44
GBR	73	73	69	65	57
RUS	60	60	60	60	60
AUS	53	50	48	48	42
FRA	63	61	58	55	43
NED	44	41	38	36	28
GER	47	45	44	44	42
KOR	32	31	29	27	25

Table 11: The number of gold medals varies with the correlation coefficient

Country	0.3	0.4	0.5	0.6	0.7
USA	55	55	55	55	55
CHN	38	38	34	32	31
JPN	25	24	24	24	18
GBR	22	22	21	19	14
RUS	22	22	22	22	22
AUS	18	18	16	16	14
FRA	17	16	15	14	10
NED	17	15	15	15	11
GER	14	13	13	13	13
KOR	15	15	13	11	10

1. Screening of Effective Information: For Task 1, our prediction model is mainly based on the Olympic data from 2000-2024. This is because the data from the 1996 Olympics and earlier are too old, and to avoid potential political factors affecting participation or countries merging/breaking up before 1996.

2. Considering the impact of newly established sports events on the host country and countries excelling in these events, we predict the medal performance in new events by gathering league rankings for these events globally and using Monte Carlo simulations to make accurate predictions.

3. Considering the different peak performance cycles of athletes in various sports categories, we categorize the Olympic events into six major categories: Athletics, Aquatics, Ball Sports, Combat Sports, Gymnastics, and Racing. This categorization allows for more accurate predictions for each country in different types of events.

4. When calculating the medals for a country in a specific sport, we consider the trend of increasing or decreasing medals and determine whether to use linear forecasting or average values based on the correlation with time.

5. Due to the large number of events, we cleverly use Poisson distribution to calculate the medal distribution function for each country and then simulate the 90% confidence interval for each country's medal count using Random Forest.

Disadvantages:

1. Political factors have not been considered, such as whether countries like Russia will participate in the 2028 Los Angeles Olympics.

2. For countries with fewer medals, the Poisson distribution's fitted distribution function may lead to significant uncertainty, making the predictions unstable.

3. Due to the large volume of data required for the six categories of sports and the performances of each category in different Olympic editions, the program can be slow, resulting



in high time costs.

4. The win-rate model used to calculate the event points for new sports does not accurately represent the real win rates.

9 Reference

References

- [1] Singh, R. K., Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales, A. J., Kalita, H., Nanda, C., Sharma, S., Sharma, Y. D., Rabaan, A. A., Rahmani, J., & Kumar, P. (2020). Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model. *JMIR Public Health Surveill*, 6(2), e19115. DOI: <https://doi.org/10.2196/19115>
- [2] Unknown Author. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, 25, 1104–1109. DOI: <https://doi.org/10.1016/j.phpro.2012.03.206>
- [3] World Lacrosse. (2023). World Lacrosse Rankings. Retrieved October 1, 2023, from <https://worldlacrosse.sport/the-game/world-rankings/>
- [4] PSA World Tour. (2023). PSA Squash Rankings. Retrieved October 1, 2023, from <https://www.psasquashtour.com/rankings/>
- [5] International Federation of American Football. (2023). IFAF World Rankings. Retrieved October 1, 2023, from <https://www.americanfootball.sport/ifaf-world-rankings/>
- [6] United Nations. (2023). United Nations Data. Retrieved October 1, 2023, from <https://data.un.org/>
- [7] International Cricket Council. (2023). ICC Cricket Rankings. Retrieved October 1, 2023, from <https://www.icc-cricket.com/rankings#mens-team-rankings>
- [8] World Baseball Softball Confederation. (2023). WBSC Rankings. Retrieved October 1, 2023, from <https://www.wbssc.org/en/rankings>

Report on Use of AI

AI Tools Usage Report

- OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4o):

- **Query1:** <Please explain the basic principles of Fisher's Exact Test in detail.>
- **Output:** <Fisher's Exact Test is a statistical test used to determine if there are nonrandom associations between two categorical variables. It is especially useful when sample sizes are small, and the chi-squared test may not be valid due to low expected frequencies in contingency tables. The basic principle of Fisher's Exact Test revolves around the hypergeometric distribution. It is based on the idea of calculating the probability of obtaining the observed data (or something more extreme) under the assumption that the two variables are independent.>

- **Query2:**<Please provide LaTeX code to generate a three-line table.>

- **Output:**

```
\begin{table}[ht]
\centering
\begin{tabular}{|c|c|c|}
\toprule
Header 1 & Header 2 & Header 3 \\
\midrule
Data 1 & Data 2 & Data 3 \\
Data 4 & Data 5 & Data 6 \\
Data 7 & Data 8 & Data 9 \\
\bottomrule
\end{tabular}
\caption{Example of a Three-Line Table}
\end{table}
```

- **Query3:**<How can I determine whether the ACF and PACF plots are tail-cut or tail-drag when using an ARIMA model? What kind of plots indicate that the ARIMA model can be used?>



- **Output:** When using an ARIMA (AutoRegressive Integrated Moving Average) model, the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots provide useful insights into the characteristics of the time series and help determine the appropriate model parameters.

ACF and PACF Interpretation: Tail-Cutting (or "Cut-off"):

ACF: If the ACF plot cuts off after a certain number of lags (i.e., significant autocorrelations at the initial lags, followed by values close to zero), it suggests a MA (Moving Average) model component. PACF: If the PACF plot cuts off after a certain number of lags (i.e., significant partial autocorrelations at the initial lags, followed by values close to zero), it suggests an AR (AutoRegressive) model component.

Tail-Dragging (or "Long Tail"):

ACF: If the ACF plot shows a gradual decay (a slow decline in the autocorrelation values), it suggests the presence of an AR (AutoRegressive) model component. PACF: If the PACF plot shows a gradual decay (a slow decline in partial autocorrelation values), it suggests the presence of an MA (Moving Average) model component.

ARIMA Model Applicability: ARIMA Model Usage: ARIMA is applicable when the time series is stationary or can be made stationary through differencing (the "I" component in ARIMA). The key is to identify if the series exhibits characteristics that can be captured by AR and MA components. If the ACF cuts off and the PACF shows a gradual decay, an AR model might be appropriate. If the PACF cuts off and the ACF shows a gradual decay, an MA model might be appropriate. A combination of both features (cut-off and gradual decay in ACF and PACF) indicates that an ARMA or ARIMA model could be a good fit. **Summary:** AR model: PACF cuts off and ACF tails off. MA model: ACF cuts off and PACF tails off. ARMA model: Both ACF and PACF tail off gradually. ARIMA model: For non-stationary data, you first difference the series and then apply ARIMA, checking if the differenced series becomes stationary. **ACF and PACF Plots for ARIMA:** When deciding if ARIMA can be used, check the plots to see if they exhibit one of the characteristics mentioned above. If neither ACF nor PACF cuts off abruptly or tails off gradually, you may need to transform or difference the data to make it suitable for ARIMA modeling. When using an ARIMA (AutoRegressive Integrated Moving Average) model, the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots provide useful insights into the characteristics of the time series and help determine the appropriate model parameters.

ACF and PACF Interpretation: Tail-Cutting (or "Cut-off"):

ACF: If the ACF plot cuts off after a certain number of lags (i.e., significant autocorrelations at the initial lags, followed by values close to zero), it suggests a MA (Moving Average) model component. PACF:

If the PACF plot cuts off after a certain number of lags (i.e., significant partial autocorrelations at the initial lags, followed by values close to zero), it suggests an AR (AutoRegressive) model component. Tail-Dragging (or "Long Tail"):

ACF: If the ACF plot shows a gradual decay (a slow decline in the autocorrelation values), it suggests the presence of an AR (AutoRegressive) model component. PACF: If the PACF plot shows a gradual decay (a slow decline in partial autocorrelation values), it suggests the presence of an MA (Moving Average) model component. ARIMA Model Applicability: ARIMA Model Usage: ARIMA is applicable when the time series is stationary or can be made stationary through differencing (the "I" component in ARIMA). The key is to identify if the series exhibits characteristics that can be captured by AR and MA components. If the ACF cuts off and the PACF shows a gradual decay, an AR model might be appropriate. If the PACF cuts off and the ACF shows a gradual decay, an MA model might be appropriate. A combination of both features (cut-off and gradual decay in ACF and PACF) indicates that an ARMA or ARIMA model could be a good fit. Summary: AR model: PACF cuts off and ACF tails off. MA model: ACF cuts off and PACF tails off. ARMA model: Both ACF and PACF tail off gradually. ARIMA model: For non-stationary data, you first difference the series and then apply ARIMA, checking if the differenced series becomes stationary. ACF and PACF Plots for ARIMA: When deciding if ARIMA can be used, check the plots to see if they exhibit one of the characteristics mentioned above. If neither ACF nor PACF cuts off abruptly or tails off gradually, you may need to transform or difference the data to make it suitable for ARIMA modeling.

- DeepSeek DeepSeek-V3 (Jul, 2024 version):

- **Query1:** <Can you give me some examples of great coaches and point out the specific data sources? Please ensure the information is authentic and reliable.>
- **Output:** <Here are a few examples of great coaches across various sports along with reliable data sources to support their achievements:
Sir Alex Ferguson (Football/Soccer)
Achievements: Sir Alex Ferguson is widely regarded as one of the greatest football managers of all time. He managed Manchester United from 1986 to 2013, winning 13 Premier League titles, 5 FA Cups, and 2 UEFA Champions League titles, among other accolades. Source: You can verify this information from reliable sports databases such as Transfermarkt and Manchester United Official Website. Phil Jackson (Basketball)
Achievements: Phil Jackson is known as one of the greatest NBA coaches, having won 11 NBA championships, the most by any coach



in NBA history. He coached teams like the Chicago Bulls (with Michael Jordan) and the Los Angeles Lakers. Source: This information is available on the official NBA website and reputable basketball databases like Basketball Reference. Bill Belichick (American Football)

Achievements: Bill Belichick is the head coach of the New England Patriots, and he has led the team to 6 Super Bowl titles, the most by any coach in NFL history. Source: The details about his coaching career are available on NFL Official Website and other sources like Pro Football Reference. Vince Lombardi (American Football)

Achievements: Vince Lombardi, after whom the Super Bowl trophy is named, won 5 NFL championships as the head coach of the Green Bay Packers in the 1960s. Source: His achievements are documented in NFL.com and historical sources such as the Pro Football Hall of Fame. Pat Summitt (Basketball)

Achievements: Pat Summitt is regarded as one of the greatest college basketball coaches in history. She led the University of Tennessee Women's Basketball team to 8 NCAA championships. Source: Verified by ESPN and Tennessee Volunteers Official Website.>

- **GitHub CoPilot (Jan 16, 2024 version):**

- Auto-completions for code used in preparing our models.