# abstract

During the 2024 Paris Summer Olympics, there was a surge in public interest in individual events and the medal rankings of various countries. Before each Olympic Games, a "virtual medal table" is created to predict the performance of different nations, but such predictions are typically based not only on historical medal data but also significantly influenced by the participating athletes. This article uses information such as the number and types of events a country participates in, project results, the host country, and specialty events to create a model.

For issues one and two, we used K-Means clustering to categorize the countries that have won medals into three types. We calculated the correlation between each country and different events using Pearson correlation coefficients, in order to identify the country's signature events. Based on these three types of countries, we applied the LSTM model for medal count training and prediction. Based on this model, we predicted the medal counts for each country in the 2028 Los Angeles Summer Olympics. The model also found that the number and types of Olympic events have a significant impact on a country's medal count.

For issue three, we used a random forest algorithm to build a model for predicting countries that will win medals for the first time. The goal was to identify the key factors influencing the change in a country's medal count. We selected three feature variables: "number of participants," "number of participations," and "number of events," with "Will\_Earn\_Medal" as the target variable. We trained the model using a random forest regression model and used GridSearchCV to select the best hyperparameters, resulting in the optimal model configuration. We then calculated the probability of each country winning a medal. Finally, the model's performance was evaluated through confusion matrices and ROC curves, showing high prediction accuracy.

For issue four, we used graph theory and network flow theory to quantitatively assess the influence of "great coaches" on a country's medal count. We established a directed network graph between coaches and countries, transforming the coach's impact on medals into node and edge flow relationships. The weight was calculated using the formula: w = 3\*△G + 2\*△S + 1\*△B. We then analyzed the coach’s contribution to medals through total flow and bottleneck flow, providing an optimized path for countries in selecting coaches and helping them formulate more precise sports development strategies.

For issue five, during the modeling process for the above issues, we identified some strategies that can help national Olympic committees increase their medal count. For example, with the increasing participation of women, national Olympic committees can focus on mixed-gender events to enhance winning opportunities. If a country is the host nation, it can apply to add domestic advantage events, optimize infrastructure, and increase financial support to leverage the home advantage and increase medal count. For events that are not monopolized by a few countries, strong nations should analyze these events and invest in cultivating the next generation of athletes, while emerging sports nations can achieve breakthroughs in these non-monopolized events by precisely targeting projects and bringing in excellent coaches.

**keyword:**Prediction, LSTM, Random Forest, Olympic Games, Performance Modeling，Graph Theory

# introduction

Faster, Higher, Stronger - Together.

## 1.1 background

The Paris Olympics attracted global attention, with the events attracting a lot of attention, especially the medal results of the athletes from different countries. Athletes from all over the world fought hard to get a place on the medals table. In addition to the traditional Olympic powerhouses and the hosts' medal race attracting much attention, there was also much discussion about some of the lower-ranked countries, such as Albania, Cape Verde, Dominica and Saint Lucia, who won their first ever medals at the Paris Games. However, there are still more than 60 countries that have failed to collect Olympic medals.

Looking back at history, countries' medal performances in the Olympics show a certain pattern. Before each Olympics, there will be a ‘virtual medal table’ to predict the performance of countries. For example, before the Paris Olympics, Nielsen Gracenote released its final Virtual Medal Table (VMT) predictions for the 2024 Olympics. So what specific factors do such predictions rely on? The fact is that medal predictions are usually made near the start of the Olympic Games by building mathematical models that take into account known athlete participation plans and analysing past gold and total medal counts, in order to predict future medal rankings. Such forecasts are not only valuable to sports analysts, researchers and policymakers, but also help countries to better grasp the trends affecting Olympic performance.

## 1.2 Restatement of the Problem

Given the background information and constraints of the problem, we must complete the following tasks:

* Task 1: Predict the medal table for the 2028 Summer Olympics in Los Angeles, USA, outputting the countries that will perform better as well as those that will perform worse.
* Task 2: Analyse the relationship between the number and type of Olympic sports and the number of medals, output the most important sports for each country and analyse the impact of the host country's choice of sports on performance.
* Task 3: For countries that have not yet won a medal, predict the probability that they will win their first medal at the next Olympic Games and provide an estimate of the probability of this prediction
* Task 4: Explore the impact of the Great Coach effect on team sports, looking for evidence of the impact of the Great Coach effect and estimating the strength of the effect. Finally, select three countries to recommend sports that are worth investing in and estimate their impact.
* Task 5: Provide original insights into Olympic medal counts and explain how these insights inform the Olympic Committee's decision-making.

## 1.3 Our Work

【总流程图】

# Assumptions and Justification

To simplify the problem, we make the following assumptions, each of which is reasonable.

* **Assumption 1**:Historical medal counts can partially reflect the pattern of future medal distribution
* **假设2**:

# Notations

# Data Cleaning

The study began with a systematic cleaning and pre-processing of the raw data.

In terms of athlete data processing, we performed operations such as removing duplicate entries, handling missing values, standardizing, and converting data types on the summerOly\_athletes.csv dataset, which contains basic athlete information (name, team, NOC, sport, events, medals, etc.). In addition, we found that the presence of spaces in the string text affected the subsequent data analyses, so we used str.strip() to remove leading and trailing spaces, and str.replace() to remove internal spaces. We also deleted invalid information and performed reverse processing on the summerOly\_programs.csv dataset for easier future use

.In the NOC (National Olympic Committee) validation session, we used the athletes dataset as the primary reference source for valid NOCs and cross-validated it with the hosts and medals datasets. To ensure data consistency, we removed NOC entries from the hosts and medals datasets that were not present in the athletes dataset. This step provided a reliable foundation for subsequent data analyses.

To ensure data quality, we performed strict validation and quality checks, including ensuring consistency of NOC codes across all datasets, verifying year ranges and temporal continuity, and checking logical constraints such as the number of gold medals not exceeding the total number of medals. Care was taken to maintain the integrity of the data at all times during the dataset consolidation process.

Finally, we performed necessary additional transformations on the data, created an aggregated view of historical performance, unified the data format of each dataset, and prepared a standardised data structure for subsequent analysis and modelling. Through this series of rigorous data cleansing processes, we ensured the consistency, completeness and normality of the datasets, laying a solid foundation for the predictive modelling phase. These steps not only improve the data quality, but also enhance the reliability of the results of subsequent analyses.

# TASK1&2: Medal Count Prediction Based on LSTM

## 5.1 Data Analysis

First, we constructed a Feature Correlation Matrix for countries that have previously won medals to identify the advantage sports of some nations. Subsequently, we created contingency tables for each country, including features such as the year of medal wins, total medal count, gold medal count, host status, number of participants, number of events participated in, total events established by the host, and events from past editions. Finally, we used K-Means clustering to classify countries that have previously won medals into three categories, which serves as the basis for building separate predictive models for different types of Olympic nations.

### 5.1.1 Correlation Matrix Analysis Based on Pearson Correlation

* **Step 1: Data Preparation**：Align the medal count dataset with the event dataset to ensure that each row represents the number of medals a country won in a specific event.
* **Step 2: Pearson Correlation**：The Pearson correlation coefficient is suitable for analyzing data that follows a continuous normal distribution. The calculation formula is as follows:
* **Step 3: Correlation Calculation**：For each country, we combined the number of medals won by that country in various events with the total number of medals awarded across different Olympic events to calculate the Pearson correlation coefficient. The Pearson correlation coefficient quantifies the strength of the relationship between variables. The intensity of variable correlations is shown in Figure [ ]. We visualized the resulting correlation matrix and, taking China as an example, generated a heatmap as shown in Table [ ].

Table: Variable Correlation Strength

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation Strength | Very Strong Correlation | Strong Correlation | Moderate Correlation | Weak Correlation | Very Weak or No Correlation |
| 相关系数绝对值 | 0.8-1 | 0.6-0.8 | 0.4-0.6 | 0.2-0.4 | 0-0.2 |

图表

描述已自动生成

Figure: Heatmap of China

图表, 条形图

描述已自动生成

Figure: Correlation Matrix Using China as an Example

* **Step 4: Analyzing Advantage Sports**: Return a series containing the top three independent variables with the highest correlation coefficients. These represent the three events that most significantly impact a country's total medal count, which are identified as the country's advantage sports. Taking China as an example, the results and the heatmap above indicate that China's advantage sports are gymnastics, table tennis, and diving.
* **Step 5: Constructing Contingency Tables**: Create a contingency table for each country, including the year of medal wins, host status, total medal count, gold medal count, total number of events established by the host, number of events participated in, advantage sports, and average medals won.

### 5.1.2 A Classification Model for Medal-Winning Countries Based on K-Means

图示

描述已自动生成

Figure: K-Means Flowchart

To achieve accurate medal predictions, it is essential to scientifically classify participating countries. Before categorizing national teams based on medal-winning stability, we first need to define stability. For this purpose, we use two key dimensions: NOC participation frequency and diversity of sports events. Based on these dimensions, we apply the K-Means clustering algorithm to classify countries.

1. **Feature Construction**

**NOC Participation Frequency**: f\_{\text{participation}} = \text{count(NOC)}

**Sports Diversity Index**: f\_{\text{diversity}} = \text{unique\\_count(Sports)}

**Feature Normalization**: X' = \text{MinMaxScaler}

1. **Clustering Implementation**

Achieved by minimizing the within-cluster sum of squares (WCSS) objective function.

First, we calculated the occurrence frequency of each NOC in the dataset athletes\_with\_gold\_medal and the number of unique sports associated with each NOC. The data for NOC and sports were normalized using MinMaxScaler, and the two features were merged into a single dataset for K-Means clustering.

During the clustering process, it was observed that data points for the United States exhibited significant outlier characteristics. This was primarily due to the U.S. far exceeding other countries in both NOC participation frequency and the number of sports, resulting in a distinct data distribution. This uniqueness reflects the comprehensive dominance of the U.S. in the Olympics but also posed challenges for clustering analysis. To ensure the rationality of the clustering results, appropriate modifications to the clustering method are required.

图表, 散点图

描述已自动生成

Figure: K-Means Clustering Diagram

We decided to use the rankings of NOC and Sport occurrence frequencies as the basis for clustering and dividing countries into different groups. Additionally, during the model selection phase, we experimented with both K-Means and DBSCAN to cluster the athletes\_with\_medal and athletes\_with\_gold\_medal datasets. The results showed that clustering the athletes\_with\_gold\_medal dataset using the K-Means model yielded the best performance.

Based on Figure [ ], we obtained our final clustering results:Category 1 includes 37 countries that have won gold medals, Category 2 includes 36 countries, and Category 3 includes 43 countries.

|  |  |
| --- | --- |
| Table: Clustering Results | |
| Group | Feature Description |
| Cluster 1 (Olympic Powerhouses) | High participation frequency Diverse event distribution Stable medal-winning ability |
| Cluster 2 (Developing Sports Nations) | Moderate participation frequency Relatively concentrated event distribution Fluctuating medal performance |
| Cluster 3 (Unstable Nations) | Low participation frequency  Highly unstable medal performance   |  |  | | --- | --- | | Strong reliance on advantage sports |  | |

## 5.2 Model Selection

Considering that the data we processed is associated with specific time points and measured at uniform intervals (every four years), we decided to use time series analysis to predict medal counts.

Traditional time series forecasting methods, such as the commonly used Auto-Regressive Integrated Moving Averages (ARIMA), rely solely on a single time series data source. However, predicting medal counts typically requires incorporating multiple variables beyond just historical medal counts.

With the development of machine learning and artificial intelligence, various deep learning algorithms have been applied to time series forecasting. For example, Long Short-Term Memory Networks (LSTM) can address the limitations of Recurrent Neural Networks (RNNs) in capturing long-term cycles and seasonal patterns. In each time step of an LSTM, there is a memory cell, which provides the network with selective memory functionality. This allows LSTM to determine which content to retain at each time step.

Moreover, LSTM can handle multivariate analysis, and its ability to analyze multiple variables aligns well with the three-category grouping of countries based on K-Means clustering. This makes it possible to adopt differentiated variable weighting strategies for different types of countries.

## 5.3 Long and Short Term Memory Network

The study employs an LSTM network as the core predictive model, consisting of three key components: the Forget Gate, Input Gate, and Output Gate. The Forget Gate $f\_t$ controls the retention of historical information, the Input Gate $i\_t$ manages the updating of new information, and the Output Gate $o\_t$ determines the output of information. Together, these three components form a complete memory mechanism:

Here, $h\_t$ represents the hidden state at time $t$, and $x\_t$ is the input vector.

图片包含 物体, 钟表, 游戏机, 火车

描述已自动生成

Figure: Workflow Diagram of the LSTM Mechanism

## 5.4 Implementation of LSTM

We performed normalization on the data to reduce the impact of scale differences between features and ensure stability during the training process. Subsequently, we created the input dataset using a sliding window approach, segmenting the original medal data based on the time window. The time step was set to 20, meaning each year's medal count is related to the data from the previous 20 years.

图示

描述已自动生成

Figure: Sliding Window

### 5.4.2 Quantifying the Impact of Host Countries

We used the Ordinary Least Squares (OLS) method, taking the total medal count as the dependent variable and host country and events as independent variables. The model yielded an R2R^2R2 of 0.758 and an adjusted R2R^2R2 of 0.756, both relatively high, indicating that the model explains over 75.8% of the variation in total medal counts, demonstrating a good fit. The p-value for the independent variable "host country" is less than 0.05, indicating high significance and statistical validity, with a substantial impact on the total medal count.

Table: Results of the Ordinary Least Squares Method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | coef | std err | t | P>|t| |
| const | 10.4972 | 5.090 | 2.062 | 0.041 |
| host | 26.9574 | 4.975 | 5.419 | 0.000 |
| sport | 1.3783 | 0.226 | 6.103 | 0.000 |

### 5.4.3 Building the LSTM Model

As mentioned earlier, we used K-Means clustering to divide medal-winning countries into three categories:

* **Category 1**: Established Olympic powerhouses with frequent participation and a wide range of events, such as the United States and China.
* **Category 2**: Countries with frequent participation but inconsistent performance, such as Luxembourg and Singapore.
* **Category 3**: Countries with infrequent participation and unstable performance, such as India and Jamaica.

For each country, we constructed a time series that includes variables such as annual medal count, gold medal count, host status, number of participants, number of events participated in, number of events established by the host, and the distribution of awards for each event in previous Olympics. Among these, medal count, gold medal count, host status, number of participants, number of events, and the number of events established by the host were treated as the country’s historical predictors of medal performance. The product of the award distribution for each event and the Pearson correlation coefficient was used to represent the country’s advantage sports.

**Weighting strategies by category**:

* **Category 1**: These countries are typically strong and have won medals across a wide range of events, without relying on advantage sports. For such countries, since the influence of being the host nation is significant, we multiplied this factor by the influence derived from the OLS regression, distributed the weights of other historical predictors evenly, and set the weight for advantage sports to zero.
* **Category 2**: For these countries, medal predictions depend on both past participation and performance, as well as certain advantage sports. For these nations, all variables were included, and the weight for advantage sports was calculated as weight=(1−w)×Pearson correlation coefficient\text{weight} = (1-w) \times \text{Pearson correlation coefficient}weight=(1−w)×Pearson correlation coefficient.
* **Category 3**: These countries have few appearances and very low historical medal counts, making predictions akin to estimating the probability of winning medals. As past participation data is less relevant, only advantage sports were considered, and the weight of historical data was set to zero.

**Model Architecture Design**:  
We employed an LSTM layer with 50 memory cells as the primary feature extractor, capable of effectively capturing both short- and long-term dependencies within Olympic cycles. Following the LSTM layer, a fully connected layer was used for feature integration. Finally, the output layer, consisting of two neurons, was designed to predict total medal count and gold medal count, respectively.

Considering that recent Olympic performance is more indicative of future outcomes, we introduced a temporal decay weighting mechanism in the model. This allowed the model to retain historical information while focusing more on recent data trends.

## 5.5 Prediction Interval from the Result

We calculated the prediction intervals through multiple training and testing iterations for each country's model. Each country's model was trained and predicted 10 times to obtain the distribution of the final results. After removing outliers from the results, the maximum and minimum values were taken as the range of medal counts or the range of winning probabilities. The average loss for a country's model was calculated as the model's loss, and the overall model performance was determined by averaging the losses of all countries' models.

图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成

Figure: Prediction Intervals for Total Medals and Gold Medals of Different Countries

Using the trained model, predictions were made for the three categories of countries. The normalized prediction results were denormalized back to their original values, and the results are shown in Figure [ ]. Green indicates an increase in medals, red indicates a decrease, and blue indicates a stable trend. The bar chart represents a range of values.

图表, 散点图

描述已自动生成

Figure: Medal Predictions for Category 1 Countries 图表, 散点图

描述已自动生成

Figure: Medal Predictions for Category 2 Countries 图表, 散点图

描述已自动生成

Figure: Medal Predictions for Category 3 Countries

As shown in the figure above, 73 countries, represented by the United States and China, are expected to see an increase in medal counts, while 47 countries, represented by Australia and Canada, are expected to experience a decrease.

We analyzed the newly added events in previous Olympic Games and conducted a Pearson correlation analysis by combining the total historical medal counts of host countries in these new events with their overall medal counts. The results showed an average Pearson correlation coefficient of 0.4489306, indicating that the selection of new events by host countries has a significant impact on their overall medal count.

图表, 条形图

描述已自动生成

Figure: Pearson Correlation Coefficient Diagram

## 5.6 the Advantage of LSTM

The prediction of final medal counts using the LSTM model is typically not solely based on the time series of historical medal counts but rather considers multiple factors comprehensively, allowing for the analysis of both historical factors and the influence of advantage sports. To evaluate the advantages of the LSTM model in Olympic medal prediction tasks, we conducted a comparative experiment with the traditional Hidden Markov Model (HMM). By comparing the mean squared error (MSE) between the predicted results and actual values on the test set for both models, we observed that the LSTM model demonstrated significant superiority, with its prediction error being notably lower than that of the HMM model, as shown in Figure [ ].

|  |  |
| --- | --- |
| 图表, 条形图, 折线图  描述已自动生成 | 图表, 直方图  描述已自动生成 |

Figure: Comparison of Medal Predictions Between HMM and LSTM Models

From the figure, it is evident that compared to the traditional Hidden Markov Model (HMM), the LSTM model demonstrates significantly better performance in predicting Olympic medals. We attribute this to the following reasons:

1. **Model Structure**:The gated mechanism of LSTM (including the Forget Gate, Input Gate, and Output Gate) allows the model to adaptively regulate the retention of historical information. This enables the LSTM model to capture the long-term stable growth trends in medal counts for Olympic powerhouses. In contrast, HMM is constrained by its Markov assumption, which only models dependencies between adjacent time steps, making it difficult to capture long-term evolutionary patterns. Additionally, HMM struggles to accurately analyze the probability of countries with very few historical medals winning again or predict the outcomes for countries with high medal volatility or reliance on advantage sports. As shown in the figure [Loss Comparison], the LSTM prediction curve aligns much more closely with the actual values compared to the HMM model.
2. **Feature Representation Capability**:LSTM, through its memory cells, can integrate multiple dimensions of information simultaneously, including historical medal counts, event participation, athlete scale, and other key factors. This dynamic multi-dimensional feature fusion mechanism allows the model to more accurately predict complex phenomena such as host country effects and the rise of emerging sports nations. In contrast, HMM’s state-space representation capability is relatively limited.
3. **Optimization Objective**:LSTM uses an end-to-end gradient descent training approach, which directly optimizes prediction errors. In comparison, HMM relies on parameter learning through maximum likelihood estimation, which introduces a mismatch between the training objective and the actual prediction task. This discrepancy is a significant reason why LSTM demonstrates superior robustness in experiments.

# TASK3：Random Forest-based prediction of first-time award-winning countries

## 6.1 Random Forest Model

To identify the key factors influencing the transition in the number of medals won by a country, two steps are required:：

* **First**, a predictive model is developed to predict the turning point where the number of medals shifts from zero to one.
* **Second,** infer key indicators based on the model results.

## 6.2 Data Preparation and Preprocessing

### 6.2.1 Data Table Construction

Construct three data tables, participation\_by\_year\_country, participation\_by\_year\_country\_count, and sport\_count\_pivot, to record the number of participants of the country each year, the cumulative number of participations by year, and the number of events participated each year, respectively.

To exclude countries that did not win medals, entries from the participation\_by\_year\_country table corresponding to countries listed in the NOC column of the medal\_counts table were deleted.

In the medals\_by\_year and medals\_by\_year\_gold datasets, the historical data of each country was traversed to determine the year in which they first won a medal. If all values in a column were zero, None was returned. The results were then stored in first\_medal\_filtered for preview. This dataset only contains countries whose first medal or gold medal year is greater than or equal to 1920.

### 6.2.2 Dataset Creation

**First Iteration:**  
For each country's NOC, corresponding information was extracted from the participation\_by\_year\_country\_count, participation\_by\_year\_country, and sport\_count\_pivot data frames:

* **participation\_count**: The number of times the country participated in the year it first won a medal.
* **events\_participation**: The number of events the country participated in during the year it first won a medal.
* **sport\_count**: The number of sports the country participated in during the year it first won a medal.

These extracted values were then added to new\_table\_data, which will serve as the basis for subsequent datasets.

**Second Iteration:**  
To obtain data from the four years prior to the first medal year, the data from the four years after the first medal year were subtracted.

**Third Iteration:**  
For each country that had never won a medal, the country's NOC, participation\_count, events\_participation, and sport\_count were added to new\_table\_data as a list.

Finally, the entire list of data stored in new\_table\_data was converted into a Pandas DataFrame. After filling missing values and cleaning invalid data, the final dataset, tree\_dataset, was formed.

## 6.3 Model Construction

**6.3.1 Variable Analysis**  
 After careful selection and consideration of the available limited data, we chose three feature variables to train the model. The detailed list of variables is as follows:

| **Indicator** | **Description** |
| --- | --- |
| **participation\_count** | The number of participants from the country |
| **Participation\_count** | The number of times the country participated |
| **sport\_count** | The number of sports in which the country participated each time |

At the same time, our target variable is **Will\_Earn\_Medal**, which indicates whether a medal will be earned (1 or 0).

**6.3.2 Model Training**  
 After standardizing the data using StandardScaler, the dataset was split into training and testing sets, with 80% used for training and 20% for testing.

We utilized the aforementioned features as input parameters for the Random Forest model to train on the training set. The parameter grid for the Random Forest model was as follows:

| **Number of Estimators** | **Max Depth** | **Min Samples Leaf** | **Min Samples Split** | **Max Features** |
| --- | --- | --- | --- | --- |
| [50, 100, 150, 200] | [None, 10, 20, 30] | [2, 5, 10] | [1, 2, 4] | ['auto', 'sqrt', 'log2'] |

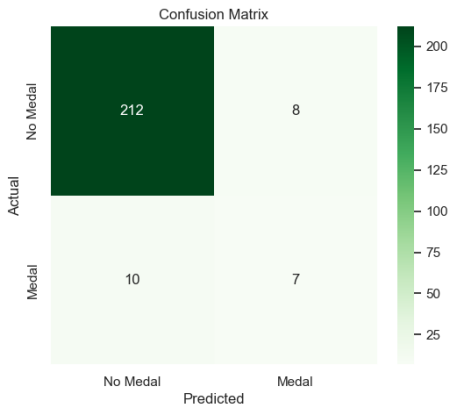
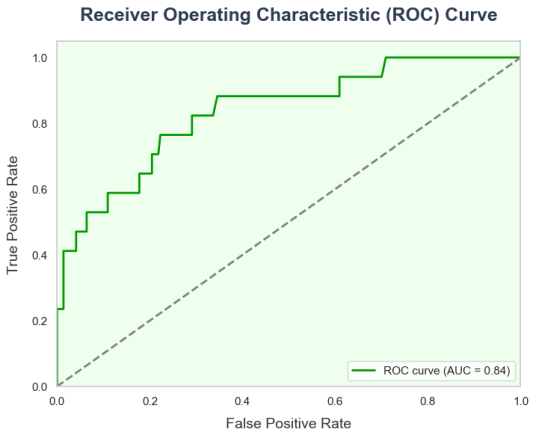
We employed the RandomForestRegressor algorithm from the scikit-learn (sklearn) machine learning library, along with the GridSearchCV method for hyperparameter tuning. The optimal hyperparameter combination obtained from GridSearchCV was as follows:

| **Number of Estimators** | **Max Depth** | **Min Samples Leaf** | **Min Samples Split** | **Max Features** |
| --- | --- | --- | --- | --- |
| 200 | 10 | 1 | 2 | sqrt |

**6.3.3 Model Performance Visualization**  
 To assess the performance of the model, we conducted an initial evaluation using a confusion matrix and ROC curve.

* **Confusion Matrix**: Displays the relationship between the model's predictions and the true labels, helping to analyze the types of errors made by the model.
* **ROC Curve**: Plots the Receiver Operating Characteristic (ROC) curve, which measures the model's performance across different thresholds.

The confusion matrix and ROC curve are shown below:



The confusion matrix indicates that the model's prediction accuracy is 94%, with a false positive rate of only 0.039 (9/233), suggesting that the model is highly reliable and that most samples are correctly classified. The match between the predicted values and the true labels in the test set is high.  
 The model's AUC (Area Under the Curve) is 0.87, demonstrating strong discriminative power and the ability to effectively distinguish between positive and negative classes. Additionally, the ROC curve rapidly rises in the low false positive rate (FPR) region, indicating that the model achieves a high recall rate while maintaining a low false positive rate. The model's true positive rate is commendable。

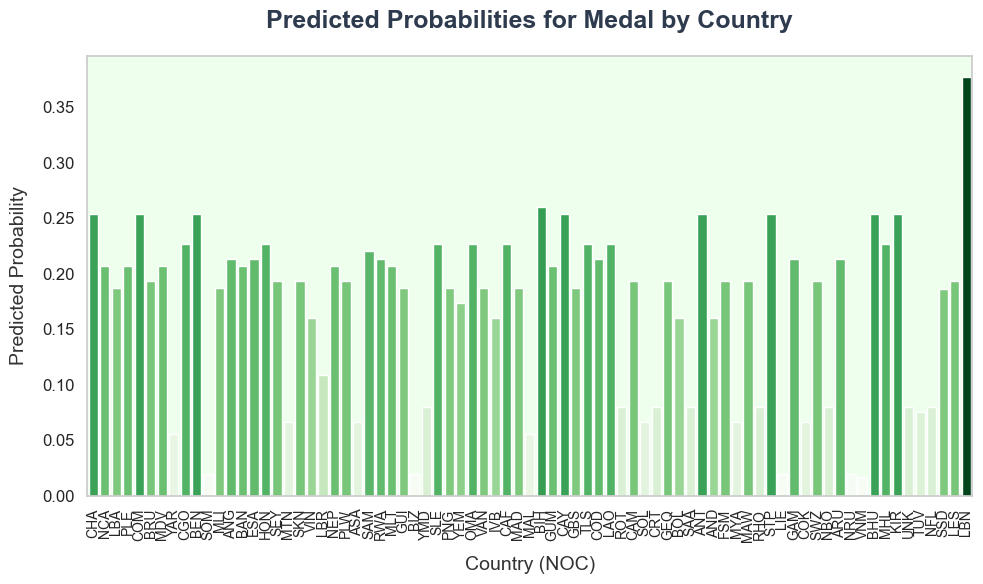
**6.3.4 Model Prediction**  
 Using the best-trained model, **best\_rf\_model**, we predicted the outcomes for new standardized data. The predict\_proba() method returns the probability of each sample belonging to each class, which allows us to calculate the probability of each country winning a medal.

**6.4 Model Testing  
6.4.1 Test Set Construction** To validate the model's performance in real-world prediction tasks, we identified and selected historical samples of countries that have never won a medal from the **athletes** dataset. We used a temporal grouping method to ensure the continuity and completeness of the data. After standardizing the feature data and applying quality control measures, we ultimately formed a high-quality test dataset.

A new dataset was generated, containing relevant statistical data for all countries that have never won a medal, for the year 2024.

**6.4.2 Testing**  
 Using the trained Random Forest model **best\_rf\_model**, we predicted the features for each country and calculated the probability of each country winning a medal. These probabilities were stored in a list, sorted, and the countries most likely to win a medal were identified. A bar chart was then generated to visualize the predicted probabilities of each country winning a medal.

Through probability distribution analysis, it was found that the prediction results exhibit distinct layering. Countries with a predicted probability greater than 0.20 are primarily concentrated in emerging market nations with well-developed sports infrastructure. These countries generally show consistent investment in sports and the specialization of sports programs.



# TASK4: Quantitative Graph Theory Model of the "Great Coach Effect" Based on Network Flow

### ****7.1 Model Background****

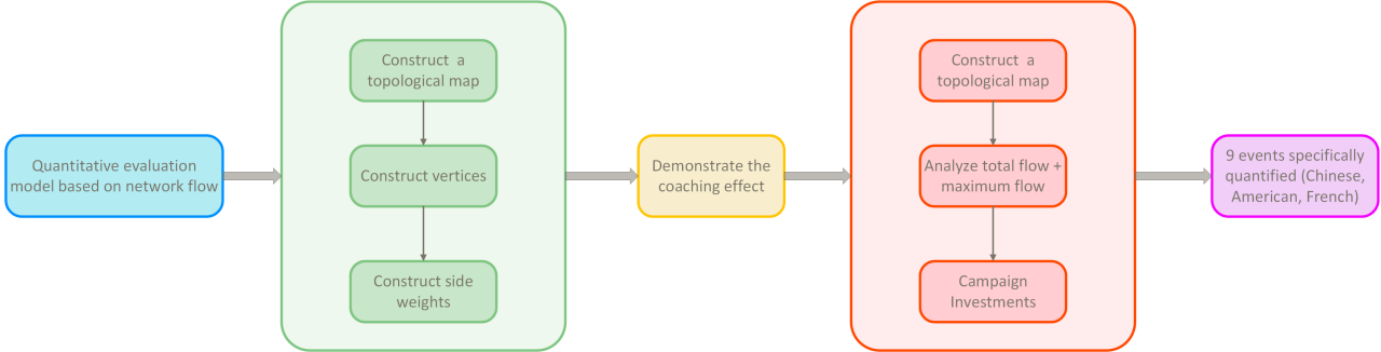
In international competitions such as the Olympic Games, athletes' nationalities are typically restricted; however, coaches do not need to be citizens of the countries they coach. This allows them to move freely between countries. In certain cases, coach changes have had a significant impact on the performance of sports teams, especially when they have a "great coach effect." Such coaches may help teams or athletes break through performance barriers and improve results. The aim of this model is to verify the existence of the "great coach effect" through data analysis and quantify its contribution to medal counts.

In this section, we need to complete:

* **Data Analysis**: Explore whether coach changes are related to changes in medal counts in different countries and sports.
* **Model Construction**: Analyze the impact of the coach effect on medal counts through a network flow model based on graph theory.
* **Impact Prediction**: Identify which sports are suitable for investing in "great coaches" and analyze their potential impact.

### ****7.2 Model Construction****

This section aims to construct a directed graph based on the network flow model of graph theory, which can effectively reflect how coach movement influences medal count changes in various countries. Through this model, we can quantify the coach’s influence, especially the impact on a country’s sports results after the coach's movement or replacement. The nodes in this graph represent different countries, while the directed edges represent the movement of coaches between countries and their contributions to medal counts. By applying the total flow and bottleneck flow algorithms, we can deeply analyze the coach's role, evaluate their impact on medal counts, and provide scientific recommendations for future coach selection and resource allocation.



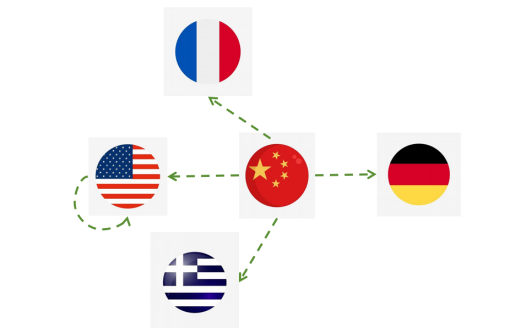
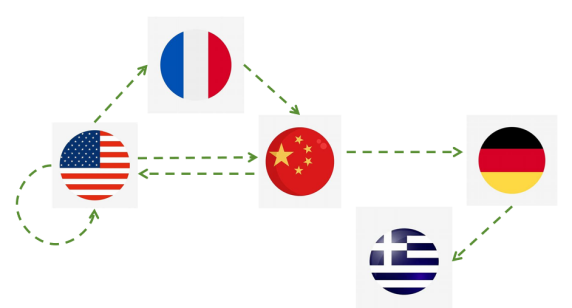
#### ****7.2.1 Construction of the Basic Directed Graph****

In the network flow model based on graph theory, we first build a basic directed graph that can reflect the movement of coaches. The nodes represent countries, while the directed edges represent the flow path of a coach moving from one country to another (with the Olympic Games as the time period). If the coach remains in the same country and participates in multiple Olympic Games, a self-loop edge can be used.

* **Nodes**: Each node represents a country, and it stores the change in gold, silver, and bronze medals (ΔG, ΔS, ΔB) for that country over a specific time period.

节点集合V：表示参与国家，  
V={v\_i\mid i\in I}  
，其中I为国家索引集

* **Directed Edges and Direction**: Directed edges represent the movement of coaches from one country to another. The direction of each directed edge is determined by the direction of the coach's transfer, from the original country (outflow country) to the target country (inflow country). The capacity of the directed edge represents the coach's contribution to medal count changes in the target country.

边集合E：表示教练流动路径，E={(v\_i,v\_j)\midi,j\inI}

#### ****7.2.2 Weight Calculation****

In this model, the weight ww of each directed edge represents the change in medal counts before and after the coach’s movement and is quantified by the increase or decrease in medal counts. The weight calculation formula is as follows:

Where:

* ΔG represents the change in gold medals;
* ΔS represents the change in silver medals;
* ΔB represents the change in bronze medals;
* wG,wS,wB are the weight coefficients for gold, silver, and bronze medals, used to quantify the impact of each medal type on the coach’s effect.

Since there is a lack of detailed coach data to quantify the specific impact of medal changes on each country’s performance, we assign fixed weight coefficients to each medal type: wG=4, wS=2, wB=1.

##### ****Data Extraction****

We use the function filter\_and\_count\_medals to filter and count Olympic data in order to analyze the performance of different countries in various sports under certain conditions. This function filters the data according to the input conditions (such as country, year, gender, and sport), and calculates the quantity of different types of medals (gold, silver, bronze).

#### ****Self-loop Edges for the Same Country****

For the coach’s movement within the same country, we calculate the change in medal counts between two consecutive Olympic Games in the same country and apply the medal weight coefficients to obtain the score. The calculation formula for self-loop edges is as follows:

w\left(e\right)=4\mathrm{\Delta G}+2\mathrm{\Delta S}+\mathrm{\Delta B}

#### ****Coach Movement Between Different Countries****

For coaches who move between different countries, we calculate the change in medal counts for both countries involved, multiply by the weight coefficients, and compute the difference between the positive growth in the target country and the negative growth in the original country to obtain the weight ww of the directed edge.

w\_A=4\mathrm{\Delta G\_{A}}+2\mathrm{\Delta S\_{A}}+\mathrm{\Delta B\_{A}}

w\_B=4\mathrm{\Delta G\_{B}}+2\mathrm{\Delta S\_{B}}+\mathrm{\Delta B\_{A}}

w = w\_A – w\_B

Where:

* A is the inflow country;
* B is the outflow country;
* wA is the weighted score of medal count changes in country A;
* wB is the weighted score of medal count changes in country B.

#### ****Lang Ping’s Coach Movement Between China and the U.S.****

In this model, Lang Ping’s coaching movement is treated as an influence transfer between countries, from the U.S. to China or vice versa. We need to calculate the weight of Lang Ping’s transfer from the U.S. to China based on the changes in medal counts.

For instance, Lang Ping coached the U.S. women’s volleyball team from 2005 to 2008, and in 2008, the U.S. team won a silver medal. In 2012, she returned to China and led the Chinese women’s volleyball team to a gold medal at the 2016 Rio Olympics.

By utilizing the data extraction method, we calculate the medal counts for the U.S. women’s volleyball team in 2004, 2008, 2012, and 2016, as well as for China in 2012 and 2016. The weight calculation method involves calculating the changes in medal counts, applying the weight coefficients, and determining the impact of Lang Ping's coaching movement on both countries.

According to the weight calculation method, since Lang Ping was coaching in the U.S. in 2008 without any coach movement, a self-loop edge is drawn for the U.S. The difference between the medal counts in 2008 and 2004 is calculated: ΔG = +0, ΔS = +1, ΔB = +0. By multiplying these changes by the medal weight coefficients, the total score is 2.

During the period of 2016, Lang Ping moved from the U.S. to China. The medal counts for China in 2016 were compared to those in 2012, resulting in ΔG = +1, ΔS = +0, ΔB = +0. By multiplying these changes by the medal weight coefficients, the total score for China is 4. Similarly, for the U.S., the medal counts in 2016 were compared to those in 2012, yielding ΔG = +0, ΔS = -1, ΔB = +1, which gives a total score of -1 after applying the weight coefficients. Finally, the positive growth for China and the negative growth for the U.S. are subtracted, resulting in a difference of 5.

The calculation process is visualized in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 运动 | 教练 | 年份 | 奖牌 | 奖牌变化 | 得分 |
| 排球 | 郎平 | 2004→2008  （美国自环边）： | 金牌  银牌  铜牌 | +0  +1  +0 | 2 |
| 2012→2016  （美国→中国） | 金牌  银牌  铜牌 | USA+0 CHN+1  USA-1 CHN+0  USA+1 CHN+0 | 5 |

#### ****7.2.3 Flow Calculation****

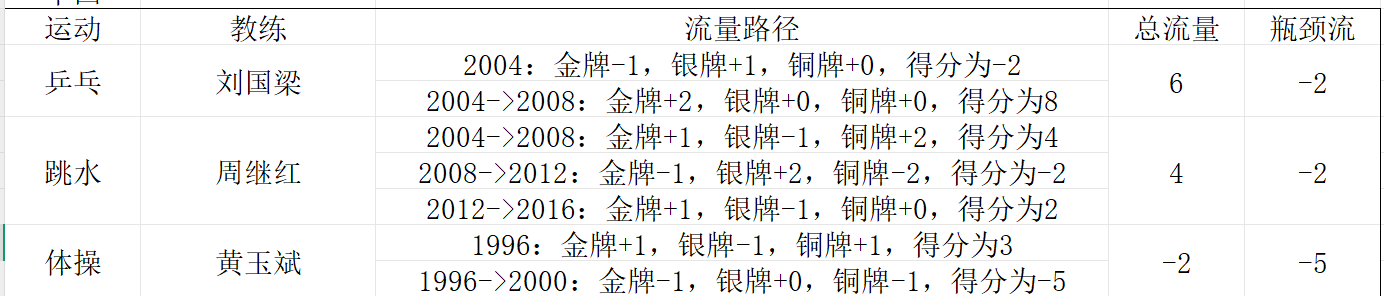
* **Total Flow**: Total flow represents the maximum flow between nodes, indicating the total contribution of coaches and other factors to the overall medal count.
* **Bottleneck Flow**: The minimum flow along each path, representing the limitations imposed by other factors (such as athletes' abilities, technical requirements, etc.) on medal counts. A low bottleneck flow means the potential for improvement in the project is limited, while a high bottleneck flow suggests that the coach’s contribution is substantial.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 运动 | 教练 | 流量路径 | 总流量 | 瓶颈流 |
| 排球 | 郎平 | 2004→2008（美国自环边）：金牌+0，银牌+1，铜牌+0，得分为2 | 7 | 2 |
| 2012→2016（美国→中国）：美国金牌+0，银牌-1，铜牌+1；中国金牌+1，银牌+0，铜牌+0，得分为5 |

### ****7.3 Model Results and Analysis****

This section analyzes the changes in medal counts for China, the U.S., and France, focusing on the impact of the "great coach" effect. Using the flow network model, we quantify the effect of coach movements on medal counts and propose investment strategies based on the model results.

#### ****China: Table Tennis, Diving, Gymnastics****



* **Table Tennis**: Under Liu Guoliang's coaching, table tennis exhibited a relatively large total flow (6), but the bottleneck flow was -2, indicating that although the coach’s contribution was evident, certain limiting factors restricted the potential for improvement.
* **Diving**: Zhou Jihong’s coaching period was relatively stable, with a total flow of 4 and a bottleneck flow of -2, suggesting that despite the coach’s contribution, the project experienced significant volatility, especially between 2008 and 2012.
* **Gymnastics**: Under Huang Yubin, the gymnastics team showed poor performance, with a total flow of -2 and a bottleneck flow of -5, demonstrating limited coach contribution and significant external restrictions on the project’s performance.

Table tennis is the most promising project. Despite the bottleneck flow, the coach's contribution is significant, suggesting that further investment in the "great coach" effect is warranted in this area. Diving, although stable, requires improvements in other factors due to the presence of bottleneck flows. Gymnastics requires more attention and enhancement, as the coach’s contribution is minimal, and both total flow and bottleneck flow indicate significant room for improvement.

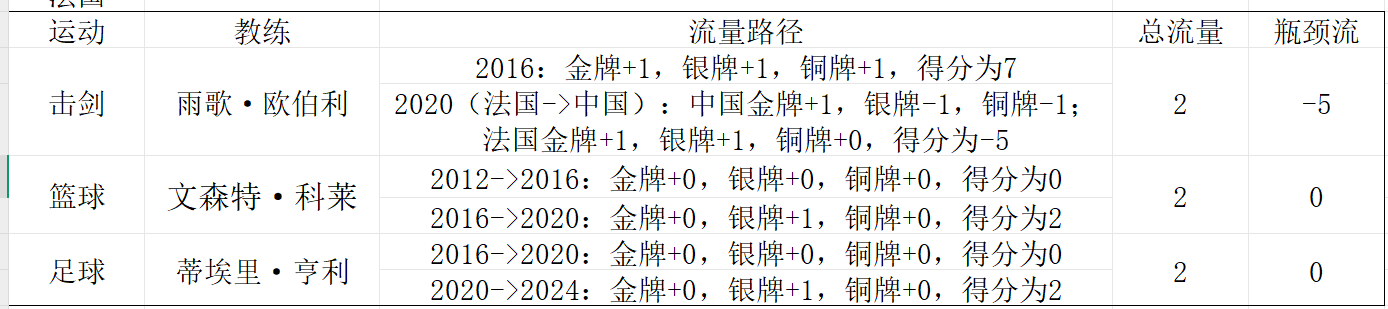
#### ****USA: Gymnastics, Basketball, Volleyball****



* **Gymnastics**: Bela Karolyi’s total flow is relatively high (14), indicating significant coach contribution, but the bottleneck flow is negative (-4), suggesting that performance improvement is constrained by external factors.
* **Basketball**: Under Gregg Popovich, the total flow is 2, and the bottleneck flow is 0, indicating stable performance with limited fluctuations. The team has won gold in multiple editions of the Olympics, and the coach’s capabilities are notable.
* **Volleyball**: Under Karch Kiraly’s coaching, the total flow is 2, and the bottleneck flow is -1, indicating limited coach contribution, with performance volatility due to external restrictions.

Basketball has shown stable performance, but the coach’s contribution is relatively small, and the project’s performance is constrained. Volleyball shows significant volatility, with the coach’s impact being limited. Gymnastics, although the coach’s contribution is substantial, requires further improvements due to external constraints. The gymnastics project is best suited for further investment in "great coach" effects.

#### ****France: Fencing, Basketball, Football****



* **Fencing**: Under Yves Guillard, the coach’s effect is complex. Although France’s performance was relatively high, the negative change after transferring to China suggests significant external constraints. The bottleneck flow is negative, indicating limitations on performance improvement.
* **Basketball**: Vincent Collet’s coaching slightly improved the French basketball team’s performance, with a bottleneck flow of 0, indicating stable performance with no significant external constraints.
* **Football**: Thierry Henry’s coaching had a positive impact on French football, especially in terms of silver medals. The bottleneck flow is 0, indicating a direct effect of the coach’s influence on the project’s performance.

Fencing shows significant volatility, with the coach’s impact being limited by external factors. Basketball performance remains stable with limited contribution from the coach. Football, however, shows positive results, with the coach’s impact being direct and meaningful. Football is most suitable for further investment in the "great coach" effect, especially in enhancing the silver medal potential.

### ****7.4 Verification and Strategy Recommendations****

After analyzing the performance of China, the U.S., and France and their respective coaches, this section explores the impact of the "great coach" effect on medal counts, particularly in the context of coach movements. Although athletes may find it difficult to change countries due to nationality requirements, coaches are free to move between countries and have a significant impact on performance. Based on the analysis, we can conclude that the "great coach" effect contributes significantly to certain sports, especially in gymnastics, table tennis, and football, where coach expertise and experience lead to substantial improvements.

# 8.Letter

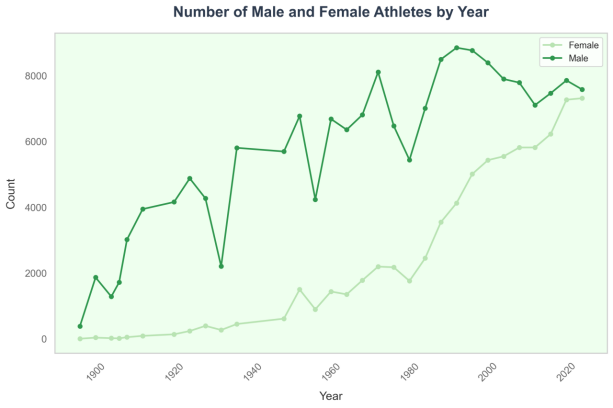
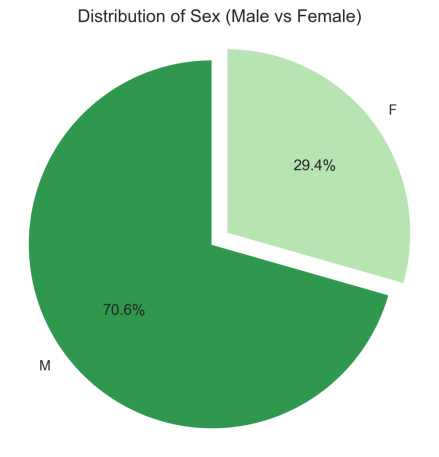
Dear sir or Madam

It is my great honor to write to you and share the results of our analysis regarding Olympic medal predictions. The Olympic Games, as a global platform for competition, not only showcase the pinnacle of athletic excellence but also foster international unity. In our mathematical modeling of Olympic medal predictions, we have identified several key findings, including trends in gender participation, the impact of host nations on medal counts, and the potential of underexplored sports disciplines.

These insights not only reflect historical and current trends but also provide actionable strategic recommendations for National Olympic Committees in their preparations for future Olympic events.

Here are the detailed analyses of these findings and the implications for future planning：

1.Through statistical analysis of Olympic participation data, it was found that gender participation has shown significant evolutionary characteristics. In the early Olympic Games, male athletes dominated (accounting for approximately 70%), while in recent years, this proportion has gradually approached balance.



As seen in the chart above, after the 2012 London Olympics, the male-to-female athlete ratio came close to 1:1 for the first time, reflecting significant progress in gender equality at the Olympics. This phenomenon offers insights for improving performance in various countries:

* National Olympic Committees (NOCs) can improve performance by increasing gender-equal events: As the participation of female athletes in the Olympics gradually increases, more mixed-gender events and 1:1 gender-equal competitions have emerged. Some countries can enhance their overall medal haul by promoting gender-equal participation.
* Gender differences in event settings impact total medal count: Some women's events, which were established later in Olympic history, have seen an increase in participation in recent years. However, when considering the number of sub-events, women’s events are still relatively disadvantaged. This gender imbalance in event settings may affect certain countries' medal performance in specialized events. NOCs should pay attention to adjusting their preparation strategies to maximize their advantages in different gender-specific events.

2.When observing the annual medal counts of each country, we found that being the host nation had a significant impact on the increase in medal counts. As a result, we used the least squares method to verify the impact of hosting the Olympics on a country's medal count. This influence mainly comes from factors such as athlete fatigue from travel, familiarity with the venue, and the influence of home crowds on athletes. Therefore, host countries can improve their medal counts in the following ways:

* Develop more detailed preparation plans based on historical experience and home advantage: Host countries can collaborate with their NOCs to add strength to their country's advantage in specific events and reduce weaker areas to boost their medal count. For instance, for the 2028 Los Angeles Olympics, the U.S. Olympic Committee added five new competition events: Baseball/Softball, Flag Football, Cricket (Twenty20), Lacrosse (Sixes), and Squash. While baseball/softball enjoys global popularity, it is the "national sport" of the U.S., with a wide base of participation and spectatorship domestically. Similarly, flag football, which is much less popular worldwide than sports like weightlifting, is a sport almost exclusively known in the U.S.
* Provide more support in infrastructure and financial resources: NOCs, taking advantage of their host status, can invest more resources in athletes’ economic welfare and competition adaptability prior to the games. Additionally, athletes from host nations receive heightened expectations from the public, and this “patriotic drive” can motivate athletes to perform at their best.

3.We conducted a statistical analysis of the medal shares for each country in various events and ranked the events by the highest share, in ascending order, to identify events that have not been monopolized, such as equestrian, sport climbing, and rowing. This result reveals that when medal distribution is more even across countries, it implies that more countries have the opportunity to compete for medals, providing valuable insights for countries of different types.

* Insights for Olympic Powerhouses: NOCs from powerhouse countries should deeply analyze these "non-monopolized" events to evaluate whether they can increase investment in these areas, nurture a new generation of athletes, and enhance performance in these events at the Olympics.
* Insights for Emerging Sports Nations: For emerging sports nations, the dominance of Olympic powerhouses in traditional major events such as athletics, swimming, and gymnastics is long-established. However, the "Great Coach" effect validated earlier provides new breakthroughs for these nations. For example, Kenya excels in long-distance running but lacks coaches; hiring a "great coach" could help break the long-standing monopoly of Olympic powerhouses. At the same time, this result provides a potential point for increasing medal counts: through precise event targeting and scientific training plans, emerging sports nations can quickly improve their competitiveness in “non-monopolized” events and achieve remarkable results.

**9 Model Assessment**  
**9.1 Strengths**

1.During the data analysis process, we comprehensively considered multiple key factors such as the number of participants, the number of events, hosts, and outstanding coaches. As a result, we conducted extensive data validation, and our predictions were not solely reliant on historical medal counts.

2.When building the model, we compared it with several other models, such as the Hidden Markov Model and Multivariate Logistic Regression, through experimental validation.

3.We applied various mathematical modeling techniques, including machine learning, neural networks, and graph theory, and flexibly adjusted variables and weights to adapt to medal predictions for different countries.

**9.2 Weaknesses**

1.During data cleaning, due to time constraints, we used the athletes dataset as the primary reference for valid NOCs and removed NOCs from the hosts and medals datasets that were not present in the athletes dataset. We did not consider historical changes in countries in greater detail.

2.The lack of data limited the ability to perform more multidimensional model validation. We excluded factors like GDP, and a multivariate linear regression could have better validated the "Great Coach" effect.