

Olympic Predictions, Grasp The Medal Trend

Summary

During the 2024 Summer Olympics in Paris, there has been 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 & two, we used **K-Means clusterin** to categorize the medal-winning countries 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** medal count training and prediction model. Using this model, we predicted the number of medals for each country at the 2028 Summer Olympics in Los Angeles. 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 to predict which countries will win medals for the first time. The aim 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. the performance of the model was evaluated using **confusion matrices** and **ROC curves**, which showed 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: $W = 3 \times \Delta G + 2 \times \Delta S + 1 \times \Delta B$. We then analyzed the coach's contribution to medals through total flow and bottleneck flow, providing countries with an optimized path for countries in selecting coaches and helping them to 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 to increase their medal count. For example, as women's participation increases, National Olympic Committees can focus on mixed-gender events to increase their chances of winning. 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 attracting in excellent coaches.

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

Contents

1	Introduction	3
1.1	Background	3
1.2	Restatement of the Problem	3
1.3	Our Work	4
2	Assumptions and Justification	4
3	Notations	4
4	Data Cleaning	5
5	TASK1&2: Medal Count Prediction Based on LSTM	5
5.1	Data Analysis	5
5.1.1	Correlation Matrix Analysis Based on Pearson Correlation	5
5.1.2	A Classification Model for Medal-Winning Countries Based on K-Means	7
5.2	Model Selection	9
5.3	Long and Short Term Memory Network	9
5.4	Implementation of LSTM	9
5.4.1	Quantifying the Impact of Host Countries	10
5.4.2	Building the LSTM Model	10
5.5	Prediction Interval from the Result	11
5.6	the Advantage of LSTM	14
6	TASK3:Random Forest-based prediction of first-time award-winning countries	15
6.1	Random Forest Model	15
6.2	Data Preparation and Preprocessing	15
6.2.1	Data Table Construction	15
6.2.2	Dataset Creation	15
6.3	Model Construction	16
6.3.1	Variable Analysis	16
6.3.2	Model Training	16
6.3.3	Model Performance Visualization	17
6.3.4	Model Prediction	18
6.4	Model Testing	18
6.4.1	Test Set Construction	18
6.4.2	Testing	18
7	TASK4: Quantitative Graph Theory Model of the "Great Coach Effect" Based on Network Flow	19
7.1	Model Background	19
7.2	Model Construction	19
7.2.1	Construction of the Basic Directed Graph	19
7.2.2	Weight Calculation	20
7.3	Model Results and Analysis	22
7.4	Verification and Strategy Recommendations	24
8	Letter	24
9	Model Assessment	25
9.1	Strengths	25
9.2	Weaknesses	25

1 Introduction

Faster, Higher, Stronger - Together.

——International Olympic Committee

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

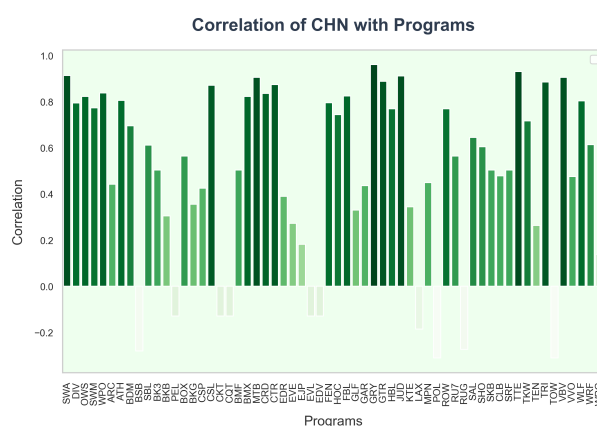


Figure 1: Correlation of CHN with Programs

2 Assumptions and Justification

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

- **Assumption 1:** It is assumed that the number and type of medals won by each country each year are allocated according to specific selection rules: these rules may include the individual performance of athletes, the historical achievements of the country, the scoring criteria for the event, etc. Each country's winnings are determined by a clear set of judging mechanisms and criteria, so the allocation of medals can be viewed as a formulaic process based on these regulations, ensuring that each country has a fair chance based on following the same rules.
- **Assumption 2:** It is assumed that potentially influential factors not included in the dataset do not play a substantial role in the production of medal counts: the model is analysed without taking into account variables that are not given and it is assumed that they do not influence the final outcome of medal counts. If these factors have the potential to introduce noise in the analysis, their interference can be eliminated through data pre-processing techniques, such as removing outliers or using standardisation methods, thus making the results more accurate and reliable.

3 Notations

Symbols	Describe
h	Convection heat transfer coefficient
N	Number of countries
Y	Year of award
M_i	Total number of medals for country i
S_i	Number of silver medals for country i
B_i	Number of bronze medals for country i
ρ	Pearson correlation coefficient

4 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.

5 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:

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma(x) \cdot \sigma(y)} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma(x) \cdot \sigma(y)} \quad (1)$$

- **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 1. We visualized the resulting correlation matrix and, taking China as an example, generated a heatmap as shown in Table 1.

Table 1: Variable Correlation Strength

Correlation Strength	Very Strong Correlation	Strong Correlation	Moderate Correlation	Weak Correlation	Very Weak or No Correlation
Absolute value of correlation coefficient	0.8–1	0.6–0.8	0.4–0.6	0.2–0.4	0–0.2

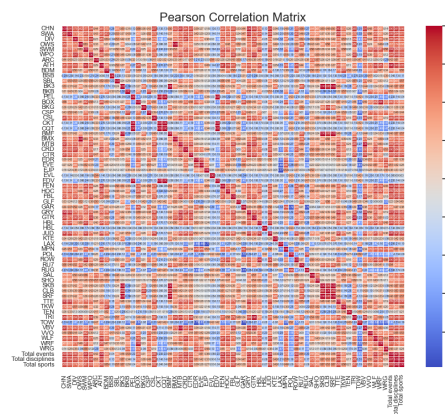


Figure 2: Heatmap of China

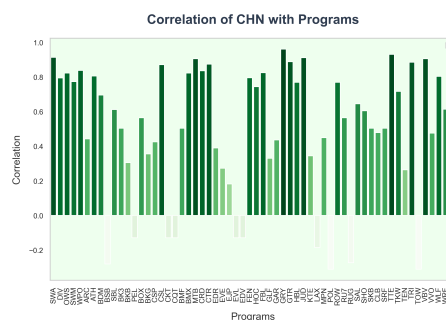


Figure 3: Correlation of CHN with Programs

- **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

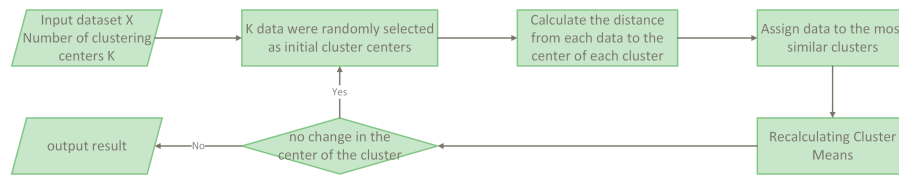


Figure 4: K-Means Flowchart

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: $f_{\text{participation}} = \text{count}(\text{NOC})$

Sports Diversity Index: $f_{\text{diversity}} = \text{unique_count}(\text{Sports})$

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

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

$$d(x_i, C_j^{(r)}) = \sqrt{\sum_{l=1}^d (x_{il} - C_{jl}^{(r)})^2} \quad (2)$$

$$C_j^{(r+1)} = \frac{1}{|S_j^{(r)}|} \sum_{x_i \in S_j^{(r)}} x_i \quad (3)$$

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.

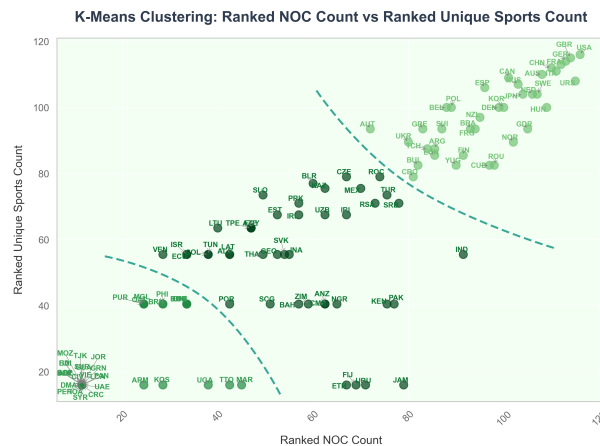


Figure 5: K-Means Clustering Diagram

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 4, 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 2: Group and Feature Description

Group	Feature Description
Cluster-1 (Olympic Power-houses)	<p>High participation frequency</p> <p>Diverse event distribution</p> <p>Stable medal-winning ability</p>
Cluster-2 (Developing Sports Nations)	<p>Moderate participation frequency</p> <p>Relatively concentrated event distribution</p> <p>Fluctuating medal performance</p>
Cluster-3 (Unstable Nations)	<p>Low participation frequency</p> <p>Highly unstable medal performance</p> <p>Strong reliance on advantage sports</p>

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:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

Here, h_t represents the hidden state at time t , and x_t is the input vector.

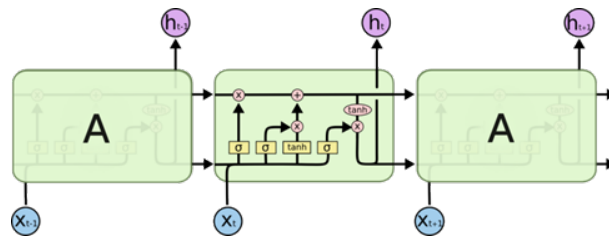


Figure 6: Workflow Diagram of the LSTM Mechanism

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.

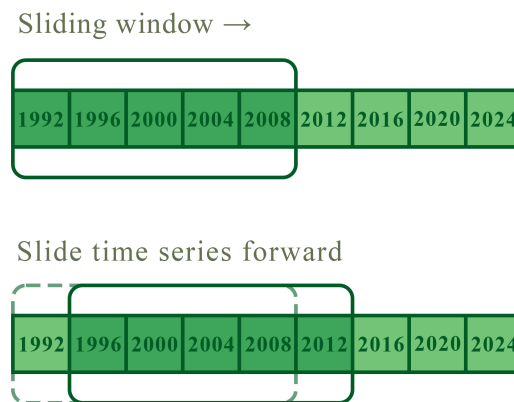


Figure 7: Sliding Window

5.4.1 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 R^R of 0.758 and an adjusted R^R 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 3: Results of the Ordinary Least Squares Method

a	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

5.4.2 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 $weight = (1 - w) \times \text{Pearson Correlation Coefficient}$
- **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.

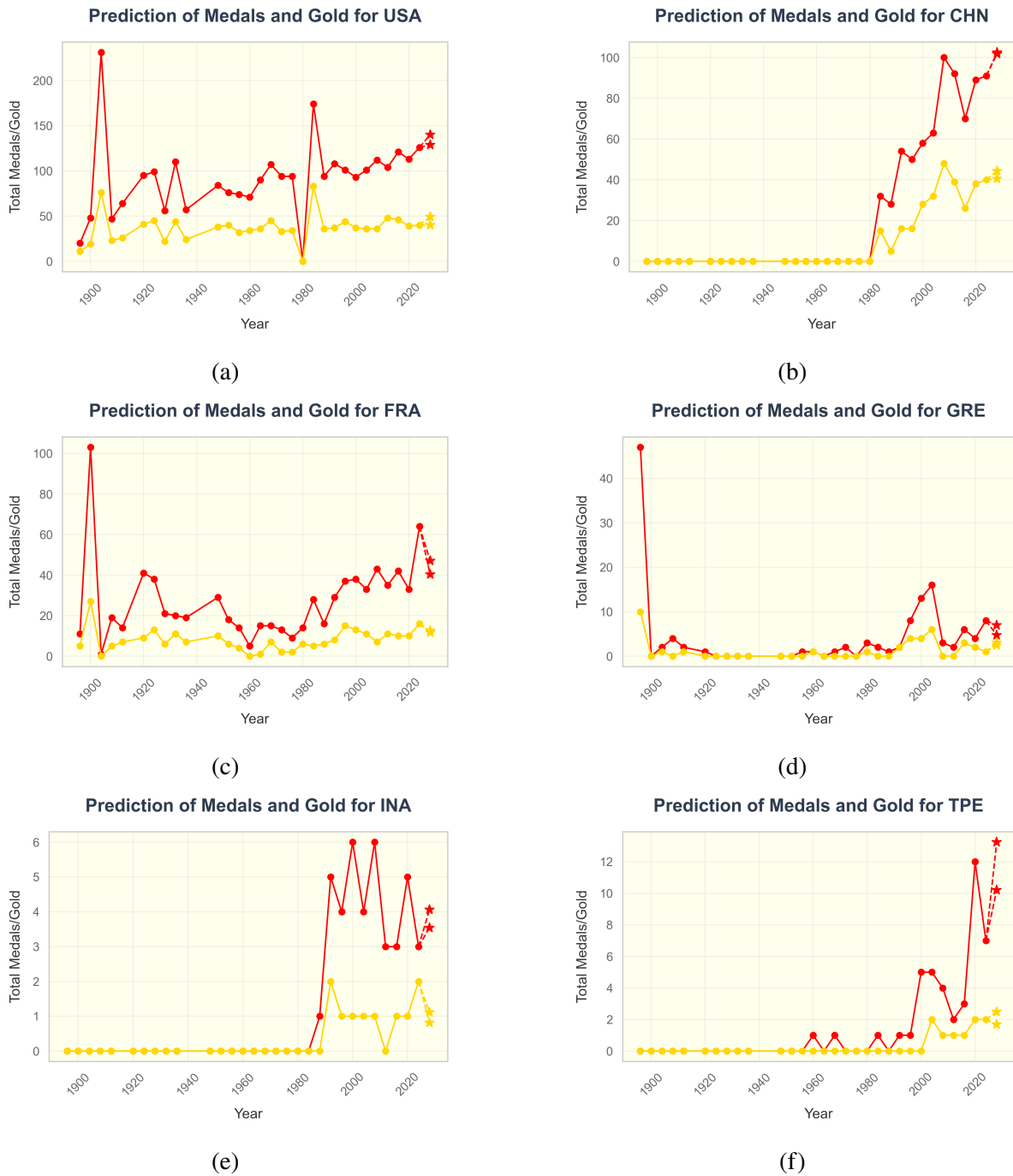


Figure 8: Prediction Intervals for Total Medals and Gold Medals of Different Countries

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 8. Green indicates an increase in medals, red indicates a decrease, and blue indicates a stable trend. The bar chart represents a range of values.

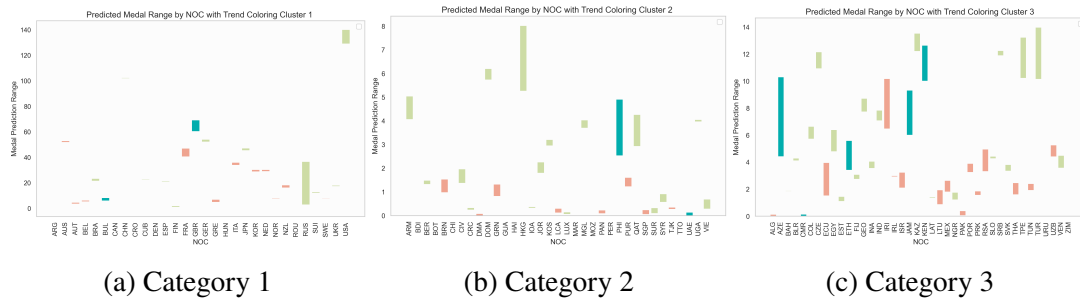


Figure 9: Total Medal Predictions for Category 3 Countries

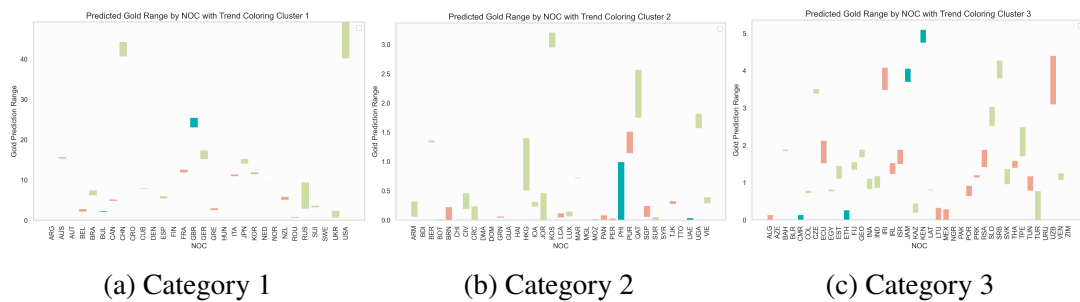


Figure 10: Gold Medal Predictions for Category 3 Countries

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.

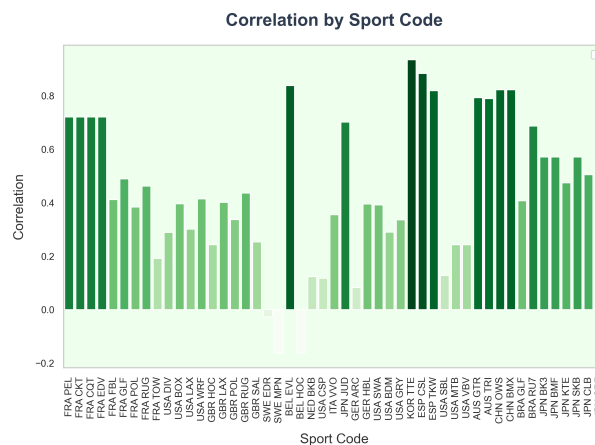


Figure 11: Correlation by Sport Code

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 11

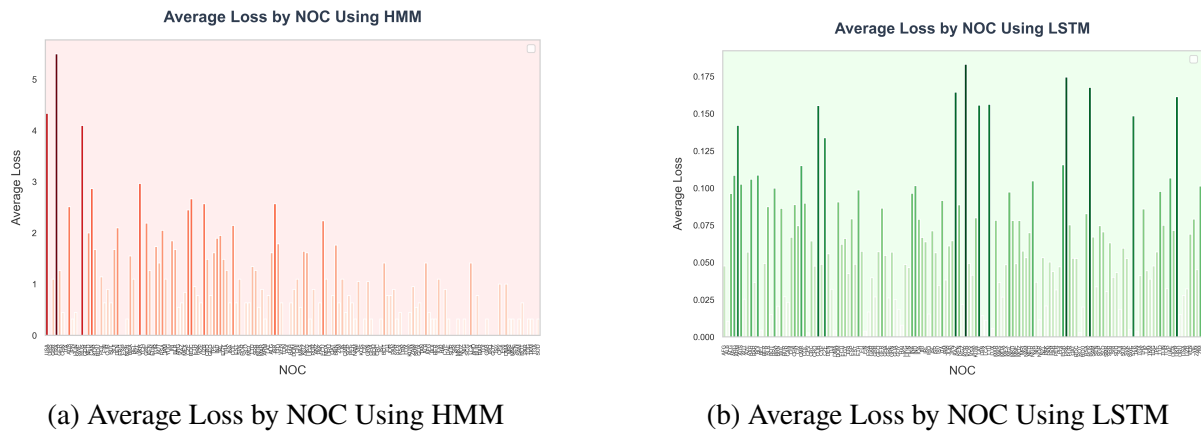


Figure 12: Comparison of Medal Predictions Between HMM and LSTM Models

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.

6 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 any 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 for 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 , which will serve as the basis for subsequent datasets.

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 in 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:

Table 4: Table of Variables

Indicator	Descriptions
<i>participation_count</i>	The number of participants from the country
<i>events_participation_count</i>	The number of times the country participated
<i>sport_count</i>	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:

Table 5: Input parameters for forest models

Indicator	Descriptions
<i>NumberOfEstimators</i>	[50, 100, 150, 200]
<i>MaxDepth</i>	[None, 10, 20, 30]
<i>MinSamplesLeaf</i>	[2, 5, 10]
<i>MinSamplesSplit</i>	[1, 2, 4]
<i>MaxFeatures</i>	['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:

Table 6: Optimal hyperparameter combination

Indicator	Descriptions
<i>NumberOfEstimators</i>	200
<i>MaxDepth</i>	10
<i>MinSamplesLeaf</i>	1
<i>MinSamplesSplit</i>	2
<i>MaxFeatures</i>	sqrt

6.3.3 Model Performance Visualization

To assess the performance of the mod

- **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:

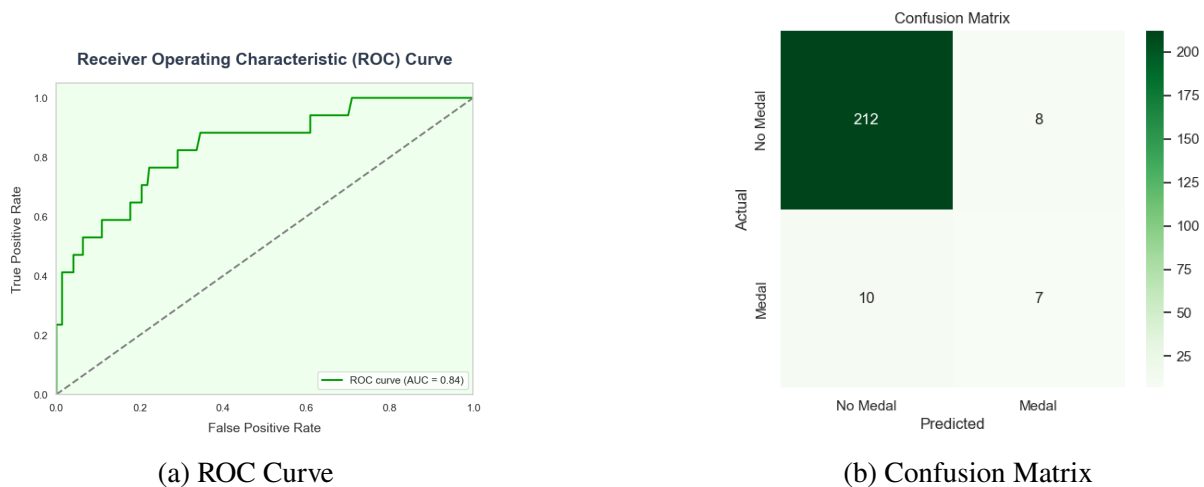


Figure 13: Model Performance Visualization

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 containing relevant statistical data for all countries that have never won a medal was generated 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.

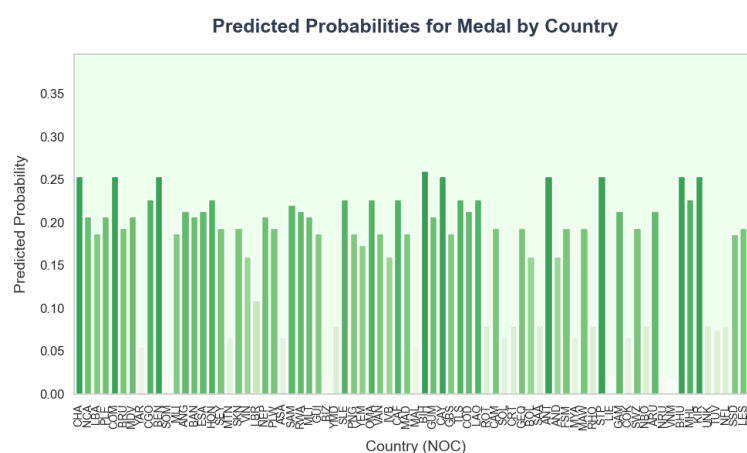


Figure 14: Predicted Probabilities for Medal by Country

7 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.

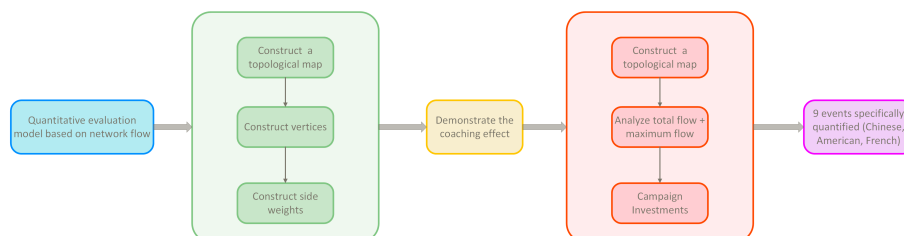


Figure 15: Quantitative Evaluation Framework Based on Network Flow Model

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 ($\Delta G, \Delta S, \Delta B$) for that country over a specific time period.

Node set V : Represents the participants and countries, $V = \{v_i | i \in I\}$ where I is the index set of countries.

- **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.

Edge set E : Represents the connections between nodes $E = \{(v_i, v_j) | i, j \in I\}$

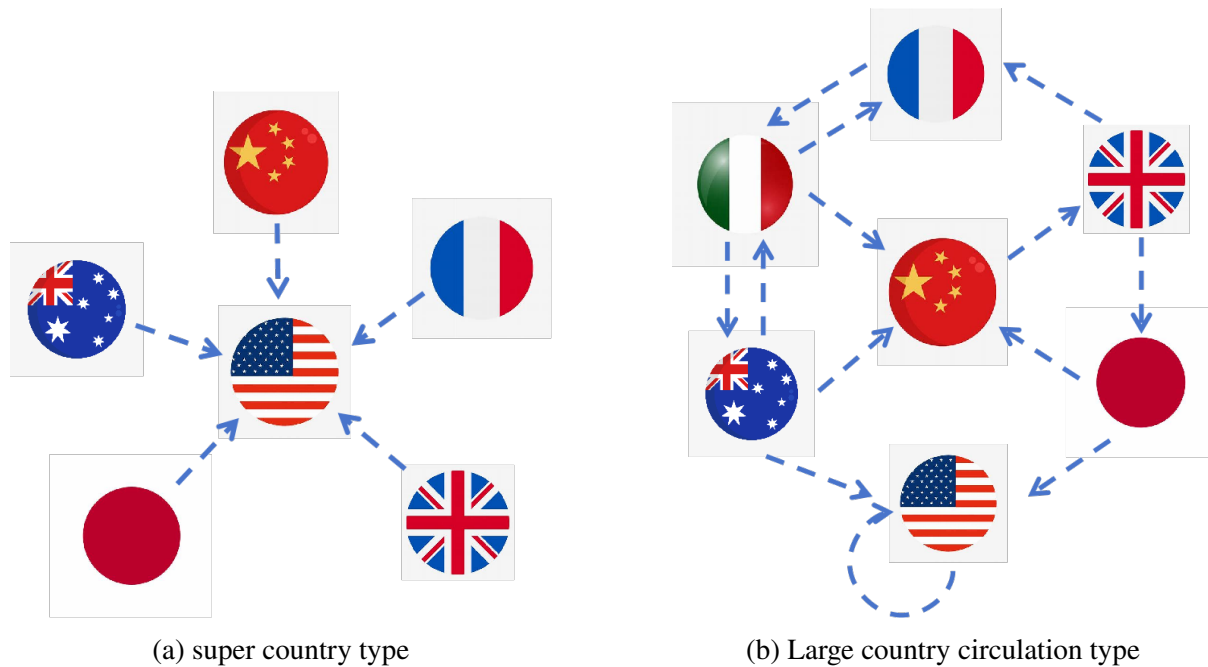


Figure 16: graph theory model

7.2.2 Weight Calculation

In this model, the weight w 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:

$$w = w_G \Delta G + w_S \Delta S + w_B \Delta B \quad (7)$$

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: $w_G=4, w_S=2, w_B=1$.

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:

$$w = 4 \times \Delta G + 2 \times \Delta S + 1 \times \Delta B \quad (8)$$

Coach Movement Between Different Countries For coaches moving 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 w of the directed edge.

$$W = W_A - W_B \quad (9)$$

- A is the inflow country;
- B is the outflow country;
- w_A is the weighted score of medal count changes in country A;
- w_B 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.

During the period of 2016, Lang Ping moved from the U.S. to China. The medal count

Sport	Coach	Flow Path	Total Flow	Bottleneck Flow
Volleyball	Lang Ping	2004→2008 (USA self-loop): Gold +0, Silver +1, Bronze +0, Score = 2	7	2

Sport	Coach	Flow Path	Total Flow	Bottleneck Flow
		2012→2016 (USA → China): USA Gold +0, Silver -1, Bronze +1; China Gold +1, Silver +0, Bronze +0, Score = 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.

(1)China: Table Tennis, Diving, Gymnastics

Sport	Coach	Traffic Path	Total Traf- fic	Bottleneck Flow
Table Tennis	Liu Guoliang	2004: Gold +1, Silver +1, Bronze +0, Score = -2 2004→2008: Gold +2, Silver +0, Bronze +0, Score = 8	6	-2
Diving	Zhou Jihong	2004→2008: Gold +1, Silver -1, Bronze +2, Score = 4 2008→2012: Gold -1, Silver +2, Bronze -2, Score = -2 2012→2016: Gold +1, Silver -1, Bronze +0, Score = 2	4	-2
Gymnastics	Huang Yubin	1996: Gold +1, Silver -1, Bronze +1, Score = 3 1996→2000: Gold -1, Silver +0, Bronze -1, Score = -5	-2	-5

Table 8: Sports Statistics

- **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.

(2)USA: Gymnastics, Basketball, Volleyball

Sport	Coach	Traffic Path	Total Traf- fic	Bottleneck Flow
Gymnastics	Bela Karolyi	2004: Gold +1, Silver +4, Bronze +0, Score = 12 2004→2008: Gold +1, Silver +1, Bronze +0, Score = 6 2008→2012: Gold +1, Silver -4, Bronze +0, Score = -4	14	-4
Basketball	Mike Krzyzewski	2004→2008: Gold +1, Silver -1, Bronze +0, Score = 2 2008→2012: Gold +0, Silver +0, Bronze +0, Score = 0 2012→2016: Gold +0, Silver +0, Bronze +0, Score = 0	2	0
Volleyball	Karch Kiraly	2008→2012: Gold +0, Silver +0, Bronze +0, Score = 0 2012→2016: Gold +0, Silver -1, Bronze +1, Score = -1 2016→2020: Gold +1, Silver +0, Bronze -1, Score = 3	2	-1

Table 9: Sports Statistics

- **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.

(3)France: Fencing, Basketball, Football

Sport	Coach	Traffic Path	Total Traf- fic	Bottleneck Flow
Fencing	Yu Ge Oubuli	2016: Gold +1, Silver +1, Bronze +1, Score = 7 "2020 (France → China): China Gold +1, Silver -1, Bronze -1; France Gold +1, Silver +1, Bronze +0, Score = -5"	2	-5
Basketball	Vincent Collet	2012→2016: Gold +0, Silver +0, Bronze +0, Score = 0 2016→2020: Gold +0, Silver +1, Bronze +0, Score = 2	2	0
Football	Thierry Henry	2016→2020: Gold +0, Silver +0, Bronze +0, Score = 0 2020→2024: Gold +0, Silver +1, Bronze +0, Score = 2	2	0

Table 10: Sports Statistics

- **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.

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 a great honour to write to you and share with you the results of our analysis of Olympic medal predictions. As a global platform for competition, the Olympic Games not only showcase the pinnacle of athletic excellence, but also foster international unity. In our mathematical modelling 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 under-researched sports.

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

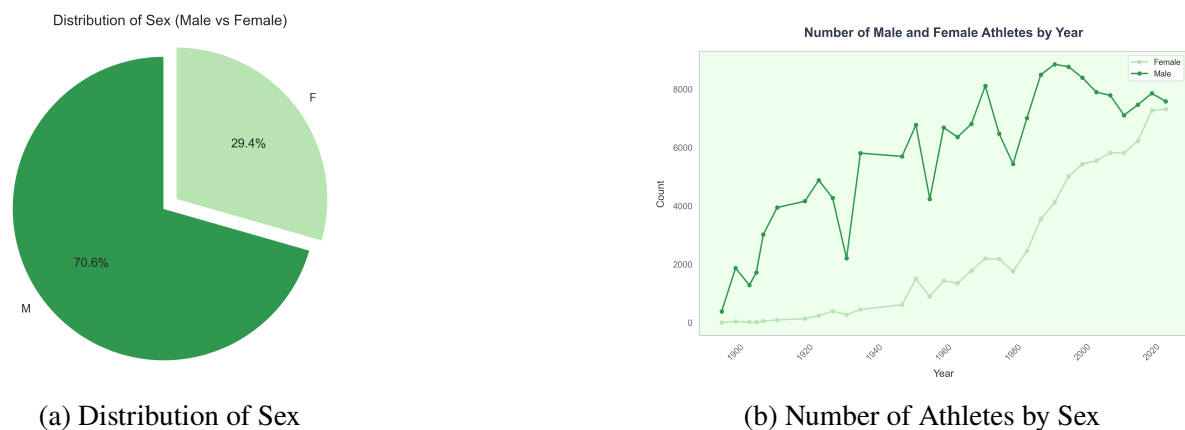


Figure 17: Player 1/2's Momentum in this set

9 Model Assessment

9.1 Strengths

1. 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. In 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.

References

- [1]Koç C K. Analysis of sliding window techniques for exponentiation[J]. Computers and Mathematics with Applications, 1995, 30(10): 17-24.
- [2]Hartigan J A, Wong M A. A k-means clustering algorithm[J]. Applied statistics, 1979, 28(1): 100-108.
- [3]Cohen I, Huang Y, Chen J, et al. Pearson correlation coefficient[J]. Noise reduction in speech processing, 2009: 1-4.
- [4]Dismuke C, Lindrooth R. Ordinary least squares[J]. Methods and designs for outcomes research, 2006, 93(1): 93-104.
- [5]Hochreiter S. Long Short-term Memory[J]. Neural Computation MIT-Press, 1997.
- [6]Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365.
- [7]Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.
- [8]Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.