Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2520861

title

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: $W = 3 \times \triangle G + 2 \times \triangle S + 1 \times \triangle B$. 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

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

2 Assumptions and Justification

3 Notations

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:

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

- **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

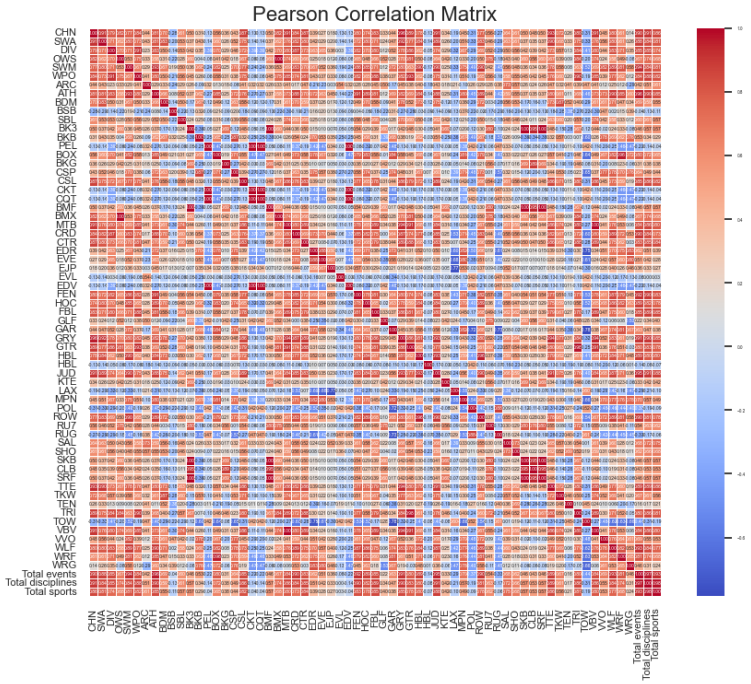


Figure 1: Heatmap of China

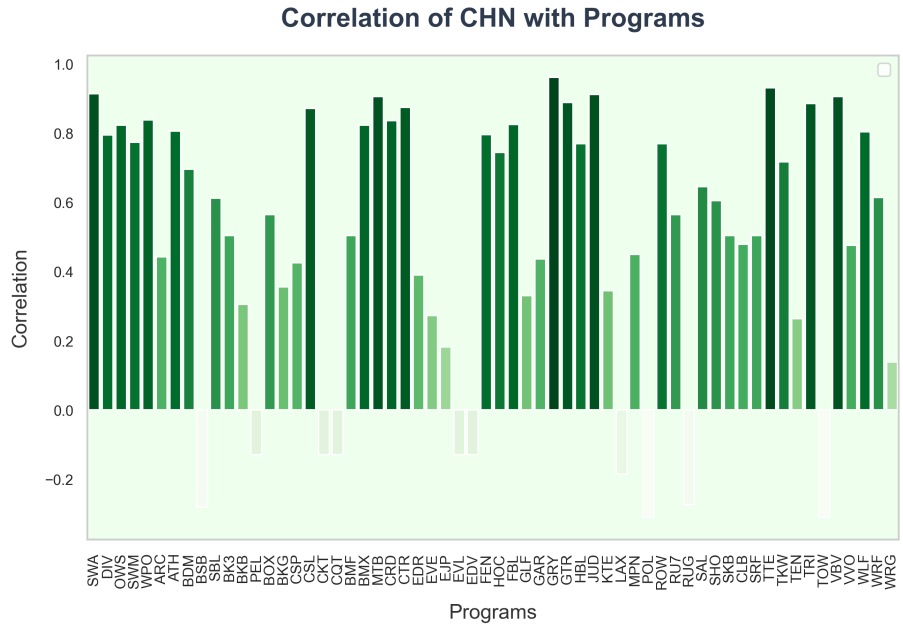


Figure 2: Correlation of CHN with Programs

- **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

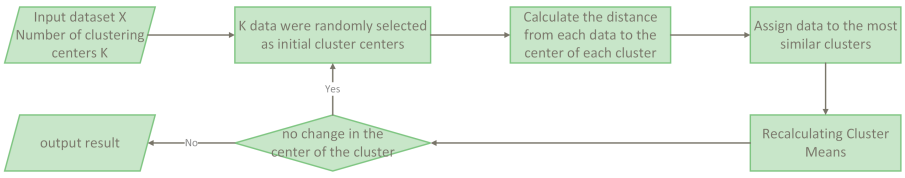


Figure 3: K-Means Flowchart

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

Feature Normalization: X' = MinMaxScaler

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

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

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

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.

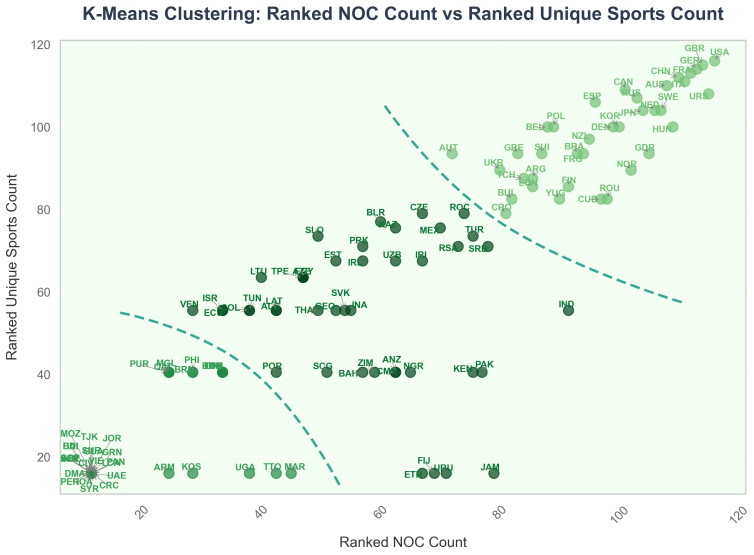


Figure 4: K-Means Clustering Diagram

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 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:

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

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

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

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

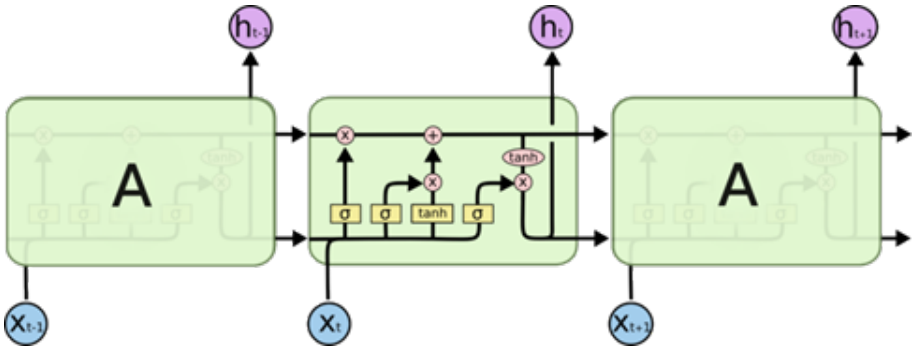


Figure 5: Workflow Diagram of the LSTM Mechanism

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.

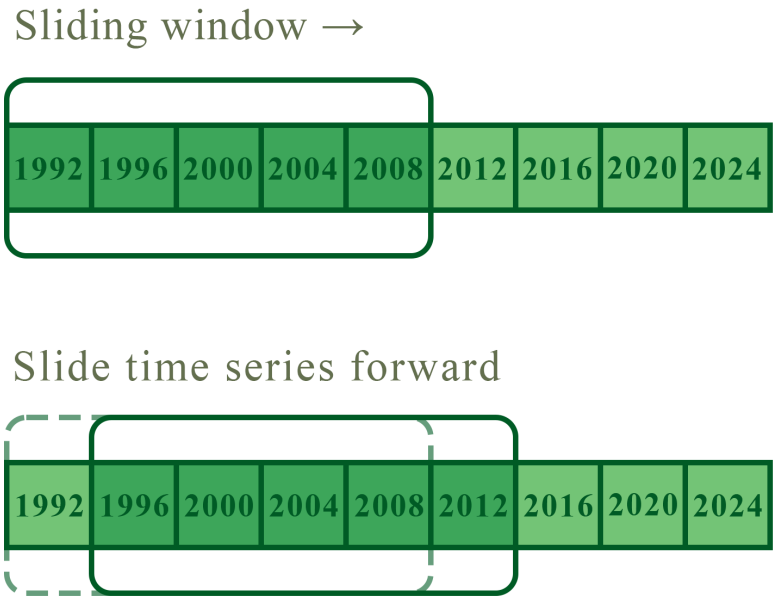


Figure 6: Sliding Window

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 $tweight = (1 - w) \times$ 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.

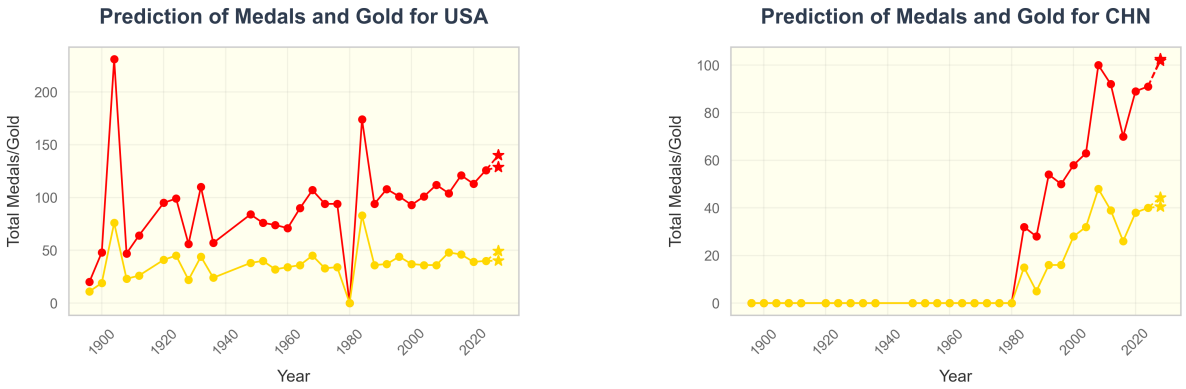
Model Architecture Design:

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.

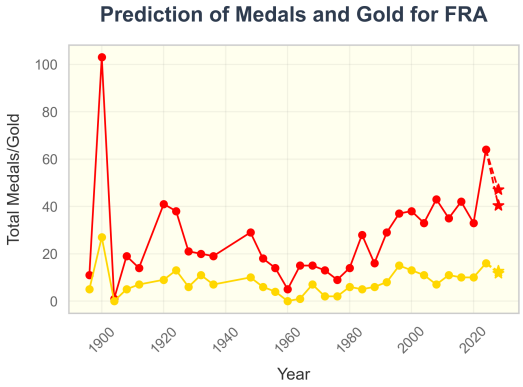
5.5 Prediction Interval from the Result

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.

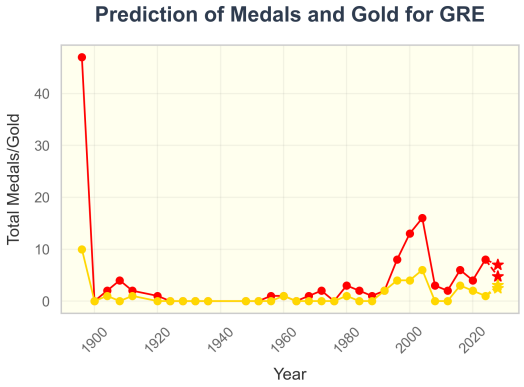


(a)

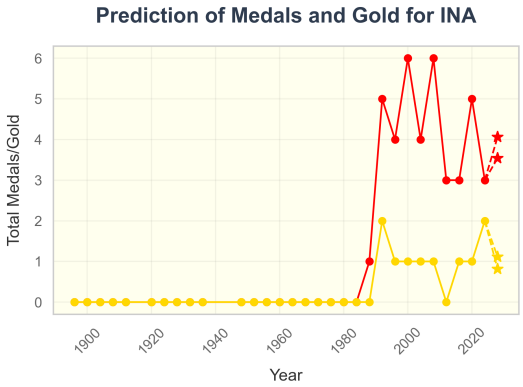
(b)



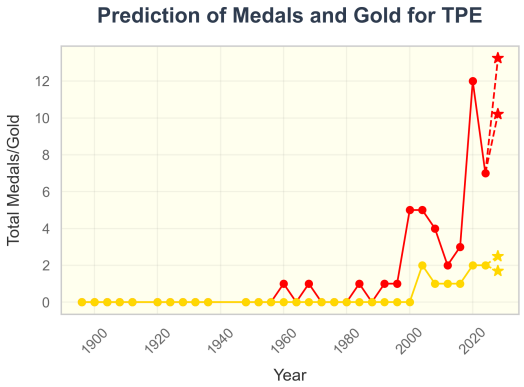
(c)



(d)



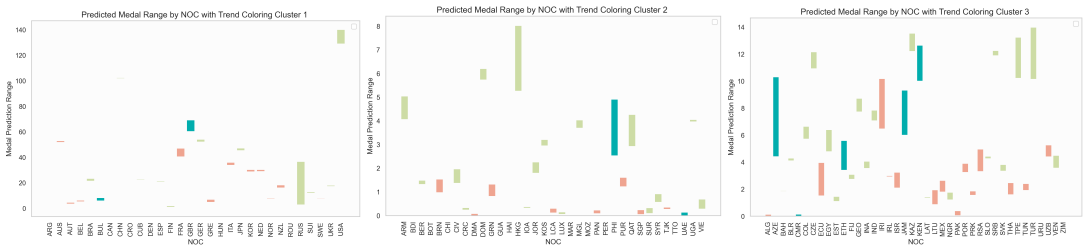
(e)



(f)

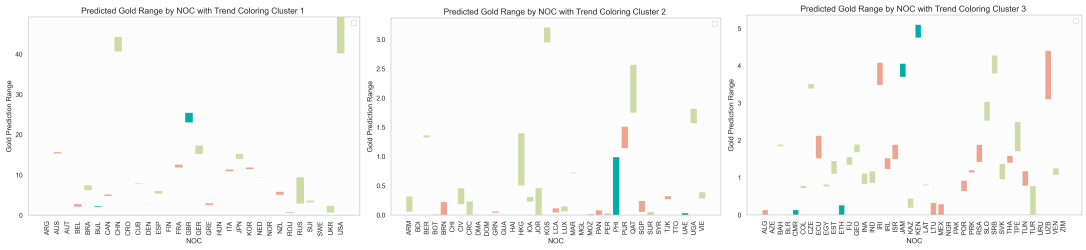
Figure 7: Prediction Intervals for Total Medals and Gold Medals of Different Countries

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.



(a) Category 1 (b) Category 2 (c) Category 3

Figure 8: Total Medal Predictions for Category 3 Countries



(a) Category 1 (b) Category 2 (c) Category 3

Figure 9: 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.

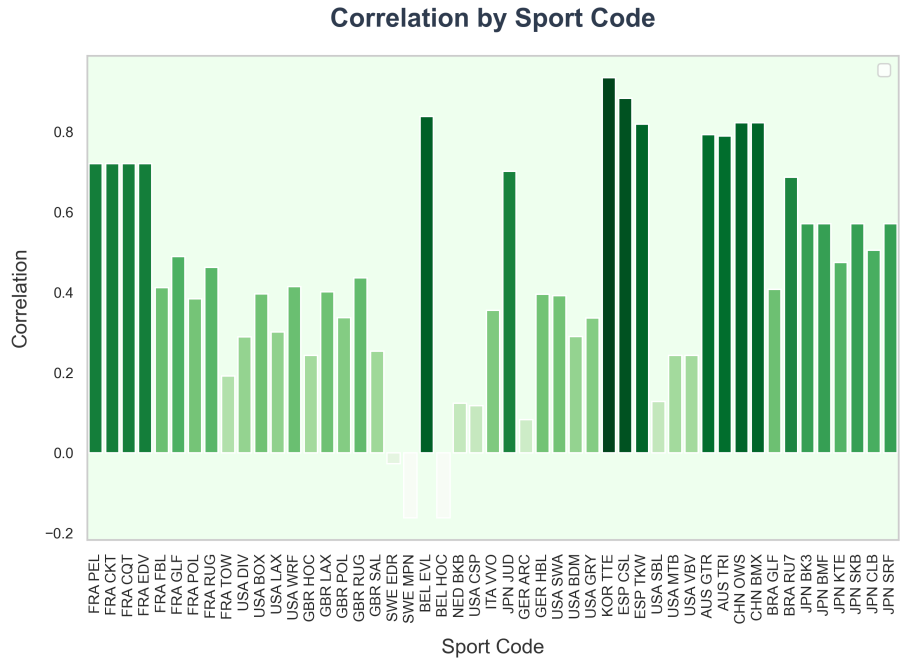
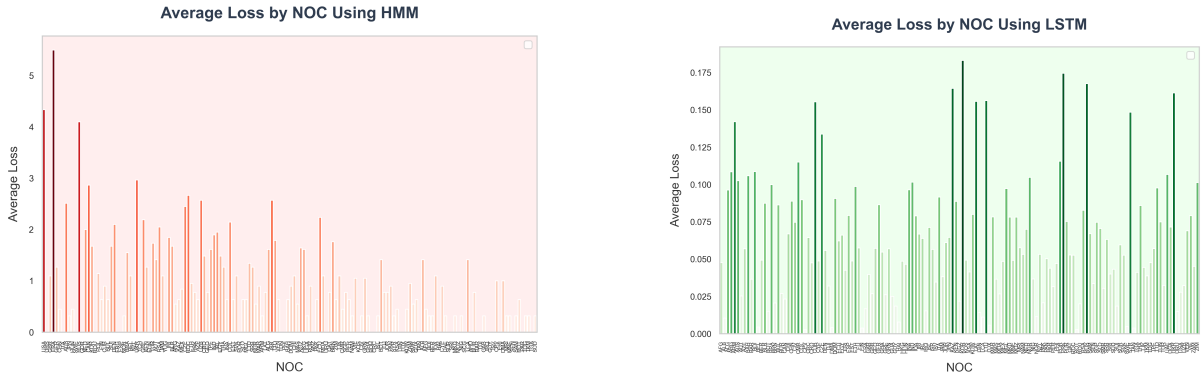


Figure 10: Correlation by Sport Code

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



(a) Average Loss by NOC Using HMM

(b) Average Loss by NOC Using LSTM

Figure 11: 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 TASK4: Quantitative Graph Theory Model of the "Great Coach Effect" Based on Network Flow

6.1 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

6.2 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.

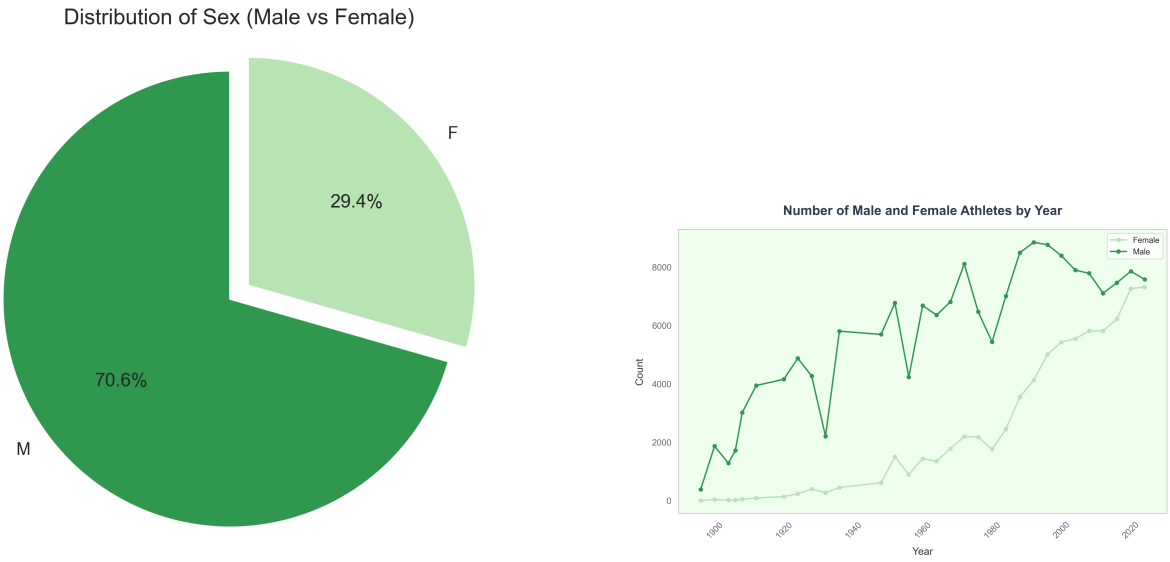
7 Letter

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



(a) Distribution of Sex

(b) Number of Athletes by Sex

Figure 12: Player 1/2's Momentum in this set

8 Model Assessment

8.1 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.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.

8.2 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.