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

title

Summary

abstract content...

Keywords: Momentum Analysis; Predictive Modeling; Random Forest; Sliding Window; Logistic Regression; Data Visualization; Generalization Capability

Contents

1 Introduction

1.1 Background

1.2 Literature Review

1.3 Restatement of the Problem

1.4 Our Work

2 Assumptions and Justification

3 Notations

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

4.1 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.

4.2 Data Preparation and Preprocessing

4.2.1 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.

4.2.2 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.

4.3 Model Construction

4.3.1 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 1: Table of Variables

Indicator	Descriptions
$participation\_count$	The number of participants from the country
$events\_participation\_count$	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).

4.3.2 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 2: Input parameters for forest models

Indicator	Descriptions
$NumberofEstimators$	[50, 100, 150, 200]
$MaxDepth$	[None, 10, 20, 30]
$MinSamplesLeaf$	[2, 5, 10]
$MinSamplesSplit$	[1, 2, 4]
$MaxFeatures$	['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 3: Optimal hyperparameter combination

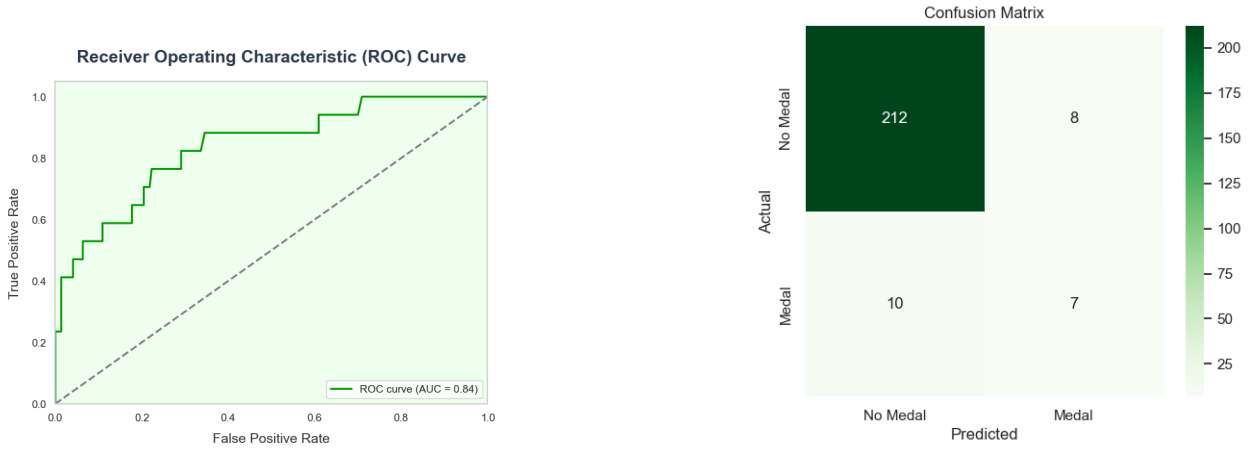
Indicator	Descriptions
$NumberofEstimators$	200
$MaxDepth$	10
$MinSamplesLeaf$	1
$MinSamplesSplit$	2
$MaxFeatures$	sqrt

4.3.3 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:



(a) ROC Curve (b) Confusion Matrix

Figure 1: 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.

4.3.4 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.

4.4 Model Testing

4.4.1 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.

4.4.2 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.

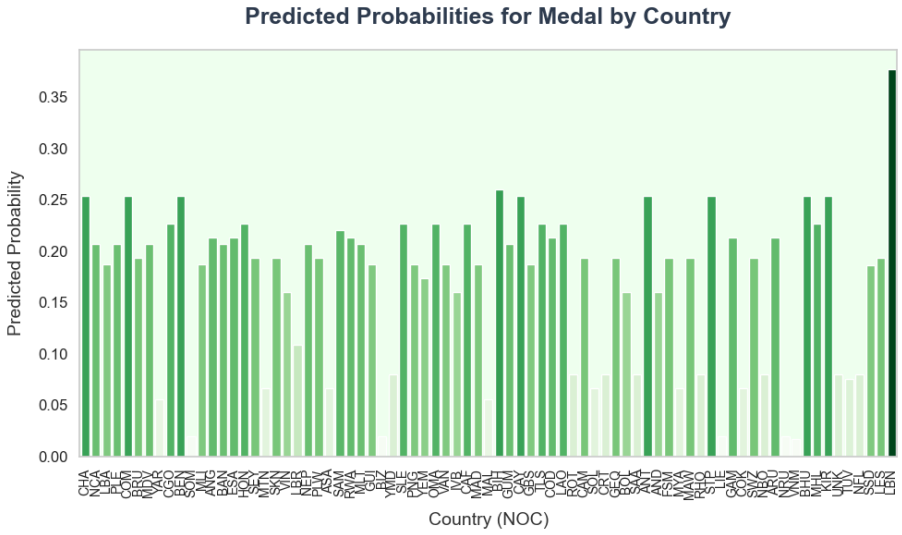
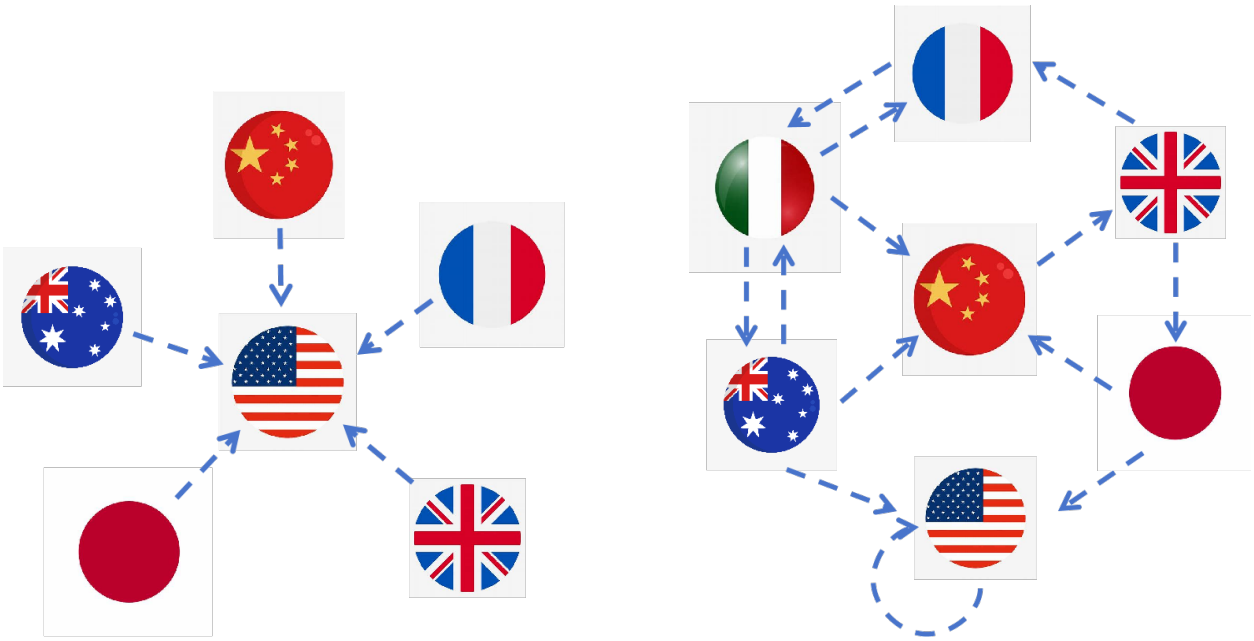


Figure 2: baseline of graphy

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

4.5.1 Model Background

4.5.2 Model Construction



(a) super country type (b) Large country circulation type

Figure 3: graph theory model

4.6 Letter

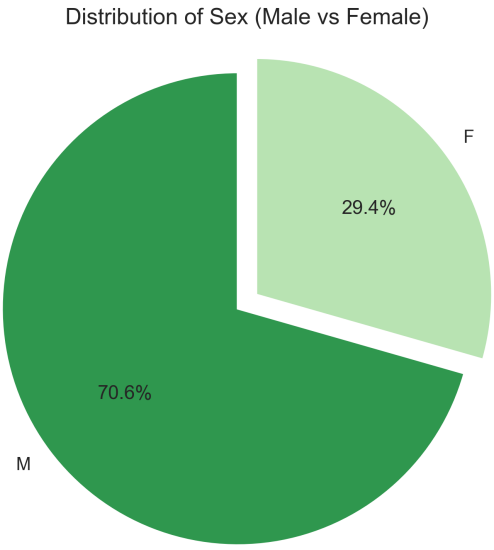
Dear Sir or Madam

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.

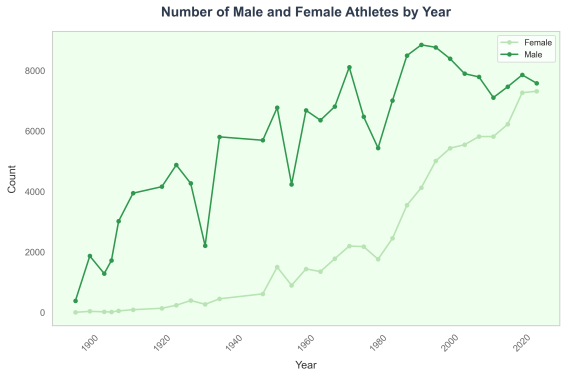
These insights not only reflect historical and current trends but also provide actionable strategic recommendations for National Olympic Committees in their preparations for future Olympic events.

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%), while in recent years, this proportion has gradually approached balance.



(a) Distribution of Sex



(b) Number of Male and Female Athletes by Year

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, the US

Olympic Committee has added five new sports for the 2028 Olympic Games in Los Angeles: 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 in depth to assess 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.

4.7 Model Assessment

4.7.1 Strengths

4.7.2 Weaknesses