Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
A		2504496

Enjoy a Cozy and Green Bath

Summary

abstract...

Keywords: Keyword one, Keyword two, Keyword three

Contents

1 Introduction

1.1 Background

The medal table of the 2024 Paris Olympics shows that the United States and China each won 40 gold medals and tied for the top spot, but the United States led with a total of 126 medals. The host country France ranked fifth in gold medals (16) and fourth in total medals (64). Dominica, Saint Lucia and other countries won their first Olympic medals, while 60 countries still have not broken through for any medals.

Figure 1: The medals of the 2024 Paris Olympics

1.2 Restatement and Analysis of the Problem

Based on the provided historical data-set of the Olympic Games from 1896 to 2024, we are employed to analyze and answer the following questions:

1. Develop a **prediction model** to forecast the number of medals each country will win in 2028, and identify countries that may progress or regress.
2. Provide **prediction intervals** and estimates of **uncertainty** and metrics to measure the model's performance.
3. Estimate the number of countries that will win their **first medal** and the probability of this happening.
4. Analyze the **relationship** between specific Olympic events (in terms of quantity and type) and the number of medals, explore which events are more important, and the impact of the host country's event selection strategy on the outcome.
5. Verify whether the **mobility of coaches** significantly enhances a country's performance in specific sports (such as Lang Ping and Bela Karolyi).
6. Quantify the contribution of **coaching effectiveness** to the number of medals, and recommend key sports for investment and expected returns for the three countries.
7. Extract the less-attended-to patterns from the model and provide strategic **suggestions** for the Olympic Committee.

For Task 1, we selected seven indicators and established an LSTM-based medal quantity prediction model, and provided interval predictions using Bayesian estimation. As for countries that have never won medals, we built an SVM-based ”first medal breakthrough” prediction model based on the new events, the number of athletes, and historical participation trends.

1.3 Overview of Our Work

2 Assumptions and Justification

To simplify the problem and make it convenient for us to simulate real-life conditions, we make the following basic assumptions, each of which is properly justified.

- **1**. ...
- **2**. ...

3 List of Notations

Symbols	**Description**
A_C, A_S	Set of country, all sports in Olympic.
A_T	$\{1, \ldots, 30\}$, representing the ordinal number of year Olympic held.
$A_H(t)$	Set of host country in year t.
$MG_{t,i,j,k}$	Number of gold medals country i won in sport j at event k in year t.
$MS_{t,i,j,k}$	Number of silver medals country i won in sport j at event k in year t.
$MB_{t,i,j,k}$	Number of bronze medals country i won in sport j at event k in year t.
$MT_{t,i}$	Number of total medals country i won in year t.
$N_{athletes}(t, i)$	Total number of athletes from country i in year t.
$N_{award}(t, i)$	Number of athletes who won medals from country i in year t.
$H(t, i)$	Host effect.
$y_{t,i,j,k}$	Logical variable of event j of sport i from i in year t.
$G_{\text{growth}}(t, i)$	Growth rate of the number of athletes from country i in year t.
$P_{Medal}(t, i)$	Probability of country i winning a medal in year t.
$P_{Gold}(t, i)$	Probability of country i winning a gold medal in year t.

Note: The Summer Olympics have been held for a total of 32 sessions.

4 Data Pre-processing

4.1 Outlier and Missing Value Handling

As the **1906 Intercalated Games** lacked the medal data of various countries and the competition results were not recognized by the International Olympic Committee, the data of 1906 is not taken into account.

In adition, **Skating** and **Ice Hockey** have been included in the Winter Olympics since 1920, so these two events are not within the scope of consideration. Otherwise, the ”·” is replaced by the number 0.

It was noticed that **Jeu de Paume** and **Roque** sports in the **summerOly_programs.csv** do not have Codes. Upon researching information from https://en.wikipedia.org/wiki/Jeu_de_ paume and https://en.wikipedia.org/wiki/Roque, it was found that only a few people are still engaged in these two sports, which have even not been held for 26 consecutive years in the Summer Olympics. Therefore, these two sports have been excluded.

5 Model 1: Prediction of Number of Medals for Medal-Winning Countries

5.1 Significance Analysis of Host Effect

Host Effect refers to the phenomenon where the host country or region performs more prominently in large-scale international events (such as the Olympic Games, the World Cup, etc.) due to its home field advantage. This phenomenon is usually reflected in a significant increase in the host country's medal count, competition results, or overall performance.

To verify the significance of the host effect, we employed a paired samples **t-Test**. Firstly, we selected the medal count of the host country in each year, denoted as MT_t, as the first sample. Secondly, to eliminate the influence of the overall growth trend in medal counts, we took the average of the medal counts obtained by the host country in the two consecutive Olympic Games as the second sample, shown as (1),

$$MT_t^H = \frac{MT_{t-1} + MT_{t+1}}{2} \tag{1}$$

where $t \in A_T^H = A_T \backslash \{1896, 2024\}$, $i \in A_C$.

Then $\{MT_t, MT_t^s\}$ constitutes paired data with a sample size of 30. Define $d_t = MT_t - MT_t^s$, and assume that

$$H_0 : \mu_d = 0, \quad vs \quad H_1 : \mu_d \neq 0.$$

Select the t-test statistic as

$$t = \frac{\bar{d}}{s_d/\sqrt{30}} \sim T(29) \tag{2}$$

where $\bar{d} = \frac{1}{30} \sum_{t \in A_T^H} d_t$ is the mean of paired samples, and $s_d = \frac{1}{29} \sum_{t \in A_T^H} \left(d_t - \bar{d}\right)^2$ is the sample variance of the differences of paired data,

For a given significance level α, the rejection domain for the hypothesis test is

$$W_\alpha = \left\{|t| \geq T_{1-\frac{\alpha}{2}}(29)\right\} \tag{3}$$

By following the described procedure, the results of the t-test were obtained and are summarized in Table **??**.

5.2 Index Analysis

5.2.1 Host effect

Define Logical Variable $H_{t,i}$ as equation (4),

$$H_{t,i} = \begin{cases} 1, & \text{Country } i \text{ is host in year } t, \\ 0, & \text{others.} \end{cases} \tag{4}$$

where $t \in A_T, i \in A_C$.

5.2.2 Events

The event vector $V(t)$ is defined as:

$$V(t) = \big(v_1(t), v_2(t), \ldots, v_M(t)\big)^T,$$

where: $v_i(t) = 1$ if event i is held in year t, $v_i(t) = 0$ if event i is not held in year t. Here, M represents the total number of distinct Olympic events considered up to year t, and the elements of $V(t)$ are binary values indicating the participation of each event in year t.

5.2.3 Formal Definition

For any event i and year t, we define $v_i(t)$ as follows:

$$v_i(t) = \begin{cases} 1 & \text{if event } i \text{ is held in year } t, \\ 0 & \text{if event } i \text{ is not held in year } t. \end{cases}$$

5.2.4 Definition of Dominant Event

Let $I_j(t)$ represent the dominance of event j in year t, where the dominance is calculated based on the medal count over the past three years and the total number of medals in year t.

$$I_j(t) = \frac{\sum_{q=t-3}^{t-1} MT_{q,i,k,j}}{\sum_{q=t-3}^{t-1} V_j(q) \cdot MTq, i, j, k}$$

where the formula calculates the dominant position of country i in year t for event j , determined by the ratio of past medal performance to the total number of medals in that year, adjusted for the presence of the event in the last three years.

5.2.5 Other indexes

5.3 LSTM Model

5.4 Uncertainty estimation Monte Carlo Dropout

In sports events, there are often unexpected incidents such as injuries, which are full of uncertainties for Olympics. The Monte Carlo Dropout (MC Dropout) can quantify uncertainty of the model Gal and Ghahramani, 2016.

The uncertainty is estimated by generating the distribution of predicted values through multiple random activation of the Dropout layer during the inference stage. By conducting multiple forward passes, the variance of the model's output is used to measure the confidence of the prediction.

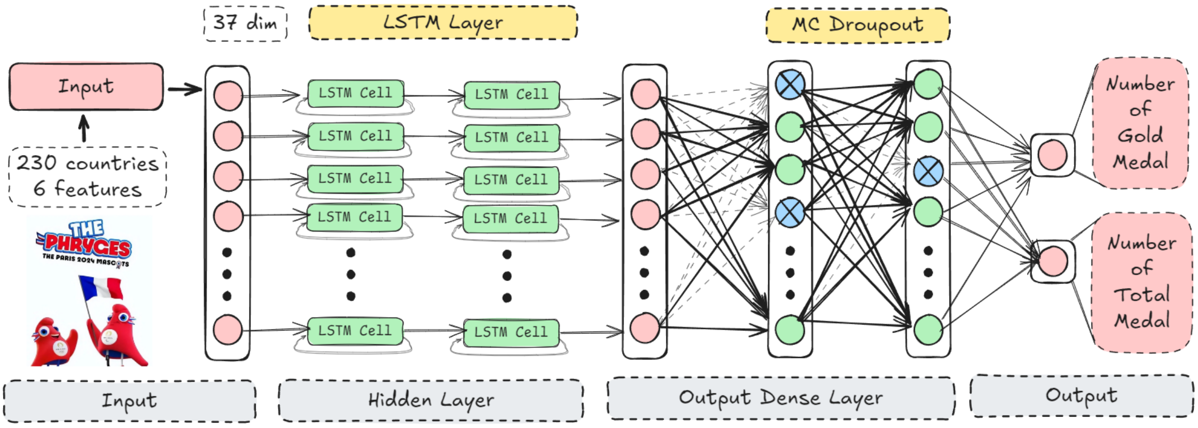

Figure 2: Flow of LSTM based on Monte Carlo Dropout

Assume $f(x;\theta)$ is the prediction model we build, x is its input and θ is parameters. In the training stage, Dropout operates with a probability p randomly dropping neurons is equivalent to sampling from the posterior distribution $P(\theta|D)$ (D are training data) of the parameters. In the inference stage, perform forward propagations 32 times before each time, generating a mask $\{f(x;\theta,m_t)\}_{t=1}^{32}$ each time. Then, the calculation of the predicted mean and variance is as follows:

$$\hat{y} = \frac{1}{32}\sum_{t=1}^{32} f(x;\theta,m_t)$$

$$Var(y) = \frac{1}{32}\sum_{t=1}^{32} \left(f((x;\theta,m_t)) - \hat{y}\right)^2$$

where $Var(y)$ reflects the predictive uncertainty of the model for the input x.

The results is shown as Figure **??**.

6 Model 2: Prediction of Maiden Medal for Medal-Less Countries

6.1 Problem Overview

The objective of this model is to predict whether countries that have never won a medal in the past (i.e., ”first-time winning countries”) will be able to win a medal in future Olympic Games. For these countries, traditional medal prediction models (which usually rely on historical medal data) may not effectively predict their future performance. Therefore, we need to consider other potential factors, such as the host country effect, athlete participation growth, and the addition of new events.

6.2 Index Analysis

6.2.1 Target variable

$$y(t,i) = \begin{cases} 1 & \text{if country } i \text{ wins a medal in year } t \\ 0 & \text{if country } i \text{ does not win a medal in year } t \end{cases}$$

6.2.2 Athlete growth rate

The growth rate of athletes from country i in year t, calculated by the following formula:

$$G_{\text{growth}}(t,i) = \frac{N_{\text{participate}}(t,i) - N_{\text{participate}}(t-1,i)}{N_{\text{participate}}(t-1,i)}$$

where $N_{\text{participate}}(t,i)$ is the number of athletes from country i in year t.

6.2.3 Participation in new events compared to previous years

$$N_{\text{new}}(t,i) = \begin{cases} 1 & \text{if } N_{\text{new}}(k,i) > 0, \quad k = t-8 \textit{ or } t-4 \textit{ or } t \\ 0 & \text{if } N_{\text{new}}(k,i) = 0, \quad k = t-8, t-4, t \end{cases}$$

where $N_{\text{new}}(t,i)$ represents the number of new events that country i participated in during year t. The term $\Delta N_{\text{new}}(t,i)$ indicates the change in the number of new events from the past three years to year t.

Prediction model: We use a Support Vector Machine (SVM) model to predict whether first-time winning countries will win medals. The model's input is the feature vector of the country $X(t,i)$:

$$X(t,i) = [G_{\text{growth}}(t,i), H(t,i), N_{\text{new}}(t,i)]$$

By training the SVM model, we can obtain the probability that country i will win a medal in year t:

$$P(\text{Medal}(t,i)) = f(X(t,i), \theta)$$

where $f(\cdot)$ is the decision function learned through SVM, and θ is the parameter learned during training.

Prediction of first-time medal probability: If country i has never won a medal (i.e., no historical medals), we use the prediction model to calculate the probability of winning a medal. If the probability exceeds a certain threshold, we predict that the country may win a medal in the future, especially a first-time medal:

$$P_{\text{first medal}}(t,i) = P(\text{Medal}(t,i)) > \text{Threshold}$$

where the threshold is determined based on the results of model evaluation.

7 Task 2: xxx

8 Task 3: xxx

9 Task 4: xxx

10 Sensitivity Analysis

11 Strength and Weakness

11.1 Strength

11.2 Weakness

12 Further Discussion

Memo

Enjoy Your Bath Time!

Sincerely yours,

Your friends

Reference

References

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. https://arxiv.org/abs/1506.02142

Appendices

Appendix A First appendix

Appendix B Second appendix

Report on Use of AI

1. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4,)

   Query1: <insert the exact wording you input into the AI tool>

   Output: <insert the complete output from the AI tool>

2. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4,)

   Query1: <insert the exact wording you input into the AI tool>

   Output: <insert the complete output from the AI tool>