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.
$G_{\text{growth}}(t,i)$	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 Task 1

5.1 Significance Analysis of Host Effect

Host Effect refers to the phenomenon where a host country tends to perform better in large-scale international events (such as the Olympic Games or the World Cup) due to the advantages associated with competing on home soil. This often manifests in a significant increase in the host country's medal count, competition results, and overall performance.

To assess the significance of the host effect, we employed a paired samples **t-Test**. First, we selected the medal count of the host country for each year, denoted as MT_t, as the first sample. To eliminate the influence of overall growth trends in medal counts, we used the average medal count from the two preceding Olympic Games as the second sample, as shown in equation (1),

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

where $t = 2, 3, \cdots, 29$, $i \in A_C$.

The data set $\{MT_t, MT_t^s\}$ then forms a paired sample with a 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

$$T = \frac{\bar{d}}{s_d/\sqrt{28}} \sim (27) \tag{2}$$

where $\bar{d} = \frac{1}{28}\sum_{t=2}^{29} d_t$ is the mean of paired samples, and $s_d = \frac{1}{27}\sum_{t=2}^{29}\left(d_t - \bar{d}\right)^2$ 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 Analysis of Key Indices

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 Event held

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 ($t = 1, 2, \cdots, 30$) and the elements of $V(t)$ are binary values indicating the participation of each event in year t.

5.2.3 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.

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

Next, define $I(t) = \big(I_1(t), I_2(t), \ldots, I_M(t)\big)^T$ as the dominance vector.

To get the modified dominance vector $I'(t)$, we set the components corresponding to the three largest values of $I(t)$ to 1, and all other components to 0:

$$\hat{I}(t) = \begin{cases} 1 & \text{if } j \in \text{Top3}(I(t)) \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{I}(t) = \mathbf{1}_{\{j \in \text{Top3}(I(t))\}},$$

where Top3$(I(t))$ refers to the indices corresponding to the three largest values in the vector $I(t)$, and $\mathbf{1}$ is the indicator function.

5.2.4 Strong Events

Let $\hat{I}(t)$ and $V(t)$ be the dominance vector and the event vector for year t, respectively. The number of strongpoints $S(t)$ can be defined as:

$$S(t) = \sum_{i=1}^{M} \mathbf{1}\{\hat{I}_i(t) = 1 \text{ and } v_i(t) = 1\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, which is 1 if the condition inside the curly brackets is true and 0 otherwise. In this context, $\hat{I}_i(t) = 1$ means that event i has a dominant position in year t, and $v_i(t) = 1$ indicates that event i is held in year t.

5.2.5 Percentage of winners

$$R(t, i) = \frac{N_{\text{award}}(t, i)}{N_{\text{athletes}}(t, i)}$$

5.2.6 Medal Distribution Concentration

$$\text{HHI}(t,i) = \sum_{j=1}^{M} \left(\frac{MT_{t,i,j}(t)}{MT_{t,i}(t)} \right)^2$$

where the closer the **HHI**(Herfindahl-Hirschmann Index)**HHI2016** is to 1, the more concentrated the distribution of medals is in a small number of sports, and the closer it is to 0, the more widely distributed the medals are.

5.2.7 historical performance

$$\widetilde{MT}(t,i) = \frac{1}{3} \sum_{q=t-3}^{t-1} MT_{q,i}$$

5.3 Prediction of Medal Count for Medal-Winning Countries Using LSTM

x x**Gal2015DropoutAA**

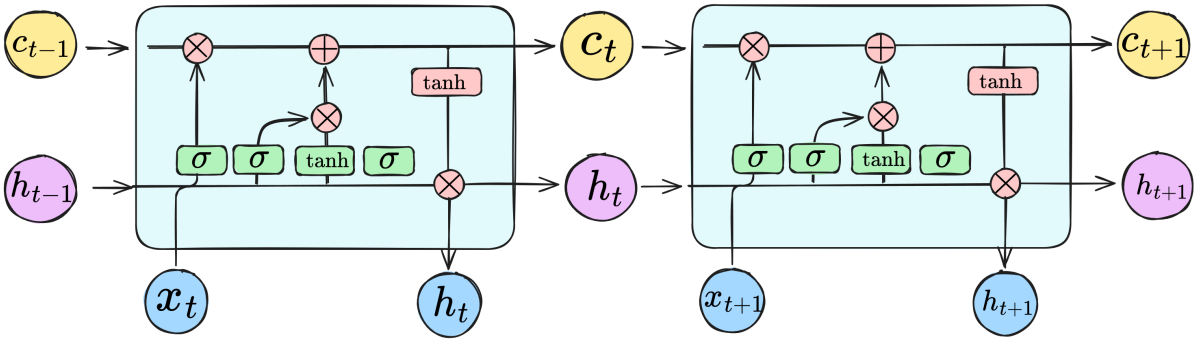

Figure 2: Flow of LSTM based on Monte Carlo Dropout

Table 1: LSTM Model Parameters Specification

Parameter	Description	Dimensions	Activation
W_f	Forget gate weight matrix	$[h_{t-1} + x_t]$	Sigmoid
W_i	Input gate weight matrix	$[h_{t-1} + x_t]$	Sigmoid
W_o	Output gate weight matrix	$[h_{t-1} + x_t]$	Sigmoid
W_c	Candidate cell state weights	$[h_{t-1} + x_t]$	Tanh
b_f, b_i, b_o, b_c	Bias vectors	$[1]$	–
h_t	Hidden state output	$[N]$	–
c_t	Cell state output	$[N]$	–
y_{pred}	Predicted medal count	Scalar	Linear
y_{true}	Actual medal count	Scalar	–
Init	Xavier initialization	$[N]$	–

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 **gal2016dropout**.

Algorithm 1 LSTM Medal Prediction with Uncertainty Quantification

Input: Historical sequence $X = [H(t,i), S(t), R(t,i), HHI(t,i), \widetilde{MT}(t,i), N_{\text{athletes}}(t,i)]$
Initialize: Parameters $\theta = \{W_f, W_i, W_o, W_c, b_f, b_i, b_o, b_c\}$
Initialize hidden state $h_0 \leftarrow \mathbf{0}$, cell state $c_0 \leftarrow \mathbf{0}$
Set dropout rate $p = 0.4$, Monte Carlo samples $M = 100$
for each $t = 1$ to T **do**
    Compute forget gate $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
    Compute input gate $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
    Compute candidate state $\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$
    Update cell state $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$
    Compute output gate $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
    Update hidden state $h_t = o_t \odot \tanh(c_t)$
end for
for each $m = 1$ to M **do**
    Enable dropout masks $\xi^{(m)} \sim \text{Bernoulli}(p)$
    Compute prediction $\hat{y}^{(m)} \leftarrow \text{LSTM}(X; \theta, \xi^{(m)})$
end for
Compute prediction mean $\mu_y = \frac{1}{M} \sum_{m=1}^{M} \hat{y}^{(m)}$
Compute standard deviation $\sigma_y = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{y}^{(m)} - \mu_y)^2}$
Compute 95% confidence interval $\text{CI}_{95\%} = [\mu_y - 1.96\sigma_y, \mu_y + 1.96\sigma_y]$
Compute loss $\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i^{\text{true}})^2$
Backpropagate gradients $\nabla_\theta \mathcal{L}$ via BPTT
Update parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ (where η is the learning rate)
Return: $\mu_y, \text{CI}_{95\%}$

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.

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 $\{f(x;\theta,m_t)\}_{t=1}^{30}$ 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 **??**.

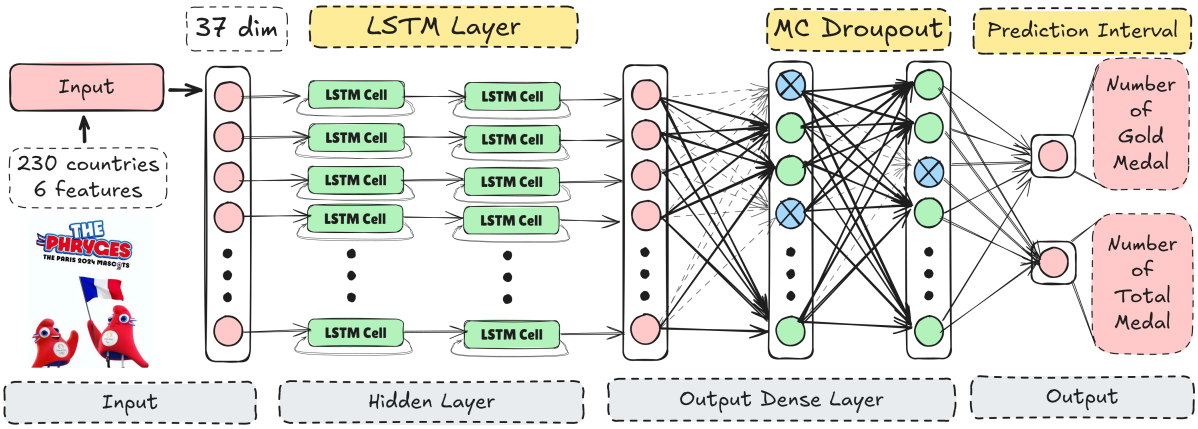

Figure 3: Flow of LSTM based on Monte Carlo Dropout

6 Task 2

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

where $N_{\text{new}}(t,i)$ represents the number of new events that country i participated in during year t, and $P_{\text{new}}(t,i)$ indicates the quantity of new events in which country i participated during year. We define $\tilde{N}_{\text{new}} = \min(N_{\text{new}}[t][i], P_{\text{new}}[t][i])$.

6.3 Prediction model:

We use a Random Forest (RF) classifier to predict whether first-time winning countries will win medals. The model's input is the feature vector of the country $X(t,i)$:

Algorithm 2 Participation in new events compared to previous years

Input: Year t, Country i, Matrices N_{new}, P_{new}
for each $k \in \{t-2, t-1, t\}$ **do**
  if $N_{\text{new}}[k][i] > 0$ **and** $P_{\text{new}}[t][i] > 0$ **then**
    Return $\min(N_{\text{new}}[t][i], P_{\text{new}}[t][i])$
  end if
end for
Return 0

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

The RF classifier consists of multiple decision trees, where each tree makes a prediction, and the final prediction is determined by majority voting from all the trees:

$$P(\text{Medal}(t,i)) = \frac{1}{T}\sum_{t=1}^{T} f_t(X(t,i), \theta_t)$$

where T is the number of trees in the forest, $f_t(\cdot)$ is the decision function of the t-th tree, and θ_t is the parameter learned during training for the t-th tree.

Prediction of first-time medal probability: If country i has never won a medal (i.e., no historical medals), we use the RF classifier 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:

$$P_{Medal}(t,i) > \text{Threshold}$$

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

Algorithm 3 Prediction with Random Forest for First-Time Medal

Input: Year t, Country i, Matrices $\tilde{N}_{\text{new}}$, P_{new}, Feature vector $X(t,i)$, RF with T trees.
Output: First-time medal prediction probability $P_{\text{first medal}}(t,i)$
Calculate feature vector $X(t,i) = [G_{\text{growth}}(t,i), H(t,i), N_{\text{new}}(t,i)]$
Initialize prediction sum: $P_{\text{sum}} = 0$
for each tree $t \in \{1, \ldots, T\}$ **do**
  Get prediction from tree t: $p_t = f_t(X(t,i), \theta_t)$
  Update prediction sum: $P_{\text{sum}} = P_{\text{sum}} + p_t$
end for
Calculate average prediction: $P_{Medal}(t,i) = \frac{P_{\text{sum}}}{T}$
if $P_{Medal}(t,i) > \text{Threshold}$ **then**
  Return 1 (Predicted to win first medal)
else
  Return 0 (Predicted not to win first medal)
end if

7 Task 3: xxx

8 Task 4: Effect of Great Coach

8.1 Test of Parallel Trend

We can see that the number of medals won by the US gymnastics team from 1896 to 1984 and 1984-2016 years in Figure **??**. To verify this conjecture, we conducted a parallel test on their medal counts.

8.2 Test of Great Coach Effect based on DiD

To seek evidence of the existence of the great coach effect, we employ the Difference-in-Differences (DiD) model to examine its impact.

The DiD model is a statistical method used to assess the causal effect of an intervention on an outcome variable. It estimates the intervention effect by comparing the performance differences between the experimental group and the control group before and after the intervention. The model equation is

$$Y_{i,t} = \alpha + \delta_t + \gamma \cdot Treat_i \cdot Post_t + \varepsilon_{i,t}, \tag{5}$$

where $Y_{i,t}$ represents team i's performance at time t (e.g., medal count). The model includes a constant term α, time fixed effects δ_t for common influences (e.g., 1980s gymnastics improvements), and an interaction term $Treat_i \cdot Post_t$ to capture **Béla Károlyi**'s impact as coach on the U.S. team. The coefficient γ measures the "great coach effect," and $\varepsilon_{i,t}$ is the error term.

By using $SPSS$, we obtained the estimated value of the regression coefficient $\hat{\gamma} = 2.1572$. To test the significance of γ, assume that

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma > 0$$

Select the test statistic:

$$T = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})} \sim t(30-4) \tag{6}$$

where $\text{SE}(\hat{\gamma}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i^2}$, $\hat{\varepsilon} = \hat{Y}_{i,t} - Y_{i,t}$. For a given significance level α, the rejection domain for the hypothesis test is

$$W_\alpha = \left\{|T| \geq t_{1-\frac{\alpha}{2}}(30-4)\right\} \tag{7}$$

The test results of regression coefficients were obtained and are summarized in Table 2. The sample falls within the rejection region $W_{0.975}$, so it can be concluded that the regression coefficient γ is significant, e.i. the impact of great coach Béla Károlyi for the USA gymnastics team is significant.

Table 2: Transposed Presentation of t-Test Results

	t-statistic	p-value	Critical value (α=0.05)	Test conclusion
Value	3.045	0.008	2.052	Reject null hypothesis

8.3

9 Task 5

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

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>