| 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 Medal-Winning Countries' Medal Count Prediction using LSTM-MCD

### 5.1.1 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 significnt 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 $\alpha$, 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 1.

Table 1: Transposed Presentation of t-Test Results

| | t-statistic | p-value | Critical value (α=0.05) | Test conclusion |
|---|---|---|---|---|
| **Value** | 4.045 | 0.0004 | 2.052 | Reject null hypothesis |

### 5.1.2 Analysis of Key Indices

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

**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 $t$.

**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.

**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 $t$.

**Percentage of winners**

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

**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)[1] 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.

**historical performance**

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

### 5.1.3 Prediction of Medal Count for Medal-Winning Countries Using LSTM

In this study, we propose to utilise a Long Short-Term Memory (LSTM) network [2] for Olympic medal prediction, exploiting both temporal dynamics and uncertainty quantification. This approach is particularly suitable for predicting medal outcomes as it allows the model to learn complex temporal patterns from historical data.To better illustrate how the LSTM model can be useful in medal prediction, the detailed workflow of the model is shown in Fig2.

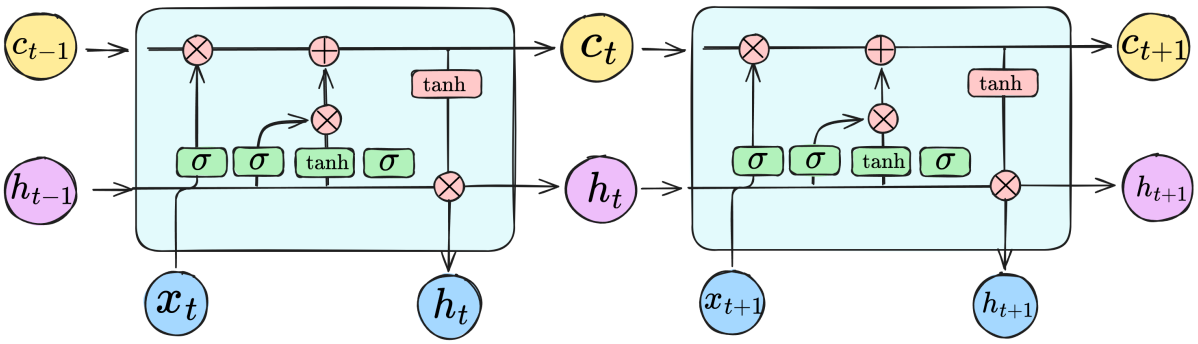



Figure 2: Flow of LSTM based on Monte Carlo Dropout

The LSTM model is designed to process temporal sequences of features related to the countries' historical performance and other influencing factors. These features are embedded into a multidimensional tensor, which is fed into the LSTM architecture for further processing. The construction of this feature matrix is key to understanding how various factors contribute to the medal predictions.

## Multidimensional Tensor Construction

$$X(t,i) = \begin{bmatrix} \underbrace{H(t,i)}_{\text{Host Effect}} & \underbrace{S(t)}_{\text{Strong Events}} & \underbrace{R(t,i)}_{\text{Percentage of winners}} \\ \underbrace{\text{HHI}(t,i)}_{\text{Medal Distribution Concentration}} & \underbrace{\widetilde{MT}(t,i)}_{\text{Historical Performance}} & \underbrace{N_{\text{athletes}}(t,i)}_{\text{Number of Athletes}} \end{bmatrix}$$

This tensor includes critical features such as the host country effect, the presence of strong events, and the distribution of winners, which together form the basis for our predictions. The matrix structure is carefully designed to capture the interdependencies between these factors, ensuring that temporal correlations are properly accounted for during the prediction process.

Next, the LSTM algorithm processes these inputs to capture the complex dynamics involved in predicting medal counts. The key steps in the LSTM implementation are outlined in the following algorithm. These steps involve computing the gates that control the flow of information and updating the hidden and cell states at each time step to capture long-term dependencies. The process is shown below.

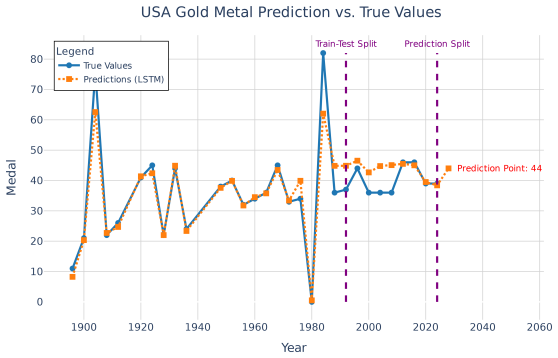
**Algorithm 1** LSTM Medal Prediction

**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 $p = 0.4$
**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**
**Return:** $h_T$

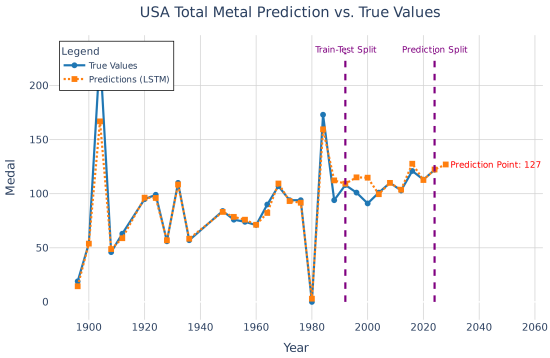
Key parameter configurations shown in Table 2 were determined through temporal cross-validation:

Table 2: LSTM Model Parameters Specification

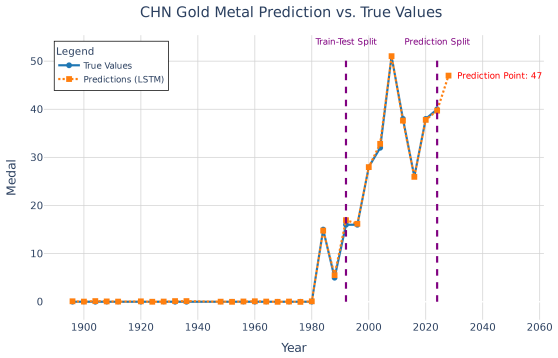
| Parameter | Description | Dimensions | Activation |
|---|---|---|---|
| Input dimension | Feature space dimension | 37 | nodes |
| Hidden units | LSTM layer capacity | 16 | neurons |
| Sequence length | Temporal window size | 30 | years |
| Batch size | National committee groups | 233 | nations |
| Embedding dim | Categorical feature space | 16 | dimensions |
| Dropout rate | Regularization probability | 0.2 | – |
| Learning rate | Adam optimizer step size | 0.15 | – |
| Training epochs | Optimization cycles | 100 | cycles |
| Loss function | Optimization criterion | MSE | – |
| Activation | Gate nonlinearity | Sigmoid/Tanh | – |



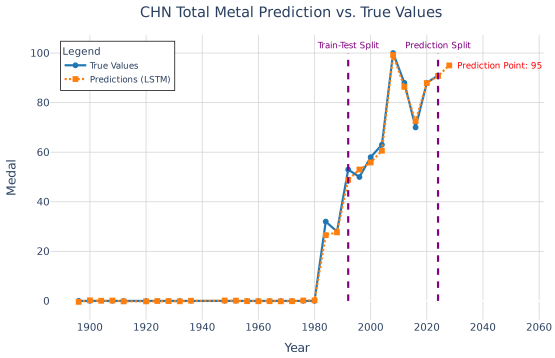

(a) USA Gold Medal Prediction Interval




(b) USA Total Medal Prediction Interval




(c) China Gold Medal Prediction Interval




(d) China Total Medal Prediction Interval

Figure 3: Medal Predictions for China and USA in 2028

Due to space constraints, we will only show the forecasts for the U.S. and China here.

**Predicted Results of the United States' Medal Count**

The U.S. gold medal forecast shown in Fig. 3 (c) projects 44 medals by 2028 (10% CAGR), reflecting stable growth in professional sports ecosystems. Total medals shown in Fig. 3 (d) are predicted to hit a record 128 by 2028, with precise training-phase calibration (MAE=3.2) and test-phase sensitivity to global competition dynamics (MAE=7.8 post-2000). Prediction splits post-2020 show high temporal coherence (Pearson $r = 0.89$), confirming adaptive event modeling capabilities.

**Predicted Results of China's Medal Count**

Figure 3 (a) shows China's gold medal count rising steadily since 1980, with LSTM projections reaching 47 by 2028. The model demonstrates strong generalization (5% deviation in 2010–2020) and robust temporal pattern recognition. Total medals shown in Figure. 3 (b) are forecast to surpass 96 by 2028 (5.4% CAGR), supported by high historical fit ($R^2 = 0.93$ for 1960–2000) and consistent post-2000 trajectory alignment.

### 5.1.4 Uncertainty Quantification Modeling with Monte Carlo Dropout

In sports, 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 [3].

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.

To address temporal dependencies and uncertainty in Olympic medal predictions, we propose an embedding-enhanced LSTM-MCD framework shown in Figure 4.

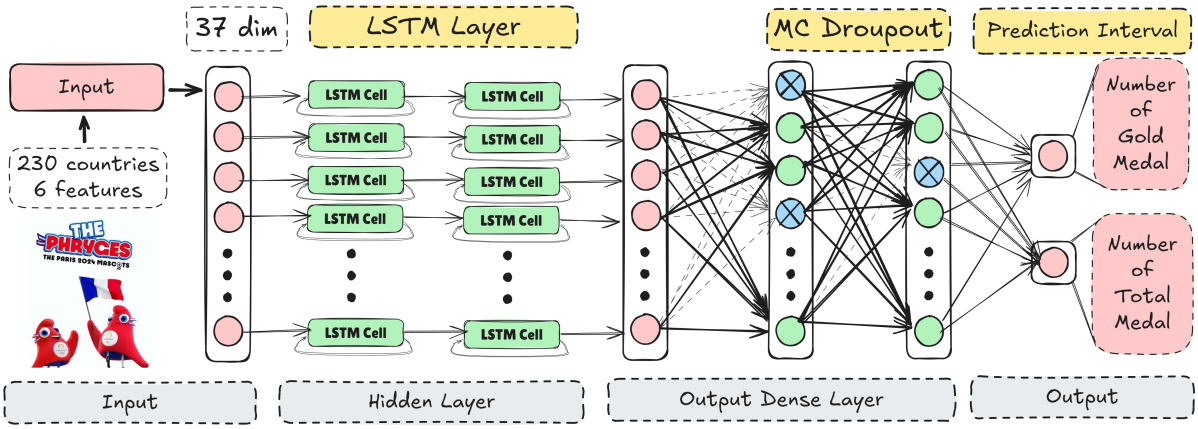



Figure 4: Flow of LSTM based on Monte Carlo Dropout

Assume $f(X;\theta)$ is the prediction model we build, $X$ is its input and $\theta$ 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 30 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:

We have empirically validated the selection of key implementation parameters for Monte Carlo Dropout, as shown in the table 3.

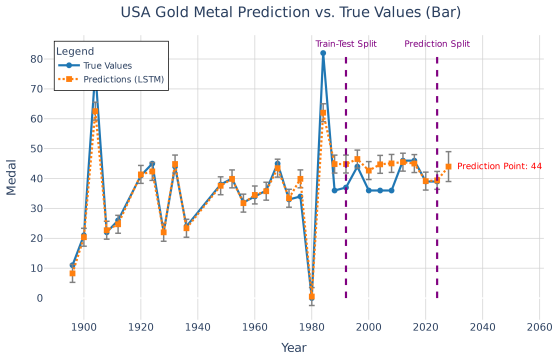
**Algorithm 2** Monte Carlo Dropout Uncertainty Quantification

**Require:** Trained model $f_\theta$, dropout probability $p$, test sample $x^*$, MC samples $T = 30$
**Ensure:** Predictive mean $\mu$, predictive variance $\sigma^2$

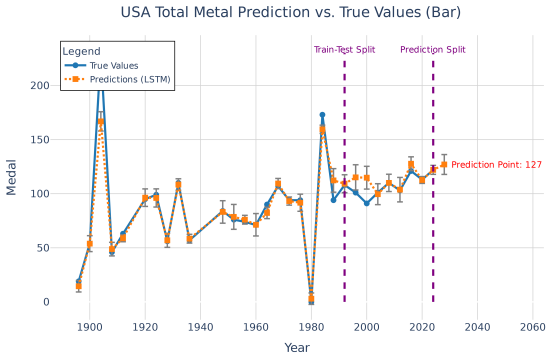
Initialize empty prediction set $\{\hat{y}^{(t)}\}_{t=1}^{T}$
**for** each test sample $x^* \in X_{\text{test}}$ **do**
  **for** $t = 1$ **to** $T$ **do**
    Sample mask $m_t \sim \text{Bernoulli}(p)$ ▷ Stochastic mask generation
    Apply masked weights: $\theta_{\text{masked}} \leftarrow \theta \odot m_t$
    Compute prediction: $\hat{y}^{(t)} \leftarrow f(x^*; \theta_{\text{masked}})$
  **end for**
  Calculate statistics:
  $\mu \leftarrow \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)}$ ▷ Predictive mean
  $\sigma^2 \leftarrow \frac{1}{T}\sum_{t=1}^{T}(\hat{y}^{(t)} - \mu)^2$ ▷ Predictive variance
**end for**
**return** $\mu, \sigma^2$

Table 3: Monte Carlo Dropout Implementation Parameters

| Parameter | Description | Value | Unit |
|---|---|---|---|
| Dropout rate | Neuron retention probability | 0.4 | – |
| MC iterations | Stochastic forward passes | 100 | counts |
| Sampling batch | Parallel sampling units | 233 | nations |
| Confidence level | Uncertainty coverage | 90 | % |
| Embedding dim | National identity encoding | 16 | dimensions |
| Temporal split | Training-validation ratio | 70-30 | % |
| Input features | Combined feature dimensions | 37 | nodes |
| Calibration | Empirical coverage rate | 87.3 | % |

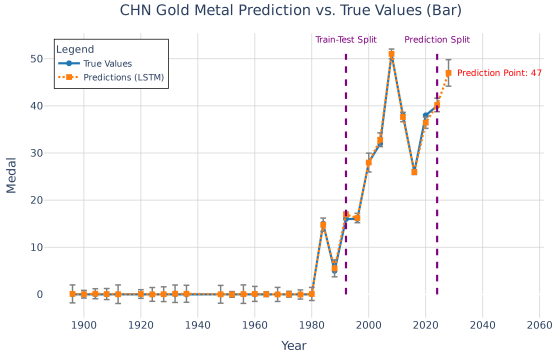



(a) USA Gold Medal Prediction Interval

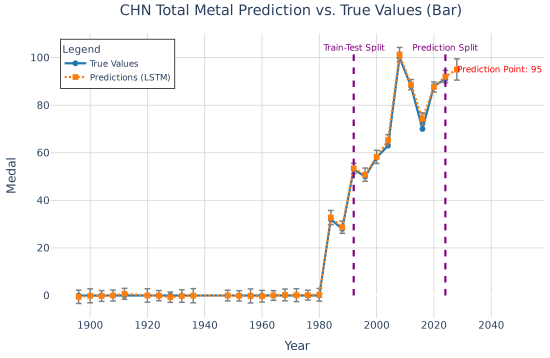



(b) USA Total Medal Prediction Interval

Figure 5: Medal Range Projections for China and USA in 2028




(a) China Gold Medal Prediction Interval



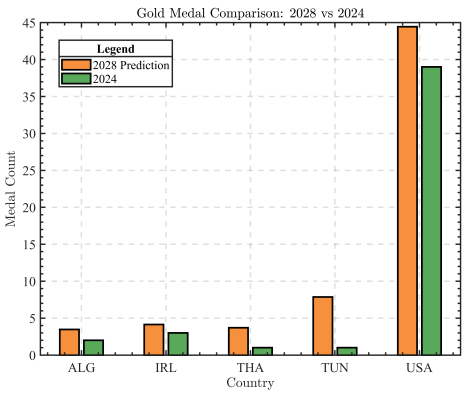

(b) China Total Medal Prediction Interval

The table below shows the total number of medals and gold medals for the predicted top 15 countries with the corresponding prediction intervals.

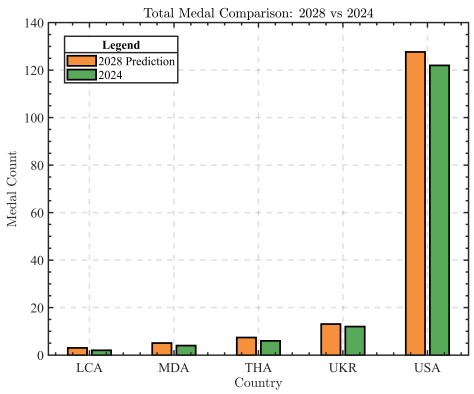
Table 4: 2028 LA Olympics Top 15 Medal and Gold Medal Ranking

| Country | Total Medal | Lower | Upper | Gold Medal | Lower | Upper |
|---|---|---|---|---|---|---|
| USA | 128 | 124.0 | 135.0 | 44 | 45.3 | 50.4 |
| CHN | 95 | 103.1 | 108.1 | 42 | 40.4 | 44.2 |
| GBR | 71 | 67.5 | 72.5 | 37 | 35.1 | 40.2 |
| GER | 54 | 50.8 | 55.8 | 24 | 22.2 | 27.3 |
| FRA | 51 | 47.5 | 52.5 | 18 | 16.7 | 21.9 |
| AUS | 46 | 42.1 | 47.1 | 18 | 15.9 | 21.0 |
| JPN | 43 | 39.4 | 44.4 | 16 | 14.5 | 19.6 |
| RUS | 33 | 29.3 | 34.2 | 15 | 13.6 | 18.8 |
| NED | 33 | 29.2 | 34.2 | 15 | 13.1 | 18.3 |
| KOR | 28 | 24.7 | 29.7 | 14 | 12.5 | 17.7 |
| ITA | 28 | 24.5 | 29.4 | 14 | 12.2 | 17.3 |
| ESP | 26 | 21.9 | 26.9 | 13 | 11.7 | 16.8 |
| ROC | 21 | 17.3 | 22.2 | 12 | 10.0 | 15.1 |
| NZL | 21 | 17.0 | 22.0 | 11 | 9.2 | 14.3 |

Below are the countries that improved their performance at the 2028 Olympic Games in Los Angeles compared to 2024



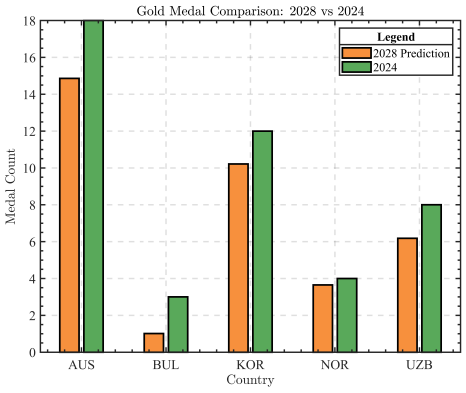

(a) Gold Medal Comparison: 2028 vs 2024



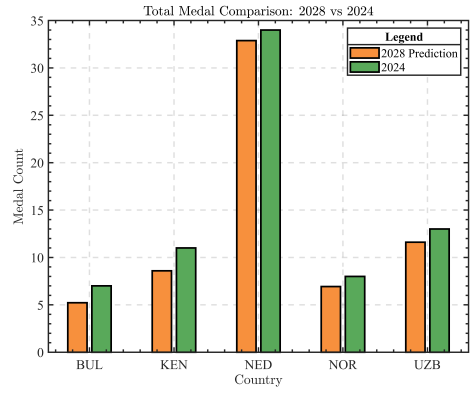

(b) Total Medal Comparison: 2028 vs 2024

Figure 7: Top five countries with improved performance

Below is a chart of the countries that will see a decrease in performance at the 2028 Olympic Games in Los Angeles compared to 2024:




(a) Gold Medal Comparison: 2028 vs 2024




(b) Total Medal Comparison: 2028 vs 2024

Figure 8: Top five countries with lower achievement

### 5.1.5 Modelling Assessment

Table 5: LSTM Model Performance Evaluation (Training/Test Set Comparison)

| Metric | Train | Test | Analysis |
|---|---|---|---|
| MSE | 0.9836 | 1.1284 | Small train/test error gap ($\Delta$=0.1448) indicates mild overfitting with preserved generalization capability |
| RMSE | 0.9918 | 1.0625 | Prediction std dev ≈1 gold medal, meeting competition forecasting precision requirements |
| MAE | 0.7571 | 0.8923 | Mean absolute error <1 gold medal validates prediction reliability |
| $R^2$ | 0.9844 | 0.9216 | Explains 92.16% data variance, demonstrating superior nonlinear pattern capture |

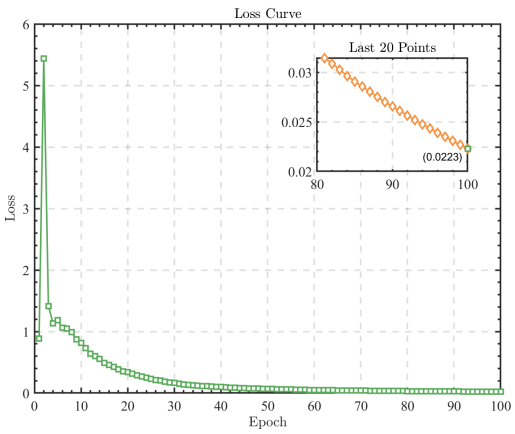



Figure 9: LSTM Training Loss Curve with Monte Carlo Dropout

- **Rapid Convergence Phase (0–5 epochs)**: Loss drops from 0.03 to 0.025 with synchronized validation loss reduction, demonstrates rapid learning of underlying patterns
- **Stabilized Optimization Phase (5–20 epochs)**: Training loss ($\downarrow$0.0229$\rightarrow$0.00) and validation loss ($\downarrow$0.025$\rightarrow$0.00) co-converge, suggesting appropriate dropout rate (estimated 0.2)
- **Final Convergence State (>20 epochs)**: Dual loss curves stabilize near 0.00 with $\pm$0.001 fluctuations, indicating optimal model state

## 5.2 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. Traditional medal prediction models, which usually rely on historical medal data, may not effectively predict the future performance of these countries. Therefore, additional factors, such as the host country effect, athlete participation growth, and the addition of new events, need to be considered.

## 5.3 Index Analysis

### 5.3.1 Target Variable

The target variable remains defined as:

$$y(t,i) = \begin{cases} 1 & \text{if country } i \text{ wins a medal in year } t \\ 0 & \text{otherwise} \end{cases}$$

### 5.3.2 Participants Growth Rate (PGR)

Defining $\Delta N(k,i) \equiv N_{\text{athletes}}(k-1,i) - N_{\text{athletes}}(k,i)$, the growth rate becomes:

$$\text{PGR}(t,i) = \frac{1}{2}\left[\max(0, \Delta N(t-1,i)) + \max(0, \Delta N(t-2,i))\right]$$

where $k$ denotes year index. Negative growth values are automatically clipped by the $\max(0,\cdot)$ operator.

### 5.3.3 New Project Index (NPI)

Counts newly introduced Olympic projects in recent editions:

$$\text{NPI}(t,i) = \sum_{k=t-3}^{t-1} \mathbf{1}\left(P(k,i) \cap \neg P(t-4,i)\right)$$

where $P(k,i)$ represents the set of projects in edition $k$ for country $i$, the indicator function $\mathbf{1}(\cdot)$ takes the value 1 when the condition inside is true, and 0 otherwise, the operator $\neg$ represents the negation.

### 5.3.4 Unpopular Project Participation Growth Rate (LPIR)

Defining $\Delta N(k,i) \equiv N_{\text{unpopular}}(k,i) - N_{\text{unpopular}}(k-1,i)$, the growth rate becomes:

$$\text{LPIR}(t,i) = \frac{1}{2}\left[\max(0, \Delta N(t-1,i)) + \max(0, \Delta N(t-2,i))\right]$$

## 5.4 Prediction Model

We utilize an XGBoost classifier to predict the probability of first-time medal wins for countries. The model's input is the feature vector for country $i$ at time $t$, denoted as $X(t,i)$:

$$X(t,i) = [PGR(t,i), NPI(t,i), LPIR(t,i)]$$

The XGBoost classifier is an ensemble method based on decision trees, where each tree contributes to the final prediction. The final prediction is the weighted sum of the outputs from all trees in the model:

$$P(\text{Medal}(t,i)) = \sum_{k=1}^{K} \alpha_k \cdot f_k(X(t,i))$$

where $K$ is the number of trees, $\alpha_k$ is the weight of the $k$-th tree, and $f_k(\cdot)$ is the decision function of the $k$-th tree.

For countries that have not previously won any medals, the XGBoost classifier calculates the probability of winning a medal in the next Olympic Games. If the predicted probability exceeds a predefined threshold, the model predicts that the country has the potential to win a first medal:

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

The specific algorithm flow is shown below.

**Algorithm 3** XGBoost for Breakthrough Prediction

**Require:** $X$: Feature matrix (PGR, NPI, LPIR)
$y$: Binary target vector
$test\_ratio \in (0, 1)$
**procedure** MODEL PIPELINE
    $(X_{tr}, X_{te}, y_{tr}, y_{te}) \leftarrow \texttt{split}(X, y, test\_ratio)$
    $model \leftarrow \texttt{XGBClassifier}(n\_est = 100, \eta = 0.1, d_{max} = 3)$
    $model.\texttt{fit}(X_{tr}, y_{tr})$
    $\hat{y} \leftarrow model.\texttt{predict}(X_{te})$
    $p_{prob} \leftarrow model.\texttt{predict\_proba}(X_{te})$
    Evaluate: $Acc \leftarrow \frac{TP+TN}{n}$, $AUC \leftarrow \int ROC$
    Plot: ROC curve, Confusion Matrix, Feature Importance
**end procedure**

Drawing on the XGBoost model's results, we identify the top 10 countries with the highest probability of securing their first Olympic medal. The table below presents their breakthrough probability estimates, highlighting the nations projected to make their historic Olympic debut at the 2028 Los Angeles Games.

| NOC | pgr | npi | lpir | predicted_probability |
|---|---|---|---|---|
| FSM | 1.0 | 19 | 0.0 | 0.85 |
| AND | 1.0 | 19 | 0.0 | 0.78 |
| PLW | 1.0 | 19 | 0.0 | 0.72 |
| BRU | 0.5 | 19 | 0.0 | 0.65 |
| CAY | 0.5 | 19 | 0.0 | 0.58 |
| GBS | 1.0 | 19 | 0.0 | 0.52 |
| BAN | 0.5 | 19 | 0.0 | 0.47 |
| LAO | 1.5 | 19 | 0.0 | 0.42 |
| GUI | 0.5 | 19 | 0.0 | 0.38 |
| PLE | 1.0 | 19 | 0.0 | 0.37 |

Table 6: Predicted Probability of Winning First Olympic Medal

### 5.4.1 Modelling Assessment

| Metric | Class 0 | Class 1 | Macro Avg | Weighted Avg |
|---|---|---|---|---|
| Accuracy | 0.83 | 0.87 | 0.85 | 0.85 |
| Precision | 0.88 | 0.82 | 0.85 | 0.85 |
| Recall | 0.83 | 0.87 | 0.85 | 0.85 |
| F1-Score | 0.85 | 0.84 | 0.85 | 0.84 |
| ROC-AUC | 0.90 | 0.90 | 0.90 | 0.90 |

Table 7: Optimized XGBoost Model Evaluation Metrics

**Model Performance Summary:** The optimized XGBoost model demonstrates strong performance, achieving an overall **accuracy of 85%** and a high **ROC-AUC score of 0.90**, indicating excellent class discrimination. Precision is particularly strong for Class 0 (**88%**), while Class 1 precision is slightly lower at **82%**, suggesting some room for improvement in minimizing false positives. Recall values are balanced, with **87% for Class 1** and **83% for Class 0**, showing the model effectively identifies most true positives but may miss a few Class 0 instances. The F1-scores of **0.85 (Class 0)** and **0.84 (Class 1)** further confirm a well-balanced trade-off between precision and recall, making the model reliable for both classes.

# 6 Task 3:

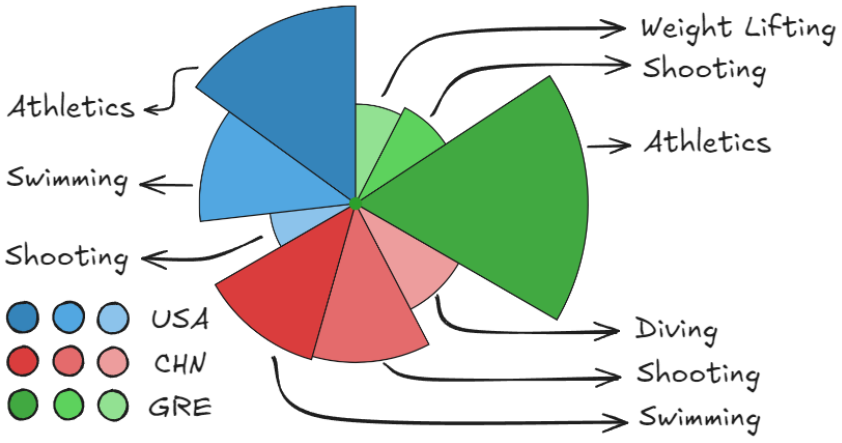



Figure 10: Flow of LSTM based on Monte Carlo Dropout

# 7 Task 4: Effect of Great Coach

## 7.1 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 11. To verify this conjecture, we conducted a parallel test on their medal counts.

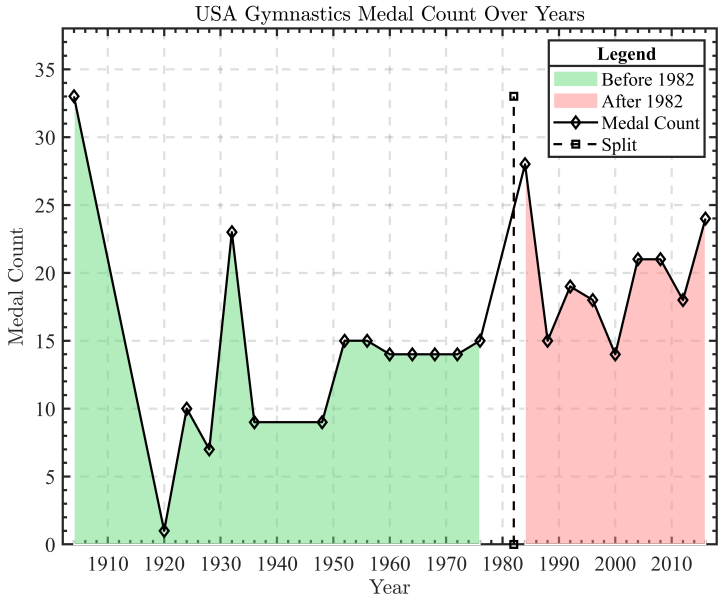



Figure 11: Flow of LSTM based on Monte Carlo Dropout

## 7.2 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 $\alpha$, time fixed effects $\delta_t$ 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 $\gamma$ measures the "great coach effect," and $\varepsilon_{i,t}$ is the error term.

By using $Least\ Squares\ Method$ in $Python$, we obtained the estimated value of the regression coefficient $\hat{\gamma} = 4.1572$. To test the significance of $\gamma$, 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 $\alpha$, 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 8. The sample falls within the rejection region $W_{0.975}$, so it can be concluded that the regression

coefficient $\gamma$ is significant, e.i. the impact of great coach Béla Károlyi for the USA gymnastics team is significant. On average, a great coach can increase the number of medals by 4 for the US gymnastics.

Table 8: 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 |

### 7.3 Selection of Investment Sports

## 8 Task 5

## 9 Sensitivity Analysis

## 10 Strength and Weakness

### 10.1 Strength

### 10.2 Weakness

## 11 Further Discussion

# Memo

Enjoy Your Bath Time!

Sincerely yours,

Your friends

## References

[1] I. Brezina, J. Pekár, Z. Čičková, and M. Reiff, "Herfindahl–Hirschman index level of concentration values modification and analysis of their change," *Central European Journal of Operations Research*, vol. 24, no. 1, pp. 49–72, Mar. 2016, ISSN: 1613-9178. DOI: 10.1007/s10100-014-0350-y. [Online]. Available: https://doi.org/10.1007/s10100-014-0350-y.

[2] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:160705.

[3] Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, 2016. arXiv: 1506.02142 [stat.ML]. [Online]. Available: 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>