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.

1.3 Overview of Our Work

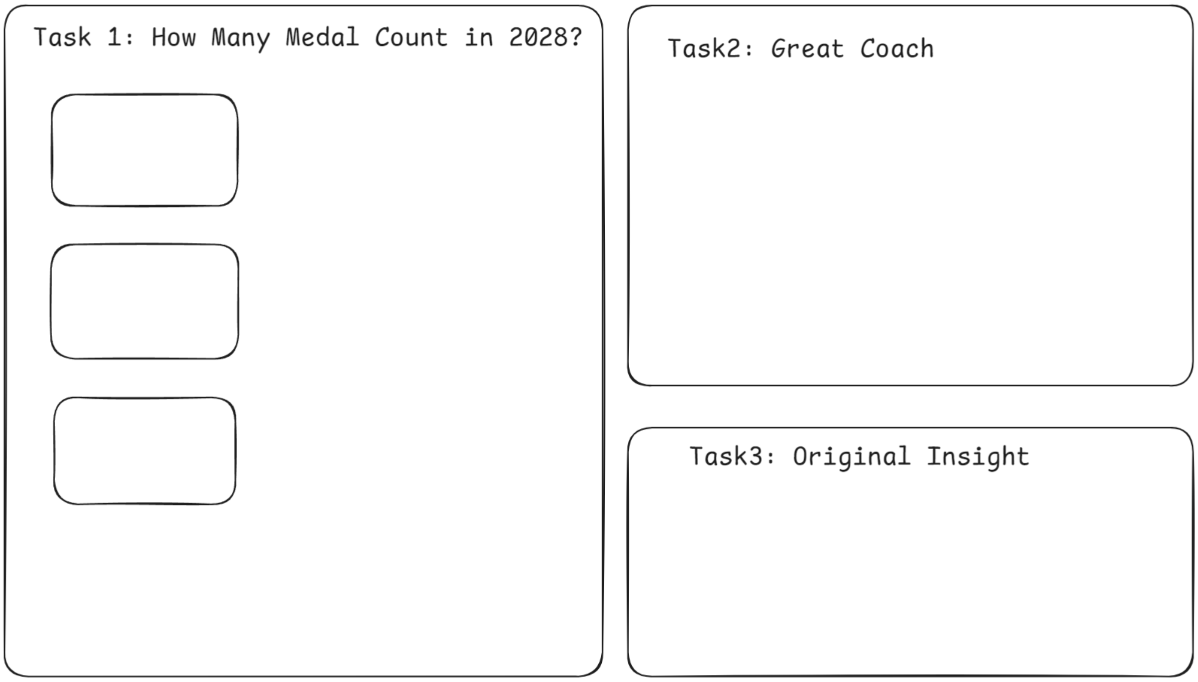


Figure 2: Overview of Our Work

2 Assumptions and Justification

1. **Historical medal data exhibits temporal dependencies that reflect future medal trends.**

   This suggests that historical performance can offer insights into future outcomes, and thus, should be treated as a time series when making predictions.

2. **Monte Carlo Dropout approximates Bayesian inference by quantifying prediction uncertainty through multiple stochastic samplings.**

   This technique provides a robust mechanism for estimating confidence intervals and is useful in scenarios with incomplete or noisy data.

3. **Historical data distributions of non-medal-winning countries align with those of future potential medal-winning nations.**

   This assumption supports the idea that non-medal-winning countries have similar characteristics to those that may perform well in future Olympics, making them a valuable reference for predicting future medal potential.

4. **The impact of coaching remains independent of confounding variables (e.g., athlete training conditions, changes in international competition rules).**

   This assumption isolates the effect of coaching from other factors that might influence performance, ensuring that coaching effects can be accurately assessed.

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.
$A_E(j)$	Represents the set of events inside the sport j.
$A_H(t)$	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: How Many Medal Count in 2028?

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 (5.1.1),

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

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

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

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

$$W_\alpha = \left\{|T| \geq t_{1-\frac{\alpha}{2}}(29)\right\}$$

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 the logical variable $H_{t,i}$ as shown in equation (5.1.2):

$$H(t,i) = \begin{cases} 1, & \text{if Country } i \text{ is the host in year } t, \\ 0, & \text{otherwise.} \end{cases}$$

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

- **Event Held**
  The event vector $V(t)$ is defined as:

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

where $v_i(t) = 1$ if event i is held in year t, and $v_i(t) = 0$ if event i is not held in year t. M represents the total number of distinct Olympic events considered up to year t ($t = 1, 2, \ldots, 30$).

- **Definition of Dominant Event**
  Let $I_j(t)$ represent the dominance of event j in year t, where 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 $I(t) = (I_1(t), I_2(t), \ldots, I_M(t))^T$ as the dominance vector.
To obtain 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:

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

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:

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

where $\mathbf{1}\{\cdot\}$ is the indicator function, which is 1 if the condition inside the curly brackets is true and 0 otherwise.

- **Percentage of Winners**
  The percentage of winners in year t for country i can be defined as:

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

where $N_{\text{award}}(t, i)$ is the number of awards won by country i in year t, and $N_{\text{athletes}}(t, i)$ is the number of athletes representing country i.

- **Medal Distribution Concentration**
  The Herfindahl-Hirschman Index (HHI) for the medal distribution concentration can be defined as:

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

where HHI approaches 1 when medals are concentrated in a small number of events, and approaches 0 when medals are distributed widely across many events.

- **Historical Performance**
  The historical performance of country i in year t can be calculated as the average medal count over the past three years:

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

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 Fig3.

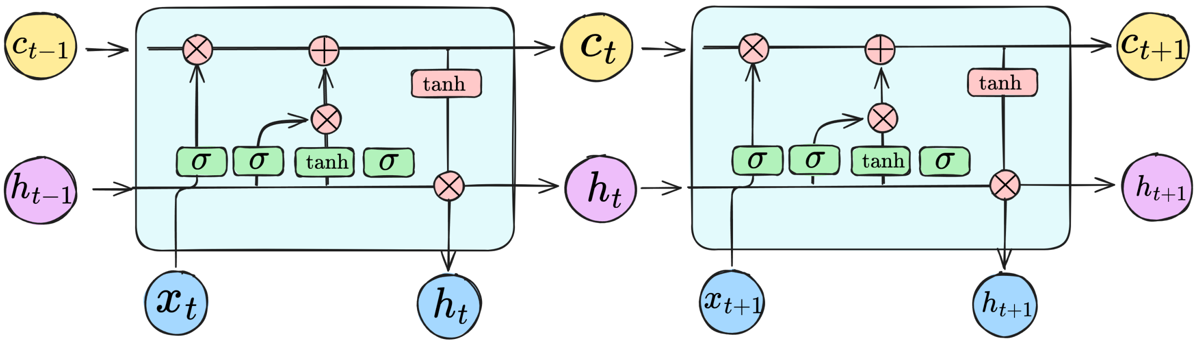


Figure 3: Flow of LSTM based on Monte Carlo Dropout

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

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

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.

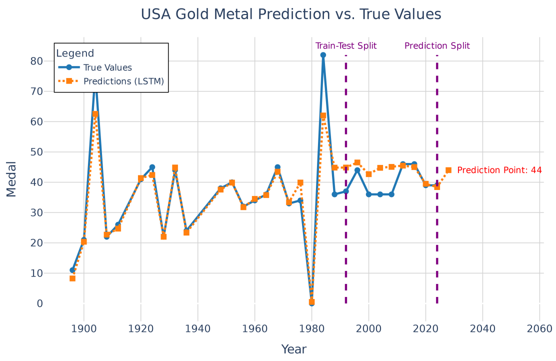
Algorithm 1 LSTM Medal Prediction

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

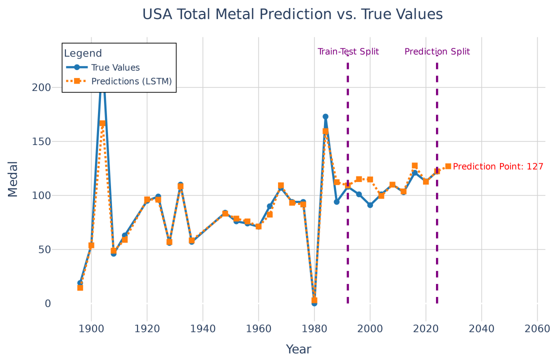
Key parameter configurations shown in Table 2 were determined through temporal cross-validation:

Table 2: LSTM Model Parameters Specification

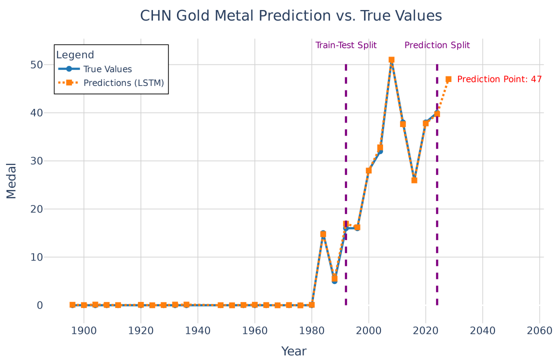
Parameter	Description	Dimensions	Activation
Input dimension	Feature space dimension	37	nodes
Hidden units	LSTM layer capacity	16	neurons
Sequence length	Temporal window size	30	years
Batch size	National committee groups	233	nations
Embedding dim	Categorical feature space	16	dimensions
Dropout rate	Regularization 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	–



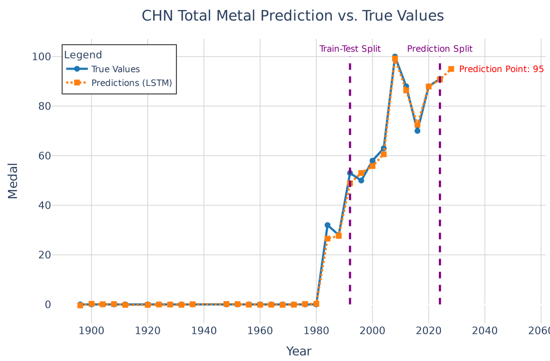
(a) USA Gold Medal Prediction Interval



(b) USA Total Medal Prediction Interval



(c) China Gold Medal Prediction Interval



(d) China Total Medal Prediction Interval

Figure 4: 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. 4a projects 44 medals by 2028 (10% CAGR), reflecting stable growth in professional sports ecosystems. Total medals shown in Fig. 4b 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 4c 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 4d 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 5.

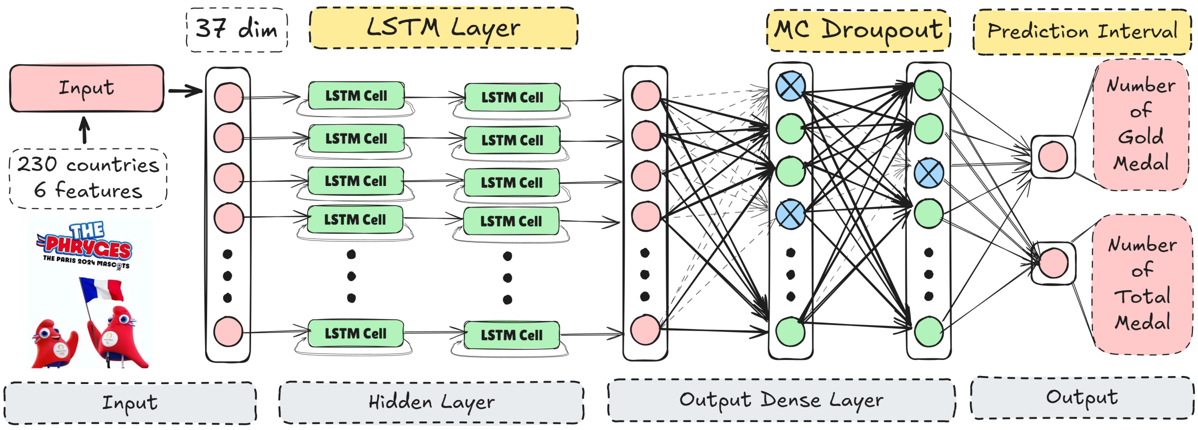


Figure 5: Flow of LSTM based on Monte Carlo Dropout

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 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:

Algorithm 2 Monte Carlo Dropout Uncertainty Quantification

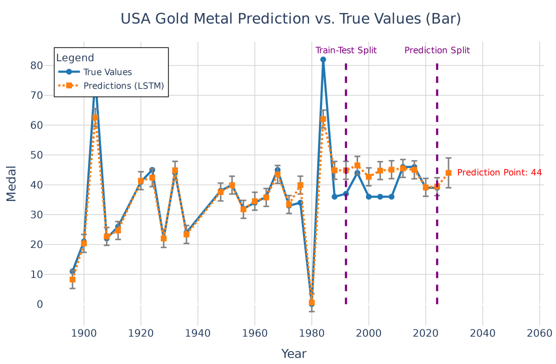
Require: Trained model f_θ, dropout probability p, test sample x^*, MC samples $T = 30$
Ensure: Predictive mean μ, predictive variance σ^2

```
Initialize empty prediction set {ŷ^(t)}_{t=1}^T
for each test sample x* ∈ X_test do
    for t = 1 to T do
        Sample mask m_t ~ Bernoulli(p)                 ▷ Stochastic mask generation
        Apply masked weights: θ_masked ← θ ⊙ m_t
        Compute prediction: ŷ^(t) ← f(x*; θ_masked)
    end for
    Calculate statistics:
    μ ← (1/T) Σ_{t=1}^T ŷ^(t)                          ▷ Predictive mean
    σ² ← (1/T) Σ_{t=1}^T (ŷ^(t) − μ)²                  ▷ Predictive variance
end for
return μ, σ²
```

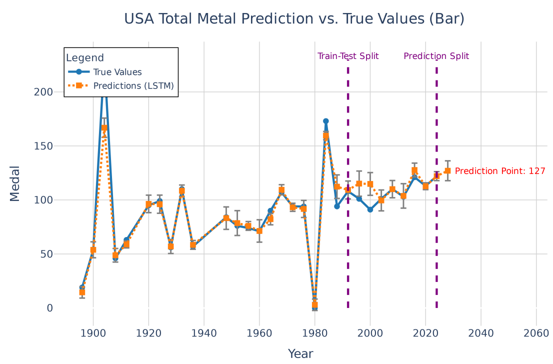
We have empirically validated the selection of key implementation parameters for Monte Carlo Dropout, as shown in the table 3.

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	%

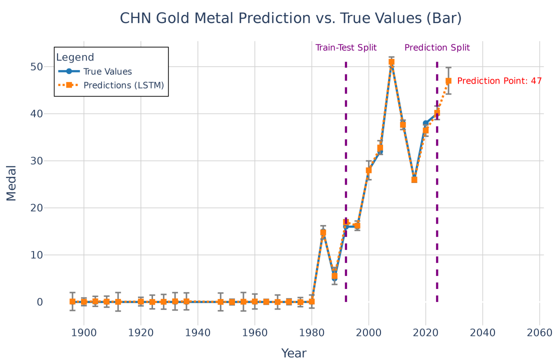


(a) USA Gold Medal Prediction Interval

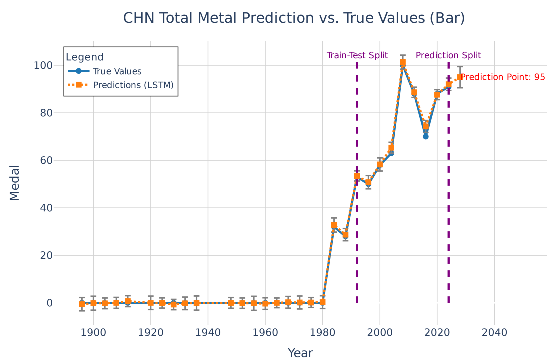


(b) USA Total Medal Prediction Interval

Figure 6: Medal Range Projections for China and USA in 2028



(a) China Gold Medal Prediction Interval



(b) China Total Medal Prediction Interval

The table 4 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	126.0	130.0	44	43.0	45.0
CHN	95	94.0	96.0	42	41.0	43.0
GBR	71	69.0	73.0	37	36.0	38.0
GER	54	53.0	55.0	24	23.0	25.0
FRA	51	50.0	52.0	18	17.0	19.0
AUS	46	45.0	47.0	18	17.0	19.0
JPN	43	42.0	44.0	16	15.0	17.0
RUS	33	32.0	34.0	15	14.0	16.0
NED	33	32.0	34.0	15	14.0	16.0
KOR	28	27.0	29.0	14	13.0	15.0
ITA	28	27.0	29.0	14	13.0	15.0
ESP	26	25.0	27.0	13	12.0	14.0
ROC	21	20.0	22.0	12	11.0	13.0
NZL	21	20.0	22.0	11	10.0	12.0

In Olympic medal prediction, we calculate the probabilities of improvement and decline based on the predicted confidence interval for the 2028 Games. Specifically, the probability of improvement refers to the likelihood that the upper bound of the predicted 2028 medal count exceeds the 2024 medal count, while the probability of decline refers to the likelihood that the lower bound of the predicted 2028 medal count is less than the 2024 medal count.

Let MT_{30} represent the medal count for a given country in 2024, and let the predicted confidence interval for the 2028 medal count be $[L_{\text{lower}}, L_{\text{upper}}]$. The probabilities of improvement and decline can be expressed by the following formulas:

Probability of Improvement

$$P_{\text{progress}} = \frac{\max(0, L_{\text{upper}} - MT_{30})}{L_{\text{upper}} - L_{\text{lower}}}$$

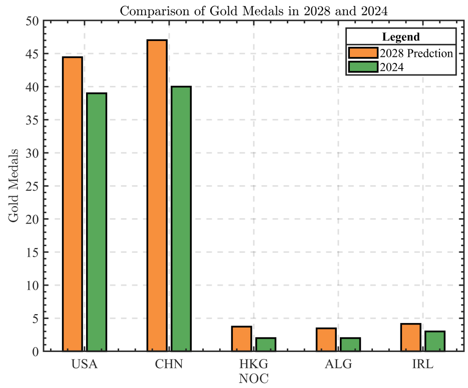
where $L_{\text{upper}} - M_{T30}$ represents the potential improvement if the 2024 medal count is below the upper bound of the 2028 prediction. If the 2024 medal count exceeds the upper bound, the probability of improvement becomes zero.

Probability of Decline

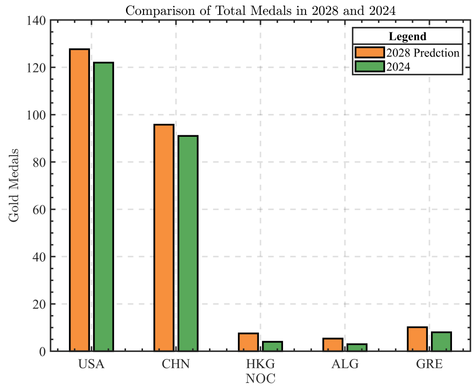
$$P_{\text{decline}} = \frac{\max(0, MT_{30} - L_{\text{lower}})}{L_{\text{upper}} - L_{\text{lower}}}$$

The following Figure 8 illustrates the top five countries that have shown an improvement in their performance at the 2028 Olympic Games in Los Angeles compared to the 2024 Games:

The following Figures 9 depict the top five countries that are projected to experience a decline in their performance at the 2028 Olympic Games in Los Angeles compared to the 2024 Games:

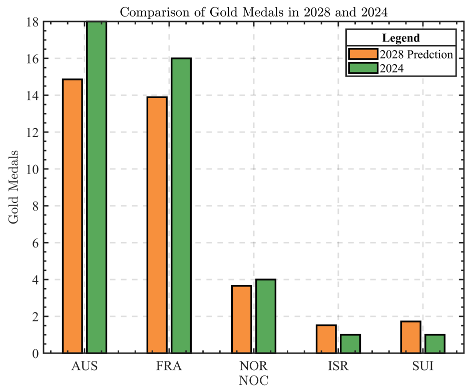


(a) Comparison of Gold Medals: 2028 vs 2024

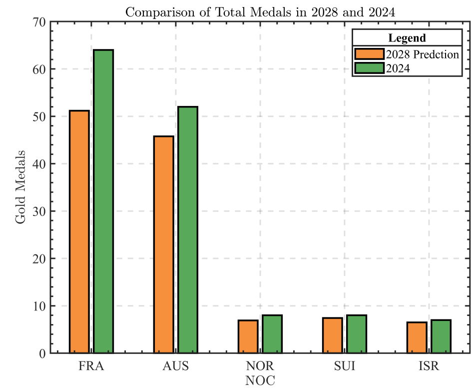


(b) Comparison of Total Medals: 2028 vs 2024

Figure 8: Top five countries with improved performance at the 2028 Olympic Games



(a) Comparison of Gold Medals: 2028 vs 2024



(b) Comparison of Total Medals: 2028 vs 2024

Figure 9: Top five countries with a decline in achievement at the 2028 Olympic Games

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 (Δ=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

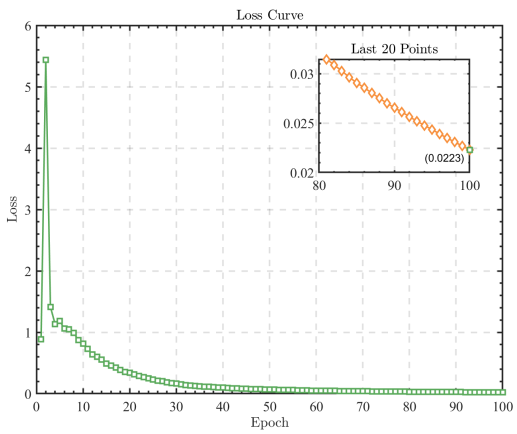


Figure 10: LSTM Training Loss Curve with Monte Carlo Dropout

- **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 (↓0.0229→0.00) and validation loss (↓0.025→0.00) co-converge, suggesting appropriate dropout rate (estimated 0.2)
- **Final Convergence State (>20 epochs)**: Dual loss curves stabilize near 0.00 with ±0.001 fluctuations, indicating optimal model state

5.2 Prediction of Maiden Medal for Medal-Less Countries

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.

5.2.1 Index Analysis

- **Target Variable**
  The target variable is defined as:

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

- **Participants Growth Rate (PGR)**
  Define $\Delta N(k,i) \equiv N_{\text{athletes}}(k-1,i) - N_{\text{athletes}}(k,i)$, the growth rate is:

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

  where k denotes the year index. Negative growth values are automatically clipped by the $\max(0,\cdot)$ operator.
- **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, and the operator $\neg$ represents negation.

- **Unpopular Project Participation Growth Rate (LPIR)**
  Define $\Delta N(k,i) \equiv N_{\text{unpopular}}(k,i) - N_{\text{unpopular}}(k-1,i)$, the growth rate is:

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

5.2.2 XGBoost 01 Breakthrough in Olympic Medal Prediction

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, α_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

5.2.3 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

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.

5.3 Analysis of Event Influence on Medal Distribution

Similar to that defined in 5.1.2

- **Event Held**

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

- **Historical Medal Rate for Country i in Event j**

$$\tilde{D}_{i,j} = \frac{\sum_{q \in \mathcal{Q}_i} V_j(q) \cdot MT_{q,i,j}}{\sum_{q \in \mathcal{Q}_i} V_j(q) \cdot \sum_{k=1}^{N} \sum_{i=1}^{M} MT_{q,i,k,j}},$$

where $\mathcal{Q}_i$ represents the set of years in which country i participated.

- **Ranking of Sports within Each Country**
Once the historical medal rates $D_{i,j}$ have been calculated for all events j for a given country i, we can rank these events for each country based on their historical medal rates. The rank $R_{i,j}$ for country i in event j can be defined as:

$$R_{i,j} = \text{Rank}(D_{i,1}, D_{i,2}, \ldots, D_{i,M}),$$

where Rank$(\cdot)$ represents the ranking function that orders the historical medal rates for country i in all events.

- **Results Visualization**
The figure below shows the ranking of sports based on historical medal rates for country i. The x-axis represents the different events, while the y-axis shows the corresponding historical medal rates. The ranking can be visualized by the height of the bars in the chart.

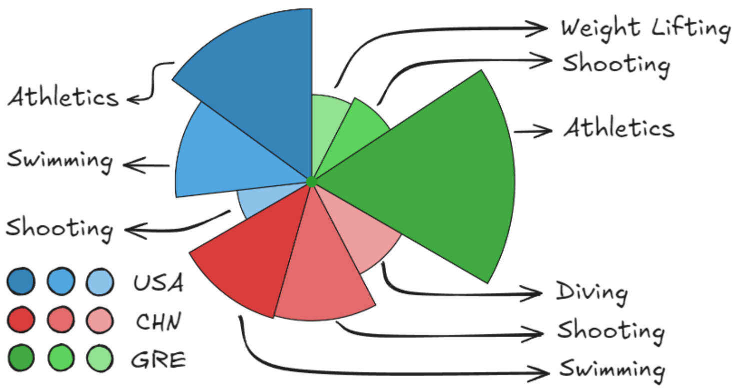


Figure 11: Ranking of Sports for Country USA, CHN, GRE Based on Historical Medal Rates

6 Task 2: Effect of Great Coach

6.1 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 12. To verify this conjecture, we conducted a parallel test on their medal counts.

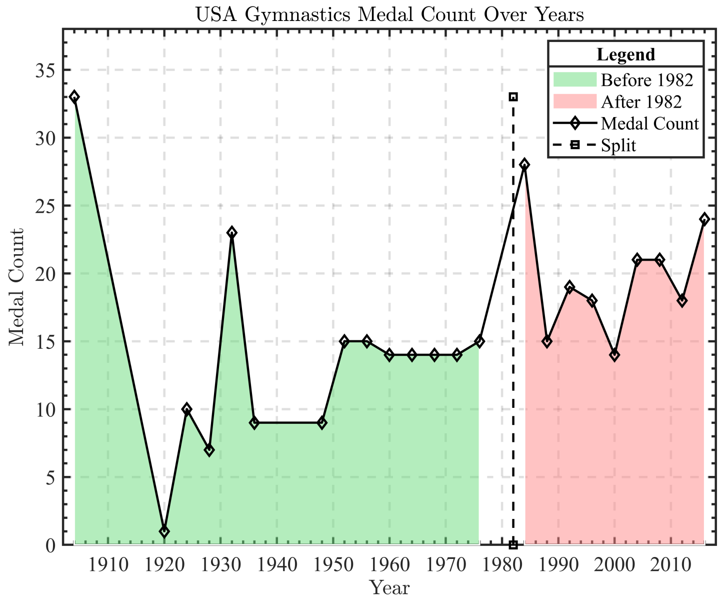


Figure 12: Flow of LSTM based on Monte Carlo Dropout

6.2 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

$$Y_{i,t} = \alpha + \delta_t + \gamma \cdot Treat_i \cdot Post_t + \varepsilon_{i,t}, \tag{1}$$

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 $Least\ Squares\ Method$ in $Python$, we obtained the estimated value of the regression coefficient $\hat{\gamma} = 4.1572$. To test the significance of γ, assume that

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

Select the test statistic:

$$T = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})} \sim t(30-4) \tag{2}$$

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

$$W_\alpha = \left\{|T| \geq t_{1-\frac{\alpha}{2}}(30-4)\right\} \tag{3}$$

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 γ 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

Table 9: Medal Rate Ranking for Different Countries

	France		Japan		Australia	
Rank	Sport	MRR	Sport	MRR	Sport	MRR
1	Weightlifting	1.00	Shooting	1.00	Artistic Gymnastics	1.00
2	Diving	1.00	Cycling	1.00	Judo	1.00
3	Badminton	1.00	Sailing	0.96	Table Tennis	1.00
4	Gymnastics	0.95	Cycling Track	0.95	Shooting	0.93
5	Wrestling	0.90	Athletics	0.95	Boxing	0.81

6.3 Selection of Investment Sports

Generally, countries with strong national power can afford to hire excellent coaches, so we'd better choose countries that have certain strength but are not the very top ones. Considering that an athlete's career usually lasts for 4 Olympic Games, the total number of medals won in the 2012, 2016, 2020 and 2024 Olympic Games was calculated and ranked, seeing Figure 13.

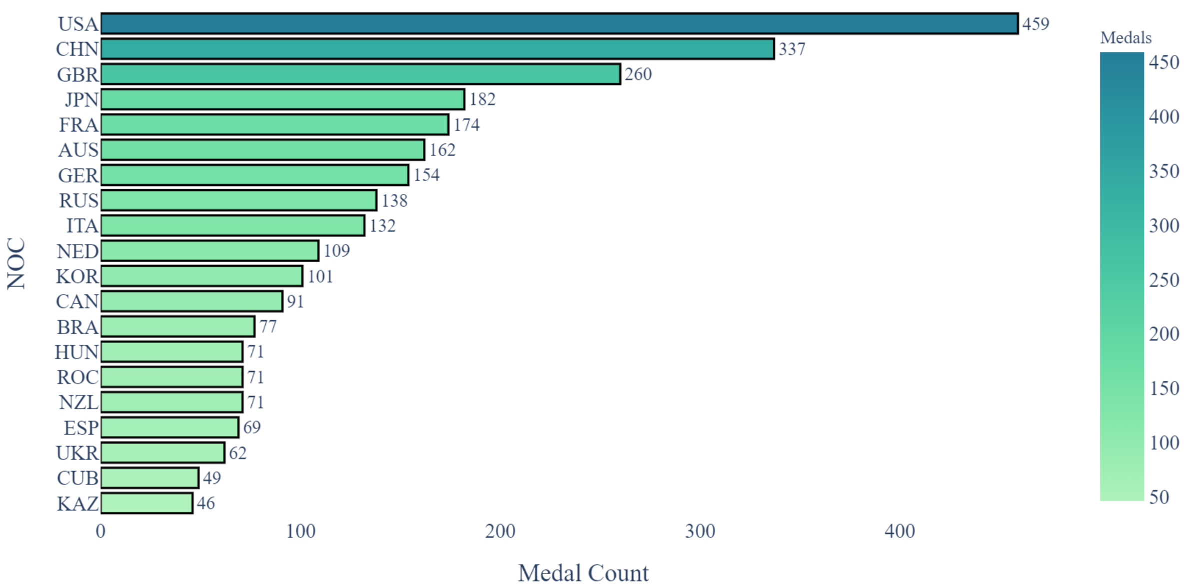


Figure 13: Sort of Various Conutries by Total Medal Count

Based on Figure 13, the United States, China, and the United Kingdom occupy the TOP 3 positions in terms of medal counts. Japan, France, and Australia rank 4-6, respectively. Given that these latter three countries possess substantial national resources and demonstrate significant potential for improvement, We choose to offer investment suggestions for Japan, France and Australia.

The item that requires the most consideration is preferably the one with the greatest "return on input", namely, to select the country with a higher medal return rate. The **Medal Return Rate** is defined as Eq(4).

$$MRR_{i,j} = 1 - \text{Normal}\left(\frac{\sum_{t=27}^{30} \sum_{k \in \tilde{A}_E(j)} MT_{t,i,j}}{\sum_{t=27}^{30} \sum_{j \in \tilde{A}_S} N_{athletes}(t,i,j)} \right) \tag{4}$$

where $\tilde{A}_E(j)$ denotes individual event of sport j, and $\text{Normal}(x) = x/(\max x - \min x)$.

The larger the MRR is, the more people are involved in the single-person sport but the fewer awards are won, indicating a greater potential for medal count improvement. After calculation, $MRR_{i,j}$ are shown in Table **??**.

7 Task 3

8 Sensitivity Analysis

9 Strength and Weakness

9.1 Strength

- **Long-Term Temporal Dependencies:** The model captures temporal dependencies, enabling future medal trend predictions.

- **Uncertainty Quantification:** Monte Carlo Dropout enhances model reliability by providing prediction uncertainty estimates.

- **High Accuracy:** Incorporating multiple features yields strong performance, with R^2 up to 0.93.

- **Efficient Classification:** XGBoost is effective for classifying first-time medal winners with an AUC of 0.90.

9.2 Weakness

- **Data Dependency:** Requires large historical datasets, limiting its applicability for countries with limited data.

- **Model Complexity:** High complexity and extensive training time are required, along with careful hyperparameter tuning.

10 Further Discussion

- **Model Generalization:** The model could be extended to predict medal trends in non-summer events like the Winter Olympics or Youth Olympic Games, testing its adaptability and generalizability across different event types.

- **Ethical Considerations:** It is important to address potential biases, such as the Matthew Effect, where wealthier nations dominate medal counts, and ensure that the model provides equitable predictions for all countries.

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>