Problem Chosen	2025 MCM Summary Sheet	Team Control Number
C		2510086

An MCM Paper Made by Team 2510086

Summary

In Task 1 we begin by implementing a sliding window approach to capture temporal dependencies and historical trends in features such as the number of athletes, events, sports, and medal counts. This data augmentation method enriches the dataset and improves the model's ability to learn from past performances. The primary model used is RidgeCV regression, which handles multicollinearity through L2 regularization and performs automatic cross-validation to select the optimal regularization parameter ($\lambda = 14$). RidgeCV outperformed other models, including Linear Regression, Lasso, ElasticNet, and Support Vector Regression (SVR), with a variance score of 0.95 and RMSE of 4.71. To quantify uncertainty in the predictions, bootstrapping was applied to estimate confidence intervals, providing a robust statistical foundation for medal predictions. The results include a detailed table of predictions for the top 20 countries, with the USA, China, and Great Britain expected to dominate the medal tally. Germany, Czech Republic, and Poland are projected to see significant improvements in their performance compared to the 2024 Olympics, while countries like France, South Korea, and Australia are predicted to experience declines.

Keywords: MATLAB, mathematics, LaTeX.

Contents

1 Introduction

1.1 Problem Background

During the most recent 2024 Paris Summer Olympics, aside from watching the competitions, spectators were also very interested in the medal tally of various countries. The nations at the top of the rankings always attract a lot of attention, and everyone hopes their own country can be among the leaders.

Over the past several decades of Olympic Games, many "sporting powerhouses" have emerged, but there are also numerous countries still striving to win their first Olympic medal in history. How will these medal-winning countries perform in the next Olympics? And do those countries without any medals have a chance to win one? These are all questions that people are concerned about.

Predictions of the final medal count are common, but they are typically not based on historical medal totals. Instead, they are often made just before the opening of the upcoming Olympics, when the athletes scheduled to compete are known[1]. However, when the information about the planned participants has not been disclosed, as is the case now with the 2028 Los Angeles Olympics where the competing athletes have yet to be announced, can we still predict the medal outcomes for various countries? This is the question that concerns us.

1.2 Literature Review

Past research has focused on exploring the application of machine learning techniques to predict medal outcomes and analyze medal distribution patterns for the 2024 Summer Olympics[2]. Leveraging a wide range of variables at both athlete and country levels, as well as event-specific metrics, various statistical models are employed to forecast medal counts and identify factors linked to Olympic success. However, such research methods cannot proceed without information on the participants of the next Olympic Games. This necessitates that we seek solutions within the limited dataset available to us.

1.3 Restatement of the Problem

Considering the background, in this paper we are required to solve the following problems:

- **Task 1:** Project the 2028 LA Olympics medal table using your model, including prediction intervals. Identify countries likely to improve or decline compared to 2024.
- **Task 2:** Predict how many countries will win their first medal in 2028 and estimate the odds of this outcome.
- **Task 3:** Analyze how event types and numbers affect medal counts. Identify key sports for different countries and the impact of host-selected events.

- **Task 4:** Investigate the "great coach" effect on medal counts. Suggest three countries and sports where hiring such a coach could have significant impact.
- **Task 5:** Share original insights from your model to guide national Olympic committees' decisions.

1.4 Our work

For ease of description and visualization, we have drawn a flow chart (Figure 1) to represent our work.

2 Assumptions

3 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
A	the first one
b	the second one
α	the last one

4 Data Pre-processing

Upon organizing the data, it becomes evident that there are inconsistencies between the format of the provided data and the required format specified in the question. To address this, it is crucial to preprocess the data accordingly. The first step involves standardizing the data headers to ensure consistency, followed by the abbreviation of the indicators to enhance clarity and streamline the subsequent listing. Detailed information regarding the specific formatting and abbreviations used can be seen as follow.

Abbreviation	**Description**
Year	Olympic Games year
NOC	Participating country
Host	Host country or not (1: yes, 0: no)
Athletes	Number of athletes
Females	Number of female athletes
Sports	Number of unique sports
Events	Number of unique events

4.1 Conflicts handling

The provided data contains several inconsistencies. For instance, countries are described in various formats, and we have standardized these descriptions using the NOC abbreviation. Similarly, different representations of the same events exist, and we have unified these as well. Additionally, when counting the total number of medals earned by athletes from a given country, we found discrepancies between the numbers derived from the "summerOly athletes.csv" file and those from the "summerOly medal counts.csv" file. To ensure accuracy, we will rely on the medal counts from the "summerOly medal counts.csv" for the final medal tally.

5 Task 1

5.1 Data augmentation using sliding windows

The organized features alone are insufficient for accurately predicting Olympic performance. Given that we are working with time series data, it is crucial to incorporate the historical context of previous performances. To achieve this, we apply the concept of a sliding window. This technique helps our model capture temporal dependencies by accounting for the uncertainty inherent in the progression of the games, while evaluating a country's performance over a recent sequence of events. Specifically, for features such as "Athletes," "Events," "Sports," "Females," "Real Gold", "Real Silver" and "Real Bronze" we use the sliding window approach to augment the dataset. In this case, we implement 9 distinct sliding windows to enrich the model's ability to learn from past data and improve its predictive power.

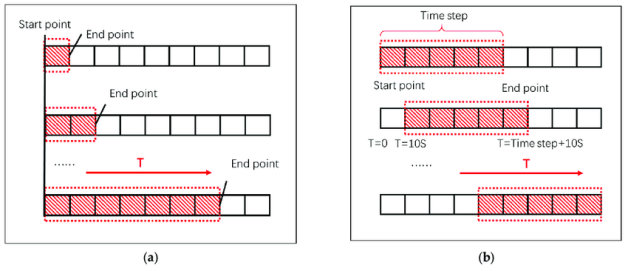

Figure 1: Sliding windows

5.2 RidgeCV Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. When the independent variables are highly correlated, the coefficients of the regression model are unstable. Ridge regression adds a small bias factor to the diagonal of the correlation matrix to stabilize the model. The bias factor is called the ridge parameter, which is a hyperparameter that needs to be tuned.RidgeCV is an extension of the basic Ridge regression model, where it performs automatic cross-validation to select the best regularization ridge parameter(alpha).

The most commonly used regression method is Ordinary Least Squares (OLS), which aims to find a line (or hyperplane) that minimizes the difference between the predicted and actual values. The objective of OLS is to minimize the sum of squared errors:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2$$

where X is the matrix of independent variables, y is the vector of the dependent variable, and β is the vector of regression coefficients. The solution to this optimization problem is:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

However, a problem arises when the independent variables are highly correlated (for example, when house size and number of rooms are strongly related). In this case, the matrix $X^T X$ may become nearly singular, with its determinant close to zero, which causes $(X^T X)^{-1}$ to become unstable. This leads to highly volatile estimates of β, resulting in high variance and poor prediction performance. Ridge Regression addresses the issue of multicollinearity by adding a regularization term $\lambda\|\beta\|^2$ to the OLS objective function. The modified objective function becomes:

$$\hat{\beta} = \arg\min_{\beta} \left(\|y - X\beta\|^2 + \lambda\|\beta\|^2 \right)$$

where λ is the ridge parameter, and I is the identity matrix. The solution to this objective function is:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

Key points to consider:

- The regularization term $\lambda\|\beta\|^2$ helps prevent large regression coefficients, thereby reducing the model's complexity.
- As λ increases, the regression coefficients shrink, lowering the model's variance but introducing some bias.
- By selecting an appropriate value for λ, Ridge Regression can find a balance between bias and variance, which improves the model's predictive performance.

Ridge regression also incorporates the second-order regularization term into the least squares objective, also known as L2 regularization, which reduces dimensionality and minimizes overfitting by controlling the model's parameters. This method is particularly useful in dealing with highly correlated datasets and abnormal samples. Our model used RidgeCV to optimize the regularization parameter λ, improving the model's fit with $\lambda = 14$.

Ridge regression is commonly preferred for handling multicollinearity, as it uses L2 regularization to shrink coefficients without eliminating variables, allowing for a more comprehensive and interpretable model. Unlike Lasso regression, which can eliminate variables by shrinking their coefficients to zero using L1 regularization, Ridge regression ensures a more stable and consistent model by retaining all predictors, making it valuable when all features contribute to the explanatory power of the model.

RidgeCV performs cross-validation over a range of alpha values to automatically select the best one, thereby preventing overfitting or underfitting.

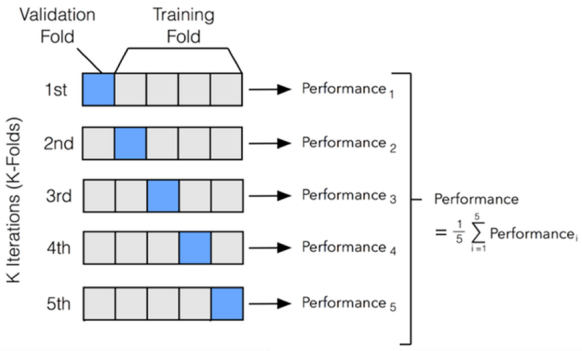

Figure 2: Cross Validation illustration

5.3 Model Evaluation of RidgeCV algorithm

We split the dataset into training and testing sets, using an 80:20 ratio. Various machine learning algorithms were applied to fit the data, including Linear Regression, Lasso Regression, ElasticNet Regression, and Support Vector Regression (SVR). The performance of these models was evaluated using the variance score and root mean squared error (MSE) on test set. Our results showed that RidgeCV consistently outperformed all other models across these metrics.

Models	Variance Score	RMSE
RidgeCV	0.95	4.71
Linear Regression	0.91	6.11
Lasso	0.88	7.05
ElasticNet	0.88	7.05
SVR	0.31	16.97
Neural Network	0.21	18.91

Table 3: Performance of different models evaluated by Variance Score and RMSE.

5.4 Applying bootstrapping to estimate confidence intervals

To estimate the confidence interval for the predicted medal numbers, we employ the bootstrapping method.The bootstrapping method is a statistical technique used to esti-

mate the distribution of a sample statistic by resampling with replacement from the observed data. It allows for the estimation of properties like confidence intervals or standard errors without making strong assumptions about the underlying data distribution.

The confidence interval (CI) for a statistic like the mean using bootstrapping can be expressed as:

$$\text{CI} = \left[\text{percentile}_{\frac{\alpha}{2}}, \text{percentile}_{1-\frac{\alpha}{2}}\right]$$

Where:

- α is the desired significance level (e.g., 0.05 for a 95% confidence interval).
- The percentiles correspond to the lower and upper bounds of the confidence interval, based on the bootstrap distribution.

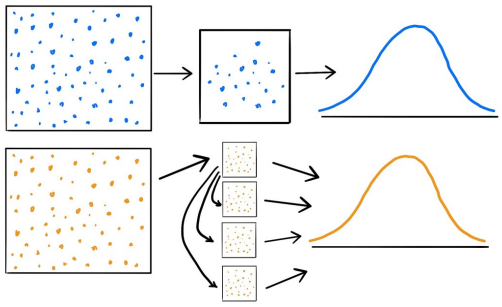

Figure 3: Bootstrapping Illurstration

5.5 The prediction and analysis of 2028 Olympics Game

By combining RidgeCV and bootstrapping, we successfully predicted the medal counts for the 2028 Olympic Games. Below are the results for the top 20 countries, sorted by their predicted total medal counts.

NOC	Golds	Gold_low	Gold_up	Silvers	Silver_low	Silver_up	Bronzes	Bronze_low	Bronze_up	Medals	Medals_low	Medals_up
USA	49	38	59	37	27	45	35	25	40	121	90	144
CHN	35	21	43	25	10	33	24	12	33	84	43	109
GBR	21	14	30	18	8	26	19	11	26	58	33	82
GER	13	2	21	14	8	23	21	15	28	48	25	72
FRA	15	11	19	14	9	18	18	13	23	47	33	60
AUS	14	7	20	12	6	18	18	13	22	44	26	60
JPN	15	9	20	12	6	16	16	12	20	43	27	56
ITA	12	10	16	12	10	15	14	12	17	38	32	48
CAN	8	4	13	10	6	14	12	8	15	30	18	42
NED	9	6	13	9	7	14	11	7	14	29	20	41
KOR	9	0	14	7	0	12	6	0	12	22	0	36
ESP	4	0	8	8	4	12	9	6	15	21	10	35
HUN	7	5	11	6	4	9	7	4	10	20	13	30
BRA	4	0	8	6	2	10	8	4	11	18	6	29
POL	5	2	8	5	2	8	7	4	10	17	8	26
NZL	6	4	9	5	2	7	6	3	8	17	9	24
UKR	3	0	9	3	0	7	8	1	13	14	1	29
SWE	3	1	6	5	1	7	5	3	8	13	5	21
CUB	4	0	7	3	2	6	5	2	8	12	2	22

Table 4: Predicted Medal Counts for the 2028 Olympic Games: Top 20 Countries

We also identified the top 10 countries expected to show improved performance in the 2028 Olympics compared to the 2024 Games, based on predicted medal counts. Conversely, we highlighted 10 countries projected to experience a decline in performance, with fewer medals anticipated in 2028 compared to 2024.

NOC	Gold_char	Silver_char	Bronze_char	Medal_char
GER	1	1	13	15
CZE	1	3	3	7
POL	4	1	2	7
KAZ	2	0	2	4
CAN	-1	3	1	3
DEN	1	2	0	3
ROU	0	0	3	3
CUB	2	2	-1	3
ESP	-1	4	0	3
DMA	1	1	1	3

Table 5: Top 10 Countries with Better Performance in the 2028 Olympics

NOC	Gold_char	Silver_char	Bronze_char	Medal_char
ARM	0	-3	-1	-4
BRN	-2	-1	-1	-4
NED	-6	2	-1	-5
KGZ	0	-2	-3	-5
USA	9	-7	-7	-5
GBR	7	-4	-10	-7
CHN	-5	-2	0	-7
AUS	-4	-7	2	-9
KOR	-4	-2	-4	-10
FRA	-1	-12	-4	-17

Table 6: Top 10 Countries with Worse Performance in the 2028 Olympics

6 Task 2: Predicting the Probability of Countries Winning Their First Medal

In this section, we identify patterns in countries transitioning from no medals to their first Olympic medal, predicting how many countries will win their first medal in the upcoming Games and their probabilities. We preprocessed the data, selected three key features, and trained an XGBoost model optimized with Bayesian tuning. Finally, we apply the model to the 2028 Olympic data to generate predictions.

6.1 Feature Engineering

In this study, we focus on identifying patterns in countries' progression from not winning any medals to winning their first medal. We pre-process the raw athlete data by retaining only records from countries that either have not won any medals or have won their first medal in a particular Olympic Games. Additionally, we remove anomalous data, such as instances where countries like Russia achieved exceptionally high performance in their first Olympic Games due to historical factors. We also exclude data from the refugee team, which began competing in the 2016 Olympics.

Next, we select N_i, P_i, and E_i as the input features. Here, N_i denotes the i^{th} country's participation number, P_i represents the number of athletes sent, and E_i is the number of events entered. The output feature, F_i, indicates whether the country won a medal (1) or not (0). Therefore, the input vector for each country is represented as:

$$\mathbf{x}_i = (N_i, P_i, E_i)$$

and the corresponding output is:

$$y_i = F_i$$

In total, we process and organize 1,244 vectors representing the various features of each country's participation and medal achievement across different Olympic Games.

6.2 Bayesian-Optimized XGBoost Medal Predictor (BO-XGMP)

Chen et al. developed the **eXtreme Gradient Boosting (XGBoost)** algorithm, which leverages an optimized distributed gradient boosting technique to quickly train datasets while maintaining efficient resource utilization and high accuracy[4]. In this problem, we aim to predict how many countries will win their first medal at the next Olympic Games, and estimate the probability of this occurrence. The dataset consists of only 1244 samples, with a significant imbalance between the number of countries that have won medals and those that have not, making the task inherently imbalanced. XGBoost has demonstrated excellent performance in addressing imbalanced and small-sample binary classification problems[5][6]. By using the binary:logistic loss function, it directly outputs probabilities, which is ideal for this task. Additionally, we employ Bayesian optimization to fine-tune the hyper-parameters of XGBoost to further enhance the model's performance[7].

6.2.1 Model ConstructionXGBoost and Bayesian Optimization

A. XGBoost

The XGBoost objective function consists of two components: the loss function and the regularization term, both of which work together to optimize the model. The objective function is expressed as:

$$L(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \Omega(\theta)$$

where $\ell(y_i, \hat{y}_i)$ represents the loss function, which measures the error for the i^{th} sample, and $\Omega(\theta)$ is the regularization term, which controls the model's complexity to prevent overfitting.

For binary classification, the loss function used is log-loss:

$$\ell(y_i, \hat{y}_i) = -y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)$$

where y_i is the actual label (0 or 1) for the i^{th} sample, and $\hat{y}_i$ is the predicted probability that the sample belongs to class 1. The loss function minimizes the difference between the predicted probabilities and the actual labels.

The regularization term is defined as:

$$\Omega(\theta) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{k} \theta_j^2$$

where T is the number of trees, γ controls the complexity of the trees, and λ controls the size of the leaf weights. This regularization term penalizes overly complex models, reducing the risk of overfitting and ensuring better generalization.

B. Bayesian Optimization

Bayesian Optimization (BO) uses historical evaluation results to select optimal hyperparameter combinations based on their distribution[8]. It uses **Gaussian Processes (GP)** to model the relationship between the model error ϑ_Z and its parameters p_Z. A set

of Z points generates a multivariate Gaussian distribution in $\mathbb{R}^Z$. With prior Gaussian processes derived from previous experiments, a posterior function $\alpha(p)$ is constructed, where the acquisition function (AC) depends on the predictive mean function $\hat{\mu}(p)$ and the predictive variance function $\sigma^2(p)$.

To determine the next sampling point, the Bayesian optimization model maximizes $p_{\text{next}} = \arg\max_p \alpha(p)$, balancing the exploration of areas with high variance and the exploitation of regions with low mean. The **Gaussian Process Upper Confidence Bound (GP-UCB)** acquisition function is used to control the exploration-exploitation trade-off, with the parameter κ adjusting the balance:

$$\alpha_{UCB} = \varpi(\mathbf{p}) - \kappa\sigma(\mathbf{p})$$

This method demonstrates strong performance in hyper-parameter tuning by efficiently navigating the search space.

6.2.2 Model Solution

- **(1) Model Training:** Since countries' first medal achievements vary significantly—for instance, San Marino (SMR) won 2 silvers and 2 bronzes in their first appearance—while many countries win only a bronze, we aim to differentiate these cases. To address this, we apply the Fibonacci Weighted Point System proposed by Sergeyev to quantify performance[3].

  The performance score for a country's team in the i^{th} Olympic Games is calculated as follows:

  $$PE_i = 3g_i + 2s_i + b_i$$

  where PE_i is the performance score for the i^{th} country's team, calculated based on the number of gold (g_i), silver (s_i), and bronze (b_i) medals won.

  For training the model, each input vector will be assigned a weight based on PE_i, the performance score of the i^{th} team. The weight for each sample is calculated using the following formula:

  $$w_i = \frac{PE_i}{\sum_{i=1}^{n} PE_i}$$

  Here, w_i represents the weight for the i^{th} team, and PE_i is the performance score. This weighting ensures that teams with stronger performance histories have more influence during model training. The dataset is split into 80% for training and 20% for testing. We utilize 5-fold cross-validation with 5 repetitions to assess the model's performance across different subsets, reducing the risk of overfitting and ensuring robustness in the results.

- **(2) Bayesian Optimization for Hyper-parameters:** Bayesian Optimization is employed to fine-tune the following hyper-parameters: learning rate, maximum tree depth, number of weak learners, column sampling rate, minimum child weight, subsample rate, and pruning parameter. We begin the optimization with 10 initial points and perform

100 iterations. The optimization goal is to maximize the **Area Under the Precision-Recall Curve (AUC-PR)**, as it provides a more informative measure for models dealing with imbalanced data, focusing on the performance of the minority class (teams winning medals)[9]. Maximizing AUC-PR ensures that the model is well-calibrated to predict rare events, such as medal wins, accurately.

- **(3) Model Evaluation:** The optimal hyperparameters and model evaluation metrics are as follows. The model demonstrates excellent performance, with an Accuracy exceeding 0.9, AUC and Precision greater than 0.8, and F1 Score and AUC-PR surpassing 0.7. These results indicate that the model has effectively balanced predictive accuracy and has the ability to correctly classify both majority and minority classes (see Table 4).

Table 7: Model Hyper-parameters and Performance Metrics

(a) Model Hyper-parameters

Parameter	Value
colsample_bytree	0.9312
gamma	0.0834
learning_rate	0.3088
max_depth	12.0
min_child_weight	1.0
n_estimators	130.0
subsample	0.737

(b) Model Performance Metrics

Metric	Value
Accuracy	0.9037
Precision	0.8341
Recall	0.8027
F1 Score	0.7667
AUC	0.8603
Log Loss	0.2022
AUC-PR	0.7777

6.3 Future Prediction

- **Step 1: Fitting the Data for 2028**

  We begin by extracting the historical data of countries that have never won a medal from the dataset. Next, we assume that the number of participants and the number of events for each country follows a linear trend. We then construct a linear regression model. For the number of participants, the model is expressed as:

$$P_{ij} = \beta_0 + \beta_i \cdot j + \epsilon_i$$

  where: P_{ij} represents the number of participants from country i in the j^{th} Olympic Games.β_0 is the intercept term. β_i is the coefficient for country ij represents the year of the Olympic Games. ϵ_i is the error term for country i.

  To estimate the coefficients β_0 and β_i, we use the least squares method to minimize the sum of squared errors:

$$\hat{\beta}_i = \arg\min_{\beta_i} \sum_{j=1}^{n} \left(P_{ij} - (\beta_0 + \beta_i \cdot j)\right)^2$$

  Finally, after fitting the model, we predict the number of participants for the year 2028 by substituting $j = 2028$ into the equation:

$$P_{i,2028} = \hat{\beta}_0 + \hat{\beta}_i \cdot 2028$$

Similarly, the predicted value of E_i in 2028 can also be obtained.

- **Step 2: Predicting with BO-XGMP**

  In this step, we use the BO-XGMP model to predict the probability of each country winning a medal in the 2028 Olympics. For each country ii, its historical data for 2028 is input into the model, providing the predicted probability p_i of winning a medal.

  Figure 4 illustrates the top thirty countries predicted by the BO-XGMP model to have the highest probabilities of winning their first medal in the 2028 Olympics. Notably, the teams from Honduras (HON), Bolivia (BOL), and Malta (MLT) have predicted probabilities of 0.28, 0.26, and 0.23, respectively, all exceeding 0.2, suggesting a strong likelihood of winning medals at the 2028 Los Angeles Olympics.

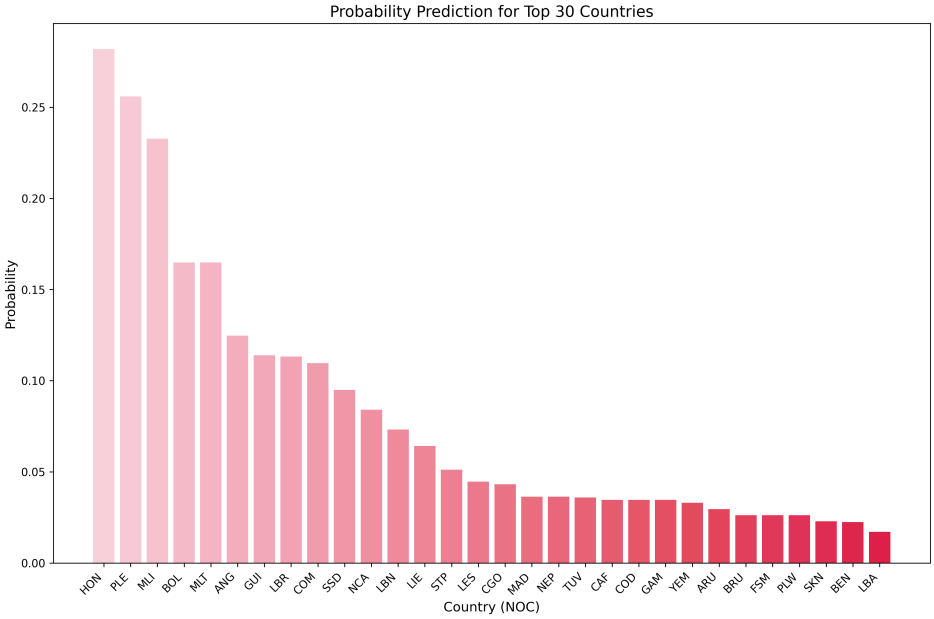

Figure 4: Top30-Country-Probability-Prediction

We assume the medal-winning probabilities for different countries are independent. The total probability distribution for k countries winning medals is modeled using the binomial distribution. The binomial probability mass function is:

$$P(X = k) = \binom{N}{k} \prod_{i=1}^{N} p_i^{k_i} (1 - p_i)^{(1-k_i)}$$

where X is the number of countries winning medals, N is the total number of countries being considered, k_i is 1 if country i wins a medal and 0 otherwise, and p_i is the predicted probability of country i winning a medal.

Using Python, we computed the probability distribution of countries winning their first medal, as shown in Figure X. Our prediction for the 2028 Los Angeles Olympics indicates that two countries have the highest probability, approximately 0.30, of winning their first medal. Following them are one country with a probability of 0.28, three countries with a probability of 0.19, and zero countries with a probability of 0.12. Four

countries have probabilities of 0.08, while the remaining scenarios have probabilities of less than 0.05, which are considered low-probability events.

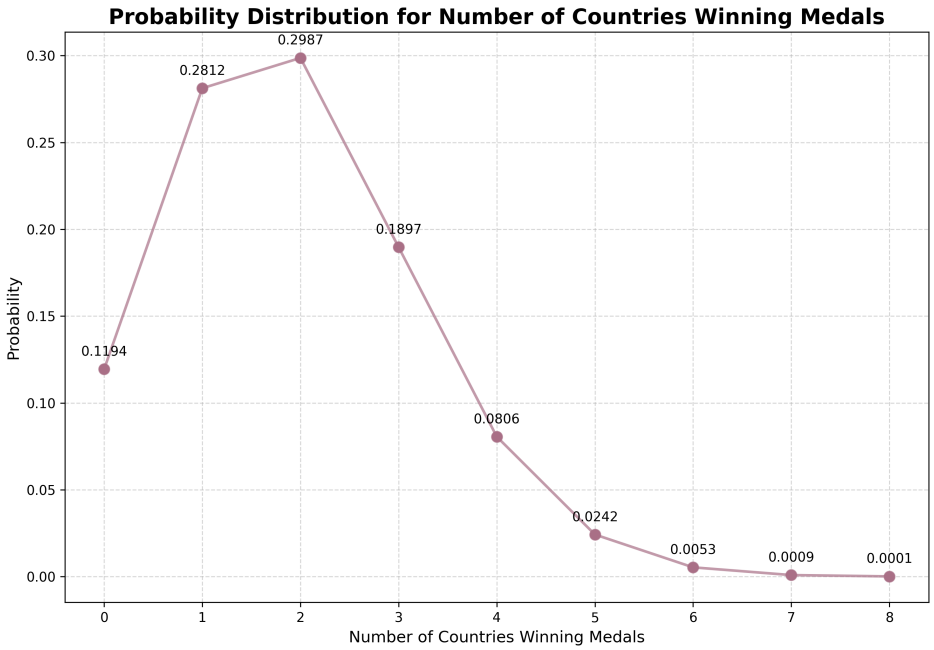

Figure 5: Probability Distribution for Number of Countries Winning Medals

7 Task 3

7.1 Explore the Relationship of Events and Medal Counts

The relationship between event types (both in terms of number and category) and medal counts is a fascinating area of research. To explore this, we conduct a statistical analysis on the number of athletes participating in various event types. Additionally, we create a correlation matrix to clearly visualize how the number of events and the event types contribute to the overall medal counts. This approach helps in identifying key patterns and understanding the factors influencing performance outcomes.There are numerous events, so we encoded different events and have only listed a few with full names as examples.

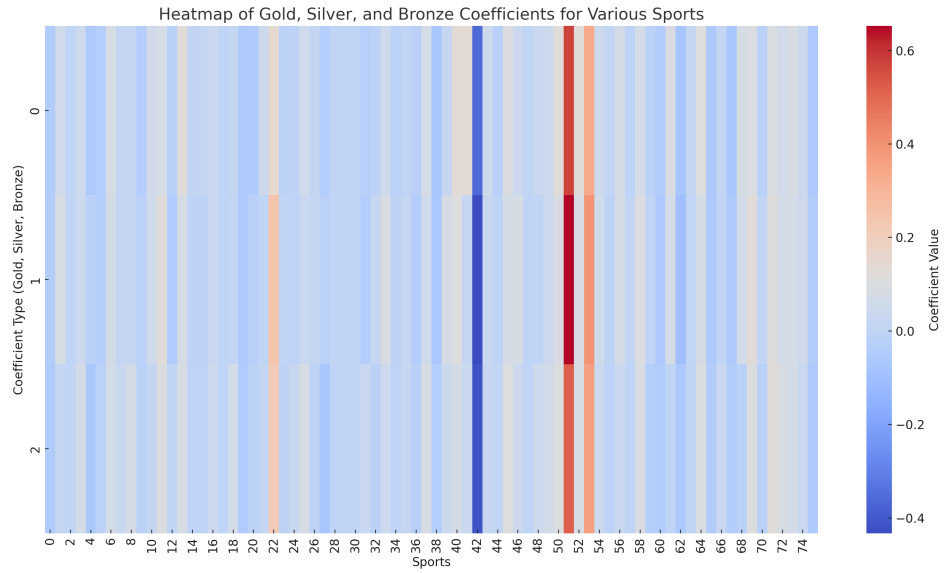

Figure 6: Heatmap of Gold, Silver and Bronze Coefficients for Different Events

Feature	Events	Sport_3x3 Basketball	Sport_3x3 Basketball, Basketball	Sport_Aeronautics	Sport_Alpinism	Sport_Archery	Sport_Art Competitions	Sport_Artistic Gymnastics	Sport_Artistic Swimming	...	Sport_Wrestling
Golds Coefficients	-0.045377	0.058666	0.046575	-0.057182	-0.034728	0.081804	0.015434	0.01263	-0.036921	...	-0.060007
Silvers Coefficients	-0.040295	0.081501	0.041639	-0.015283	-0.032828	0.069523	0.015134	0.015989	-0.016273	...	-0.039233
Bronze Coefficients	-0.047274	0.019216	0.070082	-0.067294	-0.016598	0.07101	0.035957	0.078456	-0.015647	...	-0.058405

Table 8: Coefficients Matrix

We can observe that different events may have varying positive or negative effects on medal counts, for example, 3x3 Basketball has a positive effect on medal counts while Artistic Swimming has a negative effect on medal counts

7.2 The Importance of Sports Projects for a Country

A country's investment and emphasis on sports significantly influence the number of athletes and their overall performance[10]. To identify the most important Olympic sports for a country, we must analyze both the number of athletes participating and the country's achievements in each sport.

7.2.1 Data Normalization and Proportional Analysis

However, the inherent differences between sports the number of participants and the total number of medals availablemust be taken into account. For example, in 2024, the athletics event had 1810 participants and 144 medals, while table tennis had only 172 participants and 15 medals. To eliminate the influence of these differences between sports, we process the data as follows: For each country, we calculate the proportion of participation and medal achievements in each sport. Specifically, the participation ratio for each event is defined as:

$$\text{PAratio}_{i,j} = \frac{\text{PA}_{i,j}}{\sum_i \text{PA}_{i,j}}$$

where $PA_{i,j}$ represents the total number of athletes from country i in sport j.

Similarly, we compute the PE for each sport in each country using the previously mentioned Fibonacci Weighted Point System. The performance ratio for each event is:

$$PE_{\text{ratio},i,j} = \frac{PE_{i,j}}{\sum_{i=1}^{N} PE_{i,j}} \quad \text{where } N \text{ is the number of countries.}$$

7.2.2 Entropy Weight Method (EWM)

To determine the relative importance of participation ratio and performance ratio, we use the EWM. This method allows us to quantify the relative importance of each feature (participation and performance) in assessing the importance of each sport. H_j: The entropy for feature j is computed as follows:

$$H_j = -\frac{1}{\ln(N)} \sum_i P_{i,j} \ln(P_{i,j})$$

where where $P_{i,j}$ is the normalized value of feature j, and N is the number of countries.

The weight for each feature j is then derived using the following formula:

$$W_j = \frac{1 - H_j}{\sum_{j=1}^{M}(1 - H_j)} \quad \text{where } M \text{ is the number of features.}$$

7.2.3 TOPSIS Method

Once the weights for the participation ratio and performance ratio are determined, we apply the TOPSIS method to calculate the overall importance score for each sport. The steps of the TOPSIS method are as follows:

- **Step 1: Construct and Normalized Decision Matrix:** We construct a decision matrix where each row represents a country and each column represents a specific sport. Each element x_{ij} in this matrix corresponds to the combined value of participation and performance ratios for country i in sport j.

  In order to eliminate the influence of different dimensions of indices, we standardize the matrix. The standardization formula is as follows:

  $$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}$$

  Use Z_{ij} as the element to construct the normalized matrix $Z_{nm} = [Z_{ij}], i = 1, ..., n; j = 1, ..., m;$

- **Step 2: determine the best and worst solutions:** For each column (representing a sport), we determine the best and worst solutions:

  The best solution Z^+ is the maximum value for each column:

  $$Z^+ = (max(Z_{i1}), max(Z_{i2}), ..., max(Z_{im})$$

The worst solution Z^- is the minimum value for each column:

$$Z^- = (min(Z_{i1}), min(Z_{i2}), ..., min(Z_{im})$$

Calculate the distance between Z^+ and D_i^+ and the distance between Z^- and D_i^- for each evaluation object. The calculation of the distance here needs to use the weight calculated by the EWM method in the previous section.

$$D_i^+ = \sqrt{\sum_{j=1}^{m} w_j(\max Z_{ij} - Z_{ij})^2} \qquad D_i^- = \sqrt{\sum_{j=1}^{m} w_j(\min Z_{ij} - Z_{ij})^2}$$

- **Step 3: Calculate the Closeness Index:** $C_i = \frac{D_i}{D_i^+ + D_i^-} \quad 0 \leq C_i \leq 1$

  The closer C_i is to 1, the more important sport j is for country i. A higher C_i value indicates that sport j is of more importance to that country.

7.3 Results

To illustrate the application of our model, we focused on China (CHN) and input the normalized participation and performance ratios for each sport into the model. These ratios, derived from China's historical athlete participation and medal achievements, were processed using the **EWM-TOPSIS Model** to calculate the Closeness Index for each sport. This index represents the proximity of each sport's performance to the optimal solution, with higher values indicating greater importance.

As shown in the Figure 7 below, the Closeness Index for **table tennis (Closeness = 0.90)** is the highest among all sports, underscoring its paramount importance to China in the Olympic context. Following closely are badminton,trampolining, and diving, all of which have Closeness values exceeding **0.5**. These sports are also critical to China's Olympic strategy, highlighting their significant role in the country's athletic achievements.

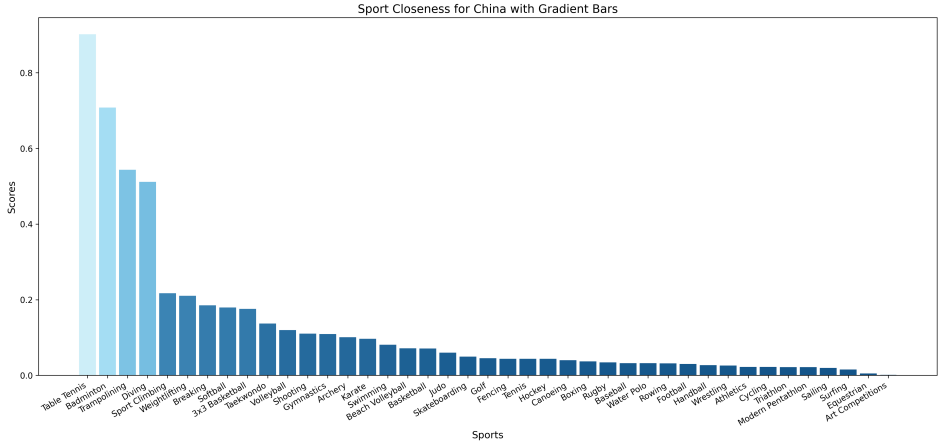

Figure 7: Sport Closeness for China with Gradient Bars

8 Task 4: Evaluating the "Great Coach" Effect

8.1 Evidence for the existence of the "great coach" effect

Given the limitations of the data available, the existence of the "great coach" effect can be substantiated by examining the careers of the two given exemplary figures: Lang Ping and Béla Károlyi. Lang Ping, renowned for her coaching prowess, has led both Chinese and American volleyball teams to championship victories, demonstrating her ability to excel across different cultural and athletic contexts. Similarly, Béla Károlyi's illustrious career includes guiding Romanian and subsequently American women's gymnastics teams to remarkable achievements, further reinforcing the notion that exceptional coaching transcends national boundaries and can yield consistent success.The analysis of the provided data reveals strong evidence of Béla Károlyi's significant coaching impact on the USA team's performance in the 2016 Olympics. The actual performance of the USA team far exceeded the predicted medal counts, highlighting the exceptional influence of his coaching on the team's success.

	Gold	**Silver**	**Bronze**	**Total**
Predicted	45	35	31	111
Actual	46	37	38	121

Table 9: Comparison of Predicted and Actual Medal Counts for the USA in the 2016 Olympics

8.2 Anticipating the effect's contribution to medal counts

To better evaluate the effect, we continue to use the Fibonacci Weighted Point System proposed by Sergeyev to quantify performance[3].

The performance score for a country's team in the i^{th} Olympic Games is calculated as follows:

$$PE_i = 3g_i + 2s_i + b_i$$

where PE_i is the performance score for the i^{th} country's team, calculated based on the number of gold (g_i), silver (s_i), and bronze (b_i) medals won.

Using the Fibonacci Weighted Point System in conjunction with the collected data, we estimate that the influence of the exceptional coaching effect increases medal counts by approximately 10

8.3 Suggestions for three countries to make use of such effect

8.3.1 India – Athletics (Track Field and Middle/Long Distance Running)

India has shown immense potential in athletics but struggles to consistently win medals in events like track and field. While Neeraj Chopra's gold in javelin at the 2020 Olympics

was historic, other areas like sprinting, middle-distance running, and high jump remain underdeveloped.

India could hire top international coaches specializing in specific disciplines, such as sprinting (e.g., a coach with experience training athletes in Jamaica or the US) or distance running (e.g., from Ethiopia or Kenya). A focus on nurturing youth and setting up world-class training facilities would amplify results.

A focused approach with a great coach could elevate India's competitiveness in track and field, targeting a significant medal haul in future Olympics or World Championships.

8.3.2 Brazil – Swimming

While Brazil is strong in team sports (soccer, volleyball, etc.), it has untapped potential in swimming. It has produced strong swimmers like César Cielo but lacks consistent medalists in major competitions.

Bringing in an elite swimming coach from the US or Australia—countries with a track record of Olympic swimming dominance—could help develop a deeper talent pool. Establishing long-term training programs with an emphasis on youth development would also be critical.

By maximizing homegrown talent with world-class guidance, Brazil could dominate not only in regional competitions but also increase their global presence in swimming.

8.3.3 South Africa – Boxing

South Africa has a rich boxing history and athletic potential but has struggled to translate it into consistent Olympic success. With the right focus, the country could regain prominence in boxing, particularly in lightweight and welterweight categories.

Hiring an internationally renowned boxing coach—such as someone with Olympic success from Cuba or Eastern Europe—could elevate the skill and strategy of South African boxers. Such countries are known for their rigorous boxing training systems.

A strategic investment in coaching and grassroots programs could make South Africa a dominant force in boxing and increase their medal count at future Olympic Games.

9 Task 5

10 Sensitivity Analysis

11 Strengths and Weaknesses

11.1 Strengths

- First one...
- Second one ...

11.2 Weaknesses

- Only one ...

References

[1] Nielsen. Nielsen's Gracenote Expects USA, China, Great Britain, France and Australia to Lead 2024 Paris Olympic Games Medal Table. https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/.

[2] Moolchandani, Jhankar, et al. "Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics." *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*. 2024. https://ieeexplore.ieee.org/document/10840553.

[3] Sergeyev, Yaroslav D. "The Olympic Medals Ranks, lexicographic ordering and numerical infinities." *arXiv preprint arXiv:1509.04313*, 11 Sep. 2015. Available: https://arxiv.org/abs/1509.04313.

[4] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[5] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6). *doi:10.1177/15501329221106935*

[6] Li, J., Liu, H., Yang, Z., & Han, L. (2021). A Credit Risk Model with Small Sample Data Based on G-XGBoost. *Applied Artificial Intelligence*, 35(15), 1550–1566. https://doi.org/10.1080/08839514.2021.1987707

[7] Shi, R., Xu, X., Li, J., & Li, Y. (2021). Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization. *Applied Soft Computing*, 109, 107538. https://doi.org/10.1016/j.asoc.2021.107538

[8] Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1), 26-40.

[9] Saito, T., and M. Rehmsmeier. *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. *PLoS One* 10.3 (2015): e0118432. *doi: 10.1371/journal.pone.0118432*.

[10] Liang, Xiaodan. *Analysis of the Development Trend of China's Competitive Track and Field Events from the Results of the 18th Asian Games Track and Field Competition*. *Journal Name* 10.2 (2019): 123-145.

Appendix A: Further on LaTeX

To clarify the importance of using LaTeX in MCM or ICM, several points need to be covered, which are ...

To be more specific, ...

All in all, ...

Anyway, nobody **really** needs such appendix ...

Appendix B: Program Codes

Here are the program codes we used in our research.

test.py

```python
# Python code example
for i in range(10):
    print('Hello, world!')
```

test.m

```
% MATLAB code example
for i = 1:10
    disp("hello, world!");
end
```

test.cpp

```
// C++ code example
#include <iostream>
using namespace std;

int main() {
    for (int i = 0; i < 10; i++)
        cout << "hello, world" << endl;
    return 0;
}
```