Problem Chosen	**2025**	**Team Control Number**
C	**MCM/ICM** **Summary Sheet**	**2518527**

Olympic Medal Modeling and Analysis

Summary

For **Task.1**, we developed an ensemble model based on gradient boosting to predict the total number of medals for each country. After multiple attempts for parameter adjustment and train , our model achieved an R^2 of 1.00 and 0.96 on the training and test sets, respectively, and controlled the `MSE` on the dataset to around 2.77, `MAPE` to around 12.93% , showing relatively good performance.

According to that, we predicted the medal table for the 2028 Los Angeles Olympics: the United States will remain in first place with 45 gold medals, 46 silver medals, and 25 bronze medals, totaling 117 medals; while China will rank second with 35 gold medals, 27 silver medals, and 21 bronze medals, totaling 83 medals.

For **Task.2**, we established a machine learning model similar to **Task.1** and used the importance of parameters to analyze the impact of projects on the number of medals in each country.

For **Task.3**, we developed a neural network learning model to identify the top 10 countries with the highest probability of winning their first medal in the next event.

To estimate the probability of previously non-award-winning countries securing a medal in the next Olympic Games, we employed a neural network to learn a classifier that predicts whether a country without a medal in the current games will win a medal in the next. The sigmoid function served as the activation function, while gradient descent optimized the propagation matrix parameters to minimize the loss function.The backpropagation algorithm was used to calculate the gradient of each element in the propagation matrices during optimization. Finally, the accuracy on the validation set and the value of the loss function were used to assess the classifier's predictive performance and ensure its reliability.

For **Task.4**, we find evidence for the existence of a 'great coach' effect by employing Bayesian change-point detection methods to quantify the contribution of the coach.

We set the weighted medal count to W_t and used Bayesian change point detection to identify change points in the time series data. Then we determine whether the change points are caused by the 'great coach' effect by combining the year of the 'great' coach's appointment. After that, we calculate the coach's contribution rate by comparing the changes in the number of medals before and after the change point. Next, we calculate the historical level of a country in a particular sport based on the countries and sports where we have obtained 'great' coaches. By analyzing the historical level and the total number of medals in a sport, three countries and their sports that should be considered for investing in 'great' coaches are selected. Finally, the contribution of the 'great coach' effect to the number of medals of these three countries is predicted.

For **Task.5**, we studied the impact of economic, population, and other factors on the number of Olympic medals and visualized the results of the analysis.

Keywords: Random Forest, Gradient Boosting, Neural Network, Gradient Descent, Bayesian statistics

Contents

1. Introduction 1
1.1. Background 1
1.2. Restatement of the Problem 1
2. Task 1: Model for Medal Prediction 2
2.1. Assumptions and Justification 2
2.2. Model Overview 2
2.3. Faucets Determination 3
2.4. Data Cleaning and Preprocessing 4
2.5. Model Training 4
2.6. Model Evaluation 4
2.7. Results 7
3. Task 2: Analysis of Important Sports 8
3.1. Determination of Parameters 8
3.2. Model Evaluation 8
3.3. Results 8
4. Task 3: Prediction of First-time Medal Winners 10
4.1. Model Construction 10
4.2. Model Solution 10
  4.2.1. Data Preparation 10
  4.2.2. Neural Network Framework Construction 10
  4.2.3. Cost Function 11
  4.2.4. Problem Transformation 11
  4.2.5. Parameter Optimization 11
4.3. Model Evaluation 12
5. Task 4: Analysis of the "Great Coach" Effect 13
5.1. Main tasks 13
5.2. Data preprocessing 13
5.3. Bayesian change point detection 13
  5.3.1. Principle 13
  5.3.2. Model Assumption 13
  5.3.3. Model Solution 13
  5.3.4. Analysis of Results 14
5.4. Quantifying the level of a country in a given project 16
  5.4.1. Calculating a Coach's Contribution 16
  5.4.2. Analysis of Results 17
5.5. Choosing to invest in a sport with a "great coach" 17
  5.5.1. Expansion of data 17
  5.5.2. Quantifying the level of a country in a given project 17
  5.5.3. Choose sports that are prone to the "great coach" effect 18
  5.5.4. Selection of countries based on historical level 18
  5.5.5. Estimating the contribution of their "great coaches" 18
6. Task 5: Analysis of other factors. 21
References 22

1. Introduction

1.1. Background

During the 2024 Summer Olympics in Paris, spectators and the media will focus on individual event performances and pay close attention to each country's overall medal table ranking. The medal table reflects not only the efforts of individual athletes and teams but also the overall strength and competitiveness of countries in the field of sports.

Before the start of the Olympic Games, many organizations and experts try to predict the outcome of the medal table. These predictions are usually based on historical data, recent event performances, athlete rosters, and the advantages of the host country. However, accurate medal predictions are not an easy task as they require a combination of complex factors such as athlete status, unforeseen circumstances during the competition, and changes in the program settings.

Medal predictions can help us to provide a basis for national sports planning, helping to rationally allocate resources, optimize project development, and enhance overall sports strength. At the same time, it can motivate athletes and coaches to set clear goals, adjust training strategies, and enhance confidence. It can promote the development of the sports industry. It also provides a reference for sports research and analysis, revealing the trend of changes in sports strength and potential influencing factors of various countries.

1.2. Restatement of the Problem

- **Task.1**: We need to develop a model of the total number of medals for each country. Based on this model, we need to predict the prediction intervals for all outcomes of the 2028 Summer Olympics medal table in Los Angeles, USA.

- **Task.2**: According to the model, we need to analyse which sports are most important to each country and the impact of the sports chosen by the host country on the outcome of the competition.

- **Task.3**: For countries that have not won medals, we need to predict how many countries will win their first medals at the next Olympics, giving the chances of this estimate being accurate.

- **Task.4**: We need to study the data for evidence of changes that may be caused by the "great coach" effect. Estimate the contribution of this effect to the number of medals. Select three countries and identify the sports in which they should consider investing in "great" coaches and estimate the effect.

- **Task.5**: Analyse what other factors may affect Olympic medal counts based on the modeling model.

2. Task 1: Model for Medal Prediction

2.1. Assumptions and Justification

Considered about the complexity of the Olympic medal count, it is obviously unrealistic to consider all factors. Therefore, after studying the given dataset, we made the following assumptions:

Deterministic: The number of medals in each country mainly depends on the strength of its athletes, and the number of medals in that country in that year can reflect the strength of that country to a certain extent.

Stability: Although the number of medals is affected by various external factors (such as international politics), we believe that the impact of these factors is stable in the short term.

Host Advantage: The host country usually performs better at the Olympics.

2.2. Model Overview

After trying lots of model pipelines and feature engineering methods, we finally chose an ensemble algorithm with **gradient boosting** as the core to predict the number of medals.

As Figure 1 describes, our training pipeline consists of the following steps:

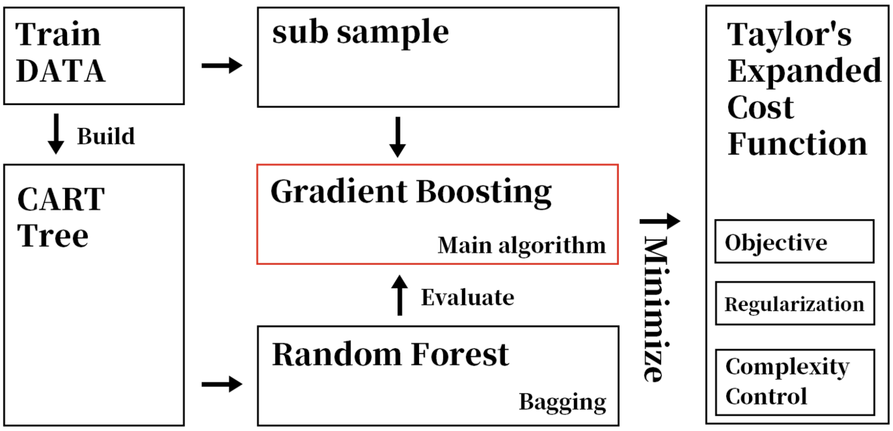


Figure 1: Modeling Pipeline

The model can be represented as:

$$\hat{y} = \sum_{t=1}^{T} f_t(X), f_t \in \mathcal{F} \tag{1}$$

Where $\mathcal{F}$ is the tree structure space of the base learner.

Before specifically describing the mathematical representation of the model, we first define the parameters:

- γ is the minimum gain of the leaf node split
- λ is the L2 regularization coefficient of the leaf node.
- $\mathbb{L}$ is the loss function, which is defined as:

$$\mathbb{L}(x, y) = \frac{(x-y)^2}{2} \tag{2}$$

Then, based on the **traditional GBDT algorithm** , we introduce the second-order Taylor expansion to approximate the loss function:[1]

$$\begin{aligned} {\mathfrak{J}_{\text{Obj}}}^{(t)} &\approx \sum_{i=1}^{n}\left[g_i f_{t(x_i)} + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda \, \|w_j\|^2, \\ g_i &= \partial_{\hat{y}^{t-1}} \mathbb{L}(y_i, \hat{y}^{t-1}), \\ h_i &= \partial^2_{\hat{y}^{t-1}} \mathbb{L}(y_i, \hat{y}^{t-1}), \end{aligned} \tag{3}$$

In the process of tree generation in gradient boosting, we adopt a greedy split strategy to select the best split point by minimizing the gain of the loss function.

$$\begin{aligned} g(I) &= \frac{1}{2}\left[\frac{g_I^2}{h_I+\lambda} + \frac{g_L^2}{h_L+\lambda} + \frac{g_R^2}{h_R+\lambda} - \frac{g^2}{h+\lambda}\right] - \gamma \\ \mathbb{G} &= \sum_{I=1}^{T}\left[\frac{g(I)^2}{h_I+\lambda}\right] + \gamma T \end{aligned} \tag{4}$$

Where I represents the current node, and L and R represent the left and right child nodes after the split.

By repeatedly adding new trees to the model according to $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_{t(x_i)}$, we finally obtained a powerful ensemble model.

2.3. Faucets Determination

As we mentioned in the model assumptions, the number of medals in a country is strongly related to its “strength”. Therefore, we need to obtain an index (or some) that can reflect the strength of a country as a feature of our model.

After analyzed the countries at the top of the medal table, we found that these countries all have some projects with strong “dominance”, such as swimming and track and field in the United States, weightlifting and table tennis in China, gymnastics and diving in Russia, etc. These projects contribute far more to their total medal list than other projects.

To incorporate this factor into the model, we introduce the following features:

- The **dominance** value of some projects with high dominance in that country.
- The **degree of Specialization** of that country.
- The number of medals that the country can provide in this session for some projects with high dominance in its history.

The **dominance** $\mathscr{D}$ is defined as:

$$\mathscr{D} = \frac{\text{gained}_i}{\text{total}_i}, \forall i \in S \tag{5}$$

Where S is the set of all projects that the country participated in this session, gained_i is the number of medals won by the country in that project, and total_i is the total number of medals in that project.

Degree of Specialization $\mathscr{V}$ is defined as the variance of the dominance of all projects:

$$\mathscr{V} = \text{var}(\mathcal{S}) \tag{6}$$

At the same time, we added the host of the Olympic Games as a parameter to deal with the impact of the host advantage on the number of medals.

2.4. Data Cleaning and Preprocessing

We cleaned and preprocessed the data in the following order:

- Merge the country codes and establish a mapping between the country codes and country names.
- Remove missing values caused by the Winter Olympics, wars, etc.
- Added project data for 2028[2].
- According to the detailed athlete data, we calculated the detailed medal data of each country and each project through the established mapping table.
- According to the above mapping table, we calculated the dominance and degree of specialization of each country.

2.5. Model Training

We used the `XGBRegressor` in the `tpot` library and grid search to automate the training process and highly encapsulated it.

Based on these encapsulations, we can automatically adjust some variables to find the optimal model.

Two parameters describing the feature combination are defined: $\mathbb{N}$ and $\mathbb{M}$, where:

- $\mathbb{N}$ represents the number of years of historical data in the parameters.
- $\mathbb{M}$ represents the number of projects with high “dominance” in the parameters.

We then plotted a heatmap of **MAPE**(Mean Absolute Percentage Error) against $\mathbb{N}$ and $\mathbb{M}$:

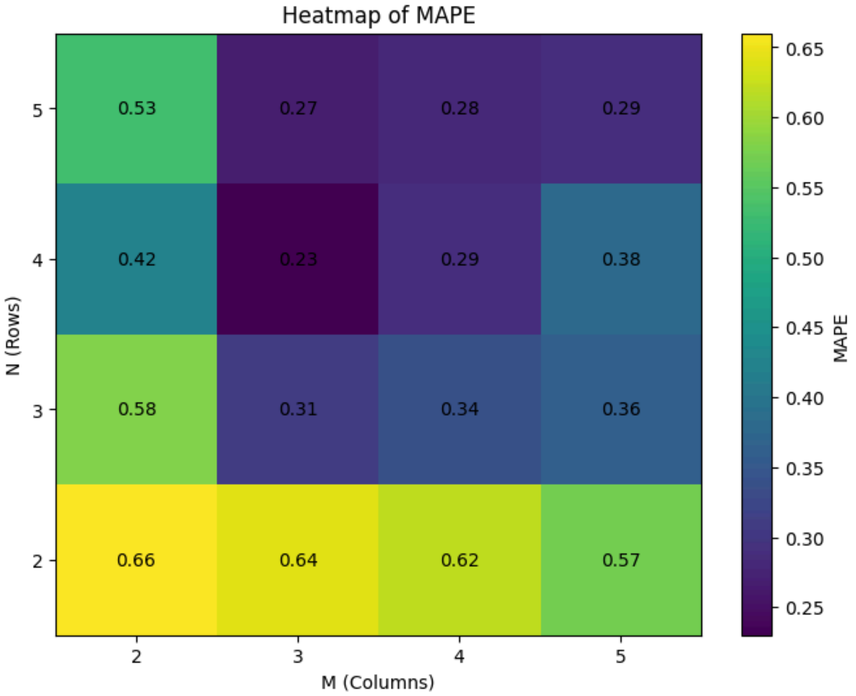


Figure 2: MAPE Heatmap

According to results above, we chose $\mathbb{N} = 3$ and $\mathbb{M} = 2$ as the final feature combination.

2.6. Model Evaluation

After some time of more in-depth training, our model achieved exciting results in predicting the total number of medals.

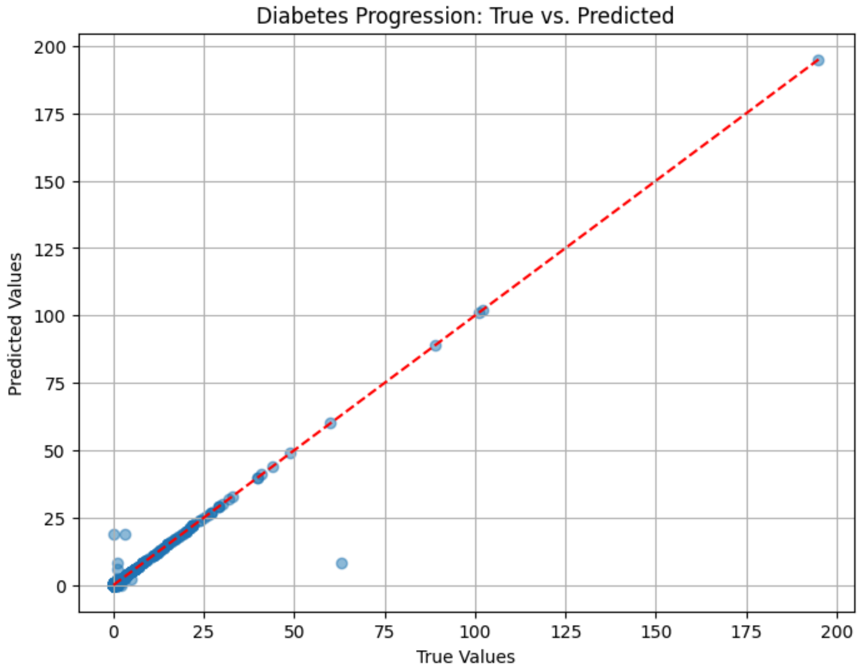


Figure 3: Prediction vs. True Value for Total Medals

Furthermore, we calculated other evaluation data of the model:

Table1: Model Performance

SSE	MSE	MAE	R^2	MAPE
0.002	2.77	0.23	0.96	12.93%

And evaluated the importance of the arrangement of various parameters:

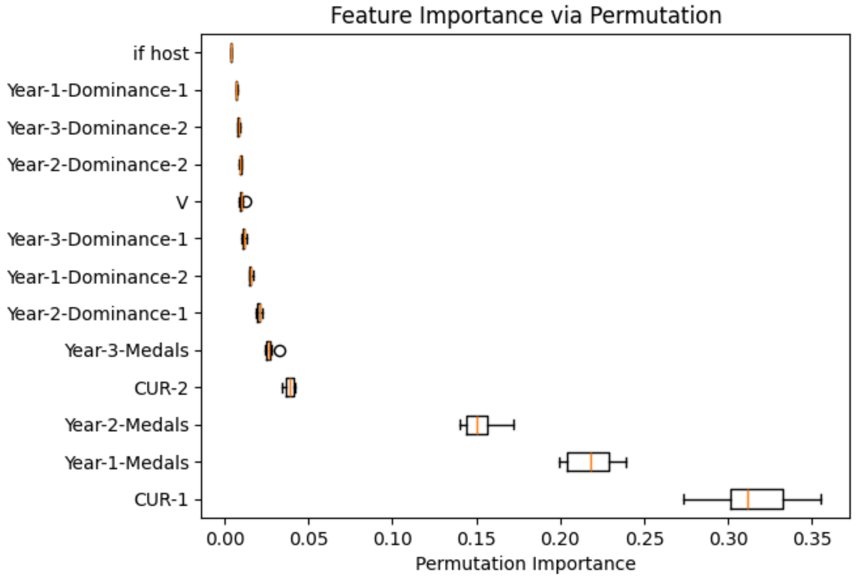


Figure 4: Permutation Importance for each Feature

The chart shows that the number of medals in the country's most dominant project and the number of medals in the country's history in the past two years play the most important role in the model.

Therefore, we believe that the number of medals in the country’s history in the past two years can better reflect the strength of the country, while the number of medals in the country’s most dominant project will determine whether the country can convert its strength into medals.

Finally, we calculated that the 95% confidence intervals for R^2 and percentage error as the score are $[0.844, 1]$ and $[0, 1.021\%]$, respectively. This means that our model has high stability in overall prediction and has a high degree of certainty to ensure the accuracy of the prediction results.

Automating the above process, we obtained the prediction results for bronze, silver, and gold medals, respectively.

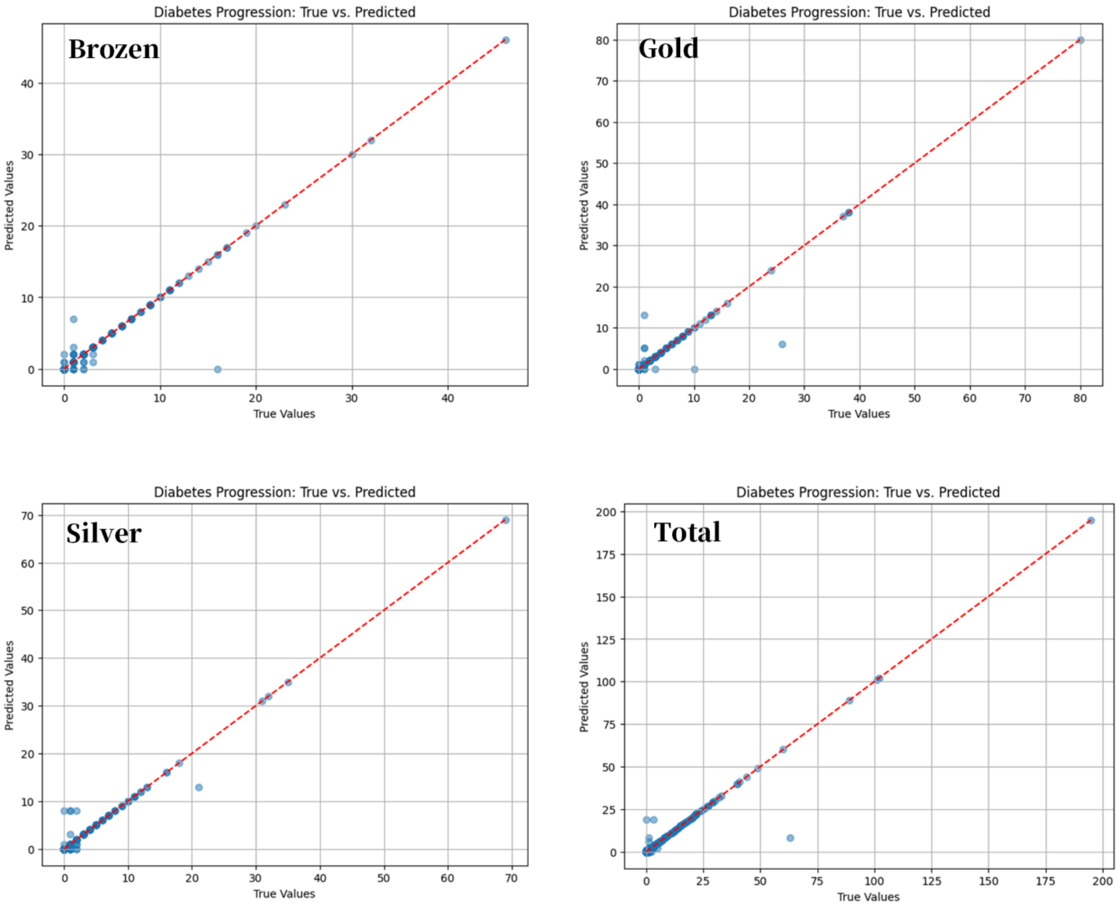


Figure 5: Intuitive Performance of each Model

All the above evaluation processes have rounded the predicted data.

2.7. Results

By substituting the data for 2028, we obtained the original data of the following prediction results:

Table2: Medal Prediction for 2028 Summer Olympics

rank	Gold	Silver	Bronze	Total	NOC	Year
1	45.990759	46.052414	25.373566	117.41674	USA	2028
2	35.17961	27.126156	21.585402	83.89117	CHN	2028
3	21.00542	21.780415	22.603785	65.38962	GBR	2028
4	25.49453	21.445782	14.94941	61.88972	AUS	2028
5	34.74799	8.688086	10.723672	54.159744	JPN	2028
6	17.675909	11.654378	13.032905	42.363194	FRA	2028
7	9.2394495	9.359339	9.370603	27.96939	KOR	2028
8	9.753317	5.24933	12.4193325	27.42198	ITA	2028
9	14.714196	4.5507064	7.990886	27.255789	ESP	2028
10	10.101428	6.1777368	10.032182	26.311348	NED	2028

3. Task 2: Analysis of Important Sports

In this task, we established a machine learning model similar to **Task.1** and used the importance of parameters to analyze the impact of projects on the number of medals in each country.

3.1. Determination of Parameters

This time, we introduced the dominance of all projects as parameters into the model.

3.2. Model Evaluation

Same as **Task.1**, we first made a brief evaluation of the effect of this model:

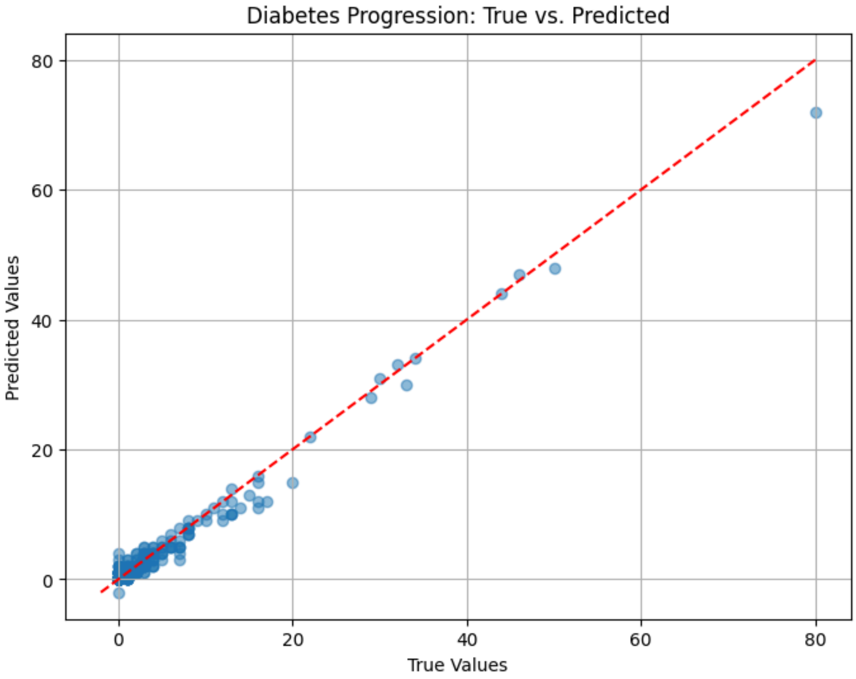


Figure 6: Prediction vs. True Value for Sport

Table3: Model Performance

SSE	**MSE**	**MAE**	**R^2**	**MAPE**
0.31	26.72	1.79	0.88	61.52%

Although the fitting results of the model have decreased compared to **Task.1**, the degree of fitting between the predicted results of the model and the true values is still relatively high from the chart. Therefore, we believe that the model can effectively extract the impact of projects on the number of medals in each country.

Therefore, we can still analyze the impact of projects on the number of medals in each country by calculating the importance of the arrangement of various parameters.

3.3. Results

By calculating the importance of the arrangement of various parameters and taking the top 10, we obtained the following results:

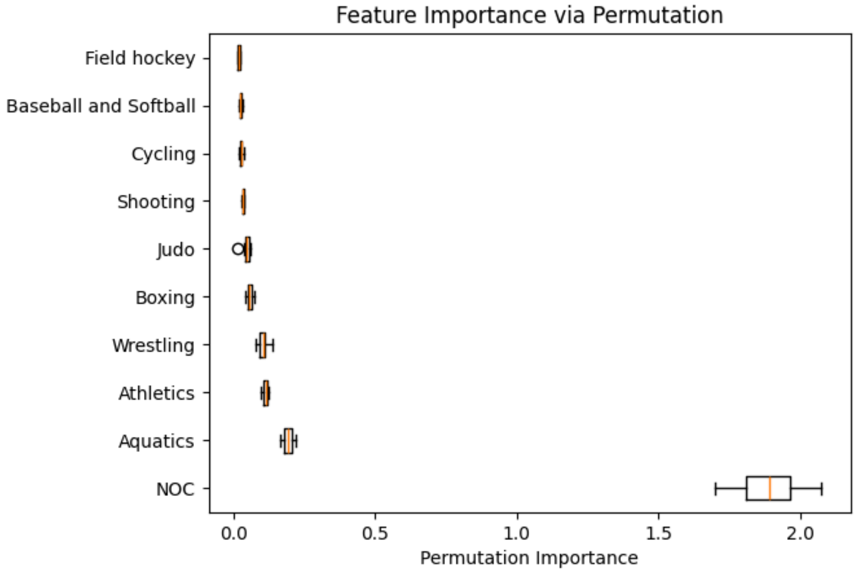


Figure 7: Permutation Importance for each Feature

It can be seen that the number of medals in each country is mainly related to its performance in traditional events such as swimming, track and field, gymnastics, weightlifting, and table tennis. This is also consistent with our expectations.

4. Task 3: Prediction of First-time Medal Winners

4.1. Model Construction

Attempt to Build a Neural Network Model

A neural network model is constructed to fit the characteristics of countries that have won their first medal in past years.

Let $X = [x_1, x_2, x_3]$. Here:

Table4: Variable Definitions

Symbols	**Description**
x_1	The number of editions in which the country has participated without winning any medals since its first participation.
x_2	The number of athletes the country has in this edition
x_3	The average number of medals awarded in this edition.

Use X as the input layer, establish two hidden layers with 4 and 3 neurons respectively, and set the output layer with 1 neuron to represent the probability of winning a medal.

Define y as the medal-winning indicator: $y = \begin{cases} 0 & \text{indicates no medal} \\ 1 & \text{indicates a medal is won} \end{cases}$

4.2. Model Solution

4.2.1. Data Preparation

The dataset is processed based on the files provided on the official website. The samples are split into training and testing sets with a ratio of 0.2.

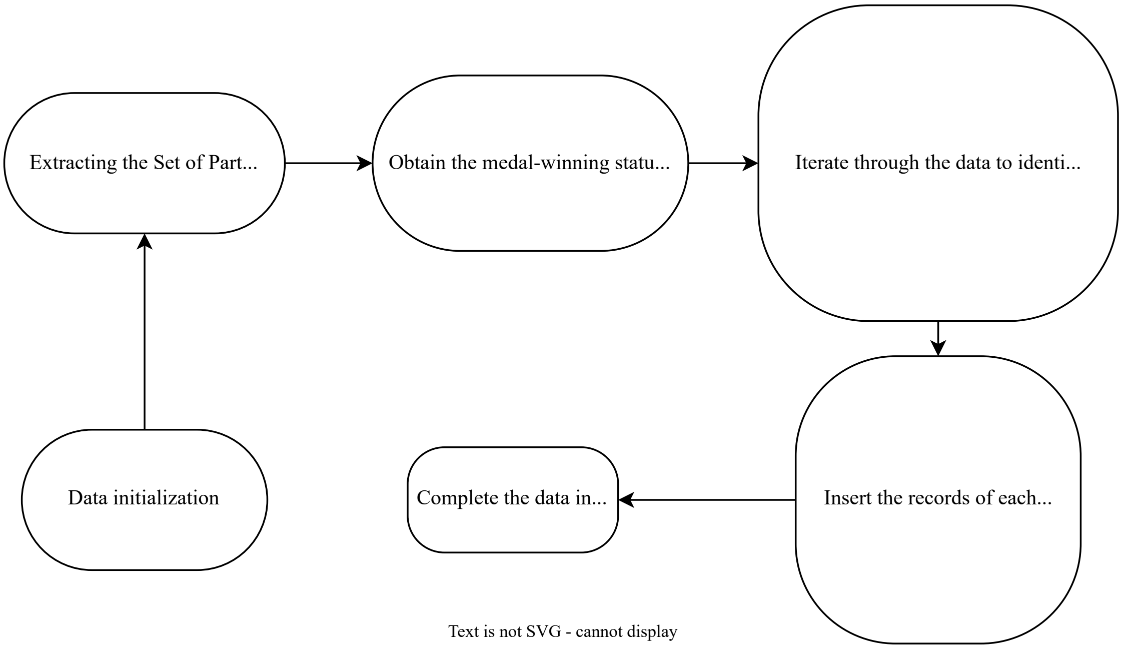

Extracting the Set of Part...
Obtain the medal-winning statu...
Iterate through the data to identi...
Data initialization
Complete the data in...
Insert the records of each...
Text is not SVG - cannot display

Figure 8: Data Preparation Flowchart

4.2.2. Neural Network Framework Construction

Let $a^{(i)}$ represent the activation value of the i-th layer:

$$\begin{aligned} a^{(2)} &= g\big(\Theta^{(1)}X\big) \\ a^{(3)} &= g\big(\Theta^{(2)}a^{(2)}\big) \\ a^{(4)} &= g\big(\Theta^{(3)}a^{(3)}\big) \\ \hat{y} &= a^{(4)} \end{aligned} \tag{7}$$

Where:

- $\Theta^{(i)}$: Propagation matrix
- g: Sigmoid activation function with output range $(0, 1)$

4.2.3. Cost Function

The binary cross-entropy loss function measures the difference between predicted and actual values:

$$J(\Theta) = -\frac{1}{n}\sum_{i=1}^{n}\big[y^{(i)}\log\big(\hat{y}^{(i)}\big) + \big(1-\hat{y}^{(i)}\big)\log\big(1-y^{(i)}\big)\big] \tag{8}$$

4.2.4. Problem Transformation

The model's learning involves minimizing the loss function by optimizing parameters through sample-based learning.

4.2.5. Parameter Optimization

4.2.5.1. Optimization Methods

Use gradient descent method with backpropagation:

Update rule:

$$\Theta^{i}_{jk} := \Theta^{i}_{jk} - \eta\frac{\partial J}{\partial\Theta^{i}_{jk}} \tag{9}$$

With grouped averaging approach:

- Let m = number of sample groups
- Process each group's average gradient per iteration

4.2.5.2. Model Training

Training Procedure:

1. Initialize parameters Θ^1, Θ^2, Θ^3 and hyperparameters
2. Forward propagation: Compute a^2, a^3, a^4
3. Split samples into m groups
4. Backpropagation: Calculate gradients per group
5. Update parameters using group averages
6. Record loss values
7. Repeat until convergence or max epochs

4.3. Model Evaluation

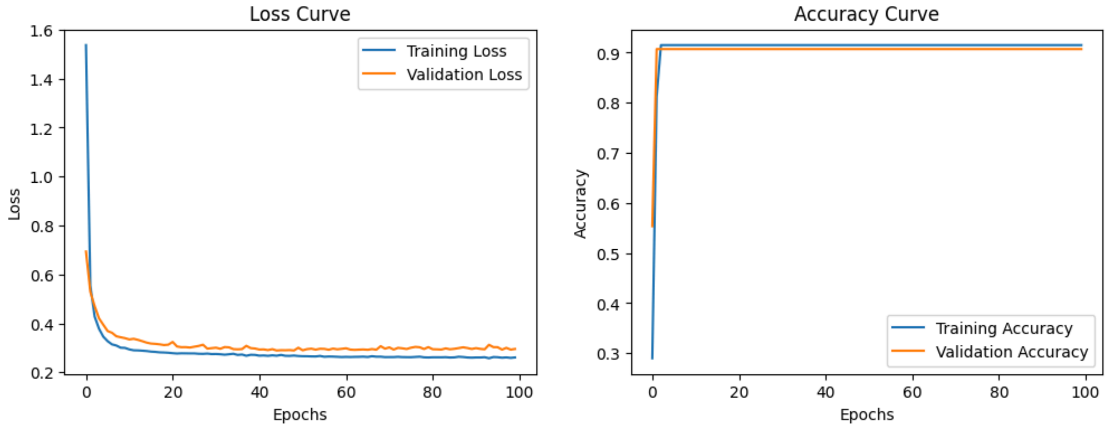


Figure 9: Learning Curve and Accuracy

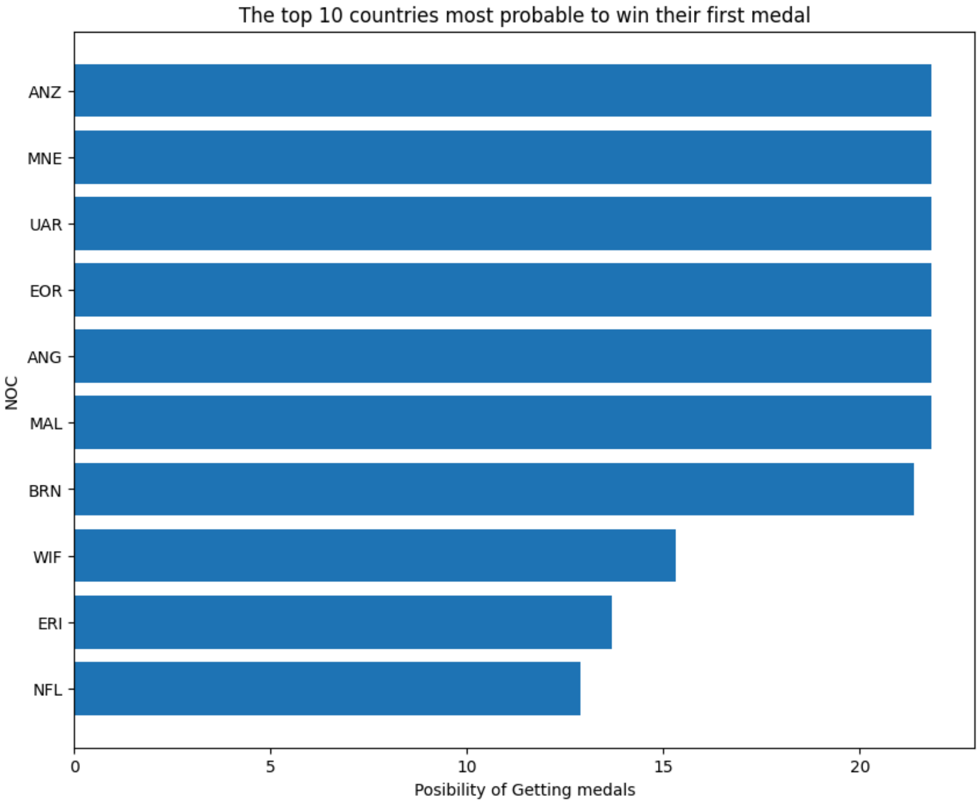


Figure 10: Top 10 Potential First-time Medal Winners

[**Note**]: Only displaying countries with probability > 0.1 for clarity.

5. Task 4: Analysis of the "Great Coach" Effect

5.1. Main tasks

We used mathematical modeling to determine the point when a coach came on board and quantify the contribution of coach turnover to the medal count based on medal data `summerOly_athletes.csv` for Olympic athletes through a change point detection method. We then recorded the country's level before the point of change in a given sport (the detailed quantification process is written in Quantifying). By analyzing the historical level as well as the total number of medals in the historical sport, three countries were selected to find the sports in which they should consider investing in "Great Coach" and to estimate their impact.

5.2. Data preprocessing

We take the medal count (gold, silver, bronze) of a country in a given sport at each Olympics. Because the types of medals (gold, silver, bronze) reflect different variations in the strength of a country in a particular sport, we assign different weights to each type of medal: Gold: 3; Silver: 1; Bronze: 0.5. then:

$$W_t = 3 * \text{Gold}_t + 1 * \text{Silver}_t + 1 * \text{Bronze}_t \tag{10}$$

Where:

- W_t is the weighted total number of medals in a sport for that country in year t.
- Bronze_t is the number of bronze medals won by the country in a sport in year t.
- Silver_t is the number of silver medals won by the country in a sport in year t.
- Gold_t is the number of gold medals won by the country in a sport in year t.

5.3. Bayesian change point detection

5.3.1. Principle

Bayesian change point detection is a method based on Bayesian statistics for identifying change points in time series data. A change point is a point where there is a significant change in the distribution of the data, such as a change in the mean, variance, or trend. Bayesian methods can deal with the change point detection problem flexibly by introducing prior and posterior distributions and providing uncertainty estimates of the location of change points.[3]

5.3.2. Model Assumption

Assume that the weighted medal count W_t in year t is distributed before and after the change point obeys a normal distribution with different means and variances:

$$W_t \propto \mathbb{N}(\mu_k, \sigma_k^2), t \in [t_{k-1}, t_k] \tag{11}$$

Where:

- μ_k is the mean of the kth interval.
- σ_k^2 is the variance of the kth interval.

5.3.3. Model Solution

By analyzing the data$W_1, W_2, ..., W_t$, estimating the location of the change point $t_1, t_2, ..., t_k$ and the mean and variance corresponding to the change point to detect the change point. According to the Bayesian formula, the posterior probability is:

$$p(t_k \mid W) = p(W \mid t_k) \frac{p(t_k)}{p(W)} \tag{12}$$

Where:

- $p(t_k \mid W)$ is likelihood function of the data at change point t_k.

- $p(t_k)$ is prior distribution of the change point t_k, assuming that the probability of the change occurring is uniform, i.e. $p(t_k) = \frac{1}{T}$.
- $p(W)$ is marginal likelihood of the data.

By calculating the posterior probability for each year, we can detect the year of change in the data. Bayesian change point detection can be modeled using the changefinder library in Python. The detection function in changefinder returns a list **Change** containing the change point detection results for each year. Each element is a boolean value indicating whether the year is a change point: True means the year is a change point; False means the year is not a change point. Using the list **Change**, it is possible to identify which years are the years in which the number of medals changed significantly, and thus infer the impact of coaching changes or other factors on the number of medals. We denote the degree of change in the medal count for that year by:

$$\text{Change}_t \times \text{Standardized}(W_t) \tag{13}$$

The standardized weighted medal count is calculated as:

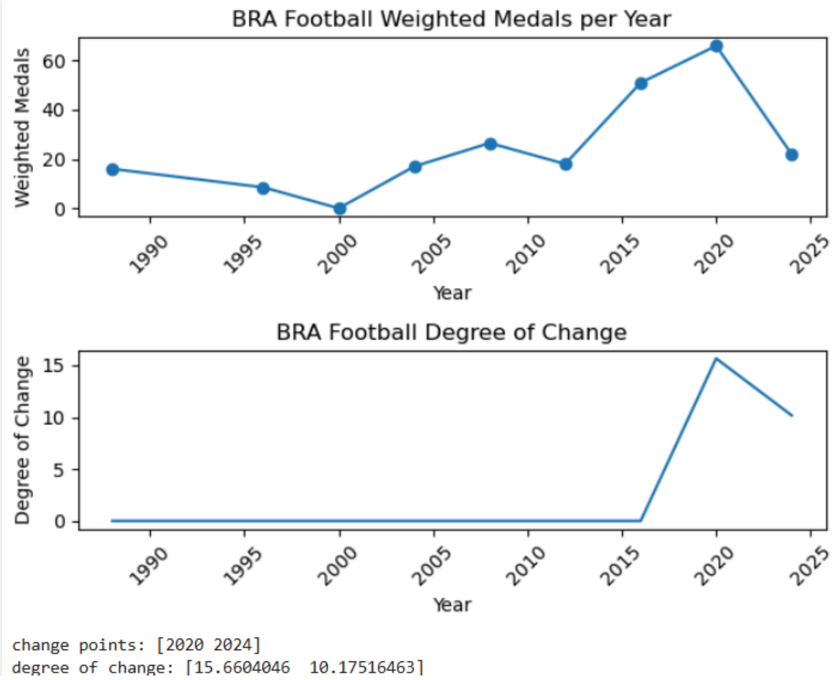
$$\text{Standardized}(W_t) = \frac{W_t - \text{mean}}{\text{std}} \tag{14}$$

Where:

- mean is the mean of W_t.
- std is the variance of W_t.

5.3.4. Analysis of Results

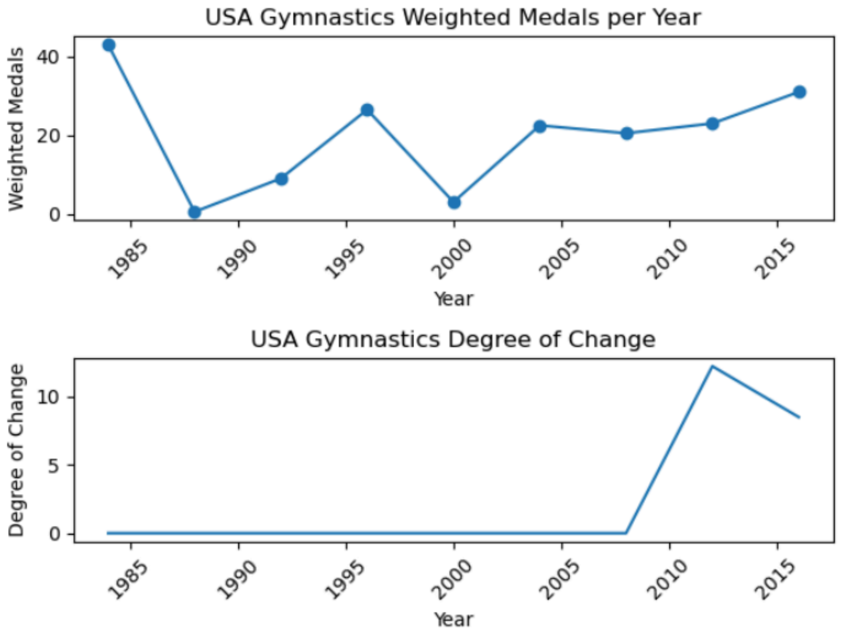
Using Python, we draw a graph of the change in the weighted number of medals for a given country, and the change in the degree of change. The point of change and the degree of change in the values were also derived. The results are given below:



```
change points: [2020 2024]
degree of change: [15.6604046  10.17516463]
```

Figure 11: BRA Football Weighted Medals & Degree of Change

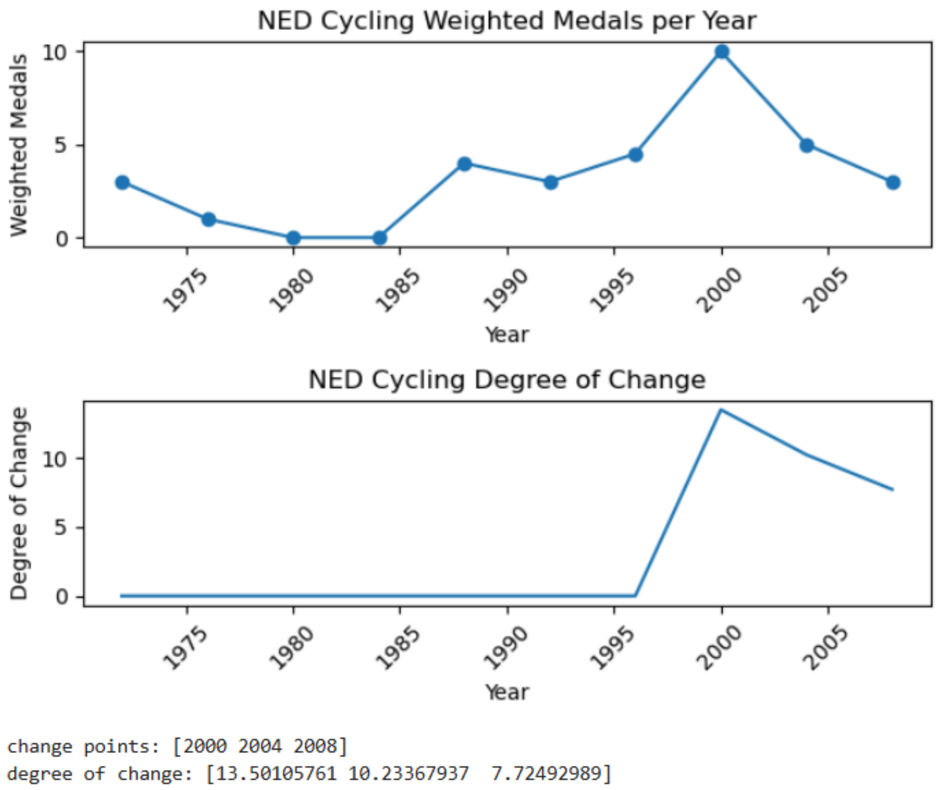
One of the change points (2020) caught our attention, which had a degree of change of 15.66. Tite became the coach of the Brazilian soccer team in 2016, and he led the team to successfully defend its title by winning the gold medal at the 2020 Olympic Games in Tokyo. Tite continues to keep Brazilian soccer competitive at the international level with his flexible tactical adjustments and precise grasp of the players' psychology.[4] As we analyze this information, we can argue that Tite's coaching is the reason for the Brazilian soccer team's surge in medals in 2020.



```
change points: [2012 2016]
degree of change: [12.20041048  8.48226102]
```

Figure 12: USA Gymnastics Weighted Medals & Degree of Change

One of the change points (2012) caught our attention, which had a degree of change of 12.20. Béla Károlyi became the coach of the USA Gymnastics team in 1999, introducing Romanian training methods and improving the overall level of the USA Gymnastics team. [5] When we analyze this information, we can conclude that Béla Károlyi's coaching is the reason for the surge in the number of medals of the U.S. Gymnastics team in 2012.



```
change points: [2000 2004 2008]
degree of change: [13.50105761 10.23367937  7.72492989]
```

Figure 13: NED Cycling Weighted Medals & Degree of Change

One of the change points 2000 caught our attention, this point has a degree of change of 13.50. Max van der Stoep became the coach of the Dutch cycling team in 2000. Under Max van der Stoep's coaching, the Dutch cycling team performed well in the 2004 Olympic Games in Athens and the 2008 Olympic Games in Beijing, winning several medals. The Dutch team has made significant progress in track cycling, becoming one of the world's strongest teams in the sport. [6] When we analyze this information, we can conclude that Max van der Stoep's coaching was the reason for the Dutch cycling team's surge in medals in 2012.

The three data sets described above are evidence of changes that may be caused by the **"great coach"** effect. Next, we will quantify the contribution of this effect to medal counts.

5.4. Quantifying the level of a country in a given project

5.4.1. Calculating a Coach's Contribution

We use weighted medal counts to measure the difference between before and after a coach takes office. The average of the weighted medal counts was first calculated.

The average of the weighted medal counts before the change point is calculated as:

$$W_{\text{average,before}} = \left(\frac{1}{T_{\text{before}}}\right) \sum_{t\in \text{ before}} (W_t) \tag{15}$$

The average of the weighted medal counts after the change point is calculated as:

$$W_{\text{average,after}} = \left(\frac{1}{T_{\text{after}}}\right) \sum_{t\in \text{ after}} (W_t) \tag{16}$$

Where:

- T_{before} is the number of years before the change point.
- T_{after} is the number of years after the change point.

We can calculate the contribution rate of a coach by comparing the change in the number of medals before and after the change point. Define the contribution rate as:

$$\text{Contribution} = \frac{W_{\text{average,after}} - W_{\text{average,before}}}{W_{\text{average,before}}} \tag{17}$$

If the contribution rate is greater than 0, it means that the arrival of the coach has had a positive impact on the number of medals. The contribution rate reflects the magnitude of the increase in the number of medals, with larger values indicating a more significant change, i.e., the greater the change caused by the "great coach" effect.

5.4.2. Analysis of Results

The results can be calculated through Python and are as follows:

Country	Sport	Contribution
BRA	Football	1.25
USA	Gymnastics	0.51
NED	Cycling	1.71

The coaching contribution rate for both the Brazilian soccer team and the Dutch cycling team is greater than 1, indicating that the weighted number of medals increased by more than a factor of one after the change point, suggesting that the impact of coaching is significant. In contrast, the U.S. Gymnastics team's coaching contribution rate was 0.51, which represents a relatively slow increase in weighted medal counts relative to the Brazilian soccer team and the Dutch cycling team, suggesting that the impact of coaching is less significant.

5.5. Choosing to invest in a sport with a "great coach"

5.5.1. Expansion of data

To better select investment projects as well as predict contribution rates, we increased the three data sets described above to eight. The added data are as follows:

NOC	Sport	ChangeYear	Contribution	GreatCoach
CHN	Volleyball	2016	0.49	郎平
USA	Gymnastics	2012	0.51	Béla Károlyi
NED	Cycling	2000	1.71	Max van der Stoep
CHN	Table Tennis	2016	0.34	刘国梁
FRA	Fencing	2000	0.26	Pierre Louaillier
BRA	Football	2020	1.25	Tite
KEN	Athletics	2008	1.33	Carlos Lopes
GBR	Swimming	2016	9.15	Daniel Jamieson&Paul Newsome

5.5.2. Quantifying the level of a country in a given project

Unlike the weighted medal count, we also need to consider the total number of participants. Because a country didn't win a medal in a certain event doesn't mean it doesn't have any competitiveness in that event. So we define the **level** as:

$$\text{Level} = 4 * \text{Gold}_t + 3 * \text{Silver}_t + 1 * \text{Bronze}_t + 0.5 * \text{No medal}_t \tag{18}$$

Where:

- No medal_t is the number of sports in which the country competed in year t but did not win a medal.

We calculate the average **Level** for the five years before the point of change as the historical level of a state in a given program, and obtain the following data:

NOC	Sport	Level
CHN	Volleyball	19.2
USA	Gymnastics	59.3
NED	Cycling	11.5
CHN	Table Tennis	34.7
FRA	Fencing	42.3
BRA	Football	44.9
KEN	Athletics	33.2
GBR	Swimming	32.6

Averaging this out to 32.6, we believe that the "great coach" effect is more likely to occur when a country's historical ability in a particular program is around 32.6.

5.5.3. Choose sports that are prone to the "great coach" effect

By counting the total number of medals for Sport in the eight data sets

$$\text{Total} = \text{Gold}_t + \text{Silver}_t + \text{Bronze}_t, t = \text{Change Year} \tag{19}$$

We get:

Sport	Number of medals
Volleyball	72
Gymnastics	72
Cycling	67
Table Tennis	72
Fencing	67
Football	72
Athletics	72
Swimming	72

With an average of 70.75, we believe that Sports with a total medal count of around 70.75 are more likely to experience the "great coach" effect. Based on the data from the 2028 Olympic Games, we selected three eligible sports: Artistic Gymnastics (total medals: 67), Water Polo (total medals: 78) and Wrestling (total medals: 72).

5.5.4. Selection of countries based on historical level

Define: $\text{Level}_{\text{Sport}_{t_1}} \sim t_2$ as the average level of a certain country in a certain program from t_1 to t_2.

We calculated each country's $\text{Level}_{\text{Gymnastics},(2008\sim2024)}$, $\text{Level}_{\text{Water Polo},(2008\sim2024)}$, and $\text{Level}_{\text{Wrestling},(2008\sim2024)}$ as their respective historical level in this program. Combining the above data, we have selected the following three countries and programs that are suitable for investing in "great coaches":

Country	Sport	Level
GER	Artistic Gymnastics	29.25
SRB	Water Polo	36.4
JPN	Wrestling	24.7

5.5.5. Estimating the contribution of their "great coaches"

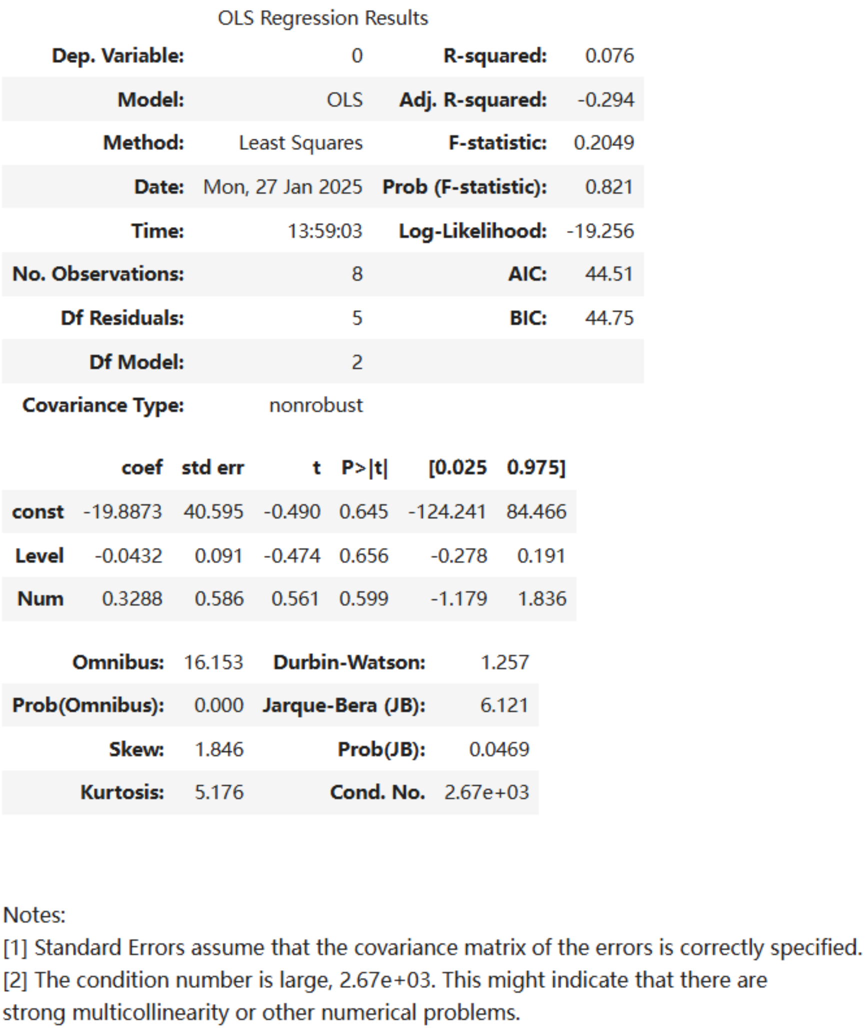
The data we have so far is shown below:

NOC	Sport	Contribution	Level	Number of medals
CHN	Volleyball	0.49	19.2	72
USA	Gymnastics	0.51	59.3	72
NED	Cycling	1.71	11.5	67
CHN	Table Tennis	0.34	34.7	72
FRA	Fencing	0.26	42.3	67
BRA	Football	1.25	44.9	72
KEN	Athletics	1.33	33.2	72
GBR	Swimming	9.15	32.6	72
GER	Artistic Gymnastics	nan	29.25	67
SRB	Water Polo	nan	36.4	78
JPN	Wrestling	nan	24.7	72

It can be seen that when the “Great Coach” effect occurs, Number of medals are all above and below 70.75, with an extreme variance of 5. Let’s try to build a multiple linear regression model：

$$\text{Contribution} = \beta_0 + \beta_1 * \text{Level} + \beta_2 * \text{Number of medals} \tag{20}$$

The model can be solved using the least squares method from the Statsmodels library in Python and the results obtained are shown below:



OLS Regression Results

Dep. Variable:	0	**R-squared:**	0.076
Model:	OLS	**Adj. R-squared:**	-0.294
Method:	Least Squares	**F-statistic:**	0.2049
Date:	Mon, 27 Jan 2025	**Prob (F-statistic):**	0.821
Time:	13:59:03	**Log-Likelihood:**	-19.256
No. Observations:	8	**AIC:**	44.51
Df Residuals:	5	**BIC:**	44.75
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-19.8873	40.595	-0.490	0.645	-124.241	84.466
Level	-0.0432	0.091	-0.474	0.656	-0.278	0.191
Num	0.3288	0.586	0.561	0.599	-1.179	1.836

Omnibus:	16.153	**Durbin-Watson:**	1.257
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6.121
Skew:	1.846	**Prob(JB):**	0.0469
Kurtosis:	5.176	**Cond. No.**	2.67e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.67e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 14: OLS Regression Results.png

From the results, the **t-test** result of **Number of medals** is not satisfactory, and the t-statistic corresponding to **Number of medals** has a P-value of 0.599, which shows that the linear relationship between **Number of medals** and Contribution is not significant. Combined with the previous results that the extreme deviation of **Number of medals** (extreme deviation = 5) is small, we believe that when the "great coach" effect occurs, the influence of **Number of medals** on the contribution of coaches is very small. The t-statistic corresponding to **Level** corresponds to a **P-value** of 0.656, and we get The linear relationship between **Level** and **Contribution** is not significant.

We conjecture that **probably** the country's technological level as well as economic level will amplify or attenuate the "great coaches" effect. This is because upgrading high-quality professional training facilities can help a coach achieve the training effect he wants. The coach can also use science and technology to analyze more appropriate training methods and research more effective countermeasures, thus increasing the number of medals. We are limited by the small amount of data available for the "Great Coach" effect, as well as the requirement of the question that our model and data analysis must use only the data set provided. Therefore, we can only predict a vague coaching contribution rate by comparing the data, and the prediction is as follows:

NOC	Sport	Contribution	Level	Number of medals
GER	Artistic Gymnastics	0.7	29.25	67
SRB	Water Polo	1	36.4	78
JPN	Wrestling	0.5	24.7	72

So we suggest that **Germany** might consider investing in "Great Coaches" on **Artistic Gymnastics** with an expected coaching contribution of 0.7, **Serbia** might consider investing in "Great Coaches" on **Water Polo** with an expected coaching contribution of 1, and **Japan** might consider investing in "Great Coaches" on **Wrestling** with an expected coaching contribution of 0.5.

6. Task 5: Analysis of other factors.

When discussing the “great coach” effect, we analyzed that the country’s economic level may enhance or weaken the “great coach” effect. So we conjectured that when the “great coach” effect does not exist, the economic level will also affect the number of medals. The following is the trend chart of the total number of medals won by the Chinese team from 1984 to 2024:

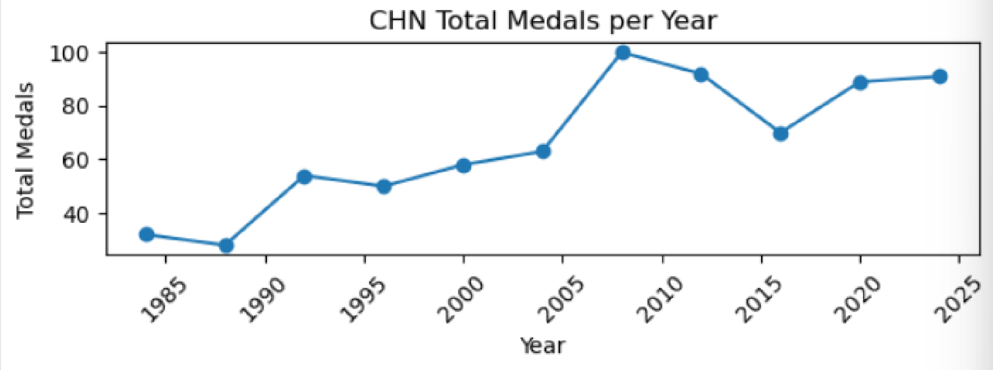


Figure 15: China Medal Trend

From 1984 to 2024, China’s economic level has been continuously improving. By excluding the data from the 2008 Beijing Olympics (to eliminate the impact of the host country), we found that China’s total medal count is also on the rise. At this point, we believe that the economic level will affect the number of Olympic medals.

The following is the trend chart of the total number of medals won by the Cuban team from 1984 to 2024:

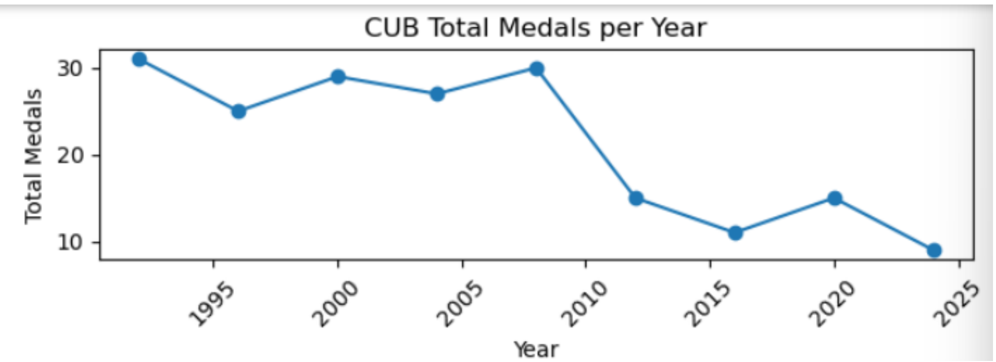


Figure 16: Cuba Medal Trend

From 1996 to 2020, Cuba’s economy showed an overall growth trend, but the total number of medals showed a downward trend. We found that Cuba’s population is much smaller than China’s, and Cuba has a phenomenon of talent loss in sports. Therefore, we believe that population size will also affect the number of Olympic medals.

The following is the trend chart of the total number of medals won by the Indian team from 1984 to 2024:

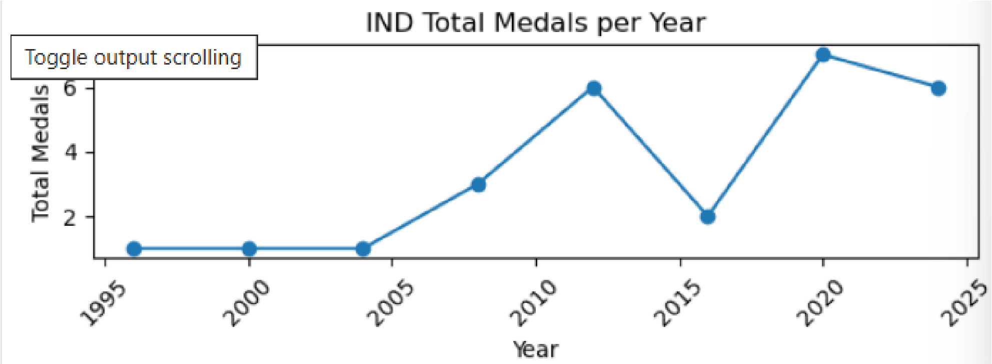


Figure 17: India Medal Trend

From 2000 to 2020, India’s economy grew rapidly, becoming one of the fastest-growing economies in the world. India also has a relatively large population, but the total number of medals has not shown a clear upward trend. We speculate that it may be due to insufficient national infrastructure and economic distribution issues that the proportion of the population converted into athletes is relatively small, thereby affecting the change in the total number of medals.

From this, we conclude that the **economic level**, **population size**, and **infrastructure level** will affect the change in the number of medals to a certain extent.

References

[1] W. Contributors, “Gradient boosting.” [Online]. Available: https://en.wikipedia.org/wiki/Gradient_boosting

[2] w., “2028 Summer Olympics.” [Online]. Available: https://en.wikipedia.org/wiki/2028_Summer_Olympics

[3] M. Adams, “The right direction?,” *SecEd*, vol. 2007, 2007, doi: 10.12968/sece.2007.11.1018.

[4] R. d. J., “The Journey to Gold: Brazil's Olympic Triumph. .” 2021.

[5] “Coaching Team - Bela and Martha Karolyi • USA Gymnastics.” [Online]. Available: https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/

[6] D. C. F., “Dutch Cycling: A Journey to Excellence..” 2008.

Report on Use of AI

1. **Deepseek Deepseek (Open score version,Deepseek-R1-665B)**

Queary 1:

```
Please translate the following markdown content into a typst code.
== 3.1 Model Construction
Attempt to Build a Neural Network Model

A neural network model is constructed to fit the characteristics of countries that have
won their first medal in past years.

Let $X=[x_1,x_2,x_3]$ Here ,
| Symbols | meaning                          |
|---------|----------------------------------|
| $x_1$   | The number of editions in which the country has participated without
winning any medals since its first participation. |
| $x_2$   | The number of athletes the country has in this edition        |
| $x_3$   | The average number of medals awarded in this edition.               |


Use $X$ as the input layer, establish two hidden layers with 4 and 3 neurons
respectively, and set the output layer with 1 neuron to represent the probability of
winning a medal.

$y$ represents whether a medal is won in the next edition, where $y=0$ indicates no
medal,and $y=1$ indicates a medal is won.

== 3.2 Model Solution

=== 3.2.1 Data Preparation
The dataset is processed based on the files provided on the official website. The
samples are split into training and testing sets with a ratio of 0.2.

The flowchart of sample data preparation is represented as follows.

![C_2data initialization.png](<attachment:C_2data initialization.png>)

=== 3.2.2 Neural Network Framework Construction
$a^{(i)}$ represents the activation value of the i-th layer.

$a^{(2)}=g(\Theta^{(1)}X)$

$a^{(3)}=g(\Theta^{(2)}a^{(2)})$

$a^{(4)}=g(\Theta^{(3)}a^{(3)})$

$\hat y=a^{(4)}$

Here, $\Theta^{(i)}$ represents the propagation matrix, and $g$ denotes the activation
function. In this case, the sigmoid function is used as the activation function, with
an output range of $(0,1)$

=== 3.2.2 Cost Function

The binary cross-entropy loss function measures the difference between the predicted
value $\hat y$ and the actual value $y$ .


$$
```

```
J(\Theta)=-\frac{1}{n}\sum^n_{i=1}[y^{(i)}log(\hat y^{(i)})+(1-\hat y^{i})]
$$

=== 3.2.3 Problem Transformation

The model's learning involves minimizing the loss function by optimizing parameters
through sample-based learning.

=== 3.2.4 Parameter Optimization
==== 3.2.4.1 Optimization Methods
Use the gradient descent method to optimize the loss function. By iterating over each
sample, calculate the gradients using backpropagation to obtain $\frac{\partial J}
{\partial \Theta^{(i)}_{jk}}$
Let $\eta$ learning rate.

The parameter update rule is given by:
$$
\Theta^{(i)}_{jk}:= \Theta^{(i)}_{jk}-\eta \frac{\partial J}{\partial \Theta^{(i)}
_{jk}}
$$
This process is repeated iteratively until convergence or the maximum number of
iterations is reached.

Since gradient descent requires repeated calculations over a large number of samples to
compute the average, to improve algorithm performance, we adopt a grouped averaging
approach for descent.Let$m$represent the number of groups into which the samples will
be divided. In each iteration, every group is processed, and the descent is performed
using the average gradient of each group.

==== 3.2.4.2 Model Training
The main steps are shown in the following:

**step1**:Initialize the parameters $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}$ the
maximum number of iterations (epochs), and the number $m$ of groups.

**step2**:Forward propagation to compute $a^{(2)},a^{(3)},a^{(4)}$

**step3**:Group the samples based on  $m$ , which represents the number of groups

**step4**: For each sample group, calculate the gradient of each parameter using the
backpropagation method.

**step5**: Perform gradient descent for each parameter after obtaining the average
gradient of each sample group, and repeat until all groups have completed their
iterations.

**step6**: Record the loss function value after each iteration.

**step7**: Determine whether the maximum number of iterations has been reached. If
satisfied, stop the training; otherwise, repeat steps 2 to 7.

== 3.3 Model evaluation
Through the trained model, the learning curve and accuracy curve are observed, and the
fitting performance is as follows:

![C_2_loss_Accuracy.png](attachment:C_2_loss_Accuracy.png)

It can be observed that the classifier trained by this model demonstrates high accuracy
and a low loss value.
```

```
The final model predicts the probability of countries that have not previously won a
medal winning a medal in the next event. The results are shown in the figure below.
***Since many countries have a probability close to zero, only the top 10 countries
most likely to win their first medal are displayed.***

![The top 10 countries most probable to win first medal.png](<attachment:The top 10
countries most probable to win first medal.png>)
```

Output 1:

```
== Model Construction

Attempt to Build a Neural Network Model

A neural network model is constructed to fit the characteristics of countries that have
won their first medal in past years.

Let $X = [x_1, x_2, x_3]$. Here:
#figure(
  table(
    columns: (1.2fr, 4fr),
    inset: 4pt,
    stroke: none,
    align: (center, left),
    [*Symbols*], [*Meaning*],
    [$x_1$], [The number of editions in which the country has participated without
winning any medals since its first participation.],
    [$x_2$], [The number of athletes the country has in this edition],
    [$x_3$], [The average number of medals awarded in this edition.]
  ),
  caption: [Variable Definitions],
)<tab-symbols>

Use $X$ as the input layer, establish two hidden layers with 4 and 3 neurons
respectively, and set the output layer with 1 neuron to represent the probability of
winning a medal.

Define $y$ as the medal-winning indicator:
$y = cases(
  0 quad "indicates no medal",
  1 quad "indicates a medal is won"
)$

== Model Solution

=== Data Preparation
The dataset is processed based on the files provided on the official website. The
samples are split into training and testing sets with a ratio of 0.2.

#figure(
  image("C_2data_initialization.png", width: 80%),
  caption: [Data Preparation Flowchart],
)<fig-flow>

=== Neural Network Framework Construction
Let $a^{(i)}$ represent the activation value of the i-th layer:

$
a^{(2)} = g(Theta^{(1)} X) \
a^{(3)} = g(Theta^{(2)} a^{(2)}) \
a^{(4)} = g(Theta^{(3)} a^{(3)}) \
hat(y) = a^{(4)}
$

Where:
- $Theta^{(i)}$: Propagation matrix
- $g$: Sigmoid activation function with output range $(0,1)$

=== Cost Function
The binary cross-entropy loss function measures the difference between predicted and
actual values:

$
J(Theta) = -1/n sum_(i=1)^n [y^{(i)} log(hat(y)^(i)) + (1 - hat(y)^(i)) log(1 -
y^{(i)})]
$

=== Problem Transformation
The model's learning involves minimizing the loss function by optimizing parameters
through sample-based learning.

=== Parameter Optimization
==== Optimization Methods
Use gradient descent method with backpropagation:

Update rule:
$
Theta_(jk)^(i) := Theta_(jk)^(i) - eta (del J)/(del Theta_(jk)^(i))
$
With grouped averaging approach:
- Let $m$ = number of sample groups
- Process each group's average gradient per iteration

==== Model Training
**Training Procedure**:
1. Initialize parameters $Theta^(1)`, $Theta^(2)`, $Theta^(3)` and hyperparameters
2. Forward propagation: Compute $a^(2)`, $a^(3)`, $a^(4)`
3. Split samples into $m$ groups
4. Backpropagation: Calculate gradients per group
5. Update parameters using group averages
6. Record loss values
7. Repeat until convergence or max epochs

== Model Evaluation
#figure(
  image("C_2_loss_Accuracy.png", width: 90%),
  caption: [Learning Curve and Accuracy],
)<fig-eval>

#figure(
  image("top_10_countries.png", width: 100%),
  caption: [Top 10 Potential First-time Medal Winners],
)<fig-results>

[*Note*]: Only displaying countries with probability > 0.1 for clarity.
```

1. **Github `Copilot` (January 17, 2025, `Release`, Copilot-v1.0)**

Auto-completions for code used in cleaning and preprocessing data for Task.**1**,.**2** models.

1. **OpenAI `ChatGPT` (January 17, 2025, `Release`, GPT-o1-min)**

Queary 1:

```
Please help me find ten examples of great coaches who can greatly impact the number of
Olympic medals.
```

Output 1:

```
1. Lang Ping and the Chinese Women's Volleyball Team
Coach: Lang Ping
Sport: Volleyball
Coaching Period: 1995-1999, 2013-2021
Achievements:
During her first tenure from 1995 to 1999, Lang Ping led the Chinese women's volleyball
team out of a trough, winning the silver medal at the 1996 Atlanta Olympics and the
silver medal at the 1998 World Championships.
From 2013 to 2021, Lang Ping returned to coach the Chinese women's volleyball team,
leading the team to win the 2015 World Cup championship, the 2016 Rio Olympics gold
medal, and the 2019 World Cup championship.
Impact: Lang Ping's coaching not only enhanced the team's technical and tactical level
but also strengthened the players' self-confidence and team cohesion. Her training
methods and management style significantly improved the competitiveness of the Chinese
women's volleyball team on the international stage.

2. Béla Károlyi and the Romanian and American Women's Gymnastics Teams
Coach: Béla Károlyi
Sport: Gymnastics
Coaching Period: Late 1970s to 1980s (Romania), 1999 to early 2000s (USA)
Achievements:
Romanian Gymnastics Team: Under the coaching of Béla Károlyi, the Romanian gymnastics
team became a world gymnastics powerhouse. At the 1980 Moscow Olympics, the Romanian
gymnastics team won the team gold medal, and Nadia Comăneci became the first gymnast to
score a perfect 10 at the Olympics.
American Gymnastics Team: After joining the American gymnastics team in 1999, he
introduced Romanian training methods, quickly improving the overall level of the
American gymnastics team. At the 2000 Sydney Olympics, the American women's gymnastics
team won the team silver medal; at the 2012 London Olympics, the American women's
gymnastics team won the team gold medal, and Gabby Douglas won the all-around gold
medal.
Impact: Béla Károlyi's strict training methods and technological innovations greatly
improved the athletes' technical level and competitive performance. His coaching style
and training system played a key role in the rise of the Romanian and American
gymnastics teams.
Max van der Stoep and the Dutch Cycling Team

3.Coach: Max van der Stoep
Sport: Cycling
Coaching Period: 2000-2008
Achievements:
Under the coaching of Max van der Stoep, the Dutch cycling team performed outstandingly
at the 2004 Athens Olympics and the 2008 Beijing Olympics, winning multiple medals. The
Dutch team made significant progress in track cycling and became one of the world's top
teams in the sport.
Impact: Van der Stoep introduced new training methods and tactical systems, improving
the athletes' physical fitness and technical level. His coaching not only enhanced the
competitive level of the Dutch cycling team but also promoted the overall development
of Dutch cycling.

4. He Zhenliang and the Chinese Table Tennis Team
Coach: He Zhenliang
Sport: Table Tennis
Coaching Period: 1980s
```

Achievements:
During the 1980s, He Zhenliang served as the head coach of the Chinese table tennis team, leading the team to achieve excellent results in multiple World Championships and Olympics, establishing the dominant position of the Chinese table tennis team in the world table tennis arena.
Impact: Zhenliang's coaching style and training methods greatly improved the overall level of the Chinese table tennis team, cultivating many world champions and Olympic champions, such as Deng Yaping and Qiao Hong.

5. Fabio Capello and the Italian and English Football Teams
Coach: Fabio Capello
Sport: Football
Coaching Period: 1990s (Italy), 2008-2011 (England)
Achievements:
Italian Football Team: During the 1990s, Capello coached AC Milan and Juventus, leading the teams to win multiple Serie A league championships and UEFA Champions League titles.
English Football Team: From 2008 to 2011, Capello coached the England national team, leading the team to achieve certain results at the 2010 World Cup.
Impact: Capello's tactical system and management style greatly improved the competitive level of the teams, cultivating many world-class players, such as Paolo Maldini and Francesco Totti.

6. Pierre Louaillier and the French Fencing Team
Coach: Pierre Louaillier
Sport: Fencing
Coaching Period: 2000-2008
Achievements:
Under the coaching of Pierre Louaillier, the French fencing team performed outstandingly at the 2004 Athens Olympics and the 2008 Beijing Olympics, winning multiple medals. The French team made significant progress in fencing and became one of the world's top teams in the sport.
Impact: Louaillier introduced new training methods and tactical systems, improving the athletes' technical level and competitive performance. His coaching not only enhanced the competitive level of the French fencing team but also promoted the overall development of French fencing.

7. Igor Kovalyov and the Russian Figure Skating Team
Coach: Igor Kovalyov
Sport: Figure Skating
Coaching Period: 1990s-early 2000s
Achievements:
Under the coaching of Igor Kovalyov, the Russian figure skating team won multiple World Championships and Olympic gold medals in the 1990s and early 2000s.
Impact: Kovalyov's training methods and technological innovations greatly improved the athletes' technical level and artistic expression. His coaching style and training system played a key role in the rise of the Russian figure skating team, cultivating many world champions and Olympic champions, such as Irina Slutskaya and Maria Butyrskaya.

8. Joe Mazurkowski and the Polish Volleyball Team
Coach: Joe Mazurkowski
Sport: Volleyball
Coaching Period: 1980s-1990s
Achievements:
Under the coaching of Joe Mazurkowski, the Polish volleyball team won multiple European Championship and World Championship medals in the 1980s and 1990s.
Impact: Mazurkowski's training methods and tactical systems greatly improved the overall level of the Polish volleyball team, cultivating many world-class players, such as Tomasz Zaremba and Marek Sadowski. His coaching not only enhanced the competitive

level of the Polish volleyball team but also promoted the overall development of Polish volleyball.

Summary

These "great coaches" have significantly improved the performance of athletes and teams by introducing advanced training methods, strict management, and technological innovations. Their success is not only reflected in the increase in the number of medals but also in the cultivation of many world champions and Olympic champions, which has had a profound impact on the development of sports projects. These examples fully demonstrate the important role of coaches in sports competitions and provide valuable references for the recruitment of coaches and the formulation of sports policies by national Olympic committees.