Problem Chosen	**2024**	**Team Control Number**
A	**MCM/ICM** **Summary Sheet**	**00000**

--

Summary

–

Keywords: --, --, --, --

Contents

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. Literature Review

1.3. 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

In this task, 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.

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:

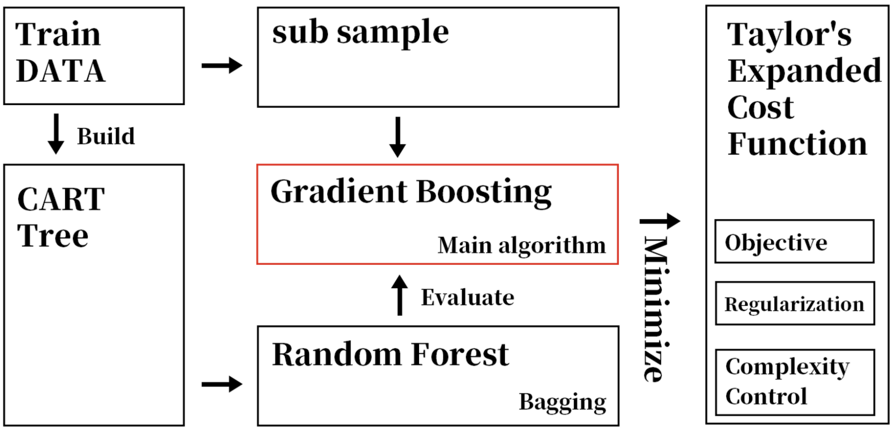

Figure 1: Modeling Pipeline

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:

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

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.

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

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 $\texttt{total}_i$ is the total number of medals in that project.

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

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

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.
- 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 $\mathbb{M}$:

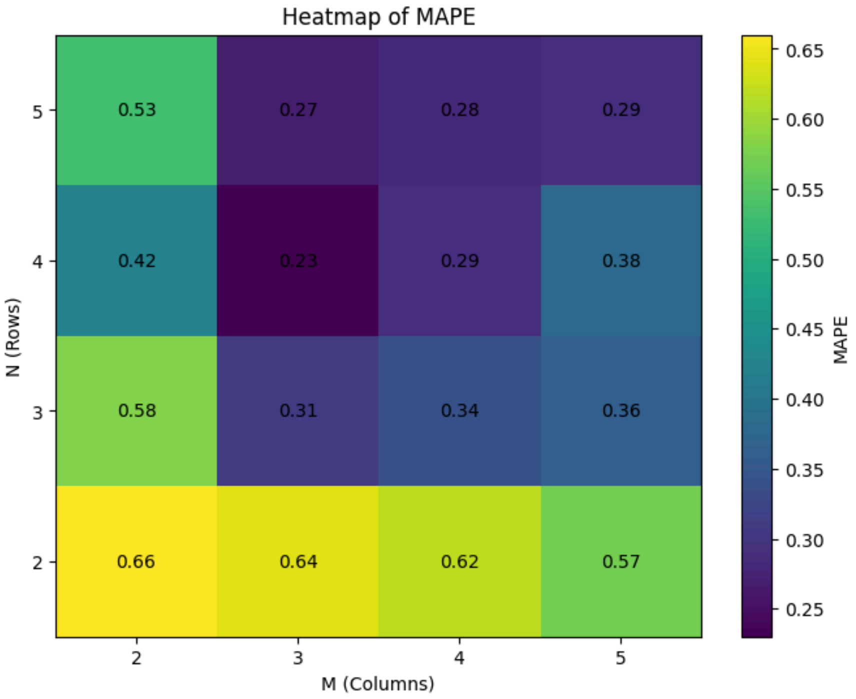

Figure 2: MAPE Heatmap

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.

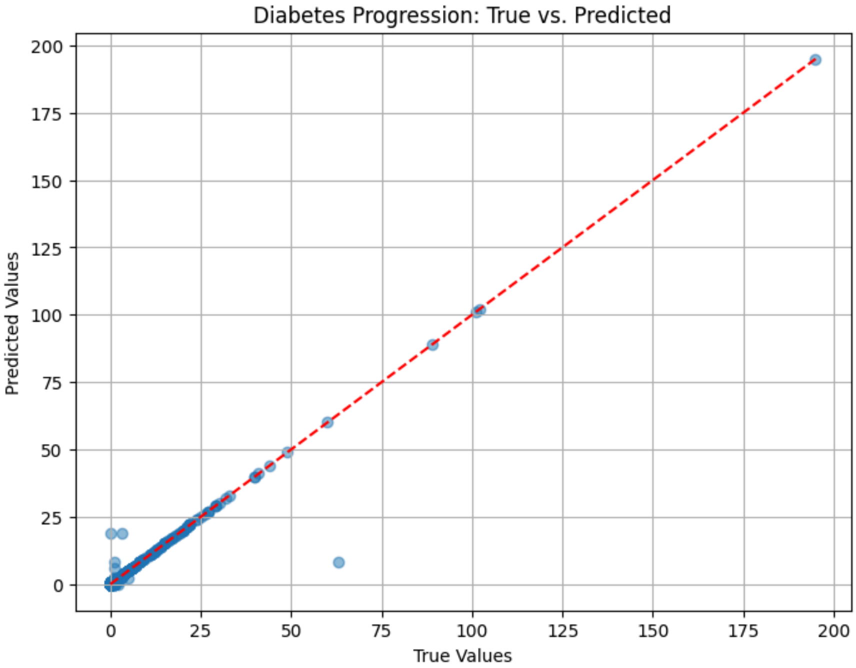

Figure 3: Prediction vs. True Value for Total Medals

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:

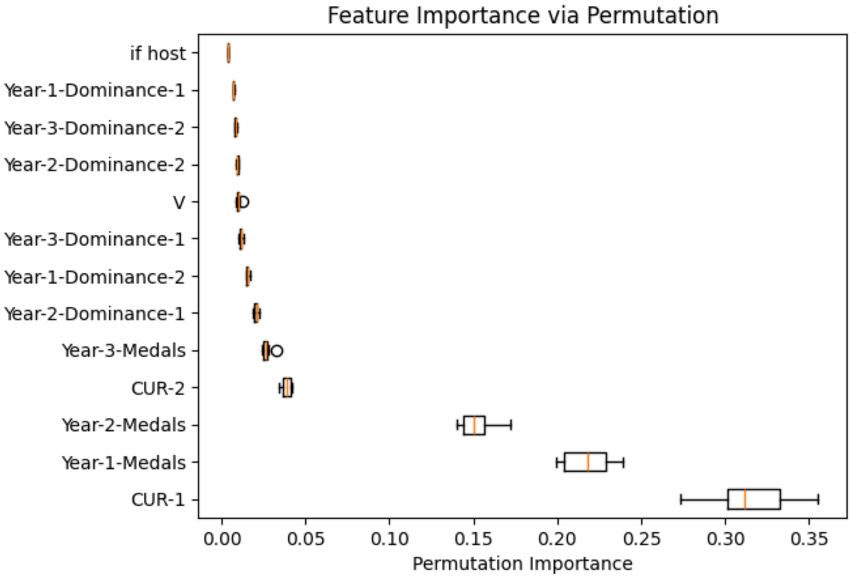

Figure 4: Permutation Importance for each Feature

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.

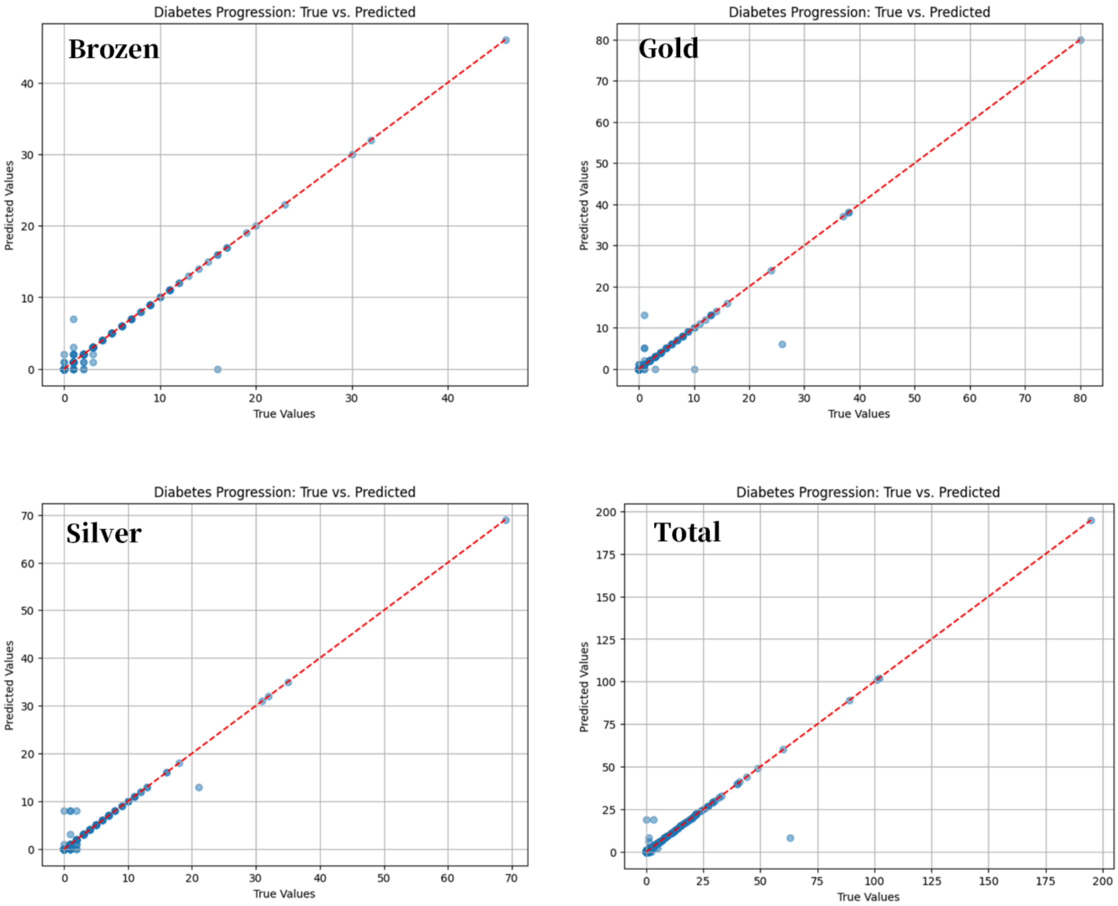

Figure 5: Intuitive Performance of each Model

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:

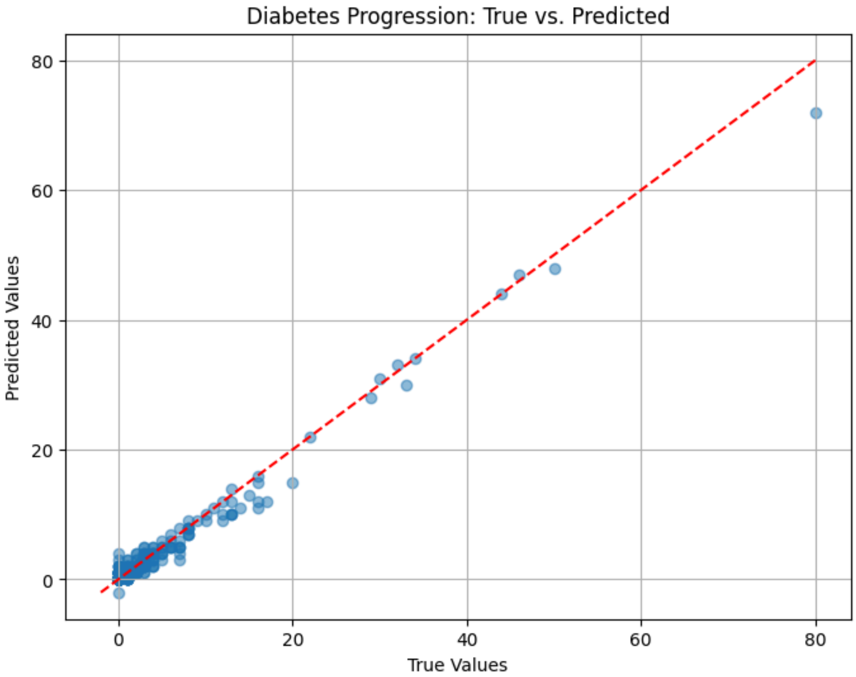

Figure 6: Prediction vs. True Value for Sport

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:

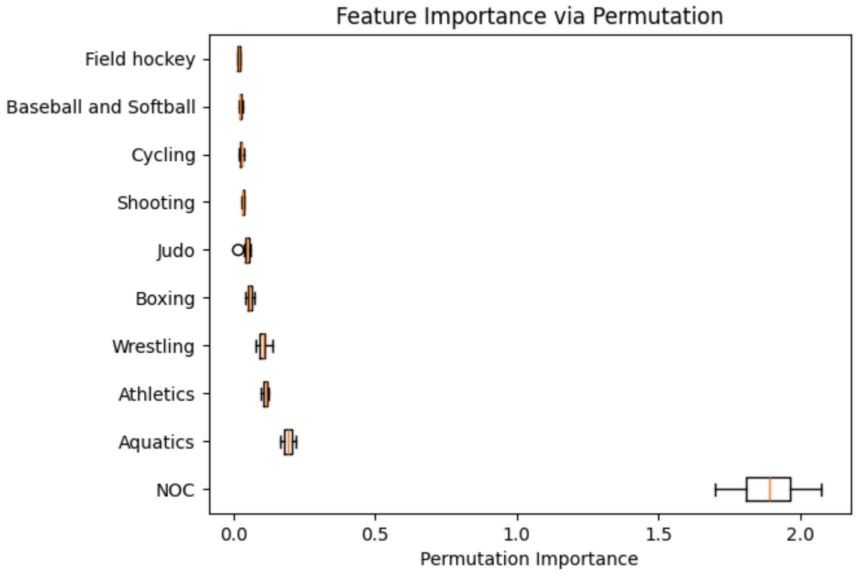

Figure 7: Permutation Importance for each Feature

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.

References