4.Task2：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 preprocess the data, select three key features, and train an XGBoost model optimized with Bayesian tuning. Finally, we apply the model to the 2028 Olympic data to generate predictions.

4.1Feature Engineering:

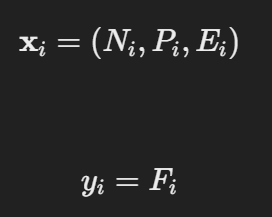
In this study, we focus on identifying patterns in countries’ progression from not winning any medals to winning their first medal. We preprocess the raw athlete data by retaining only records from countries that either have not won any medals or have won their first medal in a particular Olympic Games. 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 NiN\_i, PiP\_i, and EiE\_i as the input features. Here, NiN\_i denotes the ii-th country's participation number, PiP\_i represents the number of athletes sent, and EiE\_i is the number of events entered. The output feature, FiF\_i, indicates whether the country won a medal (1) or not (0). Therefore, the input vector for each country is represented as:

xi=(Ni,Pi,Ei)\mathbf{x}\_i = (N\_i, P\_i, E\_i)

and the corresponding output is:

yi=Fiy\_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.

4.2 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【1】. 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【2】【3】. 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 hyperparameters of XGBoost to further enhance the model's performance.【4】

4.2.1 Model Construction：XGBoost 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)

\]

Here, \(y\_i\) 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 actual labels.

The regularization term is defined as:

\[

\Omega(\theta) = \gamma T + \frac{1}{2} \lambda \sum\_{j=1}^{k} \theta\_j^2

\]

In this expression, \(T\) is the number of trees, \(\gamma\) controls the complexity of the trees, and \(\lambda\) 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)** leverages historical evaluation results to select the optimal hyperparameter combinations based on their distribution.【】 It uses **Gaussian Processes (GP)** to model the relationship between the model's error $\vartheta\_Z$ 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 $\kappa$ adjusting the balance:

αUCB=μ^(p)−κσ(p)\alpha\_{\text{UCB}} = \hat{\mu}(p) - \kappa \sigma(p)

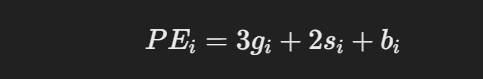
This method demonstrates strong performance in hyperparameter tuning by efficiently navigating the search space.

4.2.2Model 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.***

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



where PEi​ is the performance score for the i-th country's team, calculated based on the number of gold (gi​), silver (si​), and bronze (bi​) 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:

wi=PEi∑i=1nPEiw\_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.

**(3) Bayesian Optimization for Hyperparameters:**Bayesian Optimization is employed to fine-tune the following hyperparameters: 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).[] Maximizing AUC-PR ensures that the model is well-calibrated to predict rare events, such as medal wins, accurately.

（4）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 effectively balances predictive accuracy and the ability to correctly classify both majority and minority classes.（Table1 2）



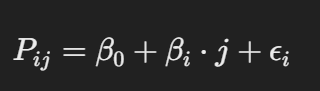


用latex画表！类似下面的（这表只是举例）

4.3Future Prediction

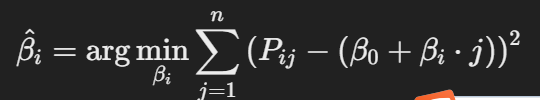
**Step 1: Fitting the Data for 2028**

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:

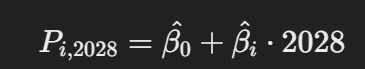


Where:Pijrepresents the number of participants from country I in the j-th Olympic Games.，β0is the intercept term.，βi is the coefficient for country i，j represents the year of the Olympic Games.，ϵi is the error term for country i

To estimate the coefficients β0\beta\_0 and βi\beta\_i, we use the least squares method to minimize the sum of squared errors:



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



同理也可以得到Ei，2028的预测值.

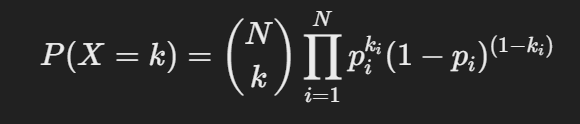
**Step 2: Predicting with BO-XGMP**

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 pip\_i of winning a medal. We assume the medal-winning probabilities for different countries are independent.

放横轴国家纵轴概率图，分析一下

The total probability distribution for kk countries winning medals is modeled using the binomial distribution. The binomial probability mass function is:

P(X=k)=(Nk)∏i=1Npiki(1−pi)(1−ki)P(X = k) = \binom{N}{k} \prod\_{i=1}^{N} p\_i^{k\_i} (1 - p\_i)^{(1 - k\_i)}



Where XX is the number of countries winning medals, NN is the total number of countries being considered, kik\_i is 1 if country ii wins a medal and 0 otherwise, and pip\_i is the predicted probability of country ii winning a medal.

放分布图和结论