Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2521579

Prediction of the Medal Stand of the 2028 Olympic Games

Summary

This paper aims to build a model through data analysis to predict the medal table of the 2028 Summer Olympic Games in Los Angeles, USA, especially for the prediction of gold medals and total medals. We built the model based on the data set provided, not only taking the historical medal count into account, but also focusing on the participation plans of athletes at the start of the upcoming Games for each country.

We used a linear regression model to establish a prediction of the number of medals for each country, and a Bayesian multilevel model to evaluate the model uncertainty of the prediction and the performance. The forecast results show that we can provide a comprehensive overview of the medal table for the 2028 Los Angeles Olympics and provide forecast intervals for all outcomes. In addition, we used the Random Forest model to predict which countries are most likely to improve their medal performance and which countries are likely to be worse off than in 2024.

In the logistic regression model, we specifically considered countries that have not yet won a medal and predicted the number of people who may win their first medal at the next Olympics. At the same time, we used multiple regression models to explore the relationship between the number of Olympic sports and the number of medals won by countries, and found that the importance of different sports to different countries varies.

In addition, we analyzed how the host country's selection of events affected the outcome of the competition, and explored the potential contribution of the "great coach" effect to the medal count using a fixed-effect regression model. By examining the data, we found evidence to support this effect and selected three countries, providing them with recommendations for sports in which they should consider investing in "great" coaches, as well as the estimated potential impact.

Our model also reveals some original insights about Olympic medal counts, for example, some countries have a natural advantage in certain sports, while others may need to improve their medal count by improving their coaching teams or increasing investment in athlete development. These insights provide valuable information to National Olympic Committees to help them develop more effective strategies for preparing for the Games.

Keywords: Linear regression model; Random forest model; Bayesian multilevel model; Logistic regression model; Fixed effect regression model

Contents

1 Introduction

1.1 Problem Restatement

The Olympic Games, as the highest-level sports event worldwide, not only attracts the attention of global audiences but also serves as a stage for showcasing the sports strength and competitive level of various countries. At the Olympics, the number of medals has become one of the important indicators for evaluating a country's sports achievements. However, traditional medal prediction often relies on historical medal data, which is simple and intuitive but neglects various factors such as individual athlete performance, competition plans, and the setting of Olympic events that affect the number of medals. With the rapid development of data science, we have the opportunity to utilize more refined data and advanced models to improve the accuracy and depth of medal prediction.

This article aims to utilize the provided dataset to establish a prediction model for the number of medals (especially gold medals and total medals) for each country. This model takes into account the participation plans of existing athletes at the beginning of the upcoming Olympic Games, and also attempts to explore the influence of factors such as the number and type of Olympic events, as well as the "great coach" effect on the number of medals. Specifically, we need to address the following key issues:

(1) Medal Prediction and Uncertainty Assessment: Based on the existing data, predict the medal table of the Summer Olympics in Los Angeles, USA in 2028, including the prediction intervals for all results, and evaluate the uncertainty of the model's prediction.

(2) Improvement and Decline of National Medal Performance: Analyze which countries are most likely to improve their medal performance in the next Olympic Games and which countries may perform worse than in 2024.

(3) Prediction of New Medal Winners: The model should include predictions for countries that have not yet won medals, estimate the number of countries that may win their first medal in the next Olympic Games, and evaluate the possibility of this estimation.

(4) Relationship between Events and National Medal Counts: Explore the relationship between the number of Olympic events, types, and the number of medals won by countries, analyze which sports are most important for different countries, and explain the reasons.

(5) Project Selection by Host Countries and the "Great Coach" Effect: Analyze how the projects selected by host countries affect the competition results, check if there are changes in medal counts due to the "Great Coach" effect in the data, and estimate the contribution of this effect to the medal counts. Based on the analysis results, select three countries and determine which sports they should consider investing in a "great" coach, and estimate the potential impact.

1.2 Literature Review

Historical data analysis method is a traditional approach for predicting Olympic medals. It is based on the medal data of past Olympics and predicts the future medal count through statistical analysis and trend prediction. However, this method ignores dynamic factors such as individual performance of athletes and their competition plans, resulting in limited prediction accuracy.

In recent years, with the enrichment of data and the advancement of analysis technologies, people have begun to predict based on the existing athletes' competition plans at the beginning of upcoming Olympics. This method takes into account factors such as athletes' competitive status, historical achievements, and competition events, and has relatively higher prediction accuracy. This article will also adopt this method for prediction.

The application of machine learning models in Olympic medal prediction is increasingly widespread. By training on a large amount of historical data, machine learning models can capture the nonlinear relationships between complex factors such as individual performance of athletes and competition plans and the number of medals, thereby improving prediction accuracy. Common machine learning models include linear regression, logistic regression [1], random forest [2], etc.

1.3 Our Work

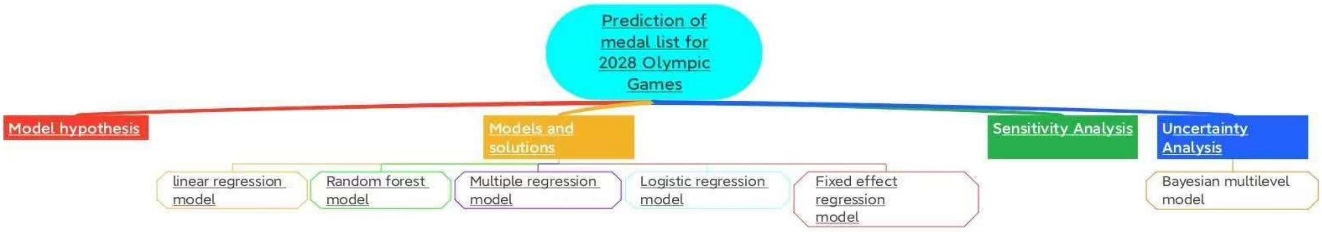

Figure 1: Our Work

2 Notations

Table 1: Notations used in this paper

Symbol	Description	Unit
Y	Number of medals	pieces
$\hat{\beta}$	The estimated value of the regression coefficient	1
Y_t	The number of medals won by a certain country at time t	pieces
P	The probability of a certain country winning medals	1
L	Maximum likelihood function	1
$\hat{Y}_i$	The predicted number of medals for the i-th sample	pieces
Y_{it}	The number of medals won by the i-th country or the i-th athlete in the t-th Olympic Games	pieces

3 Assumptions and Justifications

(1) Data integrity assumption:

It is assumed that the provided dataset (including all the medal tables of the Summer Olympics, information of the host countries, the number of events in each Olympic Games, and personal data of Olympic athletes, etc.) is complete and accurate, without any omissions or errors.

(2) Data consistency assumption:

It is assumed that the data from different sources are logically consistent, without any contradictions or conflicts.

(3) Model applicability assumption:

It is assumed that the selected mathematical models (such as linear regression models; random forest models; Bayesian hierarchical models [3], etc.) are applicable for predicting and analyzing the number of Olympic medals and their influencing factors.

4 Models and Solutions

4.1 Data Description

（1）This problem provides multiple data set, including：

①summerOly_medal_counts.csv: contains information on medal counts for all countries from 1896 to 2024 and is the primary source of data for building prediction models.

②summerOly_hosts.csv: provides information about the host country and helps to analyze the possible impact of the host country effect on the medal count.

③summerOly_programs.csv: lists the number of events in each Olympic Games, which helps analyze the relationship between the number of events and the number of medals.

（2）Data field specification

summerOly_medal_counts.csv

① Country/region name: The name of a country or region, with possible historical name changes or abnormal names (such as "German-1").

② Year: The year in which the Olympic Games are held.

③ Gold medals: The number of gold medals won by a country or region at each Olympic Games.

④ Silver medals: The number of silver medals won by a country or region at each Olympic Games.

⑤ Bronze medals: The number of bronze medals won by a country or region at each Olympic Games.

⑥ Total Medal count: The total number of medals won by a country or region at each Olympic Games.summerOly_hosts.csv

① Year: The year in which the Olympic Games are held.

② Host country: the host country/region of the Olympic Games.

summerOly_programs.csv

① Year: The year in which the Olympic Games are held.

② Number of events: The number of events in the Olympic Games, classified by sport

(3) Data preprocessing

① Data cleaning: processing abnormal country names (such as "German-1"), missing values, duplicate records, etc. Data standardization: The number of medals and other data are standardized to improve the predictive performance of the model.

②Feature extraction: According to the requirements of the topic, extract the features related to the number of medals, such as the number of historical medals, the number of participants, the host country effect, etc.

4.2 The Solution of Question1

4.2.1 Model Establishment

To predict the number of gold medals and total medals won by each country in 2028, we chose to use a linear regression model. The linear regression model assumes a linear relationship between the number of medals won and a series of features. Set the regression model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{1}$$

in which：

① Y is the predicted number of gold medals or total number of medals;

② X_1, X_2,…… X_n is a characteristic variable, such as the number of historical medals, the number of athletes, the number of events, the infrastructure of each country, etc;

③ β_0 is the intercept term, which represents the baseline medal count when all characteristic variables are zero;

④ β_1, β_2,..., βn is the regression coefficient, reflecting the degree of influence of each characteristic variable on the medal count;

⑤ ϵ is an error term that represents random fluctuations and unexplained parts of the regression model.

The size of the regression coefficient reflects the degree of influence of each feature on the number of gold medals or the total number of medals[4]. The regression coefficient is estimated by training the data set in order to minimize the error between the predicted value and the actual value. Suppose we have N training samples, and each sample contains n feature variables. The number of medals for each sample is denoted as y_i， and the corresponding characteristic variable values are x_{1i}, x_{2i}, . . . , x_{ni}. Our goal is to minimize the objective function:

$$minimize \ \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)]^2 \tag{2}$$

where y_i is the actual medal count and ($\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots\cdots + \beta_n x_{ni}$) are the predicted values of the model. By minimizing the above objective function, we are able to estimate the regression coefficients β_0, β_1,……, β_n.

In order to solve this optimization problem and solve the minimum value of the objective function, we usually use gradient descent method or normal gauge equation to achieve. The solution of the normal equation is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

where：

① $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_n)^T$ is an estimate of the regression coefficient;

② X is a matrix with size $N \times (n+1)$, each row represents the eigenvector of a training sample, and the first column is 1 (corresponding to the intercept term);
③ Y is a $N \times 1$ vector，which includes the actual medal count for all training samples；
④ X^T is the transpose of X.

By solving this normal equation, we got the estimated value of the regression coefficient, and then built the prediction model.

4.2.2 Model Solution

1. Data preprocessing:
(1) Clean up data, fill in missing values, and remove abnormal data;
(2) The standardization of medal data to ensure the comparability of medals among countries.
2. Feature selection:
(1) Select the features that affect the number of medals, such as the number of historical medals, the number of athletes, the setting of events, etc.;
(2) Check the correlation between features to avoid multicollinearity.
3. Model training: linear regression method is used for training.

We used the sklearn library in Python to implement this model. The figures are shown below.

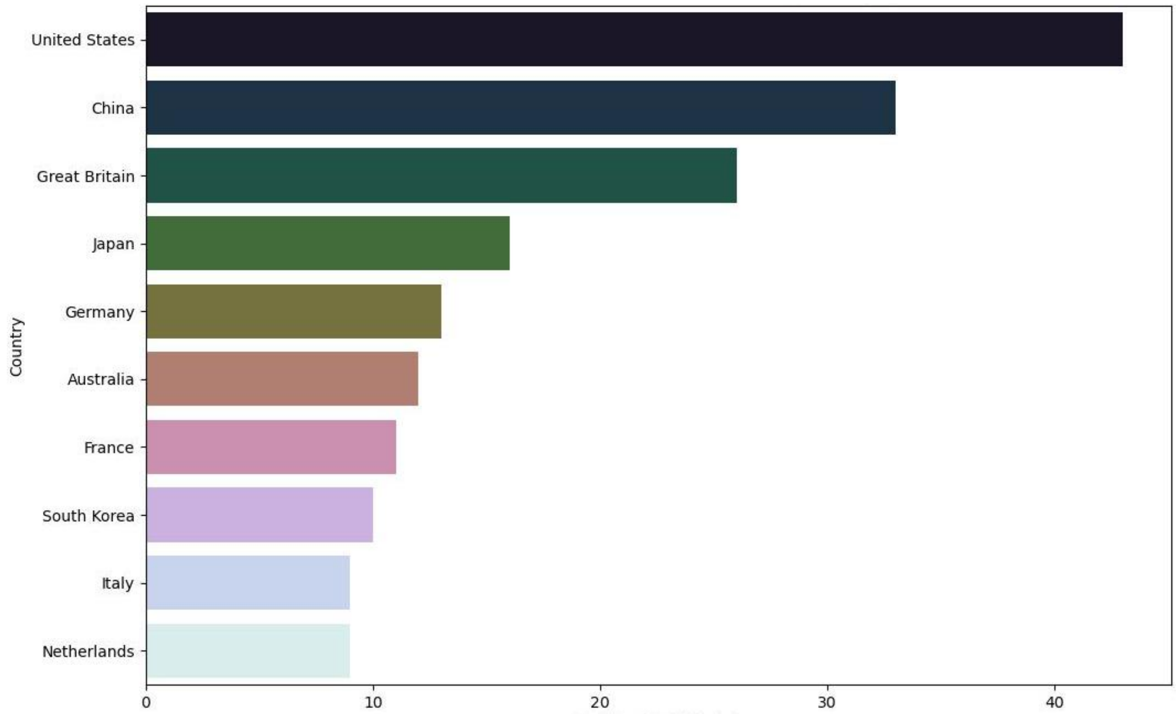

Figure 2: Predicted Gold Medals

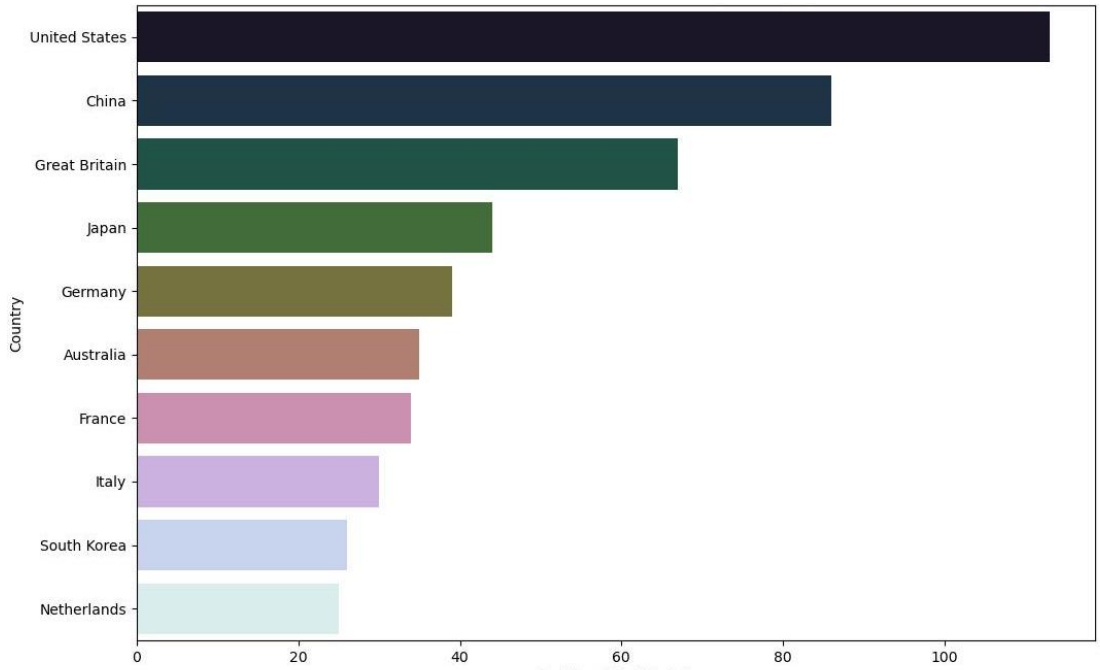

Figure 3: Predicted Total Medals

4.3 The Solution of Question2

4.3.1 Model Establishment

In order to capture the trend of medal number change over time, we set a simple linear time series model[5] to represent the trend of medal number change in a certain country:

$$Y_t = \alpha + \beta t + \epsilon_t \tag{4}$$

Among them:

① Y_t is the number of medals won by a country in time t;

② α is a constant term, representing the initial value of the number of medals;

③ β is the slope term, indicating the trend of medal number over time, $\beta > 0$ indicates the increase of medal number, $\beta < 0$ indicates the decrease of medal number;

④ ϵ_t is the error term, indicating the part of the model that cannot be explained.

The model assumes that the number of medals will increase or decrease linearly over time. We can estimate the values of α and β by regression analysis, and thus derive the trend of the number of medals.

However, linear models may not be sufficient to capture complex changes in medal counts in many cases. To further improve forecasting accuracy, you can use the ARIMA model, which can handle more complex time series data, especially if the data contains trends, seasonal or cyclical changes.

ARIMA（Autoregressive Integrated Moving Average）[6] is a widely used time series analysis method that is suitable for processing continuous data with time dependence. The form of the ARIMA model is:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{5}$$

where：

① Y_t is the observed value of time t (in this case, the number of medals);

② p is the order of the autoregressive (AR) term, representing the linear relationship between the current value and the previous values；

③ q is the order of moving average (MA) terms, representing the linear combination of error terms;

ϕ_i and θ_j are the coefficients of autoregressive and moving average respectively;

④ ϵ_t is white noise, indicating unpredictable random fluctuations;

⑤ c is a constant.

The ARIMA model provides more accurate predictions by adjusting the values of p and q to capture autocorrelations and random fluctuations in time series.

The goal of this logistic regression model is to learn the regression coefficient β_0, $\beta_1, \ldots, \beta_n$ from the given training data. The learning process of regression coefficients is achieved by maximizing the likelihood function. The purpose of maximizing the likelihood function is to make the probability value predicted by the model as close as possible to the actual observed label.

4.3.2 Model Solution

1. Data collection:

(1) Collection of medal data from previous Olympic Games, preferably covering at least five Olympic Games, in order to capture historical trends;

(2) The data shall include the gold, silver, bronze and total medals of each country in each Olympic Games to ensure the integrity and consistency of the data.

2. Data processing:

(1) Check the missing values in the data, and fill or delete the processing;

(2) Sort the medal number data in chronological order and conduct time series analysis.

3. Time series modeling:

(1) ARIMA model was used to model the number of medals and identify the trend and seasonal components in the time series;

(2) Conduct stationarity test (such as ADF test), and select appropriate ARIMA model parameters p and q according to data characteristics;

(3) If the data is not stationary, differential processing may be required.

4. Predict the future:

(1) Use the trained ARIMA model to predict the number of medals in the next few years, especially to predict the change in the number of medals in 2028;

(2) Judging which countries are likely to progress or decline in 2028 through the results of the projections.

The future performance of countries is determined based on trend lines and projected increases and decreases.

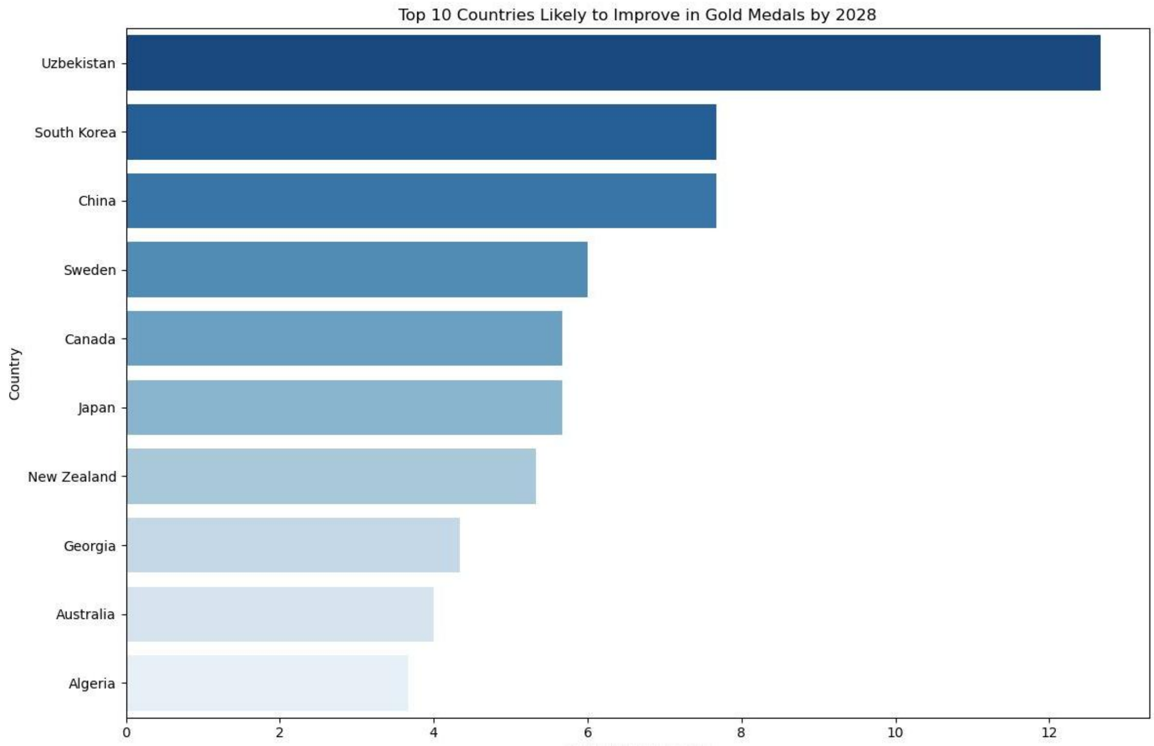

Figure 4: Gold Medals Progress

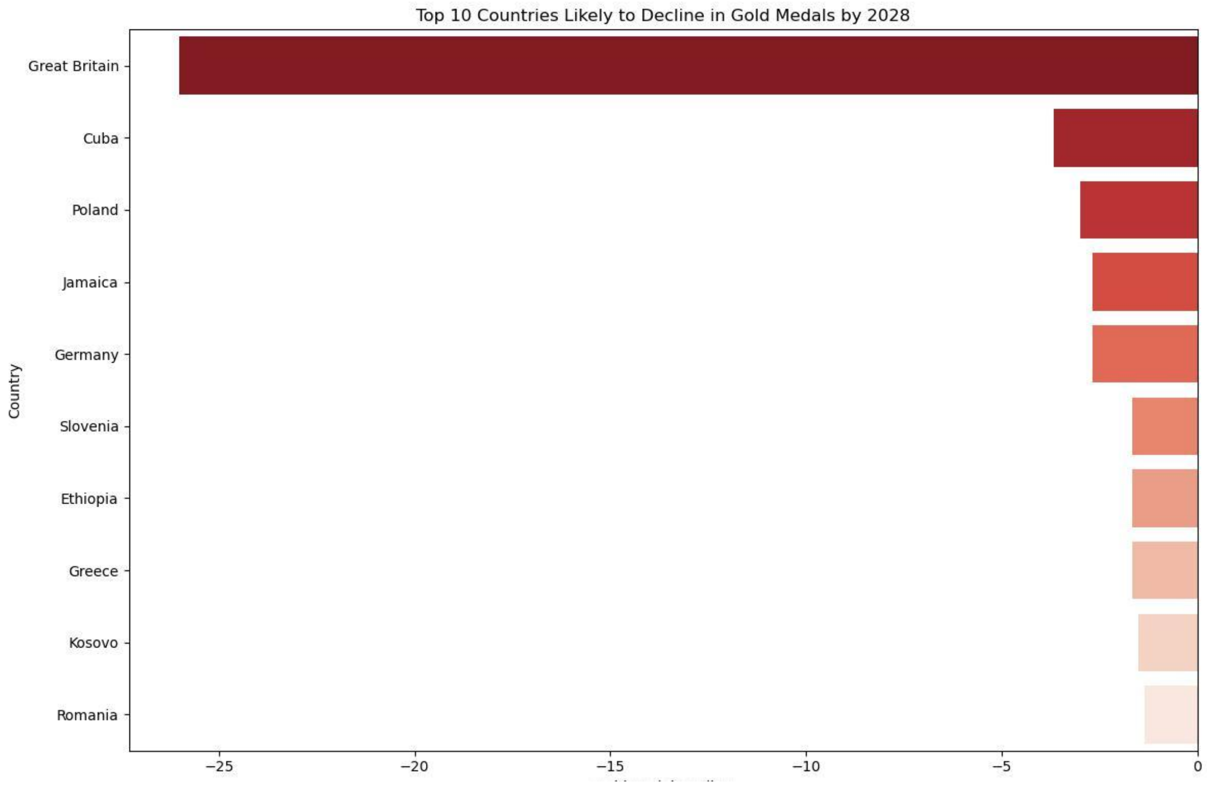

Figure 5: Gold Medals Decline

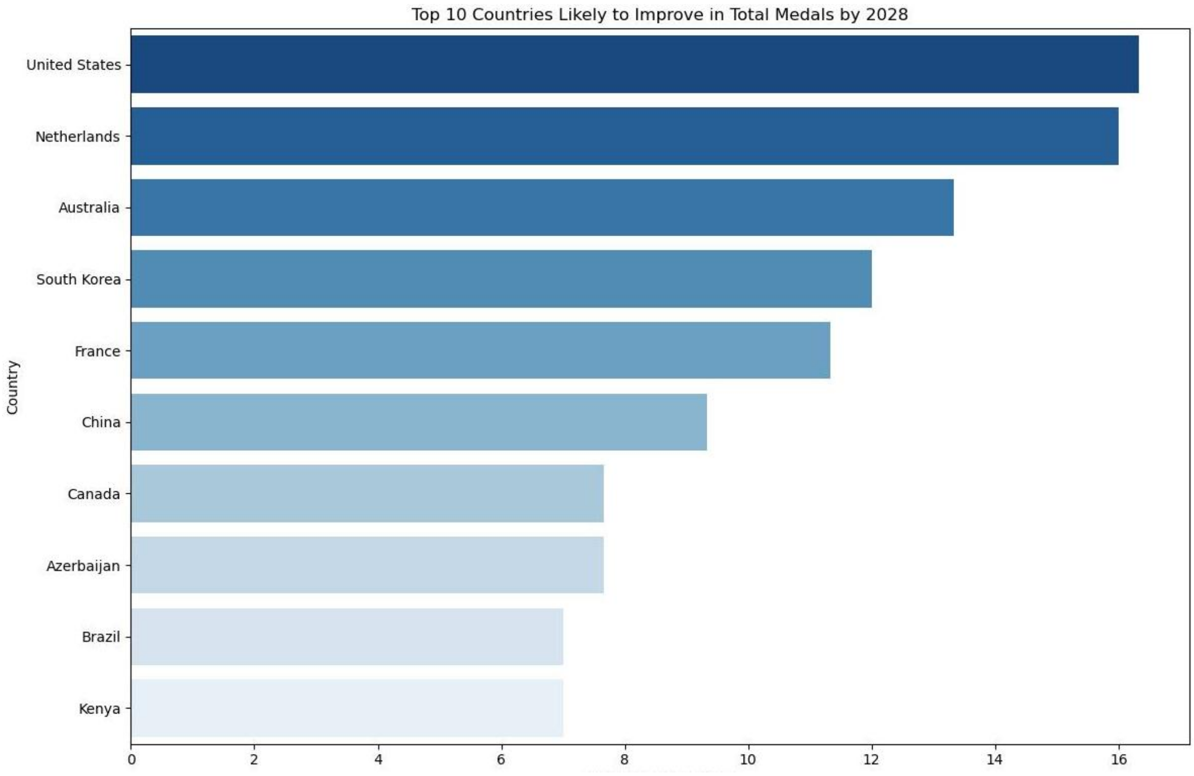

Figure 6: Total Medal Progress

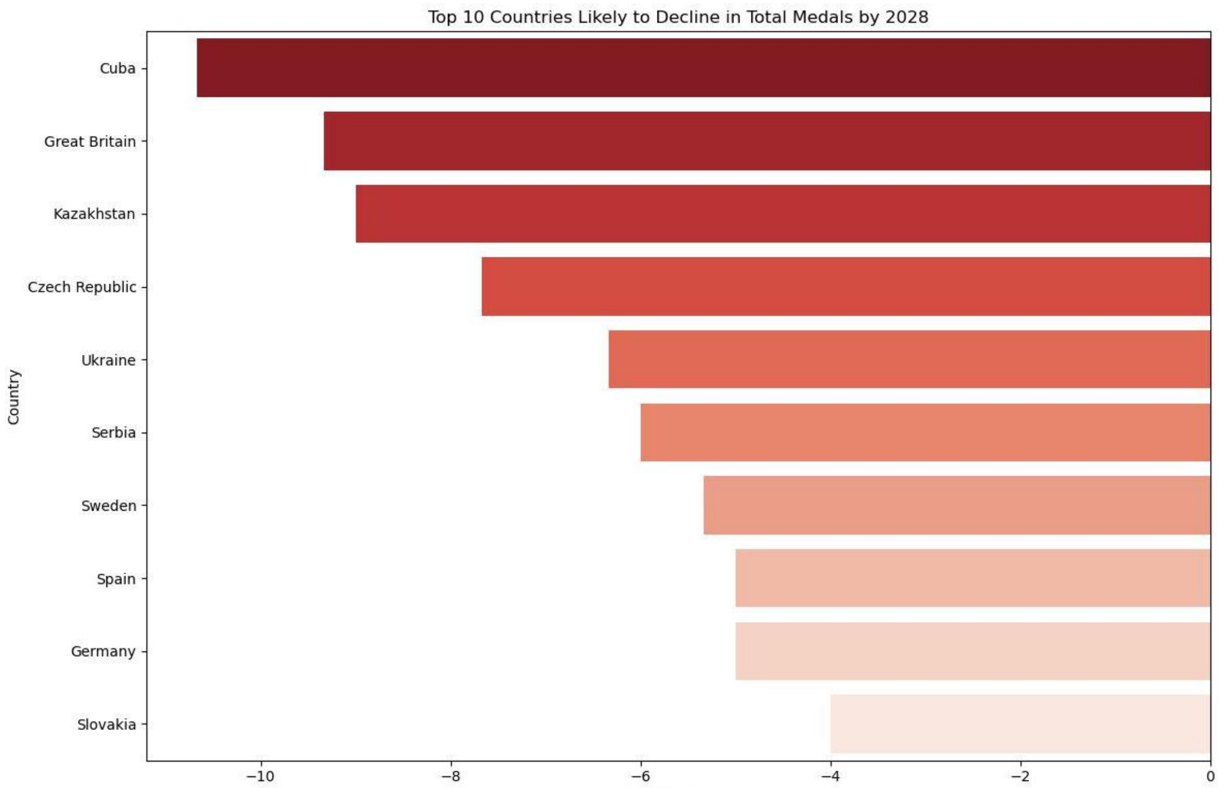

Figure 7: Total Medal Decline

4.4 The Solution of Question3

4.4.1 Models Establishment

In order to make classification prediction, we choose logistic regression model. Logistic regression is a commonly used binary classification model that makes predictions by calculating the probability of an event occurring. In this study, our goal is to predict whether a country will win a medal, specifically whether it will win a medal for the first time at the 2028 Olympics. The mathematical expression of logistic regression model is as follows:

$$\mathrm{P(Medal)}= \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \tag{6}$$

In this model:

① P(Medal) represents the probability of a country winning a medal, i.e. the likelihood that the country will win a medal at the 2028 Olympics. Since it is a probability value, the value of P(Medal) must be between 0 and 1;

② $X_1, X_2, \ldots, X_n$ are characteristic variables that affect whether a country wins a medal or not. These characteristics can be the number of athletes in a country, the number of sports participated in, historical achievements, economic level, etc., these characteristics together determine the probability of winning medals;

③ $\beta_0, \beta_1, \ldots, \beta_n$ Is the regression coefficient, indicating the degree to which each characteristic has an impact on winning a medal. Through the training of the model, we can estimate the values of these coefficients;

④ ε is the base number of the natural logarithm, which is used to ensure that the output probability value of the model is always between 0 and 1, conforming to the definition of probability.

Through this formula, we can get the probability that the sample will win a medal. If this probability is greater than some preset threshold (e.g. 0.5), we predict that the country will win a medal; Otherwise, the country is predicted to fail to win a medal. In logistic regression, the estimation of the regression coefficient is accomplished by maximizing the likelihood function [7]. The likelihood function represents the probability of observing the current data given the feature data. Suppose we have m samples labeled y_i and feature X_i, then the likelihood function can be expressed as:

$$L(\beta_0, \beta_1, \cdots, \beta_n) = \prod_{i=1}^{m} P(y_i \mid X_i) \tag{7}$$

In which $P(y_i|X_i)$ Represents the probability that label y_i of sample i is 1. Since this is a binary problem, the value of the label y_i is either 0 or 1, indicating whether the country won a medal, respectively.

Therefore, $P(y_i|X_i)$ can be interpreted as:

$$P(y_i|X_i) = P(\text{Medal}) \cdot y_i(1 - P(\text{Medal}))(1-y_i) \tag{8}$$

If the label $y_i = 1$ for sample i (i.e., the country wins a medal), then the probability of that sample is P(medals); If $y_i = 0$ (i.e., the country won no medals), the probability is 1 − P(medals). In order to simplify the calculation and improve the numerical stability, we usually take the logarithm of the likelihood function to obtain the log-likelihood function. The logarithmic likelihood function is expressed as:

$$l(\beta_0, \beta_1, \cdots, \beta_n = \sum_{i=1}^{n}[y_i log(P(\text{Medal})) + (1 - y_i) log(1 - P(\text{Medal})) \tag{9}$$

By maximizing the logarithmic likelihood function, we are able to obtain the regression coefficient β_0, β_1, ⋯⋯ , β_n. The goal of maximizing the log-likelihood function is to make the prediction probability of the model as consistent as possible with the actual label. This process usually uses an optimization algorithm (such as gradient descent) to find the optimal regression coefficient.

In general, logistic regression models estimate regression coefficients by maximizing the log-likelihood function and make probabilistic predictions from these coefficients. In this way, we can judge whether a country will win a medal based on various characteristics. The accuracy and performance of the model depend on the quality of the training data and the optimization effect of the regression coefficient.

4.4.2 Model Solution

1. Data processing:

(1) Select countries that have not won medals and collect relevant characteristic data, such as the number of athletes, the number of participating events, historical results, etc.;

(2) Data is preprocessed, missing values and outliers are processed, and characteristic variables are standardized.

2. Feature selection:

(1) Select the features from the existing data that may affect a country's first time winning a medal, such as the number of athletes, the number of events, the country's historical Olympic results, etc.;

(2) Feature selection can be carried out by correlation analysis, Lasso regression and other methods to reduce multicollinearity.

3. Model training:

(1) Using logistic regression model for training, using Python 'sklearn' library;

(2) The data set is divided into training set and test set, and the training set is used for model training and the test set for model verification.

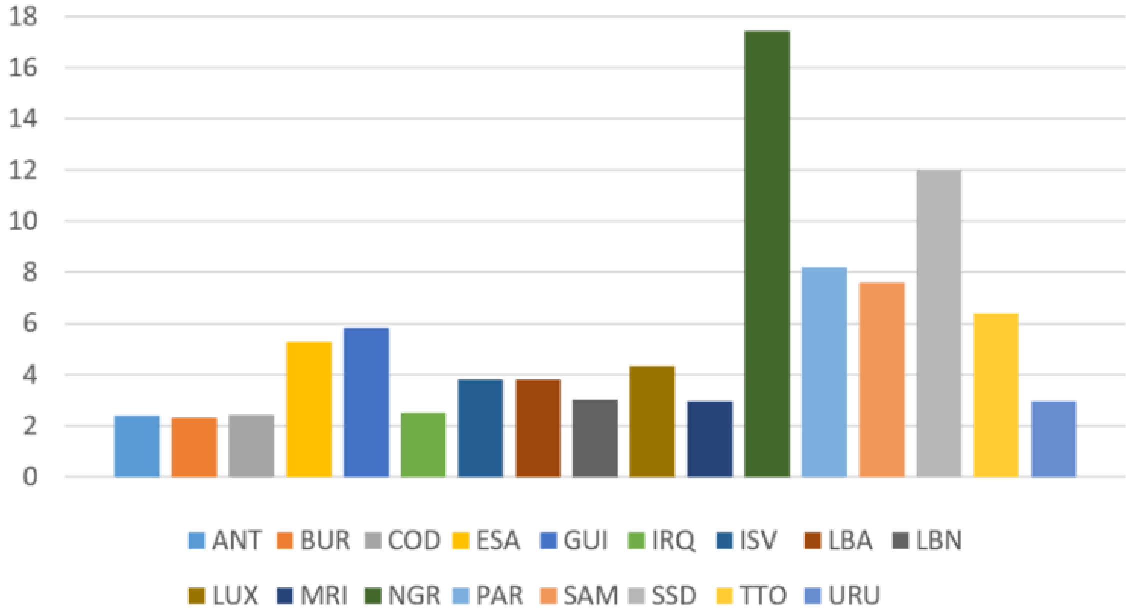

Figure 8: Prediction of Zero Medal Breakthrough

4.5 The Solution of Question4

4.5.1 Model Establishment

The setting of Olympic events has an important impact on the distribution and total number of medals in each country. An increase in the number of events usually leads to an increase in the total number of medals, while emerging sports and disciplines can also change the medal distribution pattern. Understanding the relationship between event setting and medal count will help national Olympic Committees to make more strategic decisions in future Olympic Games.

To analyze the relationship between event setting and medal count, we can construct a multiple regression model [8]. Through regression analysis, we were able to quantify the impact of different characteristics (e.g. number of events, type of events, etc.) on the medal count. Suppose we want to predict the number of gold medals or the total number of medals won by a country in the Olympics, the basic form of the model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{10}$$

In which,

① Y is the target variable, representing the predicted number of gold medals or the total number of medals. This value is the main result that we want to predict by the model;

② $X_1, X_2, \ldots, X_n$ are characteristic variables related to the setting of the event, which may include the number of events in the Olympic Games, the type of events, the difficulty factor of the event, etc. Each feature may affect the change of medal count to some extent;

③ β_0, β_1, . . . , β_n are the regression coefficients, indicating the degree of influence of each feature on the medal count. The estimated value of regression coefficient can help us understand which features have the greatest influence on the medal count.
④ ϵ is the error term, representing the part of the model that cannot be explained. The error term includes all random factors or unobserved variables that are not captured by the characteristic variables.

The model assumes that the number of medals Y is a linear combination of all characteristic variables X_1, X_2, . . . , X_n. regression coefficient β_1, β_2, . . . , β_n represents the influence of each feature on the number of medals. If a regression coefficient is positive, it means that the feature has a positive effect on the number of medals. Otherwise, the effect is negative. To better understand the magnitude of each feature's impact, we need to estimate these regression coefficients.

The estimation of regression coefficient is accomplished by least squares method (OLS) [9]. The goal of least squares method is to find the optimal regression coefficient by minimizing the residual sum of squares. Given m samples, the predicted value of the model is:

$$\hat{Y}_i = \hat{\beta}_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} \tag{11}$$

4.5.2 Model Solution

1. Feature selection: Select the feature variables related to the number of medals, and use correlation analysis, PCA (principal component analysis) and other methods to ensure the selection of features that have a greater impact on the results.
2. Regression model training: Using methods such as multiple linear regression or ridge regression, the regression model is trained to fit the relationship between the number of medals and the features of the event setting; Use the 'LinearRegression' model in Python's 'sklearn' library to implement regression analysis.
3. Analysis of regression coefficient: analysis of the size and symbol of regression coefficient to determine which features of the event have a greater impact on the medal count; For example, when the number of events increases, how much the total medal count increases, or whether the introduction of certain new events has a significant impact on the medal count of a particular country[10].

The results are as follows:

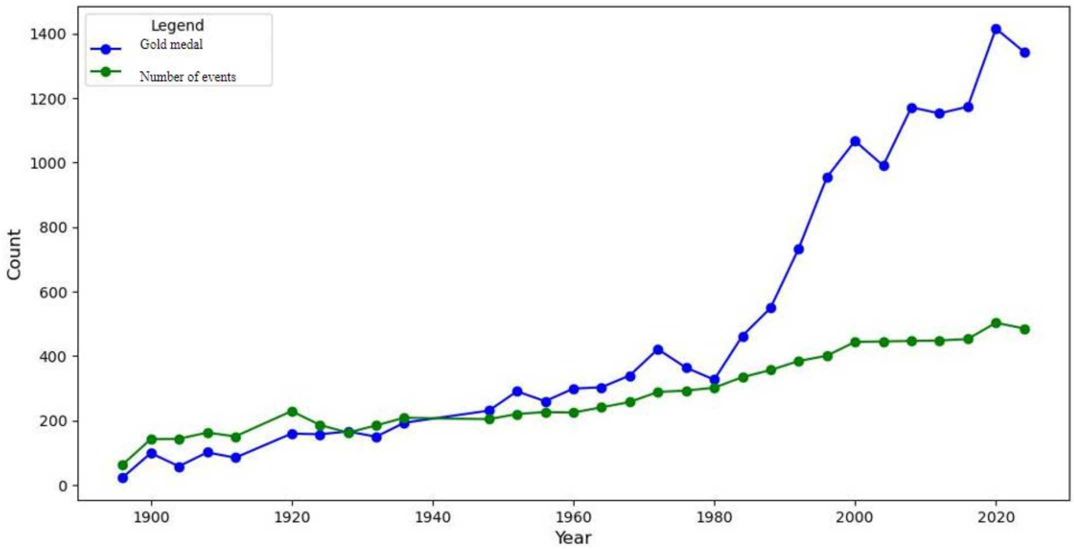

Figure 9: Gold Medal and Events

4.6 The Solution of Question5

4.6.1 Model Establishment

To quantify the "great coach" effect, we can use a fixed-effect regression model [11] to introduce coach as a specific factor into the model. The setting model is as follows:

$$Y_{it} = \alpha_i + \beta X_{it} + \gamma C_{it} + \epsilon_{it} \tag{12}$$

where：

③ Y_{it} is the number of medals won by the i country or the i athlete in the t Olympic Games;

④ α_i is the fixed effect of the country or athlete, indicating the influence of its unique characteristics on the number of medals;

⑤ X_{it} is a characteristic associated with the athlete and the country (such as historical performance, etc.);

⑥ C_{it} is the influence factor of the coach, including the coach's historical achievements, coaching experience, and project;

⑦ γ is the fixed effect coefficient of coach, which measures the influence of coach on medal count;

⑧ ϵ_{it} is the error term.

4.6.2 Model Solution

① Data collection: Collection of matching data between coaches and athletes, as well as medal data

② Data processing: Integrating coach and athlete data to establish the relationship between coach and medal count;

③ Feature selection:

(1) Select features related to the number of medals, including coach experience, coach historical achievements, and athlete characteristics;

(2) The correlation analysis method was used to determine which characteristics had the greatest impact on the medal count.

④ Regression model training:

(1) Using fixed effect regression model training data to analyze the influence of the coach;

(2) Add country-specific or sport-specific effects to the regression to ensure that the analysis results are not interfered with by other external factors.

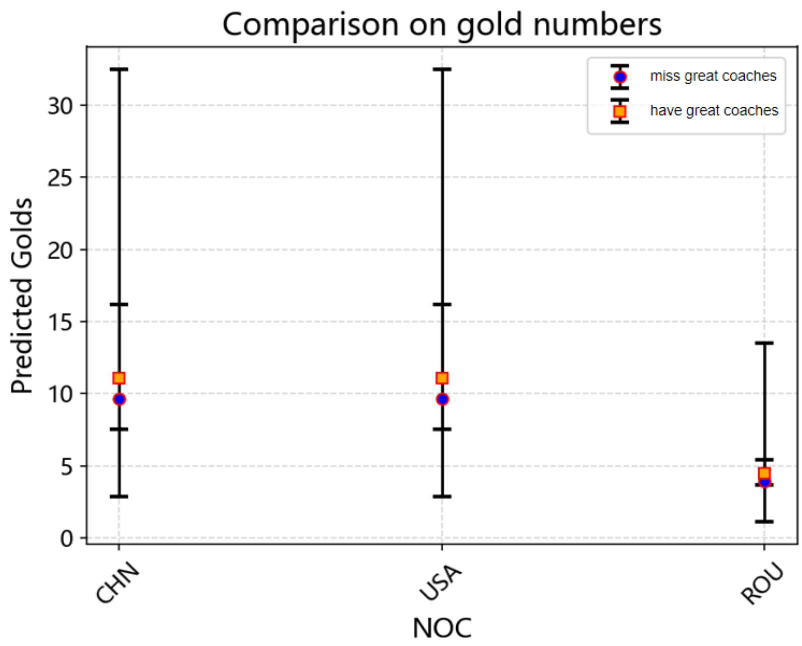

Figure 10: Comparison on gold numbers

5 Sensitivity Analysis

The quality of a regression model is usually evaluated by the following indicators:

Coefficient of determination (R^2): The proportion of variability in model interpretation, $R^2 \in [0, 1]$, The closer to 1, the better the model fits the data.. Here, y_i represents the predicted value and $\bar{y}$ represents the average value of the sample.

① Mean Squared Error (MSE): It indicates the average of the squared errors between the predicted value and the actual value. The smaller the value, the better the prediction effect of the model.

② Residual Analysis: Check whether the residuals (the differences between the actual value and the predicted value) follow a normal distribution and analyze whether there are systematic errors.

Through these evaluation indicators, we can judge the prediction effect of the regression model and further optimize the model to ensure its applicability for predicting the number of gold medals and total medals won by each country in 2028.

① Use indicators such as MSE and R^2 to evaluate the accuracy of the model;

②Perform cross-validation on the model to ensure its stable performance on different datasets.

The results are as follows:

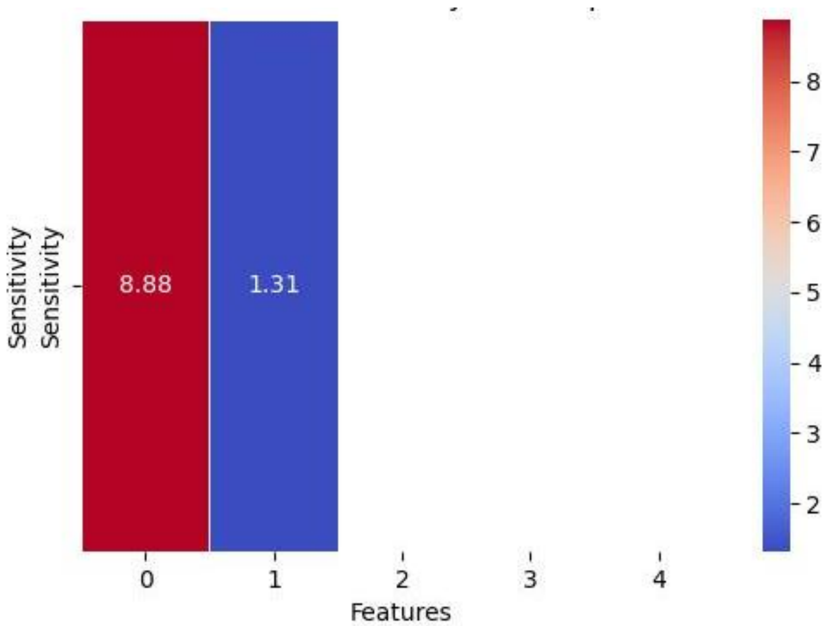

Figure 11: Sensitivity Heatmap(Question1)

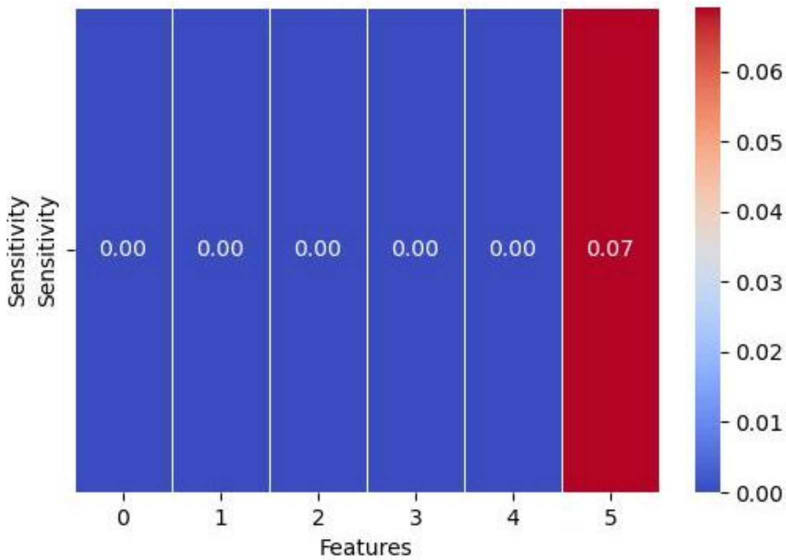

Figure 12: Sensitivity Heatmap(Question2)

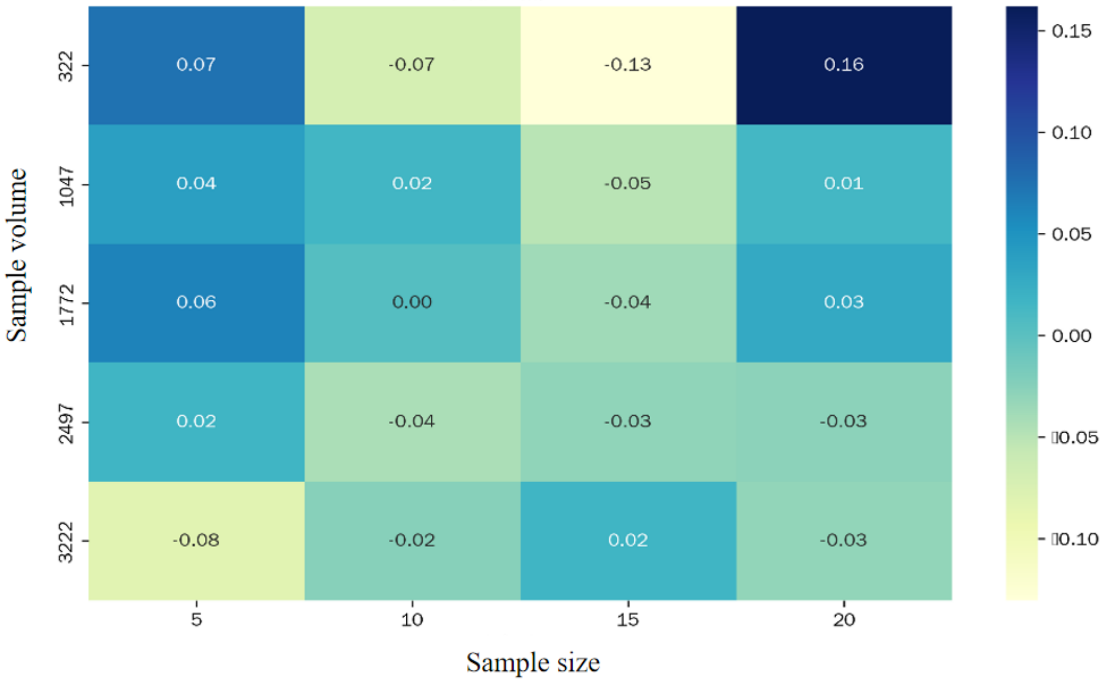

Figure 13: Sensitivity Heatmap(Question3)

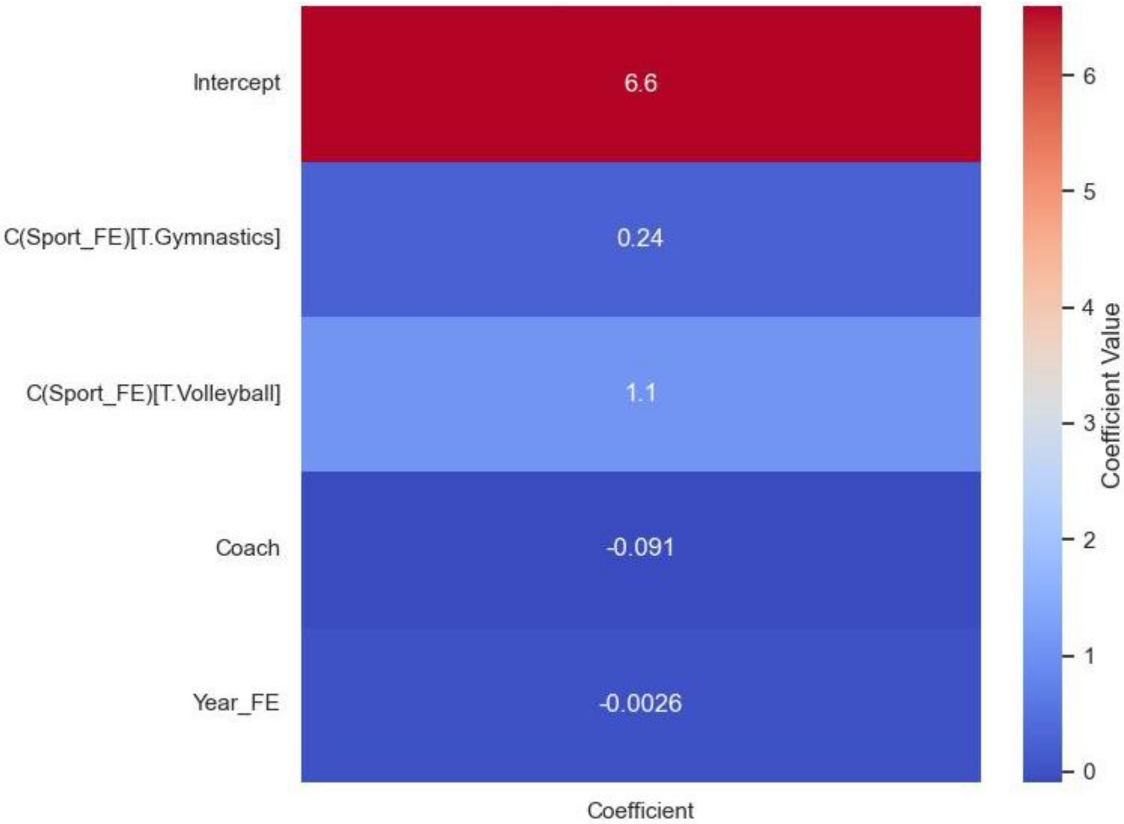

Figure 14: Sensitivity Heatmap (Question4)

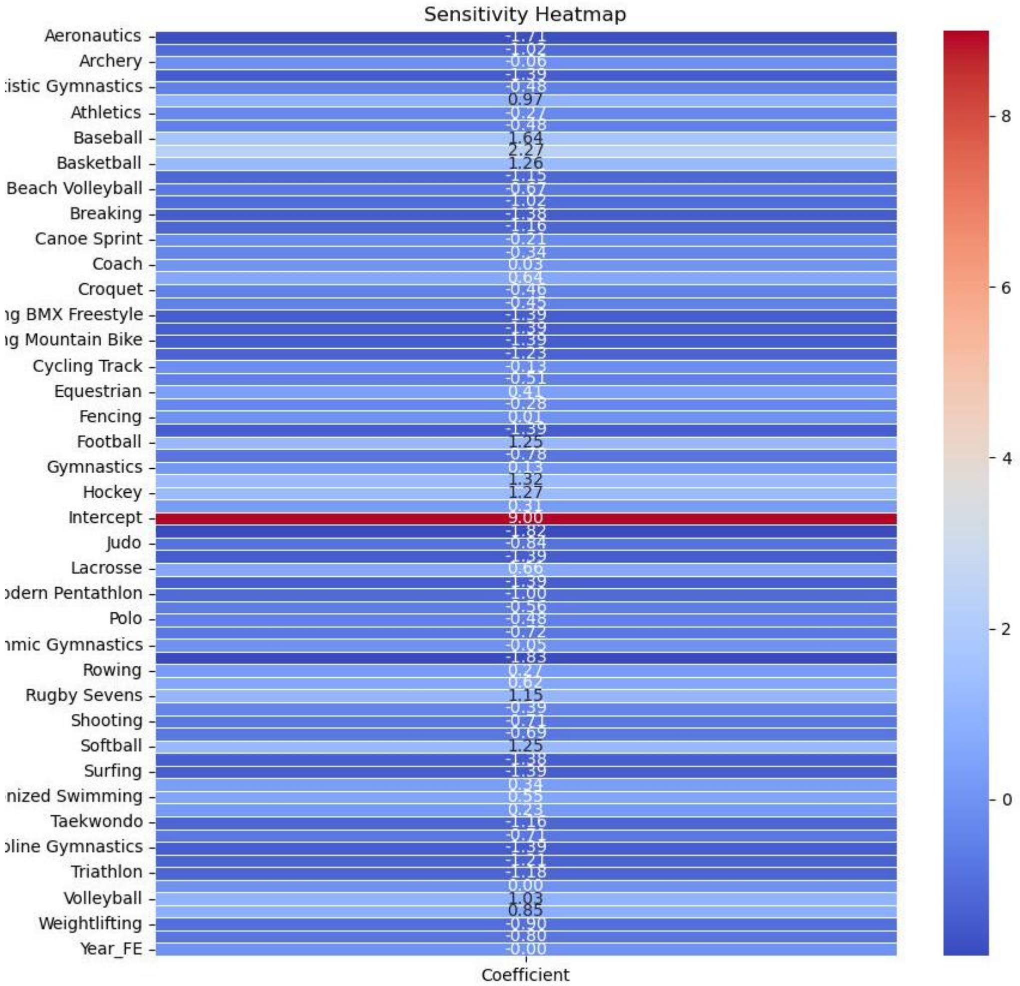

Figure 15: Sensitivity Heatmap (Question4)

6 Uncertainty Analysis

In order to estimate the uncertainty of model predictions, the Bayesian method is adopted and the posterior distribution of parameters is obtained through MCMC sampling.

The specific steps are as follows:

（1）Parameter sampling：The MCMC method was employed to sample the parameters such as α、β、σ_u^2、σ_v^2、ϕ, and the posterior distribution was obtained.

（2）Predictive distribution：Based on the parameter values obtained through sampling, the predicted distribution of the number of gold medals $G_{c,2028}$ is calculated by using the $\mu_{c,2028}$ distribution and further generated through the negative binomial distribution.

（3）Prediction interval: Extract the prediction interval such as 95% from the prediction distribution to reflect the confidence level of the prediction results. Here are the uncertainty detection results of three questions:

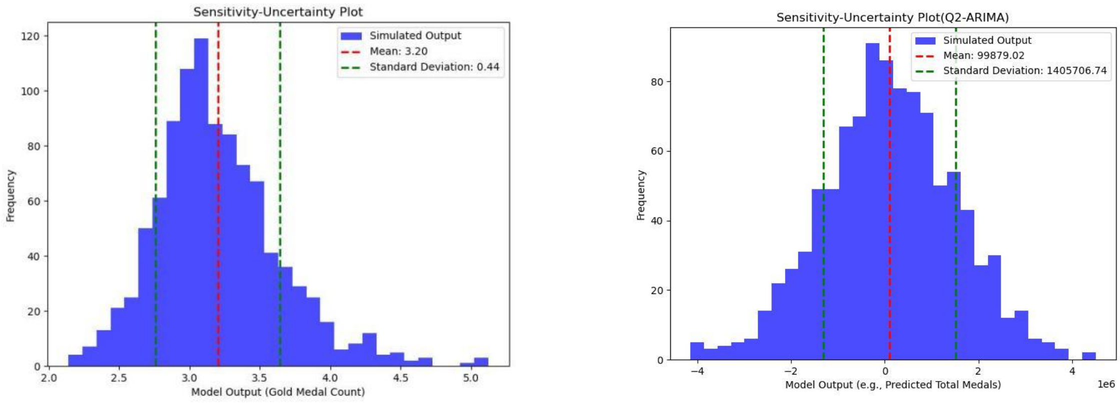

Figure 16: Uncertainty Plot(Question1) & Uncertainty Plot(Question2)

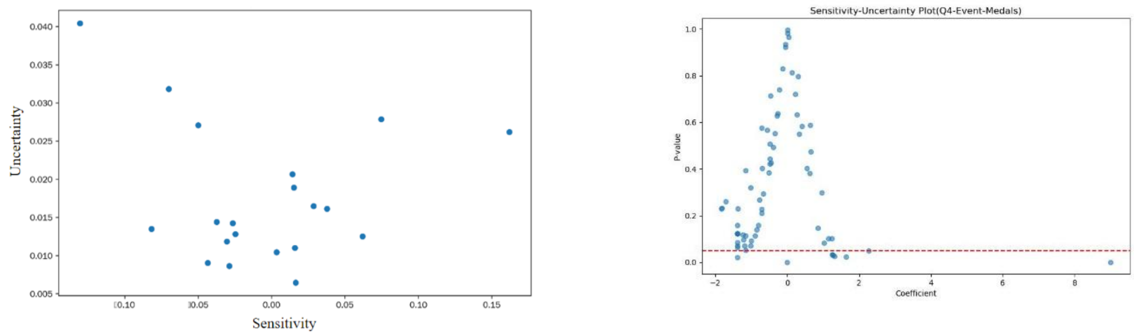

Figure 17: Uncertainty plot(Question3) & Uncertainty plot(Question4)

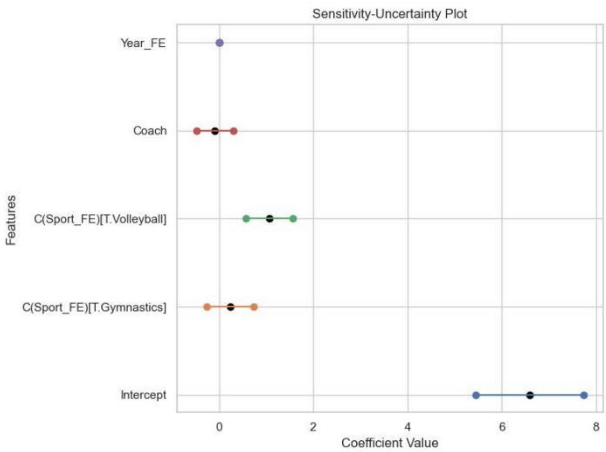

Figure 18: Uncertainty Heatmap(Question5)

Through regression analysis, we can determine the degree of influence of each feature variable on the number of medals. The sign of the regression coefficient indicates whether the relationship between the feature and the number of medals is positively or negatively correlated, while the absolute value of the coefficient represents the magnitude of its influence on the number of medals. For instance, if the regression coefficient of historical medal count is large and positive, it implies that historical medal count has a strong influence on the prediction of gold medal count.

The prediction results will provide the number of gold medals or total medals for each country in the 2028 Olympic Games. At the same time, we can quantify the uncertainty of the prediction based on the error assessment of the model (such as MSE and R^2). If the error is small, it indicates that the model's prediction is relatively accurate.

7 Model Evaluation and Further Discussion

7.1 Strengths

1. The curve established in Question (1) has a simple analytical expression and the error test MRE is very small.
2. Question (2) can be upgraded and optimized into a model by using the VARMAX method.
3. The curve established in Question (3) has a coherent and seamless effect and the simulation error is very small.
4. The curve established in Question (4) conforms to the regulations and is concise and clear, ensuring the correctness of the prediction results.
5. The curve established in Question (5) clearly contrasts the changes in the number of medals with and without a great coach.

7.2 Weaknesses

1. Further exploration is needed on the more detailed applicable conditions of each model.
2. Economic and other factors were not given detailed consideration due to the lack of data.

7.3 Further Discussion

The VARMAX model can be utilized to optimize the solution for the growth of the number of medals won.

The VARMAX (Vector Autoregressive Moving Average with exogenous variables) model [12] is a multivariate time series model that combines the characteristics of autoregression (AR) and moving average (MA), and allows the introduction of exogenous variables (X).

The following steps can be followed:

(1) Variable selection:

① Dependent variable: The number of medals won by each country (possibly including gold, silver, and bronze medals).

② Independent variables: Historical medal numbers (as time series data), the number of participants, the experience of the athletes, etc., as well as possible exogenous variables such as the "Great Coach" effect index.

(2) Setting of the VARMAX model:

① Based on the characteristics of the data and the requirements of the problem, set the structure of the VARMAX model, including the order of autoregression, the order of moving average, and the way of introducing exogenous variables.

② Ensure that the model can capture the dynamic relationships in the time series data and reflect the influence of exogenous variables on the dependent variables.

The results are as follows:

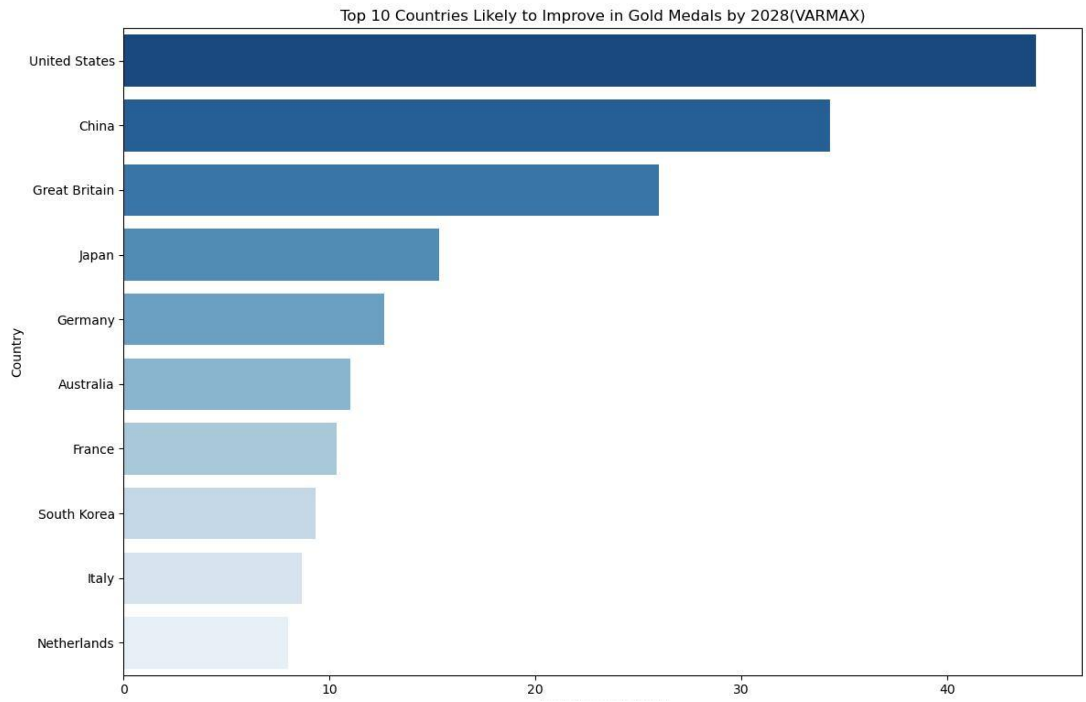

Figure 19: Gold Medal Progress

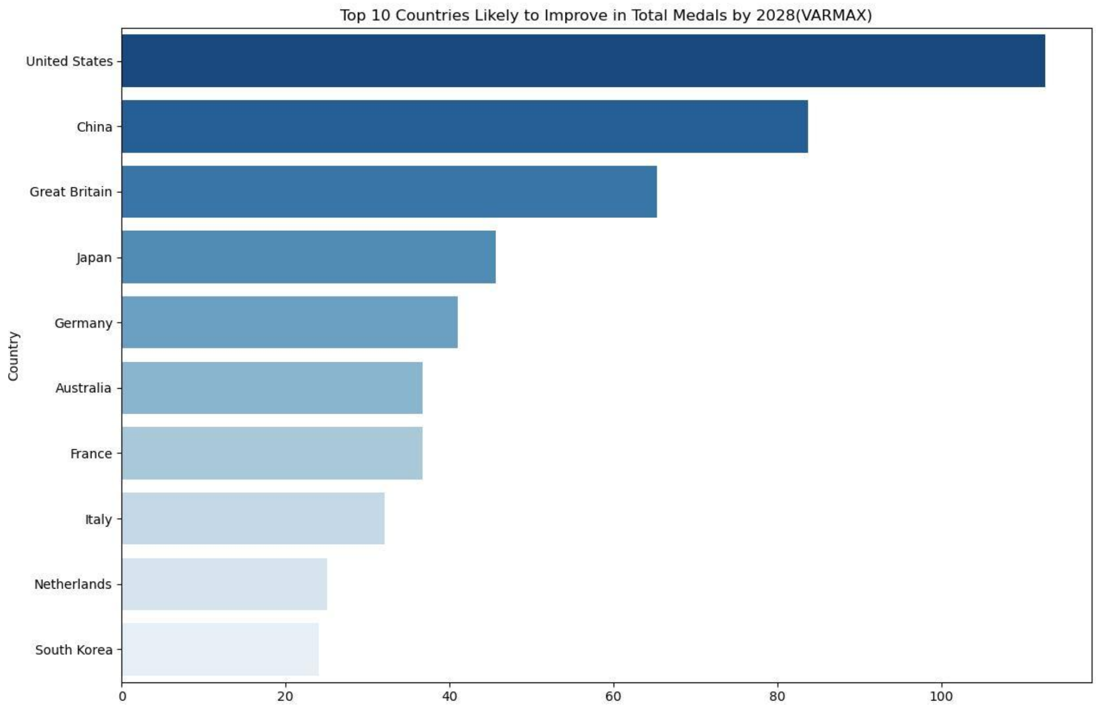

Figure 20: Total Medal Progress

8 References

[1]Suszek A ,Guze S . A Logistic Regression Model for the Analysis of Attitudes and Behaviours Towards Functional Foods Among Senior Consumers Aged 60+ Years [J]. Sustainability, 2024, 16 (24): 11015-11015.
[2]Kusanovic D ,Ayoubi P ,Seylabi E , et al. A supervised approach for improving the dimensionless frequency estimation for time‐domain simulations of building structures on embedded foundations [J]. Earthquake Engineering & Structural Dynamics, 2024, 53 (9): 2782-2804.
[3]Yuheng L . Estimating coastal premiums for apartment prices: Towards a new multilevel modelling approach [J]. Environment and Planning B: Urban Analytics and City Science, 2022, 49 (1): 188-205.
[4]Savkovic B ,Kovac P ,Dudic B , et al. Decoding of Factorial Experimental Design Models Implemented in Production Process [J]. Computers, Materials & Continua, 2022, 71 (1): 1661-1675.
[5]Amiri S ,Goulet A J ,Trépanier M , et al. Modeling Transportation Time Series using Bayesian Dynamic Linear Models [J]. Transportation Research Procedia, 2025, 82 3352-3360.
[6]Zeng Yifei, Shi Hong. Application of ARIMA model in GDP forecasting of Guangxi Zhuang Autonomous Region [J]. Economic and Trade of the Era, 2025, 22 (01): 129-133. DOI:10.19463/j.cnki.sdjm.2025.01.020.
[7]Wang C ,Li S . The construction of modern aesthetic education innovative talent cultivation mode in universities based on big data technology [J]. Applied Mathematics and Nonlinear Sciences, 2024, 9 (1):
[8]Radloff M ,Schwabe R . Locally D-optimal designs for a wider class of non-linear models on the k-dimensional ball [J]. Statistical Papers, 2019, 60 (2): 165-177.
[9]Hong S ,Wang L ,Lin T , et al. Mathematical analysis for CSI scheme with the interpolation kernel size increased [J]. IET Image Processing, 2017, 11 (8): 595-604.
[10]Zaky A Z ,Dossari A M ,Hussien M A M , et al. Modelling aluminium nitride’s refractive indices under various situations for optical simulations: a mixed research [J]. Optical and Quantum Electronics, 2024, 56 (10): 1728-1728.
[11]Wang Chuwen. Research on the Effect Measurement and Regional Heterogeneity of Digital Economy's Empowering on the Optimization of the Internal Structure of the Tertiary Industry[D]. Dongbei University of Finance and Economics, 2023. DOI:10.27006/d.cnki.gdbcu.2023.001984.
[12]K. L ,J. P . Decentralized damage diagnostic technique for tall buildings using VARMAX model [J]. Earthquake Engineering and Engineering Vibration, 2022, 21 (2): 417-439.

Report on Use of AI

1. OpenAI *ChatGPT* (May 14, 2024 version, ChatGPT-4o)
   Query1: *Request on translation of problem*
   Output1: *(problem translation)*
   Query2: In problem (2), we will use ARIMA model. Give us the formula of ARIMA model and its applicable conditions.
   Output2:
   ARIMA (Autoregressive Integrated Moving Average) is a widely used time series analysis method suitable for processing time-dependent continuous data. The form of the ARIMA model is:
   $$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$
   The ARIMA model provides more accurate predictions by adjusting the values of p and q to capture autocorrelations and random fluctuations in time series. The goal of this logistic regression model is to learn the regression coefficients β0, β1,..., βn from the given training data. The learning process of regression coefficients is achieved by maximizing the likelihood function. The purpose of maximizing the likelihood function is to make the probability value predicted by the model as close as possible to the actual observed label.
   Query3: The following is a summary of model fitting for a Poisson regression model. Use the data to generate a sensitivity heat map and a sensitivity uncertainty map for the model (use python’s matplotlib). Note that all the data you need is in this text. *Following is the model abstract*
   Output3:
   To generate sensitivity heat maps and sensitivity uncertainty maps, it is first necessary to clarify how to extract information from the provided Poisson regression model results. Sensitivity heat map: This graph usually shows the size of the influence of each independent variable on the target variable. From the regression coefficient (coef) and its standard error (std err), we can calculate the effect of each coefficient. We will use the standard error of the regression coefficient to measure the uncertainty. Sensitivity Uncertainty plot: This is a plot that shows the sensitivity of the model's output to the input parameters, and is usually plotted based on the confidence interval ([0.025, 0.975]) of the regression coefficient. Here is the Python code that generates these two diagrams:
   “”””

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

coefficients = {
        'Intercept': 6.5920,
```

```
        'C(Sport_FE)[T.Gymnastics]': 0.2366,
        'C(Sport_FE)[T.Volleyball]': 1.0672,
        'Coach': -0.0910,
        'Year_FE': -0.0026
}

std_errors = {
        'Intercept': 0.585,
        'C(Sport_FE)[T.Gymnastics]': 0.253,
        'C(Sport_FE)[T.Volleyball]': 0.256,
        'Coach': 0.197,
        'Year_FE': 0.000
}


ci_lower = {
        'Intercept': 5.446,
        'C(Sport_FE)[T.Gymnastics]': -0.259,
        'C(Sport_FE)[T.Volleyball]': 0.566,
        'Coach': -0.478,
        'Year_FE': -0.003
}

ci_upper = {
        'Intercept': 7.738,
        'C(Sport_FE)[T.Gymnastics]': 0.732,
        'C(Sport_FE)[T.Volleyball]': 1.568,
        'Coach': 0.296,
        'Year_FE': -0.002
}


df = pd.DataFrame({
        'Coefficient': coefficients.values(),
        'Std_Error': std_errors.values(),
        'CI_Lower': ci_lower.values(),
        'CI_Upper': ci_upper.values()
}, index=coefficients.keys())

sns.set(style="whitegrid")
plt.figure(figsize=(8, 6))
```

```
sns.heatmap(df[['Coefficient']], annot=True, cmap="coolwarm", cbar_kws={'label': 'Coefficient Value'})
plt.title('Sensitivity Heatmap: Coefficients')
plt.tight_layout()
plt.show()


plt.figure(figsize=(8, 6))
for feature in df.index:
        plt.plot([df.loc[feature, 'CI_Lower'], df.loc[feature, 'CI_Upper']], [feature, feature], label=feature, marker='o')
        plt.scatter(df.loc[feature, 'Coefficient'], feature, color='black')

plt.xlabel('Coefficient Value')
plt.ylabel('Features')
plt.title('Sensitivity-Uncertainty Plot')
plt.grid(True)
plt.tight_layout()
plt.show()
```

“”””

Sensitivity heatmap: seaborn.heatmap is used to draw the heatmap of the regression coefficient, showing the influence of each independent variable on the target variable. Through Coefficient column, the heat map shows the influence degree of each independent variable. Sensitivity Uncertainty plot: This is a representation of the sensitivity and uncertainty of the model output through the confidence interval of the regression coefficient. The confidence interval for each variable is represented by a line at either end, while the position of the regression coefficient is indicated by a black dot. Sensitivity thermogram: Shows the magnitude of the regression coefficient to help identify which variables have the greatest influence on the target variable. Sensitivity Uncertainty plot: Confidence intervals and actual estimates for regression coefficients are shown to help understand the uncertainty of these estimates. This code will generate the corresponding visualization based on the coefficient data of the Poisson regression model that you provided.

2. Kimi (Jan 20, 2025 version, k1.5)
   Kimi is a LLM that experts in reading long text files. Therefore, we took advantage of it to read long technological essays.
   Query1: Please analyze this article introducing the random forest model and answer my questions about the applicable conditions of the random forest model. And how to use the model correctly?
   Output1:
   - **Applicability of Random Forest Models**

Random forest models are suitable for the following scenarios:

- **Large Datasets:** Random forests perform exceptionally well when processing large datasets efficiently.
- **High-Dimensional Data:** They can handle thousands of explanatory variables, making them suitable for high-dimensional datasets.
- **Nonlinear Relationships:** Random forests can capture nonlinear relationships between variables, making them ideal for datasets with complex nonlinear interactions.
- **Variable Interactions:** They can model interactions between variables, making them applicable to data with variable interdependencies.
- **Robust to Outliers:** Random forests are insensitive to outliers and work well even when outliers are present in the data.
- **No Variable Selection Required:** They do not require variable selection, making them suitable for datasets with numerous variables where selection is challenging.

- **How to Use Random Forest Models Correctly**

1. Data Preparation:

- **Data Cleaning:** Handle missing values using the built-in methods of random forests, such as median or mode imputation.
- **Data Standardization:** Although random forests have minimal distribution requirements, standardizing data can enhance stability and prediction accuracy.

2. Model Parameter Settings:

- **ntree:** Specifies the number of decision trees to generate. Generally, a larger number improves model performance but increases computational cost. Use cross-validation or out-of-bag (OOB) error to determine the optimal number of trees.
- **mtry:** Determines the number of variables considered for splitting at each node. For classification tasks, the default is the square root of the total variables; for regression tasks, it is one-third of the total variables. Adjust mtry to optimize performance based on the specific problem.

3. Model Training:

- Train the model using a random forest algorithm. You can use tools like the randomForest package in R or other libraries that support random forests.

4. Model Evaluation:

- **Error Evaluation:** Random forests use OOB error to assess model performance. For classification, it is the classification error rate; for regression, it is the variance of residuals.
- **Variable Importance:** Random forests can evaluate the importance of each variable for predicting the target variable. Use the importance parameter to obtain these values.

5. Model Optimization:

- **Parameter Tuning:** Optimize model performance by adjusting parameters such as ntree and mtry.

- **Cross-Validation:** Use cross-validation to evaluate the model's generalization ability and avoid overfitting.

6. Model Application:

- **Prediction:** Use the trained model to make predictions on new data.
- **Interpretation:** Interpret the model results using tools like variable importance and partial dependence plots to understand the effects of variables on the target variable.

- **Considerations**
- **Class Imbalance:** Random forests may favor the majority class in imbalanced datasets. Address this by adjusting sample weights or using oversampling/undersampling methods.
- **Model Interpretation:** While random forests excel in prediction accuracy, they are less interpretable. Use tools like partial dependence plots to aid interpretation.
- **Computational Resources:** Random forests require significant computational resources, especially when processing large and high-dimensional datasets. Ensure sufficient resources are available for model execution.

By following these steps and considerations, you can effectively utilize random forest models and leverage their strengths in classification and regression analysis.

3. Tabnine
   Auto-completions for code used in preparing our models.