International Journal of Advanced Science

https://doi.org/10.70731/c3q02875

Application of Data-Driven Prediction and Strategic Optimization in Olympic Medal Distribution

Zhaoyu Zhu [a], **Jiajin Sheng** [a], **Minghao Yu** [a], **Chengtian Liang** [a,*]

[a] *School of Physics, Hangzhou Normal University, Hangzhou 311121, China*

KEYWORDS

XGboost,
AHP model,
Logistic Regression,
Multiple Information Regression Equation,
K-Means

ABSTRACT

The Olympic medal list is an important indicator to assess the competitive strength of countries, and the prediction and analysis of the distribution of the number of medals provide a scientific basis for countries to formulate sports development strategies. This paper takes the 2024 Paris Olympic Games and the previous Olympic Games as the basic data, combines the historical medal data, the distribution of each Olympic Games and the special characteristics of the host country, constructs a number of mathematical models, explores the law of medal distribution, and proposes a strategy to improve the number of medals.The model in this paper is comprehensive, flexible and practical, which provides a new way of thinking for the analysis of medal distribution in the Olympic Games, and also provides data support for the sports development strategy of each country.

1. Introduction

The Olympic medal table demonstrates the overall strength of countries in different sports[1]. At each Olympic Games, the performance of countries in gold, silver and bronze medals reflects the preparation of their athletes, the level of training and policy support. As training methods continue to innovate and competition intensifies, certain factors, such as the 'Great Coach Effect,'are emerging as having an impact on medals. In addition to the role of coaches, the selection and development of Olympic sports also directly affects the number of medals won by each country. By analyzing historical data, we are able to reveal which sports contribute the most to the medal count of each country, and how to improve Olympic performance by optimizing the selection of sports, improving training methods and bringing in great coaches[2].

Considering the background information and restricted conditions identified in the real-life problem, we determine to solve it by dividing it into following tasks:

- Modeling the number of medals for each country: A model was developed to predict the number of medals for each country, focusing on the number of gold medals and the total number of medals, and estimating the uncertainty and accuracy of the predictions.
- Predicting the 2028 Los Angeles Olympics Medal Standings: Based on the model, predict the medal performance of each country at the 2028 Summer

* Corresponding author. E-mail address: lct.lctsoft@hotmail.com

Received 9 January 2025; Received in revised from 20 February 2025; Accepted 25 February 2025; Available online 28 February 2025.

Table 1 | Notations used in this paper

Symbol	Description
N	Number of countries
$Host_{i,t}$	Host country characteristics
$Y_{i,gold}$	Number of gold medals won by country i at the t th Olympics
Y_i	Number of medals won by country i at the t th Olympics
NOC_i	National Olympic Committee of the i th country
$num_{sports,i}$	Number of sports in which the i th country participates

Olympics in Los Angeles, USA, and give a prediction interval to analyze which countries are likely to improve or decline in performance.

- Predictions for countries that have not yet won a medal: Estimates of which countries that have not yet won a medal are likely to win one for the first time at the next Olympic Games, with corresponding probabilities.
- The Impact of Olympic Programs on Medal Counts: Analyzes how the number and type of Olympic programs affect medal counts in each country and explores the significance of specific programs for different countries.
- The Great Coach Effect study: examines the potential impact of great coaches on medal counts and selects three countries to assess the likely impact of their investment in great coaches.
- Provide insights to National Olympic Committees (NOCs)

2. Assumptions and Justifications

Considering that real problems always contain many complex factors, first we need to make reasonable assumptions to simplify the model, and each assumption is immediately followed by the corresponding explanation:

- Assumption 1: Medal counts from previous Olympics are independent of each other.
- Explanation: This would make the modeling simpler even though there may actually be some long-term trends or cyclical factors.
- Assumption 2: The actual medals won are based on the data given in the title.
- Explanation: The question explicitly uses the data given, but this is somewhat different from the actual data.
- Assumption 3: Athletes from each country have a relative stability of performance in different competition disciplines, with a career of no more than four Olympic cycles.
- Explanation: This will enable better use of the regression model for prediction and enhance the accuracy and stability of the model prediction

3. Notations

The key mathematical notations used in this paper are listed in Table 1.

4. Data Processing and Analysis

4.1. Data Pre-Processing

Immediately after getting the data we realized that there are many problems with the data in many ways. First of all, we observed that there are many missing values in the data, for example, there are two missing values in columns 1928 and 1932. At the same time, there are a lot of anomalies or errors in the data, for example, in the summerOly_programs.csv file there are a lot of garbled codes like '聳' and '1906*', in the summerOly_athletes.csv has 1466 lines of duplicate data.

4.2. Descriptive Statistical Analysis

After taking care of these obvious data formatting issues we started to preprocess the data to make it fit our modeling needs. For the athlete data, we summarized it by country. We also harmonized the naming of countries in the different files and merged the different data sets (e.g., athlete data, medal data, event data, etc.). After a series of preprocessing was completed, we performed further cleaning, this time focusing on unreasonable as well as invalid data, as well as looking at missing values to see if they affected, for example, the reasons for cancellations due to the effects of World War II, and deleting countries and leagues that no longer existed in order to ensure the continuity of the data over time.

We visited https://odf.olympictech.org/project.htm to correct some of the missing values that could be modified.

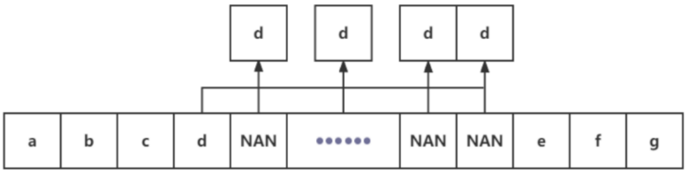

Figure 1 | Missing value insertion

Based on the data given, we analyzed the data using statistical methods and plotted some of the statistical graphs:

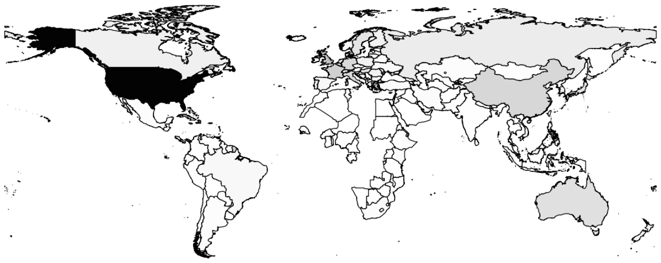

Figure 2 | All-time medals by country

4.3. Norm Normalization

Because there may also be outliers or outliers in our data, we chose to use RobustScaler to normalize the data for files with a high number of outliers in each category to ensure that there is no impact on the subsequent modeling process.

RobustScaler's normalization formula is as follows:

$$X_{\text{scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)} \tag{1}$$

In the following model description, we will use the data processed above for further operations.

5. Medal Prediction Modeling and Correlation Analysis

5.1. Forecast of the Olympic Medal Table

Notice that the context of the question explicitly asks us to predict is the number of medals (both gold and total) for each country. This task is essentially a regression problem.

The number of Olympic medals is influenced by many characteristics, including the historical performance of the athletes, the sports development of the country, and the historical medal table, and there is often a complex non-linear relationship between these factors. We cannot make predictions based only on historical medal counts, so we include athlete data, host country effects.

- **Athlete Data**

The number of athletes a country sends has a significant impact on its potential to win medals. Typically, countries with more participants will rank higher in the medal table. At the same time, an athlete's individual performance (e.g., winning a gold, silver, or bronze medal or not) directly affects the final medal distribution.

- **Host effect**

In particular, a new feature was added to the model: a host country indicator variable. For each Olympic Games: if a country is the host country for that year, that country is given a binary (0 or 1) feature in the data for that Olympic Games indicating whether it is the host country or not. Other non-host countries are assigned the feature 0.

$$Host_{i,t} = \begin{cases} 1, & \text{if country } i \text{ is the host in year } t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

5.1.1. Feature Engineering

After scanning the data, we found that there was still singular data that had a large impact on the results, and we normalized the data again in order to reduce the impact of outliers in the model.

For categorical variables such as 'sport' or 'country code' (NOC) we converted them to numerical features by unique heat coding or label coding to facilitate model processing. We further analyzed these features to determine their contribution to the prediction of medals. Therefore, a Recursive Feature Elimination (RFE) algorithm[3] was used to rank the different features. Finally, ‘historical medals’,‘athlete performance’ and ‘host country effect’ were selected as the three dominant features.

5.1.2. Construction of Forecasting Models

The prediction of the number of Olympic medals does not depend only on a single factor, but is influenced by a combination of several factors (e.g., the number of gold medals, silver medals, bronze medals, the host country effect, the performance of

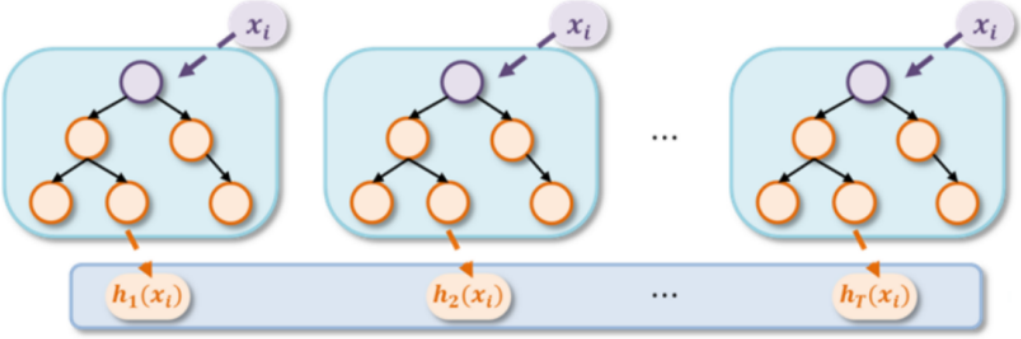

Figure 3 l Schematic diagram of the XGboost model

the athletes, and the country's historical performance). There are usually complex non-linear relationships between these factors. Here we have picked XGboost model.

- **XGBoost Model**

For a dataset containing n entries of m dimensions, the XGBoost model can be represented as:

$$y_i = \sum_{K=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1,2,\ldots,n) \tag{3}$$

$$\text{where,} \quad F = \left\{ f(x) = \omega_{q(x)} \right\} \quad \left(q : R \in R^T \right) \tag{4}$$

It is a set of CART decision tree structures, where q is the tree structure mapped from the sample to the leaf node, T is the number of leaf nodes, and w is the real fraction of the leaf node.

When constructing the XGBoost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function in order to build the optimal model.The objective function of the XGBoost model can be divided into an error function term L and a model complexity function term Ω. The objective function can be written as:

$$Obj = L + \Omega \tag{5}$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

where, γT is the L1 regular term, $\frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$ is the L2 regular term.

When training the model optimally using the training data, it is necessary to keep the original model unchanged and add a new function f to the model so that the objective function is reduced as much as possible.

- **Building Olympic Medal Forecasting Models**

The following are the steps in the construction of the model.

Step 1. Initialize the model

In gradient boosting regression, the model begins training with an initial estimate of the target variable. For regression problems, the initialization is usually the mean of the target variable Y:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{8}$$

Here, y_i is the target value for the i th sample.

Step 2. Residual Calculation

In each iteration, we compute the residual between the current model prediction and the true target value. The residual $r_i^{(m)}$ is the difference between the actual value y_i for the i th sample and the current predicted value $F_{m-1}(x_i)$:

$$r_i^{(m)} = y_i - F_{m-1}(x_i) \tag{9}$$

In the first iteration, the residual is equal to the difference between the actual target value and the initial predicted value.

Step 3. Train the weak learner

Next, we train a weak learner to fit the residuals. The training objective is to minimize the loss function of the residuals. Use the residuals of the target variable as the objective of the new model:

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} \left(r_i^{(m)} - h(x_i) \right)^2 \tag{10}$$

This is the optimal decision tree to be found when fitting the residuals $h_m(x)$

Step 4. Model update

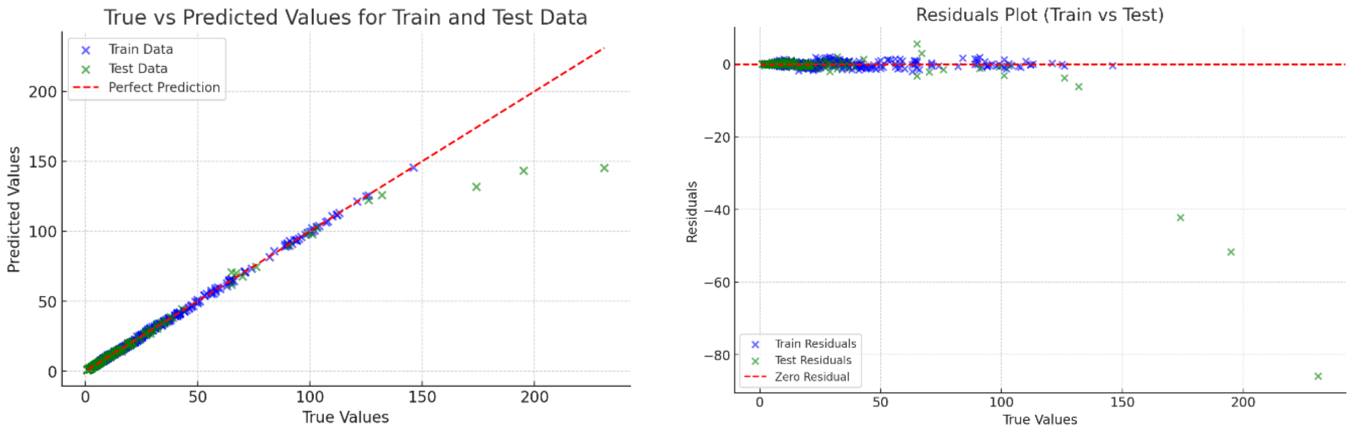

Figure 4 l Plot of true versus predicted values with residuals for the adjusted training and test sets

With each iteration, we update the model by adding the predictions of the new learner to the previous model. The update rule for the new model is:

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x) \tag{11}$$

Step 5. Final model expression

The final gradient boosting regression model is a weighted sum of multiple trees. Assuming that we have n rounds of iterations, then the final prediction model is

$$F_M(x) = F_0(x) + \sum_{m=1}^{n} \alpha_m h_m(x) \tag{12}$$

Immediately after that, we started to train the XGBoost model with the model constructed above, by dividing the dataset into a training set and a test set, where the first 80% is used as the training set for model training and the second 20% is used as the test set for validation. The coefficient of determination is found to be as high as 0.90 and the mean square error (MSE): 41.72. this means that the model predicts quite well on the test set and the model is able to explain the vast majority of the variability. The smaller MSE value also indicates that the model has less error.

5.1.3. Model Uncertainty Analysis and Optimization

- **Uncertainty analysis**

In this study, the accuracy of each model was validated using the cross-validation method, and the mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and R_2 generated from the cross-validation results were used as evaluation metrics for estimating the model and validating the accuracy of the model.

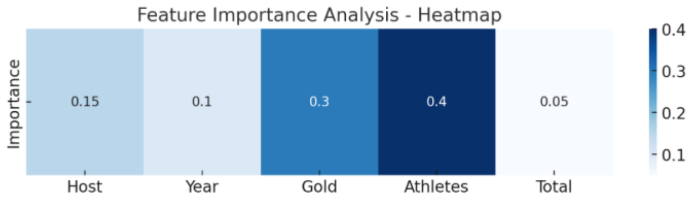

Figure 5 l Characteristic Importance Analysis

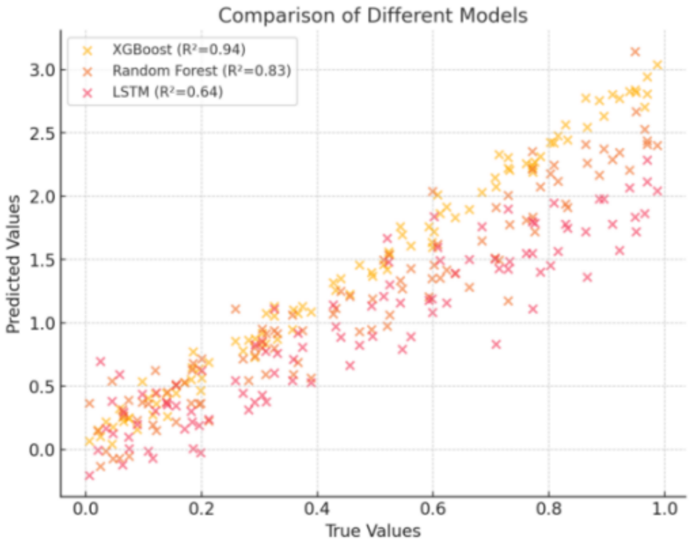

Figure 6 l Comparison chart of different models

To avoid overfitting, we usually need to tune some key hyperparameters. Subsequently, we performed hyperparameter optimization, starting from the learning rate, depth of the tree, minimum number of sample splits of the tree, and number of iterations, to improve the important hyperparameters, and succeeded in tuning the coefficient of determination on the test set by R^2 to 0.94.

Table 2 | Predicted top ten in the gold medal table and their confidence intervals

Country	Predicted Gold	Gold Lower Bound	Gold Upper Bound
United States	41.92	39.1738470180344	44.6661529819656
China	36.21	33.4638470180344	38.9561529819656
France	21.05	18.3038470180344	23.7961529819656
Great Britain	20.04	17.2938470180344	22.7861529819656
Australia	19.01	16.2638470180344	21.7561529819656
Russia	18.61	15.8638470180344	21.3561529819656
Japan	17.06	14.3171803513678	19.8094863152989
Italy	13.86	11.1138470180344	16.6061529819656

Table 3 | Predicted top ten in the medal table and their confidence intervals

Country	Predicted Total	Total Lower Bound	Total Upper Bound
United States	124.2	121.0132539	127.3867461
China	85.29	82.10325389	88.47674611
Great Britain	71.78	68.59325389	74.96674611
France	70.37	67.18325389	73.55674611
Russia	55.45	52.26325389	58.63674611
Australia	54.58	51.39325389	57.76674611
Japan	41.98	38.79325389	45.16674611
Italy	41.35	38.16325389	44.53674611

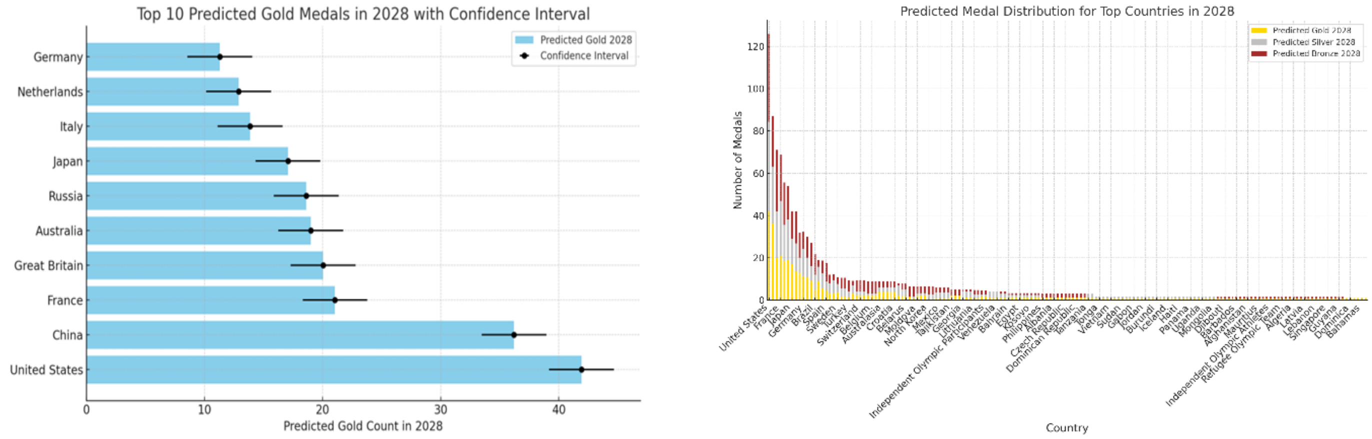

Figure 7 | Number of medals for different countries in 2028

- **Model Evaluation**

In order to more clearly represent the extent to which different features contribute to the prediction model, an analytical plot of feature importance is drawn here It is clear that the performance of athletes in recent events, as well as the number of historical medals, make a relatively large contribution to the final results of the model.

Based on the comparison with other models, we find that XGboost has a better fit in this problem, and we also find that LSTM is a poor predictor in this problem.

5.1.4. Forecast Results and Analysis

Based on the model above, we have obtained the medal table predictions, including the number of gold, silver and bronze medals and the total number of medals for the 28 years of the Olympic Games, as well as the confidence intervals for these medal predictions. Due to the large number of countries, we

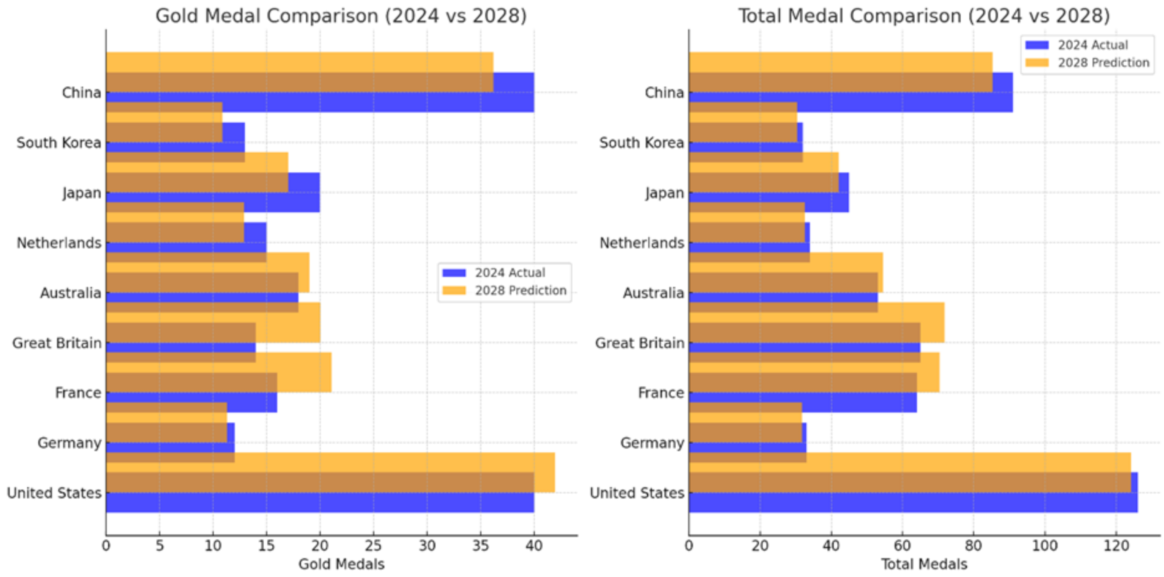

Figure 8 I Comparison of gold medals and medals for selected countries

have selected the top 10 in the gold and medal tables to show the data. (95% confidence intervals obtained here)

The following image shows the 2028 predicted gold medals with their confidence intervals and the number of gold, silver and bronze medals for all countries on the medal table.

At the same time, we compared the medal table with the 2024 Olympic medal table and the predicted 2028 medal table to analyze which countries are likely to advance and which countries are likely to regress. After comparing, we found that 50 countries, represented by Great Britain and Moldova, will improve in 2028, and 29 countries, represented by Uzbekistan, South Korea and Netherland, will decline.

After looking at the population and GDP data for these countries it was found that the predictions were very close to the results predicted using population and GDP which indirectly proves the accuracy of the model's predictions.

After looking at the population and GDP data for these countries it was found that the predictions were very close to the results predicted using population and GDP which indirectly proves the accuracy of the model's predictions.

5.2. Forecast of Countries Receiving Awards for the First Time

Predicting whether a country that has not yet won a medal will be able to win its first medal at the next Olympic Games, we based our prediction model on the Olympic medal table and considering that this is a probabilistic or classification problem, we modified the model and introduced a logistic regression model along with a machine learning approach to predict the probability of countries that have not yet won a medal will be able to win their first medal for the countries that have not yet won a medal.

5.2.1. Logistic Regression Models

In this problem, we set'whether or not to win a medal'as the categorization target, and set it to 0 (not winning a medal) and 1 (winning the first medal).

I Related formulas

The model is of the form:

$$P(y=1\mid X)=\frac{1}{1+e^{-(\omega_0+\omega_1x_1+\omega_2x_2+\ldots+\omega_nx_n)}} \quad (13)$$

I Model Training

During training, our goal is to find a set of parameters w^0,w^1,w^n, such that the model is able to predict the category y=1 or y=0 as accurately as possible For this purpose, we usually use Maximum Likelihood Estimation (MLE) to estimate the model parameters. By optimizing the loss function of the model, the probability of predicting the outcome is made as close as possible to the actual observed category labels given the features.

The loss function is calculated as:

$$\text{Log-Loss}=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right) \quad (14)$$

5.2.2. Predicting the Probability of Winning a Prize

The following is the prediction result we got after modifying the model as shown in Fig. 9, in order to

Table 4 | Demonstration of predicted probabilities for selected countries

Country Name	2028_Medal_Prob
Tuvalu	0.662356228176717
Nauru	0.662171373161028
Marshall Islands	0.661939667405879
Palau	0.661911988406352
Kiribati	0.66182316707161

Table 5 | Predicted outcomes and probabilities

Number of countries awarded Medals	Average Predictive Rate
51	62.5%

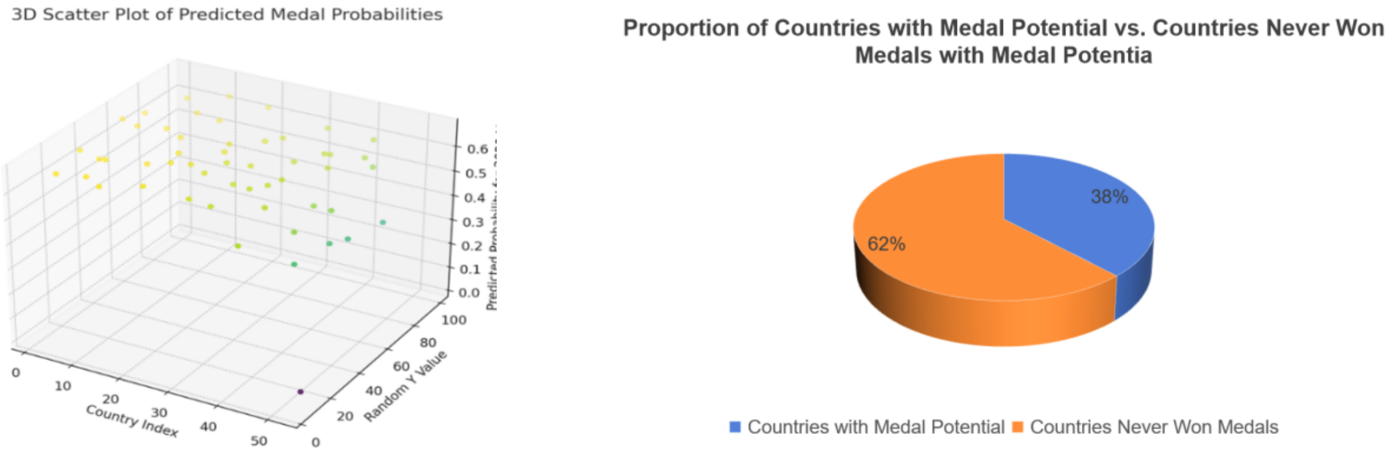

Figure 9 | Visualization of the probability of a first-time licensee country

facilitate the display we adopt the way of Country Index to show and through the three-dimensional scatter plot to more intuitively show the probability of each country that has not won the prize may win the medal. The probability of each country winning a medal is summarized to give a prediction of the total number of countries and their probabilities.

From the data in the calculations, it was obtained that the average probability of winning the first medal was 62.5 percent

- **Model measurement**

For this predicted model measure we, in this study, used the cross-validation method to validate the accuracy of each model and introduced the accuracy,precision, recall, ROC AUC value as the evaluation metrics for estimating the model and validating the model accuracy.

The following are the visualization results under some of these metrics:

The visualization shows a certainty of 0.89 for giving this prediction, i.e. there is an accuracy of 89% for predicting the countries that will be able to win medals at the next Olympic Games.

5.3. Correlation Between Projects and Number of Awards

On the previously constructed model as well as the trained data, and also with the help of cluster analysis (K-Means) to help us understand the distribution and characteristics of the data, and to discover potential data patterns, we can identify outliers or noise in the data and improve the quality of the data; and finally, we divide the analysis into the following three aspects:

- Analyze the relationship between the number of medals and the number of events: Explore whether there is a correlation between the number of events

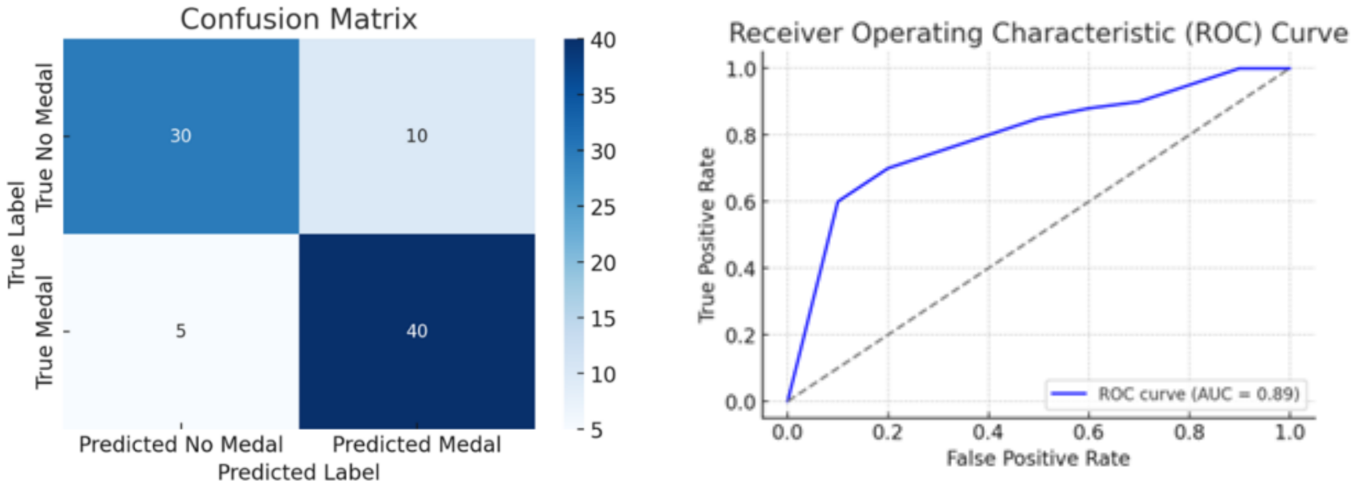

Figure 10 I Confusion Matrix and ROC Curve

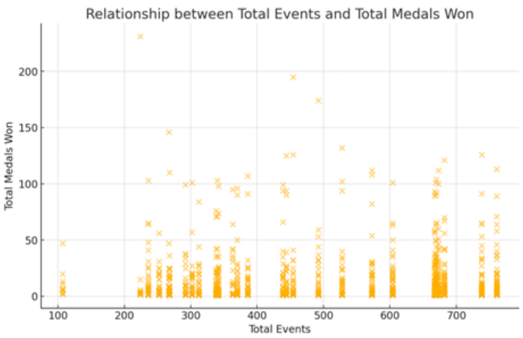

Figure 11 I Relationship between the number of medals and the number of programs

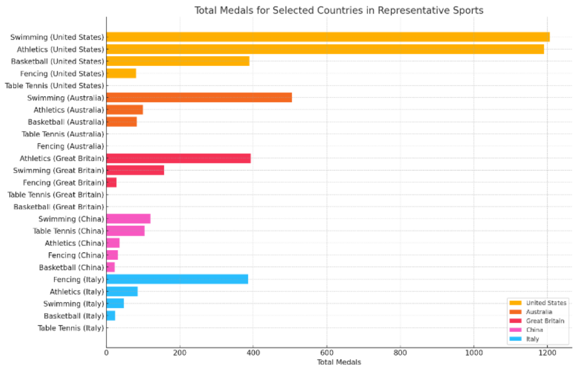

Figure 12 I Comparison of medals won by typical programs

Table 6 I Number of medals won by dominant programs

United States	Australia	Great Britain	Italy
2396	505	712	385

at each year's Olympic Games and the number of medals won by each country.

- Analyzes which programs are the most important for different countries: look at each country's performance in the different programs and identify the strengths and important programs for each country.

Figure 11 shows a scatter plot of the relationship between the number of medals and the number of projects. As can be seen from the graph, while some years have a higher number of events, this does not always mean that the country has won more medals. This may be related to the number of participating countries, the performance of each country and other variables of the Olympic Games (e.g., type of event).

As we can see from the following graph: the United States excels in swimming and track and field. China has accumulated notable medals in table tennis and diving. Italy has a strong presence in fencing.

As we can see from the graph: the United States excels in swimming and track and field. China has accumulated notable medals in table tennis and diving. Italy has a strong presence in fencing.

We also have the Great Britain dominating in badminton, croquet and other less popular disciplines, winning most of the medals in these events. The United States won a relatively large number of medals in golf, but a slightly smaller share of the overall medals (around 70%). Switzerland and Spain are also notable performers in aeronautics and

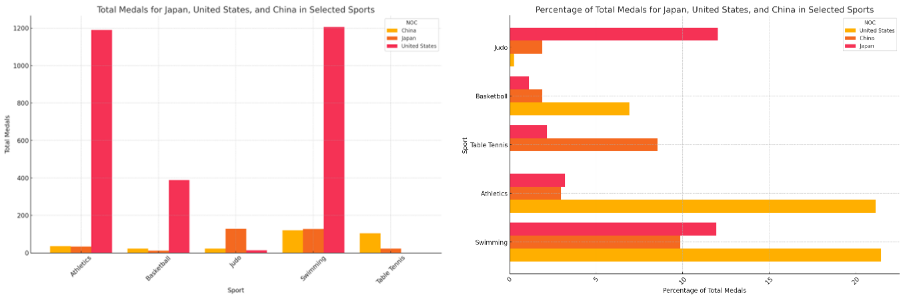

Figure 13 I Typical program's medal count and medal share

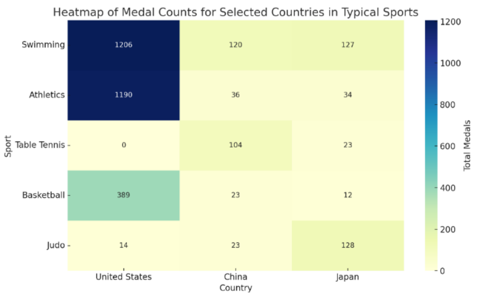

Figure 14 I Heat Map of Medal Counts vs. Typical Events

Basque ball. These programs directly contributed to the higher number of medals won by the countries.

Next we took the top three countries in the medal table and plotted their number of medals in typical events versus the percentage of those medals to the total (i.e., how much they contributed to their country's closing result). Clearly showing which events gave them an advantage in the medal table.

We conclude that the United States is the top performer in swimming, track and field, and basketball, contributing a large percentage of the medals. China is particularly strong in table tennis and diving. Japan has a prominent position in Judo.

At the same time we analyzed the impact of the program on the average country, Algeria got 60% of its medals from the strongest events. The West Indies Federation has almost all of its medals coming from the Strengths Program, at 100%. Zimbabwe, with 31.8% of its medals coming from strong events, shows that its medals are more spread out. This shows the great influence of the strongest events on the results in some countries.

6. Anfalysis of the Effect of Great Coaches

6.1. Feature Extraction

Model Assumptions: The assumption is that the women's volleyball team participates in three Olympic cycles. As volleyball, like other ball sports, is a high-intensity sport, athletes' careers are often seen to span about three Olympic cycles. Therefore, when creating the dataset, we focus on data from three Olympic Games.

The data provided in the problem covers a total of 239 countries and regions. However, not all records are relevant because, in many cases, both the medal counts (Y) and the feature data (X) are zero, which are not useful for analysis. For the model's input and output, 'rows' represent the samples (a country and an extended period spanning over 12 years), and 'columns' represent various features, with each feature taking its own column.

As a result, we have clearly defined the structure of the input and output data, which includes three main categories of indicators: Athlete Category, Medal Count Category, and Proportional Change Category, as outlined below:

Athlete Category: p

1) $p_{1,t}$, Among the participants of that year, the number of experienced athletes (those who participated 4 years ago, regardless of whether they won an award or not).
2) $p_{2,t}$, Among the participants of that year, the number of outstanding experienced athletes (those who participated and won medals 4 years ago, regardless of whether they won gold, silver, or bronze)
3) $p_{3,t}$, Among the participants of that year, the number of core athletes (those who participated 4 years ago and 8 years ago, regardless of whether they won an award or not).
4) $p_{4,t}$, Among the participants of that year, the number of outstanding core athletes (those who participated and won medals 4 years ago and 8 years

Table 7 I Key items as a percentage of overall medals

Algeria	West Indies Federation	Zimbabwe
60%	100%	31.8%

Table 8 | Comparison of Medal Counts With and Without the 'Great Coaches' Effect

The probability of obtaining the corresponding medal	Without the 'Great Coaches' effect	Actual results
Gold Medal	35.78	0.43
Silver Medal	27.05	98.91
Bronze Medal	37.17	0.66

ago, regardless of whether they won gold, silver, or bronze).

5) $p_{5,t}$, The number of participants from 4 years ago.

6) $p_{6,t}$, The number of participants from 8 years ago.

Medal Count Category: q

7) $q_{7,t}$, The number of gold medals won by athletes 4 years ago.

8) $q_{8,t}$, The number of silver medals won by athletes 4 years ago.

9) $q_{9,t}$, The number of bronze medals won by athletes 4 years ago.

10) $q_{10,t}$, The number of gold medals won by athletes 8 years ago.

11)$q_{11,t}$, The number of silver medals won by athletes 8 years ago.

12)$q_{12,t}$, The number of bronze medals won by athletes 8 years ago.

Proportional Change Category: s

13)$s_{13,t}$, Change in participation from 8 years ago to 4 years ago

14)$s_{14,t}$, Change in medal count from 8 years ago to 4 years ago

15)$s_{15,t}$, Change in participation from 4 years ago to the current year

16)$s_{16,t}$, Change in medal count from 4 years ago to the current year

In summary, the multiple linear regression multi-output prediction model established for the second question, without considering the 'great coach' effect, is as follows:

$$\begin{cases} u_t^{\text{gold}}, u_t^{\text{silver}}, u_t^{\text{bronze}} = & \lambda_1 P_{1,t} + \lambda_2 P_{2,t} + \lambda_3 P_{3,t} + \lambda_4 P_{4,t} + \lambda_5 P_{5,t} + \lambda_6 P_{6,t} \\ & -\lambda_7 q_{7,t} + \lambda_8 q_{8,t} + \lambda_9 q_{9,t} + \lambda_{10} q_{10,t} + \lambda_{11} q_{11,t} + \lambda_{12} q_{12,t} \\ & +\lambda_{13} s_{13,t} + \lambda_{14} s_{14,t} + \lambda_{15} s_{15,t} + \lambda_{16} s_{16,t} + \epsilon \end{cases} \quad (15)$$

Where,

- $\lambda_1 - \lambda_{16}$ represents the coefficient in front of the linear regression first-order term.
- ϵ is the error term.
- The constant term is not considered.
- The gymnastics model is the same as volleyball.

In the above multiple linear regression multi-output model prediction, only records where both 'number of participants 4 years ago' and 'number of participants 8 years ago' are not zero are selected for subsequent research. Because: if a team did not participate 4 years ago (last edition) and 8 years ago (second last edition), we believe they are very likely not to participate in this edition, so they are not included in the consideration. Taking volleyball as an example, the Olympic Games start with 16 strong teams (depending on the year, the specific participating countries and regions may be around 16). Countries that did not even enter the top 16 are not included in the consideration.

Based on the previous dataset, by fitting and calculating the data of the US women's volleyball team, the following results were obtained.

Based on the above analysis, it can be observed that the actual number of silver medals obtained is significantly higher than the predicted value without the 'Great Coaches' effect. The 'Great Coaches' effect increased the number of silver medals by 265.66%. This indicates that excellent coaches play a significant positive role in improving the number of silver medals obtained. The 'Great Coaches' effect may have enhanced the athletes' performance through methods such as developing scientific training plans, devising tactics based on opponents' characteristics, and improving the athletes' psychological resilience, which led to better performances in competitions. Therefore, based on the above analysis, we will quantify the content of the analysis in the upcoming discussion of the 'Great Coaches' effect.

6.2. Quantitative Indicator Analysis

Considering that the impact of coaches on a team is difficult to quantify, we choose to use the subjective judgment-based AHP (Analytic Hierarchy Process) method[4]. Through expert consultation and by combining the training conditions of Olympic athletes, we conclude that when 'Great Coaches' are present in a team, they can influence areas such as improving athletes' confidence, increasing training intensity, and

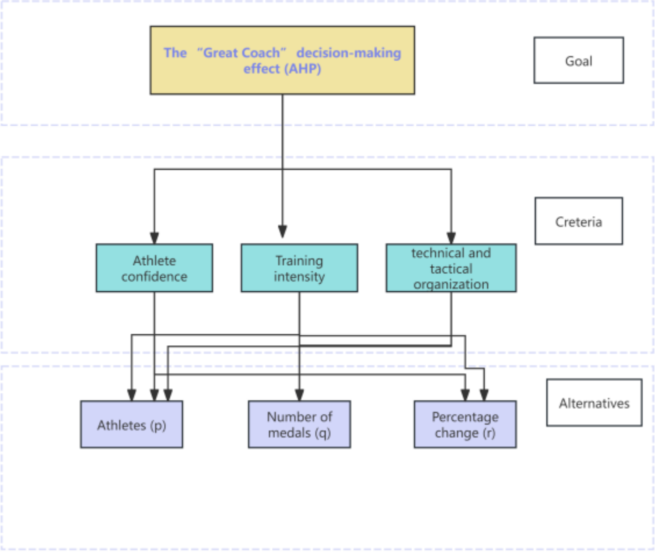

Figure 15 | The process of establishing the AHP model

providing more effective tactical arrangements in competitions. These three indicators are used with weighted importance as our features, namely p, q, and s. Therefore, we will analyze the interaction between the three major category indicators and the criterion layer through AHP.

When applying AHP to solve a problem, the first step is to organize the problem and construct a hierarchical structure model. These layers can be divided into three categories:

1) Goal Layer: 'Great Coaches' Effect Decision
2) Intermediate Layer: Athlete Confidence Improvement, Training Intensity, Tactical Arrangements
3) Scheme Layer: Athlete Category, Medal Count Category, and Proportional Change Category

Once the subordinate relationships of the elements are defined, a fuzzy judgment matrix can be established from top to bottom. The following table shows the scale of the nine-point method. The nine-point scale compares the importance of elements within the same layer of the hierarchy in pairs to construct the required judgment matrix.

Step.1 Construct the criterion layer judgment matrix

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & t_{12} & \cdots & t_{1n} \\ \frac{1}{t_{12}} & 1 & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{t_{12}} & \frac{1}{t_{12}} & \cdots & 1 \end{bmatrix} \qquad (16)$$

By comparing two adjacent indicators and assigning values using the scaling method, the hierarchical analysis judgment matrix can be obtained, as shown in the above formula. In the formula, R represents the hierarchical analysis judgment matrix; r_{ij}(i=j=1, 2, …, n) is the importance level of comparison between two adjacent indicators. We obtained the judgment matrix based on expert ratings, as follows:

$$R = \begin{bmatrix} 1 & \frac{1}{2} & 4 \\ 2 & 1 & 6 \\ \frac{1}{4} & \frac{1}{6} & 1 \end{bmatrix} \qquad (17)$$

Step.2 Weight of the criterion layer relative to the goal layer

$$W_i = \frac{\sum_{j=1}^{n} a_j + \frac{n}{2} - 1}{n(n-1)} \qquad (17)$$

In the formula, W^i represents the single-layer ranking weights of the elements in the criterion layer judgment matrix, which are used to obtain the combined weight vector $\mathbf{W} = (W_1, W_2, W_3, \ldots, W_n)$.In the formula, the element value in the i-th row and j-th column of the constructed judgment matrix is represented, and n is the order of each judgment matrix.

Step.3 Establish the judgment matrix for the scheme layer

Table 9 | Numerical Scale of the Nine-Point Method

Point	Definition
1	a_i and a_j are equally important.
3	a_i is slightly more important than a_j
5	a_i is significantly more important than a_j
7	a_i is much more important than a_j
9	a_i is extremely more important than a_j
1/3,1/5,1/7,1/9	Comparing a_j with a_i (reverse comparison)

$$B_1 = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1 & 2 & 3 \\ \frac{1}{2} & 1 & 2 \\ \frac{1}{3} & \frac{1}{2} & 1 \end{bmatrix} \quad B_3 = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ 3 & 1 & \frac{1}{2} \\ 5 & 2 & 1 \end{bmatrix} \tag{18}$$

In the formula, B^1, B^2, and B^3 represent the judgment matrices for each criterion layer in the scheme layer, that is, athlete information p (athlete confidence, training intensity, tactical arrangements), medal count q (athlete confidence, training intensity, tactical arrangements), and change rate s (athlete confidence, training intensity, tactical arrangements).

Step.4 Perform consistency testing for the weight vector and the combined weight vector

1) To calculate the consistency index (CI) in AHP, the formula is as follows:

$$CI = \frac{\lambda_{\text{max}} - n}{n - 1} \tag{19}$$

where:

- λmax is the maximum eigenvalue of the judgment matrix.
- n is the number of elements in the matrix.

2) Find the corresponding average random consistency index (RI).
3) Calculate the consistency ratio(CR)

$$CR = \frac{CI}{RI} \tag{20}$$

Let the pairwise comparison judgment matrix of the factors related to the B layer undergo consistency testing in the single ranking. The single ranking consistency index is denoted as CI(j) (where j = 1, ..., m), and the corresponding average random consistency index is RI(j). (CI(j), RI(j)) have been obtained during the hierarchical single ranking. Then, the total ranking random consistency ratio for layer B is calculated as:

$$CR = \frac{\sum_{j=1}^{n} CI(j)a_j}{\sum_{j=1}^{n} RI(j)a_j} \tag{21}$$

When CR < 0.1, the overall ranking results are considered consistent, and the analysis is accepted. The preprocessed data is then input into the model. First, the number of elements in the criterion and scheme layers is set, and the random consistency index (RI) data is initialized. Pairwise comparison matrices for both layers are entered, followed by consistency checks and weight calculations. The weighted vector relative to the overall goal is calculated, and consistency is verified. With a CR value of 0.0019, the consistency check passes. The weight calculations for p, q, and r are 0.5006, 0.3210, and 0.1784, respectively.Taking the previously established model as an example, multiply the corresponding weights with the indicators, which results in the following model:

$$\begin{cases} \partial_t^{\text{gold}}, \partial_t^{\text{silver}}, \partial_t^{\text{bronze}} = 0.5006 \times \left(\lambda_1 P_{1,t} + \lambda_2 P_{2,t} + \lambda_3 P_{3,t} + \lambda_4 P_{4,t} + \lambda_5 P_{5,t} + \lambda_6 P_{6,t}\right) \\ \qquad +0.3210 \times \left(\lambda_7 q_{7,t} + \lambda_8 q_{8,t} + \lambda_9 q_{9,t} + \lambda_{10} q_{10,t} + \lambda_{11} q_{11,t} + \lambda_{12} q_{12,t}\right) \\ \qquad +0.1784 \times \left(\lambda_7 q_{7,t} + \lambda_8 q_{8,t} + \lambda_9 q_{9,t} + \lambda_{10} q_{10,t} + \lambda_{11} q_{11,t} + \lambda_{12} q_{12,t}\right) + \epsilon \end{cases} \tag{22}$$

In the formula ∂_t^{gold}, $\partial_t^{\text{silver}}$ and $\partial_t^{\text{bronze}}$ represent the number of gold, silver, and bronze medals obtained after considering the 'Great Coaches' effect. The models for women's volleyball and gymnastics are the same.

6.3. Validate the 'Great Coaches' Effect for the Other Three Countries

By applying the weights (0.5006, 0.3210, 0.1784) obtained from the AHP method for the three major types of indicators, the data is validated. In the volleyball model, we select Japan, Serbia, and Brazil, and in gymnastics, we choose Romania, East Germany, and Poland. The results are shown in the table.

Looking at the table, the probability of Japan winning a gold medal without the 'Great Coaches' effect is only 0.2%, but with the 'Great Coaches' effect, it increases significantly to 17.29%. This clearly shows that the coach played a crucial positive role in helping Japan compete for the gold medal. For Serbia, the

Table 10 | Results of the validation of the coaching effect in women's volleyball

	Coaching effect	Japan	Serbia	Brazil
Gold Medal	TRUE	17.29	23.47	29.22
	FALSE	0.12	7.41	14.63
Silver Medal	TRUE	9.64	64.01	37.78
	FALSE	0.01	89.74	9.24
Bronze Medal	TRUE	73.07	12.53	33
	FALSE	99.87	2.78	76.4

Table 11 | Results of the validation of the coaching effect in women's gymnastics

	Coaching effect	Japan	Serbia	Brazil
Gold Medal	TRUE	23.02	12.09	33.46
	FALSE	6.86	0.24	28.16
Silver Medal	TRUE	25.90	17.6	35.05
	FALSE	12.48	0.97	45.79
Bronze Medal	TRUE	51.07	70.30	31.05
	FALSE	10.67	98.79	26.03

probability of winning the gold medal without the 'Great Coaches' effect is 7.48%, and with the 'Great Coaches' effect, it rises to 23.47%. This suggests that the coach had a positive impact on Serbia's chances of winning the gold medal, though the increase is smaller compared to Japan. For Brazil, without the 'Great Coaches' effect, the probability of winning the gold medal is 14.62%, and with the 'Great Coaches' effect, it increases to 29.22%. When comparing the coach effect on bronze medals, the probability of winning a bronze is relatively low, while the overall probability of winning a gold medal increases, indicating that the 'Great Coaches' effect contributes to converting bronze medals into gold, thus enhancing the team's ability to compete for gold or silver.

For East Germany, the probability of winning the gold medal without the 'Great Coaches' effect is only 0.24%, but with the 'Great Coaches' effect, it significantly increases to 12.09%. This clearly shows that the coach played a crucial positive role in helping East Germany compete for the gold medal. For Romania, the probability of winning the gold medal without the 'Great Coaches' effect is 6.86%, and with the 'Great Coaches' effect, it increases to 23.02%. This suggests that the coach had a positive impact on Romania's chances of winning the gold medal, although the increase is slightly smaller than that of Serbia. For Finland, the probability of winning the gold medal without the 'Great Coaches' effect is 28.16%, and with the 'Great Coaches' effect, it increases to 33.46%. When comparing the coach effect on bronze medals, the probability of winning a silver medal is relatively low, while the overall probability of winning a gold medal increases, indicating that the 'Great Coaches' effect is turning silver medals into gold, enhancing the team's ability to compete for gold.

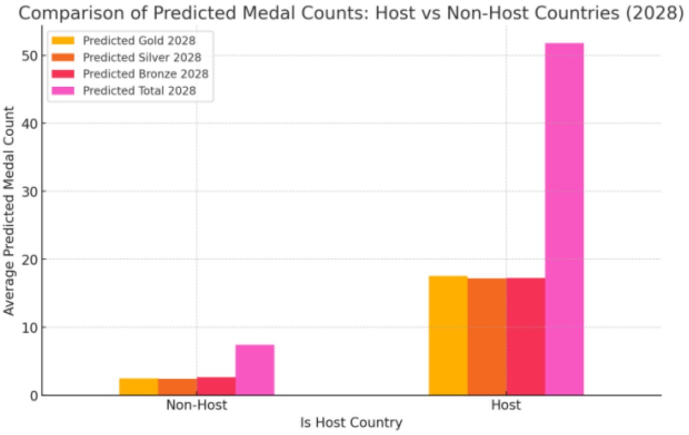

Figure 16 | Host effect impact visualization

7. Unique Insights Into Predictive Modeling

7.1. Unique Insights and Applications

Unique insights were made into many aspects of our model and the information they provide to NOCs is explained. Below are some of the unique insights:

- **Host effect**

Insight: National Olympic Committees can assess their own country's influence in international sports organizations and its economic and infrastructural readiness by looking at which countries or regions host the Olympic Games more frequently. In addition,

Table 12 | Host country impact factor

Is Host	Gold 2028	Silver 2028	Bronze 2028	Total 2028
FALSE	2.4633070149252	2.397123287671	2.6308904109589	7.44537253515577
TRUE	17.552470654824	17.18206896551	17.242758620689	51.7886832985023

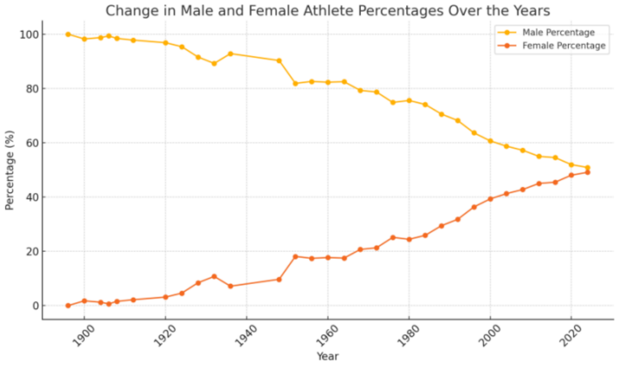

Figure 17 | Sex ratio over time

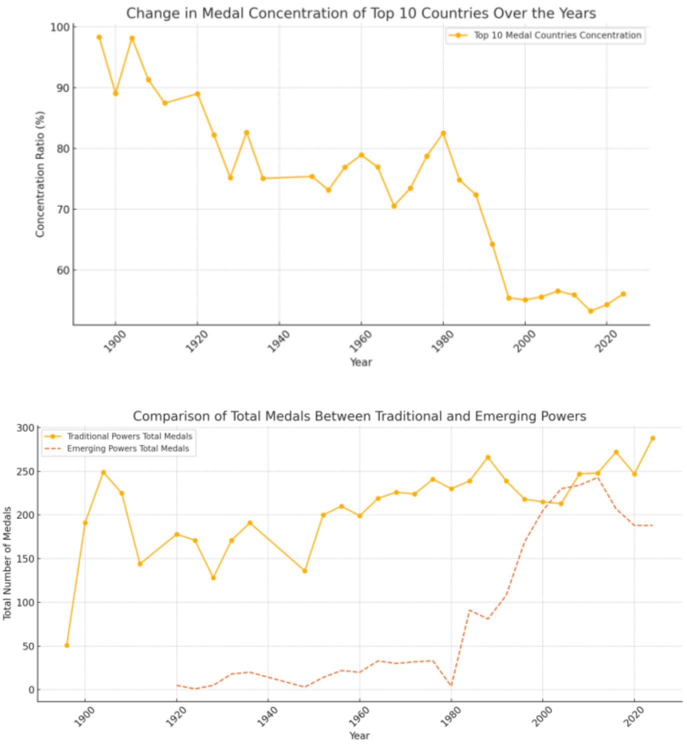

Figure 18 | Medal Trends at the Olympics for Traditional vs. Emerging Sports Powers

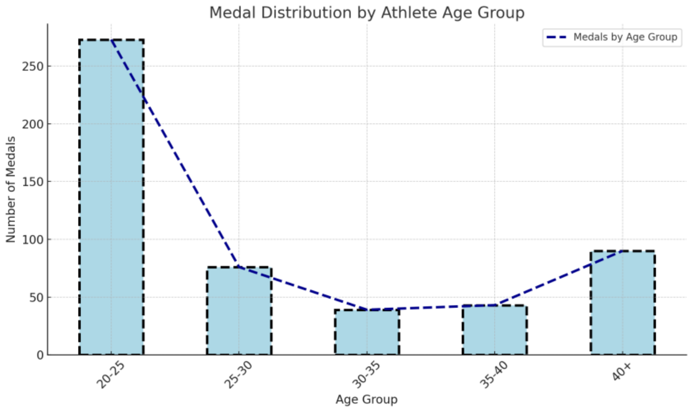

Figure 19 | Medal Distribution by Athlete Age Group

this can help predict future hosts of the Olympic Games and their impact on sport in their countries.

- **Gender balance and participation**

Insight: National Olympic Committees can follow this trend and take more measures to encourage women's participation in competitive sports, especially for those countries with a more imbalanced gender ratio. By promoting gender balance, the Olympic Committee will not only improve the country's Olympic performance, but also better reflect social values and the Olympic spirit.

- **Traditional and Emerging Sports Powers(Medal concentration)**

Insight: This provides guidance to National Olympic Committees on the dynamics of competition. While years of high concentration indicate that some traditional powers may dominate medal distribution, years of lower concentration indicate a gradual intensification of international competition and the potential for more countries to break the monopoly of the major powers. This has prompted NOCs to focus on ways to improve their position in international competition and to increase support for emerging athletes and programs.

- **Golden Age of Exercise**

Insight:Prime age group: 25-30 years old is the prime age for most athletes and many medals come from this age group.

Potential of young athletes: Athletes in the 20-25 age group also excel to a certain extent. Although they may not have the same physical strength and experience as athletes in their 30s, they have greater potential.

Balance of experience and physical strength: Athletes may lose physical strength as they age, but their experience and skill can still make up for this, especially in specific sports where experienced athletes still have the potential to win medals.

The Olympic Committee and national sports organizations may take this factor into account in the selection, training and support of athletes.

7.2. Demographic and Economic Impacts

Based on the insights above, this study also collected population and GDP data for all countries for all years to add them to the subsequent model and optimize the model. The model was allowed to train on these two new features and it was found that GDP and population also had a large impact on the forecasting results, with the model reaching a coefficient

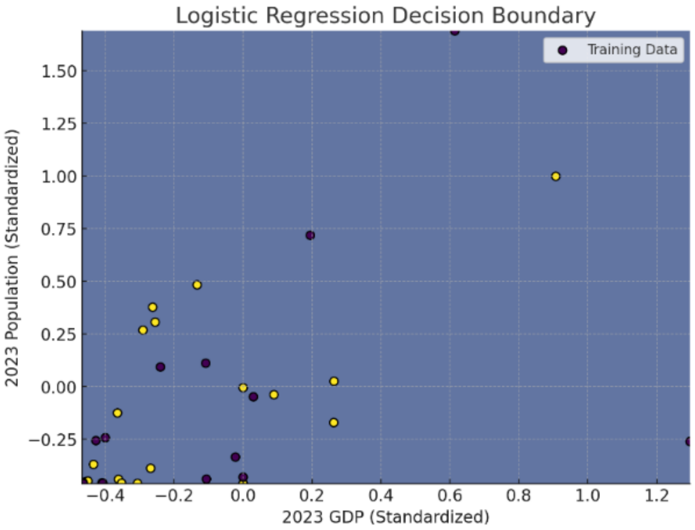

Figure 20 | GDP and population training

of determination of 0.95 after the addition of these two features.

8. Sensitivity Analysis

In this study, we conducted a sensitivity analysis of a model for predicting the number of Olympic medals with the aim of assessing the extent to which different input features affect the model output (total number of medals).

In order to take into account the interactions between the input features and to assess the global impact of the features on the model output, we used methods such as analysis of variance (ANOVA) and Sobol's index[5,6]. By calculating the contribution of each feature and its interaction to the variance of the model output, we were able to identify the features that were most important for the prediction of the total number of medals. Global sensitivity analysis takes into account the complex relationships between features and helps us to understand the model's response to different features in a more comprehensive way.

where Sobol's total sensitivity index is formulated as[7]:

$$S_{\text{total.i}} = 1 - \frac{\text{Var}(y \mid x_{\sim i})}{\text{Var}(y)} \tag{23}$$

Athletes were identified as the most important feature through the global sensitivity analysis, and their contribution to the variance of the total number of medals accounted for more than 40% of the variance. The number of medals and the number of gold medals were ranked second and third, respectively, and their effect on the total number of medals was second to that of the athletes. The host country effect and year have a weaker contribution than the previous ones in the global analysis, indicating that they have a limited contribution to the prediction, but still have a larger role in some specific contexts (e.g., home field advantage of the host country).

9. Model Evaluation

9.1. Strengths

The model fits well with reality: the model looks at the athlete's recent performance, the country's historical medal trends, and the host effect, simplifying the complex constraints while taking into account a number of important factors.

High authenticity: This model does not add too many assumptions to the original data, which ensures the authenticity and diversity of the data to better tap into the deeper features.

Algorithmic superiority: The XGboost model used in this paper is highly accurate, robust, capable of handling complex nonlinear relationships, and maintains high generalization ability on complex training data by controlling hyperparameters such as tree depth, number of leaf nodes, and learning rate.

Practical and stable results: Through visual analysis and quantitative forecasting, the model not only provides an accurate estimate of the number of medals but also offers considerable advice to national Olympic committees.

9.2. Weaknesses

The scope of application is limited due to the fact that the prediction is constructed on Olympic medals and has not been validated for other aspects or different events.

The cultural, political context of the actual situation also has an influence on it, as no external information has been introduced.

10.Conclusion

This paper improves the traditional regression forecasting model to adapt to a more complex Olympic environment.

- **Results 1**

We made predictions based on the XGboost model and then sent these results to the decision model. We successfully predicted the medal table for 2028, including every every type of medal and its confidence interval, with a coefficient of determination of 0.94 and very factual. To prove that our model is the best model, we compared it with some other strategies. In the second subquestion we get that 51 countries will win their first medal in the next Olympics with a probability of 62.5% and a certainty of 89%. Finally we identified the relationship between many events and the number of medals, and verified the strengths of many countries and the contribution of these to the total medals.

- **Results 2**

The analysis of the 'great coach' effect showed that it had a significant impact on medal outcomes. For example, Japan's probability of winning a gold medal increased from 0.2% without the 'Great Coach' effect to 17.29% with the 'Great Coach' effect. Serbia's probability of winning a gold medal increased from 7.48% to 23.47% and Brazil's from 14.62% to 29.22%. These results show that good coaching plays a crucial role in increasing the probability of a gold medal.

- **Results 3**

For the conclusion of the Last Task, the insights of the model on the Olympic Games include host effect, gender balance and participation, medal concentration, and the change of traditional and emerging sports powers, etc. Due to the space limitation, only four insights are shown and relevant suggestions are given. Meanwhile, in order to provide unique insights, we plan to add population and GDP features to improve the accuracy of the model, and after re-training the model, we successfully improve the coefficient of determination of the optimization model to 0.95.

References

1. J. Moolchandani, V. Chole, S. Sahu, R. Kumar, A. Shukla and A. Kumar, "Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics," 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2024, pp. 1987-1992.
2. Cook, Gillian M., David Fletcher, and Christopher Carroll. 2020. “Psychosocial Functioning of Olympic Coaches and Its Perceived Effect on Athlete Performance: A Systematic Review.” International Review of Sport and Exercise Psychology 14 (1): 278–311.
3. A. M. Priyatno and T. Widiyaningtyas, “A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS”, jitk, vol. 9, no. 2, pp. 196–207, Feb. 2024.
4. Olson, D.L. (1996). The Analytic Hierarchy Process. In: Decision Aids for Selection Problems. Springer Series in Operations Research. Springer, New York, NY..
5. Sobol, I.M. (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. MATH COMPUT SIMULAT,55(1–3),271-280.
6. Sobol’, I. (1990). Sensitivity estimates for nonlinear mathematical models. Matematicheskoe Modelirovanie 2, 112–118. in Russian, translated in English in Sobol’, I. (1993). Sensitivity analysis for non-linear mathematical models. Mathematical Modeling & Computational Experiment (Engl. Transl.), 1993, 1, 407–414.
7. Homma, T. and A. Saltelli (1996). Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering and System Safety, 52, 1–17.