Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2518035

Science on the Sports Field: Olympic Medal Prediction and Strategic Analysis

Summary

Against the background of intensified global sports competition, Olympic medal prediction and sports resource allocation have become the key concerns of national sports committees. In this paper, indicators such as the number of athletes, the number of event participants and other indicators are defined and modeled to provide a basis for predicting the distribution of Olympic medals as well as optimizing sports resources.

The provided dataset was first thoroughly preprocessed and feature engineered. After extracting the relevant eigenvalues, we applied **Kendall correlation analysis** and **principal component analysis** to reveal the intrinsic structure of the data. Then, the principal components were analyzed by **ARIMA** model for prediction. This prediction result was further trained by an integrated learning model with **Random Forest** and **Support Vector Machine**. The model **R^2** was **0.8908**, which was a good fit. We predicted the **top ten** in the **medal table** for the 2028 Olympics. The resultant predictions show that **Germany and South Korea** will **improve** their results, while the **Netherlands'** results are likely to **decrease**. In addition, based on the above model, a predictive analysis of countries likely to win medals for the first time was conducted. The results showed that five countries were expected to win their first Olympic medals. Then, the effect of program settings on medals was explored, with the value of the sport's effect on the number of medals being around 4%. It was found that **Swimming, Swimming, and Athletics** were the three most important aspects of medal winning for the **USA, China, and Great Britain**, respectively. Subsequently, the correlation between featured programs and increased medals was analyzed, and it was found that the results of the information gain between featured programs and increased number of medals were all **>0.5**, which can be considered as a necessity for countries to select and develop featured programs.

Subsequently, here, we conducted an analysis using **Ridge Regression**. The results show that great coaches can significantly improve a country's performance at the Olympics by at least **1.69 bronze medals**. In particular, **Mexico's diving, South Africa's swimming, and Brazil's gymnastics programs** are ideal for investing in great coaching resources. Simulated projections using data from 2024 show that great coaches could bring an average increase of **1.47 gold medals** to these three countries.

In addition, this study analyzes the **historical trends** in the number of medals won by the **United States** and explores the reasons behind them. Meanwhile, the actual impact of the **host effect** is verified. Further, this study explores **the effect of outstanding athletes**, taking Michael Phelps as an example, and finds that there is a significant correlation between the presence of top athletes and the number of medals. The study suggests that the great athlete effect does have a significant impact on a country's medal tally.

Finally, this study tested the sensitivity of the model by **implementing ten-fold cross-validation** and **adding noise disturbances**, and the results showed that the model has good robustness and stability.

Keyword：ARIMA； SVM ；RF ；Stacking ；Shap analysis；Ridge Regression

Contents

1 Introduction

1.1 Problem Background

As a long-standing and large-scale sporting event, the Olympic Games have consistently attracted global attention. Each edition captivates audiences worldwide with its unique appeal, and predicting Olympic medal counts and rankings has long been a subject of interest due to its complexity and unpredictability.

This study develops a forecasting model based on multiple influencing factors to predict Olympic medal counts and rankings. By analyzing key variables' impact on medal distribution, the study challenges the traditional view of randomness in Olympic results. With the advent of big data and artificial intelligence, scientifically grounded predictions are now feasible, offering a more accurate understanding of medal dynamics.

The findings deepen our understanding of the competitive landscape of the Olympics and provide a theoretical foundation for improving national sports policies, resource allocation, and training strategies. This research holds significant academic and practical value.

1.2 Restatement of the Problem

Based on a thorough analysis of the problem, we have identified the following key issues to address:

- Develop a forecasting model to predict the number of gold medals and total medals each country will win at the Olympic Games.
- Predict the medal standings for the 2028 Summer Olympics in Los Angeles, providing prediction intervals for all results.
- Identify which countries are likely to show improvement in the next Olympics and which may perform worse than in 2024, by analyzing historical data and key performance indicators.
- Predict which countries are likely to win their first Olympic medal in the upcoming Games, and calculate the confidence intervals for these predictions.
- Analyze the relationship between the number of medals won and the sports events chosen by each country. Identify which sports are most crucial for each country and explore the reasons behind this.
- Examine the impact of the host country's selection of events on overall medal outcomes and how this influences participating countries' results.
- Investigate the influence of exceptional coaches on medal performance, selecting three countries to identify which sports would benefit most from hiring outstanding coaches.
- Provide unique insights into Olympic medal predictions, revealing underlying patterns and factors that affect medal counts.

1.3 Our Work

Figure.1 shows the process of our work:

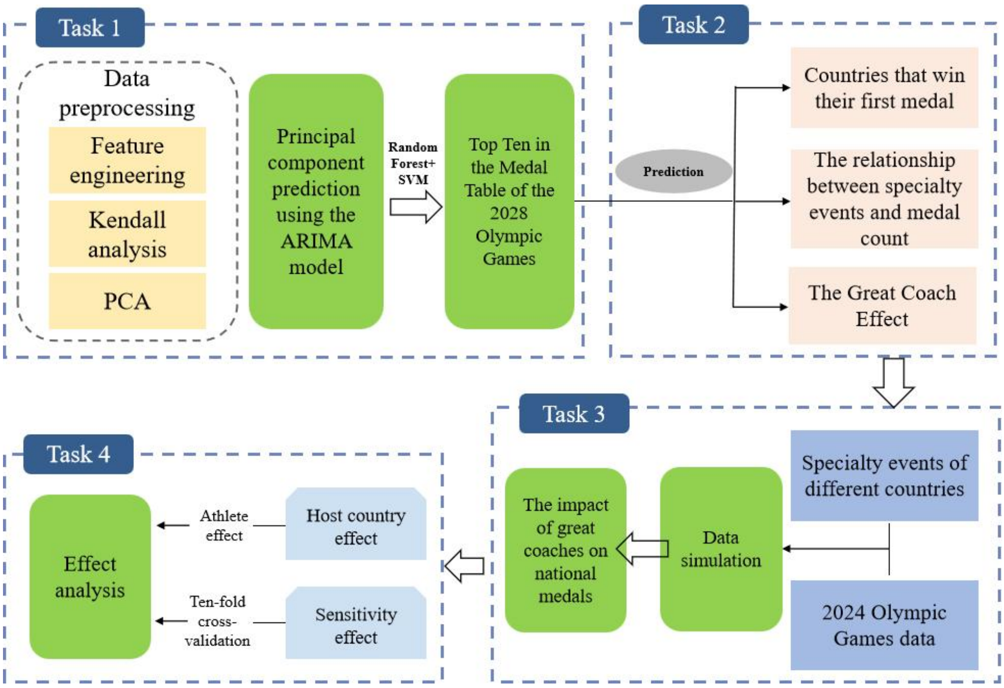

Figure 1: The process of Our Work

2 Assumptions

Due to the numerous factors influencing the number of medals won at the Olympic Games, in order to ensure the accuracy of the study, the following assumptions have been made for this research:

- Athletes perform at their normal level, without exceptional performances or major mistakes.
- The competition is fair, with no bias or favoritism from referees towards any athlete.
- Athletes do not engage in doping or other forms of cheating.
- Athlete performance is not significantly affected by external factors such as weather or the presence of spectators.
- Each country prepares for the events based on historical experience, with no sudden major changes.
- No country will suddenly produce an athlete with exceptional talent capable of changing the overall outcome of the competition.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
Gold	Number of gold medals
Silver	Number of silver medals
Bronze	Number of bronze medals
Total Medals	Number of medals

***Note:** Here is a partial notation.Detailed explanations for each symbol can be found in the corresponding text.

4 Data Preprocessing and Feature Engineering

4.1 Data Mapping

The model construction requires extracting the gold medal count and total medal count for each country from the provided files for past Olympic Games, as well as the number of athletes from each country participating in the Olympics. The two files represent the concept of a country using both the full name (e.g., China) and the IOC country code (e.g., CHN). To ensure data consistency, we referred to the 'List of IOC Country Codes' and created a mapping table to map each country accordingly.

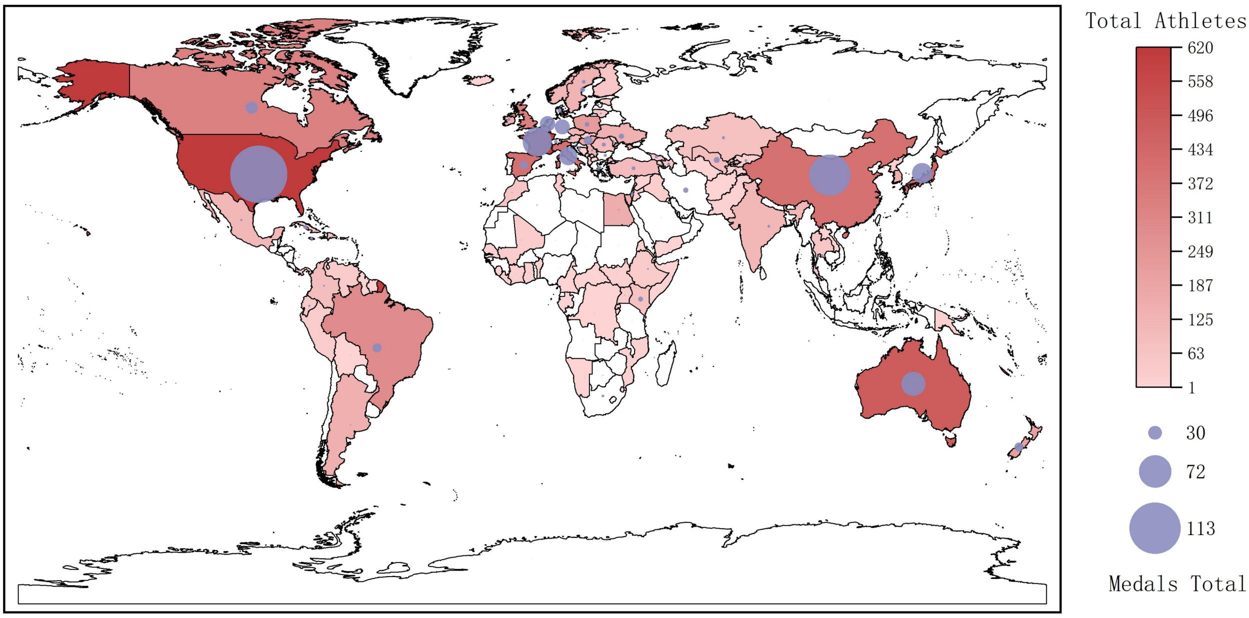

Figure 2: 2024 Olympics:Athlete Numbers by Country & Medal Counts

4.2 Outlier handling

The model building requires extracting the number of gold medals and total medals for each country from the *summerOly_medal_counts.csv* file, and the number of athletes representing each country in the Olympics from the *summerOly_athletes.csv* file. In these two files, the concept of a country is represented using either the full country name (e.g., China) or the IOC country code (e.g., CHN). To ensure data consistency, we referred to the List of IOC Country Codes and created a mapping table, mapping each country to its corresponding IOC code. For countries that no longer exist (e.g., East Germany and West Germany), we preserved their historical IOC codes to maintain data integrity.

4.3 Missing value handling

After performing data filtering and statistics on the *summerOly_programs.cs*v file, we found that some cells contained missing values. These missing values represented the number of events in certain sports for specific years. Using the functions in the table, we compared these cells with the Total events column and concluded that the data in these cells should be 0. Therefore, we filled the missing values in the table by assigning them a value of 0.

4.4 Feature Engineering

To evaluate the trends in medal acquisition by various countries, we have developed a comprehensive evaluation system. This system consists of three main aspects: the country level, the athlete level, and the event impact. Each primary indicator is further divided into several secondary indicators, as shown in the figure below.

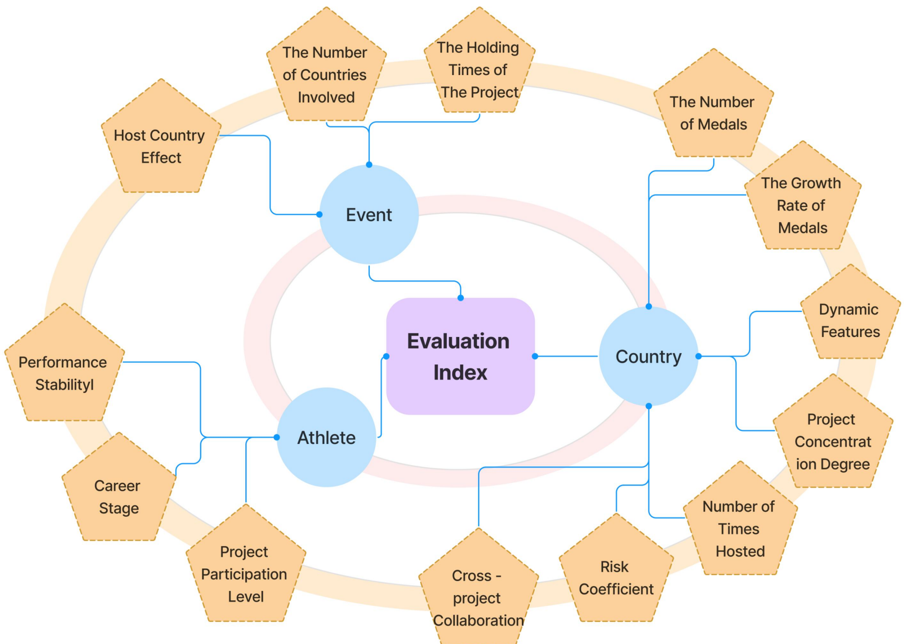

Figure 3: Evaluation Index

We categorize and organize the data into three levels. The following provides detailed descriptions of all indicators and their corresponding formulas.

Table 2: Metrics and Formulas for Feature Engineering

	Indicator	formula set		Indicator	formula set
Country	Medal count	$T_n = \sum_{k=1}^{m} \left(g_k - s_k + b_k\right)$		Cross-competition program synergies	$S_{XY} = \frac{Cov\left(X,Y\right)}{\sigma_X \sigma_Y}$
	Medal growth rate	$v_n = \frac{T_n^t - T_n^{t-1}}{T_n^{t-1}}$	Athletes	Project participation	$D_a = -\sum_{j=1}^{m} \left(\frac{P_j}{P_a}\right) \ln\left(\frac{P_j}{P_a}\right)$
	Accelerated medal growth	$a_n = \frac{v_n^t - v_n^{t-1}}{\Delta t}$		Career stage	$L_a(t) = \begin{cases} Rookie\ phase & if\ \ t \le T_1 \\ Peak\ phase & if\ \ T_1 < t \le T_2 \\ Decline\ phase & if\ \ t > T_2 \end{cases}$
	Dynamic Characterization Changes	$\Phi_n = \int \frac{d^2 T}{dt^2} dt + \alpha \cdot \Delta T$		Performance Stability	$W_a = 1 - \frac{\sigma\left(s\right)}{\mu\left(s\right)}$
	Number of Olympic Games hosted	$H_n = \sum_{t=1}^{T} \delta_n^t$	Event	Number of times the competition program	$F_i = \sum_{t=1}^{T} h_i^t$
	Competition Program Concentration	$C_n = \sum_{i=1}^{m} \left(\frac{p_i}{p_n}\right)^2$		Number of participating countries	$N_i = \left\|\left\{n \middle\| \exists a \in n \text{ participates in Program } i\right\}\right\|$
	Risk factor	$R_n = \frac{\sigma\left(p\right)}{\mu\left(p\right)}$		Host country effect	$M_n = e^{\beta \cdot H_n}$

- **Medal count**: T_n represents the total number of medals. g_k, s_k, and b_k represent gold, silver, and bronze medals
- **Medal growth rate**: v_n represents the growth rate of the number of medals, and a_n represents the rate of change in the growth rate.
- **Dynamic Characterization Changes**: ΔT represents the acceleration of the number of medals, and α represents the trend weight coefficient.
- **Number of Olympic Games hosted**: H_n represents the number of times the country has hosted the Olympic Games.
- **Competition Program Concentration**: p_i represents the number of medals won in event i, and p_n represents the total number of medals.
- **Risk factor**: $\sigma(p)$ represents the standard deviation of the number of medals, and $\mu(p)$ represents the mean of the number of medals.
- **Cross-competition program synergies**: X and Y represent the time series of the number of medals.
- **Project participation**: P_j represents the number of times event j has been participated in, and P_a represents the total number of participations.
- **Career stage**: t represents the number of times an athlete has participated in the Olympics, and L_a represents the career stage label.
- **Performance Stability**: $\sigma(s)$ represents the standard deviation of the performance, and $\mu(s)$ represents the mean of the performance.
- **Number of times the competition program was held**: F_i represents the total number of times event i has been held.
- **Number of participating countries**: N_i represents the total number of countries that have registered for event i. i represents the competition event.
- **Host country effect**: H_n indicates whether country n is the host country.

5 Question 1: Modeling the number of gold medals as well as the number of medals

The prediction of the medal rankings is typically not based solely on historical medal counts, but needs to take into account multiple factors, which often affect the actual number of medals a country ultimately wins. Therefore, in the previous section, we developed feature engineering to assist in building the prediction model.

To present our work more intuitively, we have created a flowchart of the model construction process. Below is the display of the flowchart.

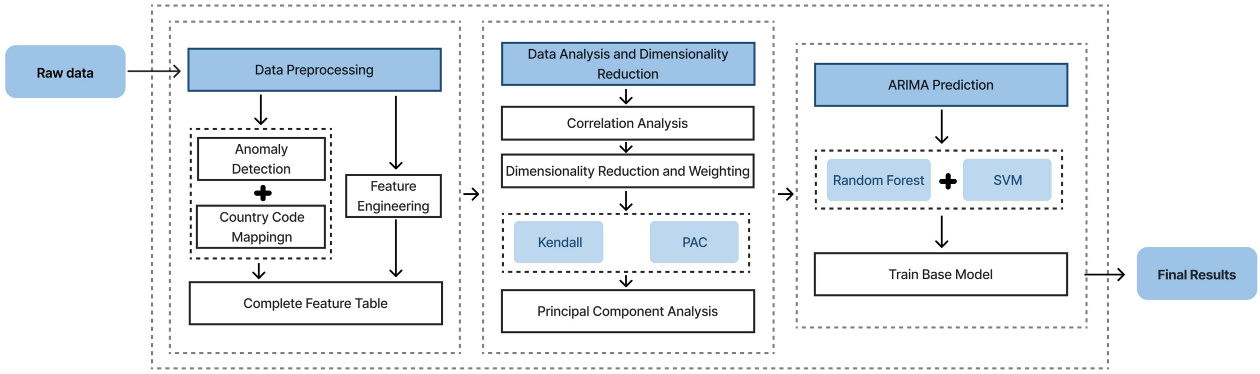

Figure 4: Flowchart for model construction

5.1 Construction of ARIMA-Random Forest-SVM models

This study constructs a prediction model by improving the traditional TOPSIS method. To address its shortcomings, such as subjective weighting and the neglect of indicator correlations, we innovatively introduce the Kendall correlation coefficient to analyze the relationship between features and objectives. Additionally, we combine Principal Component Analysis (PCA) for dimensionality reduction to determine the indicator loadings and compute the principal component scores. This hybrid method objectively quantifies the impact of features and effectively selects variables that significantly influence the number of medals/gold medals.

Based on data from the top ten countries prior to 2024, we establish a two-stage prediction framework: first, the ARIMA model is used to predict the principal component scores for 2028, and then a stacked ensemble model (Random Forest + SVM meta-model) is built for the final prediction. This framework combines the ability to capture time series evolution with the advantages of nonlinear and heterogeneous feature integration. Through the collaboration of both models, it significantly enhances data adaptability, enabling the deep exploration of the complex mapping relationship between principal components and medal counts, thus ensuring prediction accuracy.

5.1.1 Kendall correlation analysis

This study uses the Kendall correlation coefficient to assess the relationship between features and objectives. This non-parametric test calculates the correlation coefficient by evaluating the probability of coordinated variation between variables, offering advantages such as distribution-free properties and strong robustness to outliers. It is particularly suitable for capturing nonlinear monotonic relationships and analyzing ordinal features. By ranking the correlations between all features and the two target variables (number of gold medals and total medal count), we complete the selection of key influencing factors. The calculation process is as follows: first, each feature value is paired with the two sets of target variables to form observation pairs. For each set of observations, comparisons are made with all other observations to identify the following three scenarios:

Table 3: different scenarios

Circumstances	Condition
Consistency	$X_i > X_j$ *and* $Y_i > Y_j$ or $X_i < X_j$ *and* $Y_i < Y_j$
Inconsistency	$X_i > X_j$ *and* $Y_i < Y_j$ or $X_i < X_j$ *and* $Y_i > Y_j$

Difference	$X_i = X_j \ and \ Y_i = Y_j$

Finally, based on the number of consistencies and inconsistencies, Kendall's coefficient is calculated as follows：

$$\tau = \frac{C-D}{\sqrt{(C+D+T_1)-(C+D+T_2)}} \tag{1}$$

The next step is to explain the parameters involved in the above equation. In the formula C is the number of concordant pairs in the observations of the eigenvalues. D is the number of discordant pairs in the observations of the eigenvalues. T_1 and T_2 are the average number of pairs present in the respective observations. The denominator in the formula is used to normalize the number of pairs of possible ranks in the data to make it more stable. Hypothesis testing and significance analysis were performed by verifying whether Kendall's coefficient was significant or not.

Calculate the standard error：

$$SE_\tau = \sqrt{\frac{4(C+D)}{n(n-1)}} \tag{2}$$

Calculate t-value：

$$t = \frac{\tau}{SE_\tau} \tag{3}$$

Based on the results of Kendall's analysis, we get the following figure：

Figure 5: Correlation of Principal Eigenvalues and Scatter Plots

As illustrated in Figure, the findings indicate a robust correlation between the total number of athletes participating, the three primary categories of award-winning programs, the efficiency of winning awards, and other characteristics of the values with the ultimate number of medals and gold medals. This suggests a strong correlation between the selected characteristics of the values and the target variable.

5.1.2 Principal component analysis

In order to perform important feature extraction as well as dimensionality reduction of the eigenvalues, we have taken the approach of **Principal Component Analysis (PCA)**.Principal Component Analysis is a dimensionality reduction technique that reduces the variability of the data by transforming the original data to a new coordinate system.PCA helps to identify patterns in the eigenvalues as well as important feature mining.The following steps are taken to perform PCA principal components:

Initially, the data must undergo normalization. In this study, **Z-Score normalization** was employed, a technique that eliminates the effect of scale by converting each eigenvalue to a mean of 0 and a standard deviation of 1. This, in turn, ensures the consistency of the weights of individual features.The following formula is used to calculate Z-Score normalization:

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

In the formula, X denotes the original data (i.e., an observation of a feature), μ is the mean of the feature, and σ is the standard deviation of the feature.

Subsequently, the normalized data is employed to calculate the covariance matrix. The resulting covariance matrix can then be utilized to reflect the correlation between dimensions, thereby facilitating comprehension of the relationship between disparate features. The specific formula is as follows:

$$C = \frac{1}{n-1} X^T X \tag{5}$$

In the formula, C denotes the covariance matrix and X^T is the transpose of the data matrix.

Subsequently, the rationalization of the covariance matrix's eigenvalues and eigenvectors is imperative for the identification of the principal components. The eigenvectors signify the predominant directions of the data, while the eigenvalues denote the significance of each direction. The subsequent step involves the implementation of an eigen decomposition of the covariance matrix, which results in the identification of the eigenvalues, denoted by I , and the eigenvectors, denoted by V .The selection of the size of the eigenvalues is then performed to achieve a ranking of the eigenvectors. The eigenvectors that correspond to the largest eigenvalues are then selected as the new axes, with the directions of the eigenvectors aligning with the principal components of the data.

Finally, the normalized data and the selected principal component eigenvectors are multiplied to obtain the dimensionality-reduced data. The formula for the transformed data is as follows:

$$Z = X \cdot V \tag{6}$$

where V is the matrix containing the principal component eigenvectors. Finally we get the following result:

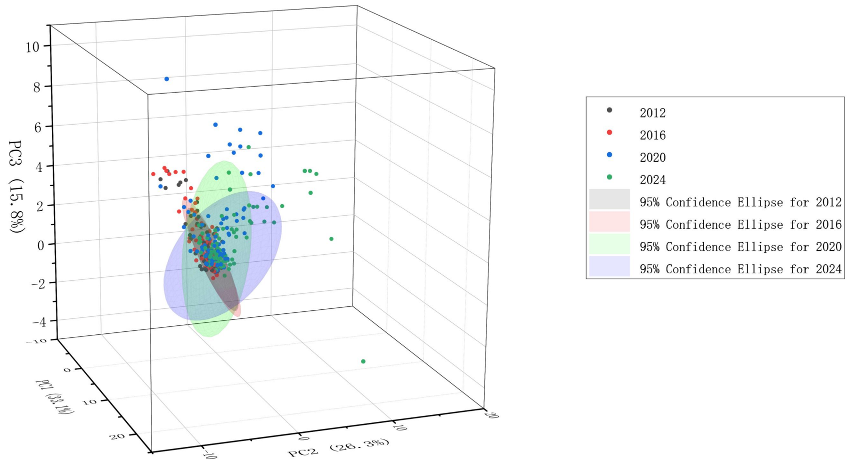

Figure 6: Principal component analysis results

The findings of the concluding analysis are presented in Figure 1. The uppermost three principal components (PC1, PC2, PC3) from 2012 to 2024 are displayed. The distribution of principal components is evident in the figure, resulting in a more concentrated pattern. The dimensionality reduction effect is discernible, as evidenced by the final 53 principal component values and their corresponding loadings. This analysis suggests that PCA is a more effective method.

5.1.3 ARIMA Model

In this section, we propose a forecasting method for the principal components of the ten countries by 2028. This forecasting method is based on time series analysis, and we have selected the **ARIMA model** for this purpose.The ARIMA model is a type of statistical model known as the Differential Integrated Moving Average Autoregressive Model. The selection of this model is predicated on its capacity to address non-stationary time series, its aptitude for comprehensive time series characterization, and its relative interpretability.[1] The subsequent steps in the solution process are outlined as follows.

Initially, the processed principal components must undergo testing for smoothness and differencing using the ADF test. This is done to ascertain whether the data necessitates differencing. Subsequent to this determination, the difference operation is performed until the sequence attains smoothness. To determine the order of the difference, the following formula is employed:

First difference：

$$\Delta y_t = y_t - y_{t-1} \tag{7}$$

Second difference：

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \tag{8}$$

Subsequently, the orders of AR and MA must be determined. This is followed by parameter estimation via the maximum likelihood method, as well as model fitting for residual test operations. Finally, model prediction is performed. The specific formulas are as follows:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{9}$$

Finally, temporal cross-validation of the model was conducted using data from

2020-2024, resulting in the following findings: a range of 6.8 ± 8.2 for the MAE of total medals, 12.2% for the MAPE of gold medals, and 0.4 ± 0.38 for the MAE of the principal components.The MAE is thus concluded to be less than 8, the MAPE is less than 15, and the MAE is less than 0.5%. It can thus be considered that the error of this model is acceptable and that it is a model with a good prediction effect.

5.1.4 Random Forest + SVM Modeling

In light of the aforementioned findings, the predicted values of the principal components for the ten countries by 2028 have been obtained. A combination of Support Vector Machines (SVMs) and Random Forests (RFFs) has been selected to construct a meta-model employing the stacking method. This method integrates the predictions of the two base models, SVMs and RFFs, to yield the final prediction outcomes. The following is the model construction process.

Initially, the **SVM model** was constructed. The SVM is a potent supervised learning algorithm extensively employed for classification and regression tasks. The objective of the SVM is to classify or fit (regress) different classes of data by finding an optimal hyperplane and maximizing the interval. [2] The subsequent subsections delineate the specific workflow of the SVM model.

Initially, the dataset must be collected, which contains feature vectors $X = \{X_1, X_2, \ldots, X_N\}$ and target vectors $Y = \{Y_1, Y_2, \ldots, Y_N\}$. For Problem 1, the feature vector is the data PC1-PC53 processed by principal component analysis, and the target vector is the number of gold medals and the total number of medals.

Subsequent to this, feature normalization is performed to scale the features to the same scale, and the formula is：

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

In this equation, μ denotes the mean and σ denotes the standard deviation.

The data set is then divided into a training set (X, Y) and a test set (X, Y).The training set is used for model training, and the regularization parameter C, the kernel function type, and the kernel function parameters are set.The appropriate kernel function $K(X_i, Y_j)$ is selected according to the data characteristics.

For the regression task：

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{11}$$

restrictive condition：

$$\begin{gathered} y_i - (w \cdot x_i + b) \le \varepsilon + \xi_i \\ (w \cdot x_i + b) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{gathered} \tag{12}$$

The variable C denotes the spacer bandwidth.

The prediction of the test set using the trained model is given by the following formula:

$$\hat{y} = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x_{text}) + b \tag{13}$$

The Random Forest algorithm utilizes a Bagging integration framework, employing the construction of multiple decision trees to achieve prediction through the self-sampling method. The fundamental mechanism underlying this process involves the generation of K training subsets through the process of putative back sampling. In this procedure, each subset is trained independently on the base decision tree model. Subsequently, the prediction results of each tree are integrated through the averaging method. The method exhibits dual overfitting characteristics (row sampling + column sampling), which can effectively capture nonlinear feature interactions and improve model stability through variance reduction. This renders it particularly suitable for regression prediction tasks in high-dimensional feature space. In comparison with a single decision tree, the integration mechanism of the proposed method exhibits a substantial enhancement in noise robustness and generalization ability.

The ensuing prediction by Random Forest is the average of all tree predictions.

$$\hat{y}_{RF}(x) = \frac{1}{N}\sum_{j=1}^{N} T_j(x) \tag{14}$$

where $\hat{y}_{RF}(x)$ is the prediction of the random forest for the input x. $T_j(x)$ is the predicted value of the j tree. N is the number of decision trees.

Previously, we have identified two base models, Support Vector Machine and Random Forest, and we choose to take the stacking method for subsequent processing. The stacking method is an integrated learning method, by using the prediction results of multiple base models as inputs and training a meta-model to generate the final prediction results, the use of Support Vector Machines and Random Forest models combined with the stacking method is a powerful integrated learning method, Random Forests are suitable for dealing with high-dimensional and non-linear data, and SVM is suitable for dealing with small-scale and non-linear data, and the stacking method is capable of combining the advantages of the two and improving the prediction performance.

Based on the advantages and disadvantages of various models, we choose the simpler linear regression model, which can be computed quickly when the data volume is large and has strong interpretability. The following is the specific process:

The stacked model will train the base model first, and then use the prediction results of the base model to train the meta-model. The prediction formula is as follows:

$$\hat{y}_{final} = w_1\hat{y}_{RF} + w_2\hat{y}_{SVM} + b \tag{15}$$

where w_1 and w_2 are the weights of the base model predictions. b is the bias term.

In order to ensure the accuracy of the predicted values, we added a confidence interval to the predicted values and calculated a confidence interval of 95% for each predicted value. We performed a visualization to get the following figure.

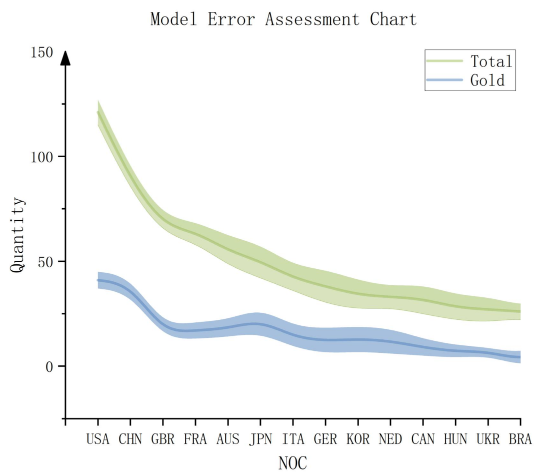

Figure 7: Predicted Valus and Confidence Intervals by Country

The final results are shown in the figure, we get the prediction errors of some countries, from the final results, the confidence interval of the medal count is in a suitable range, and it can be considered that the model's prediction results are better

For the total medal count prediction model, **the MSE is 16.5853 and the RMSE is 6.0486**, which means that the average error between the predicted value and the real value is about 4.0486, and judging from the actual range of the total medal count variable, this result is relatively reasonable and a relatively small error; the MAE is 2.3203, which means that the average absolute error between the predicted and real values is about 2.3203; R 2.3203; **R^2 is 0.8908**, which is a high value, indicating that the model has a strong ability to predict the total number of medals.

We have done visualization of the above to visualize the result.

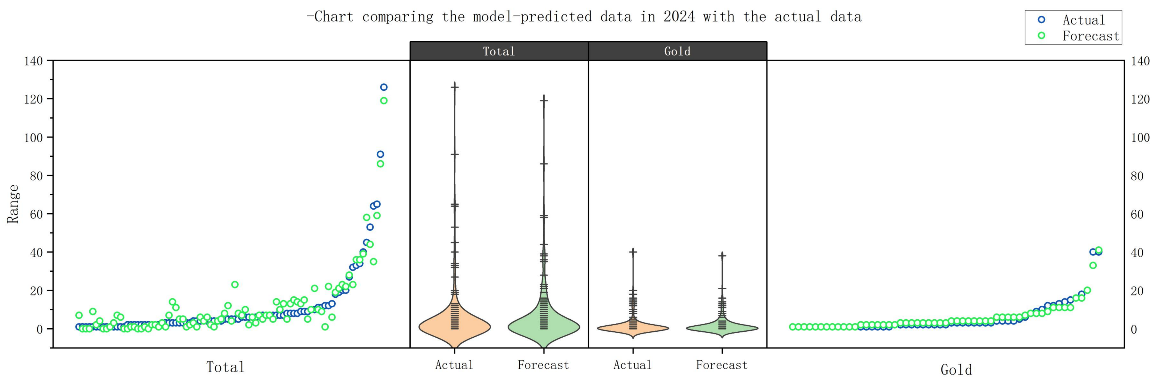

Figure 8: Model fitting results

The final results are shown in Fig. It can be seen that the prediction results of the model are denser, the model fit is better and the error is relatively small. For the prediction model of the number of gold medals, the MSE is 5.7518 and the RMSE is 2.3983, which indicates that the average error between the predicted value and the real value is about 2.3983, and judging from the actual range of the variable of the number of gold medals, the result is relatively reasonable and a relatively small error; the MAE is 0.9347, which indicates that the average absolute error between the predicted value and the real value is about 2.3203; R^2 is 0.8634, which is a high value, indicating that the model has a strong ability to predict the number of gold medals.

In general, this model has a better prediction effect on medals and gold medals, and can predict the number of medals and gold medals in 2028 more effectively.

5.2 Model prediction results

5.2.1 Projected results of the Los Angeles Summer Olympics medal table

The predicted values for each nation were obtained based on the medal prediction model constructed from the aforementioned requirements. In order to better evaluate the performance of each nation, the top ten nations in the ranking were selected, and the predicted results of the medal table of the 2028 Summer Olympics in Los Angeles, USA, were analyzed in comparison with the performance of each nation in 2024.

NOC	Country	Gold	Total	NOC	Country	Gold	Total
United States		43	131	Japan		22	50
China		39	89	Italy		14	42
Great Britain		16	67	Germany		12	38
France		17	64	South Korea		13	34
Australia		18	55	Netherlands		12	33

Figure 9: 2028 Predictions for Gold and Total

The medal prediction model we constructed yielded predicted values for each country, and the final predicted medal table is shown in the figure.

5.2.2 Progress or decline

The United States, China, Great Britain, France, Australia, Japan, Italy, England, Germany and South Korea still maintain their power as major sporting nations and are firmly in the top ten. However, the number of medals won by the United States and China has declined, possibly due to the increasing level of overall athleticism around the world, resulting in a more competitive Olympics and the possibility that they may have suffered setbacks in their respective traditionally dominant sports.

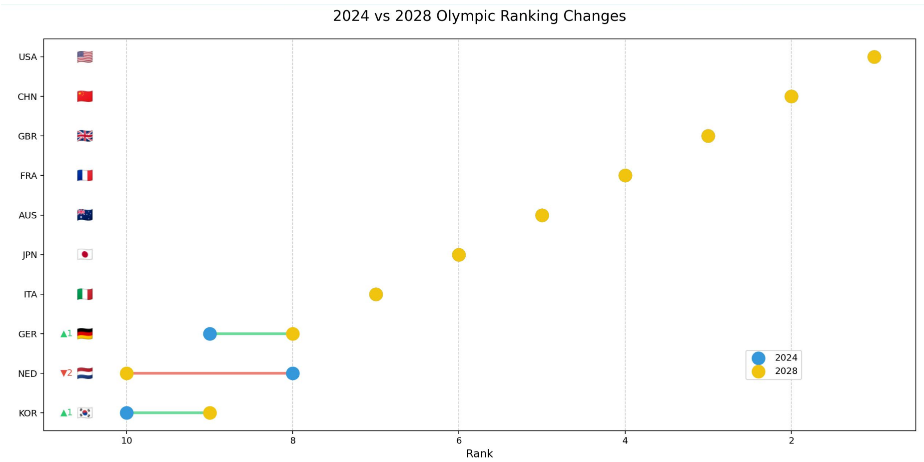

Figure 10: Change in ranking of major countries

The number of medals won by the United Kingdom, France, Australia, Japan, Italy, the Netherlands, Germany and other countries is on the rise, which is very importantly related to the continuous improvement of the level of sports competition in the world. The projected number of medals won by countries in the world in the year 2028 shows that the number of medals won by countries with a medium level of sports competition is on the rise, whereas the number of medals won by countries that used to be the top countries in terms of sports competition is declining, although they still manage to maintain their position. The number of medals won by the former great sports powers, while maintaining their position, is declining.

We can visualize through the above chart that for the top ranking countries, there is no big change in the medal table ranking. The top seven countries are USA, China (CHN), Great Britain (GBR), France (FRA), Australia (AUS), Japan (JPN) and Italy (ITA). We predict that **Germany** and **South Korea** will improve their Olympic results, finishing in eighth and tenth place in the medal table respectively. The **Netherlands** will slip in their Olympic results, slipping to ninth place in the medal table.

5.2.3 Zero breakthrough in the number of Olympic medals

In the first question of problem one we have constructed the ARIMA-Random Forest-SVM model. For the second question asked, we modified the parameters and changed the output to whether the award was won in 2028 and the probability of winning the award. And we use four metrics totaling Accuracy, Precision, Recall and F1-Score for evaluation. Here are the formulas for each metric.

Table 4: Indicator & formulas

Indicator	Formula	Indicator	Formula
accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	recall rate	$\text{Recall} = \frac{TP}{TP + FN}$
precision	$\text{Precision} = \frac{TP}{TP + FP}$	F1-Score	$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

In the above equation, TP(True Positives) denotes the number of samples in the training set that were predicted to win the prize and actually did so.TN(True Negatives) denotes the number of samples in the training set that were predicted to fail to win the prize and actually did not win the prize.FP(False Posi-tives) denotes the number of samples in the training set that were predicted to win the prize but did not actually win the prize.FN (False Negatives) denotes the number of samples in the training set that were predicted not to win the prize but actually did.

By calculation, we obtain the following results.

Table 5: Results of indicator calculations

	Accuracy	Precision	Recall	F1-Score
value	0.8915	0.8679	0.9427	0.8459

The metrics show that the overall prediction correctness of the model is **89.15%**, which is a good performance; **86.79%** of the samples predicted to be in the positive category are actual correct samples, with a low false positive rate; it is able to capture **94.27%** of the actual correct samples, with a very low underreporting rate; and the balance between precision and recall is good, with an excellent overall performance.

Based on the parameters of the above four metrics, we conducted a prediction on the question of how many countries will win medals for the first time in the next Olympic Games. The prediction results are as follows.

Table 6: Projected Probability of Winning First Time Chart

Country	Probability
Aruba(ARU)	0.3868
Angola(ANC)	0.3304
El Salvador(ESA)	0.1969
Maldives(MDV)	0.1668
Mali(MLI)	0.1513

The results of the model predictions show that Aruba and the Republic of Angola have the highest probability of winning a medal for the first time in 2028. However, this is also relative, most of the strong countries have already won medals many times, and the remaining countries are relatively weak, do not have advantageous programs, and have limited training conditions in their countries, making it difficult to win gold medals, so the probability is lower. For these countries, the emergence of an exceptionally talented athlete will directly determine whether or not they can win the award. And this factor is more complicated, we will not consider it.

5.2.4 Impact of program settings on medals

In this section, we explore the impact of the Olympic program setup on the distribution and number of medals for each country. Typically, as the number of events increases, the total number of medals rises accordingly, and emerging events may change the pattern of medal distribution. For traditionally strong countries (e.g., the United States), they may dominate in the added events, while other countries may lose their original medals. In addition, the type of programs hosted can also have an impact on medal distribution. These analyses can help countries make more appropriate decisions in future Olympic Games. In order to study the impact of changes in the number and type of sports on medals, we used **Shap analysis** and used **chi-square tests** to verify the impact of specific sports on some countries.

Shap is a machine learning method based on game theory that quantifies the contribution of each feature to the prediction result of a single model. Based on our previously established ARIMA-Random Forest-SVM model, we use Shap for our analysis.Shap not only has the ability to explain globally and locally, but also supports complex models and high-dimensional interactions. It can effectively explain the relationship between the predicted number of medals and gold medals and the input eigenvalues, thus revealing the impact of changes in the number and type of sports on countries and the importance of specific sports for certain countries.

For Shap analysis, the computational process can be divided into the following key steps:

First we need to define the baseline value, here we have chosen the average of the eigenvalues defined in the first question as the baseline value. The formula is as follows：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!}[f(S \cup \{i\}) - f(S)] \tag{16}$$

The following is an explanation of the notation in formulas. $S \subseteq F \setminus \{i\}$ denotes an arbitrary subset that does not contain feature i . $|S|!$ and $(M-|S|-1)!$ denote the weighting factors of the permutations that ensure that the subsets are fairly weighted. $f(S)$ denotes the result of the prediction using only the features in subset S . The result of the prediction using only the features in subset S is shown in the table below.

Next, all feature subsets are traversed, and for each subset S , the prediction difference after feature i is added is computed. The expression is：

$$\Delta_i(S) = f(S \cup \{i\}) - f(S) \tag{17}$$

Subsequently, the contributions of all subsets are averaged and a weighted summation of S is performed with the summation formula：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!}\Delta_i(S) \tag{18}$$

Finally, the numerical results are obtained. In order to make the representation of the data more intuitive, we carried out the visualization of the data. A graph below is obtained.

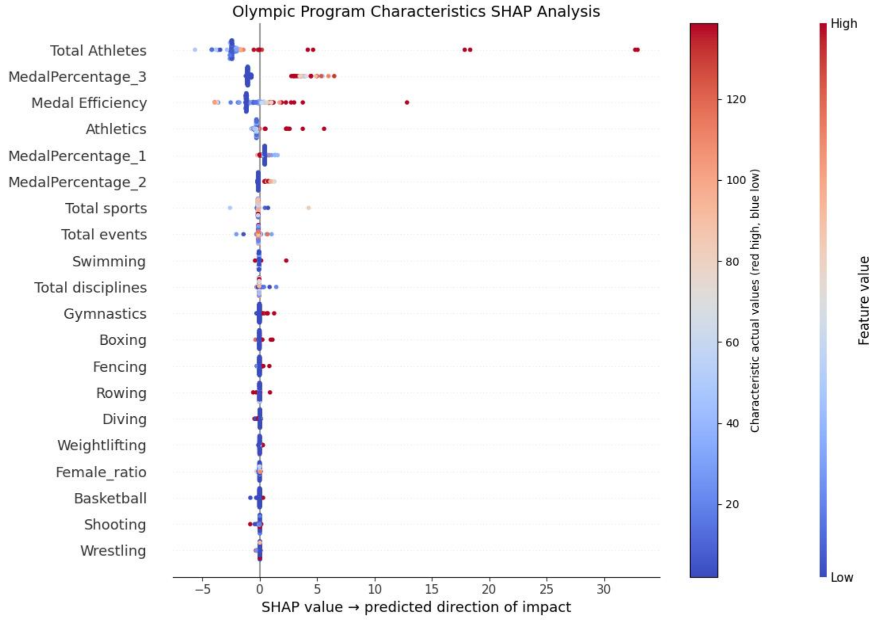

Figure 11: Olympic Program Characteristics SHAP Analysis(1)

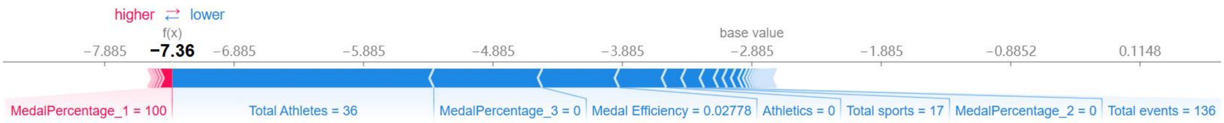

Figure 12: Olympic Program Characteristics SHAP Analysis(2)

The final results are shown in Fig.11. From Fig. 12, Total sports, Total events and Total discipline are in the top ten in terms of changes in the number of medals and gold medals, so it can be assumed that an increase in the number and type of sports has a certain influence on the increase in the number of medals and gold medals. From Figure 12, the number and types of sports are distributed on the right side of the chart, and it can be assumed that increasing the number of sports tends to increase the number of medals and gold medals won by the country.

5.2.5 Information Entropy Increase

In this section, we analyzed the most important sports in each country. The data preprocessing resulted in the top three most awarded sports in each country, which were considered to have a significant impact on the outcome of the competition. Next, we performed an information gain analysis by collecting the changes in the number of sports participated in each country and the changes in the total number of medals, proving that the selection of specific sports by certain countries can significantly increase the number of medals.

Due to the large number of countries, three countries in the 2024 Olympics were selected for analysis in this study: the United States, China, and the United Kingdom. For the United States, the top three medal-winning sports are Swimming, Athletics, and Volleyball, with 21.50%, 18.07%, and 8.10% of the medals respectively. For China, Swimming, Hockey and Diving were the three most awarded sports with 21.43%, 10.12% and 8.93% respectively. For the UK, Athletics, Rowing and Cycling Track were the top three, accounting for 25.64%, 23.08% and 15.38% respectively. This shows that for the US, China and the UK, Swimming, Swimming and Athletics are the most important sports respectively.

Subsequently, we used information gain to measure the relationship between the number of specific participants and the total number of medals. Information gain is a measure of the reduction in the amount of information a feature imparts to a dataset, which quantifies the importance of a sport to a country's medal performance. By measuring the “information contribution” of a sport's features to the medal prediction, the information gain reflects the correlation between the sport and the number of medals. The next section describes the procedure for calculating the information gain.

First, we need to determine the information entropy, which indicates the degree of confusion in the medal sorting data set, and is calculated as follows:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{19}$$

The following is an explanation of the symbols in the formula. S denotes the data set. c denotes the number of categories (e.g., $c = 2$ in the second question). p_i denotes the proportion of samples in category i in the data set.

We need to select the significant number of entries indicator as the basis for categorization.

Subsequently, we need to calculate the conditional entropy, which is the weighted average entropy of each subset of participants after the value is divided by the number of participants using the classification basis. Finally, the original entropy is subtracted from the conditional entropy to get the information gain, the expression is as follows:

$$IG(S,\ A) = H(S) - H(S \mid A) \tag{20}$$

We ended up with the following information gain graph based on the results, using Team USA's gymnastics program as an example.

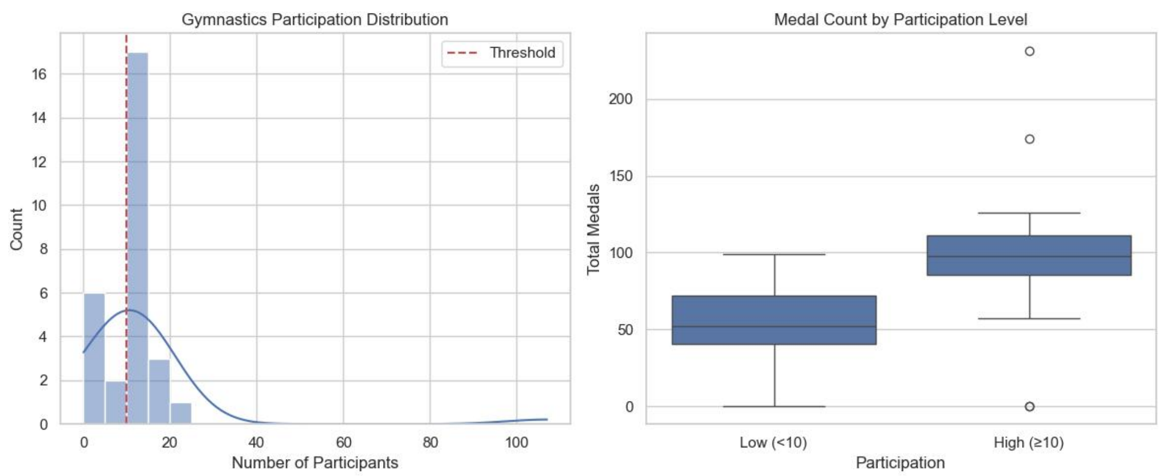

Figure 13: U.S. Team Information Gain

We calculated the information gain of Gymnastics in the US, Volleyball in China, and Rowing in the UK, and the results were **0.5215, 0.6234**, and **0.6023,**respectively, and we can think that these programs have a medium degree of correlation with the final medals won. Therefore, it can be concluded that each country selects and actively participates in its own representative and characteristic programs can effectively contribute to the increase in the number of medals.

6 Question 2: Analyzing the "great coach" effect

6.1 Possibility of a "great coach" effect

In the Olympic arena, an athlete's performance is often closely linked to the influence of his or her coach. Great coaches can help athletes enhance their performance in all aspects such as technique, tactics and psychology, which directly affects the number of medals. Therefore, studying the effect of "great coaches" on the number of medals can help us identify the key role of coaches in training and competition.

Considering the high correlation of feature data, the coefficients estimated by ordinary least squares method have great variance and the model is unstable. Therefore, we used ridge regression analysis in linear regression to compress the coefficients by adding the L2 regularization term to improve the generalization ability of the model, and then evaluated the role of the coach in the variation of the number of medals.[3]

linear regression equation:

$$y=\theta_1 x_1+\theta_2 x_2+\cdots+\theta_{n-1} x_{n-1}+\theta_n x_n \tag{21}$$

L2 regularization term:

$$\lambda \sum_{j=1}^{p} \beta_j^2 \tag{22}$$

The overall formula is:

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right\} \tag{23}$$

Firstly, the medal contribution needs to be defined, i.e. the change in medals after the arrival of a great coach, after which the weights of the medals were assigned, assigning the weights of gold, silver, and bronze medals as 3, 2, and 1 respectively for those who did not win the medals as 0.

Next, the parameters involved in the above formula are explained. In the formula, y_i represents the medal contribution, β_j represents the regression coefficient of the j th feature, β_0 is the intercept term, which represents the initial medal contribution, and X_{ij} represents the j th feature value of the i th sample, which includes the number of athletes, the efficiency of winning awards and so on. λ is the strength parameter of regularization, which controls the compression of the coefficient.

Find the derivative of the objective function and make the derivative zero:

$$-2x_{ij}^T(y_i-x_{ij}\beta)+2\lambda\beta=0 \tag{24}$$

The solution of the coefficient is:

$$\beta_{\text{ridge}}=(x_{ij}^T x_{ij}+\lambda I)^{-1}x_{ij}^T y_i \tag{25}$$

λI is the unit matrix.

Finally, the regression coefficient θ of a great coach is 1.6856. From the results, it can be considered that a great coach can bring at least one bronze medal and above to a country, and then the effect of "great coach" is considered to exist.

In addition, we conducted an error analysis and model evaluation of this model. The R^2

of the model is **0.781**, and the adjusted value is **0.775,** which proves that the model fits well and there is no overfitting. The **p-value** of the item Great Coach is approximately equal to **0**, which is very significant for the results. Therefore, it can be concluded that the "great coach" effect exists.

6.1 Recommendations for investing in "Great Coach"

In this section, we need to choose the three most important countries as well as sports that should consider investing in great coaches and advise them. We chose the following three countries and programs:

Mexico - Diving:Mexican athletes have excellent flexibility and aerial perception, but the coefficient of difficulty of the movement has long been lower than that of Chinese and British athletes. There is already a certain foundation, but needs a more advanced training system.

South Africa - swimming: South Africa has a long coastline and suitable climate, but swimming competitive performance has long lagged behind. The success of local Chad Le Clos (2012 Olympic gold medal) proves the potential, but the successor is weak.

Brazil - Gymnastics:Brazil has achieved good results in gymnastics, but there is a gap in talent and a lack of innovative choreography, which could improve the training system to enhance the difficulty and stability of the movements. In addition, there are many individual gymnastics events with intensive medal opportunities, so Brazil can focus on breakthroughs to realize the growth of the number of medals.

In the simulation prediction of the intervention of great coaches in these three programs of the above three countries, the medal weights of these three countries have increased from **3**, **5** and **12** to **22.7252, 21.7438** and **42.3077** respectively, and each great coach can bring **1.47 gold medals** to these three countries on average. So it is reasonable to think that the above countries should choose to invest in great coaches in specific programs.

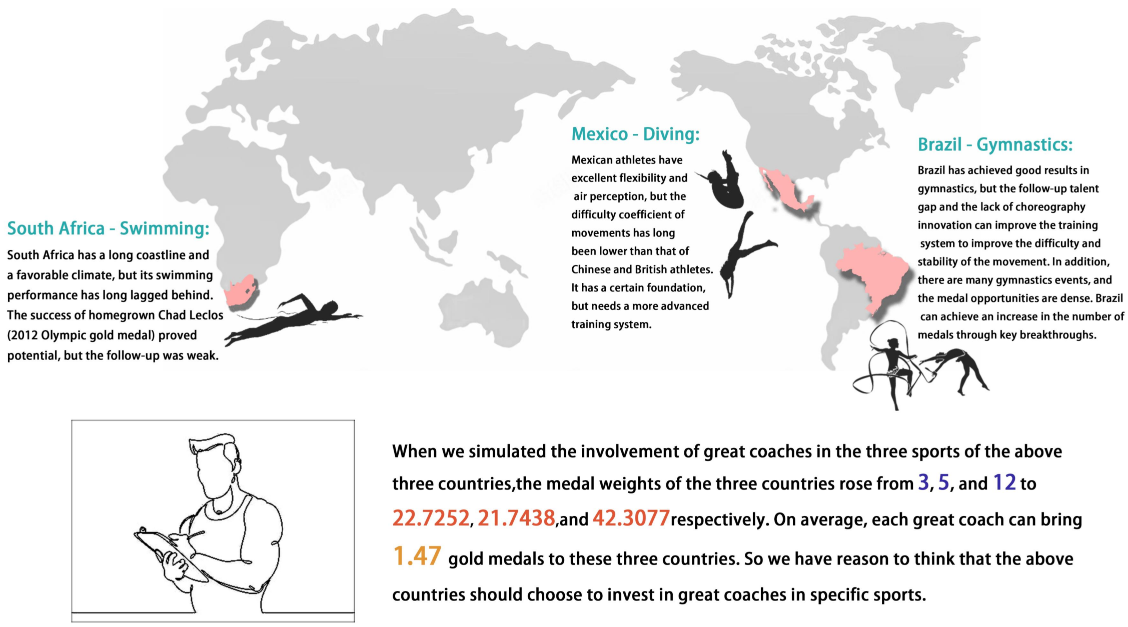

Figure 14: Advice on investing in great coaches

We have simulated the intervention of great coaches in these three programs of the above three countries, and the medal weights of the three countries have increased from 3,5,12 to 22.7252,21.7438,42.3077 respectively.On average, each great coach can bring 1.47

gold medals to these three countries. So we have reason to believe that the above countries should choose to invest in great coaches in specific programs.

7 Question 3: Unique insights on the number of Olympic medals

7.1 Analysis of historical trends in traditional sports powers

In this section, we need to analyze unique insights about the number of medals won in the Olympics, and we have selected three traditional sports powerhouses - the United States, China, and the United Kingdom - to analyze, revealing some interesting changes by analyzing historical information about the number of medals won by these three countries.

First, we visualized the historical trend changes for the United States as follows:

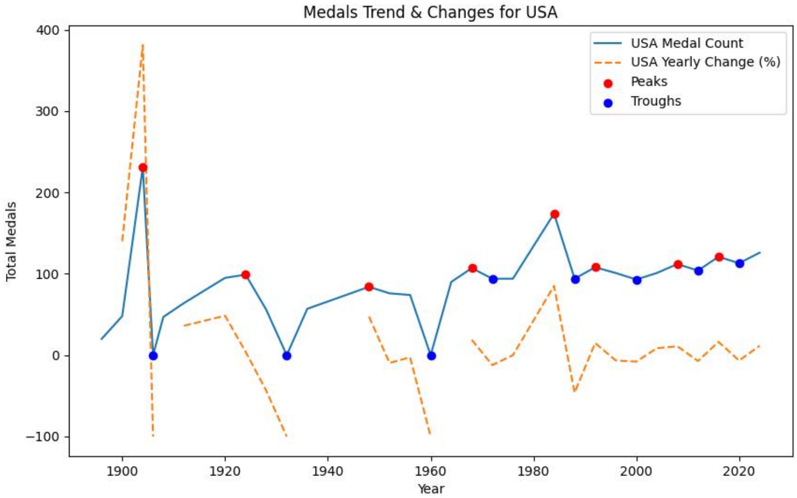

Figure 15: Medals Trend & Changes for USA

There are a couple of more critical time points in terms of visualizing the results.The first is the dramatic rise and fall in medal counts in the 1900s. Analyzing from a historical perspective, the 1900s saw a dramatic increase in the number of medals as the U.S. team performed well at the Paris Olympics, the first time the U.S. team participated in the modern Olympics, and both the size of the field and the number of athletes grew. This period was the early stage of industrialization, the United States gradually rose to economic prominence, and sporting events gradually gained social importance, which was also related to the enhancement of the concept of sports and the improvement of training methods for players.There were large fluctuations in the number of U.S. medals from the 1910s to the early 1920s, especially during World War I. The U.S. participation in the war during this period affected all areas of society in the Second World War, and many athletes enlisted in the military, which led to the competition participation decreased and the wartime economy was depressed and sporting events were greatly affected. Secondly, from the 1990s to the 2000s, the number of U.S. medals steadily increased and remained at a high level. During this period, the U.S. made it possible for athletes to improve their performance through scientific training methods. As sports became an important part of society, the training mechanism of athletes became more perfect, and the U.S. sports career entered a golden age. International competition was also an important reason to urge the U.S. to progress, and countries such as China as well as Russia gradually became strong competitors.

7.1 Host effect

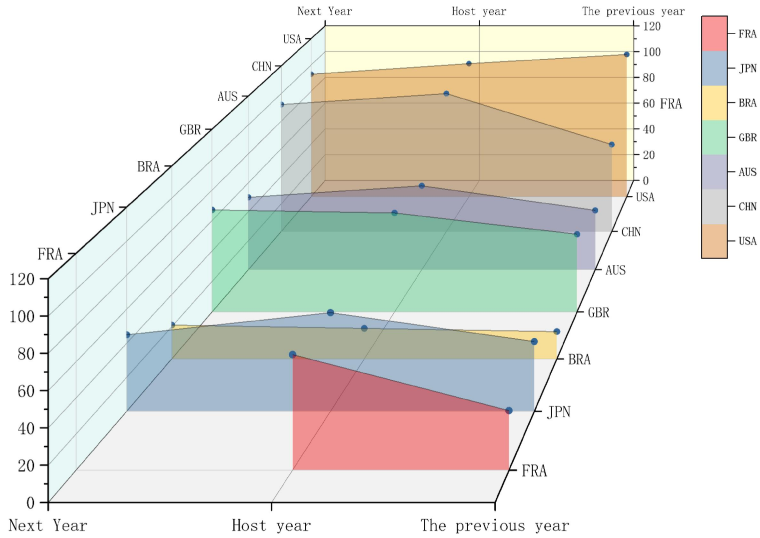

Figure 16: Host effect

Host countries can significantly affect the distribution of medals through program adjustments. According to the above figure, it can be clearly seen that the number of medals of host countries is basically more than the achievements in recent years, and we can consider that the effect of host benefits on the number of awards is significant. This study provides a theoretical basis for research targeting the host effect and suggests that the important influence of tournament rules on medal distribution is further highlighted by adjusting strategic programs.

In order to further expand the advantage at the 2028 Olympic Games in Los Angeles, the following section proposes a series of options for the U.S. to maximize its chances of winning medals in terms of program setting and strategic layout.

1. Optimize the competition system: Adjust the competition rules or system to make it more in line with the characteristics of American athletes.

2. Utilizing home field advantages: for example, taking advantage of stable ocean currents and wind conditions to hold surfing and sailing competitions.

3. Layout in advance: After identifying new programs, invest resources in training reserve talents several years in advance to ensure competitiveness when the event is held.

8 Sensitivity Analysis

In this section, we use ten-fold cross-validation with noise perturbation to analyze the sensitivity of the model for Problem 1 Problem 2. Ten-fold cross-validation is a statistical method used in machine learning to evaluate the performance of a model by estimating more reliably the generalization ability of the model over a limited data set. Noise perturbation can test the robustness of the model and help to detect the presence of overfitting.

We first performed ten-fold cross-validation on the ARIMA-Random Forest-SVM model constructed in Problem 1, and finally obtained the following results:

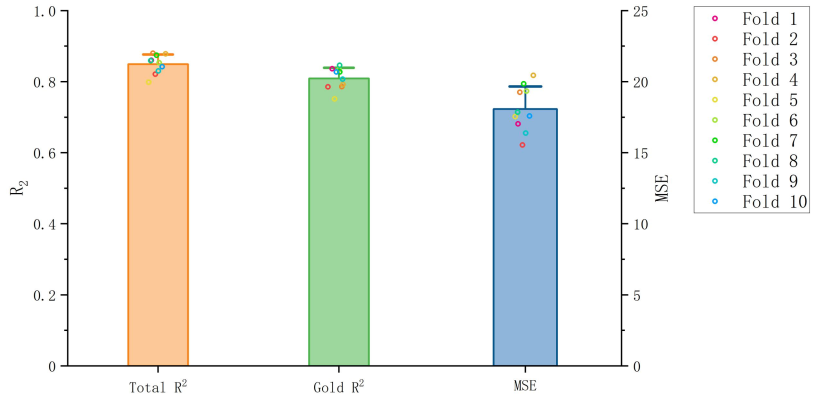

Figure 17: Result of the error assessment

By analyzing the final results, we can find that after several iterations of training, the R^2 of the two models is still stable at about 0.8, and the value of MSE is also stable, we can consider that the models have good robustness as well as a certain degree of goodness of fit.

Subsequently, we carried out the same validation of the linear regression model constructed in Problem 2 and obtained the following results:

Table 7: Model Calculation Results

	Original model coefficients	Model coefficients after adding noise
Const	0.7810	0.7724
Coach	1.0899	1.0887
Number of people	0.0097	0.0097
Medal Efficiency	-1.5968	-1.6041
Total Medals	0.0032	0.0033

From the final results, the difference in coefficients between the original model and the model with noise perturbation is minimal, indicating that the model is not sensitive to small data disturbances and has high robustness.

9 Model Evaluation and Further Discussion

9.1 Strengths

- In the feature engineering selection process, we used diverse evaluation features, which laid a solid foundation for the precise identification of influencing factors in the subsequent analysis.
- Before constructing the prediction models, we performed Kendall correlation analysis and PCA analysis. These methods helped us analyze the correlations between all the features and reduce dimensionality, which facilitated the accurate identification of feature contributions in the next steps.
- For predicting the number of medals and gold medals, we employed a combination of ARIMA, SVM, and Random Forest models, followed by a weighted averaging of the results. This enhanced the model's generalization ability and enabled precise selection of classification features.
- In the process of validating the great coach effect, we used a technique of inserting ridge regression analysis into linear regression to improve the model's generalization ability, thus enabling a more accurate assessment of the coach's role in medal variations.
- During the sensitivity analysis, we applied 10-fold cross-validation and noise perturbation methods, then calculated the changes in R-squared and MSE. This approach

made our results more reliable.

9.2 Weaknesses

- In future research, we need to increase the training iterations of the combined SVM and Random Forest model to find more optimal parameters, thereby improving the model's performance.
- When constructing the feature engineering, the scope we considered was too narrow. In reality, there are deeper influences on the number of medals and gold medals, which are far more complex than our selected features. For example, factors such as national strength and economic power have a significant impact. Therefore, we should consider using a more comprehensive model to describe these influences.

10 Conclusion

In this paper, we applied different models to analyze the distribution of medals. Overall, we reached the following conclusions:

In the section on predicting the number of medals and gold medals, we forecasted that the top ten countries on the medal table will be **USA, CHN, GBR, FRA, AUS, JPN, ITA, GER, KOR, and NED**. Additionally, we concluded that **Germany and South Korea** are expected to **improve** their performance, while **Netherland**'s Olympic results are likely to **decline**.

In the section on predicting which countries will win their first medal, we identified four countries that are likely to earn their first medal: ARU, ANC, ESA, MDV, MLI, with probabilities of **0.3868, 0.3304, 0.1969, 0.1668, and 0.1513,** respectively.

In exploring the impact of event selection on medal outcomes, we concluded that an increase in the number and variety of sports events has a **certain influence** on the overall number of medals, particularly gold medals. Furthermore, we believe that countries selecting and actively participating in sports that are representative and characteristic of their national identity can effectively **increase their medal counts**.

In the section on the potential influence of great coaches, our findings suggest that the role of a great coach could at least lead to an improvement of one bronze medal or higher for a country. We also believe that **Mexico's diving program, South Africa's swimming program, and Brazil's gymnastics program** are areas where investment in great coaches would be particularly beneficial.

In the section on unique insights, we analyzed the **trend of the USA's medal count** throughout history and provided explanations for these trends. We also validated the **host country effect**. Additionally, we explored the **effect of great athletes**, taking Michael Phelps as an example. We concluded that the great athlete effect exists and can influence the overall change in a country's medal count.

References

[1] Liang Z ,Ismail T M .Advanced CEEMD hybrid model for VIX forecasting: optimized decision trees and ARIMA integration[J].Evolutionary Intelligence,2024,18(1):12-12.

[2] Antoine G ,Martin K ,Gaurav B , et al.Comparison of random forest and support vector machine regression models for forecasting road accidents[J].Scientific African,2023,21

[3] VarinC ,FirthD .Tractable Ridge Regression for Paired Comparisons[J].Stat,2024,13(4):e70028-e70028.

Report on Use of AL

- **OpenAI ChatGPT(GPT4o)**

Query1:<What factors influence the number of gold medals and medals won by countries at the Olympics?>

Output:<Several factors influence the number of gold medals and overall medals won by countries at the Olympics:

1. **Economic Resources**: Wealthier countries often have better funding for training, facilities, and athlete support.

2. **Sports Infrastructure**: Countries with advanced sports programs and elite training centers tend to produce more top-tier athletes.

3. **Talent Pool**: Larger populations generally have more athletes to choose from, increasing the chances of winning medals.

4. **Historical Performance**: Countries with a history of success often build strong sports cultures and have more experienced coaching staffs.

5. **Government Support**: Strong government investment in sports programs can directly impact performance.

6. **Focus on Certain Sports**: Some countries excel in specific sports, often due to cultural emphasis or specialized training programs.

7. **Coaching and Technolog

These factors combined help determine a country's Olympic medal tally.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<Please tell me the advantages of arima model>

Output:<The ARIMA (AutoRegressive Integrated Moving Average) model has several advantages:

1. **Simplicity**: It’s easy to understand and implement for time series forecasting.

2. **Flexibility**: ARIMA can handle various types of time series data, including those with trends and seasonality (with additional modifications like SARIMA).

3. **Stationarity Handling**: The model can transform non-stationary data into stationary data through differencing.

4. **Forecasting**: ARIMA provides good short-term forecasts for stationary or nearly stationary data.

5. **Model Interpretability**: The components (AR, I, MA) are interpretable and provide insight into the underlying data patterns.

These advantages make ARIMA a popular choice for time series analysis and forecasting.>

Query2:<How to adjust parameters to optimize the prediction of arima models>

Output:<To optimize ARIMA model parameters (p, d, q), follow these steps:

1. **Stationarity Check**: Ensure the data is stationary (use tests like ADF test). If not, apply differencing (adjust d) until stationarity is achieved.

2. **Autocorrelation Analysis**: Use the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to identify suitable values for p (AR order) and q (MA order).

   - Look for significant lags in the PACF plot to set p.
   - Look for significant lags in the ACF plot to set q.

3. **Grid Search**: Try different combinations of p, d, and q values and evaluate model performance using criteria like AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

4. **Model Diagnostics**: Check residuals for randomness (using Ljung-Box test) to ensure no further patterns are left in the data.

5. **Iterative Refinement**: Adjust parameters based on model diagnostics and forecast accuracy.

Optimizing ARIMA involves fine-tuning these parameters to improve forecast accuracy and residual behavior. >

- **OpenAI ChatGPT(GPT4o)**

Query1:<What are the assessment metrics of Kendall's correlation analysis and what characteristics of the data are expressed>

Output:<Kendall's correlation analysis is assessed using the **Kendall's tau** coefficient, which measures the strength and direction of the association between two variables. The key characteristics expressed by Kendall's tau are:

1. **Strength of Association**:

- **+1**: Perfect positive correlation (both variables increase together).

- **-1**: Perfect negative correlation (one variable increases while the other decreases).

- **0**: No correlation (variables are independent).

2. **Direction of Association**:

- **Positive tau**: Indicates a direct relationship (both variables move in the same direction).

- **Negative tau**: Indicates an inverse relationship (variables move in opposite directions).

Kendall's tau is particularly useful for small sample sizes or when data contains ties (duplicate values), as it is less sensitive to outliers compared to other correlation methods like Pearson’s correlation.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<How to combine two models, SVM and Random Forest, for model building>

Output:<To combine SVM and Random Forest for model building, you can use **ensemble learning** techniques. Two common methods are:

1. **Stacking**:

- Train both SVM and Random Forest models separately on the same dataset.

- Use the predictions from both models as input features to train a meta-model (e.g., logistic regression or another classifier) that makes the final prediction.

2. **Voting**:

- Train both models independently.

- For classification, use **majority voting** (the class predicted by the most models is chosen).

- For regression, use **average voting** (average the predictions of both models).

Both methods combine the strengths of SVM's margin-based decision and Random Forest's ensemble of decision trees, potentially improving overall predictive performance.>

Query2:<Detailed description of the Stacking method>

Output:<**Stacking** is an ensemble learning technique where multiple base models (e.g., SVM, Random Forest) are trained on the same data, and their predictions are combined using a meta-model to make the final prediction. The process includes:

1. **Train Base Models**: Train several different models (e.g., SVM, Random Forest, etc.) on the same training dataset.

2. **Generate Predictions**: Use the trained base models to make predictions on a validation set (or on the training set using cross-validation).

3. **Meta-Model**: The predictions from the base models are treated as features for training a second-level model (the meta-model), often a simple model like logistic regression or a neural network.

4. **Final Prediction**: The meta-model makes the final prediction based on the predictions of the base models.

Stacking improves predictive performance by leveraging the diverse strengths of different models, reducing bias and variance.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<Give me some of your country's strengths in the Olympics.>

Output:<Some of the strengths of the United States in the Olympics include:

1. **Diverse Talent Pool**: With a large population, the U.S. has a vast number of athletes across various sports.

2. **Strong Sports Infrastructure**: World-class training facilities and support systems for athletes.

3. **History of Success**: The U.S. has consistently been a top performer, especially in swimming, athletics, gymnastics, and basketball.

4. **Investment in Elite Sports Programs**: Extensive government and private sector funding for athlete development.

5. **Focus on Team Sports**: The U.S. excels in team sports like basketball, soccer, and volleyball.

These factors contribute to the U.S.'s dominance in the Summer and Winter Olympics.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<Help me polish the following sentence.

We have also found that host countries have higher medal counts because they have the power to add or adjust new sports, which directly affects the distribution of medals.

The host country usually chooses the sports that its athletes are good at to increase the probability of winning medals. For example, Japan has added karate, skateboarding, rock climbing and surfing to the 2020 Tokyo Olympics, which have a wider popular base and competitive strength in Japan than in other countries.

At the same time, the host country can also reduce or adjust certain sports to weaken the competitors' chances of winning medals. For example, by reducing the number of events in track and field, diving and table tennis, the number of medals won by the host country can be increased indirectly.>

Output:< Here's a polished version of your paragraph:

"We have also observed that host countries tend to have higher medal counts, as they have the ability to add or adjust sports, directly influencing the distribution of medals. Typically, the host country selects sports in which its athletes excel, thereby increasing the likelihood of winning medals. For instance, Japan introduced karate, skateboarding, rock climbing, and surfing at the 2020 Tokyo Olympics—sports that have a broader popularity and stronger competitive presence in Japan than in other countries. Additionally, the host country can reduce or modify certain events to limit competitors' chances of winning. For example, by decreasing the number of events in track and field, diving, and table tennis, the host country can indirectly boost its medal tally.">

- **OpenAI ChatGPT(GPT4)**

Query1:<Linear regression is poorly fitted and how to improve it>

Output:<If linear regression is poorly fitted, you can improve it by:

1. **Checking for non-linearity**: If the relationship between variables is not linear, consider using polynomial regression or other nonlinear models.

2. **Feature engineering**: Add relevant features, create interaction terms, or transform variables (e.g., log, square) to improve model fit.

3. **Handling outliers**: Remove or handle outliers that can disproportionately affect the model.

4. **Addressing multicollinearity**: Use techniques like variance inflation factor (VIF) to identify and remove or combine highly correlated features.

5. **Regularization**: Apply regularization techniques like Lasso or Ridge regression to improve generalization and prevent overfitting.

6. **Increasing data quality**: Ensure data is clean, has sufficient samples, and has minimal noise.

These steps can help improve the model's fit and predictive accuracy.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<The Great Coach Effect is reflected in more ways than medal counts>

Output:< The "Great Coach Effect" goes beyond just medal counts and is reflected in:

1. **Athlete Development**: A great coach helps athletes reach their full potential, improving their skills and performance over time.

2. **Team Morale**: Coaches foster team spirit, motivation, and mental resilience, enhancing overall performance.

3. **Training Innovation**: Great coaches often introduce new training methods, strategies, and techniques that elevate the team's competitiveness.

4. **Long-term Success**: Effective coaching builds a lasting foundation, creating a legacy of excellence that extends beyond individual Olympics.

5. **Talent Identification**: Top coaches can spot and nurture emerging talent, ensuring sustained success for future competitions.

This broad influence contributes to overall sporting success and the development of athletes, not just the immediate medal count.>

- **OpenAI ChatGPT(GPT4o)**

Query1:<Help me write a piece of python code that uses ten-fold cross validation>

Output:<Here’s a concise Python code that uses **10-fold cross-validation** with a machine learning model (e.g., SVM):

```python
from sklearn.model_selection import cross_val_score
from sklearn.svm import SVC
from sklearn.datasets import load_iris
# Load a dataset (Iris dataset as an example)
data = load_iris()
X = data.data
y = data.target
# Initialize the model (SVM in this case)
model = SVC()
# Perform 10-fold cross-validation
cv_scores = cross_val_score(model, X, y, cv=10)
# Print the cross-validation scores and their mean
print("Cross-validation scores:", cv_scores)
print("Mean CV score:", cv_scores.mean())
```

This code loads the Iris dataset, trains an SVM classifier, and performs 10-fold cross-validation to evaluate its performance. The `cross_val_score` function returns the scores for each fold, and the mean score is displayed at the end.>

- **OpenAI ChatGPT(GPT4o)**

  Query1:<How to use Latex for better typography>

  Output:<Using LaTeX for better typography involves several key strategies:

1. **Font Choice**:
   - Use `\\usepackage{mathptmx}` for a Times-like font or `\\usepackage{helvet}` for Helvetica.
   - Use `\\usepackage{fontspec}` for custom fonts if using XeLaTeX or LuaLaTeX.
2. **Proper Spacing**:
   - Use `\\usepackage{microtype}` for better character and word spacing (improves readability).
   - Adjust line spacing with `\\linespread{1.3}` for a comfortable reading experience.
3. **Hyphenation and Justification**:
   - Control hyphenation with `\\hyphenpenalty=5000` to reduce unwanted breaks.
   - Enable or fine-tune justification with `\\raggedright`, `\\raggedleft`, or `\\centering`.
4. **Mathematics**:
   - For high-quality math typesetting, use `\\usepackage{amsmath}` and `\\usepackage{amsfonts}` for better alignment and symbols.
   - Use `\\text{}` within math environments to integrate text smoothly.
5. **Widows and Orphans**:
   - Avoid widows and orphans (isolated lines) with `\\clubpenalty=10000` and `\\widowpenalty=10000`.
6. **Sectioning**:
   - For well-structured documents, use `\\section`, `\\subsection`, and `\\subsubsection` for clarity and hierarchy.
   - Customize section fonts with `\\usepackage{titlesec}`.
7. **Lists and Indentation**:
   - Use `\\begin{itemize}` or `\\begin{enumerate}` for clean and easy-to-read lists.
   - Customize indentation with `\\setlength{\\parindent}{0pt}` for no indentation.
8. **Tables and Figures**:
   - Use `\\usepackage{booktabs}` for elegant tables with proper horizontal lines.
   - Use `\\caption` and `\\label` for well-positioned figures and tables.

By applying these practices, LaTeX will give you a professional, polished output with high typographic standards.>