
OlymPrc Integrated Olympic Prediction Model

Summary

In order to forecast the medal distribution in the forthcoming Olympic Games, grasp the sports standards of various countries, assist nations in formulating pertinent sports policies, refine targeted training in diverse sports, and boost national sports competitiveness, we would like to develop a comprehensive projection system based on a time-series model and a supervised learning algorithm to predict the medals for the next Olympic Games.

Several models are constructed as a component of our system:

For task 1, firstly, a National Sports Level Classification model, Host Effect Evaluation, and an Athlete Evaluation System are established to help process the raw data and construct a feature set for machine learning, and dig into the depth relation between those variables and actual medal counts. After residual analysis, MSE, RMSE and other metrics evaluation, we chose ARIMA and GB as the main prediction models, which have the lowest error rate and higher robustness. By the ARIMA-GB model we constructed, we obtained the prediction results as in Figure 5 and Figure 6, with a range of error within plus or minus 2.3 medals for a single species. Also, we find countries like Libya and Angola are quite likely to achieve their first Olympic medal in 2028 (2).

For task 2, we first used the Linear Regression model to fit the data on "great coaches" coaching and medal counts, analyzed the model parameters with OLS, and found a strong correlation between the model and the coaches. Subsequently, we screened for countries and projects that needed "great coaches" using indicators such as standard deviation. Applying the model to these projects, we predicted the changes in medal counts for these projects in 2028 with the help of "great coaches," and ultimately concluded that there is a strong correlation between "great coaches" coaching and the increase in medal counts.

keyword: ARIMA Gradient Boosting Linear Regression K-means Olympic medal

Contents

1	Abstract	2
2	Background	2
3	System Overview	3
4	Modeling and Projections for Olympic Medal Counts	3
4.1	Data Cleaning	3
4.2	Data Process	3
4.2.1	National Sports Level Classification	3
4.2.2	Athlete Evaluation System Formula	4
4.3	Feature Selection	6
4.4	Model Selection	7
4.5	Model Training and Evaluation	8
4.5.1	Data Input	8
4.5.2	Parameters Selection	9
4.5.3	Model Evaluation	9
4.6	ARIMA-GB Model	10
4.6.1	Core Idea	10
4.6.2	Error and Weight Calculation	10
4.6.3	Combination Formula	11
4.7	Result	11
5	Events, Country and Medal Counts Analysis	12
5.1	Correlation Analysis Of Events, Country and Medal Counts	13
5.1.1	Heat-map Visualization	13
5.1.2	”Host Country” Effect Analysis	14
6	The “Great Coach” Effect Analysis	17
6.1	Data Process	17
6.2	Feature and Model Selection	19
6.3	Model Training and Validation	19
7	Conclusion	20
7.1	Task 1	20
7.2	Task 2	21
8	Discussion	21
8.1	Model Insights	21
8.2	Strategic Implications	21
A	AI Use Report	22

1 Abstract

In order to predict the distribution of medals in the next Olympic Games, to understand the level of sports in each country, to guide countries in designing relevant sports policies, to improve the training of different sports in a targeted manner, and to enhance the competitive level of the national sports, we designed a system based on a time-series model and a supervised learning algorithm to predict the medals in the next Olympic Games.

With comprehensive evaluation, finally the number of medals in history, the level of national sports, the level of athletes (we construct a evaluation system for athletes) participating in the competition, and the bonus of the host country are calculated as the feature information to make the projection; the ARIMA model and the Gradient Boosting model are selected as the main prediction component, and the results of the two are integrated to obtain the final prediction results.

At the same time, we mine the correlation between the medal data and the information of great coaches, and then filter out the countries that might need 'great coaches' to improve their medal performance through ordinary least squares (OLS) and other indices, and predict the changes in medal counts of these countries after having 'great coaches' by the linear regression model, so as to get a more perfect conclusion.

2 Background

Every four years, there is always a sporting event that attracts the attention of people from all over the world, who may travel from all over the world to be on the spot or just stay in front of the screen to cheer on the athletes. Since it is Olympic, a competition, there is inevitably a medal competition, and during the Olympic Games, people around the globe are constantly watching every change in the medal table, and there are also many people who conduct layers of analyses, trying to predict the outcome of the competition.

More than just for entertainment purposes, predicting the medal table for the next Olympics can help countries to plan their sports strategies more strategically and allocate resources more efficiently to the sports where they are more likely to win medals. For example, forecasts can figure out areas where the country is falling short and then can consider introducing new training programmes or investing in more relevant facilities; or where there is strong potential and then country can maintain the resources invested in the project to ensure that performance in the coming year is at a consistently high level. In addition, knowing the number of medals expected can set a reasonable goal for the country's athletes and coaching staff to strive for, motivating athletes to perform at their best and achieve better results.

To sum up, predicting the number of medals in the next Olympic Games is very important for the development of national sports and the competitiveness of future events.

This system aims to construct an accurate and robust medal prediction model based on information from previous years' Olympics Games, including historical medal counts, competition programme information, and relevant participation information of athletes from various countries in each year.

3 System Overview

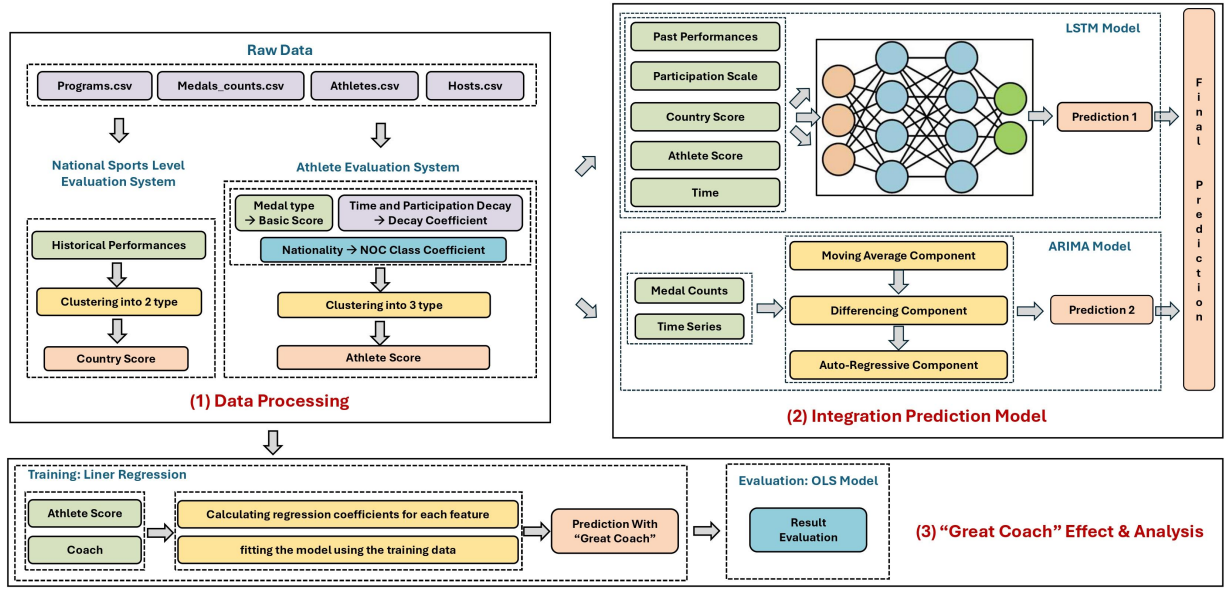


Figure 1: Overview of this work.

4 Modeling and Projections for Olympic Medal Counts

4.1 Data Cleaning

The data cleaning process involved standardizing country names, NOC codes, and event categories to ensure consistency and accuracy in the dataset. The steps are summarized as follows:

- 1. Standardizing Country Names:** Certain regions were unified under their respective country names for consistency. For example, *British West Indies* was mapped to *Great Britain*.
- 2. Cleaning NOC Codes:** NOC codes representing the same country were standardized. For instance, *LBA* and *LBI* were both unified as *Libya*.
- 3. Categorizing Sports:** Similar sports were grouped under a single category for simplification. For example, *Synchronized Swimming* was categorized as *Swimming*.

These cleaning steps ensured that the dataset was consistent, accurate, and ready for subsequent analysis.

4.2 Data Process

4.2.1 National Sports Level Classification

Clustering medal counts based on historical Olympic performance can provide valuable insights into the competitive level of different countries. Different categories of countries may be characterised

by different trends in terms of athletic ability and medal winning, and categorising countries can provide more information for subsequent medal predictions and improve the accuracy of the system. For example, sports resources associated with strong sports countries are usually more abundant and concentrated, with a higher likelihood of breakthroughs in athlete development, leading to an increase in medal counts.

In our system, we use the sum of historical gold, silver, bronze and total medals as features, using Kmeans Clustering to do the country classification. In addition, we standardized the feature data so that each feature has a mean of 0 and a standard deviation of 1 to ensure that the contribution of different features is equal, which can avoid bias in the classification results due to the difference in magnitude. The Classification result are shown in Figure 2, divided into sport powerhouses (cluster 1) and sport developing countries (cluster 0).

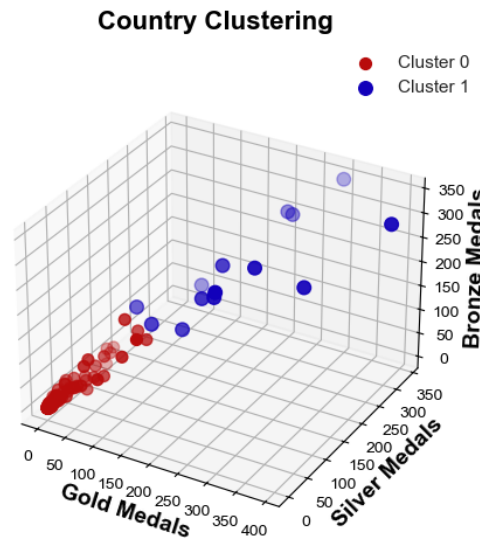


Figure 2: Country Classification Base on Medal Counts

4.2.2 Athlete Evaluation System Formula

To comprehensively evaluate an athlete's performance, we combine the following factors:

- Basic Score

$$\text{Raw Score} = \{ \text{'Gold': 3, 'Silver': 2, 'Bronze': 1, 'No medal': 0} \}$$

- Decay Coefficient

– Participation Decay

- * Formula: The effect of the athlete's number of participation, reflecting the balance between experience accumulation and performance stabilization/decline:

$$f_{\text{participation}}(n) = \frac{1}{1 + e^{-k_1 \cdot (n - t_1)}}$$

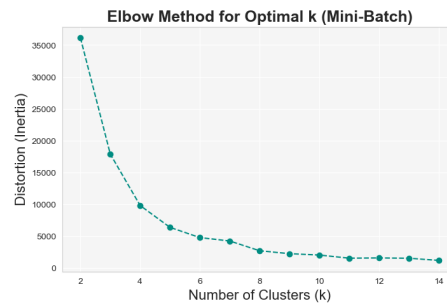
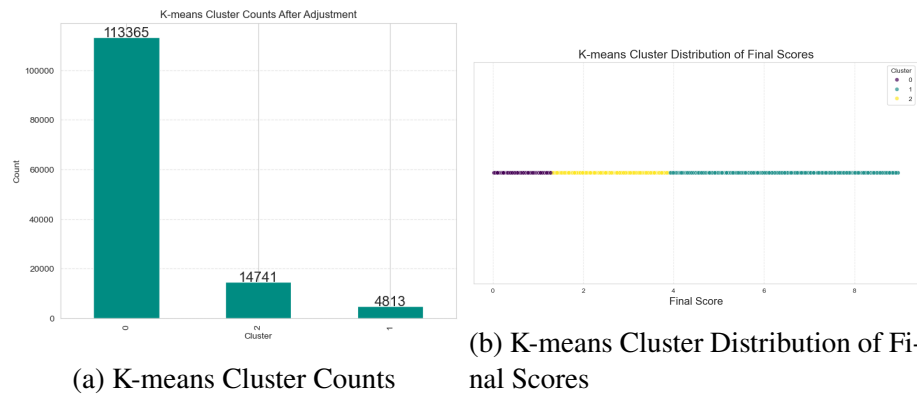


Figure 3: Elbow Method for Optimal k (Mini-Batch)



* Explanation:

- n : The total number of participation by the athlete.
- k_1 : A parameter controlling the steepness of the decay.
- t_1 : The baseline participation value (experience inflection point).

– Time Decay

- * Formula: Reflects the athlete's career peak and decline over time:

$$f_{\text{time}}(\text{Year}) = \frac{1}{1 + e^{-k_2 \cdot (\text{Year} - t_2)}}$$

* Explanation:

- Year: The year of the competition.
- k_2 : A parameter controlling the steepness of the decay.
- t_2 : The baseline time point (e.g. the year 2016).

• Country's Sports Development Effect

- Definition Reflects the advantage of strong sports nations in training and resources:

$$w = \begin{cases} w_1 & \text{Strong sports nation (Cluster = 1)} \\ w_0 & \text{Weak sports nation (Cluster = 0)} \end{cases}$$

Final Score Formula

Combining the above factors, the final score is calculated as:

$$\text{Final Score} = \text{Raw Score} \times f_{\text{participation}}(n) \times f_{\text{time}}(\text{Year}) \times w$$

Finally, K-means clustering was applied to the athletes' scores, and the optimal number of clusters was determined using the Elbow method. As a result, the data was divided into three clusters.

4.3 Feature Selection

Although, as a rule, we don't speculate on the medal table for the next tournament based solely on the number of medals a country has historically won, because the number of medals is a combination of the level of athletes competing in that year's tournament, as well as a number of other factors. The emergence of new high-level athletes, the retirement of old athletes, and changes in physical condition of athletes are all related to the medal counts, and these information seems cannot be shown in the pure medal table data. However, by analyzing the datas, we can still find that the change in the number of medals in history reflects the overall level of the country's sports to a certain extent, and has a certain correlation with the number of medals in the next competition.

The number and ratio of different medals is still considered to be a direct indicator of a country's or its athletes' performance at the previous Olympic Games in our system. A higher number of gold (or silver and bronze) medals may indicate that the country is strong in certain sports, which may continue to play a role in the next Olympic Games.

Therefore, we calculate and include the following statistic informations in our feature set.

- **Past_Gold:** Number of gold medals in the last Olympics
- **Past_Silver:** Number of silver medals in the last Olympics
- **Past_Bronze:** Number of bronze medals in the last Olympics
- **Past_Total:** Number of all medals in the last Olympics
- **Gold_Ratio:** Percentage of gold medals in the previous Olympic Games
- **Silver_Ratio:** Percentage of silver medals in the previous Olympic Games
- **Bronze_Ratio:** Percentage of bronze medals in the previous Olympic Games

In addition to the above characteristics, the addition of the following characteristics (more relative to the athletes' and participation, instead of medal counts) can provide a more comprehensive understanding of past participation and performance of athletes and countries, leading to a more accurate prediction of future medal counts

- **Events_Count & Sports_Count:** The greater the number of events a country participate in, the greater chances of winning medals. Greater number of participation may also indicate that the country has a higher level of competitiveness in more than one sports type, which may also reflected in the number of final medals.

- **isHoster:** The host country usually has the bonus of home advantage, including crowd support, familiarity of the competition environment, etc. These factors may lead to an increase in the number of medals for the host country. And as we known, the host country has some power to decide to add or delete some sports events, therefore they can add and participation in more dominant sports to increase the chances of winning medals.
- **Country_Type:** Base on the sum of historical medal counts, we use Kmeans Cluster to classify the countries into two group, sporting powers and sporting developing countries. Sporting powers countries that have won a lot of medals in previous years usually have more effective experience and a mature system in training their athletes and are able to maintain a consistently high number of medals.
- **Ath_Score:** For every athlete, we calculate a performance score, considering medals he/she won, participation decay, time decay and the country's sport development Effect. Athletes with higher score are more likely to win medals in competitions. Then sum up the past years athletes' scores, indicating country's overall strength and performance, which directly affects the awarding of medals.
- **Good_Ath_Ratio:** Use Kmeans Cluster to classify the athletes into three levels, and then calculate the highest level athletes ratio of that country in past years. The higher the percentage of outstanding athletes, the greater the chance of winning a medal, reflecting the overall strength of the country.
- **Year:** The Olympics may have different trends and changes from year to year, so considering the year help capture these changes over time.

4.4 Model Selection

Considering we have the historical medal data and athlete participation data, when predicting the number of medals of different countries for the next 2028 Summer Olympic Games, the following models are considered in our trial:

- **Autoregressive Integrated Moving Average Model (ARIMA)**

Frist we consider the ARIMA model, because the past medal records is partly a reflection of the trend of the country's level of sports competitiveness. So we would like to predict a country's next year performance base on the historical medals. ARIMA model is a time series forecasting model which is commonly used when handling data with temporal dependencies. It can capture the trend, the seasonal or cyclical patterns inside the given historical medal counts over time to forecast future medal counts.

Comparing with the Machine Learning methods, ARIMA can maximise retention of trend characteristic and seasonality in the time series data, and will be effective in modeling the temporal dependencies in historical medal data.

- **Linear Regression Model**

Linear Regression is a simple and easy way to interpret and understand the linear relations hidden behind the numerous datas, and it is fast in computation, making it suitable for handling large-scale dataset.

To find out whether there are linear relationships between our features, we take Linear Regression model into account. In our predicting model, historical medal counts (including gold, silver, bronze and total number), and athletes' basic participation information (such as evaluation score, competitive level etc.) are all used as independent variables to predict the medal counts for the next Olympic Games.

- **Decision Tree Model**

Since the complexity of the dataset, the linear correlation of data cannot be guaranteed, therefore we still need some other models as a supplement.

The decision tree model is a tree-structured classification and regression model that can handle non-linear relationships in data. When predicting Olympic medal counts, it can use a series of decision rules based on athletes' information and medal count records to predict the medal counts for the next Olympic Games. Better still, this model is robust to noise and outliers in the data, which might help improving the accuracy of forecasts.

- **Random Forest Model**

The random forest model is an ensemble learning model, which can improve the accuracy and stability of predictions by constructing multiple decision trees and combining their prediction results. So it is also less prone to overfitting than a single decision tree.

In predicting medals, the random forest can comprehensively consider the impact of multiple factors on medal counts, handle high-dimensional data and non-linear relationships (which are quite likely to be seen in our data), and is still robust to noise and outliers in the data.

- **Gradient Boosting Model**

The gradient boosting model is also an ensemble learning model based on decision trees, while it is constructed by gradually optimizing the prediction errors of the model. The core process is iteratively train weak learners to correct the errors of the previous models, thereby improving the overall performance. This model can also achieve quite high prediction accuracy and stability, and is suitable for high-dimensional data and non-linear relationships.

4.5 Model Training and Evaluation

4.5.1 Data Input

For ARIMA, the time series model simply acquires medal records, for all countries, the numbers of gold, silver, and bronze medals in each of the previous years, and then feeds them into the ARIMA model to predict the medals winning for the next Olympics.

For Linear Regression model, Decision Tree model, Random Forest model and Gradient Boosting model, we extract numbers of features from the raw dataset (what we mentioned in Part 3.2, `X_train = ['Past_Gold', 'Past_Silver', 'Past_Bronze', 'Past_Total', 'Gold_Ratio', 'Silver_Ratio', 'Bronze_Ratio', 'Events_Count', 'Sports_Count', 'isHoster', 'Country_Type', 'Year']`).

4.5.2 Parameters Selection

1. ARIMA: According to the ACF figure and PACF figure, we choose the ARIMA model's order.
2. Machine learning models: GridSearchCV() is used to find the parameter, taking both the lower error rate and smaller time complexity into consideration.

The final models are constructed as follows:

```

1  model = ARIMA(train, order=(2, 1, 1))
2  model = LinearRegression()
3  model = DecisionTreeRegressor( max_depth=None,
4                                min_samples_split=5,
5                                max_features=None,
6                                max_leaf_nodes=15,
7                                random_state=42),
8  model = RandomForestRegressor( n_estimators=250,
9                                max_depth=5,
10                               random_state=42,
11                               min_samples_split=6,
12                               min_samples_leaf=1,
13                               max_features=5),
14  model = GradientBoostingRegressor(random_state=42,
15                                    max_depth=2,
16                                    min_samples_split=9,
17                                    n_estimators=250,
18                                    learning_rate=0.072,
19                                    min_samples_leaf=3)

```

4.5.3 Model Evaluation

We split the dataset in the ratio of training set and test set as 8:2 to evaluate the previously mentioned models, and the final results computed on the test set are shown in Figure 4. And we also calculate the Mean Squared Error(MSE), Root Mean Squared Error(RMSE), R-squared(R^2) and Mean Absolute Percentage Error(MAPE), the performance of those models can be seen in Table 1, from both the figures and table, we can find that the Gradient Boosting Model gives out the most outstanding performance from the others (with lowest MSE, RMSE and R^2 closest to 1).

Therefore, in our prediction model, Gradient Boosting is finally chosen and will be combined with the ARIMA model to predict the 2028 Olympic Medal Tally.

Model	MSE	RMSE	R^2	MAPE
Linear Regression	7.4594	2.7312	0.7836	10e+14
Decision Tree	7.9236	2.8148	0.7702	6.8e+14
Random Forest	6.9320	2.6328	0.7989	5.5e+14
Gradient Boosting	5.3087	2.3041	0.8460	4.8e+14

Table 1: Models' Accuracy Evaluation

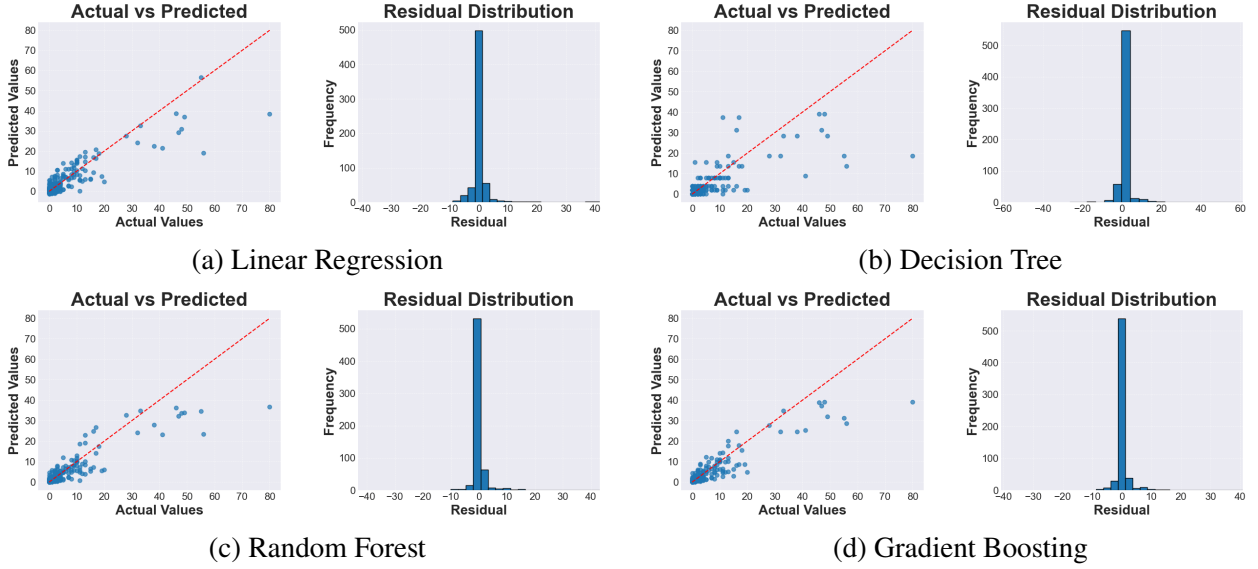


Figure 4: Models Evaluation

4.6 ARIMA-GB Model

4.6.1 Core Idea

As the ARIMA model is specifically designed to deal with time-series data, it can efficiently capture the time-dependence and trends in the data, and is suitable for data with a more pronounced linear trend, which can quickly respond to changes in the data. The Gradient Boost model, on the other hand, captures non-linear relationships in time series data and is suitable for complex non-linear scenarios such as Olympic medals prediction, mining subtle changes and trends in the data, and enabling better long-term prediction. Therefore, we decided to combine the two models to better utilize their respective advantages and improve the accuracy and robustness of the projection system.

4.6.2 Error and Weight Calculation

First calculate the error of both model:

- **ARIMA Error:** $error_{ARIMA} = |y - \hat{y}_{ARIMA}|$

- **Gradient Boost Error:** $error_{GB} = |y - \hat{y}_{GB}|$

where y is the actual medal counts, \hat{y}_{ARIMA} is the projection of single ARIMA model and \hat{y}_{GB} is the projection of single Gradient Boosting model.

Then we define the weights by the ratio of the errors:

- **ARIMA Weight:** $weight_{ARIMA} = \frac{error_{GB}}{error_{ARIMA} + error_{GB}}$

- **Gradient Boost Weight:** $weight_{GB} = \frac{error_{ARIMA}}{error_{ARIMA} + error_{GB}}$

4.6.3 Combination Formula

Final projection result is:

$$\hat{y} = weight_{ARIMA} * \hat{y}_{ARIMA} + weight_{GB} * \hat{y}_{GB}$$

. The final weights are calculated from the data training prediction results as $weight_{ARIMA} = 13\%$, $weight_{GB} = 87\%$. And according to the MSE and RMSE value of this ARIMA-GB model, the medal count projections range is approximately $[\hat{y} - 2.3, \hat{y} + 2.3]$.

4.7 Result

Putting all the data from 2024 and before into the model training and predicting the number of medals for the 2028 Olympics, the top 10 countries in the overall medal table are shown in Figure 5.

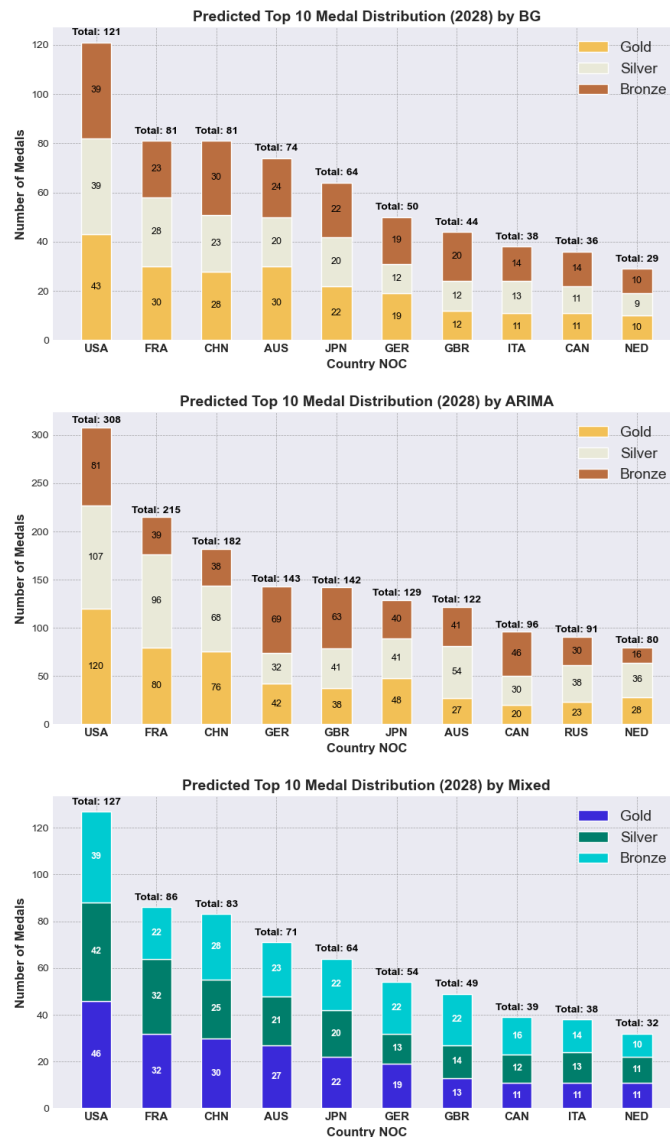


Figure 5: Total Medal Projection

Figure 6 is a look at the top 20 countries in the 2024 medals table, and the predicted changing trends in performance in 2028 Olympic Games.

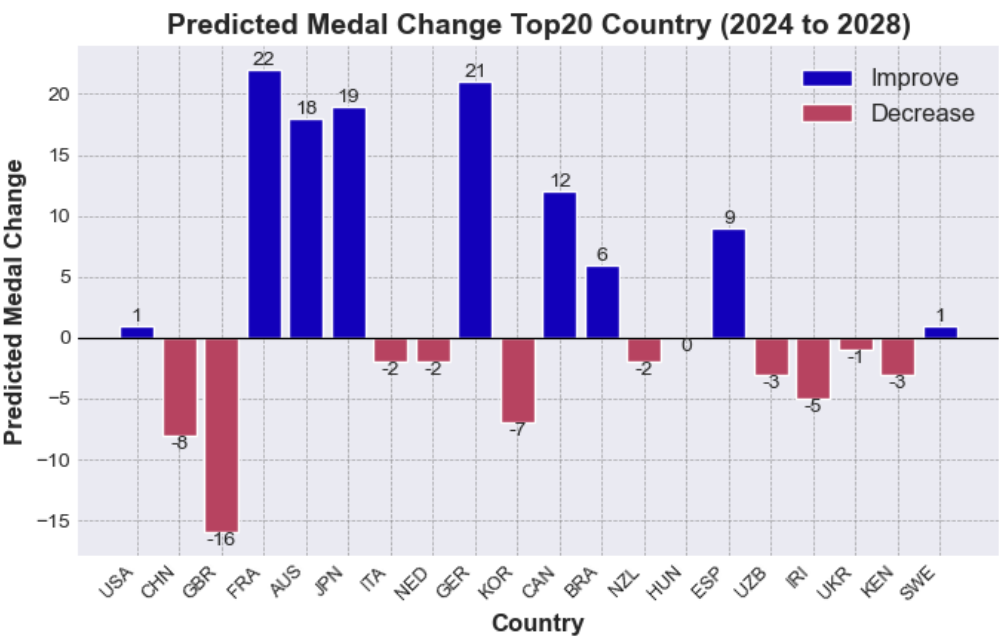


Figure 6: Predicted Performance Change of Top 20 Countries

Notably, of all the countries who has not yet won a medal in Olympics, Angola and Libya are predicted to win its first medal in 2028 with a possibility of approximately 50%. The following Table2 shows the projection of some countries which has no medal records.

Country	LBI	ANG	MYA	BIH	SOL	CGO	HON
Medal Projection	0.52	0.49	0.34	0.34	0.34	0.33	0.33

Table 2: Projection of some No_medal Countries

5 Events, Country and Medal Counts Analysis

The dominance of countries such as the United States in gymnastics and swimming raises questions about the relationship between sports events, countries, and medal counts. By analyzing the correlation between event-specific medal distributions and overall national medal performance, we can identify which events contribute the most to the success of a country. This analysis helps us understand how certain countries gain a competitive edge in specific sports and provides information on the strategic value of event selection in future Olympic Games.

5.1 Correlation Analysis Of Events, Country and Medal Counts

5.1.1 Heat-map Visualization

Heat-maps are a powerful tool to visualize data, utilizing color gradients to emphasize patterns and trends. They make complex datasets easier to interpret by providing a clear overview of value distributions. In the context of Olympic medal counts across countries and events, heat-maps effectively highlight which countries excel in specific events, allowing us to quickly identify areas of strength through high medal count combinations.

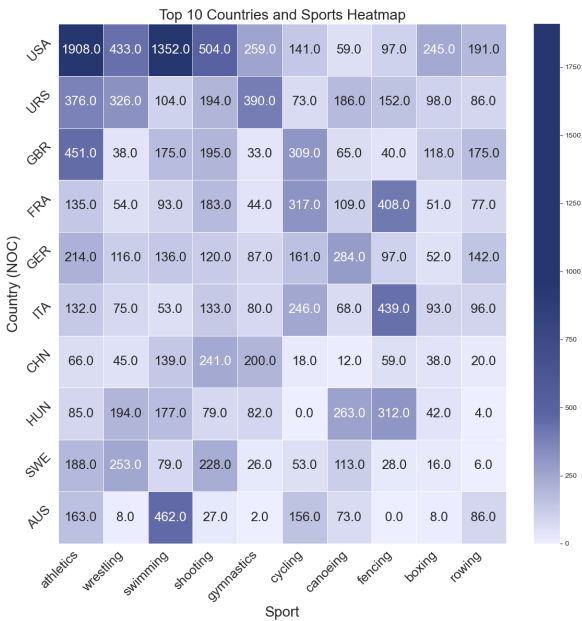


Figure 7: Top 10 Countries and Sports Heat-map with Medal Counts

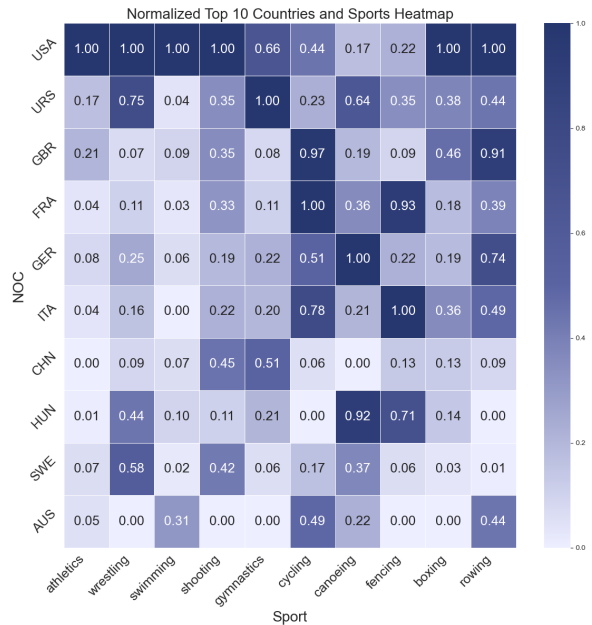


Figure 8: Top 10 Countries and Sports Heat-map with "Min-Max-Scaler" Medal Counts

Heat-maps introduction Figure 5 displays the raw medal counts for the top 10 countries across selected sports. Figure 6 normalizes the data using a Min-Max scaler, emphasizing the relative importance of each sport for each country rather than absolute dominance. The darker the cell, the higher the absolute number of medals earned by a country in a particular sport. Normalization highlights relative strengths, revealing how smaller nations like Hungary excel in niche sports despite having fewer total medals. It also de-emphasizes the overwhelming dominance of large nations in absolute medal counts, providing a clearer picture of specialization. **Sport-Specific Importance Across Countries**

- **United States (USA):** Consistently high performance in swimming, athletics, and gymnastics, as shown by the darkest cells in both heat-maps. These sports offer multiple events, providing the USA with opportunities to secure numerous medals.
- **Soviet Union (URS):** Excels in wrestling and gymnastics, as evident in both heat-maps. These sports reflect historical investments in athlete training and development.
- **Great Britain (GBR):** Shows relative strength in cycling and rowing, which aligns with their investments in Olympic programs for these sports.

- **China (CHN):** Notable in gymnastics and shooting, reflecting China's emphasis on precision and technical sports.
- **Australia (AUS):** Excels in swimming, which is in line with its strong swimming culture and extensive national programs.
- **Hungary (HUN):** Dominates in fencing and canoeing, highlighting its traditional strengths in niche sports.

Factors Influencing Sports Dominance

- **Cultural and Historical Factors:** Countries tend to excel in sports that are culturally significant or historically prioritized. For example, Hungary's dominance in fencing is deeply rooted in tradition.
- **Government Investment:** Nations like China and Great Britain have systematically invested in targeted sports programs, leading to specialization in specific events.
- **Medal-Rich Sports:** Sports such as swimming and athletics, which include multiple events, disproportionately favor countries like the USA with large athlete contingents and resources.
- **Geographical Influence:** Australia's strength in swimming and Great Britain's dominance in rowing can be attributed to their geography and access to natural training environments.

5.1.2 "Host Country" Effect Analysis

To analyze whether host countries tend to select events in which they have a competitive advantage, and how these selections impact their overall medal tally, we first need to differentiate the data based on whether the country is the host. This can be achieved by using the "isHoster" field. Next, we will count the events selected and the total medal counts for host countries during their hosting years to observe any patterns or biases in their event selection and performance.

• Correlation Analysis

– Chi-Square Test

To examine whether there is a significant association between the types of events selected by host countries and their total medal counts, we employed a chi-square test. The Chi-Square test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i represents the observed frequencies and E_i represents the expected frequencies. Although the large degrees of freedom (848) might suggest that the dataset's size could contribute to the chi-square value, the extremely small p-value (7.41e-171) and the high chi-square statistic (2558.69) strongly indicate a significant association between the variables, making it clear that the observed and expected frequencies deviate substantially.

Statistic	Value
Chi-Square Statistic	2558.6901303909285
p-value	7.406758772945909e-171
Degrees of Freedom	848

Table 3: Chi-Square Test Results

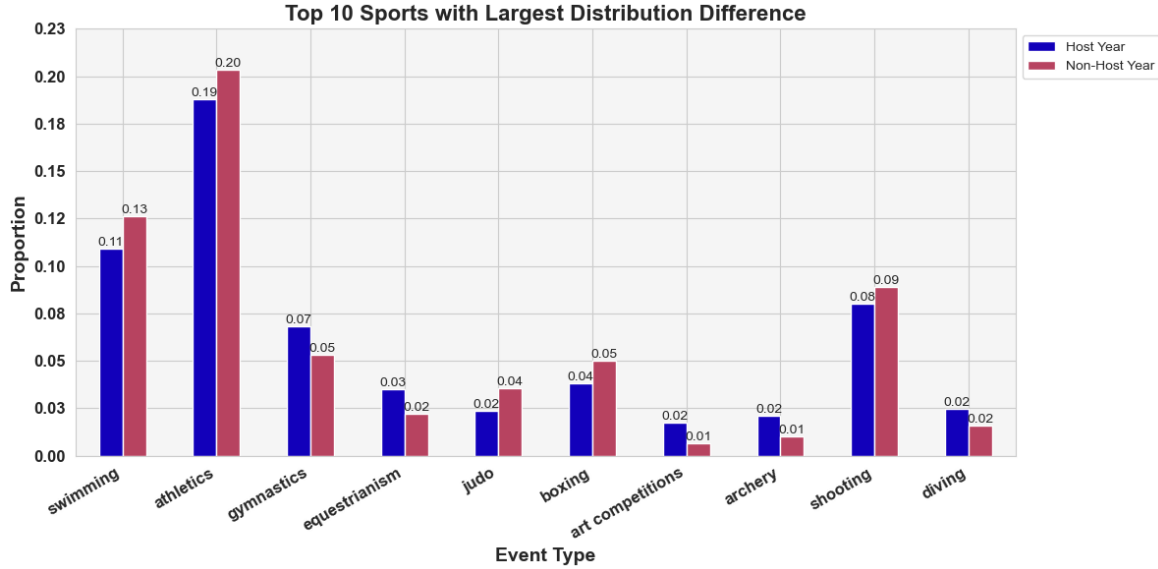


Figure 9: Top 10 Sports with Largest Distribution Difference between Host and Non-Host Years

– Regression Analysis

To investigate whether being a host country significantly impacts the total number of medals, we conducted a linear regression analysis. The independent variable (*isHoster*) indicates whether a country is the host (1 for host, 0 otherwise), while the dependent variable represents the total number of medals (*Medals*), calculated as the sum of Gold, Silver, and Bronze medals. The Ordinary Least Squares (OLS) Regression Analysis is calculated as:

$$\text{Medals}_i = \beta_0 + \beta_1 \cdot \text{isHoster}_i + \epsilon_i$$

Where:

- * $\beta_0 = 0.1572$: The intercept, representing the expected number of medals when a country is not the host.
- * $\beta_1 = 0.2022$: The coefficient for *isHoster*, representing the additional medals a host country is expected to win on average.
- * ϵ_i : The error term accounting for variability not explained by the model.

The regression analysis shows an R-squared of 0.012, indicating limited explanatory power, yet the model is statistically significant (F-statistic = 1527.0, $p < 0.001$). The host country status (*isHoster*) has a significant positive effect on total medal counts, with a coefficient of 0.2022 ($p < 0.001$) and a precise 95% confidence interval of [0.192, 0.212]. While residual

Table 4: OLS Regression Results: Impact of Host Country Status on Medal Counts

Metric	Value	Metric	Value
R-squared	0.012	Coefficient (const)	0.1572
Adjusted R-squared	0.012	Coefficient (isHoster)	0.2022
F-statistic	1527.0	Standard Error (isHoster)	0.005
Prob (F-statistic)	0.00	t-value (isHoster)	39.080
AIC	1.106e+05	p-value (isHoster)	0.000
BIC	1.107e+05	95% Confidence Interval (const)	[0.155, 0.159]
Coefficient (isHoster)	0.2022	95% Confidence Interval (isHoster)	[0.192, 0.212]
Durbin-Watson	1.383	Omnibus	37138.225
Prob (Omnibus)	0.000	Jarque-Bera (JB)	77893.556
Skew	1.781	Kurtosis	4.268

tests indicate slight autocorrelation (Durbin-Watson = 1.383) and deviation from normality, the results affirm that being a host country significantly impacts medal counts.

– Host and Non-Host Event Prioritization

To investigate the impact of events chosen by host countries on their performance, we analyzed the proportional distribution of sports events between host and non-host years. The analysis compared the relative frequency of each event during host years and non-host years, with a focus on the top 10 events that showed the largest distribution difference. The proportional difference for each event was calculated as:

$$\text{Difference} = |\text{Proportion (Host Year)} - \text{Proportion (Non-Host Year)}|$$

The results are visualized in Figure 9, which highlights the sports where host countries show the greatest deviations in proportional representation.

As shown in Figure 9, sports such as *Athletics*, *Swimming*, and *Gymnastics* demonstrate notable proportional differences, with host countries allocating greater focus to certain events. For instance:

- * *Athletics*: Proportion during host years (0.19) is slightly lower than non-host years (0.20), indicating high competition across all contexts.
- * *Gymnastics*: Shows a higher proportion in host years (0.07) compared to non-host years (0.05), suggesting strategic emphasis by host countries.
- * *Shooting*: A higher proportion in host years (0.08) versus non-host years (0.09), potentially reflecting preparation advantages or strategic prioritization.

Conversely, events such as *Archery* and *Diving* exhibit minimal proportional differences, indicating limited impact of host country selection on these events. The analysis supports the hypothesis that host countries strategically focus on events where they have competitive advantages, while globalized events maintain consistent representation.

The results suggest that host countries exhibit a strategic bias toward certain events, which may be influenced by cultural preferences, resource allocation, or competitive strengths. These findings underscore the role of event selection in shaping host countries' performance outcomes.

6 The “Great Coach” Effect Analysis

6.1 Data Process

When studying the impact of “great coaches” on medal counts in sports competitions, it is evident that coaches have a more significant influence on team sports than on individual sports. This is because individual sports often rely more on the athletes’ talent and hard work, while team sports emphasize the coordination and cooperation among athletes. The presence of a coach can help athletes better find their positioning. Therefore, in this study, we focus primarily on the impact of great coaches on team sports.

We first extracted all combinations of Sport+Event from the athletes’ CSV file and used keywords such as “single” to eliminate individual sports. However, due to the complexity of the data, not all individual sports have obvious keywords in their names. Therefore, we extracted team sports for the following Sport types:

- Archery
- Athletics
- Canoe Slalom
- Canoe Sprint
- Canoeing
- Cycling
- Cycling Track
- Diving
- Gymnastics
- Sailing
- Shooting
- Swimming

For other Sport types not mentioned, such as Volleyball and Boxing, we can simply infer that these Sports are all team/individual sports. Therefore, we performed a simple full exclusion/preservation treatment on them, resulting in a CSV file containing all team sports.

Through data analysis, we noticed that in different Olympic years, the same Event often has different Event names. For example, in Sport = Volleyball, the women’s team has three Event names: “Women,” “Women Team,” and “Volleyball Women’s Volleyball,” which clearly refer to the same Event. Therefore, if we want to systematically analyze a country’s project, it is necessary to unify these names. Therefore, we unified the Events through script and manual adjustment and recorded them in `Events_final.csv`. The table data example is as follows:

A coach often only acts on a specific country’s Sport+Event, and the medals for this Event only have three possibilities: gold, silver, bronze and No medal. To better study the effect of medals, we converted these medals into weights:

```
medal_map = {'Gold': 3, 'Silver': 2, 'Bronze': 1, 'No medal': 0}
```

Using the above methods, we can draw medal line charts for any project. Taking Sport = Gymnastics, the image example of the drawn chart is as follows:

To introduce the influence of coaches, we added a new column of data “coach” to represent whether there is the help of a “great coach,” with values of 1/0 (yes/no). Since the problem statement does not provide exact start and end times for the coaches Lang Ping[1] and Béla Károlyi[2], we have looked up the information and completed the coaching information of the two coaches, and used it to build the “coach” data.

Sport	Event	Event_update
Volleyball	Volleyball Men's Volleyball	Men Team
Volleyball	Volleyball Women's Volleyball	Women Team
Volleyball	Men Team	Men Team
Volleyball	Women Team	Women Team
Volleyball	Men	Men Team
Volleyball	Women	Women Team
Swimming	Swimming Men's 4 x 100 metres Freestyle Relay	Men's 4 x 100m Freestyle Relay
Swimming	Swimming Women's 4 x 100 metres Freestyle Relay	Women's 4 x 100m Freestyle Relay
Swimming	Swimming Women's 4 x 200 metres Freestyle Relay	Women's 4 x 200m Freestyle Relay
Swimming	Swimming Men's 200 metres Team Swimming	Men's 200m Team Swimming
Swimming	Men's 4 x 100m Freestyle Relay Team	Men's 4 x 100m Freestyle Relay
Swimming	Women's 4 x 100m Freestyle Relay Team	Women's 4 x 100m Freestyle Relay
Swimming	Women's 4 x 200m Freestyle Relay Team	Women's 4 x 200m Freestyle Relay

Table 5: partial Event_final.csv: Unified Event Names for Volleyball and Swimming

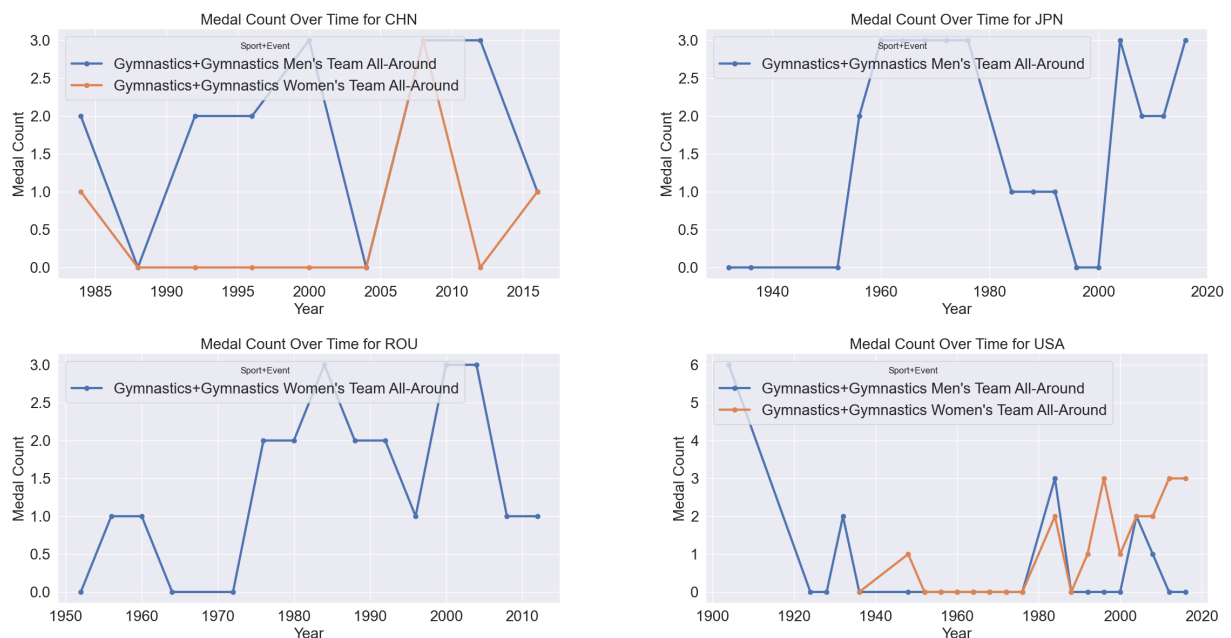


Figure 10: Example of a medal line chart for the Gymnastics event

6.2 Feature and Model Selection

To detect the impact of a coach's addition on medal counts, the medal counts from previous years and the presence of a coach significantly influence the model. Therefore, after experimentation, we selected the following as input parameters: `row['Count']`, `next_year_row['coach'].values[0]`, `next_year_row['Count'].values[0]`, and `next_next_year_row['coach'].values[0]`. We chose `next_next_year_row['Count'].values[0]` as the target variable y .

Given the limited data, choosing a quadratic polynomial would lead to redundant model parameters and overfitting. Therefore, we opted for a first-degree polynomial LinearRegression model for training.

6.3 Model Training and Validation

Taking Karoyli as an example, we selected the women's gymnastics data from ROU and USA, where Karoyli coached, as the original data. We extracted three consecutive sets of data in a loop to input into the model for training, supplemented by other gymnastics data to ensure the training dataset reached an appropriate size.

After training, we evaluated the model using the `staticmodel` library, with the following results:

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.768			
Model:	OLS	Adj. R-squared:	0.731			
Method:	Least Squares	F-statistic:	20.66			
Date:	Mon, 27 Jan 2025	Prob (F-statistic):	1.26e-07			
Time:	12:19:45	Log-Likelihood:	0.73950			
No. Observations:	30	AIC:	8.521			
Df Residuals:	25	BIC:	15.53			
Df Model:	4					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-0.0440	0.087	-0.508	0.616	-0.222	0.134
x1	0.3200	0.116	2.767	0.010	0.082	0.558
x2	0.1040	0.196	0.532	0.600	-0.299	0.507
x3	0.6200	0.116	5.361	0.000	0.382	0.858
x4	0.4200	0.245	1.712	0.099	-0.085	0.925
=====						
Omnibus:	29.546	Durbin-Watson:	1.900			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	62.770			
Skew:	-2.169	Prob(JB):	2.34e-14			
Kurtosis:	8.604	Cond. No.	8.82			
=====						

The model is sensitive to the third and fourth parameters, indicating that the current medal count is significantly related to the previous medal count and the presence of a coach, confirming the existence

of the "Great Coach" effect.

By examining the differences and slopes, we identified the national project combinations with a recent decline in medals, selected the top four, and applied the model to these four combinations to detect the impact of a "Great Coach" joining in 2028 on performance. The results are as follows:

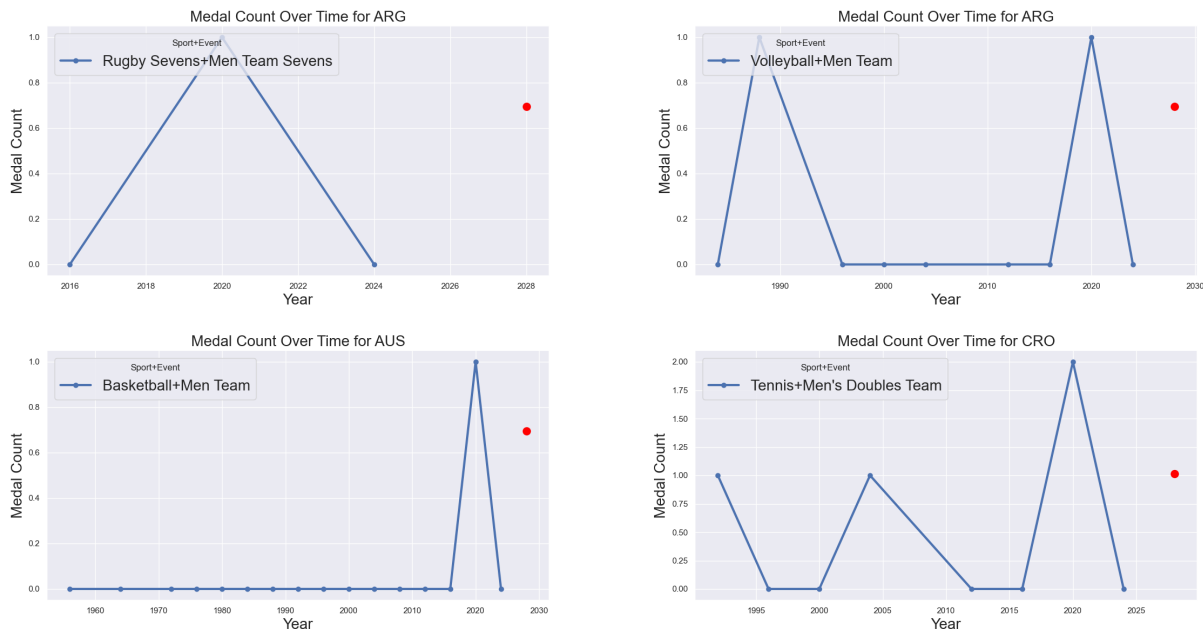


Figure 11: Example of a medal predict chart for the Gymnastics event

Among them, the red points represent the model-predicted medal values. We can observe that when a "great coach" joins, it can lead to a certain degree of increase in the medal values.

7 Conclusion

7.1 Task 1

We firstly constructed a National Sports Level Classification Model and Athlete Evaluation System to help extract effective features from the raw dataset, and then put all the data we summarized into our ARIMA-GB joint model. Measured results: USA, ITA, NED, NZL, etc. will remain basically unchanged in terms of the number of medals (up and down by about 1-3 medals); FRA, AUS, JPN, GER, etc. will see a big increase in the number of medals (up by about 15-20); while GBR, CHN, KOR, etc. will see a more significant decrease in the number of medals (down by about 5-10). With correlation analysis, these sports powerhouses usually excel and dominate in a few specific sports, which is one of the reasons for their consistently high medal counts. And the host country's preference for the selection of events will, to some extent, affect the distribution of events and the distribution of medals.

7.2 Task 2

We concluded through model fitting that great coaches significantly boost medal counts. Then, using indicators like standard deviation, we screened for [Country, Sport, Event] projects that might benefit from a "great coach" to improve performance. We estimated the performance improvement these projects would see with a "great coach" using the fitted model. The conclusion is that after initially introducing a "great coach," all projects see a significant performance boost, especially those that have recently performed well but are showing a downward trend.

8 Discussion

8.1 Model Insights

Our model reveals several original insights about Olympic medal counts: **Economic Factors:** Countries with higher GDP and greater investment in sports infrastructure and coaching staff tend to have more consistent and higher medal counts over time. **Diverse Sports Culture:** Nations with a more diverse sports culture, participating in a wider range of sports, also see an increase in their overall medal tally. **Host Country Advantage:** The model indicates that host countries often experience a boost in medal counts due to the "home-field advantage". **Specific Sports Dominance:** Certain sports, like swimming and athletics, have a higher medal distribution, and countries that excel in these sports tend to have a higher overall medal count. **Coach Investment Return:** The model estimates that investing in a "great coach" can significantly increase medal counts in specific sports, with some sports showing higher returns on coach investment than others.

8.2 Strategic Implications

These insights can inform country Olympic committees in the following ways: **Prioritize Investment:** Committees can prioritize investment in sports infrastructure and coaching staff to ensure consistent and high medal counts. **Encourage Diversity:** By supporting a wider range of sports, committees can foster a diverse sports culture, leading to an increase in overall medal tally. **Leverage Host Advantage:** Committees can strategically plan to maximize the benefits of hosting the Olympics, such as improved facilities and increased public support. **Focus on Dominant Sports:** Committees can focus on sports with higher medal distributions, like swimming and athletics, to maximize their medal potential. **Coach Investment:** Committees can identify sports where coach investment has a high return and target these areas for coach recruitment and retention to boost medal counts.

References

- [1] Wikipedia, Lang Ping, <https://zh.wikipedia-on-ipfs.org/wiki/>
- [2] Wikipedia, Béla Károlyi, <https://en.wikipedia.org/wiki/B>

A AI Use Report

We use Grammarly to help check the grammar error in our report, and DeepL to help translate some academic vocabulary.

We use ChatGPT4 to deal with some coding exceptions, but the whole model is constructed and evaluated by ourselves.

1. **Query 1: What linear regression models are available? Please provide some viable options.**

Output: The following are some common linear regression models and their applicable scenarios:

- **Ordinary Least Squares Regression (OLS Regression):** Suitable for simple cases where there is a linear relationship between the target variable and features, without regularization.
- **Ridge Regression:** Introduces $L2$ regularization to the ordinary least squares method, ideal for situations with many features and multicollinearity.
- **Lasso Regression:** Employs $L1$ regularization, enabling feature selection by driving less important features' coefficients to zero.
- **Elastic Net Regression:** Combines $L1$ and $L2$ regularization, balancing sparsity and handling multicollinearity effectively.
- **Weighted Least Squares Regression (WLS Regression):** Assigns different weights to data points, addressing cases with heterogeneous variance.

2. **Query 2: My feature engineering is quite complex, and I might need nonlinear multi-objective regression. Can you provide some viable models?**

Output: Here are several common nonlinear multi-objective regression models suitable for complex feature and target relationships:

- **Support Vector Regression (SVR with RBF Kernel):** Handles nonlinear problems using the radial basis function (RBF) kernel. Suitable for small to medium-sized datasets.
- **Random Forest Regressor:** A tree-based ensemble method capable of capturing nonlinear relationships and handling high-dimensional data while providing feature importance evaluation.
- **Gradient Boosting Regressor (e.g., XGBoost):** A tree-based ensemble method effective for large datasets and capable of capturing complex nonlinear patterns, supporting multi-objective regression.
- **Neural Networks for Regression:** Useful for high-dimensional data with complex nonlinear relationships, such as Multilayer Perceptron (MLP Regressor).
- **Gaussian Process Regression (GPR):** Provides regression predictions and uncertainty estimates through Gaussian process modeling, ideal for small-sample scenarios.

3. **Query 3: My KMeans elbow function is running too slowly. What are some optimization methods?**

Output: The following strategies can help optimize the KMeans elbow method:

- **Use Mini-Batch KMeans:** The MiniBatchKMeans algorithm, based on small-batch data optimization, significantly reduces computation time while maintaining clustering effectiveness.
- **Reduce the range of cluster counts:** Narrow the range of K , for example, from 2-15 to 3-10.
- **Lower feature dimensions:** Use Principal Component Analysis (PCA) or feature selection to reduce feature dimensions and lower computational complexity.
- **Increase initialization efficiency:** Set `init="k-means++"` to improve the selection of initial cluster centers, reducing the number of iterations needed.
- **Parallelize computations:** Use `n_jobs=-1` to leverage multi-core CPUs for faster computation (applicable to non-Mini-Batch KMeans).
- **Apply sampling techniques:** Perform random sampling on the dataset to reduce the scale of clustering computations.

These optimization methods can be applied individually or combined to significantly improve the runtime of the KMeans elbow method.