Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	**2514461**

Sketch an Olympic Vision

Summary

The Olympic Games is no doubt the most grand and prevalent sporting event on this planet. The medal counts represent the countries' dedication and development in sports and physical wellbeing. Specifically, outstanding coaches have led many teams to great success. However, their presence are typically hidden behind the numbers. In this essay, we mainly discussed the methods of utilizing limited data, predicting the countries' performance in 2028, honoring the coaches and predicting future numbers based on the effect of potential great coaches.

In model I, we predicted the number of events in every discipline at the 2028 Games. We first preprocessed the data, grouping the players by their teams and countries. We chose to analyze the data on disciplines, summing up their events since events in the same discipline are similar and tend to develop simultaneously, also avoiding overgeneralizing to the level of sports. We employed Support Vector Regression (SVR) instead of Linear Regression, which cannot interpret the complex structure of the data. SVR captured the nonlinear patterns in the data, with MSE equals 1.744 and R-squared 0.634, doing nicely for recent years. We also analyzed the importance of different sports for different countries using weighted sums of medals, and visualized it using rose plots.

In Model II, based on the event number predictions in Model I, we predicted the medal numbers in 2028. We first attempted at a plain Long Short Term Memory (LSTM) model. We employed the technique of the Adam optimizer and plateau-detection schedulers. We also added punishment terms, forcing the model to predict equal numbers for all medal types. However, results showed clear underfitting and overfitting. Therefore, we turned to an advanced XGBoost model. In this model, we predicted fractions instead of numbers, and multiplied the fractions with results in model I to yield final medal counts. Noting that 90% of host countries witnessed significant improvement, we averaged the increase as 15%, multiplied to the 2028 host USA, and normalized the results. USA topped the board with 47 golds, 45 silvers and 36 bronzes, China followed with 35, 24, 15, and then Japan with 18, 9 and 10. We evaluated the soundness of this model on year 2024. The 95% confidence intervals were calculated for 9 typical countries, and all predicted results lied in the intervals, proving high reliability.

In Model III, we addressed the cases where a country that had never earned an Olympic medal earned its first at 2028. We employed the Cox Proportional Hazard Model, with its risk function representing the possibility that a country switched from the no-medal status to first-medal status. By fitting the model, normalizing the results, and Monte-Carlo simulations, we predicted the potential number of first-time medalists by 2028. 3 is the most likely with probability 22.73%, and the three most likely first-time medalists are Monaco, Liechtenstein and Myanmar.

In Model IV, we addressed the Great Coach Effect (GCE). We developed a Threshold Method, filtering first-order difference and first-order ratio of score, where scores are calculated by a weighted sum of medal counts. An ensemble-learning anomaly detection method, Isolation Forest, was also employed to cross-validate the choice of weights. We ended up choosing 3, 2 and 1. There was a 89% correspondence between the two methods. We selected the thresholds accordingly. Detection was an OR-logic with the two thresholds. Summing up all years and teams, we measured GCE influence on every sport. We chose USA, CHN and TPE for analysis, and identified the most GCE-influenced sports that they play. We yielded that Taekwondo for USA, Archery for CHN and Swimming for TPE, are their most coach-needing sports. We then used our own ensemble learning by averaging the prediction results of Linear Regression and Random Forest. The data for fitting are all the threshold-activated points for the sport. We predicted that CHN Archery would go from 2 in 2024 to 7 in 2028, USA Taekwondo from 5 to 8, and TPE Swimming from 0 to 9, after hiring great coaches.

Finally, we looked into the male-female score ratio to provide insights on gender-equality since data showed that the stronger countries are more gender-balanced. We visualized our results, showing the predicted medals, the fitted lines, random forests and coach effects. Meanwhile, strengths and weaknesses of the models were analyzed, insights were visualized and discussed, and sensitivity analysis was done.

Keywords: Threshold Method, SVR, XGBoost, LSTM, Mann-Kendall, Cox, Isolation Forest, Linear Regression, Random Forest

Contents

1 Introduction

1.1 Background

Following the conclusion of the 2024 Paris Olympics, nations are turning to the 2028 Los Angeles Games. As a globally celebrated sporting event, the Olympics serves as not only a world stage for athletes to showcase their exceptional sportsmanship, but also a platform for cultural exchange and fostering friendship among nations. Consequently, accurately forecasting a nation's medal count at the Olympics is of paramount importance to sports governing bodies and athletes alike.

Historically, Olympic medal projections have largely relied on the final athlete roster.However, with globalization and the evolution of sports, numerous unforeseen factors, such as the emergence of a great coach or new talent, regional sports culture and home-field advantage can significantly impact medal tallies. Despite these challenges, early predictions are crucial for informed strategic planning and resource allocation within the sporting landscape.

1.2 Literature Review

The Olympic Games is an international sporting event, the medal tally of which has always been a focal point of attention and analysis. There now exists an extensive body of research pertaining to medal and result projection. For example, *Zhang Bo* predicted the gold-medal result of women's shot put in 2012 based on GM(1,1) prediction model in Gray System Theory, a more applicable solution when there is lack of data [**1**]. More methods on the basis of machine learning algorithms were provided by *Jhankar Moolchandani et al.* [**2**] including Linear Regression, Random Forest, Support Vector Machines and Neural Networks. They are more useful for forecasting the medal count according to the athlete's attributes and country information. Among them, Random Forest and SVM stood out. *Noviyanti T M Sagala* and *Muhammad Amien Ibrahim* compared XGBoost, LightGBM and CatBoost and found that XGBoost had the highest accuracy [**3**]. Since Python has numerous libraries that facilitate machine learning tasks, it is convenient for predicting events and medals. XGBoost, which outperforms Random Forest through iterative optimization of trees, can be a reliable choice for our medal ranking prediction.

For the prediction of first medal wins, work by *Nisreen Osman Abdelsalam et al.* on applying Cox proportional hazards models to time-to-event data provided profound inspiration. Similar to survival time, there are numerous factors contributing to the duration of a country going without winning a medal, which can be analyzed by multivariable Cox regression [**4**].

For the problem of Great Coach Effect, we conceptualize differential points to detect the occurrence of the effect. This can be understood through Anomaly Detection. There exist many calculation methods to do so when handling time series data [**5**]. *Yu Qin* and *YuanSheng Lou* used Isolation Forest algorithm to detect the anomaly points of the hydrological time series data, which is more efficient and stable method than traditional ones [**6**].

1.3 Problem Restatement

1. **Data Preparation**

   Preprocessing the 5 datasets about all the history Olympic Games based on appropriate assumptions and decisions for different predictions.

2. **Model Construction**

Building prediction models for medal counts for each country, including prediction of event count and types and first-time medalists.

Also, establishing a model to evaluate Great Coach Effect. Using this quantifier to build another model predicting the results of introducing a coach.

3. **Model Application**

   Applying our models to the history data and making projections for the event count and medal table in the Los Angeles, USA 2028 Olympics. Analyzing countries that will do better or worse and the prevailing sports for different countries, which leads to different possible results due to the varying of the events chosen by the home country and their numbers. Determining how many countries are likely to pick up their first Olympic medal in 2028.

   For Great Coach Effect, choosing three countries and applying our model to find sports where coaching investment pays off. Predicting their performance improvement brought by a coach.

   Offering innovative insights found in our model to country Olympic committees.

4. **Model testing and sensitivity analysis**

   Evaluating the precision and reliability of our models with prediction intervals, probabilities, errors, etc. Performing the sensitivity analysis.

1.4 Our Work

1. For data preparation, we set some basic assumptions to mathematize the real-life problem and conducted data curation, including cleaning, filtering, and establishing relationships between events, disciplines, and sports in different ways for further usage.

2. For model construction, we
   (a) built a SVR model to predict event count, projected medal count from LSTM baseline model to XGBoost model, and established Mann-Kendall-based method for performance change prediction.
   (b) constructed a Cox proportion hazard model for first-time medalists prediction.
   (c) quantified medal score system and Contribution Factor and Influence Factor of coach effect, identified threshold method and Isolation Forest method to capture the evidence of Great Coach Effect and used both Linear Regression and Random Forest to estimate the impact on specific sports.

3. For model application, we applied our models to predict the event count, medal count, number of first-medalists, and the probable outcomes of the three countries we chose to introduce a coach in 2028 Olympic Games. We found some insights offering practical reference based on our model and gender ratio analysis.

4. For model testing and sensitivity analysis, we assessed the precision of our models by mean squared error, R-squared value and prediction intervals. Sensitivity analysis on threshold of coach effect exhibited robustness.

5. For further discussion and evaluation, we discussed the strengths and weaknesses of our models and found room for improvement.

2 Assumptions and Justification

- **A country's medal count is only determined by the trend of its performance in past Olympic Games.**

  Abrupt changes caused by external factors such as sudden events are not taken into account.

- **We attribute all performance improvements exceeding a specific threshold to the coaches.**

  For a country's Olympic team, improvements within a certain range are attributed to individual athlete performance, while those exceeding this threshold are attributed to additional efforts, specifically the introductions of coaches.

- **The probabilities of countries that have yet to earn a medal becoming a first-time medalist in a given year are mutually independent.**

  Each country's performance is influenced by a unique set of factors like athletes and training. Additionally, the outcome of one particular country has little to no bearing on the outcome of another particular country given the multitude of countries participating.

- **The differences between great coaches from various countries can be neglected.**

  When a coach who moves internationally is considered great, his ability is independent to his nationality.

- **The impact pattern of a coach on a specific sport is consistent across different nations.**

  The enhancement of athletic performance due to the introduction of a coach is solely historical, meaning that as long as athletes from different countries have the same baseline, the improvement results will be identical.

3 Notations

Symbol	Definition	Symbol	Definition
N'_E	predicted number of events	ω_G	weight of gold medal
N'_M	predicted total number of medals	ω_S	weight of silver medal
R'_G	predicted ratio of gold medals	ω_B	weight of bronze medal
R'_S	predicted ratio of silver medals	S	score
R'_B	predicted ratio of bronze medals	$\triangle S$	first-order difference of score
P	probability of winning the first medal	γS	first-order ratio of score
X_a	number of athletes	$S_{T\triangle}$	threshold of $\triangle S$
X_p	number of participations	$S_{T\gamma}$	threshold of γS
X_e	number of events	C_i	Contribution Factor at the i^{th} Games
N_G	number of gold medals	I_s	Influence Factor of sport s
N_S	number of silver medals	S_i	score at the i^{th} Games
N_B	number of bronze medals	S_{ij}	score of sport j at the i^{th} Games.

4 Model I : Event Prediction based on SVR

As a preliminary step in predicting the overall medal count at the 2028 USA Olympic Games, we first forecast the number of events using the Support Vector Machine , utilizing its ability of handling complex and nonlinear data.

4.1 Model Overview

In this model, we employed **Support Vector Regression** to predict the number of events in various disciplines at the 2028 Olympic Games based on history event counts. Given the nonlinearity of the time series data, SVR was selected due to its ability to effectively capture nonlinear patterns. By training and tuning the model on the dataset, we successfully forecast the event count for each discipline in 2028. Evaluation results proved outstanding performance, especially for disciplines with relatively stable event counts in recent years. We also calculated medal weighted score S to measure the significance of different sports to each country.

4.2 Model Establishment

Given the varying event number in each Games shown in the provided dataset, regression methods can be employed to forecast the number of events in the 2028 Games. While ARIMA models are commonly used for time series forecasting since they're suitable for linear data, Support Vector Regression (SVR) model offers a more robust approach with its ability to capture nonlinear patterns. [7]

Support Vector Regression (SVR), an extension of Support Vector Machines (SVM) for regression problems, aims to find a function that minimizes the prediction error on the target variable while satisfying certain constraints. The optimization problem for constructing the model is:

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

subject to:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \tag{2}$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \tag{3}$$

$$\xi_i, \xi_i^* \geq 0 \quad \forall i \tag{4}$$

By solving the optimization problem outlined above, the optimal weight vector w and bias b are determined. For non-linear SVR, a kernel function is employed to map the input data into a higher-dimensional feature space, thereby allowing for a linear regression model to be fitted. This can be expressed as:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{5}$$

Due to the obvious nonlinear characteristics of the prediction, we adopt a Gaussian kernel, expressed as:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \tag{6}$$

Subsequent parameter tuning and model training enable the prediction of the event count for the 2028 Olympics.

We conducted regression analysis on the number of past events in `summerOly_programs.csv` to predict the event count in certain discipline for the 2028 Olympics. The modeling steps are shown in Fig. 1. For each discipline, a time series of event counts was constructed, saved as a two-dimensional array. This series was divided into training and testing sets in an 8:2 ratio and subsequently standardized. An SVR model wass instantiated and its hyperparameters tuned via Grid Search coupled with 5-fold Cross-Validation. The trained model was then employed to predict future event counts.

4.3 Results

Taking artistic gymnastics as an example, the SVR model predicted 14 events in 2028, with a **mean squared error(MSE)** of **1.744** and an **R-squared value** of **0.634** as shown in Fig. 2. This relatively high error is due to significant fluctuations in the number of events for this discipline in early years. Despite this, this SVR model performs well when predicting disciplines with a more stable number of events in recent years.

Ultimately, We can roll up the results to determine the overall event count for each sport. They are shown in table 1.

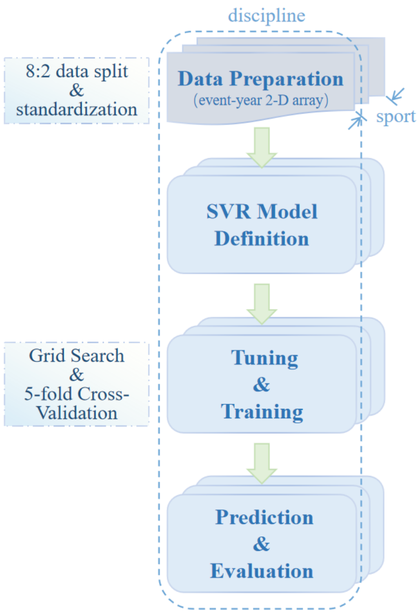

Figure 1: Flow chart of event count prediction

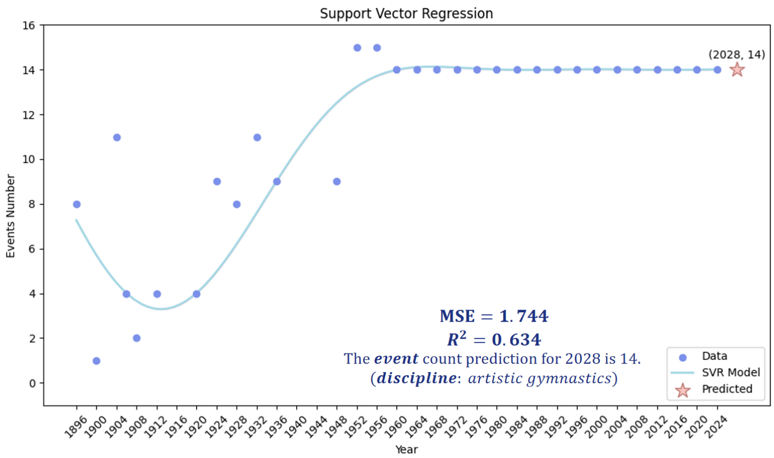

Figure 2: Event count prediction of artistic gymnastics

Table 1: Predicted event count in 2028 Olympics

Sport	Discipline	Code	Discipline Event	Sport Event
Aquatics	Artistic Swimming	SWA	2	49
	Diving	DIV	8	
	Marathon Swimming	OWS	2	
	Swimming	SWM	35	
	Water Polo	WPO	2	
Archery	Archery	ARC	5	5
Athletics	Athletics	ATH	48	48
Badminton	Badminton	BDM	6	6
Baseball and Softball	Baseball	BSB	1	2
	Softball	SBL	1	
Basketball	3x3	BK3	1	3
	Basketball	BKB	2	
Basque Pelota	Basque Pelota	PEL	0	0
Boxing	Boxing	BOX	13	13
Breaking	Breaking	BKG	0	0
Canoeing	Sprint	CSP	11	16
	Slalom	CSL	5	
Cricket	Cricket	CKT	0	0
Croquet	Croquet	CQT	0	0
Cycling	BMX Freestyle	BMF	1	21
	BMX Racing	BMX	2	
	Mountain Bike	MTB	2	
	Road	CRD	4	
	Track	CTR	12	
Equestrian	Dressage	EDR	2	6
	Eventing	EVE	2	
	Jumping	EJP	2	
	Vaulting	EVL	0	
Fencing	Fencing	FEN	11	11
Field hockey	Field hockey	HOC	2	2
Flag football	Flag football	AFB	0	0
Football	Football	FBL	2	2
Golf	Golf	GLF	1	1
Gymnastics	Artistic	GAR	14	18
	Rhythmic	GRY	2	
	Trampoline	GTR	2	
Handball	Indoor	HBL	2	2
	Field	HBL	0	
Jeu de Paume	Jeu de Paume		0	0
Judo	Judo	JUD	16	16
Karate	Karate	KTE	0	0
Lacrosse	Sixes	LAX	0	0
	Field	LAX	0	
Modern Pentathlon		MPN	2	2
Polo	Polo	POL	0	0
Rackets	Rackets	RQT	0	0
Roque	Roque		0	0
Rowing	Coastal	ROC	0	14
	Rowing	ROW	14	
Rugby	Sevens	RU7	1	0
	Union	RUG	0	
Sailing	Sailing	SAL	10	10
Shooting	Shooting	SHO	15	15
Skateboarding	Skateboarding	SKB	0	0
Sport Climbing	Sport Climbing	CLB	2	2
Squash	Squash	SQU	0	0
Surfing	Surfing	SRF	0	0
Table Tennis	Table Tennis	TTE	5	5
Taekwondo	Taekwondo	TKW	8	8
Tennis	Tennis	TEN	6	6
Triathlon	Triathlon	TRI	3	3
Tug of War	Tug of War	TOW	0	0
Volleyball	Beach	VBV	2	4
	Indoor	VVO	2	
Water Motorsports		PBT	0	0
Weightlifting	Weightlifting	WLF	16	16
Wrestling	Freestyle	WRF	12	18
	Greco-Roman	WRG	6	

4.4 Analysis of the importance of different sports

Firstly, we quantified a country's performance in a specific sport at a certain Games as score S, which is the sum of gold, silver and bronze medal counts N_G, N_S and N_B assigned with different weights ω_G, ω_S and ω_B.

$$S = N_G\omega_G + N_S\omega_S + N_B\omega_B \tag{7}$$

For the sake of simplicity, we let

$$\omega_G = 3, \quad \omega_S = 2, \quad \omega_B = 1. \tag{8}$$

We calculated the proportion of scores in various sports for different countries in 2024 relative to their total scores, to measure the significance of different sports to each country. We

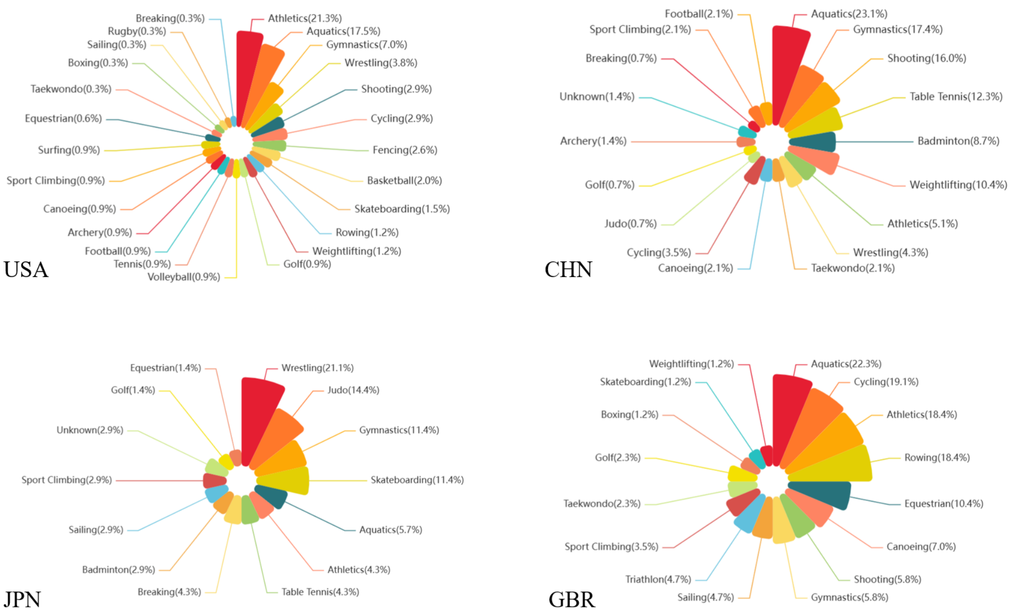

Figure 3: The score percentage of various sports for different countries at the 2024 Olympic Games

assumed that the higher the proportion of a sport in a country, the more important that sport is to that country.

From Fig. 3, it can be seen that for the United States, Athletics and Aquatics have the largest proportions and are the most significant; for China, Japan, and the United Kingdom, Aquatics, Wrestling, and Aquatics are the most important, respectively.

5 Model II : Medal Prediction based on LSTM and XGBoost

5.1 Model Overview

To capture the intricate relationship between a country's medal count and the specific disciplines, we first tallied the gold, silver and bronze medals won by **each country** in **each discipline** every year, then divided by the total number of medals awarded in that discipline for that year to obtain the country's medal winning percentage. We then used the ratios from previous years to train an **LSTM** model as baseline and conducting a more accurate model based on **XGBoost algorithm**. The results we predicted were the percentages of medals won by each country out of the total medals awarded in each discipline in 2028. The predicted number of gold, silver, and bronze medals can be calculated by multiplying them by the predicted event numbers and suming up the products. Furthermore, we calculated the **95% confidence intervals** for the predicted medal counts by using residuals from the XGBoost model and the properties of the normal distribution. Analysis based on **Mann-Kendall test** forecast whether a country would do better or worse depending on its performance of the past 5 Games.

5.2 Data Preparation

A country's medal count is closely tied to its proficiency in various sports. Accounting for the relationship between events and medals, we proposed predicting the probability of a country winning a medal in each discipline and then multiplying it by the number of events so that we could provide a more granular and realistic analysis prediction.

First, we tried every means to categorize each event within dataset `summerOly_athletes.csv` under its respective discipline. Subsequently, the gold, silver, and bronze medal counts for each country within each discipline were calculated and converted into percentages, representing the likelihood of each country achieving correlative medal success. This approach considered the overall performance within each discipline while also acknowledging the significant differences between disciplines within the same sport (e.g., Swimming and Diving under Aquatics), thus avoiding overgeneralization.

5.3 Baseline: LSTM model

5.3.1 Model Overview

In this section, a baseline LSTM model was built and trained using Pytorch. The model took as input all the gold / silver / bronze medal counts of all countries before the 2024 Games, and was fitted to the medal counts in 2024. For the loss function, instead of the typically-used MSE loss, we added a punishment term, punishing deviations of the sum of medal counts of all countries for a certain discipline from the discipline's known event count. In the predict step, all past years' data were fed into the model to predict the medal counts for all countries in all disciplines. Summing up the disciplines, we yielded the medal count for all countries.

5.3.2 Training Process

The training process used the **Adam** optimizer with the following parameters: learning rate $\text{lr}_{\text{initial}} = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The **One Cycle Learning Rate Scheduler** was applied with a maximum learning rate max_lr $= 0.001$ for a total of 100,000 epochs. The learning rate was gradually increased and then decreased in a cyclical manner during training.

5.3.3 Results

The final loss dropped to 0.0058, indicating good convergence, however, plain LSTM yielded poor results, as indicated in the following two figures. Fig. 4 repeated the medal count in 2024, and is a clear overfit, while Fig. 5 predicts that China will earn 8 Olympic Gold medals in Table Tennis in 2028, while we know that there are only 5, and China can earn at most 5 (as it did in 2024). Therefore, we must take into consideration varying in the number of events held in each discipline, and avoid overfitting. Also, we should consider host effect. All of these will be addressed in our advanced XGBoost Model in the following section.

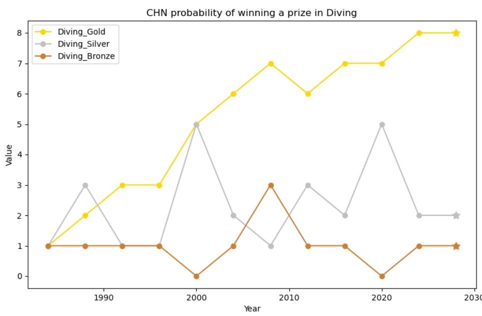

Figure 4: China's Diving, Overfitting

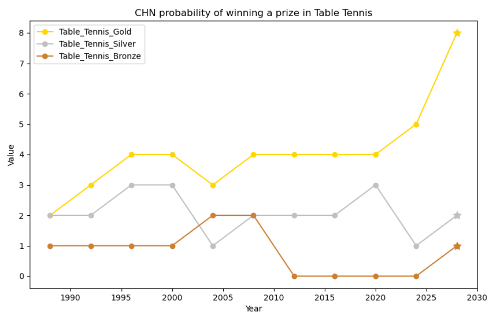

Figure 5: China's Table Tennis, Underfitting

5.4 Advanced: XGBoost model

5.4.1 Model Establishment

Since we had got all the history medal winning ratios, we then employed regression analysis using XGBoost model, combined with cross-validation and randomized search optimization.

XGBoost, an ensemble learning algorithm based on gradient boosting, is renowned for its efficiency in making predictions and generating rankings on small to medium-sized structured datasets. The inclusion of 5-fold cross-validation and grid search optimization can further enhance the model's performance and ensures robust results. [**3**]

XGBoost is an efficient implementation of the gradient boosting algorithm that iteratively adds decision trees to optimize an objective function composed of a loss function for prediction errors and a regularization term for model complexity. In the t^{th} iteration, the model attempts to learn a decision tree f_t that fits the residuals $r_i^{(t-1)} = y_i - \hat{y}_i^{(t-1)}$ from the previous iteration, where y_i is the true value and $\hat{y}_i^{(t-1)}$ is the predicted value from the previous iteration. The objective function is defined as:

$$\text{Obj} = \sum_{i=1}^{n} l(r_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)) \tag{9}$$

Here, l is the loss function, and $\Omega(f_t)$ is the regularization term, which penalizes large leaf node weights and an excessive number of leaf nodes to prevent overfitting. In this way, XGBoost can control the complexity of the model while maintaining its performance.

We first collected statistics on the number of gold, silver and bronze medals each country won in each discipline every year, then divided by the total number of medals awarded in that discipline for that year to obtain the medal ratio of that country. By running an XGBoost model with the ratios from past years, we predicted medal ratios in 2028.

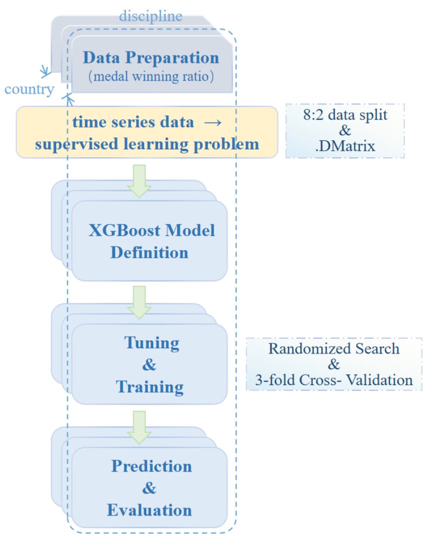

Figure 6: Flow chart of medal count based on XGBoost

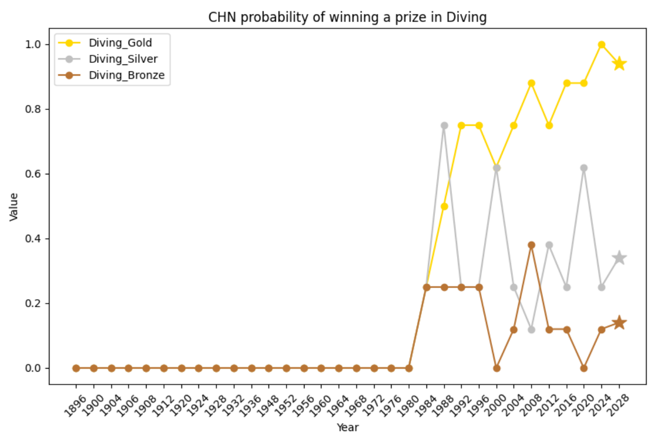

Figure 7: Probability prediction of Chinese diving gold and Silver MEDALS in 2028 Olympic Games

Based on predictions, we took into account the host country effect, where athletes may perform better when competing in their own country due to familiarity with the environment, reduced travel fatigue, easier adjustment to time zones and climate, and other factors. We calculated the score increases of the host countries in recent Olympic Games, measuring the host country effect by the ratio of this Game's scores to last Game's scores.

We observed that approximately 90% of countries exhibit an improvement in their performance during the Olympics hosted in their own nation. Specifically, 77% of these countries have seen an increase exceeding 20%, and 44% have experienced an increase of more than 50%. However, the countries with increases over 50% tend to be those with relatively fewer points (the Soviet Union is excluded from this analysis as it has since transitioned into Russia and other countries). Therefore, it is unreasonable to anticipate a substantial increase for the United States, a major scoring country and the host of the 2028 Olympics. Considering that the U.S. has previously observed a decrease of 7% in scores when hosting, we deemed it reasonable to predict that the U.S. will have a hosting nation increase between 0% and 20% for the year 2024. In our predictive model, we incorporated a scoring increase of 15%.

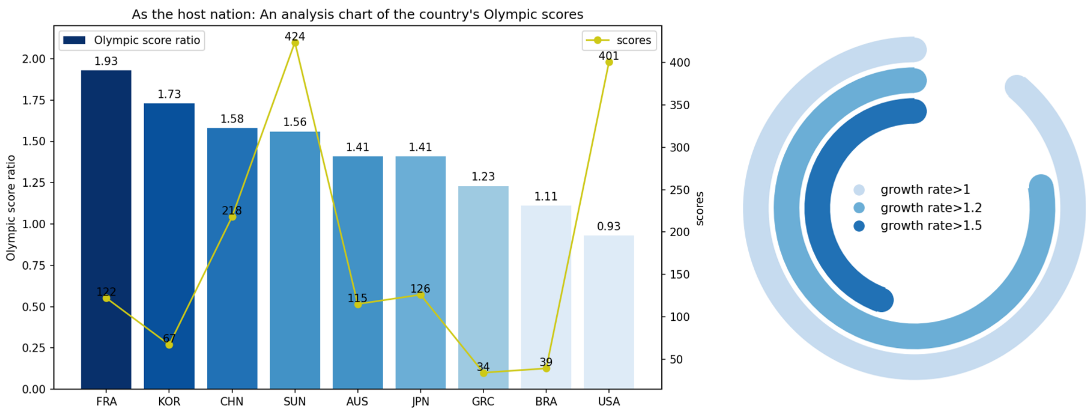

Figure 8: Analysis of the advantages of the host country. The light yellow score indicates the score when the country was the host country, and the score ratio indicates the score when the country is the host country divided by the score of the previous Olympics Game.

After obtaining the medal winning ratios for each country in 2028 using the XGBoost model and considering the host country effect, we normalized these ratios and multiplied them by the number of events in each discipline as predicted by the SVR model discussed in the fourth section. This process ultimately yielded the predicted medal count, which we then displayed in a medal ranking table as Fig.10 shows.

5.5 Examining the confidence of the prediction interval

Since the XGBoost model can only return predicted values and not directly provide confidence intervals, we can use the predicted values given by XGBoost and the real number of medals to calculate the standard deviation of the residuals between the predicted and real values in the first step. In the second step, we can estimate the prediction interval by calculating the confidence interval using the predicted and actual values. Below is a detailed explanation of the approach.

We used the sampling distribution theory in statistics to estimate the confidence intervals of prediction results. Due to the randomness of sampling, different samples may produce different

2028 USA summer Olympics

Country	1	2	3	Total
USA	43	41	32	116
CHN	36	25	15	76
JPN	19	9	10	38
GBR	16	21	22	59
AUS	15	11	16	42
FRA	13	16	14	43
GER	11	11	12	34
ITA	11	9	15	35
KOR	9	7	6	22
NED	9	8	10	27

Figure 9: Top 10 of predicted 2028 medal table

2028 USA summer Olympics

Country	1	2	3	Total
USA	47	45	36	128
CHN	35	24	15	74
JPN	18	9	10	37
GBR	16	21	21	58
AUS	14	11	16	41
FRA	13	16	14	43
ITA	11	9	15	35
GER	11	11	12	34
NED	9	8	10	27
KOR	9	7	6	22

Figure 10: Top 10 of predicted 2028 medal table with host country effect

statistics. The confidence interval provides a range within which we can say with certainty that the population parameter is included, usually at a 95% level. For this problem, we predicted the number of gold medals for each country in 2024, and compared it with the actual gold medal results of each country in 2024, we can calculate the mean and standard deviation of the residuals of the gold medal results in 2024.

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{10}$$

$$\sigma_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2} \tag{11}$$

Assuming that the residuals e_i follows a normal distribution $N(\bar{e}, \sigma_e^2)$, we can use the properties of the normal distribution to calculate the prediction intervals at different confidence levels. For a given confidence level α (for example, 95%), the corresponding z-score $z_{\alpha/2}$ can be found (for a 95% confidence level, $z_{0.025} \approx 1.96$). For each predicted value $\hat{y}_i$, its prediction interval can be expressed as:

$$\left(\hat{y}_i \,-\, z_{\alpha/2}\cdot\sigma_e,\ \hat{y}_i \,+\, z_{\alpha/2}\cdot\sigma_e\right) \tag{12}$$

As shown in the Fig. 11, the pink dots represent the estimates of the number of gold medals for the top 9 countries in the 2024 Olympics, predicted by the XGBoost model with France as the host country in 2024. These are the predicted values for the number of gold medals for each country in 2024. Since we already knew the actual number of gold medals for each country in 2024, we can calculate the standard deviation of the residuals between the predicted and actual values. By combining the Z-score for a 95% confidence interval, we can calculate the upper and lower bounds of the 95% confidence interval for the prediction interval, as shown by the lavender legend in the figure.

5.6 A Mann-Kendall-based analysis to project significant change in medal acquisition

Mann-Kendall test, a non-parametric method for identifying trends in time series data, is used to analyze the performance of each country in the past 5 Olympic Games [**9**]. The aim is to forecast which nations will exhibit substantial fluctuations in their medal tallies.

For a specific country's time series of Olympic medal counts $X = \{x_1, x_2, ..., x_n\}$, Mann-

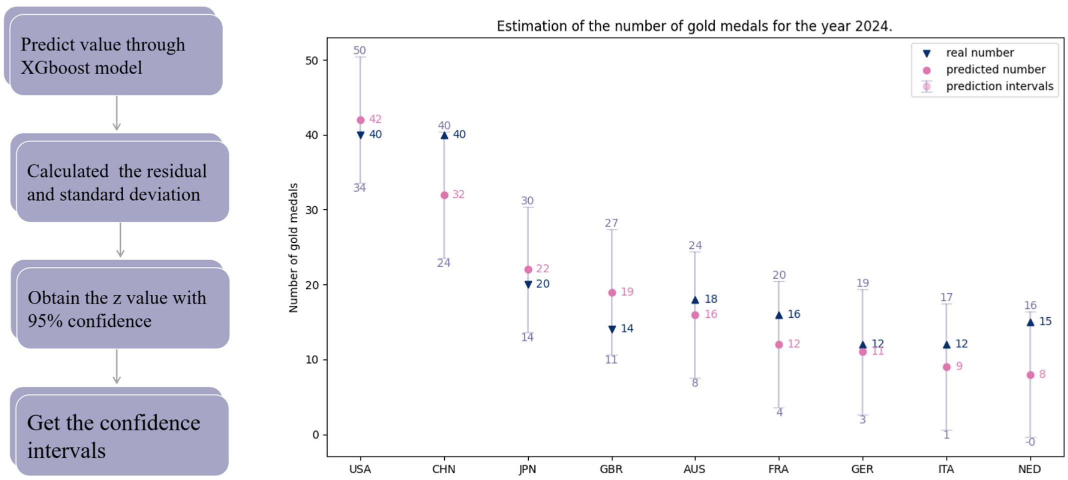

Figure 11: Taking the 2024 gold medal predictions as an example, the confidence interval of the XGBoost model is analyzed.

Kendall test first calculate the S statistic:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_j - x_i) \tag{13}$$

where sgn() is the sign function.

When there are ties in the data (i.e., repeated values), the variance of the S statistic needs to be adjusted. The variance of S is calculated as:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{m} t_i(t_i-1)(2t_i+5)}{18} \tag{14}$$

where m is the number of unique values in the data and t_i is the number of times the i^{th} unique value appears.

To standardize the S statistic, a Z statistic is calculated as:

$$Z = \frac{[S - E(S)]}{\sqrt{Var(S)}} \tag{15}$$

where $E(S)$ is the expected value of S. For a trend-free time series, $E(S)$ is typically close to 0. Then, assuming that the z-value follows a standard normal distribution, we can perform a p-parameter test on the z-value that follows a standard normal distribution to identify data with a p-value less than 0.05, which allows us to determine whether there is a significant upward or downward trend.

$$\text{if } p < 0.05 \begin{cases} \text{may improve} & \text{if } (z > 0) \\ \text{may worse} & \text{if } (z < 0) \end{cases}$$

Based on the analysis of the past twenty years, that is, the performance of the past five Olympic Games, we analyzed whether these countries would have a significantly better or worse performance in the 2028 Olympic Games. Since we considered that data from 20 years ago has less impact on the next Olympic Games than data from the past 20 years, which can better reflect the trend of a country's Olympic level in recent years.

We converted the gold, silver, and bronze medals of the countries in the past five Olympic Games into scores. Then we used the Mann-Kendall model to analyze the upward and downward trends of the countries. Finally, it is analyzed that four countries have a significant trend of improvement: NZL, ITA, NED, SWE, and one country has a significant downward trend, which is BLR.

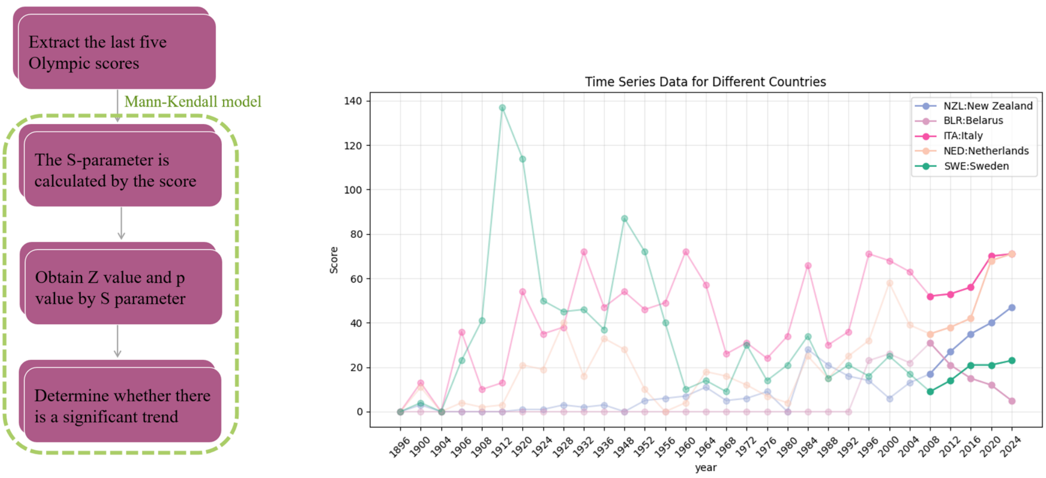

Figure 12: The scores of countries with obvious upward or downward trend in the past 5 Olympic Games

6 Model III: First-time Medalist Prediction based on Cox Proportional Hazard Model

6.1 Model Overview

This part utilizes a **Cox proportional hazards model** to predict the probability of a country winning its first Olympic medal. Key predictors taken into account include the number of athletes developed, Olympic participation history, and the diversity of sports culture. The model treats a country's journey towards its first medal as a survival process, excluding data from countries that have already won medals. By fitting the model and conducting Monte Carlo simulations, we predicted the potential number of first-time medalists by 2028.

6.2 Theoretical Basis

The objective of Cox proportional hazard model is to estimate the hazard function, which represents the instantaneous risk of an event occurring at a specific time, given a set of predictor variables. It incorporates the proportional hazards assumption, which posits that the hazard ratios between any two individuals remain constant over time. Here, we can conceptualize probability for countries to picking up their very first medal as a measure of risk. The model utilizes a semi-parametric approach, modeling the baseline hazard function non-parametrically while estimating the effects of predictor variables through coefficients in a linear predictor. This approach allows for flexibility in capturing complex relationships between covariates and the risk of the event.

6.3 Key Predictors of First-time Medalists

For countries that have yet to earn medals, their chance to make a landmark moment can be derived according to the features listed as below:

- **Quantity of athletes developed** The total number of athletes a country has sent to the Games through history is a strong indicator of its sports development levels, reflecting the depth and breadth of its sports talent pool.
- **Olympic participation history** How many times a nation has participated in the Olympics reflects its engagement with the global sporting community and experience in competing with other teams in such a grand sport event.
- **Diversity of sports culture** The number of event categories a country has participated in over the past Games showcases the abundance of coaching resources and how rich its sports culture is, also contributing to its medal prospects.

6.4 Model establishment

The moment a country makes its first breakthrough and graduates from the Olympic Village for newcomers represents the time when our event of interest occurs, similar to a risk event studied in clinical medicine. Based on the hazard function, we defined the probability for a country to win the first medal at the Games in year t as $P(t)$. Given the covariates we have discussed above, a Cox model can be established to estimate the probability of a country achieving its first medal win at a specific time point.

$$P(t, X) = P_0(t)e^{\beta_a X_a(t)+\beta_p X_p(t)+\beta_e X_e(t)} \tag{16}$$

$X_a(t)$ is the total number of **unique athletes** registered for Olympics from the first Games in 1896 to the Games before year t. $X_p(t)$ is the number of the country's participations in Olympics. $X_e(t)$ is the total number of **unique events** the country participated in from the first Games to the Games before year t. $P_0(t)$ is the baseline hazard function reflecting the risk of an event occurring in the absence of other risk factors, namely the probability for a country that has never competed in a single Olympic Games to win the first medal in year t.

6.5 Model Fitting and Prediction

The function $P(t, X)$ aims to model the trend in first-medal-winning probability for countries and then predict this probability for a certain country based on its characteristics in a certain year. Therefore, the data we used for fitting is the feature vector $X = [X_a(t), X_p(t), X_e(t)]$ of each country up to and including the year of its first medal win, namely the process of the country stays in *Olympic Village for newcomers*, rather than including the data after a country wins its first medal. So we marked countries that had already won medals in each Games and conduct a survival analysis, treating Year as duration time and winning as the event. After excluding the data of countries with prior wins, we fitted a model and successfully forecast the likelihood of first-time winning for various countries in a given year such as 2028.

6.6 Monte Carlo Simulation

Since we assumed independence between countries' probabilities of obtaining their first medal, if we conducted n repeated Olympic trials in year t and Y represents the number of times this country wins a medal for the first time, then Y will follow a binomial distribution.

$$Y(t) \sim (n(t),\ P(t)) \tag{17}$$

To simulate potential outcomes, we applied Monte Carlo simulations based on the predicted probabilities. After 100 million trials, we obtained the following distribution of the possible number of first-time medalists by 2028.

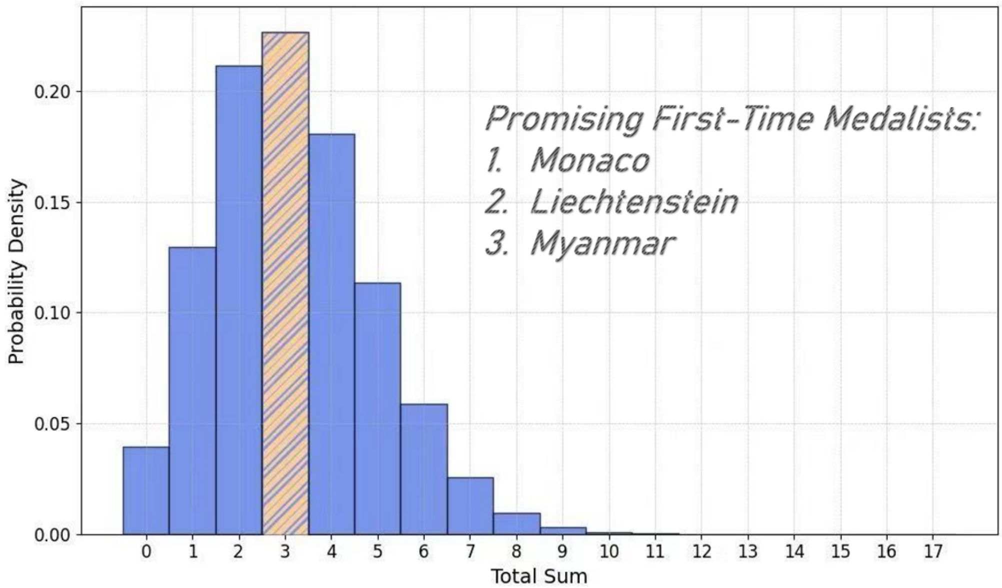

Figure 13: The distribution of the possible number of 2028 first-time medalists

As Fig. 13 shows, with a probability of 0.2273, 3 nations are projected to earn their first medal in 2028. The countries with the highest probability of achieving their first-ever medals in 2028 are Monaco (MON), Liechtenstein (LIE), and Myanmar (MYA) respectively.

Table 3 provides a breakthrough probability ranking of countries that have never won a medal. The right column is the corresponding probability after exponential normalization.

7 Model IV: Great Coach Effect

7.1 Model Overview

This section quantifies the impact of coaches on a country's Olympic performance. By developing a scoring system and applying a threshold method and Isolation Forest, we identified time points where significant performance improvements occur due to the introduction of a coach, namely differential points. The contribution and influence of coaches on specific countries and sports are quantified as factors to reflect the impact of coaches directly. These serve as the basis for us to select sports benefiting most from coaching for chosen countries. By applying both Linear Regression and Random Forest to handling the data of differential points we captured, we predicted the prospects of introducing coaches for specific sports in specific countries.

Table 3: Extrapolated and Exponential Values for top 20 first-time medalists in 2028

NOC	Extrapolated Values	Exponential Values
MON	0.6995	0.0503
LIE	0.6434	0.0476
MYA	0.6297	0.0469
MLT	0.6215	0.0465
BOL	0.5682	0.0441
MLI	0.5519	0.0434
NEP	0.5464	0.0432
CHA	0.5452	0.0431
LBR	0.5343	0.0427
BIZ	0.5309	0.0425
CGO	0.5203	0.0421
AND	0.5163	0.0419
MAD	0.5066	0.0415
LES	0.5045	0.0414
GUI	0.4990	0.0412
BEN	0.4957	0.0410
NCA	0.4946	0.0410
SLE	0.4851	0.0406
CAY	0.4747	0.0402
CAF	0.4733	0.0401

It's worth noting that **only** in this part, **sport** is essentially a **discipline**, given the significant differences between various disciplines in practice and the fact that a coach often specializes in a single discipline. While the analysis at this level shares the same principles as at the sport level, it will be more accurate and detailed.

7.2 Effect Evidence

A coach plays a pivotal role in a country's Olympic performance. If a country introduces a coach in a specific sport, it is highly likely that the country's performance in that will see a significant improvement in the following years. For instance, in China in 2003, after introducing Coach Guoliang Liu to table tennis, the country's performance improved markedly. Therefore, we can estimate the impact of coaches (Great Coach Effect, GCE) by examining the changes in performance over time. We can identify **differential points** as those where we conclude that a coaching effect took place.

7.2.1 Threshold Method

The first-order difference and first-order ratio of score S are defined as

$$\triangle S = S_i - S_{i-1} \tag{18}$$

$$\gamma S = \frac{S_i}{S_{i-1}} \tag{19}$$

The larger $\triangle S$ and γS are, the greater progress the country have made in i^{th} Olympics. The first method to determine whether GCE has happened is to set a threshold value $S_{T\triangle}$ for the first-order difference and $S_{T\gamma}$ for the first-order ratio of scores between consecutive Olympic Games and compare the actual $\triangle S$ and γS with them.

$$GCE = \begin{cases} 1 & \text{if } \triangle S \geq S_T \mid\mid \gamma S \geq S_{T\gamma}, \\ 0 & \text{else.} \end{cases} \tag{20}$$

Points with a GCE value of 1 are defined as **differential points**.

These threshold values are established to minimize the variance in a country's total score due to the exceptional performance by certain individuals and consider the progress of small nations. A $\triangle S$ or γS exceeding the threshold indicates a statistically significant improvement in the country's overall performance, suggesting a substantial coaching effect.

7.2.2 Isolation Forest

We also employed the Isolation Forest method to capture the differences in Score S.

Isolation Forest is an ensemble-based method for anomaly detection. It identifies outliers by randomly selecting features and feature values to build multiple isolation trees, each of which isolates data points by randomly choosing features and split points. The anomaly score for a data point is typically given by the negative exponential of its average path length in the isolation tree, i.e., $\text{score}(x) = -2^{-\frac{E(h(x))}{c}}$, where $E(h(x))$ is the average path length of data point x in the isolation tree, and c is the average path length of the tree. The shorter the path length, the higher the anomaly score, making it easier to identify outliers.

For instance, if country's differential values over the years are represented as $[X_1, X_2, \ldots, X_n]$, where X_n is the largest, then X_n will be closest to the root node in the isolation tree. By extracting these points, we can identify some of the larger differential points in Score.

7.3 GCE contribution to medal counts

To measure the contribution of great coaches to a country's performance at the i^{th} Games, we defined a contribution factor C_i.

$$C_i = \frac{\sum_{j \in sports} \alpha_{ij} \triangle S_{ij}}{\sum_{j \in sports} S_{ij}} \tag{21}$$

If $\triangle S_{ij} > S_T,\ \alpha_{ij} = 1; else\ \alpha_{ij} = 0.$

This factor quantifies how much the great coaches of a whole country contributes to its medal counts. It's different from country to country. The data we calculated are listed in table 4 in order.

7.4 GCE influence on certain sports

To measure the influence of a coach to any country's performance in a certain sport s, we defined an influence factor I_s.

$$I_s = \frac{\sum_{k \in years} \sum_{countries \triangle S_{sk} > S_T} \alpha_{sk} \triangle S_{sk}}{\sum_{k \in years} \sum_{countries \triangle S_{sk} > S_T} S_{sk}} \tag{22}$$

If $\triangle S_{sk} > S_T,\ \alpha_{sk} = 1; else\ \alpha_{sk} = 0$. This factor mirror the relationship between coach and sports. It's different from sport to sport. Because the mobility of coaches is global, we assumed the I_s is the same for each country. We ranked I_s in table 5.

Table 4: NOC and their corresponding Contribution Factor

NOC	C_{2024}
VIE	1.0000
SGP	1.0000
CIV	1.0000
TTO	1.0000
ISL	1.0000
MAR	0.9318
CHI	0.9167
IRL	0.8635
POR	0.8590
SMR	0.8333
ANZ	0.8182
EST	0.8034
BOT	0.8000
BRN	0.8000
BAH	0.7894
MEX	0.7843
MGL	0.7500
NGR	0.7159
PHI	0.6875
URU	0.6778
ETH	0.6691
SLO	0.6666
HKG	0.6666
ESP	0.6611
DOM	0.6538
RSA	0.6427
ECU	0.6306
KEN	0.6266
PAK	0.6242
CRO	0.6306

Table 5: Sports and their corresponding Influence Factor

Sport	I_s
Cycling Mountain Bike	1.0000
3x3 Basketball	1.0000
Cycling BMX Racing	1.0000
Cycling BMX Freestyle	1.0000
Golf	0.8690
Surfing	0.8333
Triathlon	0.8161
Sport Climbing	0.8125
Trampolining	0.8077
Tug-Of-War	0.8068
Canoe Slalom	0.8065
Modern Pentathlon	0.8063
Cycling Road	0.8000
Art Competitions	0.7850
Beach Volleyball	0.7692
Handball	0.7590
Trampoline Gymnastics	0.7500
Football	0.7399
Water Polo	0.7358
Tennis	0.7324
Softball	0.7238
Volleyball	0.7211
Hockey	0.7124
Rhythmic Gymnastics	0.7063
Sailing	0.7058
Baseball	0.6989
Archery	0.6987
Artistic Swimming	0.6970
Shooting	0.6814
Taekwondo	0.6777

7.5 Application and Estimation

7.5.1 choosing countries and sports

In order to specify the sports where coaching investment would yield the better returns for them, we ranked sports based on their coaching influence factor I_s to identify those most susceptible to coaching effects. Then, we determined whether their historical performance in each sport falls within a promising range(S_{min}, S_{max}), which means they have enough potential and room for improvement both. If so, we would consider introducing a great coach.

7.5.2 Great Coach Effect Prediction

Our goal is to build a model to forecast the impact of introducing a coach in a particular sport. By extracting differential data points that exhibit a coaching effect and plotting the previous Game's scores against the subsequent Game's scores, we discovered a discernible pattern.

So we employed both Regression Analysis and Random Forest to fit a function, which we denoted as F, where the previous Game's score S_{i-1} is the independent variable and the following Game's score S_i is the dependent variable.

$$F(S_{i-1}) = S_i \tag{23}$$

F reflects the pattern of score improvement under Great Coach Effect. Once the sport and the previous score S_{i-1} are determined, the following year's score S_i can be predicted regardless of the country or year.

7.5.3 Results

We choose CHN, USA and TPE for case study and recommend 2 sports for each of them in table 6. The rank is the I_s rank of corresponding sports.

Table 6: Sports choosen for specific countries

NOC	**Sport**	I_s **Rank**
CHN	Sailing	26
	Archery	28
USA	Canoe Slalom	12
	Taekwondo	31
TPE	Surfing	7
	Artistic Swimming	29

By inputting each country's score S_{2024} of this sport into the fitted models (linear regression and random forest respectively), we can predict its winning outcomes S_{2028} if they introduce a great coach. The results are shown as follows:

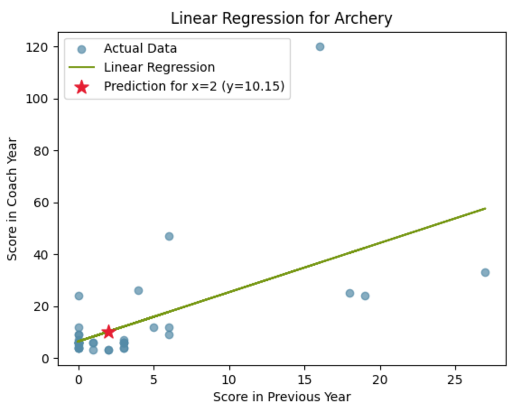

Figure 14: CHN Archery S prediction -Linear Regression

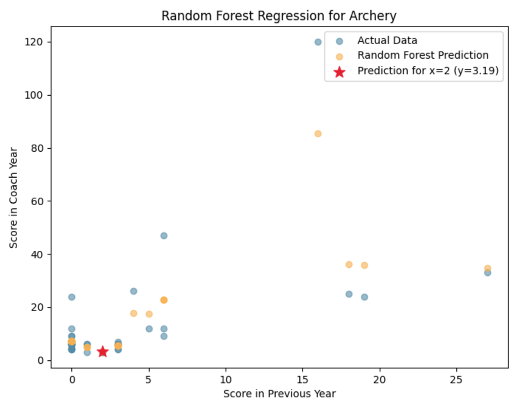

Figure 15: CHN Archery S prediction -Random Forest

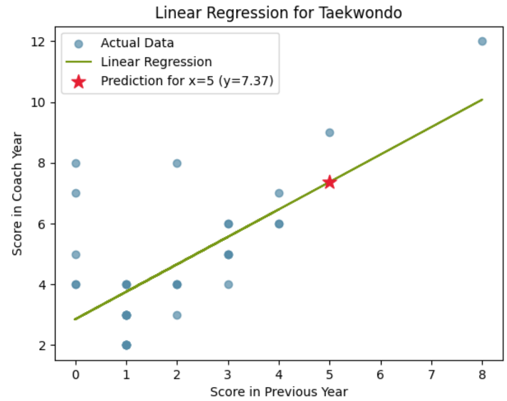

Figure 16: USA Taekwondo S prediction -Linear Regression

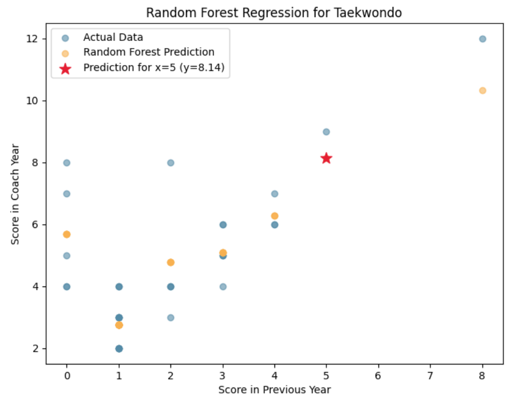

Figure 17: USA Taekwondo S prediction -Random Forest

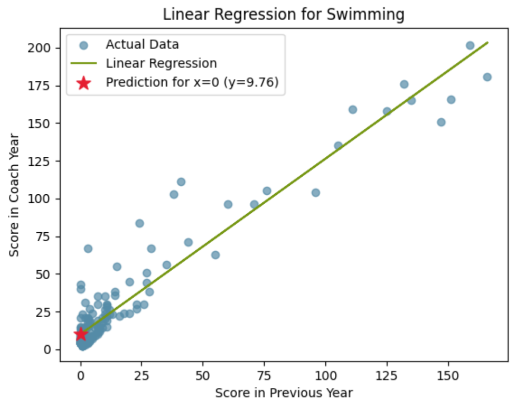

Figure 18: TPE Swimming S prediction -Linear Regression

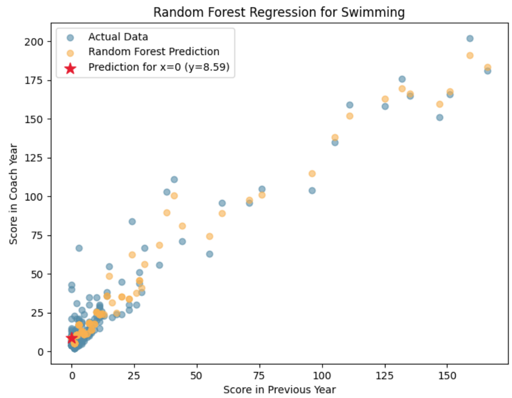

Figure 19: TPE Swimming S prediction -Random Forest

By averaging the two results, we obtained the predicted score S_{2028}.In the case of CHN archery, the score increases from 2 to 7. For USA Taekwondo, the score increases from 5 to 8, and for TPE swimming, it increases from 0 to 9.

8 Insights

8.1 From Models Above

- We have successfully predicted the percentage of medals each country will win in each discipline based solely on historical data. This can serve as a reference for each country to set breakthrough goals. The extent to which the number of medals changes under the influence of the host country's venue has been predicted, which can inspire each country's Olympic Committee to determine the extent to which they should enhance their training to adapt to the venue.

- For countries that have not won any medals, we have provided the probability of achieving a breakthrough in the next Olympic Games, which can to some extent boost the confidence of countries with a high probability of winning medals, as they have provided a breakthrough trajectory from zero to one.

- In the study of GCE, we provided the Olympic Committees with a list of sports that would most easily improved with the help of coaches through impact factors. It is important to select coaches for introduction based on the specific circumstances of different countries. Blindly choosing to improve weaknesses may not yield good returns. In addition, our fitted prediction model can also map out a blueprint for their progress.

8.2 Extra Study: Impact of Gender Ratio

We analyzed the proportion of women's points in the Olympics from 1986 to 2024. It was found that from 1896 to 2014, the proportion of women's points in the Olympics continued to increase, from 5% in 1986 to nearly 50% in 2024, reflecting the increasingly important role of female athletes in international Olympic competitions.

We then analyzed countries that performed well at the Olympics, such as the United States and China, and found that the proportion of points scored by female athletes in their Olympic scores was close to 60%, higher than the world average. This indicates that they place great emphasis on the role of female athletes in the overall Olympic performance of their countries.

We also analyzed Japan and the UK, two countries with decent Olympic performances, and found that the proportion of points scored by female athletes in their Olympic scores was slightly below 50%. For CUB and UZB, two countries with less than ideal Olympic performances, the proportion of female athletes was only 30%. Therefore, our suggestion is to pay more attention to and cultivate female athletes, which can help to further explore the potential for better national performance in Olympic events.

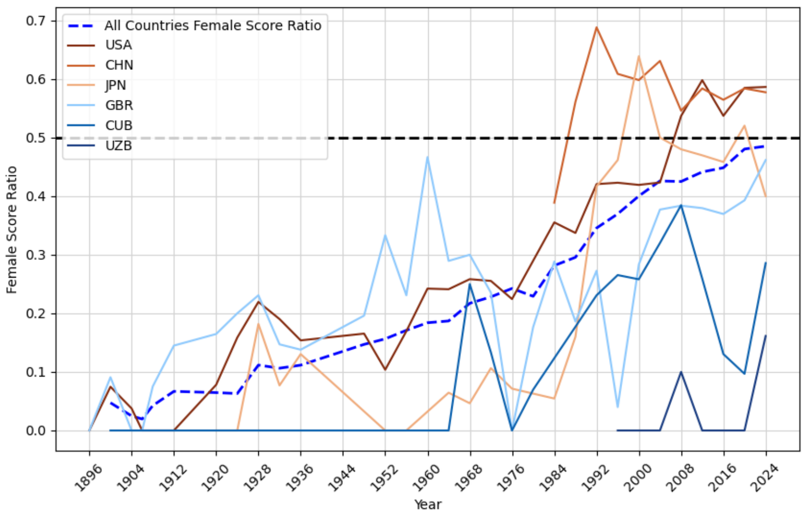

Figure 20: Relationship between predicted score after hiring a coach and the thresholds

9 Sensitivity Analysis

We performed sensitivity analysis for the two thresholds, namely the score difference threshold and the score increase rate threshold, which we calculated and employed to detect where great coaches played an important role in sports.

We yield the following key observations:

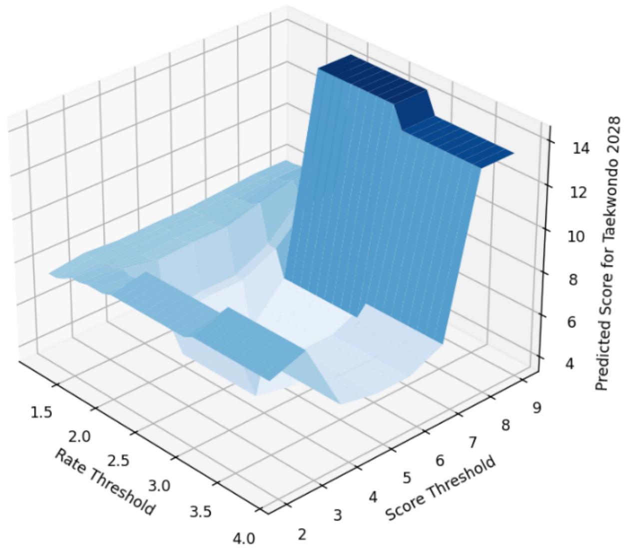

Figure 21: Relationship between predicted score after hiring a coach and the thresholds

- As the *rate threshold* increases, there is a slight increase in the predicted score, especially when the *score threshold* is low. Nevertheless, our model is relatively robust to rate threshold.
- The *score threshold* has a significant impact on the predicted score, especially when set to higher values. A sharp increase in the predicted score is observed as the score threshold approaches higher values, particularly above 7. This suggests that the model is highly sensitive to changes in the score threshold at this range. However, we set rate threshold to 1.2 in the calculations, where the model is not that sensitive to the score threshold.
- The shape of the 3D surface agreed to our intuitions, that is, as the threshold is increased, "greater" coaches remain, and these coaches would help the athletes win more medals, and also the more precious ones, leading to a greater increase in predicted scores.
- Overall, given that one Gold medal would contribute 3 to the score, our model is rather stable in reasonable ranges of thresholds, which is, $1.2 \leq$ rate-threshold ≤ 4, and $3 \leq$ score-threshold ≤ 7.

10 Strength and Weakness

10.1 Strength

- **Reasoning**. Our model is constructed from a logical foundation, allowing us to provide more granular details such as the medal count of each country in different disciplines and the winning probability of each first-time medalists, enhancing the accuracy and reliability of our overall predictions.
- **Comprehensiveness**. We took many factors into account like the impact of the host country effect and the association between event count and medal count. In addition, we finished comparative analysis and found better model.
- **Stability**. Our model demonstrates robust stability when thresholds are set within the optimal range, highly adaptable to potential changes in various history data.

10.2 Weakness

- **Idealization**. Despite the fact that we have already conducted a detailed analysis of datasets and fitted a prediction model relying solely on historical by using relatively advanced methods,a few, such as GDPs, population resources and geographical locations were still neglected, leading to the limitations of our model.

10.3 Room for Improvement

- There are more details worth exploring such as the growth rate of performance. Extra information about economic development and population can be used to optimize the model for more practical use.

References

[1] Z. Bo, Q. Chaoling, X. Xiaoli and Z. Fanbo, "GM (1,1) Model Gray Prediction for the Gold-Medal Result of Women's Put Shot in the 30th Olympic Games," 2011 International Conference on Future Computer Science and Education, Xi'an, China, 2011.

[2] J. Moolchandani, V. Chole, S. Sahu, R. Kumar, A. Shukla and A. Kumar, "Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics," 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2024.

[3] N. T. M. Sagala and M. Amien Ibrahim, "A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal," 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, 2022.

[4] N. O. Abdelsalam, M. K. Elbashir and S. M. SaadEldeen, "Applying Cox Regression in Time to Event Data," 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2018.

[5] Z. Zhao, Y. Zhang, X. Zhu and J. Zuo, "Research on Time Series Anomaly Detection Algorithm and Application," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019.

[6] Y. Qin and Y. Lou, "Hydrological Time Series Anomaly Pattern Detection based on Isolation Forest," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019.

[7] Poongadan, S., Lineesh, M.C. Non-linear Time Series Prediction using Improved CEEMDAN, SVD and LSTM. Neural Process Lett 56, 164 (2024).

[8] Kamal, Neel and Sanjay Pachauri. "Mann-Kendall Test - A Novel Approach for Statistical Trend Analysis." International Journal of Computer Trends and Technology (2018): n. pag.

Appendices

Country	Gold	Silver	Bronze	Total	Country	Gold	Silver	Bronze	Total
USA	47	45	36	128	COL	0	4	0	4
CHN	35	24	15	74	KAZ	0	1	3	4
GBR	16	21	21	58	RSA	1	2	1	4
FRA	13	16	14	43	TUR	0	1	3	4
AUS	14	11	16	41	ECU	1	1	1	3
JPN	18	9	10	37	GDR	1	1	1	3
ITA	11	9	15	35	GRE	1	0	2	3
GER	11	11	12	34	INA	1	0	2	3
NED	9	8	10	27	IND	0	1	2	3
KOR	9	7	6	22	IRL	2	0	1	3
ROC	6	10	6	22	MEX	0	1	2	3
CAN	6	6	9	21	THA	1	1	1	3
HUN	6	8	5	19	ARM	0	2	0	2
NZL	7	5	4	16	AZE	0	1	1	2
BRA	3	4	5	12	BLR	0	1	1	2
ESP	3	5	4	12	ETH	0	1	1	2
SWE	3	5	3	11	FIN	0	0	2	2
POL	2	4	4	10	FRG	0	1	1	2
SUI	2	3	5	10	HKG	1	0	1	2
CUB	4	2	3	9	ISR	1	1	0	2
UKR	1	4	4	9	SLO	2	0	0	2
DEN	2	3	3	8	AIN	0	1	0	1
KEN	3	2	3	8	ARG	0	0	1	1
ROU	2	4	1	7	DOM	0	1	0	1
BEL	2	1	3	6	EGY	0	0	1	1
CZE	3	1	2	6	GRN	0	0	1	1
IRI	2	3	1	6	KGZ	0	1	0	1
TPE	2	1	3	6	MAS	0	0	1	1
UZB	4	1	1	6	MDA	0	0	1	1
BUL	2	1	2	5	MGL	0	1	0	1
CRO	2	2	1	5	PHI	1	0	0	1
GEO	2	3	0	5	POR	0	1	0	1
JAM	2	1	2	5	PRK	0	0	1	1
NOR	3	1	1	5	RUS	0	0	1	1
SRB	2	1	2	5	UGA	1	0	0	1
URS	2	1	2	5	VEN	0	1	0	1
AUT	1	1	2	4	others	0	0	0	0

Table 7: All the countries that are predicted to earn medals in 2028 USA Olympic Games based on XGBoost model (considering home country effect

Report on Use of AI

1. OpenAI ChatGPT (Jan 27, 2025 version, ChatGPT-4)

Query1: Is there a contradiction between satisfying the proportional hazards assumption and the independence of two individuals?

Output: The proportional hazards assumption and the independence of individuals are different concepts and are not contradictory. Let's explain these two concepts separately:

In the Cox regression model, the proportional hazards assumption requires that the hazard ratios between different individuals remain constant over the entire observation period. This means that, relative to a baseline individual, the hazard ratio (i.e., the rate at which an event occurs) between different individuals is fixed and does not depend on time. This assumption does not require the independence of individuals. In other words, even if there are dependencies between individuals, the proportional hazards assumption can still hold.

Independence of Individuals: Independence of individuals typically refers to the assumption that the survival time of one individual is not related to the survival time of another individual. For example, the survival time of two individuals does not influence each other, and the occurrence of an event for one individual does not alter the probability of the event occurring for another individual. In the Cox regression model, individuals in the dataset are typically assumed to be independent, meaning that the occurrence of an event for one individual is independent of others, even though their survival times may be affected by some common risk factors.

Relationship between the Two: The proportional hazards assumption and the independence of individuals are not mutually exclusive. In the Cox regression model, the assumption is that the hazard ratio for each individual is fixed, but this does not require complete independence between individuals. In practice, individuals may share common factors (e.g., environment, treatment, genetic factors) that influence their survival risk, but as long as the influence of these factors remains consistent and the proportional hazards assumption holds, complete independence between individuals is not necessary.

If there is some form of dependency between individuals in the data (e.g., family members or individuals in the same experimental group), caution is needed in modeling. In such cases, adjustments may be required, such as using more complex models like group or cluster analysis methods to account for these dependencies.

Therefore, the proportional hazards assumption and the independence of individuals do not conflict. The Cox regression model mainly relies on the proportional hazards assumption, which requires that the hazard ratios within the same group remain constant, while the independence between individuals depends on the actual data situation.

Query2: Summarize the article A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal.

Output: The article "A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal" explores various boosting algorithms used to predict the likelihood of countries winning Olympic medals. The main points of the article are:

Boosting Algorithms Overview: The study compares different boosting algorithms, including AdaBoost, Gradient Boosting, and XGBoost, highlighting their ability to improve prediction accuracy by combining weak learners (models that perform slightly better than random guessing) into a stronger, more robust model.

Data and Features: The authors use historical Olympic data, such as athletes' performance, country-specific features, and external factors like economic conditions and political stability, to train the models. They preprocess the data by handling missing values and normalizing features to enhance model performance.

Model Evaluation: The article assesses the performance of each boosting algorithm using various evaluation metrics such as accuracy, precision, recall, and F1-score. The results show that XGBoost outperforms other algorithms in terms of prediction accuracy, likely due to its advanced regularization techniques and tree-based model optimizations.

Key Findings: XGBoost was found to be the most effective algorithm for predicting Olympic medal outcomes, while AdaBoost showed slightly lower performance. The study emphasizes the importance of tuning hyperparameters and cross-validation for boosting models to avoid overfitting and improve generalizability.

In conclusion, the article demonstrates the effectiveness of boosting algorithms, particularly XGBoost, in predicting Olympic medal outcomes based on a combination of performance data and country-related features.

Query3: Is the Basis of Monte Carlo Simulation the Binomial Distribution?

Output: The basis of Monte Carlo simulation is not specifically the binomial distribution. It is a statistical method that uses random sampling to simulate and compute the solutions to complex problems. Monte Carlo simulations can use various probability distributions, including the binomial distribution, but it is not limited to it.

Basic Concept of Monte Carlo Simulation Monte Carlo simulation is a technique used to estimate the statistical properties of complex problems or systems by generating a large number of random samples. The core idea of this method is to use a computer to generate random data and approximate solutions to real-world problems based on that data. Its foundation does not rely on any specific probability distribution but instead utilizes various distributions to simulate the behavior of systems.

Binomial Distribution and Monte Carlo Simulation The binomial distribution can serve as one possible random distribution in Monte Carlo simulation, especially when simulating Bernoulli trials (e.g., events of success or failure). For example, if we needed to simulate multiple independent binomial trials (such as multiple coin tosses), we can use the binomial distribution to model the random variables.

However, the strength of Monte Carlo simulation lies in its flexibility. It is not confined to a single distribution. Any suitable probability distribution can be chosen for the simulation, and common distributions include:

Uniform Distribution: When each outcome has an equal probability. Normal Distribution: When the variation of the variables follows a normal distribution (e.g., stock price movements). **Exponential Distribution:** Used to model the time intervals between events (e.g., the arrival time of phone calls).

Poisson Distribution: Used to model the number of events happening within a fixed period (e.g., machine failures).

Binomial Distribution: Used to model a fixed number of Bernoulli trials (e.g., the result of tossing a coin). Steps in Monte Carlo Simulation

Define the Model and Objective: First, define the random process that needs to be simulated and the objective to be estimated. For example, this might involve calculating the expected value, probability, etc., of a complex problem.

Choose the Appropriate Probability Distribution: Based on the nature of the problem, select an appropriate probability distribution to generate the random variables. For instance, if simulating a coin toss, the binomial distribution might be used to model the result of each toss.

Generate Random Samples: Use a computer to generate a large number of random samples, each following the selected probability distribution.

Run the Simulation: Run multiple simulations based on the generated random samples and perform the necessary calculations. Each simulation represents a possible state of the system.

Analyze the Results: Perform statistical analysis on the simulation results, typically calculating averages, variances, probabilities, etc., to estimate the performance of the model or solve the problem.

Applications of Monte Carlo Simulation Monte Carlo simulation can be widely applied in various fields, such as:

Financial Modeling: Estimating stock prices, option pricing, etc. Physical Simulations: Modeling the movement of particles in a physical field.

Engineering Design: Assessing the reliability of system designs. Operations Research: Optimizing decision-making processes in complex systems. Risk Assessment: Evaluating the risks of decisions under uncertainty.