Problem Chosen C | **2025 MCM/ICM Summary Sheet** | **Team Control Number** 2523765

The Long River of Competition: A Multi-Dimensional Coupling Analysis Framework for a Century of the Olympics

Summary

The distribution of Olympic medals has been a key focus in sports research. Accurate prediction of medal counts is crucial for strategic planning and resource allocation. This paper proposes a **multidimensional coupling analysis model** combining modern statistical methods and machine learning techniques to uncover patterns and predict future medal counts, with a particular focus on improving prediction accuracy and model interpretability.

First, for the **interval prediction** task, we cleaned the raw data, addressing **outliers** and **structural issues**. We then identified key **multidimensional features** through **visualization techniques**, revealing the impact of **economic development** and **host country effects** on medal counts. We also introduce the **Knee-NSGA2-XGBoost model**, which combines **multi-objective optimization** with **XGBoost** to solve the trade-off between **prediction coverage** and **interval prediction**. The model improves prediction accuracy by introducing a **knee point selection mechanism** that overcomes limitations in traditional optimization methods. Experimental results showed the model's superiority, with the United States, China, and the United Kingdom ranking in the top three for the 2028 Los Angeles Olympics.

To improve **interpretability** and **transparency**, we used **SHAP (Shapley Additive Explanations)** analysis, identifying the importance of features like **host country effects** and **gender ratio of athletes**. This also helped identify **feature interactions**, providing valuable insights for further model optimization.

For the **probability prediction task** of countries winning their first Olympic medal, we treated it as an extreme event **binary classification task**. Using a **support vector machine (SVM)** model with **Focal Loss**, we addressed **small sample sizes** and **data imbalance**. This led to outstanding performance, identifying **eight** countries likely to win their first Olympic medal at the 2028 Games.

In the task of identifying **great coaches** and developing **investment strategies**, we used the **PELT (Pruned Exact Linear Time)** algorithm to identify potential great coaches and employed **LightGBM** for prediction. Through this process, we successfully identified **six potential great coaches**, including **Lang Ping** of China's women's volleyball team. Based on these predictions, we recommended specific teams for future coaching hires, helping guide strategic decisions in sports development.

Finally, **robustness testing** of the **Knee-NSGA2-XGBoost model** demonstrated its stability and reliability under various conditions, validating its **feasibility** and **credibility** in practical applications for Olympic medal prediction, offering valuable insights for future strategic planning.

Keywords: Olympic Medal Prediction; Multi-Objective Optimization; Knee-NSGA2-XGBoost Model; Explainable Machine Learning; Multidimensional features

Contents

1 Introduction

1.1 Problem Background

Throughout the long history of competitive sports, the Olympic Games have been hailed as the most brilliant gem. Since the revival of the modern Olympic Games in 1896, this globally renowned event has undergone more than a century of remarkable progress. In recent years, the competitive allure of the Olympics has become increasingly captivating, with the medal table undoubtedly being the most prominent focal point. The medal table is often viewed as a key indicator of national strength and international prestige. However, accurately predicting the number of Olympic medals for each country remains a challenging task. This is not only due to the non-linear distribution of medals but also because the outcomes are influenced by various scales and complex interacting factors. Therefore, a deeper understanding of how multi-scale attributes affect medal distribution is crucial for comprehending and forecasting the Olympic medal landscape.

1.2 Restatement of the Problem

We aim to develop a model that, based on past medal distributions, host country information, event settings, and athlete data, thoroughly analyzes the potential impacts of various factors on medal distribution, and ultimately forecasts the number of medals in future Olympic Games. Given the complexity and background of this issue, our focus will be on the following aspects:

- Developing a prediction model to estimate the number of medals for each country in the Olympics and evaluating its performance for the 2028 Games.
- Building a model focused on countries that have won medals for the first time, extracting potential patterns to predict the likelihood of countries that have not won medals yet obtaining medals in the upcoming Los Angeles Olympics, and conducting a country-specific medal prediction analysis for particular events and athletes.
- Identifying the impact of "great coaches" in historical data, evaluating their contributions to medal counts, and further analyzing which countries should invest in great coaches for specific events to increase their medal counts.
- Providing unique insights into the Olympic medal landscape and explaining these findings.

1.3 Literature Review

In the field of medal prediction, it is widely acknowledged that relying solely on past medal data as the primary input for statistical models is insufficient for accurately predicting the total number of medals in the Olympics. Bernard proposed that factors such as economic conditions and population size should be considered to capture and determine the final distribution of medals [1]. With the advancement of statistical techniques and machine learning, some researchers have attempted to characterize the distribution patterns of medals using collected data. For instance, Forrest developed a statistical regression model using GDP and population density as covariates [2]; Schlembach introduced a socio-economic model based on a two-stage random forest to predict Olympic medal distributions accurately during the pandemic [3]; Csurilla analyzed three major countries and identified the key factors for Olympic success, offering insights for countries that have yet to win medals [4].

However, there are some limitations to the aforementioned studies:

- These studies typically analyze only a single year or country, neglecting the spatiotemporal information between years and countries in Olympic data;
- The studies lack a multidimensional consideration of the factors affecting medal counts, which is crucial for deconstructing the spatiotemporal coupling characteristics of medal counts and improving prediction accuracy;
- Despite using machine learning methods, the analysis of model interpretability is often lacking, which hinders the understanding and application of these models by the Olympic Committee and related personnel;
- Very few studies address the modeling of first-time medalists, a topic that is often one of the highlights of the Olympic Games.

1.4 Our Work

In response to the analysis above, we propose using modern statistical methods and machine learning techniques to construct a data-driven multidimensional coupling analysis model. This model will deeply analyze historical Olympic data, revealing potential patterns in the development of the Olympics. Specifically, our work can be summarized as shown in Figure 1

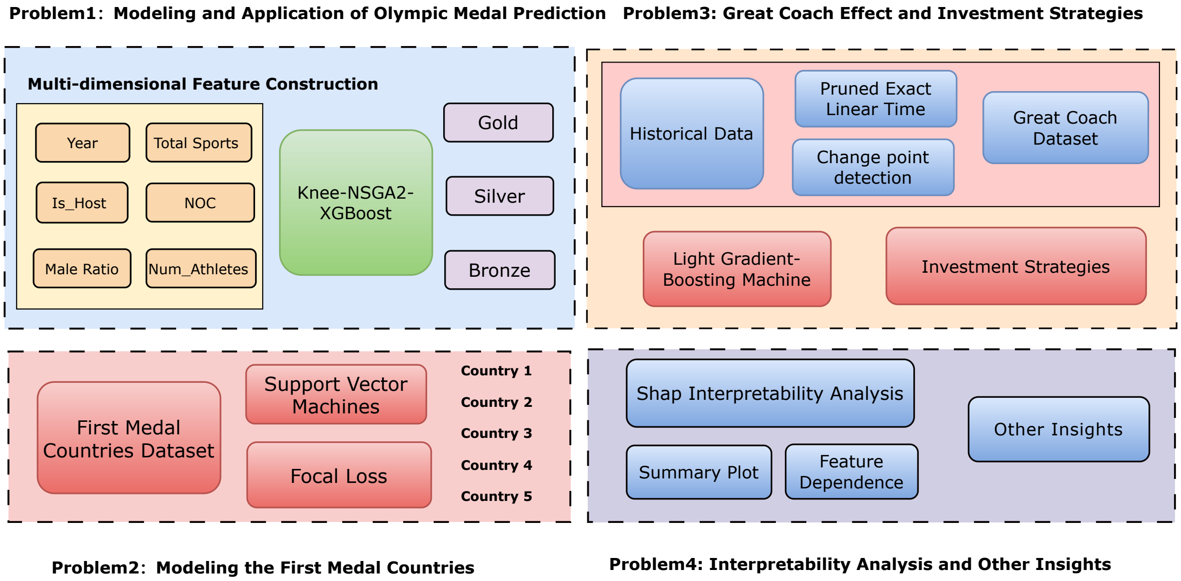

Problem1: Modeling and Application of Olympic Medal Prediction
Multi-dimensional Feature Construction
Year
Total Sports
Is_Host
NOC
Male Ratio
Num_Athletes
Knee-NSGA2-XGBoost
Gold
Silver
Bronze
Problem3: Great Coach Effect and Investment Strategies
Historical Data
Pruned Exact Linear Time
Change point detection
Great Coach Dataset
Light Gradient-Boosting Machine
Investment Strategies
First Medal Countries Dataset
Support Vector Machines
Focal Loss
Country 1
Country 2
Country 3
Country 4
Country 5
Problem2: Modeling the First Medal Countries
Shap Interpretability Analysis
Summary Plot
Feature Dependence
Other Insights
Problem4: Interpretability Analysis and Other Insights

Figure 1: Overview of the proposed framework

- For Problem 1, we first organize multidimensional feature information from a large volume of Olympic data. Then, we improve the XGBoost model using a knee-enhanced Pareto optimization mechanism, which ensures prediction coverage while reducing the width of prediction intervals, thus providing more accurate uncertainty analysis. Finally, we analyze the model's interpretability using Shap values to identify key features.
- For Problem 2, we first determine the conditions for first-time medalists and construct a medal-winning dataset based on the multidimensional information from Problem 1. To address the issue of class imbalance, we adopt an improved cross-entropy loss function to guide

the support vector machine in capturing nonlinear features. Finally, we visualize the probability distribution of first-time medalists for the Los Angeles Olympics, along with the dependency relationships between countries in specific events and athletes.
- For Problem 3, we first apply the Pelt change point detection algorithm to identify potential “great coaches” and establish relevant datasets. Then, we calculate a "great coach index" using multidimensional information and analyze its potential patterns using the LightGBM model.
- For Problem 4, the model will provide new perspectives to help national Olympic committees better understand medal predictions and offer valuable insights for strategic decision-making.

2 Assumptions and Notations

2.1 Assumptions and Explanations

In reality, Olympic medal prediction analysis is complex, and reasonable assumptions must be made to simplify the model. Each assumption is followed by its corresponding explanation:

- **Assumption 1:** The number of medals a country wins is closely related to various multidimensional factors, including the country’s economic strength, sports development level, the number of participating athletes, age and gender structure, the types and number of events, etc. These factors can be used to infer the number of medals, and such inference patterns have historical continuity.
  - **Explanation:** This assumption is based on the correlation between a country’s medal count and specific factors, which serves as the foundation and premise for modeling.
- **Assumption 2:** After the reunification of East and West Germany, both are considered as Germany, and after the dissolution of the Soviet Union, its successor is Russia. This assumption is commonly adopted in historical data.
  - **Explanation:** Since data for the 2028 Olympics has not been fully disclosed, it is reasonable to infer this assumption based on historical data inertia.
- **Assumption 3:** The participation data for the 2028 Olympics can be inferred and adjusted appropriately by analyzing the 2022 and 2020 data and considering the relevant information published by various countries.
  - **Explanation:** Since 2028 data has not been fully disclosed, it is feasible to make predictions based on historical data inertia.
- **Assumption 4:** New events added to the 2028 Olympics will not be included in the model’s consideration.
  - **Explanation:** There is insufficient historical data for the new events, and the model cannot effectively capture their characteristics, leading to uncertainty in the prediction results.

In order to simplify the analysis of each section, additional assumptions have been made. These assumptions will be discussed at appropriate points in the following sections.

2.2 Notations

Table **??** shows the necessary symbols and notations used in this paper. Other symbols and notations will be defined when used.

Table 1: Notations for Knee-NSGA2-XGBoost Optimization Algorithm

Notation	Description
P	Population size
G	Number of generations
θ	Knee threshold
$H = \{\eta, \gamma, \lambda, \text{max\_depth}, n\_estimators\}$	XGBoost hyperparameter space
$p_{crossover}$	Crossover rate for NSGA2
$p_{mutation}$	Mutation rate for NSGA2
k	Tournament size for NSGA2
D_{train}	Training dataset
D_{val}	Validation dataset
$(1-\alpha)$	Confidence level for prediction interval
h_i	Individual in the population
M_i	XGBoost model trained with hyperparameters h_i
$\hat{y}_{lower}, \hat{y}_{upper}$	Lower and upper bounds of the prediction interval
f_1	Coverage objective
f_2	Width objective
N	Number of data points in validation set
P_g	Population at generation g
Q_g	Mating pool at generation g
O_g	Offspring population at generation g
$F_1, F_2, \ldots$	Pareto fronts after non-dominated sorting
ϕ_i	Angle between consecutive points in Pareto front
h^*	Knee point, optimal solution in Pareto front
ε	Neighborhood size for knee selection
P_{g+1}	New population selected for next generation
F_1^{final}	Final Pareto front from last generation
$PI(\cdot \mid h^*)$	Prediction interval generator with optimal h^*

3 Data Processing and Feature Engineering

3.1 Data Pre-processing

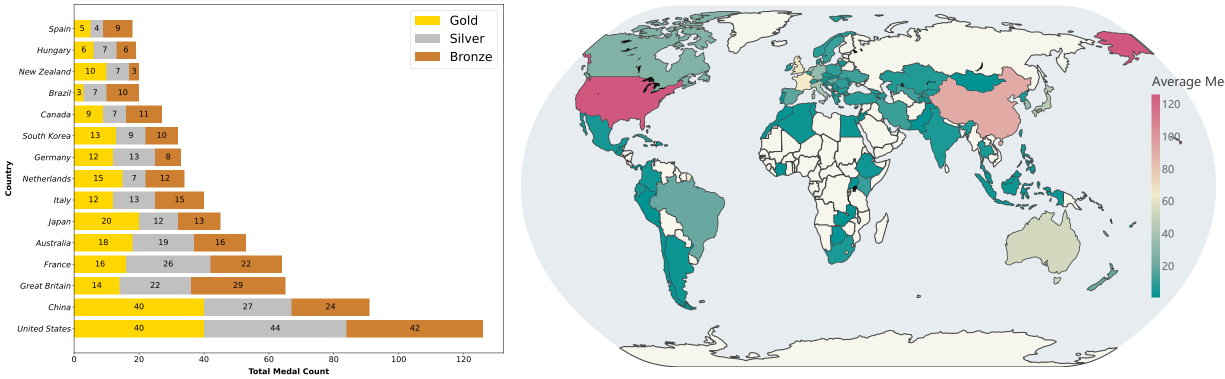
First, we carefully examined the provided data and found that the summerOly_programs dataset contained 49 missing values and included textual data. According to the actual situation, we filled these missing values with 0, indicating that the event was not held in the corresponding year. Next, we noticed that the summerOly_athletes dataset contained 1466 duplicate entries, which were removed after processing. These data preprocessing steps help improve the data quality and the accuracy of the analysis.

3.2 Feature Engineering

3.2.1 Data Characteristics

Before performing feature engineering, it is necessary to analyze the medal distribution. Taking the 2024 Paris Olympics as an example, we plotted the medal distribution for the top 15 countries, as shown in Figure 2a. The United States leads in gold, silver, and bronze medals, while China is on par with the US in gold medals. The competition for bronze medals is more balanced, while gold medals are mainly concentrated among major powers. Focusing on this aspect will help improve prediction accuracy.

On the other hand, we plotted the global medal distribution heatmap for 2024, as shown in Figure 2b. This heatmap effectively displays the geographical distribution of countries that won medals during the Paris Olympics. From the map, it can be observed that medal winners are concentrated in regions with higher economic strength, such as North America, East Asia, Western Europe, and Australia. In Africa, medal-winning countries are primarily located along the coasts, which may be closely related to the infrastructure development, sports development level, and national economic strength of these regions. Therefore, considering geographical location and regional characteristics in subsequent medal prediction analysis will help improve the model's accuracy and predictive capability.



(a) 2024 Top 15 Medal Countries

(b) 2024 Global Medal Distribution Heatmap

Figure 2: Paris 2024 Olympic medals

3.2.2 Host Country Effect

The host country typically performs better in the medal tally during the Olympic Games due to factors such as home advantage, national image and motivation, and financial support. This phenomenon is defined as the host country effect. Capturing the influence of the host country effect on medal distribution helps in selecting key features during the modeling process. To observe this effect, we plotted Figure 3. The results show that when a country is the host, the diagonal elements, i.e., the number of medals, are significantly higher compared to normal conditions.

Multi-dimensional Feature Construction We have established a multi-dimensional feature dataset, as referenced in Table 2.

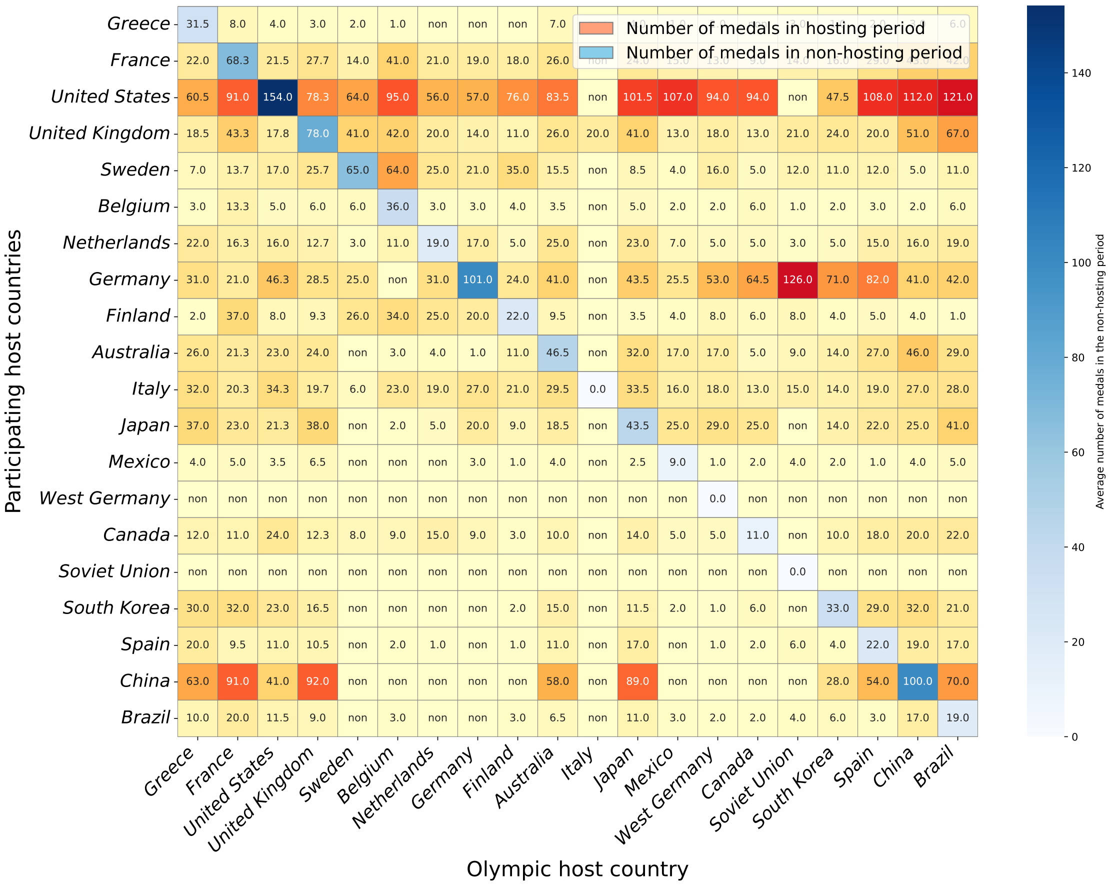


Figure 3: Host Country Medal Performance Comparison

4 Modeling and Application of Olympic Medal Prediction

4.1 Knee-NSGA2-XGBoost

In this section, we propose a medal range prediction method combining the NSGA-II multi-objective optimization algorithm with the XGBoost model. The goal is to optimize the hyperparameters of XGBoost to both minimize the width of the prediction intervals and maximize the coverage of the prediction intervals. Additionally, we use a knee mechanism to process the obtained Pareto set, effectively avoiding subjective biases. Specific details are provided in Appendix A.

4.1.1 XGBoost Model and Quantile Regression

The XGBoost model trains a regression model by optimizing the loss function. For quantile regression problems, XGBoost predicts the regression result for a specified quantile by minimizing the quantile regression loss. For a set of training data $\{(X_i, y_i)\}_{i=1}^{N}$, where X_i is the multi-dimensional feature information of the i-th sample and y_i is the true medal distribution, the objective of the XGBoost model is to minimize the following loss function:

Table 2: Multi-dimensional Feature Dataset

Feature	Description
Year	The year of the Olympic Games
Is_Host	Whether the country is the host (1 for host, 0 for non-host)
NOC	National Olympic Committee code
Total_Discipline	The total number of disciplines in the Olympic Games
Total_Sports	The total number of sports in the Olympic Games
Num_Sports	The number of sports the country participates in
Num_Events	The number of events in which the country participates
Num_Athletes	The number of athletes representing the country
Male_Ratio	The proportion of male athletes
Rank	The country's rank in the medal tally
Gold	The number of gold medals won
Silver	The number of silver medals won
Bronze	The number of bronze medals won
Total	The total number of medals won (including gold, silver, and bronze)

$$L_\theta(q) = \sum_{i=1}^{N} \rho_q(y_i - f_\theta(X_i)) \tag{1}$$

$$\rho_q(u) = (q-1)\cdot u \cdot 1(u<0) + q \cdot u \cdot 1(u \geq 0) \tag{2}$$

where $f_\theta(X_i)$ is the predicted value of the XGBoost model for the feature X_i, u is the difference between the predicted value and the true value, and q is the specified quantile, with $q = 0.05$ and $q = 0.95$ corresponding to the 5% and 95% quantiles, respectively, for calculating the lower and upper bounds of the prediction.

4.1.2 NSGA-II Multi-objective Optimization

Through NSGA-II optimization, the hyperparameters of the XGBoost model (such as maximum tree depth, learning rate, and subsample ratio) are effectively adjusted. The model can provide a higher prediction interval coverage while ensuring a smaller interval width. This makes the final prediction results more stable and reliable.

Objective 1: Maximize coverage. Given the sample set $\{(X_i, y_i)\}_{i=1}^{N}$, the coverage C is expressed as equation 3.

Objective 2: Minimize interval width. For each sample X_i, the interval width is defined as $\hat{y}_i(U) - \hat{y}_i(L)$, so the interval width W is expressed as equation 4.

$$C = \frac{1}{N}\sum_{i=1}^{N} 1\left(y_i \in [\hat{y}_i(L), \hat{y}_i(U)]\right) \tag{3}$$

$$W = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{y}_i(U) - \hat{y}_i(L)\right) \tag{4}$$

where $1(A)$ is the indicator function, which equals 1 if A is true and 0 otherwise. Coverage measures the frequency with which the prediction interval covers the true value.

4.1.3 Knee Point Selection of Pareto Optimal Solutions

In practical applications, decision-makers often need to select a balanced solution from the obtained X_p in the Pareto optimal set. The traditional methods for balancing objectives can be divided into the following two categories:

- Use a random algorithm (such as roulette) to select a set of solutions X_p from the Pareto optimal set.
- The decision-maker balances the weights of multiple objectives based on specific experience and needs, and selects the balanced solution from the Pareto optimal set based on the fitness of each objective.

As shown in Figure 4, the most convex part of the Pareto frontier is intuitively described as the knee point. Any one-dimensional objective value (such as Δf_1^{AC} and Δf_2^{AB}) will lead to significant increases in the objective values in other dimensions, such as Δf_1^{AB} and Δf_2^{AC}, around these knee points. Knee point A represents: $\Delta f_1^{AB} >> f_2^{AB}$ and $\Delta f_2^{AC} >> f_1^{AC}$. By using the knee point, the subjective bias of selecting a solution from the Pareto set can be effectively avoided [5].

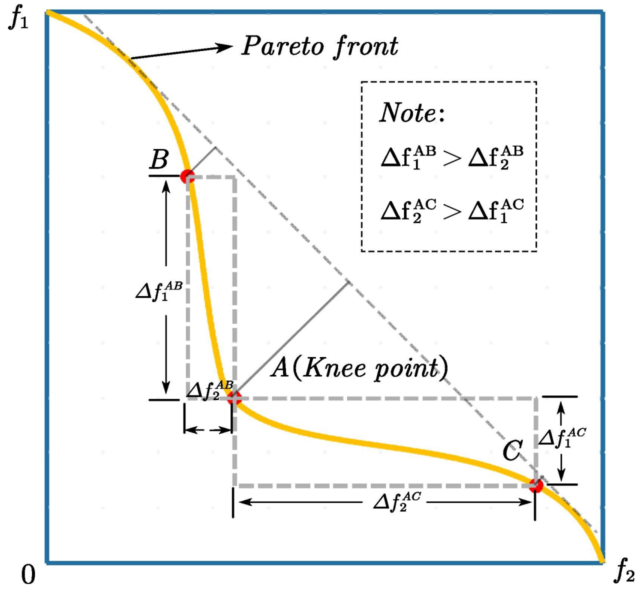


Figure 4: The Knee Mechanism of the Pareto Frontier

4.2 Results Analysis

4.2.1 Model Comparison Performance

In this subsection, we divide the training and testing sets in an 8:2 ratio and compare the performance of the proposed K-NSGA2-XGB model with common machine learning models and statistical methods on the test set. The compared models include Random Forest, XGBoost, LightGBM, and Linear models. The results show that K-NSGA2-XGB outperforms the others on multiple metrics. Meanwhile, Figure 5 demonstrates the fitting performance of the proposed model on the test set, showing good results. To provide more uncertainty information, we also establish a quantile regression version of K-NSGA2-XGB, and Figure 6 displays the predictive interval results from the model for the gold medal perspective.

Table 3: Model Comparison Performance

Type	Indicator	K-NSGA2-XGB(ours)	Random Forest	XGBoost	LightGBM	Linear
Total	RMSE	**4.27**	4.57	4.33	5.79	8.53
	R2	**0.93**	0.92	0.92	0.87	0.71
	MAPE	**20.90**	34.55	33.80	51.61	91.45
Gold	RMSE	**1.68**	2.03	1.91	2.64	3.89
	R2	**0.92**	0.89	0.90	0.81	0.60
	MAPE	**24.60**	40.06	34.87	53.70	89.80

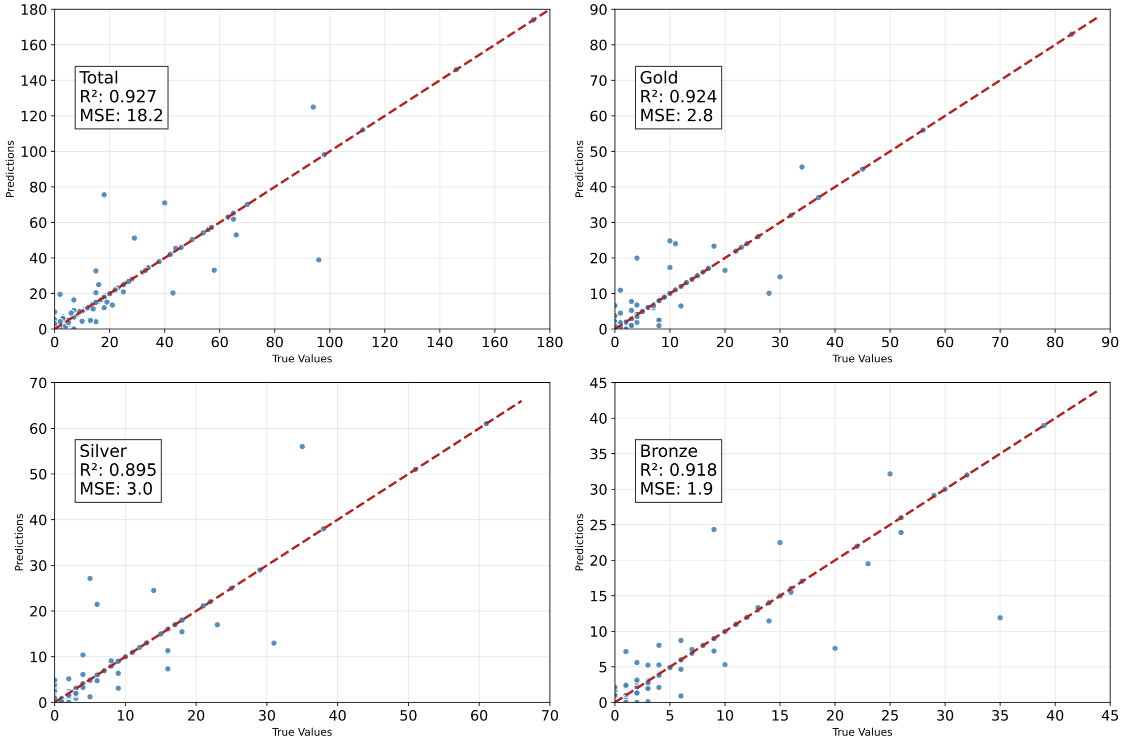


Figure 5: Point Prediction Performance

4.2.2 Shap Interpretability Analysis

SShap (SHapley Additive exPlanations) is a powerful tool for interpreting machine learning models, which helps to understand the contribution of each feature to the model's predictions. In this study, SHAP is used to analyze the key features that affect the model's prediction results and to explain their importance and relationships. Figure 7a shows the SHAP Summary Plot, and Figures 7b and 7c display the dependence plots for Num Athletes and Male Ratio, respectively.

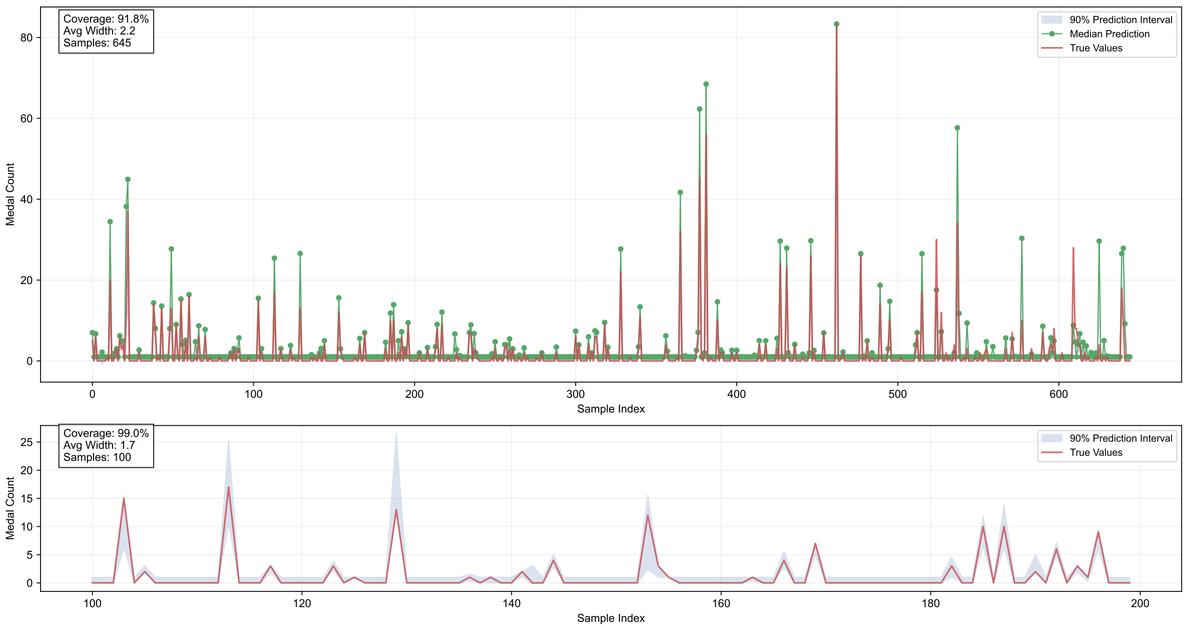
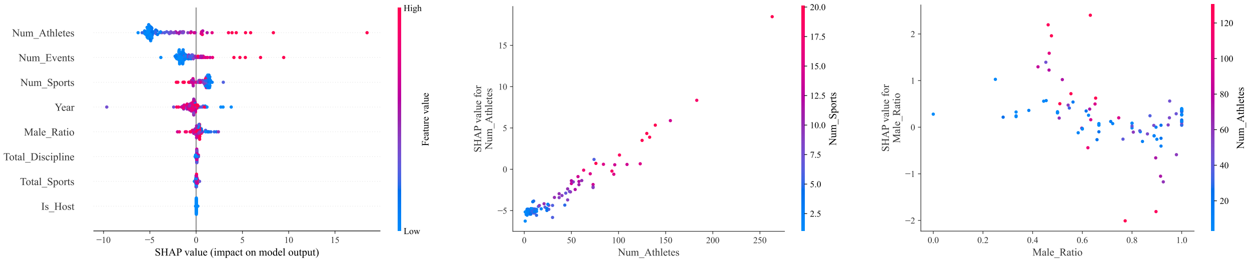


Figure 6: Predictive Interval Results

The results show that the Num Athletes feature has the most significant influence on the model's predictions. SHAP interpretability analysis is of great importance for improving the transparency and credibility of the model. By deeply analyzing the driving factors of the model's predictions, we can enhance the model's explainability.



(a) Shape Summary Plot (b) Dependence of Num Athletes (c) Dependence of Male Ratio

Figure 7: Shap Result

4.2.3 2028 Los Angeles Olympics Medal Prediction Analysis

In this subsection, we apply the proposed model K-NSGA2-XGB to predict the medal outcomes for the 2028 Los Angeles Olympics. The results for the top 15 countries are shown in Figure 8. From the figure, it is clear that the United States, benefiting from the host country effect, is leading by a wide margin. Additionally, an interesting point emerges: the top ten countries almost monopolize the medals, with significant differences between them. We believe this is due to the recent economic downturn, which has led many countries to reduce spending on sports. Furthermore, we present the interval predictions for the total number of medals and gold medals for the top five countries in Table 4.

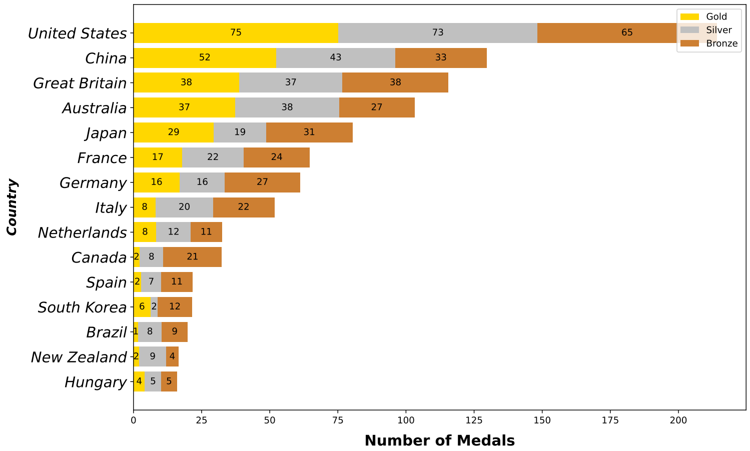


Figure 8: Top 15 Countries by Predicted Medals in 2028

Table 4: Interval Prediction Results for Different Countries

Type	Indicator	United States	China	Great Britain	Australia	Japan
Total	Lower Bound	215.54	129.89	114.32	106.43	82.32
	Point Prediction	213	128	113	102	79
	Upper Bound	211.23	126.54	110.76	100.65	76.96
Gold	Lower Bound	76.02	52.99	38.78	38.31	30.41
	Point Prediction	75	52	38	37	29
	Upper Bound	74.36	51.33	37.28	36.78	27.64

5 Modeling the First Medal Countries

Analyzing the probability of first-time medal-winning countries in the Olympics holds significant value and meaning. It can reveal the rise of emerging sports powerhouses, help countries optimize sports policies and resource allocation, and promote international sports cooperation. This analysis can also inspire athletes and the public's enthusiasm for sports, driving the development of the sports industry. To this end, we selected past events of first-time medalists and constructed corresponding datasets using the multi-dimensional feature information mentioned earlier for subsequent modeling.

5.1 Support Vector Machines for Imbalanced Classification

The prediction of first-time medal events can be modeled as a binary imbalanced classification task. The difficulty in handling such tasks lies in capturing the complex nonlinear relationships between features and events and addressing the challenges posed by imbalanced samples. Therefore, we use Support Vector Machines (SVM) along with a novel loss function to handle this issue. Specifically, SVM works by finding a hyperplane that separates different classes. However, in imbalanced datasets, most of the data comes from the dominant class, which may lead to the decision boundary being dominated by the dominant class. By using Focal Loss, we can adjust the model's focus on different class samples to mitigate such issues.

SVM Hinge Loss and Focal Loss

$$L_{\text{hinge}}(y, f(x)) = \max(0, 1 - yf(x)) \tag{5}$$
$$FL(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t) \tag{6}$$
$$L_{\text{focal}}(y, f(x)) = -\alpha_t (1-p_t)^\gamma \log(p_t) \tag{7}$$

where $y \in \{-1, 1\}$ is the label, $f(x)$ is the predicted output from the SVM model, $p_t = \sigma(f(x)) = \frac{1}{1+e^{-f(x)}}$ is the predicted probability, α_t is the class balancing factor, and γ is the focal parameter used to control the attention on difficult-to-classify samples.

Improved SVM Optimization Objective Function

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i f(x_i)) \tag{8}$$

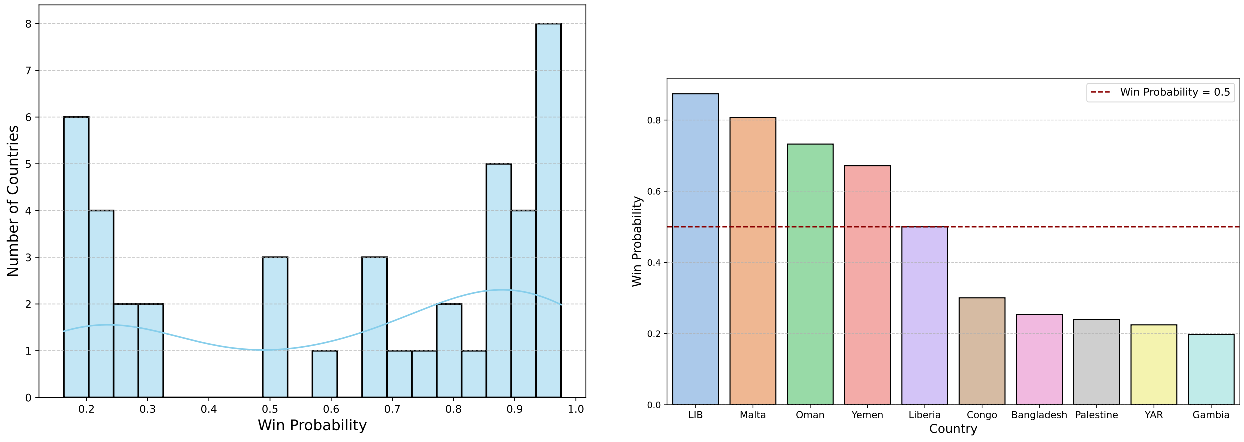
$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \alpha_t (1-p_t)^\gamma \log(p_t) \tag{9}$$

where $p_t = \sigma(f(x_i))$ is the predicted probability for sample x_i.

5.2 Results Analysis

SVM Model's Predicted Win Probability Distribution Histogram: Figure 9a shows the distribution of win probabilities predicted by the SVM model for each country. The histogram and KDE curve visually display the probability density of each country's chances of winning. It is clear from the figure that the model provides a good separation, identifying approximately 8 countries that are highly likely to win their first medal in 2028.

Top Ten Countries Likely to Win Their First Medal: Figure 9b presents the top ten countries predicted by the SVM model to win their first medal. The bar chart is sorted by the winning probability from high to low, with a dashed line indicating a winning probability of 0.5.



(a) Win Probability Distribution

(b) Top 10 Countries Likely to Win First Medal

Figure 9: Prediction of First Medal Countries

5.3 National Dependence on Top Athletes and Sport-Specific Performance Analysis

In Figure 10, we present an analysis of national dependence on top athletes and sport-specific performance. The left subfigure, Figure 10a, shows the degree to which each country relies on its top athletes, who contribute significantly to their nation's medal count with their exceptional performances. The right subfigure, Figure 10b, focuses on each country's performance in specific sports, highlighting nations that depend heavily on certain events.

The analysis presented in these figures is closely tied to the prediction of first-time medal-winning countries. For some nations, their top athletes and specific sports could be the key to earning their first Olympic medal. By analyzing these figures, we can better understand which countries are more reliant on certain areas, which provides strong support for predicting nations likely to win their first medals.

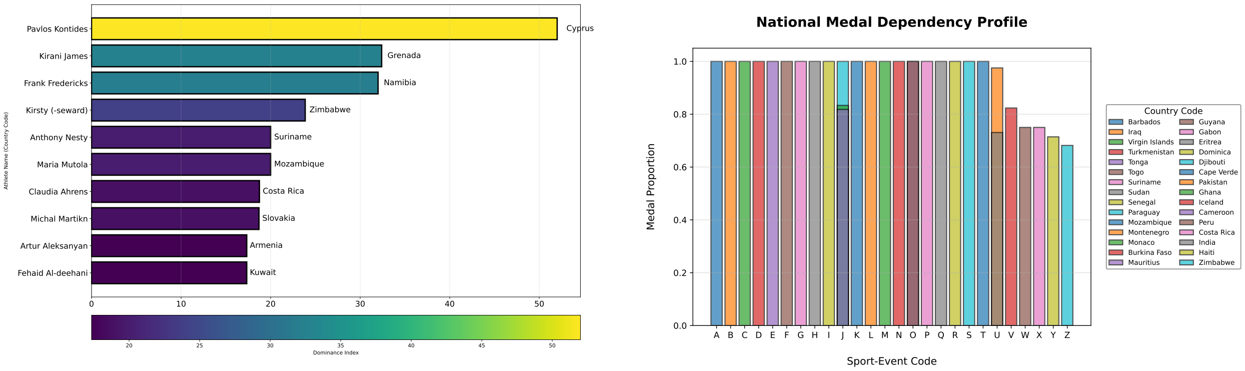
To measure an athlete's dominance in their field, we designed the Dominance_Index, which is calculated as follows:

$$\text{Medal\_Proportion} = \frac{\text{Medal\_Count}}{\text{Country\_Medals}} \tag{10}$$

$$\text{Weighted\_Score} = 3 \times \text{Gold} + 2 \times \text{Silver} + 1 \times \text{Bronze} \tag{11}$$

$$\text{Year\_Span} = \text{Last\_Year} - \text{First\_Year} + 1 \tag{12}$$

$$\text{Dominance\_Index} = \text{Weighted\_Score} \times \text{Year\_Span} \times \text{Medal\_Proportion} \tag{13}$$



(a) National dependence on Top Athletes

(b) National dependence on Specific Sports

Figure 10: National dependence on Top Athletes and Specific Sports

6 Great Coach Identification and Investment Strategies

In this section, we first identify anomaly change points in large-scale data using a change point detection algorithm. We hypothesize that these points are influenced by great coaches or similar effects. Next, we build a LightGBM model to explore the potential relationship between multi-dimensional information and great coaches. Finally, we provide investment recommendations for countries predicted in 2028, which are expected to regress the most compared to the average level of the previous three Olympics and lack great coaches.

6.1 Great Coach Identification

The Pelt algorithm (Pruned Exact Linear Time Algorithm) is an efficient change point detection method designed to optimize the data analysis process by detecting significant change points in time series.Specific details are provided in Appendix A.The basic idea is to minimize the cost function and partition the time series into multiple subsequences with similar behavior. Given a time series $x = \{x_1, x_2, \ldots, x_T\}$, the Pelt algorithm uses the cost function $C(i, j)$ to measure the difference between points i and j in the sequence. A common cost function is the Mean Squared Error (MSE), defined as:

$$C(i, j) = \sum_{k=i}^{j} (x_k - \hat{x})^2 \tag{14}$$

where $\hat{x}$ is the mean or fit value of the subsequence. The Pelt algorithm recursively calculates each possible split point, aiming to find a series of change points that minimize the total cost of the time series. Specifically, the recursive relationship for the algorithm is:

$$\text{cost}(k, T) = \min_{1 \le t < T} \left(\text{cost}(k-1, t) + C(t, T) + \text{penalty}\right) \tag{15}$$

where penalty is a term used to control the number of change points. To improve efficiency, the Pelt algorithm employs a pruning strategy that discards change points with higher costs, thus reducing unnecessary computations. When the algorithm detects a change point, if the cost of a particular change point plus the penalty exceeds the current minimum cost, the algorithm will skip that point. Ultimately, the Pelt algorithm efficiently detects significant change points in time series by minimizing the cost function and is suitable for large-scale data analysis.

The test results are shown in Figures 11-17. A total of six potential country-event-year combinations that may involve great coaches were identified.

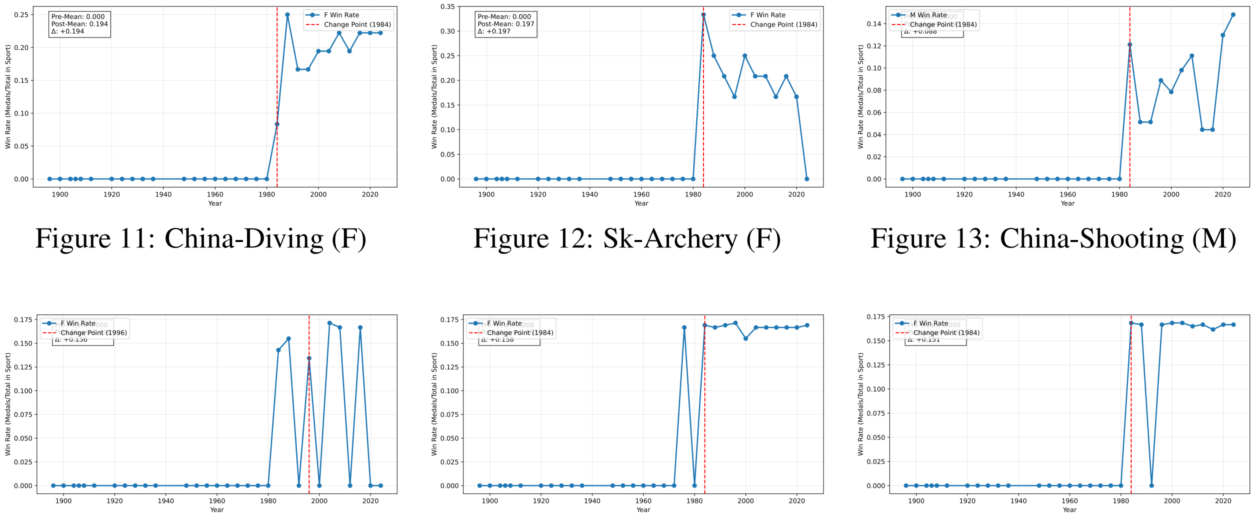


Figure 11: China-Diving (F)

Figure 12: Sk-Archery (F)

Figure 13: China-Shooting (M)

Figure 14: China-Volleyball (F)

Figure 15: US-Basketball (F)

Figure 16: Netherlands (F)

Figure 17: Chnge point detection result

6.2 Investment Strategies

To maximize investment benefits, the investment strategy for great coaches proposed in this study is based on two points: 1. The country-event combinations lacking great coaches in 2028. 2. A significant decline in performance in these events in 2028 compared to the average of the previous three Olympics. Based on these rules, we suggest the following three country-event combinations to hire great coaches:

- **China - Women's Volleyball** Although China's women's volleyball team achieved excellent results at the 2016 Rio Olympics, their recent performance has been less satisfactory. From the actual situation, key figures such as Lang Ping and Zhu Ting have left the team, which aligns with the model's identification of a lack of a great coach.
- **Germany - Men's Football** The German men's football team has historically won multiple World Cups and European Championships, but their recent performance in major tournaments has been unsatisfactory. In the 2018 World Cup, Germany failed to advance past the group stage, indicating a decline in overall team strength. With changes in the head coach and the retirement of some core players, Germany may need to find a great coach to help them recover from this low point.
- **Brazil - Men's Volleyball** Brazil's men's volleyball team has been a powerhouse in the sport, but after the 2016 Rio Olympics, although their performance remained strong, there was a slight decline in the following years. Changes in the head coach and adjustments in the team could affect their overall performance. Therefore, Brazil's men's volleyball team may be another event in need of a great coach.

7 Other Insights

The distribution of Olympic medals and the changes in performance are influenced by various factors, especially the number of countries participating, the growing strength of major countries, GDP levels, and geographical locations. Below are several angles of analysis:

- **Changes in Participation:** The increasing number of countries participating in the Olympics diversifies the competition. Emerging nations, especially from Africa and Asia, are making breakthroughs in events like track and field and weightlifting, spreading out medal distribution and reducing the dominance of traditional strong nations.
- **The Growing Strength of Major Countries:** As countries like the United States and China invest more in sports, their Olympic performance improves. The increase in resources and athlete development has led to a steady rise in their medal counts, enhancing their competitive edge.
- **The Impact of GDP:** Higher GDP countries can afford better sports infrastructure and training, leading to superior performance and more medals. Economically stronger nations tend to dominate several events, as seen in the correlation between GDP and Olympic success.
- **Geography and the Advantage of Coastal Regions:** Coastal regions often experience faster economic development and better sports infrastructure, attracting more talent. This gives countries in these areas a competitive advantage.
- **Conclusion:** Factors like participation, the strength of major countries, GDP, and geography shape Olympic medal distribution. These elements highlight global shifts in competition and offer insights for future predictions and athlete development.

8 Model Robustness Analysis

In this experiment, we conducted a robustness verification of the Knee-NSGA2-XGBoost model's hyperparameter combinations, primarily through experiments on different combinations of learning rate η and maximum tree depth $d_{\max}$. The model's performance was evaluated under different data splits, Gaussian noise interference, and class imbalance conditions. The results showed that the model exhibited good robustness for most hyperparameter combinations, especially the combination of low learning rate and shallow tree depth, which demonstrated stability under noise perturbation and data splitting. In contrast, the combination of high learning rate and deep trees increased model complexity and accuracy but was more sensitive to class imbalance and noise. Bias-variance analysis explains that the high bias of shallow trees effectively suppresses the increase in variance, while low learning rates reduce the sensitivity of estimates through iterative averaging. Overall, the model is robust to most hyperparameter combinations.

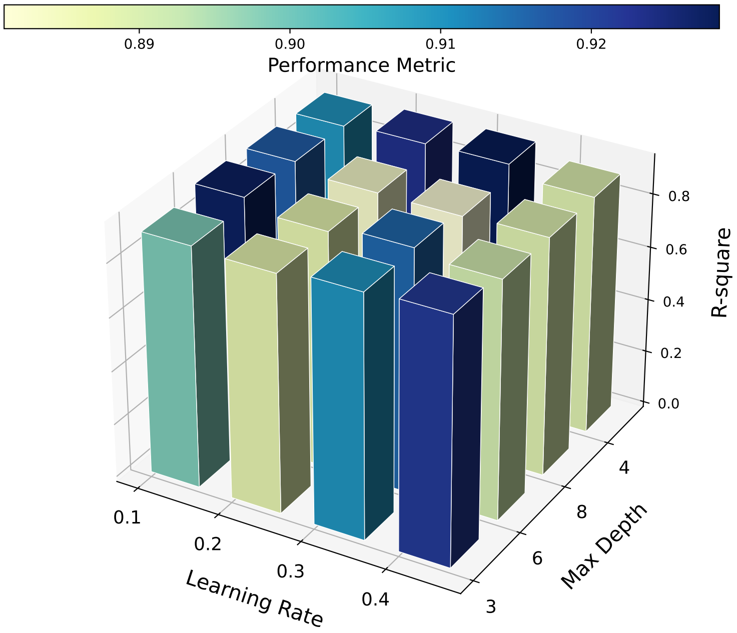


Figure 18: Knee-NSGA2-XGBoost Hyperparameter Combinations

9 Model Advantages and Disadvantages

9.1 Advantages

- This study systematically explores the complex relationships between medals and various factors from multiple perspectives. By combining historical data with multidimensional feature information, we built a detailed feature library, providing comprehensive data support for medal prediction. This multidimensional feature library can effectively uncover potential influencing factors, enhancing the model's interpretability and prediction accuracy.

- We proposed the innovative Knee-NSGA2-XGBoost model for medal prediction. This model employs a multi-objective optimization mechanism during interval prediction, effectively balancing coverage and interval width, optimizing the selection of prediction intervals. Additionally, the introduction of the Knee mechanism eliminates subjective factors when selecting Pareto optimal points, improving the model's decision-making objectivity and stability.
- To address the class imbalance issue in medal prediction, we used the Focal Loss function in combination with Support Vector Machines (SVM) for optimization. This approach has significant advantages when predicting extreme events (such as rare occurrences during first-time medal predictions), effectively improving prediction accuracy and reducing bias.
- From the perspective of change point analysis, we deeply understood the effect of great coaches. By using the Pelt algorithm, we identified potential great coaches from past competition data, revealing their potential impact on medal prediction. This method allows us to better capture subtle changes between coaches and athlete performance, enriching the feature layer of the prediction model.

9.2 Disadvantages

- Although the Knee-NSGA2-XGBoost model achieves a good balance between multiple objectives, the optimization process is computationally complex, especially when dealing with large-scale datasets, which may result in longer training times. Moreover, the results of multi-objective optimization may also be affected by parameter selection, requiring fine-tuning in different application scenarios.
- Although Focal Loss has significant advantages in addressing class imbalance, its optimization process may introduce strong bias, especially when dealing with highly imbalanced data, leading to overfitting for the minority class. Additionally, the training time for SVM may be long, particularly in high-dimensional feature spaces.
- Focal Loss is very effective in handling class imbalance, but it could cause the model to focus too much on the minority class, leading to overfitting in such classes. In extreme cases of imbalance, the model's performance might require further tuning.
- When using the Pelt algorithm for change point detection, although the algorithm effectively captures change points in the data, it may require more time to adjust parameters and improve robustness when dealing with very large or noisy datasets.

References

[1] Andrew B Bernard and Meghan R Busse. Who wins the olympic games: Economic resources and medal totals. *Review of economics and statistics*, 86(1):413–417, 2004.

[2] David Forrest, Ismael Sanz, and J de D Tena. Forecasting national team medal totals at the summer olympic games. *International Journal of Forecasting*, 26(3):576–588, 2010.

[3] Christoph Schlembach, Sascha L Schmidt, Dominik Schreyer, and Linus Wunderlich. Forecasting the olympic medal distribution during a pandemic: a socio-economic machine learning model. *arXiv preprint arXiv:2012.04378*, 2020.

[4] Gergely Csurilla and Imre Fertő. How to win the first olympic medal? and the second? *Social Science Quarterly*, 105(5):1544–1564, 2024.

[5] Jianzhou Wang and Yuyang Gao. An integrated forecasting system based on knee-based multi-objective optimization for solar radiation interval forecasting. *Expert Systems with Applications*, 198:116934, 2022.

A Knee-NSGA2-XGBoost Algorithm and PELT Algorithm

Knee-NSGA2-XGBoost Algorithm

Algorithm 1 Knee-NSGA2-XGBoost Optimization with XGBoost Interval Prediction (Part 1)

```
Input:
    1. Population size $P$, generations $G$, knee threshold $\theta$
    2. XGBoost hyperparameter space $H = \{\eta, \gamma, \lambda, max\_depth, n\_estimators\}$
    3. NSGA2 parameters: crossover rate $p_{crossover}$, mutation rate $p_{mutation}$, tournament size $k$
    4. Training data $D_{train}$, validation data $D_{val}$
    5. Prediction interval confidence level $(1-\alpha)$
Initialize:
    1. Generate initial population $P_0 = \{h_1, \dots, h_P\}$ where $h_i \sim U(H_{min}, H_{max})$
    2. Build prediction intervals $\forall h_i \in P_0$:
        a. Train XGBoost model $M_i$ on $D_{train}$ with $h_i$
        b. Calculate quantile predictions $\hat{y}_{lower}, \hat{y}_{upper}$ on $D_{val}$
    3. Compute objectives $\forall h_i$:
        a. Coverage: $f_1 = \frac{1}{N}\sum_{j=1}^{N} I\{\hat{y}_j \in [\hat{y}_{lower}(j), \hat{y}_{upper}(j)]\}$
        b. Width: $f_2 = \frac{1}{N}\sum_{j=1}^{N}(\hat{y}_{upper}(j) - \hat{y}_{lower}(j))$
    4. Non-dominated sort $P_0$ using fast nondominated sort
    5. Set $g = 0$
```

Algorithm 2 Knee-NSGA2-XGBoost Optimization with XGBoost Interval Prediction (Part 2)

```
While $g < G$ do
    1. Selection:
        a. Tournament selection with size $k$ on $P_g$
        b. Select mating pool $Q_g$ of size $P/2$
    2. Variation:
        a. Apply SBX crossover to $Q_g$ with $p_{crossover}$
        b. Apply polynomial mutation to offspring with $p_{mutation}$
        c. Generate offspring population $O_g$ of size $P$
    3. Evaluation:
        a. For each $h_i \in O_g$, compute $f_1(i)$ and $f_2(i)$ as in lines 9-10
        b. Combine parent and offspring: $R_g = P_g \cup O_g$
    4. Knee Identification:
        a. Perform non-dominated sort on $R_g$ to get Pareto fronts $F_1, F_2, \ldots$
        b. For first front $F_1$, compute knee point:
            i. Normalize objectives to $[0, 1]$ range
            ii. Calculate angle $\phi_i = \arctan\left(\frac{\Delta f_2}{\Delta f_1}\right)$ for consecutive points
            iii. Identify $h^* = \arg\max\left(\phi_i(\phi_i - \theta)\right)$
        c. Select solutions within $\varepsilon$-neighborhood of $h^*$
    5. Environmental Selection:
        a. Apply crowding distance sorting to $R_g$
        b. Select new population $P_{g+1}$ of size $P$ from $R_g$
        c. Preserve knee solutions in elite archive
    6. $g \leftarrow g + 1$
```

Algorithm 3 Knee-NSGA2-XGBoost Optimization with XGBoost Interval Prediction (Part 3)

```
Output:
    1. Final Pareto front $F_1^{final}$ from $P_G$
    2. Optimal hyperparameters $h^*$ from knee selection
    3. Prediction interval generator $PI(\cdot \mid h^*)$
```

PELT Algorithm Pseudocode (Part 1)

Algorithm 4 PELT - Pruned Exact Linear Time Change Point Detection

```
Input:
    1. Time series data $y_{1:n} = \{y_1, y_2, ..., y_n\}$
    2. Penalty term $\beta$ (e.g., BIC/MDL penalty)
    3. Loss function $\mathscr{L}(\cdot)$ (e.g., MSE, negative log-likelihood)
    4. Minimum segment length $l_{\min}$
Initialize:
    1. Set $F(0) = -\beta$ {Initial cost}
    2. $cp = [0]$ {Change point candidates}
    3. $R = \{0\}$ {Set of last change points}
    4. seg_cost$(i, j) = \mathscr{L}(y_{i:j})$ {Precompute segment costs}
```

Algorithm 5 PELT - Pruned Exact Linear Time Change Point Detection (Part 2)

```
**Main Procedure:**
**for** $t = 1$ **to** $n$ **do**
    1. Calculate candidate costs:
       $F(t) = \min_{\tau \in R} [F(\tau) + \text{seg\_cost}(\tau + 1, t) + \beta]$
    2. Update change points:
       $\tau^* =_{\tau \in R} [F(\tau) + \text{seg\_cost}(\tau + 1, t) + \beta]$
       $cp = cp \cup \{\tau^*\}$
    3. Pruning step:
       $R = \{\tau \in R \cup \{t\} \mid F(\tau) + \text{seg\_cost}(\tau + 1, t) \leq F(t)\}$
    4. Enforce minimum segment length:
       Remove $\tau$ from $R$ where $t - \tau < l_{\text{min}}$
**end for**
**Backtracking:**
1. Initialize $m = 0$, $\hat{cp} = []$, $last\_cp = n$
**while** $last\_cp > 0$ **do**
    $m =_{\tau \in cp} [F(\tau) + \text{seg\_cost}(\tau + 1, last\_cp)]$
    $\hat{cp} = [m] \cup \hat{cp}$
    $last\_cp = m$
**end while**
**Output:**
  1. Detected change points $\hat{cp} \setminus \{0\}$
  2. Optimal cost $F(n)$
```