Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2524635

Integrating Athlete Probabilities for National Olympic Medal Forecasting

Summary

To enhance the accuracy of Olympic medal predictions and to explore the relationships between Olympic medals and various influencing factors, we have innovatively proposed a powerful national medal prediction model and computational framework based on athlete probability aggregation, called APINet.

APINet itself, as a general and effective probability aggregation computation framework, includes the Athlete Medal Prediction Model (APM), the National Medal Expectation Regression Model, and the Monte Carlo-Poisson Distribution Sampling Model. Together, these three components form a national medal prediction model. Not only did we effectively solve the prediction of medal counts and the prediction of the first medal-winning country for the 2028 Olympics based on APINet, but we also addressed the project-country association issue and the ”Great Coach Influence”.

Before modeling, we conducted rigorous screening and data processing of past Olympic medal data and athlete data, performed targeted feature engineering, and established multiple datasets to help us train different models. At the same time, when constructing the computational framework of APINet, we designed APM and the expected regression model as pluggable models to adapt to various mainstream algorithms, further enhancing the generality of APINet. On this basis, we compared the advantages and disadvantages of different specific algorithms and selected the optimal model.

For APINet, we first established the APM, an athlete probability prediction model, which was trained on data to output the probability of athletes winning medals, with XGBoost yielding the best results.

Keywords: APINet, APM, SHAP, KMeans, Causal model

Contents

1 Introduction

1.1 Background

The charm of competitive sports lies in the unpredictability of their outcomes. As the world's grandest sporting event, the modern Olympics has consistently garnered widespread attention and love since its inception. In each Olympic Games, medal predictions often become a hot topic of discussion. Through these predictions, people not only look forward to the events but also witness moments where athletes challenge themselves. At the recent 2024 Olympic Games, 206 countries and regions gathered in Paris, where athletes devoted their energies and strived for excellence. In the end, the United States, China, and Japan secured the top three spots on the medal tally.

Although competitions have uncertainties, the strong strength of some teams in certain specific events ensures a higher probability of winning, making the outcomes of these events more predictable. Using data from the modern Olympics [1], we calculated the cumulative rankings of all countries' medal counts since 1896. We can see the top five in the medal table in Figure 1:

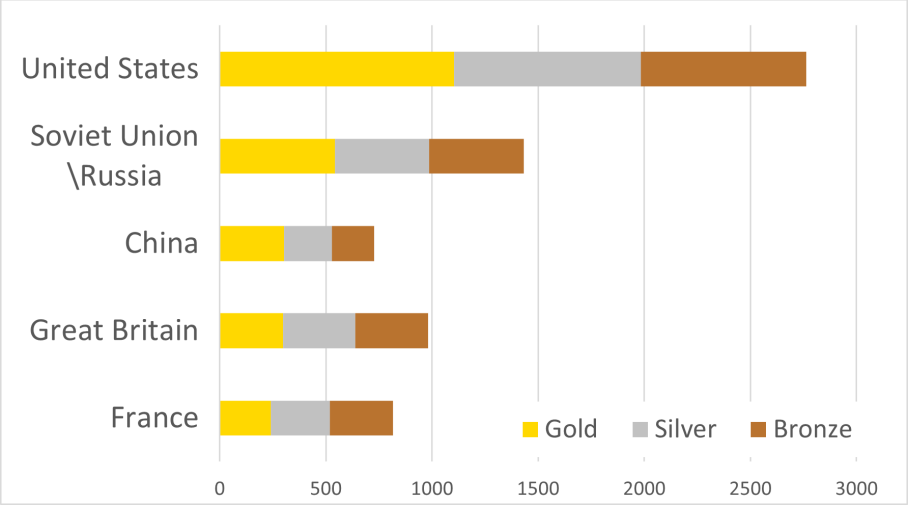

Figure 1: The top five nations in the all-time medal standings

1.2 Problem Restatement

Any match is influenced by many complex factors. Through in-depth analysis and research of the problem background, combined with the given specific requirements, the problem is restated as follows:

- Establish a mathematical model to accurately predict the number of gold medals and total medals for each country, taking into account the probability of winning the first medal for countries that have never won medals. This model needs to extract patterns and discover trends from various complex variables and data. Ultimately, use this model for precise predictions.
- Establish a mathematical model to discuss the relationship between the number and types of events and the number of medals won by a country, and use this model to illustrate the importance of different events for different countries.

- Establish a mathematical model to discuss the impact of ”great coaches” on the number of national medals, and use the model to select specific examples to illustrate the influence of ”great coaches.”
- By establishing new models, we can find new possible insights into the number of Olympic medals.

1.3 Literature Review

This issue mainly concerns predictions related to the number of Olympic medals. In recent years, research on predicting various Olympic medal counts has exhibited great popularity. It is mainly divided into time series prediction models, empirical models, and intelligent prediction models, among others.

Based on time series forecasting models, past behaviors of the phenomenon are used to predict the future. This method is centered around time series and has high computational efficiency. However, this ”predicting by analogy” approach has a significant degree of randomness.

Empirical models do not analyze the actual process; they simply derive mathematical relationships between various parameters and variables based on the principle of least error, are data-driven, and are highly practical. For example, Oyebanke Oyeyinka [3] found that the main factors affecting the number of athletes participating in the Olympics and their performance are economic, political, and religious, but these models lack theoretical explanatory power and cannot explain fundamental laws.

The intelligent prediction model is based on the use of intelligent algorithms, such as simulated annealing algorithms and swarm intelligence technology [4]. It does not require a predetermined formula form, can adapt and dynamically update, but the entire model is not entirely within people's expectations, making it prone to being out of control, with challenges of overfitting and generalization. The strengths and weaknesses of the three algorithms can be visually presented and is shown below:

1.4 Our Work

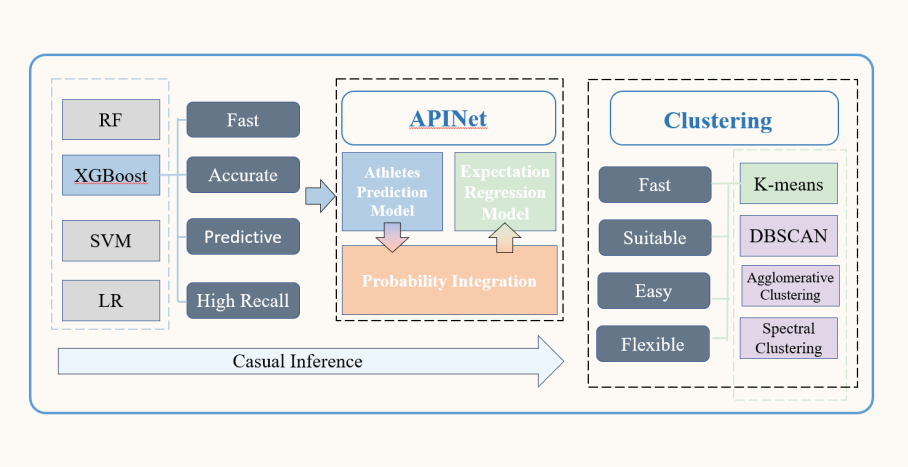

Figure 2: Our Model

2 Assumptions and Justifications

Considering that the real-world problem involves many complex factors, we first need to make reasonable assumptions to simplify the model. Each assumption is followed by an explanation:

- **Assumption 1:** We assume that all factors affecting the number of national medals only influence the athletes themselves. The strength accumulated by athletes through long-term training and their performance on the day of the competition determine the outcome of the match, and each match determines the number of national medals.
- **Assumption 2:** Based on Assumption 1, the number of medals for a country is determined by each match. We continue to assume that each match follows a Poisson distribution, with its lambda being the expected overall strength of the team.
- **Assumption 3:** We assume that by 2025, the global situation has improved, wars have ended, and peace comes to the world. Belarus is allowed to compete as a nation, and Russia competes as the ROC (Russian Olympic Committee).

Additional assumptions are made to simplify analysis for individual sections. These assumptions will be discussed at the appropriate locations.

3 Notations

Symbols	Description	Unit
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6

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

4 Data Processing

Since the founding of the modern Olympic Games in 1896, the participating countries and the events held have been constantly changing, and even three editions could not be held due to war. In order to ensure the completeness and consistency of the data from the 1st to the 33rd Olympic Games, we repeatedly compared the data before and after and processed the original data as follows:

1. We cleaned up all the invisible or non-printable characters, removed the earlier data with significant fluctuations and fewer participating countries (before 1924), deleted the ice sports that were moved to the Winter Olympics, and removed the older sports that are no longer part of modern events.

2. We have handled the missing and garbled values in the project, using the default value of 0 to fill in the missing project data.

3. We extracted the host country information, parsed the host location string into country names, and added data indicating whether each athlete is from the host country.
4. For historical changes or unusual NOCs, such as the Soviet Union and Russia, we refer to the latest information from the IOC
5. We map the projects to standardized project codes for easier searching.
6. Based on the aforementioned years and country codes, we have consolidated the medal data.

By conducting a preliminary sorting of the raw data, we indexed it by year and country code, combined the obtained medals and related events, merged some datasets, and established an initial dataset called the "Year-Country-Host-Medal Database," which contains elements such as year, country, is_host, gold, silver, bronze, and total_medal.

We have established a unified coding system. We uniformly code countries as NOC and sports events as sport event codes.

Since many events have both individual and team competitions, such as table tennis, and the number of participants in a particular event also reflects the team's strength, we need to calculate how many people from that country participate in that event, denoted as TeamNum. Additionally, to facilitate the establishment of a unified model, we can consider the same person as one individual when calculating past experiences, but as two individuals when predicting different events. In other words, we treat them as different individuals with the same past experiences.

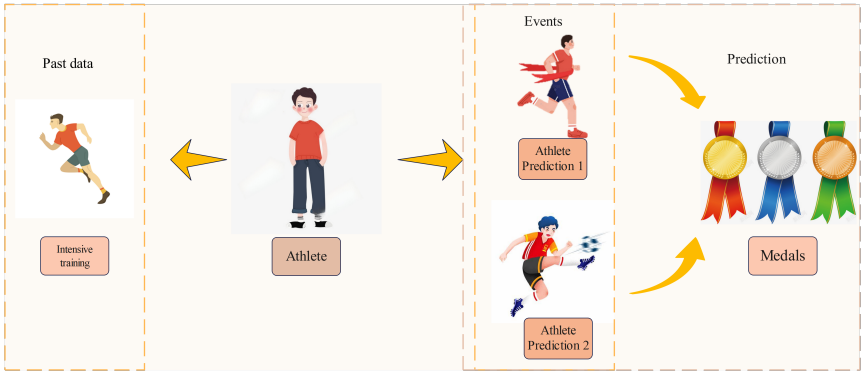

Figure 3: Explanation Diagram for Athletes

After removing the repeated influence of athletes, we can calculate the average gold medal winning rate ratio for different countries and different events. Based on the research of the given data, we know that most athletes participate in the Olympics at most twice. Therefore, when we establish the model, we introduce empirical numbers and only consider the number of gold medals in the first two Olympics:

$$AvgGoldRate = \frac{\text{Number of gold medals won by the nation in the most recent two editions}}{\text{Total participants from the nation in that sport across the same two editions}} \quad (1)$$

Among them, if the event is held for the first time, the average number of medals won by the country in the previous two editions of all events will be considered. If it is held for the second time, only the previous edition will be taken into account. Similarly, there is also the ratio of medal participants:

$$AvgMedalRate = \frac{\text{Number of total medals won by the nation in the most recent two editions}}{\text{Total participants from the nation in that sport across the same two editions}} \tag{2}$$

For the above two new definitions, we processed the data again, calculated the AvgGoldRate and AvgMedalRate for each event in each country, and then added them to the database of the relevant athletes in that country.

In addition, we also calculated what medal each athlete won in their last Olympic appearance, to directly reflect the athlete's experience and strength. If there is no Olympic participation experience, it is recorded as "No Medal."

The final database includes medal data and host country information for the valid years (1924-2024), covering 220 countries (NOC) and 52 events.Contains 11 parameters, see Table 1 for specific parameters.

Name	Year	NOC	Sport	Sex	isHost	TeamNum	AvgGoldRate	AvgMedalRate	LastMedal	Medal
Arvo Aaltonen	1924	FIN	SWM	M	0	2	0	1	Bronze	No medal
Viljo Wiklund	1924	FIN	SWM	M	0	2	0	1	No medal	No medal
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

5 From Individual to Nation: Olympic Medal Prediction Model

Conducting medal predictions, we plan to compare common excellent algorithms to select the one with the best results. Through the database, we can obtain the mathematical expectation of each athlete's medal count for 2028. Additionally, we can use a trained regression model to derive the final Olympic medal predictions based on the data of mathematical expectations.

5.1 Selection of Athlete Medal Prediction Models: Random Forest, XGBoost, SVM, Logistic Regression

Through our in-depth analysis of the data and feature engineering, we selected cases based on each athlete from each participating country and each event in every Olympic Games to construct a structured machine learning dataset, thereby building our athlete medal prediction model. Based on the existing data, we transformed the probability of athletes winning medals into a classification problem of athletes' medal-winning situations, and then obtained accurate probabilities through the trained classifier.

We know that although we have made various assumptions and conducted extensive and tedious processing of the data, there are still profound patterns hidden within the data that we cannot directly discover. Therefore, we decided to use machine learning methods to build a classifier.

Next, I will discuss some details about how we converted the database into a dataset suitable for machine learning.For the missing values in the data, we chose the forward fill method. Additionally,

we mapped NOC, Sport, Sex, Medal, and LastMedal to numerical values using categorical encoding.In feature selection, we chose NOC, Sport, isHost, TeamNum, AvgGoldRate, AvgMedalRate, LastMedal, and set Medal as the classification target.

In the context of having a large amount of data, in order to accurately predict the number of medals, our primary goal is to determine which features have the greatest impact on the model's predictions.Considering the constraints of time and resources, we decided to choose a model with the best predictive performance among the four algorithms: Random Forest, XGBoost, SVM, and Logistic Regression.To ensure the fairness of the experiment, we standardized the training process, controlled the same training dataset, and used K-fold cross-validation as the evaluation method.

Through the training of the model, we obtained evaluations of different aspects for each model. Initially, we received a very high accuracy evaluation, but based on the analysis of classification metrics related to data distribution, we found that the model had high precision, recall, and F1 scores for the non-medal cases, while the recall for gold, silver, and bronze medals was relatively low. We discovered that this was due to the excessive number of non-medal samples in the data, causing the model to tend to

However, the reason we trained the athlete medal model is that we hope the model can achieve a well-performing classifier under massive data, accurately identifying the situations of gold, silver, and bronze medals, and obtaining a probability distribution of the athlete's award situation as realistically as possible. Moreover, due to the inherent uncertainty of the Olympic Games, sacrificing a certain degree of accuracy is acceptable. Therefore, we did class weighting. Although the model's accuracy and precision decreased after weighting, the recall improved by 21

Furthermore, for the selection of classification models, we created a radar chart (Figure 5) to compare and display various aspects of the four models.After the experiments, it is evident that Random Forest and XGBoost perform well in various aspects, while SVM and Logistic Regression show significantly poorer performance in terms of Recall and other metrics. In comparison to Random Forest, XGBoost has a slight advantage in various aspects. Overall, XGBoost is the best-performing model among the four, excelling in all areas. Additionally, XGBoost's Log Loss is 0.3752, indicating good performance and highlighting its predictive accuracy.

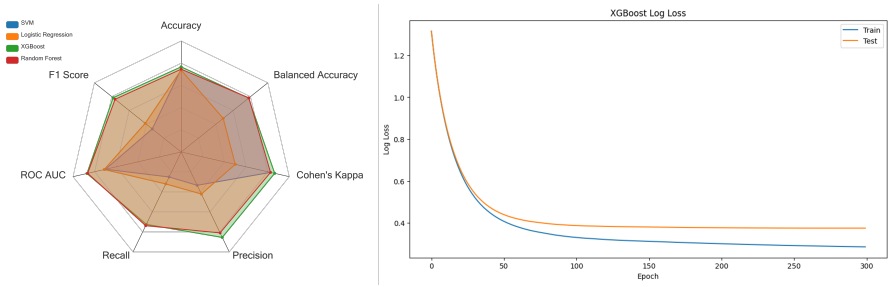

Figure 4: Model Comparison Radar Chart and XGBoost LogLoss Curve

Then, we calculated the importance of each feature for XGBoost, and the results are shown in Figure 6. It is not difficult to see that AvgMedalRate is the most important for the model, followed by AvgGoldRate, is_Host, and Sport. Each feature has a certain weight, which aligns with prior knowledge, making the model more interpretable. Another reason not shown is that we found

the Random Forest model assigned very low importance to the isHost feature, indicating that the model might have learned incorrect patterns.In addition, we also explored the correlations between the various features of XGBoost, as shown in the heatmap in Figure 6.We found that the correlation coefficients between most variables are below 0.1, which can be considered uncorrelated, while the correlation between AvgGoldRate and AvgMedalRate is 0.71, indicating a high correlation. Therefore, we conducted an analysis of the interaction effect model predictions between the two, and found that when AvgGoldRate is higher, the probability of a higher AvgMedalRate is greater, as shown in Figure 7. In this figure, (a), (b), (c), and (d) correspond to the validation analysis under the conditions of predicting no medals, winning bronze medals, winning silver medals, and winning gold medals.

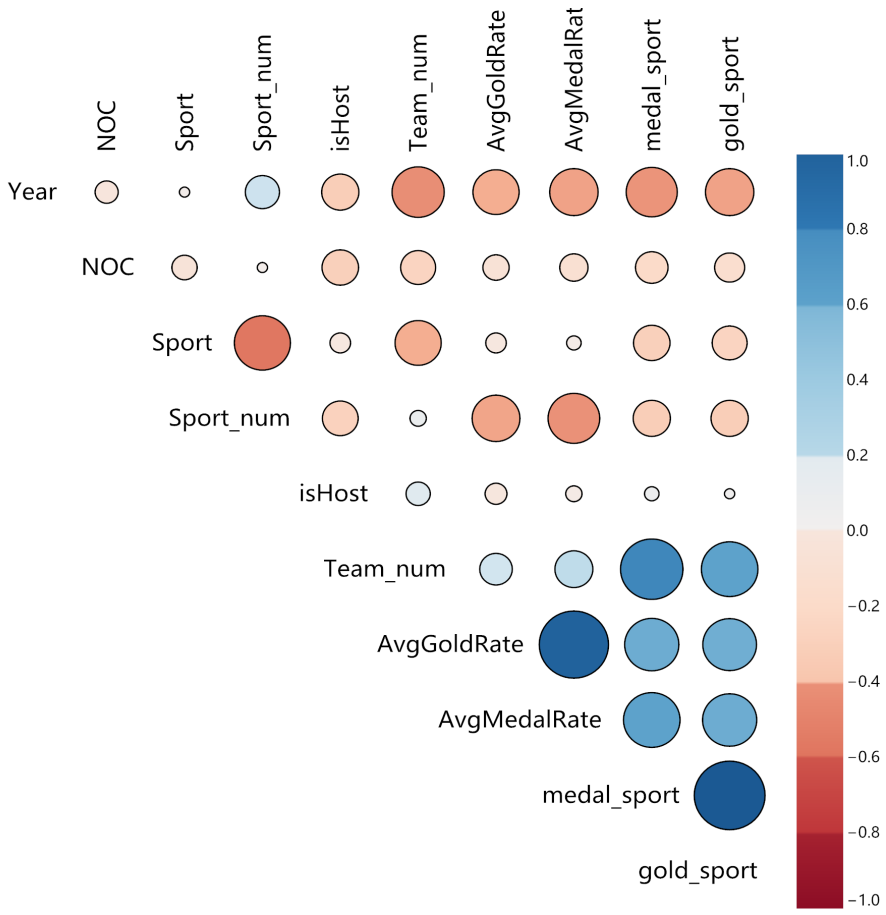

Figure 5: Heatmap of Feature Relationships in XGBoost

To better understand the model's performance, we delve into the workings of XGBoost. XGBoost, or Extreme Gradient Boosting, is an advanced implementation of the Gradient Boosting Machine (GBM) algorithm. It builds an ensemble of decision trees sequentially, where each tree aims to correct the errors (residuals) of the previous tree. The key feature of XGBoost lies in its use of a regularized objective function, which combines both the loss function and a penalty term to control model complexity and prevent overfitting. The objective function is given by:

$$Obi(f) = L(\theta) + \omega(f) \tag{3}$$

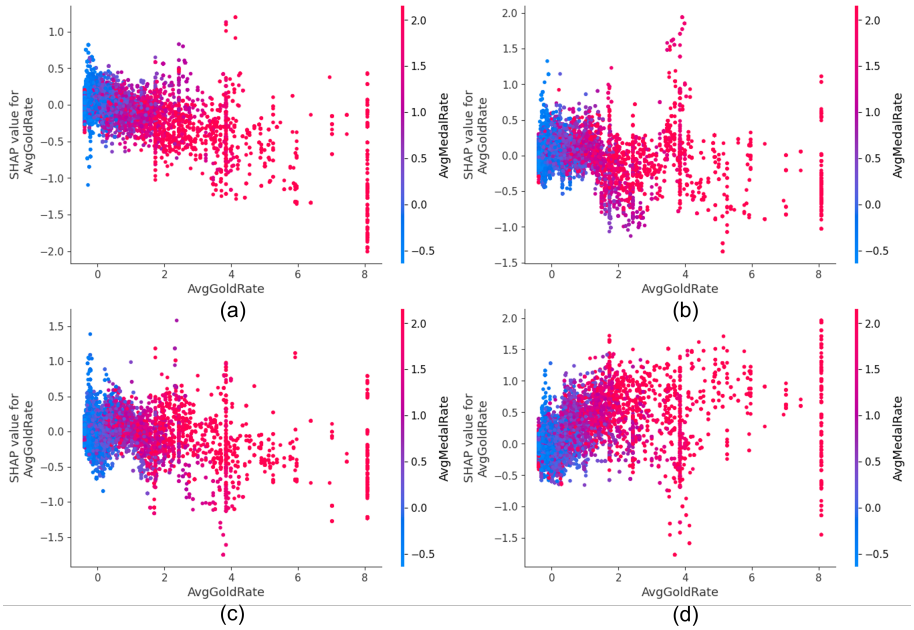

Figure 6: Relationship between AvgGoldRate and AvgMedalRate for Different Medal Predictions

where $L(\theta)$is the loss function, which measures the error between predicted and actual values, and $\Omega(f)$is the regularization term that penalizes overly complex trees. In each iteration, XGBoost minimizes the following expression:

$$f_t = \arg\min_f \sum_{i=1}^{n} \left[\hat{y}_i^{(t-1)} + f(x_i) - y_i\right]^2 + \Omega(f) \tag{4}$$

This helps the model fit the residuals while controlling complexity, which contributes to its high performance in tasks like multi-class classification, as is the case in our analysis.

Through another round of theoretical analysis, we know that XGBoost can effectively handle multi-class classification problems, directly outputting the probability predictions of athletes winning gold, silver, bronze medals, and other awards. Therefore, we decided to use this model and the SHAP model to help us analyze the importance of each feature in the predictions. SHAP is a method used to interpret machine learning model predictions. Its core is to calculate SHAP values to assess the importance of each feature in the model for the prediction results. The calculation method of SHAP is as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f(S \cup \{i\}) - f(S)\right] \tag{5}$$

Among them, $S \subseteq N \setminus \{i\}$ represents all subsets of N that do not include feature i;, $|S|$ denotes the size of subset S, M is the total number of features; $f(S)$ is the model's output when only using the feature subset S;$f(S \cup \{i\})$ represents the model's output after adding feature i to S;the weight term $\frac{|S|!\,(M-|S|-1)!}{M!}$ ensures fairness among features.

Through the SHAP model, we obtained the SHAP values for various features when predicting different medals, as shown in Figure 8. Here, (a), (b), (c), and (d) correspond to the predictions for no medal, bronze medal, silver medal, and gold medal, respectively.Due to space constraints, we focus on analyzing the predictions for the gold medal case, as the predictions for the other cases are generally similar.In the gold medal prediction model, we can see that AvgMedalRate and AvgGoldRate have the most significant contribution to high probability predictions, and the host identity (isHost) also shows a strong positive impact, which further validates the correctness of our choice of XGBoost.

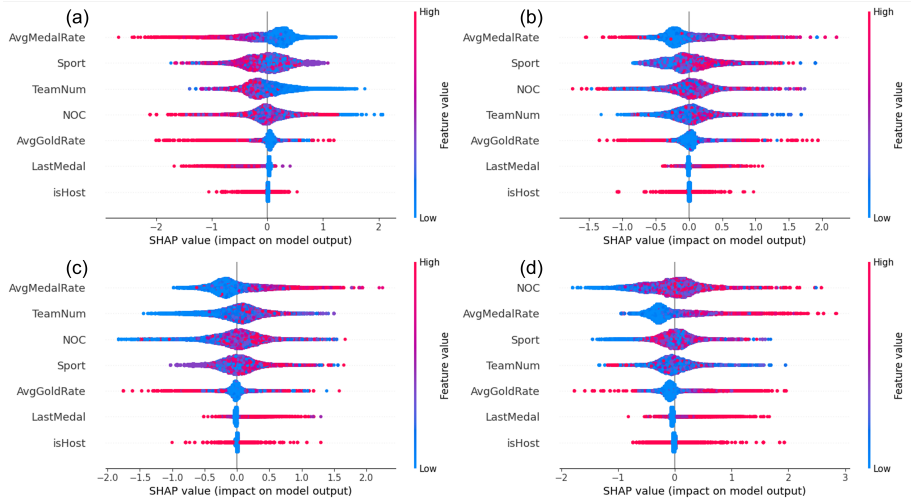

Figure 7: Feature Map for Different Medal Predictions

Through the analysis of the relevant features, we gained an understanding of the approximate weights of each feature in XGBoost, and completed the probability predictions for athletes winning various medals.

5.2 Regression Model Training

After establishing the athlete medal prediction model, we want to extend it to a national medal prediction model to help us solve practical problems. Through the athlete medal prediction model based on XGBoost, we can obtain the probabilities of each athlete winning gold, silver, bronze medals, or not winning any medals. Subsequently, by aggregating the winning probabilities of each athlete, we can derive the aggregated probabilities of all athletes in the country winning gold, silver, bronze medals, or not winning any medals.

Furthermore, we can obtain the mathematical expectation of each athlete winning various awards. We will perform a weighted sum of the mathematical expectations of each athlete in team sports events to calculate the mathematical expectation of winning various awards for each event, denoted asE_{iMedal} By summing the mathematical expectations of each country's participation in the events, we can approximate the mathematical expectations of the entire country winning medalsE_{Medal}

$$E_{Medal} = \sum_i E_{iMedal} = \sum_i p_{iMedal} + \sum_j \sum_t \alpha_t p_{jtMedal} \tag{6}$$

Among them, Medal can be Gold, Silver, Bronze, None;p_{iMedal}represents the probability of the individual event winning that Medal,$t$represents the number of players in the$j$team event,α_t is the weighted sum coefficient, and$p_{jtMedal}$ represents the probability of the$j$player in the t team event winning that Medal.

However, we cannot simply take E_{Medal}as the result of the entire national medal prediction model.Because when we consider these probabilities as the expected number of medals for the country, under the simple assumption that the aggregation of individual expectations equals the national expectation, we overlook the emergence effect from individuals to the nation, where quantitative changes lead to qualitative changes. Basically speaking, it is a kind of collective emergence effect.So we can only say that individual expectations are closely related to national expectations, but there is a certain deviation, and they cannot be directly obtained.In order to fit the expected number of gold medals and total medals for a country, bypassing the individual probabilities and the deviation from the total number of medals, we need to train a regression model based on aggregated probabilities to regress the expectations. Considering that there may be complex interwoven effects between various features that traditional linear regression finds difficult to capture, we decided to use deep learning techniques based on multilayer perceptrons to train a regression model for E_{Medal} and the actual results Y_{Medal}

BBefore training, we followed our usual data processing routine. We merged the final database, placed all the features into a single instance, converted the athletes' dataset into a national dataset, and then performed data reconstruction. The final dataset obtained only contains Year, NOC, Features, and Medals.

And we have built the APINet data pipeline, allowing the earlier APM model to maintain predictions while only training the later regression model.

During training, we chose a cosine annealing learning rate and performed error analysis on MAE, MSE, HuberLoss, and SmoothL1Loss, resulting in the outcomes shown in Figure 9. We found that the loss for MSE was too high, while the loss for MAE remained consistently low.We believe this is because MSELoss is very sensitive to outliers; even a small number of extreme values can significantly increase the loss value, whereas L1Loss (MAE) is less sensitive to outliers and does not amplify the impact of large errors.Overall, HuberLoss or SmoothLoss, which is a special case of HuberLoss, combines MAE and MSE, making it more robust to outliers and converging faster.

Specifically, SmoothL1Loss combines the advantages of MAE and MSE, and the mathematical expression is:

$$E_{Medal} = \sum_i E_{iMedal} = \sum_i p_{iMedal} + \sum_j \sum_t \alpha_t p_{jtMedal} \tag{7}$$

Among them, x represents the error between the true value and the predicted value, and β is a hyperparameter. The SmoothL1 function uses a quadratic function for small errors, approximating MSE, and a linear function for large errors, approximating MAE. Therefore, it possesses the advantages of both, positively impacting our model training.

Therefore, we ultimately selected SmoothL1Loss as the final loss function for the model based

on experimental performance (MAE, RMAE, MSE, RMSE, R^2 metrics) from HuberLoss and SmoothLoss.

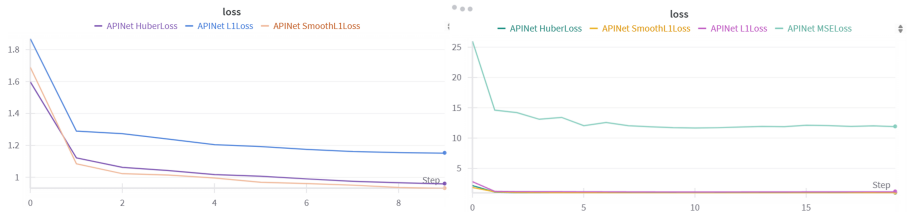

Figure 8: Loss of Different Methods

5.3 Predictions for the 2028 Olympic Games medals

After completing the training of the regression model, we began to consider how to predict the 2028 Olympics.

First, we need to collect information related to the 2028 Los Angeles Olympics and simulate data for 2028 Los Angeles, such as the host, the number and types of events, participating countries, athlete information, etc. Then, we will process the data as we did previously.Specifically, based on the Olympic spirit, we optimistically assume that world peace is imminent. Therefore, we have included Russia, which could not participate in 2024 due to the war, among the participating countries. Considering the CAS sanctions, we believe that Russia will use the ROC participation code as it did in 2020. As for the participating athletes, since the countries have not yet provided their lists, we assume that the strength of each country in various events will not change drastically. Thus, taking into account the normal updates and replacements of athletes, the overall team strength remains largely unchanged. Consequently, we have removed factorsSpecifically, Russia and Belarus will participate as AIN in 2024, and there is no relevant data, so we will use the team characteristics of the 2020 athletes as a reference.

Based on our thoughts on the Olympic Games, we have Hypothesis 2, knowing that the number of prizes is discrete data and all are positive integers. Since the competition events occur independently in sequence, we can use the Poisson distribution to simulate.The Poisson distribution is determined by a parameter λ, known as the average rate of occurrence or average frequency of events. It represents the average number of times an event occurs within a given time or space. In our model, the predicted mathematical expectation is used as this parameter.Through it, we know the probability of the event occurring$k$ times ($k > 0$) is :$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Monte Carlo simulation is a method that uses random sampling to approximate the probability distribution of complex systems or processes. By conducting a large number of random trials or simulations, the behavior and outcome probability distribution of the system can be estimated. Thus, to predict the number of medals for all countries in 2028, as well as the number and specific countries that will win medals for the first time, we only need to perform one Monte Carlo sampling in this probability model. By repeatedly performing Monte Carlo sampling within a Poisson

Figure 10 illustrates the overall framework and computational workflow of our APINet model.

Using APINet, we predicted the number of gold medals and total medals for each participating country in the 2028 Olympics. The results are visualized in the medal heatmap shown in Figure

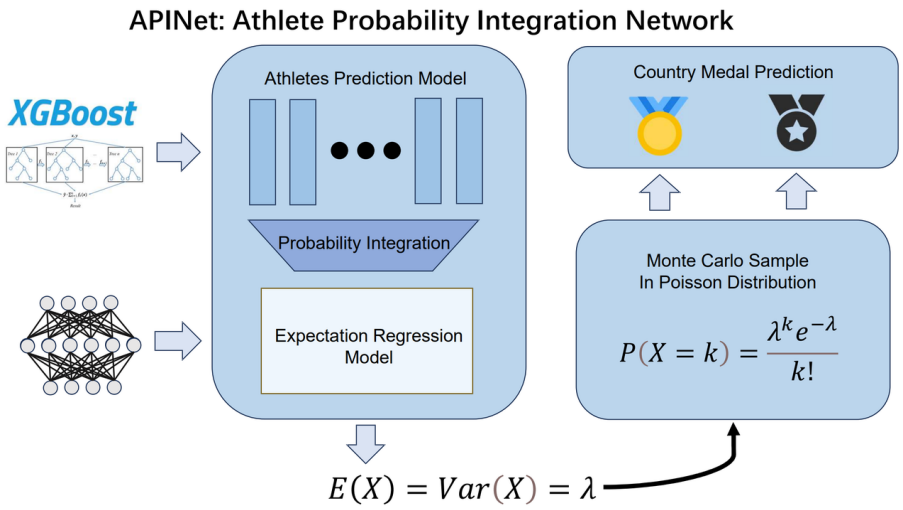

Figure 9: Framework of Model 1

11.

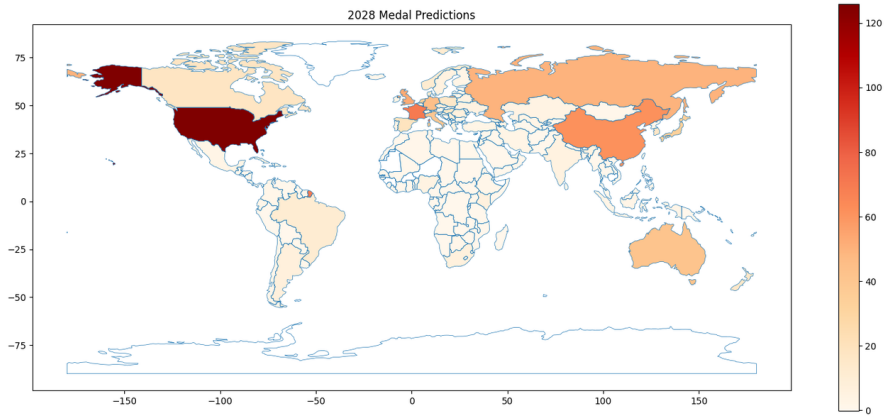

Figure 10: Heatmap Based on Prediction Results

We selected the most representative countries to compare the predicted results for 2028 with the medal standings from 2024. Additionally, we used box plots to present the predicted medal count ranges for 2028. The results are shown in Figure 12.

To more clearly highlight the changes in the medal standings between 2028 and 2024 and to directly observe the progress or decline of each country, Figure 13 visualizes the top ten countries with the most significant improvements and declines.

Our analysis shows that countries with the greatest improvements include Russia, Germany, China, and the United States, while countries with the most significant declines include France,

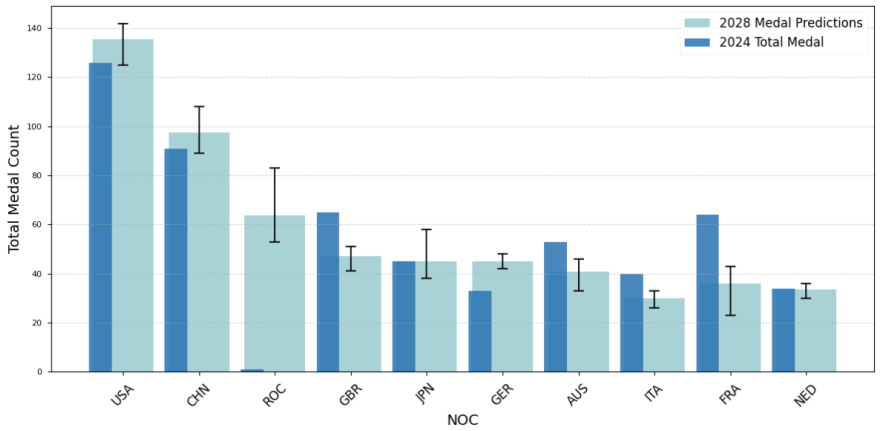

Figure 11: Prediction Range for Top Ten Countries

Germany, and Australia. The top ten countries with the largest degrees of improvement and decline are detailed in Figure 13.

Russia's dramatic improvement is primarily due to the 2024 Olympics, where athletes from Russia and Belarus participated as neutral athletes, and Russian athletes won only one medal. With a hopeful vision for global peace and the spread of the Olympic spirit, we formulated Assumption 3 in our analysis. For the 2028 simulation, we replaced the incomplete 2024 data for Russia with data from the 2020 Russian Olympic Committee, enabling a more comprehensive prediction of the country's athletic potential.

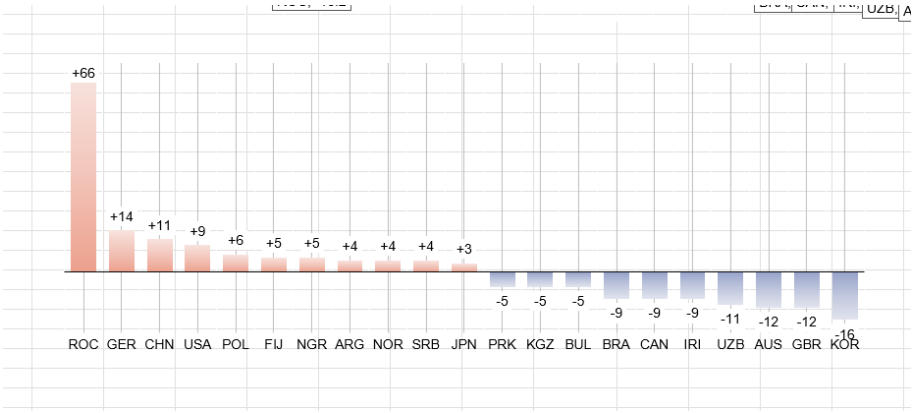

Figure 12: Top Ten Countries with Progress and Decline

As described above, for countries that might win a medal for the first time, we conducted extensive sampling to obtain the distribution shown in Figure 13. It can be observed that this distribution follows a Gaussian distribution, which aligns with the characteristics of Monte Carlo sampling.

We selected the point with the highest probability density function (PDF) value, 13, as the most likely prediction for the number of medals. For the prediction range from 0 to 63, the probabilities are presented according to the Gaussian distribution in the figure. This provides a clear and probabilistic basis for forecasting the medal outcomes for these emerging medal-winning countries.

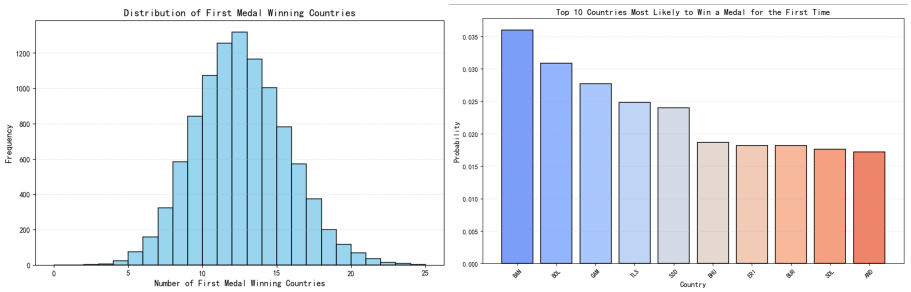

Figure 13: Top Ten Countries with Breakthrough Prediction and Probability

We also ranked all previously non-winning countries based on their expected total medal counts and calculated the probability of each country winning a medal. The top 10 countries most likely to win their first medal in 2028, along with their probabilities, are presented in Figure 15.

When the number of simulations is relatively small, deviations may arise. This is because the expected medal counts for non-winning countries differ only slightly in the model, leading to variations in results across multiple sampling iterations. To address this, in addition to the expectation-based approach, we conducted extensive Monte Carlo simulations. By approximating probabilities through frequencies, we validated the consistency of these results, providing a robust basis for identifying countries most likely to achieve their first Olympic medal.

To further explore the relationships among sports events, countries, and medal counts, we first filtered out country-sport pairs where the number of gold medals exceeded 10 in the past decade. Using this data, we constructed a Sankey diagram (Figure 14) to visualize the strong associations between certain countries and specific sports. Observing these connections, we then applied clustering methods to uncover deeper patterns within the data.

We constructed a clustering dataset to analyze the relationships among sports events, countries, and medal counts. For categorical features, we applied one-hot encoding to NOC and Sport, while for numerical features, we standardized gold (number of gold medals) and total_medal (total number of medals). Several clustering methods were tested, including KMeans, DBSCAN, Agglomerative Clustering, and Spectral Clustering. After tuning parameters, evaluating, and comparing the methods, KMeans achieved the highest Silhouette Score of 0.26, outperforming the others. Its clustering results were satisfactory, with fast computation and good scalability for large datasets. Therefore, we selected KMeans as the most effective clustering method. Using the elbow method (Figure 16), we determined the optimal parameter, n_cluster=3n_cluster = 3. After clustering, we performed PCA to reduce dimensionality and plotted the results as a 2D PCA scatter plot (Figure 17) and a 3D PCA scatter plot (Figure 18), providing a clear visualization of the clustering outcomes.

From the visual analysis of the clusters formed by the KMeans algorithm, we observed significant differences among the three clusters. Statistical analysis of the clustering results revealed the following patterns:

- Cluster 1 has the lowest average gold medal count, 1.933875.
- Cluster 2 has the highest average gold medal count, 41.25.
- Cluster 3 has a moderate average gold medal count, 13.439838.

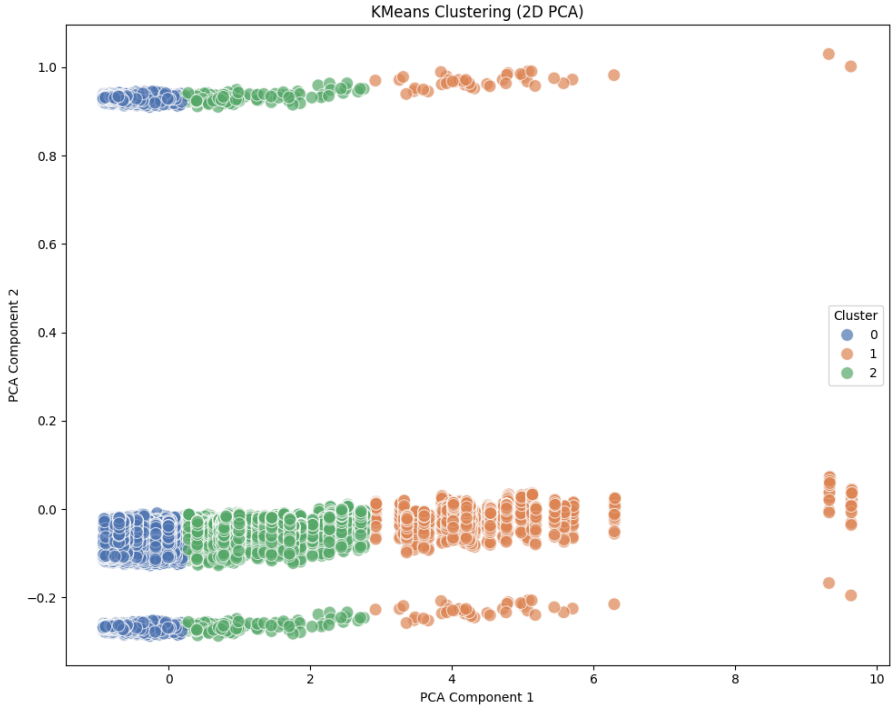

Figure 14

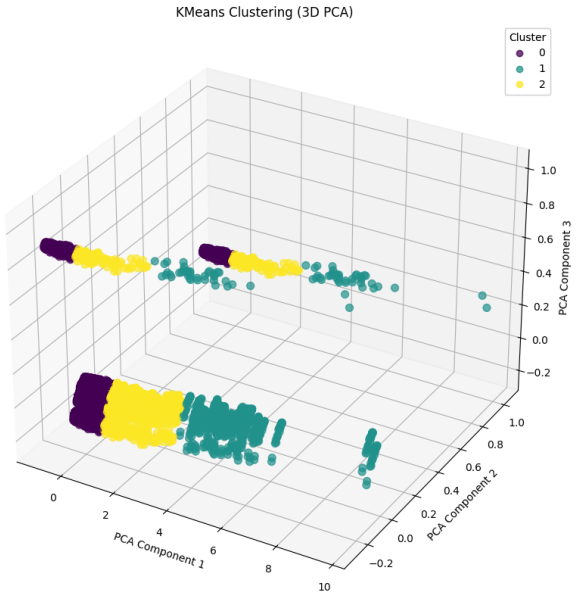

Figure 15

By identifying the five most prominent sports and countries in each cluster, we found:

1. Cluster 2: This cluster includes countries and sports with the highest average medal counts, corresponding to nations that dominate specific major sports fields. Examples include China in diving, the United States in swimming, and the former Soviet Union and Russia in gymnastics. These countries possess near-monopolistic strength in these events, securing a significant share of the medals.

2. Clusters 1 and 3: Both clusters feature similar sports, such as athletics, swimming, boxing, and canoe slalom, which are highly competitive and have large numbers of medals available. However, their characteristics differ:
    - Cluster 3: Includes major sporting nations like Australia, Italy, France, Japan, and Germany. These countries, while not dominant, can occasionally secure medals from the monopolistic countries, typically achieving about one-third to one-fourth of their medal count.
    - - Cluster 1: Primarily consists of smaller European countries with limited populations and resources. These countries struggle to compete in resource-intensive sports requiring extensive training systems and infrastructure.

This analysis suggests that dominant nations should focus on maintaining their monopoly in specific major sports, while major sporting nations should prioritize cost-effective and high-profile sports such as athletics, swimming, and boxing, leveraging their population base and resource advantages. Meanwhile, smaller, wealthy European countries may find opportunities in niche sports with high entry barriers, such as equestrian events, to establish competitive advantages.

For the impact of host country status, we utilized the APINet model by setting the isHost variable to 0 or 1, representing whether a country is the Olympic host. Using a causal inference framework, we constructed a Causal Directed Acyclic Graph (Causal DAG), where isHost was treated as the treatment variable and Medal Count (Medal) as the outcome variable. To enhance the scientific rigor and interpretability of the model, we introduced additional key variables as effect modifiers, including the country's historical medal count (LastMedal), number of sports participated (NumSports), number of athletes (TeamNum), and medal performance metrics such as gold medal rate and total medal rate (AvgGoldRate). Additionally, we identified and excluded confounding variables that do not directly impact medal counts but may influence the analysis results, such as gender. While such variables are correlated with other features, they lack a direct relationship with medal acquisition. By maximizing the exclusion of confounder interference, we achieved a more accurate investigation into the relationship and potential causality between isHost and medal counts.

he weighted path diagram derived from path analysis provides insights into the constructed Causal Directed Acyclic Graph (Causal DAG). By combining domain knowledge with data distribution analysis, we examined the direct and indirect effects among variables. The visualization of the path analysis and weighted causal graph clarifies how host country status (isHost) influences medal counts (Medal) through various pathways.

Direct Path:

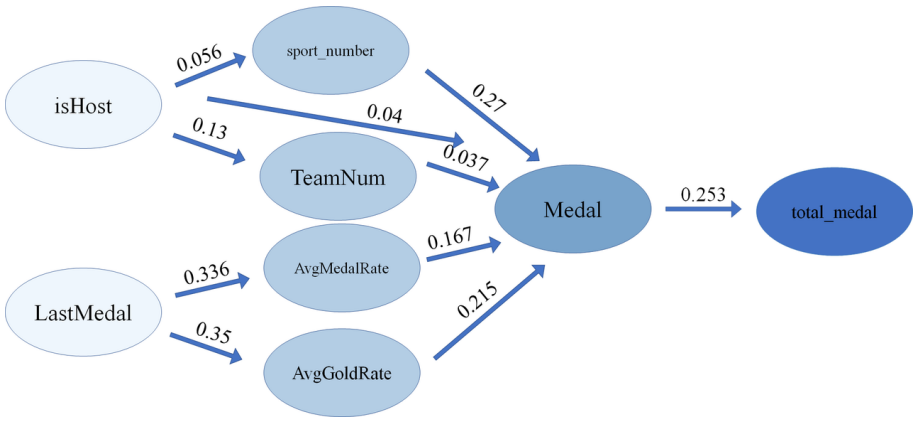

Figure 16

The path isHost → Medal has a high path weight, indicating that host country status directly enhances medal counts. This effect likely arises from home-field advantages, including familiarity with competition environments, potential referee bias, and support from local spectators.

Indirect Paths:

1. isHost → NumSports → Medal: Host countries tend to add or strengthen events in which they excel, providing their athletes with more opportunities to compete for medals.

2. isHost → TeamNum → Medal: Host countries often allocate more resources to increase the number of participating athletes, which significantly raises the likelihood of their athletes winning medals.

These pathways demonstrate how host country status, through both direct and indirect mechanisms, can substantially influence a nation's medal outcomes in the Olympics.

6 "Great Coaches" Influence Model

In the previous models, we successfully predicted the 2028 Olympic medal standings and the potential performance of various countries. We also conducted relevant studies and comparisons of event-related features. However, these models were primarily based on athletes. On the Olympic stage, an athlete's performance is not only related to personal ability but also closely tied to the guidance of their coach. Great coaches can help athletes reach higher levels, optimize strategies, and provide psychological support, all of which significantly impact medal outcomes. Therefore, great coaches also have an influence on the medal standings. Studying the "Great Coaches" influence model can positively enhance the accuracy of our predictions.

To explore the impact of "great coaches" on a country's medal count, we plan to use Python's DoWhy library to construct a causal effect model and perform counterfactual reasoning to analyze the intervention effects of coaches.

6.1 Causal Effect Analysis

We plan to use Python's DoWhy library to determine the causal relationships between sports disciplines and countries. First, we map some of the factors in our model to the components in the DoWhy library. We treat each country as a Unit. It is important to note that countries in different years are considered distinct Units. Next, we treat various country attributes as Variables, and interventions on countries, such as whether to invest in great coaches, are treated as Treatments. Variables that influence both the Treatment selection and the outcome are referred to as Confounders.

The primary goal of this model is to explain causal effects. Specifically, applying a Treatment to a Unit results in an outcome that differs from what would occur if the Treatment were not applied. The difference between these two outcomes represents the causal effect.

We define the outcome when no Treatment is applied as the counterfactual outcome, denoted as Y(0)Y(0), and the outcome when the Treatment is applied as the factual outcome, denoted as Y(1)Y(1). Thus, the individualized treatment effect (ITE) is expressed as Y(1)−Y(0)Y(1) - Y(0).

To evaluate the impact of "great coaches," we plan to utilize counterfactual reasoning. This involves hypothesizing a scenario opposite to the actual situation to explore the potential outcomes through logical reasoning. By comparing the factual and counterfactual results, we can better assess the influence of "great coaches" on a country's medal performance.

6.2 Dataset Construction

To support the model analysis, we constructed a specialized dataset that includes the following key variables:

- NOC (National Olympic Committee Code): A unique identifier for each country.
- Presence of "Great Coaches": A 0-1 label indicating whether a country employed "great coaches" for a specific sport in a specific year.
- Years: The Olympic year.
- TeamNum (Number of Athletes): The total number of athletes participating from each country.
- Sports: The types of sports events in which the country participated.
- AvgGoldRate (Average Gold Medal Rate): The rolling average of the country's historical gold medal rate.
- AvgMedalRate (Average Medal Rate): The rolling average of the country's historical total medal rate.
- LastMedal (Medals in the Previous Olympics): The total number of medals the country won in the previous Olympics.
- Medal (Current Medal Count): The actual number of medals won by the country in the current Olympics.

This dataset serves as the foundation for evaluating the causal effects of employing "great coaches" on a country's medal performance, enabling robust analysis and counterfactual reasoning.

6.3 Model Construction and Causal Inference

To explain the impact of "great coaches" on a country's Olympic medal count, we first define the core variables involved in the model. Medal count (including gold medals and total medals) serves as the outcome variable, while the decision to invest in "great coaches" is the treatment variable. Additionally, a country's sports resources (e.g., number of athletes, historical medal counts), the host country effect (whether the country hosts the Olympics), and the number and types of sports events are considered confounders, as they may simultaneously influence the choice of employing great coaches and the resulting medal count. The relationships among these variables can be visually represented using a causal graph, where nodes represent variables and edges illustrate causal pathways.

Next, we need to determine whether the causal relationship between the treatment variable and the outcome variable can be identified through the identification step of the causal inference framework. Using the DoWhy library, we analyze causal effects by making assumptions based on the causal graph and applying principles such as the backdoor criterion or instrumental variables to determine identifiability.

6.3.1 Backdoor Criterion

The backdoor criterion states that if all common causes of the treatment variable AA and the outcome variable BB are observed, the causal effect can be identified by adjusting for these confounders. The causal effect is expressed as:

$$P(B|do(A)) = \int P(B|A,W)P(W)dW \tag{8}$$

where W represents the confounders (i.e., common causes of A and B).

6.4 Instrumental Variables

Instrumental variables are special variables that influence the treatment but do not directly affect the outcome. They are unaffected by any unobserved confounders impacting the outcome. When unobserved confounders exist between the treatment and outcome variables, instrumental variables can be used to estimate the causal effect:

$$\text{Effect} = \frac{E[B|Z=1] - E[B|Z=0]}{E[A|Z=1] - E[A|Z=0]} \tag{9}$$

where Z represents the instrumental variable.

6.5 Implementation with DoWhy

Using the DoWhy framework, we verified whether the causal effect is identifiable. The results showed that the backdoor paths were fully blocked, confirming that the causal effect can be identified.

This analysis provides a rigorous foundation for estimating the influence of "great coaches" on Olympic medal outcomes, leveraging both theoretical principles and computational tools to ensure robust causal inferences.

6.5.1 Causal Effect Estimation

Due to the non-linear relationships between variables, traditional methods cannot directly identify causal effects. Therefore, in the causal effect estimation phase, we integrated XGBoost with DoWhy. The estimation process was completed in three steps:

1. Propensity Score Modeling: We trained a classification model using XGBoost to predict the probability of each country hiring a "great coach" based on confounding variables such as GDP, the number of athletes, and the host country effect. This propensity score is denoted as $P(Coach = 1 Confounders) P(\text{Coach} = 1 \mid \text{Confounders})$

2. Counterfactual Prediction: Two regression models were trained using XGBoost to predict medal counts for the treatment group (countries hiring "great coaches") and the control group (countries not hiring "great coaches").
   - The first model predicted the medal count under the treatment condition (Y(1)Y(1), i.e., hiring a "great coach").
   - The second model predicted the medal count under the control condition (Y(0)Y(0), i.e., not hiring a "great coach"). Using these two models, the individualized treatment effect (ITE) was calculated as:

3. $3. ITE = Y(1)Y(0) \text{ITE} = Y(1) - Y(0)$

4. Average Treatment Effect (ATE): By averaging the ITE values across all samples, we obtained the average treatment effect (ATE), representing the average improvement in medal counts attributable to hiring a "great coach."

6.6 Results

After controlling for confounding variables, the ATE indicated a modest average increase in medal counts due to hiring "great coaches." However, notable improvements were observed in specific scenarios:

- India: The shooting team showed a significant improvement, with an estimated increase of 4.0 medals.
- France: Athletics was expected to see an increase of 3 medals.
- China: Swimming was projected to gain an additional 2 medals. These results highlight the positive impact of hiring "great coaches" in specific sports and countries, offering actionable insights for optimizing coaching investments in future Olympic preparations.

1. National Economic and Cultural Strength, Population, and Other National Factors

   From the analysis of the athlete database features in Model 1, it is clear that national factors have a significant impact on the number of medals a country wins. Important indicators of a country's power include its economic, cultural, technological, and military strength, as well as its population size. In today's world, where people increasingly desire peace, the most prominent factors among these are economic and cultural strength, and population size.

A combined consideration of these two factors—per capita GDP and population size—can provide a more accurate picture. According to related studies, countries with higher per capita GDP and larger populations tend to have a higher likelihood of winning medals.

For example, countries with high per capita GDP, such as Luxembourg, despite its high GDP, have only won six medals in total due to its small population of only a few hundred thousand. In contrast, countries like Switzerland, Norway, the United States, and Sweden, which have high per capita GDP and larger populations, have better records. Even though China's per capita GDP ranks around 70th globally, its large population has helped it achieve good results in the Olympics.

Recommendation: Strengthen sports infrastructure, leverage the demographic dividend, and large population countries can select and cultivate outstanding talents to tap into their athletic potential. At the same time, relevant policies should be established, and long-term sports strategies should be implemented to promote economic and cultural development.

2. Host Country Effect

In Model 1, we observed that the variable is_host has a significant impact on the probability of athletes winning various medals. This is because host countries compete in their familiar environment, which allows athletes to adjust their mindset more easily and receive local support. To further explore this, we can compare the average number of medals won by countries when they host the Olympics versus when they do not. To minimize errors, we will select the top three countries that have hosted the most Olympic Games (the United States, France, and Australia). The data is shown below.

It is evident from the graph that these three countries have seen a noticeable increase in the number of gold, silver, and bronze medals when they act as host countries.

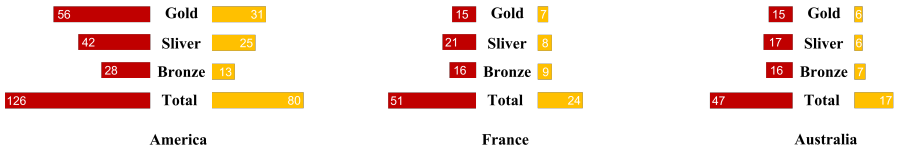

Figure 17: Literature Review Framework

Recommendation: Maximize the home advantage by developing detailed training plans in advance. Strengthen audience support by creating a better atmosphere for spectators, which can provide more psychological support to athletes. Additionally, analyze historical data and draw on the successful experiences of previous host countries. For example, focus on investing in local advantage sports, which can help improve medal prospects for host countries.

3. Distribution of Event Numbers

In the previous analysis, we clearly observed the connection between the number of events and the number of medals for each country. Since every country has its own strong events, the number of these advantageous events will also impact the total number of medals won. Therefore, we plotted the average number of medals and the average number of events for different categories of events in recent years.

Recommendation: Countries should optimize the allocation of resources across different events by increasing investment in their strong areas while also identifying and fostering events with growth potential. Encouraging diversified development in sports can help countries build a broader competitive edge. Additionally, establishing a development system for new events, with specific plans in place, can help these events gain more medals in future competitions.

7 Conclusion

7.1 Strength and Weakness

Through shifting the focus of medal prediction from country-level analysis to athlete-centered factors, we proposed the APINet model for predicting Olympic medals. By exploring features associated with Olympic medals, we leveraged the powerful machine learning algorithm XGBoost as the Athlete Prediction Model (APM), successfully capturing complex data relationships and achieving outstanding performance. Additionally, we examined influential factors such as the role of "great coaches." However, the model still has limitations that require further optimization. Below, we analyze the strengths and weaknesses of the model in detail:

Advantages:

1. Human-Centric Approach

   Unlike most existing models that focus on medal predictions at the country level, we broke through traditional thinking by creating a human-centered model. By aligning with the Olympic spirit of pushing human limits, this model places athletes at its core.

2. High Precision

   Our model strictly adheres to the competition's data scope without introducing highly correlated but external factors, such as GDP, commonly used in prior studies. Instead, we utilized all relevant variables within the dataset to ensure precise predictions.

Disadvantages:

1. Limited training due to time constraints

   Due to time limitations, we were unable to train the model to its full potential, and it did not achieve the best performance in all areas. However, this is something that can be overcome.

7.2 Future Work

In previous research, we identified key factors influencing Olympic medal counts, such as GDP, which correlates with gold medal wins. Future models could reassess GDP and include additional factors like host country status, population, geographical area, and healthcare expenditures, moving beyond the simplistic use of NOC as a feature.

An athlete's performance on the day, especially if they are from the host nation, is crucial. We can quantify this "performance index" and integrate it into our model for a more realistic approach.

Machine learning algorithms play a vital role in our evaluation. APINet's plug-and-play framework supports various machine learning and deep learning libraries (scikit-learn, XGBoost, CatBoost, PyTorch, TensorFlow, Keras) and allows easy swapping of the Athlete Prediction Model (APM) to explore better-performing algorithms with accurate probability distributions.

Recent advancements in machine learning have achieved significant results in feature extraction and representation learning. We plan to introduce advanced feature fusion mechanisms and multi-scale attention mechanisms to enhance APINet's representational capacity, improving prediction accuracy and robustness. Leveraging these innovations will further optimize our model and enhance prediction accuracy.

References

[1] Olympics.com, https://olympics.com/en/paris-2024/medals

[2] https://en.wikipedia.org/wiki/2028_Summer_Olympics

[3] WANG Guofan, XUE Erjian, TANG Xuefeng. Prediction of the Number of Medals in International General Games: With Beijing Olympic Games as an Example [J]. Journal of Tianjin University of Sport, 2010, 25 (01): 86-90. DOI:10.13297/j.cnki.issn1005-0000.2010.01.007.

[4] BAL MER NIGELJ, NEVILL ALAN M, WILLIAMS A MARK. Home advantage in the Winter Olympics (1908-1998) [J]. J Sport Sci 2001, 19:129-139.

[5] HOFFMAN ROBERT GI NG et al. Public policy and Olympic success [J]. Appl Eco Letters 2002, 9:545-548.

[6] X. Bian, "Predicting Olympic Medal Counts: the Effects of Economic Development on Olympic Performance," *Honors Projects*, vol. 13, Jan. 2005.

[7] P. Y. K. Ho, J. S. Lee, and T. K. W. Lee, "Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics," *IEEE Access*, vol. 11, pp. 10840553-10840553, 2024. doi: 10.1109/ACCESS.2024.10840553.