Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2519878

Olympic Medals: What Information can We Get from Historical Data?

Summary

Due to the widespread popularity of the Summer Olympics, the event attracts a large number of fans each year. Beyond following the performances of individual athletes, fans are also highly interested in the medal counts of each country. Since the number of medals is influenced by numerous factors, we aim to model the medal counts for 2028 using various historical data. Specifically, our model predicts the medal counts for each sport for each country in 2028, this model is more flexible, and can get the medal counts for each country by summing up the separate medal counts predictions in different sports.

We first make feature engineering for training dataset, important features include but not limited to host country, athlete counts, mean athlete score, events number, etc.

Then we did elementary data analysis (eda), which visualized the trend of features and medal counts, and we did two types of correlation test, Pearson correlation and Mutual Information score, to statistically test the correlation between each feature and medal counts. This answers **question 1.3**, as the features mean_athletes_score and medal_ewa shows the importance of sports in each country, there is a strong correlation between the host country and medal counts, and the number of events is added as a feature to limit the range of medal count, as well as to interact with other features like mean_athletes_score. It also answers **question 3**, as we explore more features that have a high correlation with medal counts, such as athlete counts, countries' rank k-means, etc.

Then we use random search to find the best parameters for the neural network, the R^2 for Total Medals for each sport, country, and year is 0.6795, and summing up different sports, the R^2 is 0.7929. We predict the results for 2028, in 2028, the USA has Gold 67, Silver 54, Bronze, 54, and Total 175. We use the difference between 2024 and 2028 medal counts to find the top countries with the most improvement on gold medal counts or total medal counts, are the USA and ROC, with the most decline are CHN, UZB, and FRA, and we plot the trends. This answers **question 1.1**.

Since we want to explore whether the countries that have yet to earn medals can earn at least one medal in 2028, we trained a NN classifier, the target is binary, about whether the country has at least one medal in the sport or not, and calculate the probability of a country can earn at least one medal in 2028 by using 1 minus product of all probability of losing the specific sport. The model predicts that GAM has a probability of 0.3688 to get the first medal, which has odds of 58.4%. This answers **question 1.2**.

Finally, we explore potential great coach effect data by finding the data that has a low predicted value, but has a high true value, which is caused by unexpected influences that are not reflected by the features, which can be the potential great coach effect. The results are 1980 country GDR in sport ROW, 1984 USA in sport SWM, and 1976 URS in sport WRG. This answers **question 2**.

Key Terms: Correlation Tests, Time Series, Partial Autocorrelation, Neutral Network

CONTENT

1 Introduction......3
1.1 Background......3
1.2 Restatement the problem......3
1.3 Modeling Workflow......4
2 Data preprocessing......4
2.1 Data Cleaning......4
2.2 Feature Engineering......5
2.2.1 Data Integration......5
2.2.2 Medal Score......6
2.2.3 Categorical Features Encoding......6
3 EDA and Correlation Analysis......8
3.1 Athlete Count Effect......8
3.1.1 Data Visualization......8
3.1.2 Pearson Correlation......9
3.1.3 Mutual information......10
3.2 Host Effect......10
3.2.1 Exploratory Data Analysis (EDA)......10
3.2.2 Statistic Test......12
4 Handling Data Leakage and Creating Time Series Features......13
4.1 Lagging the feature......13
4.2 Sequential Analysis......14
4.3 Partial Autocorrelation Function......16
4.4 Correlation heatmap......17
5 Neural-Network Model......17
5.1 Data normalization......17
5.2 Hyperparameter......18
R² Results:......19
6 Result......20
6.1 Performance change......20
6.2 Predict First-Time Medal-Winning Countries......21
6.3 Great Coach Effect......22
7 Model Assessment......23
7.1 Strength......23
7.2 Future improvement......23

1 Introduction

1.1 Background

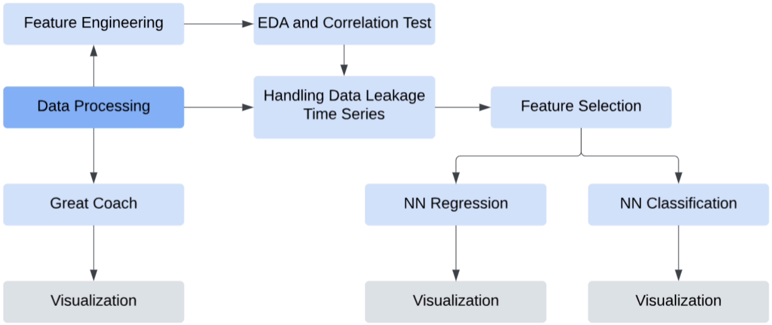
The Olympic Games are a globally celebrated event held every four years, bringing together the most outstanding athletes from around the world to compete for medals at the host city of the year. However, these medals represent more than just the personal achievements of individual athletes—they reflect shifts in the global sports landscape and embody the economic strength and technological investments of different nations. Moreover, the factors influencing an athlete's ability to win medals go beyond personal skill, including the cultural background, economic investment, and training systems of their country. By predicting the medal counts for future Olympic Games, we can inspire public interest in sports and promote nationwide fitness initiatives. Therefore, fans always pay special attention to the medal table, which can also be roughly analyzed through some historical data.

1.2 Restatement the problem

Based on the provided data and information, we need to solve the following problems:

- Develop a model to predict medal counts for each country at the 2028 Los Angeles Summer Olympics. Focusing on gold, silver, bronze and total medals, and estimates of uncertainty and precision.
- Analyze how the number and type of events affect medal counts, identify key sports for different nations, and assess how the host country's event choices influence results.
- Use the model to project the 2028 medal table, identify countries likely to improve or decline compared to 2024, and predict which countries might win their first medal along with associated probabilities.
- A "great coach" effect that may influence medal counts. Analyze data to evaluate this effect, estimate its contribution.
- Select three countries and pinpoint specific sports where hiring an exceptional coach could enhance their medal prospects, along with an estimate of the potential impact.
- Explain the unique insights revealed my our model about Olympic medal counts, and how these insights may offer help for Olympic committees.

1.3 Modeling Workflow



We follow the process to build the neural network models. The prediction of the model is for each year, each country, and each sport. Total medals for each country in each year can be obtained by summing up the values in different sports. The purpose of predicting each sport's medal count for each country, each year instead of the medal count for each country each year is for future wider use.

2 Data preprocessing

2.1 Data Cleaning

Due to certain historical events, the Olympic Games were not successfully held, or data is missing, such as during World War I and World War II, when no athletes participated. Therefore, we need to clean up this missing data to ensure the accuracy of the model. The specific data processing steps are as follows:

1. Remove the data from the periods of World War I and World War II to avoid impacting the overall model.
2. Since 1906 was not a standard Olympic year, it is marked as 1906*. To maintain consistent formatting for easier calculations later, we will change 1906* to 1906.
3. Since Russia is banned from participating in the 2024 Olympics, and Russian athletes can only compete as neutral athletes. We will merge the data of athletes under AIN in 2024 into the data for Russian athletes. We also account for RUS and ROC as the same country, which stands for Russian.

2.2 Feature Engineering

2.2.1 Data Integration

Features: Host, Medal_EWA, Events, Athletes_Counts, Athletes_Proportion, NOC(replaced with K-means)The data we need to use is scattered across different tables. To unify them and facilitate subsequent calculations, we performed the following processing:

1. Create a dataframe *train* that includes each medal count for each sport in each country in each year.
2. Merge the data frame *hosts* into *train*, in it, we map the host country's name to its corresponding NOC code, for example, {'United States': 'USA'}.
3. To facilitate model fitting later, we create a new column named Noc_if_host to mark whether the NOC is the host country in that year. If yes, it is marked as 1, otherwise it is marked as 0.
4. Add 2028 events in the dataframe program, then merge it with train. It is used to get the number of events set up for each sport in the Olympics each year, and name it Number_of_Events.
5. Create a column called athlete_count to count the number of athletes of each type in each country each year. Divide the athlete count by the total athlete count in the sport to get the athlete proportion of the NOC in the sport, named athlete proportion.
6. Create a feature that shows the trend of medal count. We balance historical and recent data by assigning higher weights to recent performance using ewa.

$$EWA_t = \frac{(1-\beta)*x_t+\beta*EWA_{t-1}}{1-\beta}$$

Where:

- EWA_t: exponentially weighted average at time t
- β :year trend weight
- $1 - \beta$:bias correction
- x_t: current data at time t
- EWA_{t-1}: EWA data of the previous year

We set β=0.8, which approximately looks back 5 times(20 years). To calculate current year performance, we give gold, silver, and bronze medals different weights: gold medals account for 0.5, silver medals for 0.3, and bronze medals for 0.2. And we neutralized the effect of the number of events on medal count.

$$x_t = \frac{gold\ weight\ *\ gold\ num\ +\ silver\ weight\ *\ silver\ num\ +\ bronze\ wright\ *\ bronze\ num}{Number\_of\_Event}$$

2.2.2 Medal Score

Score expected for each nation:

$$p_c = \sum_{year \le c} \frac{1}{s}$$

$$Score = \frac{p_c * e^m - h*(1-EDA)}{a}$$

where:

- p_c: the cumulative frequency of participation at current year c
- s: the sum over all years of all times that athletes participate
- a: the current participation time of a given athlete
- e: expectation for winning a medal
- m: total medal earned by the athlete
- h: whether the athlete's nation is hosting the Olympics
- EDA: the hosting effect 55.72/27.18 calculated in 3.2.1

We use participation frequency to represent the athlete's age. The expected score formula integrates participation frequency, medal achievements, and hosting effects. The cumulative frequency factor pcpc emphasizes frequent participation while penalizing inactivity, reflecting sustained competitiveness. Medal weighting e^m amplifies the impact of higher medal counts, prioritizing achievements over participation. The hosting adjustment −h*(1−EDA) accounts for the host nation's performance boost due to home advantage, where h=1 if hosting, balancing inflated scores. Finally, normalizing by participation time ensures fairness across career lengths, highlighting consistent performance over extended durations.

The score captures an athlete's potential to earn medals in future events, including the 2028 Olympics. Athletes with frequent participation, significant past achievements, and lower reliance on hosting advantages will likely score higher. Thus, this model enables accurate medal predictions by factoring in both historical performance and environmental conditions.

2.2.3 Categorical Features Encoding

Most models cannot handle categorical features directly. Although neural networks can use embedding, we decided not to use it due to the small data size. We encode each categorical feature based on its properties.

We replace NOC with numeric representations using K-means clustering based on key features like Gold, Silver, Bronze medals, and their Total.

- Group and calculate the mean values of Gold, Silver, and Bronze medals per NOC.
- Apply K-means clustering to generate cluster labels based on these features.
- Replace the original NOC with cluster-specific mean values (Cluster_Gold, Cluster_Silver, etc.).

Objective function: $J(x, \mu_i) = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$

Centroid update formula:$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$

Where:

- k: Number of country clusters.
- C_i: Set of data points in cluster i.
- x: Data point (Gold, Silver, Bronze, Total).
- μ_i: Centroid of cluster i.

Using K-means clustering, The objective function, $J(x, \mu_i)$, represents the total sum of squared distances between each country's medal counts (Gold, Silver, Bronze, Total) and its assigned cluster's centroid.

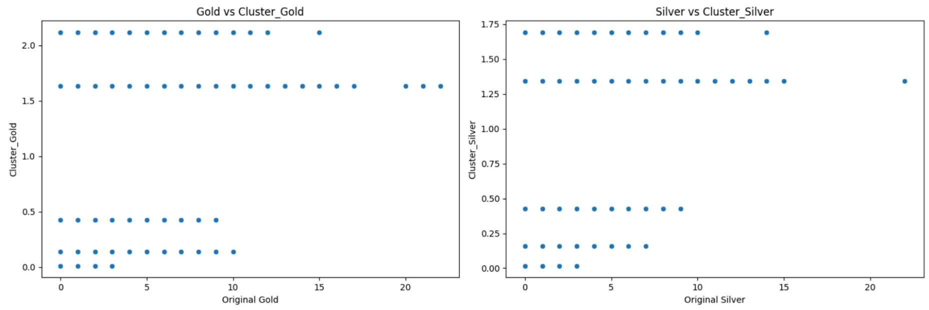


Figure 1: Difference Between Cluster and Countries medal

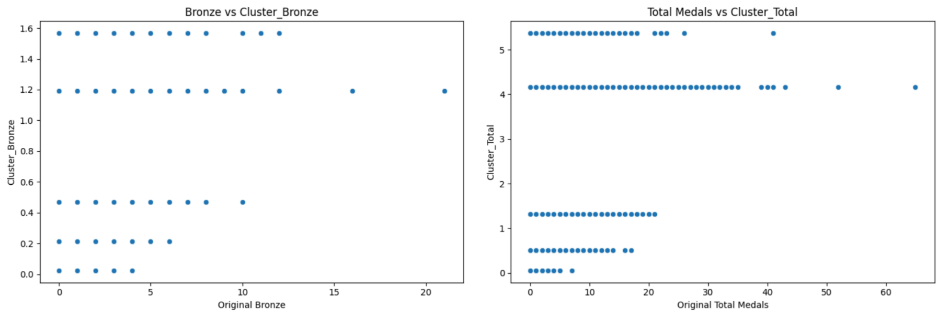
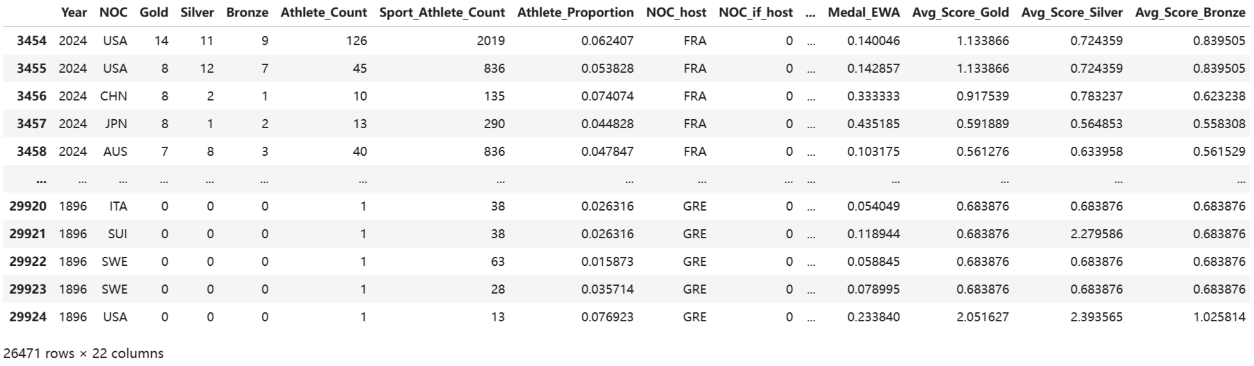


Figure 2: Difference Between Cluster and Countries medal

The differences between cluster mean values and original country-level statistics are minor, as demonstrated by scatter plots.This approach can be extended to other non-numeric representations based on country rank.



	Year	NOC	Gold	Silver	Bronze	Athlete_Count	Sport_Athlete_Count	Athlete_Proportion	NOC_host	NOC_if_host	...	Medal_EWA	Avg_Score_Gold	Avg_Score_Silver	Avg_Score_Bronze
3454	2024	USA	14	11	9	126	2019	0.062407	FRA	0	...	0.140046	1.133866	0.724359	0.839505
3455	2024	USA	8	12	7	45	836	0.053828	FRA	0	...	0.142857	1.133866	0.724359	0.839505
3456	2024	CHN	8	2	1	10	135	0.074074	FRA	0	...	0.333333	0.917539	0.783237	0.623238
3457	2024	JPN	8	1	2	13	290	0.044828	FRA	0	...	0.435185	0.591889	0.564853	0.558308
3458	2024	AUS	7	8	3	40	836	0.047847	FRA	0	...	0.103175	0.561276	0.633958	0.561529
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29920	1896	ITA	0	0	0	1	38	0.026316	GRE	0	...	0.054049	0.683876	0.683876	0.683876
29921	1896	SUI	0	0	0	1	38	0.026316	GRE	0	...	0.118944	0.683876	2.279586	0.683876
29922	1896	SWE	0	0	0	1	63	0.015873	GRE	0	...	0.058845	0.683876	0.683876	0.683876
29923	1896	SWE	0	0	0	1	28	0.035714	GRE	0	...	0.078995	0.683876	0.683876	0.683876
29924	1896	USA	0	0	0	1	13	0.076923	GRE	0	...	0.233840	2.051627	2.393565	1.025814

26471 rows × 22 columns

Figure 3: Train Table

3 EDA and Correlation Analysis

3.1 Athlete Count Effect

3.1.1 Data Visualization

We analyzed the relationship between athlete participation, athlete proportion, and total medals won for each Year, NOC, and Sport, and we specifically visualized the trend of three features for the United States (USA) and China (CHN) in the Olympics over different years.

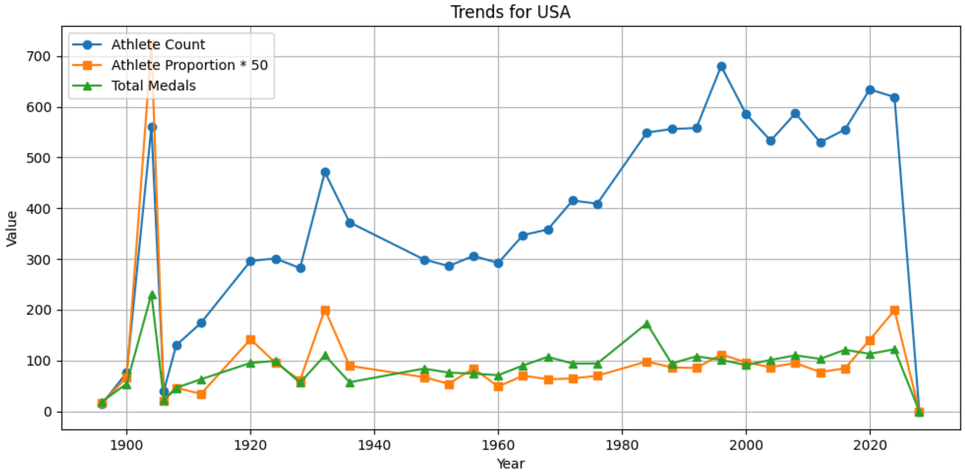


Figure 4: Trends for USA

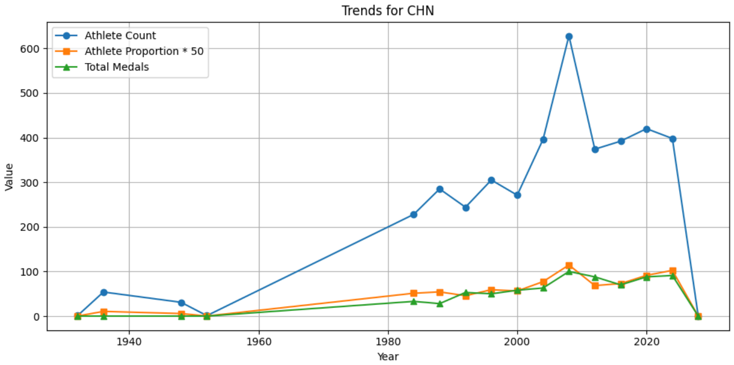


Figure 5: Trends for CHN

3.1.2 Pearson Correlation

Pearson correlation is used to test the linear relationship between two variables.

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}}$$

Where:

- r: correlation coefficient
- x_i: values of the x-variable in the sample
- $\overline{x}$: mean of the values of the x_variable
- y_i: values of the y-variable in a sample
- $\overline{y}$: mean of the values of the y-variable

The value range of r should be between -1 and 1, the absolute value of r closer to 1 means a stronger linear relationship between variables.

Pearson Correlation Matrix For Averaged Data

	Athlete_Count	Athlete_Proportion	Total_Medals
Total_Medals	0.406542	0.965225	1.000000

The result shows there is correlation between Athlete Count and Total Medals, and an extremely strong correlation between Athlete Proportion and Total Medals. Note that since we do not have access to the Athlete Count for 2028, we will time series data to avoid data leakage in the later part.

3.1.3 Mutual information

Mutual Information (MI) measures the dependency between two variables. It indicates how much information about one variable is gained by knowing the other.

$$I(X;Y) = H(X) - H(X|Y)$$

$$= \sum_{x\in\chi}\sum_{y\in Y} p(x,y)log_2\frac{p(x,y)}{p(x)p(y)}$$

Where:

- p(x,y): Joint probability distribution of X and Y.
- p(x), p(y): Marginal probability distributions of X and Y.

Higher MI values indicate stronger dependence between variables. The intuition of this formula is the relative entropy of p(x,y) and p(x)p(y), in simpler words, testing whether p(x,y) and p(x)p(y) are independent, that is, if they are independent of each other, p(x,y)=p(x)p(y), making I(X;Y)=0.

Mutual Information for Average Data

Athlete_Count →Total_Medals	0.4759
Athlete_Proportion →Total_Medals	1.3620

The graph, Pearson Correlation, and MI score all illustrate there is correlation between Athlete_Count/Athlete_Proportion and Total_Medals.

3.2 Host Effect

3.2.1 Exploratory Data Analysis (EDA)

We believe the host country will win more medals in the Olympics
We will use data visualization to test whether there is a host effect. First, we select Year, NOC, Gold, Silver, Bronze and Noc_if_host from the train table, and add up the number of Gold, Silver and Bronze to get Total Medals. Second, we group by NOC and calculate the average number of gold, silver and bronze medals for each country in history. After that, we filter out the results of the NOC, if the country is hosting, then we calculate the average number of medals for each country when it was the host country. Finally, we aggregated the data and visualized the number of medals each country had won historically and as a host country, resulting in the following picture.

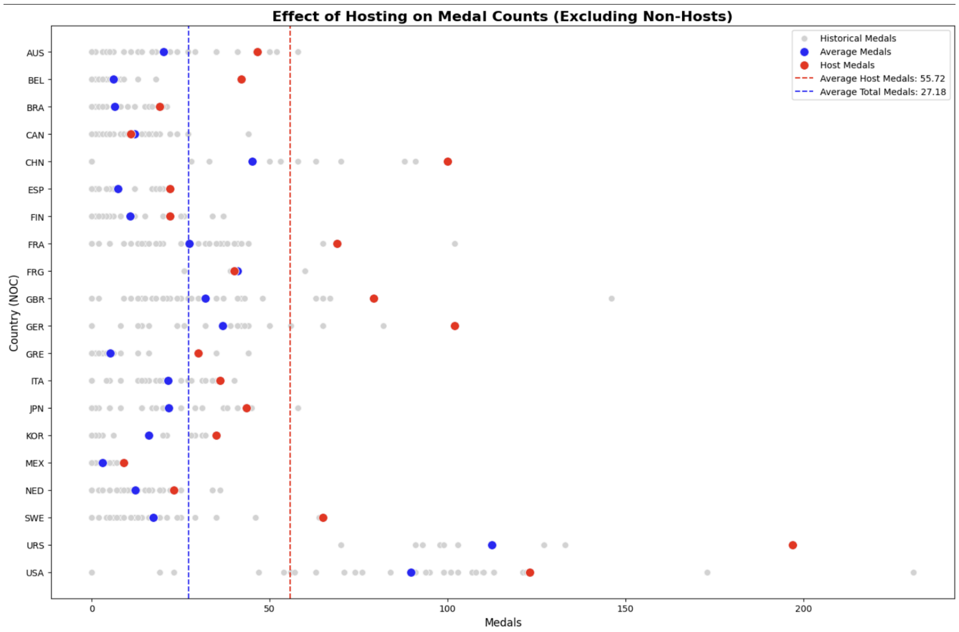


Figure 6: Effect of Hosting on Medal Counts (Excluding Non-Hosts)

- Average Host Medals: 55.72
- Average Total Medals: 27.18

In order to make the data more accurate and avoid the influence of too many bronze medals, we subdivided the number of gold, silver and bronze medals won based on the medal data of China and the United States, and also visualized it:

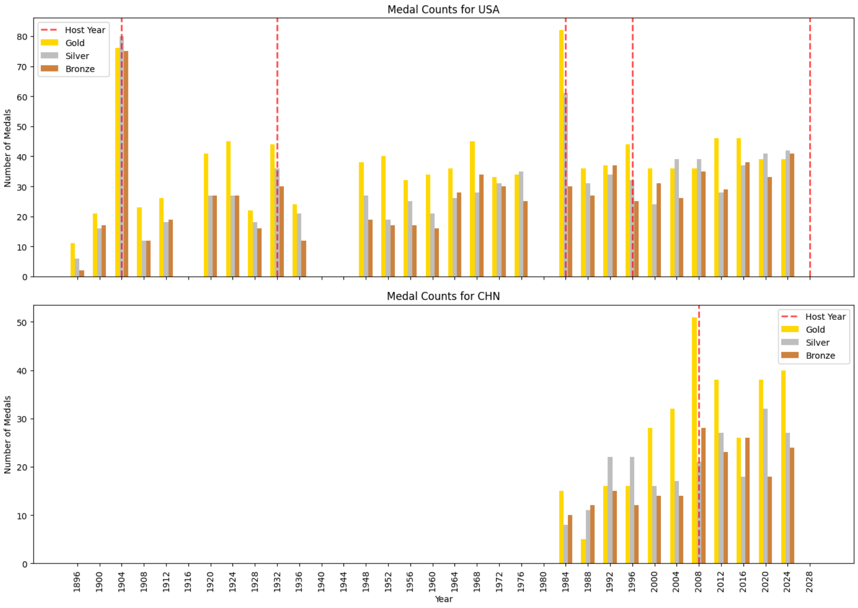


Figure 7: Num of Gold, Silver, Bronze by Year

The host effect impacts Olympic medal counts as demonstrated by host year peaks. For example, the USA achieved its highest medal count in 1904 (St. Louis) with 76 gold, 78 silver, and 77 bronze medals and experienced another spike in 1984 (Los Angeles). Similarly, China recorded its highest medal count in 2008 (Beijing), with 48 gold, 22 silver, and 30 bronze medals, reflecting targeted investments during their host year. Hosting advantages, such as increased athlete participation, familiarity with venues, and enhanced sports funding, provide measurable boosts in performance. Including the host country as a predictive feature for 2028 (Los Angeles) is critical to account for these systemic advantages. Additionally, incorporating the National Olympic Committee (NOC) feature enables the model to leverage historical trends, such as the USA's consistent average total medals of 95.67, alongside socio-economic and geopolitical factors, ensuring a comprehensive and accurate prediction.

3.2.2 Statistic Test

We aim to determine whether hosting the Olympics has a statistically significant impact on the number of medals won. To test this hypothesis, we conducted a two-sample t-test for the mean.

Since the total medal count data might exhibit skewness—where a small number of countries have extremely high medal counts—we applied a logarithmic transformation. This transformation helps reduce the influence of extreme values and ensures the data is closer to a normal distribution, making it more suitable for the t-test. Additionally, to avoid the impact of countries with zero medals, we filtered the dataset to include only countries that have hosted the Olympics at least once. This ensures a fair comparison between the medal counts in host years and non-host years for the same group of countries.
We calculated the total number of medals won in non-host years and host years separately. The Standard Error (SE) was computed using the following formula.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where:

- SE=standard error of medals earned among countries with and without hosting
- σ=standard deviation of medals earned among countries
- n=number of countries

We then performed the t-test with the following hypothesis.

- **Null hypothesis (H0)**: The mean number of medals won in host years is equal to that in non-host years, implying no host country effect.
- **Alternative hypothesis (H1)**: The mean number of medals won in host years is significantly higher than in non-host years, indicating a host country effect.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{{SE_1}^2 + {SE_2}^2}}$$

$\overline{x_1}$, $\overline{x_2}$ = mean of country total medals with and without hosting

n_1, n_2 = number of countries with and without hosting

SE_1, SE_2= standard error of medals earned among countries with and without hosting

With a significance level (α) of 0.05, the resulting p-value was 3.10e−11, which is far below 0.05. Therefore, we rejected the null hypothesis, concluding that there is a statistically significant difference in the number of medals won between host and non-host years. This result supports the existence of a host country effect, where the mean number of medals won in host years is significantly higher.

Finally, we visualized the distribution of total medal counts, with the blue section representing non-host years and the red section representing host years. The visualization shows a clear rightward shift in the distribution for host years, further illustrating that the total number of medals won in host years is typically higher than in non-host years.

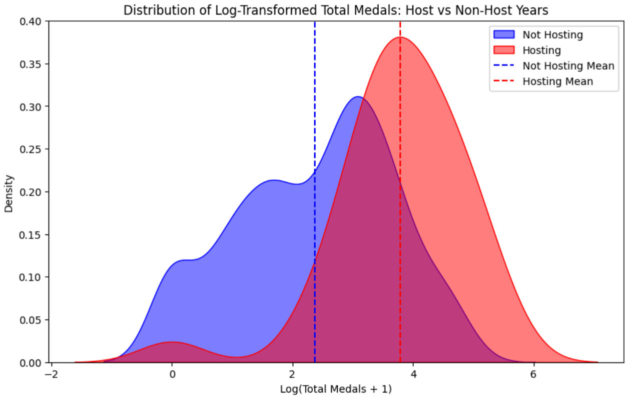


Figure 8: Distribution of Log-Transformed Total Medals: Host vs Non-Host Years

4 Handling Data Leakage and Creating Time Series Features

4.1 Lagging the feature

Since some features for 2028 are missing, such as athlete count, medal ewa, etc, we need to handle these features properly to avoid data leakage. We will use lag values to replace these values, for example, lag_1 of 2024 is 2020. Then we test the correlation between lag values with the current values, and the correlation between lag values with the medal counts. For example, between the number of athletes (Prev_Athlete_Count) and athlete proportion (Prev_Athlete_Proportion) in the previous year and the total number of medals (Total_Medals) in that year. We visualize the results, as follows:

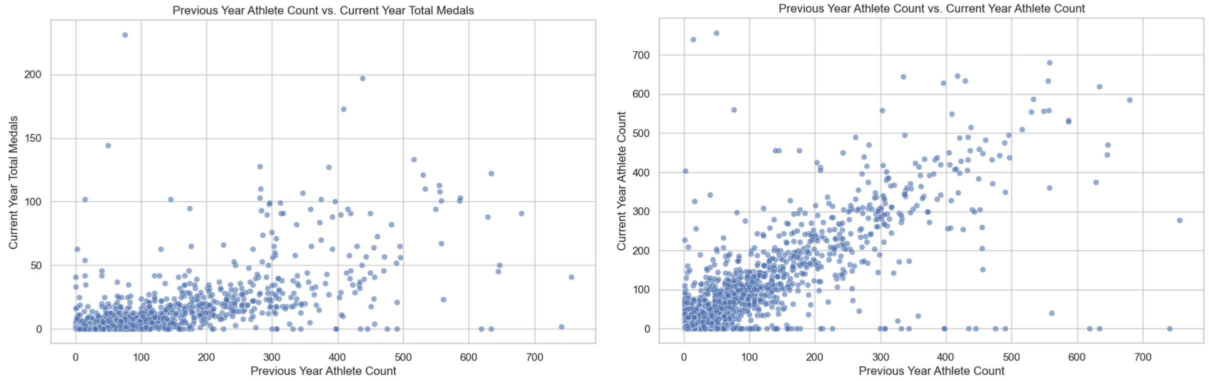


Figure 9 and 10: Previous Year Vs. Current Year Athlete Count

Similarly, we will test it through Pearson Correlation Matrix and Mutual information

Correlation Matrix for Shifted data

	Prev_Athlete_Count	Prev_Athlete_Proportiton
Total_Medals	0.696175	0.504940

Mutual Information for Shifted Data

Prev_Athlete_Count→Total_Medals	0.4411
Prev_Athlete_Proportion→Total_Medals	0.4579

Based on the graphs and correlation tests, we conclude that there is a correlation between the number of athletes and the proportion of athletes in the previous session and the total number of medals in this session.

4.2 Sequential Analysis

Then, we lag 'Athlete_Count', 'Athlete_Proportion', 'Medal_EWA', 'Avg_Score_Gold', 'Avg_Score_Silver', 'Avg_Score_Bronze', 'Gold', 'Silver', 'Bronze', focusing trends in the earning medal.

We plotted the values of the following three features: Gold count, Athlete_proportion, and EWA, calculated the correlation coefficient between the Lag data of each year and the data of the current year, and visualized them. The results are as follows:

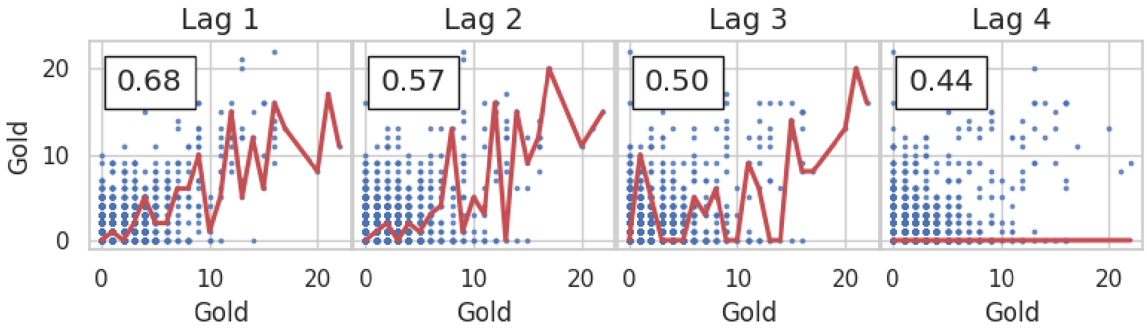


Fig 11: Lag of Gold for 4 lags

Features with higher correlation are typically more informative for models. In the above graph, lag 1 is the most significant feature

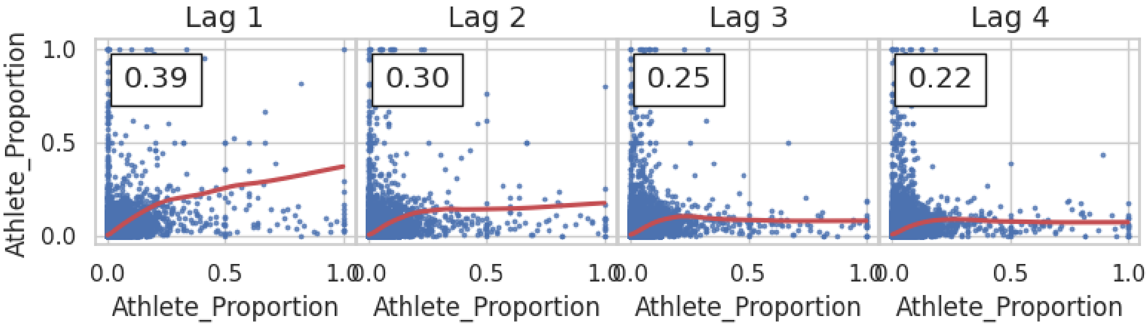


Figure 12: Lag of Athlete Proportion for 4 lags

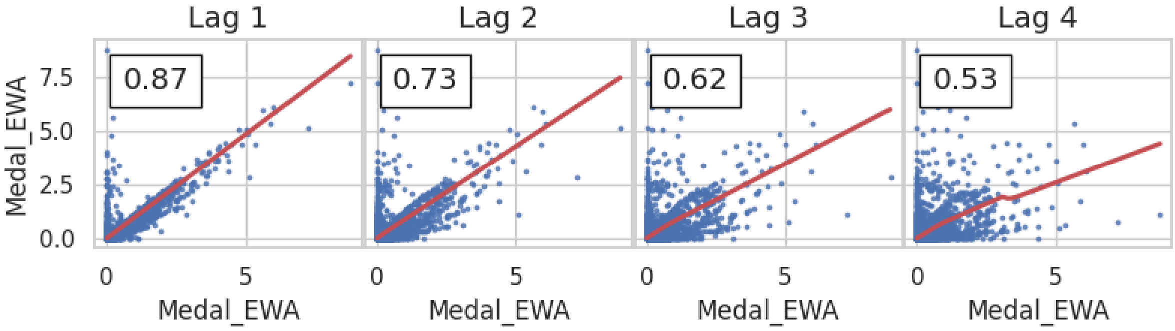


Figure 13: Lag of Medal EWA for 4 lags

From the figure, we can see that Gold and EWA have strong time dependence, that is, we can use lag to predict the data in 2028 with a high probability. Relatively speaking, the Athlete proportion has weaker time dependence, but it can also help us predict the data in 2028. Therefore, we believe that it is feasible to predict the data in 2028 through lag.

4.3 Partial Autocorrelation Function

The Partial Autocorrelation Function (PACF) helps us determine which lagged features are most useful for predicting medal counts by identifying the direct relationship between the target variable (e.g., Medal_EWA or Athlete_Proportion) and its lagged versions, while removing the influence of intermediate lags.

$$\phi_{k,k} = \rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}$$

where:

- ρ_k: The autocorrelation coefficient for training parameters at lag k.
- $\phi_{k,k}$: The partial autocorrelation for training parameters at lag k.
- $\phi_{k-1,j}$: The partial training parameters autocorrelation for prior lags.

The PACF isolates each lag's unique impact on the target by removing indirect effects from intermediate lags. For medal prediction (e.g., Medal_EWA), it identifies lags with significant predictive power by adjusting total correlations (ρ_k), ensuring only meaningful lags are included. The lag 2's PACF excludes lag 1's influence. Applying this to variables like Athlete_Proportion helps build interpretable models by focusing on significant lags and avoiding redundancy.

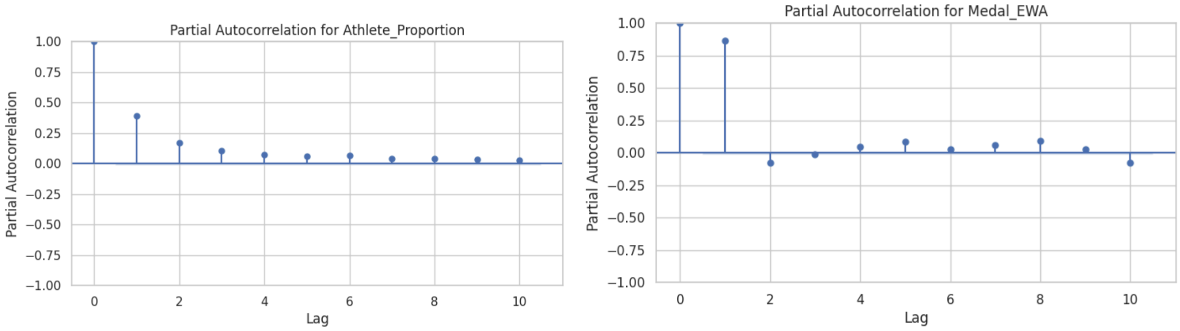
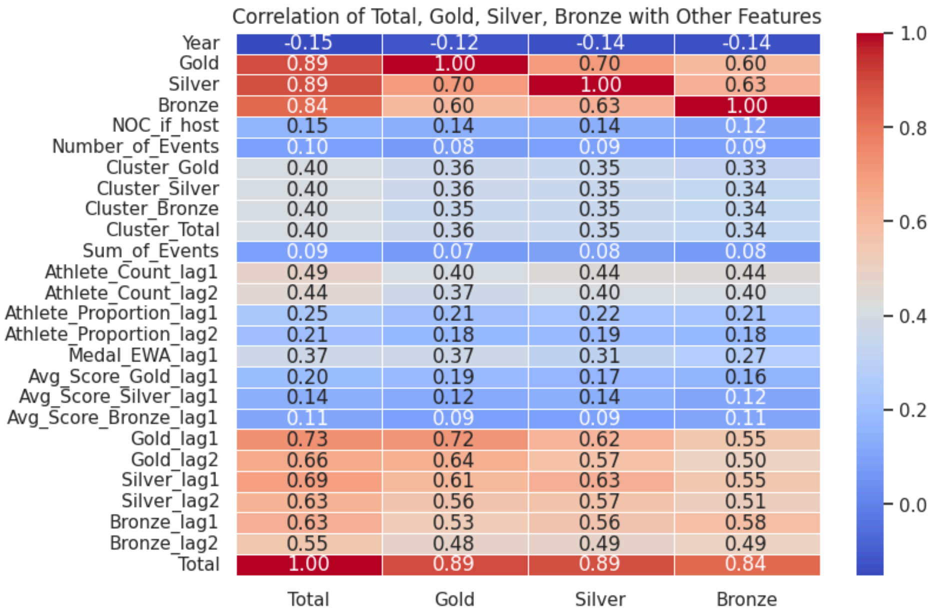


Figure 14,15: Partial Autocorrelation

By examining the lagged values, we determine which lags to include. If the value of a lag is minimal, it indicates that its information is already captured by other lags, and thus it can be excluded. Building on this foundation, we incorporated additional features to explore the correlation between medal counts and these variables.

4.4 Correlation heatmap



	Total	Gold	Silver	Bronze
Year	-0.15	-0.12	-0.14	-0.14
Gold	0.89	1.00	0.70	0.60
Silver	0.89	0.70	1.00	0.63
Bronze	0.84	0.60	0.63	1.00
NOC_if_host	0.15	0.14	0.14	0.12
Number_of_Events	0.10	0.08	0.09	0.09
Cluster_Gold	0.40	0.36	0.35	0.33
Cluster_Silver	0.40	0.36	0.35	0.34
Cluster_Bronze	0.40	0.35	0.35	0.34
Cluster_Total	0.40	0.36	0.35	0.34
Sum_of_Events	0.09	0.07	0.08	0.08
Athlete_Count_lag1	0.49	0.40	0.44	0.44
Athlete_Count_lag2	0.44	0.37	0.40	0.40
Athlete_Proportion_lag1	0.25	0.21	0.22	0.21
Athlete_Proportion_lag2	0.21	0.18	0.19	0.18
Medal_EWA_lag1	0.37	0.37	0.31	0.27
Avg_Score_Gold_lag1	0.20	0.19	0.17	0.16
Avg_Score_Silver_lag1	0.14	0.12	0.14	0.12
Avg_Score_Bronze_lag1	0.11	0.09	0.09	0.11
Gold_lag1	0.73	0.72	0.62	0.55
Gold_lag2	0.66	0.64	0.57	0.50
Silver_lag1	0.69	0.61	0.63	0.55
Silver_lag2	0.63	0.56	0.57	0.51
Bronze_lag1	0.63	0.53	0.56	0.58
Bronze_lag2	0.55	0.48	0.49	0.49
Total	1.00	0.89	0.89	0.84

Figure 16: Correlation of Total, Gold, Silver, Bronze with Other Features

5 Neural-Network Model

5.1 Data normalization

The features provided for the dataset include a combination of categorical, numerical, and lagged variables, each serving specific roles in the analysis. Starting with 'NOC_if_host', this binary variable indicates whether a National Olympic Committee (NOC) is hosting the event, as host nations typically exhibit improved performance. The 'Number_of_Events' captures the total number of events a country competes in, influencing medal counts significantly. The 'Cluster_Gold', 'Cluster_Silver', and 'Cluster_Bronze' represent clusters of gold, silver, and bronze medals won, respectively, while 'Cluster_Total' aggregates these totals, offering a consolidated performance view. Similarly, 'Sum_of_Events' provides the total participation across all events by a country.

Lagged variables offer historical context, which can be predictive of future performance. 'Athlete_Count_lag1' and 'Athlete_Count_lag2' reflect the number of athletes sent to the last two events, while 'Athlete_Proportion_lag1' and 'Athlete_Proportion_lag2' highlight the relative proportion of athletes in these previous events. The 'Medal_EWA_lag1' uses an exponentially weighted average to emphasize recent medal performances more heavily than distant ones. Additionally, 'Avg_Score_Gold_lag1', 'Avg_Score_Silver_lag1', and 'Avg_Score_Bronze_lag1' capture the average scores for each medal type from the previous event.

Finally, 'Gold_lag1', 'Gold_lag2', 'Silver_lag1', 'Silver_lag2', 'Bronze_lag1', and 'Bronze_lag2' represent the exact medal counts of each type from the last two competitions. These lagged medal counts and scores provide insights into past performance trends, helping to forecast future results. Together, these features enable a robust model capable of capturing host advantages, historical performance, and team participation dynamics.

We normalized the data using the standard scaler to compute the z-scores of the training data and Olympic medal counts (Gold, Silver, and Bronze). By transforming the data to a standard normal distribution, we ensured consistency across variables and reduced the risk of skewing the model due to disproportionate scales.

The model we used is a neural network, since all features have correlation with the targets, NN can capture the pattern.

5.2 Hyperparameter

We use random search for hyperparameter optimization, providing a balance between computational efficiency and coverage. Key parameters and hyperparameters included:

- different number of layers configured with ReLU activation functions to introduce non-linearity
- Batch normalization for standardized intermediate outputs
- Dropout rates and early stopping to prevent overfitting
- Different neuron values.
- Adam optimizer (Moment + RMSprop)

5.3 Model evaluation

The training results demonstrated strong performance, achieving balanced accuracy while avoiding overfitting. At epoch 20, the training loss was 0.2728, and the validation MAE becomes 0.2422, while the validation loss and MAE remained stable at 0.2723 and 0.2326, confirming that the model performed reasonably well in explaining the variance in medal outcomes without signs of overfitting. The final R2 values for validation data were 0.6416 for Gold, 0.5698 for Silver, and 0.4395 for Bronze.

To understand the performance of the model, we use the R2 value of directly using the last time value as prediction and the true value as a standard.

R² Results:

Metric	Gold	Silver	Bronze	Total Medals
Lag_1 Values as Prediction	0.4520	0.2885	0.1812	0.6036
Model Prediction for each Year, NOC, Sport_Code	0.6157	0.5180	0.4362	0.6795
Model Prediction for each Year, NOC	0.7327	0.7650	0.7504	0.7929

Comparing the R2 of the model with the standard, there is a clear gap between the values, which means the model has good performance. We also visualize the true medal count and predicted medal count for USA,

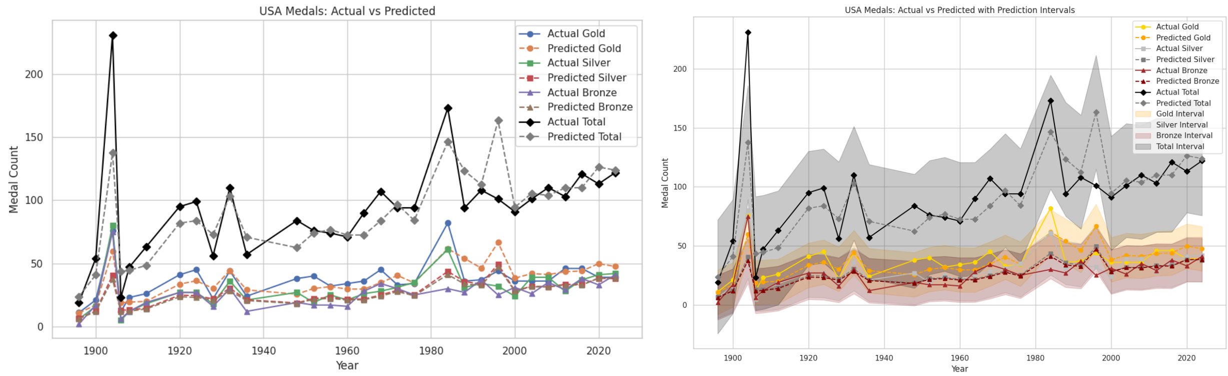


Figure 17: USA Medal prediction from 1986 to 2028 with confidence interval

As we can see, the prediction is generally good. The difference between true values and predicted values can be due to the limited data we have access to and a lot of uncertainties, like funding, athlete performance at a specific time, global events, etc. The confidence interval indicates the range within which we are 95% confident the true values for gold, silver, bronze, and total medal counts lie. Here, "z" represents the normalized z-score used to calculate this interval.

$$Confidence\ Interval = Mean\ Residual \pm z * SE$$

$$z = \frac{x_i - \bar{x}}{\sigma}$$

where:

- x_i: The medal count for a given year
- $\bar{x}$: The average medal count across all years
- σ: The standard deviation of the medal count distribution

6 Result

6.1 Performance change

Using the model to predict the Gold, Silver and Bronze medals for each country, each sport, and sum up the values in the same country but different sports.

We calculate the medal difference for gold medals between 2024 and 2028, then select top 2 countries with the most improvement or decline for Gold Medal and Total Medal, and we plot the graph to visualize. The result is

- Top 2 improvement in Gold: USA, ROC

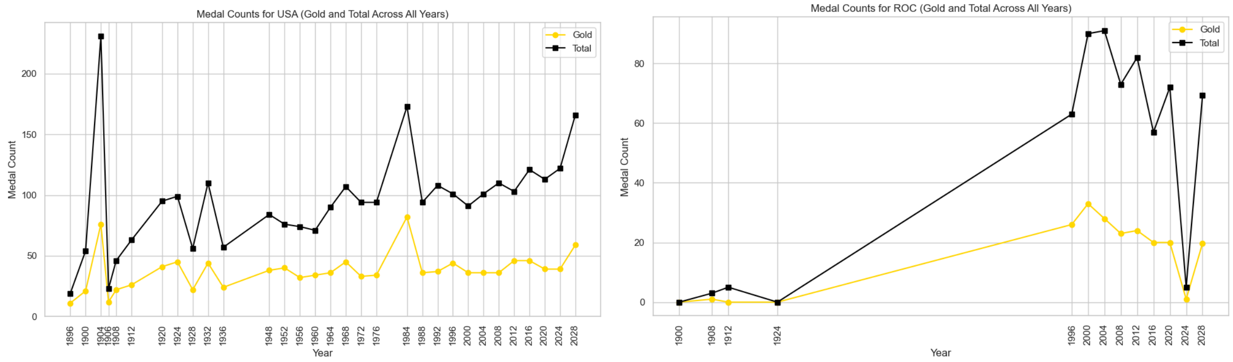


Figure 18,19: improvement in gold medal

- Top 2 decline in Gold: CHN, UZB

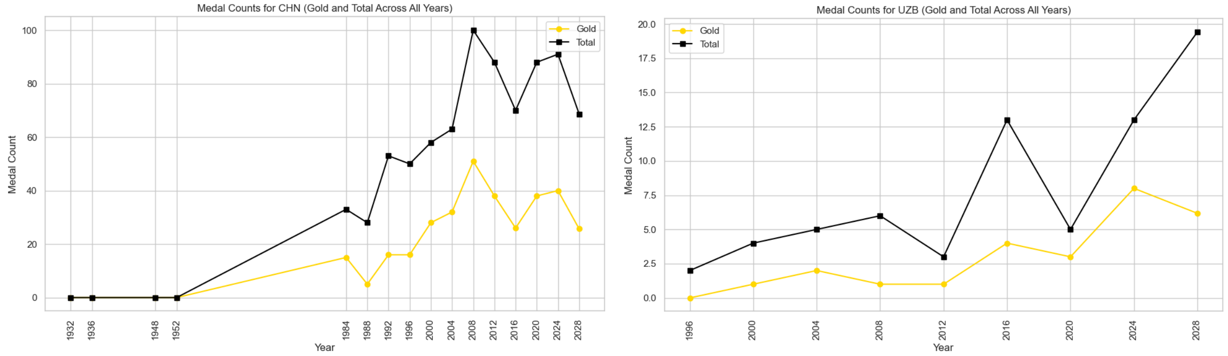


Figure 20, 21: decline in gold metal

- Top 2 improvement in Total: ROC, USA
- Top 2 decline in Total: CHN, FRA

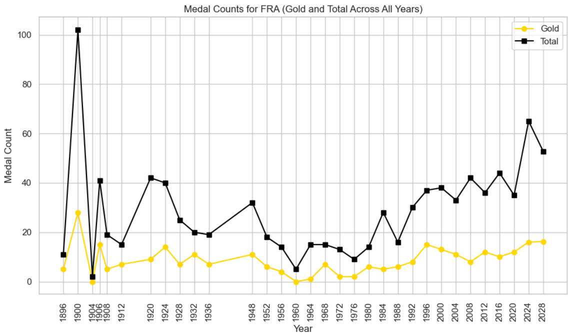


Figure 22: decline in total medal

The results are reasonable, for example, USA has a large improvement due to the host effect and overall trend, ROC has a large improvement due to being banned from the 2024 Olympics.

6.2 Predict First-Time Medal-Winning Countries

For data creation, we preprocessed the labeling countries based on whether they have earned medals previously and calculated the probabilities for countries participating in the 2028 Olympics based on historical data. The historical data is used to train where the target is whether a country earns a medal (1) or not (0). The model is trained using features extracted from past performance, with the data normalized and stratified to ensure balanced training.

We trained a 3 layer binary classification model with a sigmoid activation function. The accuracy for validation data is 0.88.

Once the model is trained, predictions are generated for the test data. For each country, the predicted probabilities for earning a medal in each sport they participate in are calculated. The probability of earning at least one medal is derived by aggregating the probabilities of not winning across all sports the country participates in, and then subtracting this aggregated probability from 1. This computes the complement, representing the likelihood of winning in at least one sport.

$$P = 1 - \prod_{i=1}^{n} Pr(E_i)$$

Where:

- P: probability of a not yet earned medal country will earn at least one medal during the current year
- $Pr(E_i)$: the probability of the country will not earn medal in i-th sport it participate it
- n: the total number of sport that the country is going to attend

The countries that have historically participated but never won a medal are filtered, and their aggregated probabilities are analyzed to determine their overall likelihood of earning at least one medal during the current year. The odds are further refined into percentages for better interpretability. This approach allows ranking countries based on their likelihood of breaking their no-medal streak, leveraging historical data and predictive modeling to provide actionable insights.

The result is none of them will earn their first medal in the next Olympics. The one with

NOC	Overall Probability	Odds in Percent
GAM	0.186227	22.884350
MAD	0.094842	10.477897
SAM	0.086376	9.454251

6.3 Great Coach Effect

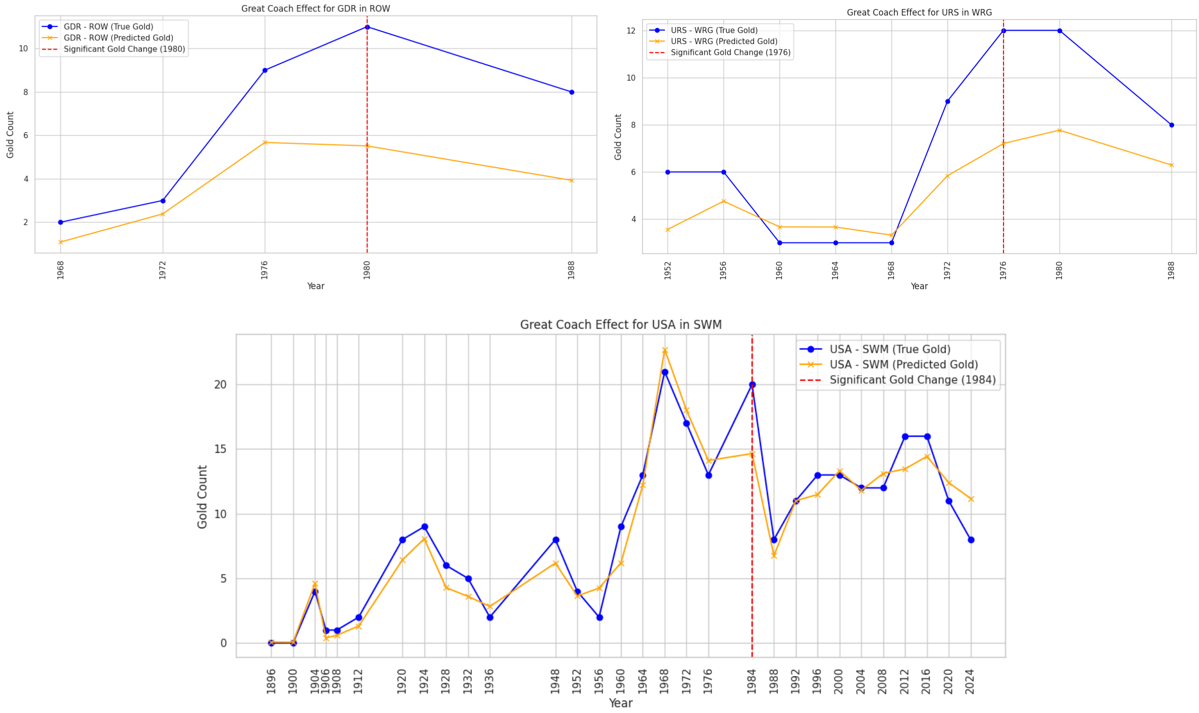


Figure 23,24,25: Great coach effect in rowing, wrestling, and swimming

We use a feature set without lagging as the training dataset, focusing on real-time analysis. To identify previous "great coaches," we evaluate their impact based on the number of gold medals won under their coaching. While it is difficult to definitively predict whether a specific success is directly due to a coach, we estimate their contribution by calculating the difference between actual results and predicted results. The maximum difference represents the data for evidence of changes potentially attributable to the "great coach" effect.

The results show the potential "great coach" effect in three cases. In the first graph, a significant spike in gold medals occurred around 1980, suggesting an external factor like hosting change led to improved performance in that sport. For swimming (USA - SWM), a notable increase in gold medals in 1968 aligns with the possibility of a transformative coach introducing innovative training methods. Similarly, in shooting (USA - SHO), a dramatic rise in gold medals in 1920 reflects the impact of a skilled coach or improved techniques, though the effect appears short-lived. These patterns highlight how influential coaching can dramatically enhance performance, especially in key Olympic years.

7 Model Assessment

7.1 Strength

- Our model effectively captures the strong correlations between features and medal counts. This ensures more accurate predictions.
- Since we have a strong feature correlation, neural networks will do a good job for us. Moreover, using a rectified linear unit activation function in neural networks allows the model to identify complex, non-linear patterns in the data.

7.2 Future improvement

- We can Incorporate more advanced features, such as multiplying the number of events by the mean athlete score, to better capture the relationships influencing medal counts.
- Explore models like XGBoost and LightGBM, which are highly effective for tabular data and can complement neural networks by offering improved interpretability and performance.
- Optimize model parameters and hyperparameters through advanced techniques like grid search, or Bayesian optimization to enhance model accuracy and robustness.

Work Cited

“2028 Summer Olympics.” *Wikipedia*, Wikimedia Foundation, https://en.m.wikipedia.org/wiki/2028_Summer_Olympics#Sports.