Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2514552

The 2024 Summer Olympics in Paris highlighted the significance of national medal standings, not only at the top of the table but also among countries earning their first medals. Albania, Cabo Verde, Dominica, and Saint Lucia achieved historic milestones with their inaugural Olympic medals. Meanwhile, over 60 countries remain without a medal, raising the question of how to predict medal counts for both dominant nations and those striving for their first success. Building on historical data from past Summer Olympics, this study addresses the challenge of projecting medal outcomes and evaluating the impact of strategic investments, such as hiring elite coaches.

Before developing specific models, an exploratory data analysis was conducted to identify key patterns in Olympic medal counts, participation rates, and other contributing factors. This analysis uncovered valuable insights into medal distributions and participation trends over time, which informed the feature selection and model design. Details of these findings can be found in the Appendix.

This study employs a multifaceted modeling approach to predict medal counts and analyze the "great coach" effect. A Cox Proportional Hazards Model was used to estimate the likelihood of countries winning their first medals. By leveraging survival analysis, the model effectively handles censored data, highlighting the importance of long-term participation and diversification of events. The model demonstrated strong predictive performance, validated through proportional hazards testing and Brier Score evaluation.

To forecast medal counts for countries at the 2028 Summer Olympics in Los Angeles, a LightGBM model was developed. By integrating features such as medal growth rates, athlete statistics, and host country advantages, the model achieved high accuracy, with R^2 values exceeding 0.8. Predictive intervals provided a measure of uncertainty, revealing countries poised for improvement and others likely to face setbacks. Additionally, the analysis explored how sports specialization influences national performance, offering strategic insights into event prioritization.

The Difference-in-Differences (DID) method was utilized to examine the causal effects of elite coaches. Case studies of Lang Ping (volleyball) and Béla Károlyi (gymnastics) demonstrated the potential of great coaches to enhance medal outcomes, as evidenced by statistically significant increases in the U.S. volleyball team’s performance under Lang Ping. DID analysis quantified these effects by comparing medal trends before and after coaching interventions across experimental and control groups, ensuring robustness through t-tests and confidence interval estimation.

This comprehensive approach addresses the complexity of Olympic medal prediction and provides actionable recommendations for national Olympic committees. By aligning data-driven strategies with historical trends and the unique impacts of coaching, this study offers a framework for optimizing medal success at future games.

Keywords: Cox Proportional Hazards, LightGBM, DID

content

1. Problem Interpretation

During the interpretation of the problem, a key issue was identified through analysis of the provided datasets: a conflict between the prediction target and the temporal scope of the datasets. The ultimate goal of the competition is to predict the medal counts (gold/total medals) or the possibility of a country earning its first medal at the 2028 Olympic Games. However, the provided datasets, excluding the host country information, only extend to 2024. This means that crucial information, such as the event schedule or participating athletes for 2028, is not directly available. Consequently, when constructing datasets or setting up prediction tasks, it is essential to strictly adhere to the principle of temporal causality. For instance, when predicting the medal outcomes for 2024, the features and data used must be derived solely from the Olympic Games held in 2020 or earlier. This approach aligns with real-world forecasting scenarios and effectively prevents data leakage. For example, when predicting outcomes for 2024, the model must not access any information related to the 2024 Games during the training process.

Additionally, the competition requires thorough data cleaning and preprocessing to ensure high-quality input data for the models. When analyzing the datasets, particular attention must be paid to aspects such as the temporal span of the data, its structure, and potential issues like feature redundancy or missing values.

Given the diversity of the problem, a hierarchical modeling approach was designed. One model is used to predict the total number of medals or gold medals, while another model is used to predict whether a country will earn its first medal. The rationale behind this layered design lies in the differing prediction goals and task natures. The former focuses on predicting medal counts, which is fundamentally a regression problem, while the latter predicts whether a previously medal-less country will win a medal at a given Olympic Games, which can be framed as a binary classification problem. Moreover, the feature engineering requirements for these two tasks differ significantly. For instance, the LightGBM model, used for medal prediction, leverages historical medal data and related features for countries with prior medal histories. On the other hand, the Cox proportional hazards model, used for predicting first-time medalists, focuses on countries with no prior medals, meaning it cannot directly utilize medal-related features. Instead, it emphasizes temporal and event-specific characteristics.

In terms of dataset construction, each data entry represents a (country, year) pair, which corresponds to a country's participation in a specific Olympic Games. However, the data selection criteria and temporal scope vary between models. For the Cox model, the dataset includes only the period from a country's first participation to the time it wins its first medal, disregarding subsequent medal records. For the LightGBM model, the dataset is limited to countries that have already won medals, as only their data is relevant for training and prediction. This hierarchical and targeted data selection approach not only meets the competition's requirements but also ensures the scientific validity and accuracy of the models.

2. Cox Proportional Hazards Model

2.1 Principles

The Cox proportional hazards regression model is a widely used method in survival analysis, modeling the hazard function $h(t, X)$, which represents the instantaneous risk of an event occurring at time t, given predictors *X*. The model is semi-parametric and assumes that the hazard can be expressed as the product of a baseline hazard $h_0(t)$, representing the risk when all covariates are zero, and a partial hazard function $exp\left(\sum_{i=1}^{n} \beta_i * x_i\right)$, where β_i quantifies the effect of each covariate x_i on the hazard [1]. This formulation relies on the proportional hazards assumption, which posits that the hazard ratio remains constant over time and depends linearly on the covariates.

The coefficients β_i are estimated using maximum likelihood estimation (MLE), while the baseline hazard $h_0(t)$ is left unspecified. To calculate $h_0(t)$, the Breslow estimator is commonly employed, which determines the hazard at each time point t_j based on the number of events d_j and the risk set R_j (individuals at risk at t_j).

The cumulative baseline hazard $H_0(t)$ is then obtained by summing risks over time, and the survival probability is calculated as $S_0(t) = exp\left(-H_0(t)\right)$. The model's final prediction formula, $P(t) = 1 - S_0(t)^{exp(\sum \beta_i * x_i)}$, provides the probability of an event occurring by time t, accounting for the influence of covariates.

This approach allows for robust survival analysis without making strict assumptions about the baseline hazard's form, making the Cox model both flexible and effective for analyzing time-to-event data in various fields.

2.2 Modeling Motivation and Feature Engineering

2.2.1 Modeling Motivation

The Cox proportional hazards model is a classical method for survival analysis, particularly suited for studying event occurrence times and their influencing factors. In this competition, the problem of "a country winning its first Olympic medal" can naturally be modeled as a survival analysis task: countries that win their first medal can be regarded as experiencing a "failure" event in survival analysis, while countries that have not yet won a medal correspond to a "surviving" state. Furthermore, time can be defined as the interval between a country's first participation in the Olympics and the year it wins its first medal. These problem characteristics align perfectly with the typical application scenarios of the Cox proportional hazards model.

Compared to traditional regression methods, the Cox model offers several advantages. First, it handles right-censored data flexibly; for example, some countries may not win a medal within the study period. Data from these countries can be naturally incorporated into the model without being discarded. Second, the Cox model allows the inclusion of multiple covariates (such as the number of participants, the number of events entered, and the history of hosting the Olympics) to quantify their impact on the "time to first medal" and generates interpretable hazard ratio estimates. Most importantly, the Cox model makes no specific assumptions about the baseline hazard function, making it well-suited for analyzing complex event times.

These characteristics make the Cox model a reasonable choice for addressing this problem. It not only provides a comprehensive interpretation of the process by which countries win their first medal but also offers valuable data-driven insights to support policy-making.

2.2.2 Feature Engineering

Table 2: Features of Cox Model

Features	Meaning
NOC	Country code
Year	Year of participation (current Olympics)
duration	Time since the first participation
event	Whether a medal was won (1 for won, 0 for not won)
total_appearances	Total number of Olympic Games participated in historically
total_athletes	Total number of athletes historically
unique_events	Number of unique events participated in historically
unique_sports	Number of unique sports participated in historically
athletes_per_games	Average number of athletes per Olympic Games historically
events_per_games	Average number of events participated in per Olympic Games historically
last_HHI	Concentration of events (HHI) in the previous Olympics
last_athletes_count	Number of athletes in the previous Olympics

last_unique_events	Number of events in the previous Olympics
last_unique_sports	Number of sports in the previous Olympics
last_female_ratio	Gender ratio in the previous Olympics
last_veteran_ratio	Ratio of veteran athletes in the previous Olympics
athlete_growth_rate	Average annual growth rate of participating athletes for the country
event_growth_rate	Average annual growth rate of events participated in for the country
historical_events_per_athlete	Average number of events participated in per athlete historically
athlete_per_event	Average number of athletes per event in the previous Olympics

The dataset construction in this project aims to explore the progression of countries from having no medals to winning their first Olympic medal and uncover the underlying patterns in this process. Specifically, each data entry is uniquely identified by the combination of a country's National Olympic Committee (NOC) code and the year (Year), reflecting the performance of a country during the period from its first participation to its first medal win. The dataset construction process is as follows:

First, countries that have never won a medal up to their first medal win were selected as the subjects of study. Since these countries remain in a no-medal state during the study period, the feature construction focuses primarily on participation-related information, such as athletes and events, rather than medal results. This design ensures that the dataset includes features that are predictive of winning a first medal, such as the number of athletes, the number of events participated in, gender ratios, and veteran ratios among athletes.

Second, the calculation of features strictly adheres to the principle of no data leakage. For each data entry (NOC, Year), all features are computed based solely on the country's historical data prior to the given year (Year), ensuring that no future information is used. For instance, if the current year is Year, only data from the country's participation in Olympics prior to Year (e.g., the number of athletes and event distribution in the previous Olympics) is used to calculate the relevant features. This principle effectively avoids data leakage and ensures consistency and reliability in model training and testing.

Furthermore, the dataset construction not only considers the performance in each Olympic Games but also incorporates trend-related features, such as the growth rates of athlete numbers and event numbers, to capture the dynamic changes in a country's participation strength. These trend features help reveal the underlying changes during the progression from having no medals to winning medals. Additionally, to reflect both short-term and long-term performance, the dataset includes cumulative historical features as well as features from the previous Olympics, such as the number of athletes, veteran ratios, and event concentration (HHI) in the previous Games.

2.3 Experimental Design and Model Performance

2.3.1 Experimental Design

Table3: Training Set Event Distribution (n=468)

Event	Count
0	349
1	119

Table4: Testing Set Event Distribution (n=121)

Event	Count
0	91
1	30

To ensure the generalization ability of the model, the dataset was split into training and testing sets using a country-level random partitioning method. Specifically, the data for each country includes features from the period between its first Olympic participation and its first medal win. In both the training and testing sets, the ratio of event status 1 (indicating the first medal win) to event status 0 (indicating no medal yet) is approximately 1:3. This partitioning approach ensures dataset balance, thereby avoiding prediction bias caused by class imbalance.

For this study, the Cox Proportional Hazards model from the lifelines library (CoxPHFitter) was used for training, and the model performance was evaluated using the Brier Score. After training the model, the proportional hazards assumption was tested using the cph.check_assumptions function. This function employs statistical tests (such as the Schoenfeld residual test) and visualizations of lowess linear regression results to verify the validity of the proportional hazards assumption.

2.3.2 Feature Importance Analysis

To intuitively demonstrate the impact of each feature on the risk of an event occurring, we plotted a feature importance chart. The horizontal axis represents the log hazard ratio (log(HR), i.e., log(Hazard Ratio)), which quantifies the magnitude of a feature's effect on the event risk. Each point on the chart represents the estimated value of a feature, while the horizontal line around the point indicates the 95% confidence interval (CI) for that estimate. If the confidence interval crosses 0 (the midpoint of the horizontal axis), it suggests that the feature may have no significant impact on the risk. Conversely, if the confidence interval lies entirely on one side of 0, it indicates a significant effect, with points to the right of 0 denoting an increased risk (positive correlation) and points to the left denoting a reduced risk (negative correlation).

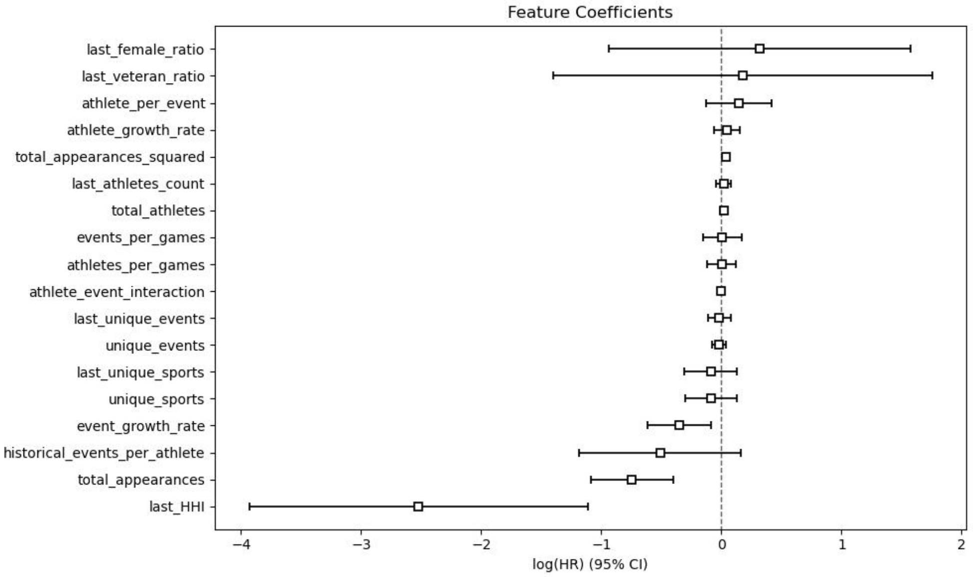


Figure1: Feature coefficients

As shown in the Figure1, two variables have the most significant impact: last_HHI and total_appearances. Their effects are explained as follows:

- **last_HHI**: The log(HR) for this feature is significantly below 0, with its confidence interval entirely in the negative range. This indicates that higher event concentration (higher HHI values) reduces the risk of a country winning its first medal. In other words, countries tend to increase their chances of winning their first medal by diversifying their event participation.
- **total_appearances**: The log(HR) for this feature is also significantly below 0, and its confidence interval does not cross 0. This suggests that the more frequently a country participates in the Olympics, the lower the risk of not winning its first medal. This aligns with intuition, as countries with long-term Olympic participation are more likely to secure their first medal.

2.3.3 Hypothesis Testing

To validate the model's underlying assumption—the proportional hazards assumption—we used the check_assumptions method from the lifelines library, setting the significance level at 0.01. The results indicated that most covariates satisfied the proportional hazards assumption, with only two variables (**total_appearances** and **total_appearances_squared**) failing the test. Such occurrences are not uncommon in real-world data analysis, as almost all real-world datasets violate the proportional hazards assumption to some extent.

In this study, the primary objective of the model is to enhance prediction performance rather than to achieve entirely precise interpretation of the hazard ratios for covariates [2]. Therefore, we believe that the minor violations of the proportional hazards assumption by a small number of variables have a negligible impact on the model's predictive performance. Additionally, the fact that most variables passed the assumption test demonstrates the model's overall robustness and applicability.

It is worth noting that, according to the research by Stensrud and Hernán, strictly meeting the proportional hazards assumption is not always necessary [3]. They point out that forcefully adjusting a model to satisfy the assumption under non-proportional hazards may alter the scientific interpretation of the research question, thereby affecting the interpretability of the results. This further highlights that prioritizing the model's predictive ability for the target event is more meaningful than overly adhering to theoretical assumptions.

In conclusion, while some variables violate the proportional hazards assumption, this reflects the complexity of real-world data and does not significantly affect the model's predictive performance. In this context, the model remains capable of providing reliable support for predicting first-time Olympic medal wins, demonstrating strong practical value.

2.3.4 Model Performance

We used the Brier Score as the evaluation metric, with the formula as follows:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}_i - y_i)^2$$

Where $\hat{y}_i$ is the predicted probability of the event occurring at instance *t*, y_i is the actual outcome at instance *t*, where 1 represents the event occurring, and 0 represents it not occurring.

Table 5: Brier Score

Train Brier Score	Test Brier Score
0.186	0.219

From the results in the table, it can be observed that the Brier Score is very low on both the training set and the testing set, indicating that the model achieves a good fit.

2.4 Prediction Results

Table 6: Prediction Results

NOC	Duration	Years_Waiting	First_Year	Probability
MON	27	108	1920	0.929626
MYA	20	80	1948	0.904626
MLT	25	100	1928	0.813601
MLI	16	64	1964	0.747054

LIE	23	92	1936	0.735107

To perform the prediction, we constructed a prediction dataset consisting of countries that participated in the 2024 Olympics but have never won a medal in their history. The results indicate that Morocco has a high probability of winning its first Olympic medal at the 2028 Games.

3. LightGBM Model

3.1 Principles

Gradient Boosting Decision Trees (GBDT) is a powerful machine learning algorithm, but it faces scalability challenges with high-dimensional and large-scale data [4]. LightGBM addresses these challenges by introducing a histogram-based split-point search and optimizing instance and feature handling for improved scalability.

One key innovation in LightGBM is Gradient-based One-Side Sampling (GOSS), which prioritizes instances with large gradients and randomly samples from those with smaller gradients. To maintain data distribution, GOSS scales the information gain calculation by adjusting the weight of small-gradient instances using a factor of (1−a)/b, where aaa and bbb are sampling proportions. This approach focuses on under-trained instances without distorting the original data distribution.

LightGBM also reduces feature dimensionality using Exclusive Feature Bundling (EFB). EFB bundles mutually exclusive features by treating them as graph vertices and connecting them with edges if they are not mutually exclusive. The bundling problem is simplified as a graph coloring task, solved using a greedy algorithm to minimize conflicts. During feature bundling, LightGBM ensures that bundled features remain distinguishable by assigning offsets to their values. This process, executed once before training, creates dense features while preserving the integrity of the original feature values. The EFB method significantly reduces redundant computations for zero-value features, enhancing efficiency.

3.2 Modeling Motivation and Feature Engineering

3.2.1 Modeling Motivation

Gradient Boosting Decision Trees (GBDT) is a powerful machine learning algorithm, but it faces scalability challenges with high-dimensional and large-scale data. LightGBM addresses these challenges by introducing a histogram-based split-point search and optimizing instance and feature handling for improved scalability.

One key innovation in LightGBM is Gradient-based One-Side Sampling (GOSS), which prioritizes instances with large gradients and randomly samples from those with smaller gradients. To maintain data distribution, GOSS scales the information gain calculation by adjusting the weight of small-gradient instances using a factor of (1−a)/b, where aaa and bbb are sampling proportions. This approach focuses on under-trained instances without distorting the original data distribution.

LightGBM also reduces feature dimensionality using Exclusive Feature Bundling (EFB). EFB bundles mutually exclusive features by treating them as graph vertices and connecting them with edges if they are not mutually exclusive. The bundling problem is simplified as a graph coloring task, solved using a greedy algorithm to minimize conflicts. During feature bundling, LightGBM ensures that bundled features remain distinguishable by assigning offsets to their values. This process, executed once before training, creates dense features while preserving the integrity of the original feature values. The EFB method significantly reduces redundant computations for zero-value features, enhancing efficiency.

3.2.2 Feature Engineering

Table 7: Features of LightGBM model

Features	Meaning
NOC_Code	Country code (abbreviation)
Year	Year of the current Olympics to be predicted
gold_predict	Predicted number of gold medals for the current Olympics
metal_predict	Predicted total number of medals for the current Olympics
gold_3_games_avg	Average number of gold medals won by the country in the past three Olympics
total_3_games_avg	Average total number of medals won by the country in the past three Olympics
gold_std	Standard deviation of gold medals won by the country in the past three Olympics
total_std	Standard deviation of total medals won by the country in the past three Olympics
gold_growth_noc_rate	Growth rate of gold medals won by the country in the previous Olympics (compared to the one before)
metal_growth_noc_rate	Growth rate of total medals won by the country in the previous Olympics (compared to the one before)

gold_growth_game_rate	Average growth rate of gold medals per Olympics (calculated as the average of growth rates between consecutive Olympics)
metal_growth_game_rate	Average growth rate of total medals per Olympics (calculated as the average of growth rates between consecutive Olympics)
gold_num	Number of gold medals won by the country in the previous Olympics
metal_num	Total number of medals won by the country in the previous Olympics
best_rank_last_3	Best rank achieved by the country in the past three Olympics
rank_trend	Average rank of the country in the past three Olympics
participation_times	Total number of Olympic Games participated in by the country
best_gold	Highest number of gold medals won by the country in a single Olympics
best_total	Highest number of total medals won by the country in a single Olympics
athletes_count	Total number of athletes representing the country throughout Olympic history
veteran_ratio	Proportion of veteran athletes in the previous Olympics (veterans are defined as athletes who have participated in at least one prior Olympics)
star_athlete_num	Total number of star athletes in the country’s history (stars are defined as athletes who have won more than three medals)
athletes_growth_rate	Average growth rate of athletes per Olympics for the country (calculated as the average of growth rates between consecutive Olympics)
medals_per_athlete	Average number of medals won per athlete in the country’s history (total medals divided by total athletes)
gold_per_athlete	Average number of gold medals won per athlete in the country’s history (total gold medals divided by total athletes)

unique_sports_last	Number of unique sports participated in by the country in the previous Olympics
unique_events_last	Number of unique events participated in by the country in the previous Olympics
event_growth	Average growth rate of events participated in per Olympics (calculated as the average of growth rates between consecutive Olympics)
sport_growth	Average growth rate of sports participated in per Olympics (calculated as the average of growth rates between consecutive Olympics)
hhi_sport	Herfindahl-Hirschman Index (HHI) for the concentration of sports participated in by the country historically
dominant_sport	Number of dominant sports for the country historically (a sport is defined as dominant if the medals won in that sport exceed 50% of the total medals won in that sport by the country)
gold_event_rate	Efficiency of winning gold medals in events historically (number of events where gold was won divided by total events participated in by the country)
metal_event_rate	Efficiency of winning any medal in events historically (number of events where any medal was won divided by total events participated in by the country)
gold_event_avg	Average number of gold medals won per event historically (total gold medals divided by total events participated in by the country)
metal_event_avg	Average number of total medals won per event historically (total medals divided by total events participated in by the country)
consistent_sports	Number of sports consistently participated in by the country for at least three consecutive Olympics
years_since_last_medal	Number of years since the country last won a medal
is_host_history	Number of times the country has hosted the Olympics
ishost_hostrate_gold	Whether the country is hosting the current Olympics (1 if hosting, 0 otherwise) multiplied by the average historical growth rate of gold medals for host countries (calculated as the average increase in gold

	medals for host countries compared to the previous Olympics)
ishost_hostrate_metal	Whether the country is hosting the current Olympics (1 if hosting, 0 otherwise) multiplied by the average historical growth rate of total medals for host countries (calculated as the average increase in total medals for host countries compared to the previous Olympics)
ishost	Whether the country is hosting the current Olympics (1 if hosting, 0 otherwise)

To analyze the trends in gold and total medals for countries that have previously won medals at the Olympics and to explore the patterns of medal count changes, this study constructed a dataset organized by country and year. Each data entry is uniquely identified by the combination of the National Olympic Committee (NOC) code and the Olympic Year, where NOC represents the country’s committee code, and Year denotes the Olympic year. The dataset includes countries that have won medals in the past, ensuring the ability to analyze their medal performance trends.

In terms of feature selection, the dataset incorporates multidimensional information, including historical medal performance, athlete characteristics, event participation, and the host country effect. For example, features include the average and standard deviation of gold and total medals over the last three Olympics, growth rates in medal counts, the number and growth rate of athletes, the number of star athletes, veteran ratios, the number and growth rate of events participated in, the concentration of gold and medal distributions (measured by the HHI index), and host country status along with its historical performance. Additionally, statistics related to participation rankings and historical best achievements were extracted to comprehensively capture key factors influencing a country’s medal performance.

To prevent information leakage from future data, strict temporal causality was maintained during the data construction process. Specifically, for any data entry identified by (NOC, Year), only historical data from years prior to the target year were used to compute the features. For instance, gold or medal counts from the target year were not used in the construction of any features. This design ensures the validity of the data and the authenticity of model training, thereby providing a robust foundation for the accuracy of subsequent predictive models.

3.3 Experimental Design and Model Performance

3.3.1 Experimental Design

Table 8: Parameters Of Model

name	value
objective	regression
metric	rmse
boosting_type	gbdt
num_leaves	31
learning_rate	0.05
feature_fraction	0.9
bagging_fraction	0.8
bagging_freq	5

In this study, we used the LightGBM model to predict the number of gold medals and total medals for each country. In the experimental design, the dataset was split into training and testing sets based on the year. All data prior to 2020 were used for training (training set size: 1153), while data from 2020 and later were used for testing (testing set size: 170). This division ensures temporal independence between the training and testing data, better simulating real-world prediction scenarios.

For feature importance analysis, we used the LightGBM library to calculate feature importance scores based on split importance. This metric counts the number of times each feature is used as a splitting node across all decision trees in the model. The more frequently a feature is used, the greater its contribution to model splits, and thus, the higher its importance score. Split importance is an intuitive and efficient measure for feature importance, providing a quick reflection of the role each feature plays in the model.

During the prediction process, to comprehensively evaluate the uncertainty of the model, we introduced the Bootstrap method to generate prediction intervals. Specifically, this method involves repeatedly selecting random subsets of the test set features (e.g., randomly masking approximately 20% of the features) and running multiple rounds of predictions. A distribution of prediction results is then obtained, from which the mean value is taken as the expected prediction, and the 2.5th and 97.5th percentiles are used as the lower and upper bounds of the prediction interval, respectively, forming a 95% confidence interval. These prediction intervals provide a quantitative description of the uncertainty associated with the predictions, making the results more reliable and interpretable.

3.3.2 Feature Importance

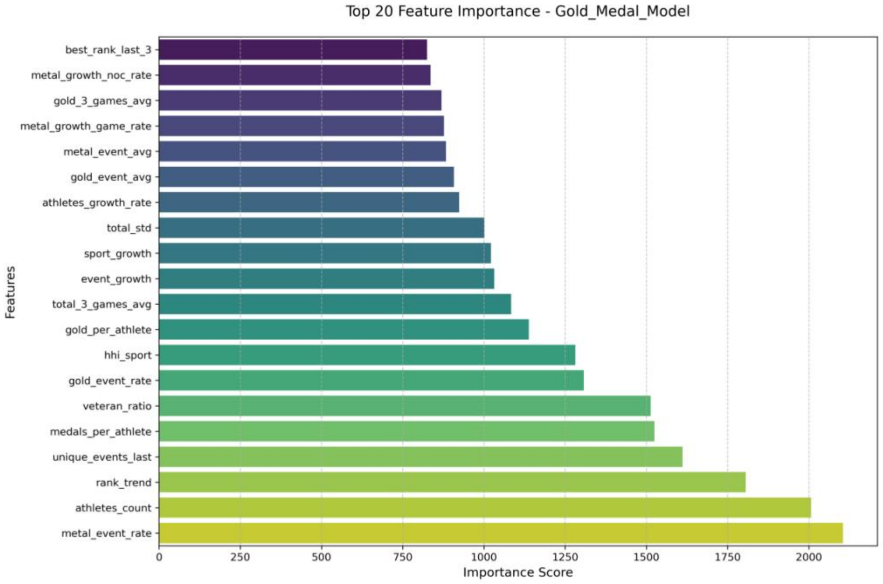


Figure 2: Top 20 features importance of Gold Medal Modal

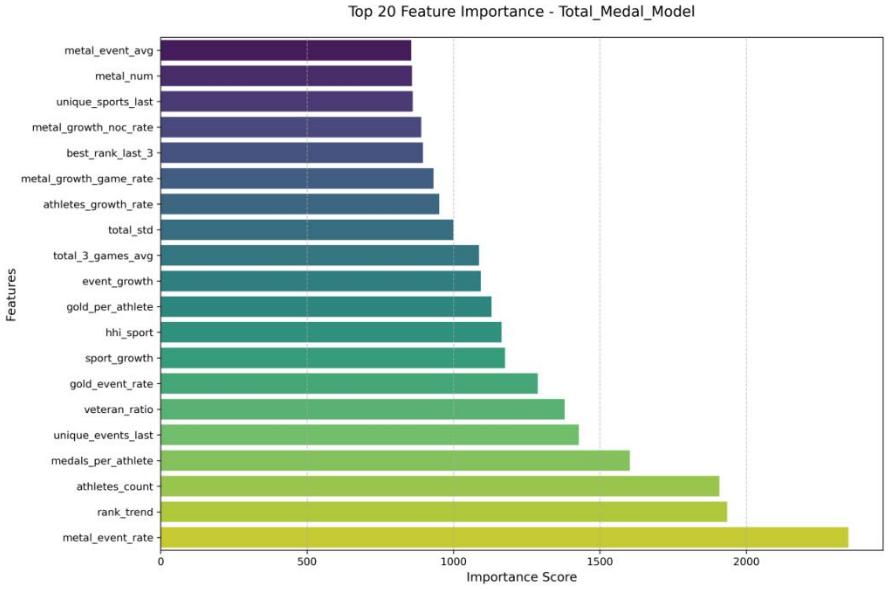


Figure 3: Top 20 features importance of Total Medal Modal

For gold medal prediction, the five most influential features are metal_event_rate, athletes_count, rank_trend, unique_events_last, and medals_per_athlete.

For total medal prediction, the five most influential features are metal_event_rate, rank_trend, athletes_count, medals_per_athlete, and unique_events_last.

3.3.3 Model Performance

We used three metrics to evaluate the performance of the model: **RMSE**, **MAE**, and **R²**, with the formulas as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i|$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2}$$

where y_i is the actual number of medals, $\hat{y}_i$ represents the predicted number of medals and $\bar{y}_i$ is the mean of the actual medals.

Table 9: Model Performance

	RMSE	MAE	R2
Gold Medal	3.10	1.77	0.81
Total Medal	7.50	3.85	0.86

The table presents the experimental results. For gold medal prediction, the model achieved an RMSE of 3.10, an MAE of 1.77, and an R²of 0.814. This indicates that the model can accurately predict the number of gold medals won by countries at the Olympics with a high degree of precision.

For total medal prediction, the model achieved an RMSE of 7.50, an MAE of 3.85, and an R² of 0.858, further demonstrating the model's strong predictive capability when handling a more complex target variable.

Overall, the R² values for both models exceed 0.8, indicating that the models can effectively explain the variance in the target variables and possess strong predictive power and practical utility.

3.4 Prediction Results

For predicting the number of gold medals and total medals at the 2028 Olympics, we assumed that all countries that won medals in 2024 would participate in the next Olympics. The top 10 predictions are as follows:

Table 10: Prediction Results

NOC	Predicted Gold	Gold Lower	Gold Upper	Predicted Total	Total Lower	Total Upper	Is Host
CHN	33	20.6	41.7	66	35.3	88.8	0
USA	29	11.3	39.8	79	26.6	104.5	1
GBR	23	13.5	29.7	53	28.0	71.9	0

JPN	19	7.3	26.4	49	19.6	66.4	0
AUS	13	6.1	19.0	37	17.1	50.4	0
FRA	10	3.3	14.2	30	10.4	43.1	0
GER	8	3.0	11.0	25	7.4	36.7	0
KOR	7	2.0	9.7	20	3.2	33.2	0
ITA	7	2.8	10.7	23	7.5	36.2	0
NED	6	2.0	8.0	16	2.8	28.7	0

4.Difference-in-Differences Method (DID)

4.1 Principles

The Difference-in-Differences (DID) method is commonly used to evaluate the effects of randomized experiments (e.g., changes in laws or regulations), where the effects of such interventions often take time to manifest [5]. Therefore, data from both before and after the intervention over several years is typically required. The implementation of the DID method requires two samples: a treatment group and a control group. Additionally, the treatment and control groups must exhibit a common trend prior to the intervention, and the estimated policy effect must account for the differences between the treatment and control groups before the policy was implemented.

An intuitive form of the DID model is as follows:

$$y_{it} = \alpha + \gamma D_t + \beta Treat_i + \delta D_t \times Treat_i + \varepsilon_{it} (i = 1, \ldots, n; t = 1,2)$$

Where D_t is a dummy variable for the experimental period, taking a value of 0 or 1 (0 represents the pre-experiment period, i.e., t=1; 1 represents the post-experiment period, i.e., t=2). Similarly, the policy dummy variable $\mathrm{Treat_i}$ takes a value of 0 for the control group and 1 for the treatment group. The coefficient δ of the interaction term represents the policy effect.

4.2 Experimental Design

4.2.1 Discovering the "Great Coach" Effect

Taking the United States as an example, it is known that coach Lang Ping began coaching the U.S. volleyball team in 2005, while Béla Károlyi started coaching around 1980 [6][7]. To examine the impact of coaches on team performance, we selected the U.S. team (USA) in specific sports (e.g., volleyball or gymnastics) and analyzed their medal performance by observing the total medal count changes for each Olympic Games.

To better capture the trends in medal changes, we introduced two features: Absolute Change and Relative Change. The former represents the actual increase or decrease in the total medal count, while the latter represents the relative percentage change in medal count. The Absolute Change and Relative Change are defined as follows:

- Absolute Change = Current year's medal count − Previous year's medal count
- Relative Change = $\frac{\text{Absolute Change}}{\text{Previous year's medal count}+1}$

For each type of change (Absolute_Change and Relative_Change), the first quartile (Q1) and the third quartile (Q3) are calculated. The interquartile range (IQR) is defined as: $\text{IQR} = \text{Q3} - \text{Q1}$. The upper and lower bounds for outliers are defined as follows:

$$\text{Lower bound} = \text{Q1} - 1.5 \times \text{IQR}, \text{Upper bound} = \text{Q3} + 1.5 \times \text{IQR}$$

Table 11: OutLiters in Volleyball for USA

Year	PreYear	Medals	PreMedals	Absolute_Change	Relative_Change
1988	1984	12	24	-12	-0.48
1992	1988	24	12	12	0.92
1996	1992	0	24	-24	0.96
2008	2004	24	0	24	24.00
2012	2008	12	24	-12	-0.48
2016	2012	24	12	12	0.92
2020	2016	12	24	-12	-0.48
2024	2020	16	12	14	1.08

Table 12: OutLiters in Gymnastics for USA

Year	PreYear	Medals	PreMedals	Absolute_Change	Relative_Change
1936	1932	0	20	-20	-0.95
1948	1936	8	0	8	8.00
1952	1948	0	8	-8	-0.89
1984	1976	26	1	25	12.5

1988	1984	1	26	-25	-0.92
1992	1988	11	1	10	5.00
2000	1996	6	11	-5	-0.42
2004	2000	19	6	13	1.86

Next, we identified and filtered out outliers, as shown in the table. It is clear that the U.S. volleyball team had a Relative Change of 24 in 2008, indicating a dramatic increase in the number of medals won that year. Similarly, the U.S. gymnastics team had a Relative Change of 12.5 in 1984, also reflecting a significant surge in medal counts. Notably, these two Olympic Games coincided with the first Games where the respective coaches began their tenure. This provides strong evidence for the existence of the "great coach" effect.

4.2.2 Experimental Design

In this study, we designed experiments to evaluate the coaching effect of "great coaches," selecting volleyball and gymnastics as the research subjects.

For volleyball, the experimental years range from 2000 to 2012, specifically covering two Olympic Games before Lang Ping began coaching the U.S. national team (2000 and 2004) and two after (2008 and 2012). The treatment group consists of the U.S. team (coached by Lang Ping), while the control group includes the Italian and Chinese national teams (not coached by Lang Ping).

For gymnastics, the experimental years range from 1976 to 1988, covering two Olympic Games before Béla Károlyi began coaching the U.S. national team (1976 and 1980) and two after (1984 and 1988). The treatment group consists of the U.S. team, while the control group includes the Romanian and Japanese national teams.

We used the medal count as the dependent variable and the treatment group indicator (treated), post-intervention indicator (post), and their interaction term (did) as independent variables to construct a Difference-in-Differences (DID) model to assess the intervention effect. The treated variable indicates whether a team belongs to the treatment group, post indicates whether the experimental year falls within the post-intervention period, and the interaction term did = treated × post captures the potential causal impact of the intervention on the treatment group.

To estimate the intervention effect, we fitted the data using the OLS (Ordinary Least Squares) model from the statsmodels library, obtaining the regression coefficient for did along with its standard error. To test the significance of the intervention effect, we conducted a t-test on the regression coefficient of did, calculating the t-statistic and two-tailed p-value to determine whether the intervention had a statistically significant effect on the medal count. Additionally, we calculated the 95% confidence interval to further quantify the credible range of the intervention effect. If the confidence interval does not include 0 and the p-value is below a specified significance level (e.g., 0.05), the intervention effect can be considered significant.

4.3 DID Regression Results and Effect Interpretation

Table 13: Volleyball Analysis

Metric	Result
Raw DID Estimate	21 medals
95% Confidence Interval	[4.59, 37.41]
P-value	0.0032

Table 14: Gymnastics Analysis

Metric	Result
Raw DID Estimate	11 medals
95% Confidence Interval	[-15.64,37.64]
P-value	0.3411

The analysis of the effect of Lang Ping on the USA volleyball team reveals a positive and statistically significant effect. The raw DID estimate suggests that Lang Ping's coaching increased the medal count by 21 medals, with a 95% confidence interval of [4.59, 37.41]. The P-value of 0.0032 confirms the significance of this result. This implies that Lang Ping likely had a substantial impact on improving the USA volleyball team's performance, as evidenced by the dramatic increase in medal counts from 0 in the pre-treatment period to 18 in the post-treatment period.

This result is likely due to Lang Ping's exceptional coaching skills and ability to implement strategies that directly impacted team performance. Furthermore, the control groups (China and Italy) show relatively stable medal counts across the same time period, strengthening the argument that the observed improvement for the USA team is attributable to Lang Ping's coaching.

The analysis of Béla Károlyi's effect on the USA gymnastics team provides a raw DID estimate of 11 medals, with a 95% confidence interval of [-15.64, 37.64]. However, the P-value of 0.3411 indicates that the result is not statistically significant. This means we cannot conclusively attribute the change in medal count to Károlyi's coaching.

One possible reason for this result could be the large variability in medal counts for the USA team during this period, coupled with relatively consistent performances by the control groups (Romania and Japan). The USA's dramatic increase in 1984 (26 medals) may be influenced by other external factors [8], such as changes in competition structure, team composition, or geopolitical events (e.g., boycotts in the 1984 Olympics), rather than coaching alone.

4.4 Strategic Recommendations for Hiring Coaches and Their Impacts

India has demonstrated a solid foundation in shooting, highlighted by its gold medal at the 2020 Tokyo Olympics. While individual athletes have performed exceptionally, the team's overall consistency remains an issue. Shooting, being a relatively niche sport, presents an opportunity for focused investment, as the competition is less saturated compared to other high-medal sports.Given the individualistic nature of shooting, the impact might be more conservative—potentially 2–4 additional medals per Olympics—through improved mental resilience, tactical planning, and precision training.

South Africa has a rich history in swimming, exemplified by athletes like Chad le Clos [9]. However, recent years have seen a decline in performance due to the absence of a systematic training infrastructure and top-tier coaching. Swimming, as a high-medal sport, presents an excellent opportunity for South Africa, given its talented pool of athletes, particularly in short-distance freestyle events.Drawing parallels to the gymnastics DID result, where Béla Károlyi's influence showed a raw estimate of 11 additional medals, we can estimate that hiring a world-class coach from Australia or the United States could yield 4–6 additional medals per Olympics.

Colombia has inherent advantages in cycling, such as its high-altitude geography and athletes' natural endurance [10]. However, its Olympic cycling performance has fallen short of its potential due to the lack of international-grade tactical guidance. Cycling is an ideal sport for focused investment by smaller nations, as it demands relatively lower resources compared to medal-heavy sports like swimming or athletics.Based on the volleyball DID analysis, introducing a European coach from cycling powerhouses like the Netherlands or France could increase Colombia's cycling medal count by 3–5 additional medals per Olympics.

5. Strengths and Weaknesses

5.1 Cox Proportional Hazards Model

- **Advantages**: The survival analysis characteristics of the Cox model are highly suitable for the task of "predicting countries winning their first medal" in this competition, as it directly models the "time to event." For countries that have not yet won a medal, the Cox model can fully utilize this right-censored data, unlike regression models, which require complete observations to be included in the modeling process.

- **Disadvantages**: The Cox model cannot capture nonlinear relationships between features, which may limit its ability to model complex factors effectively.

5.2 LightGBM Model

- **Advantages**: LightGBM is a gradient boosting model based on decision trees, capable of effectively capturing nonlinear relationships between features and target variables. This is particularly useful since the total number of gold and overall medals a country wins may be influenced by multiple complex factors (e.g., participation history, number of athletes, host country effects). LightGBM can flexibly model these relationships. In this competition, many complex features (such as athlete growth rate and event concentration) were constructed. LightGBM can efficiently handle high-dimensional data, automatically selecting important features while ignoring unimportant ones, thereby reducing the burden of manual feature selection.

- **Disadvantages**: For countries winning their first medal or countries with relatively few medals, the data may be sparse, and LightGBM may struggle to accurately fit the data distribution for these small-sample cases.

5.3 Difference-in-Differences Method

- **Advantages**: The DID model directly quantifies the coach's effect through interaction terms, providing a clear causal interpretation.

- **Disadvantages**: If the medal trends of the control group and the treatment group are not parallel before the intervention (e.g., the control group countries experience faster sports development), the inference results of the DID model may be unreliable.

Referances

[1] Cox D R. Regression models and life-tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187-202.

[2] Lifelines Documentation. (n.d.). Proportional hazard assumption. Read the Docs. Retrieved January 28, 2025, from https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Do-I-need-to-care-about-the-proportional-hazard-assumption?

[3] Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? JAMA. Published online March 13, 2020. doi:10.1001/jama.2020.1267

[4] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[5] Cox D R. Regression models and life-tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187-202.

[6] Olympics.com Biography, Lang Ping, https://olympics.com/en/athletes/ping-lang

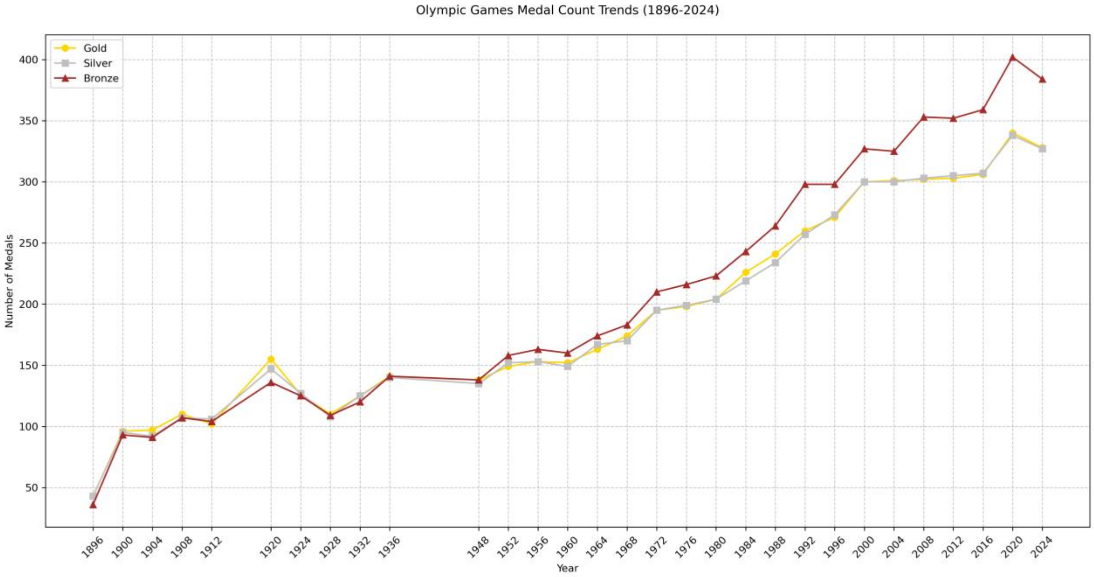
[7] USA Gymnastics Hall of Fame,https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/

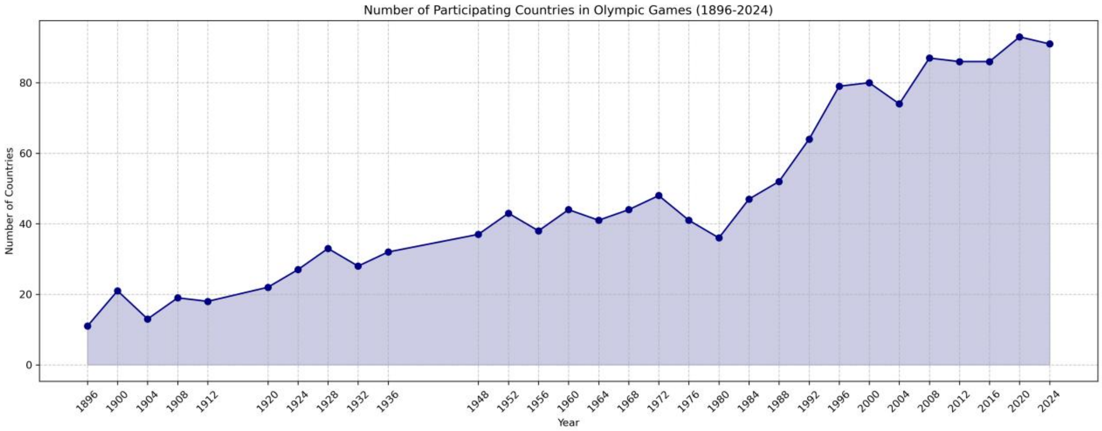
[8] Wikipedia contributors. (n.d.). 1984 Summer Olympics. In Wikipedia. Retrieved January 28, 2025, from https://en.wikipedia.org/wiki/1984_Summer_Olympics

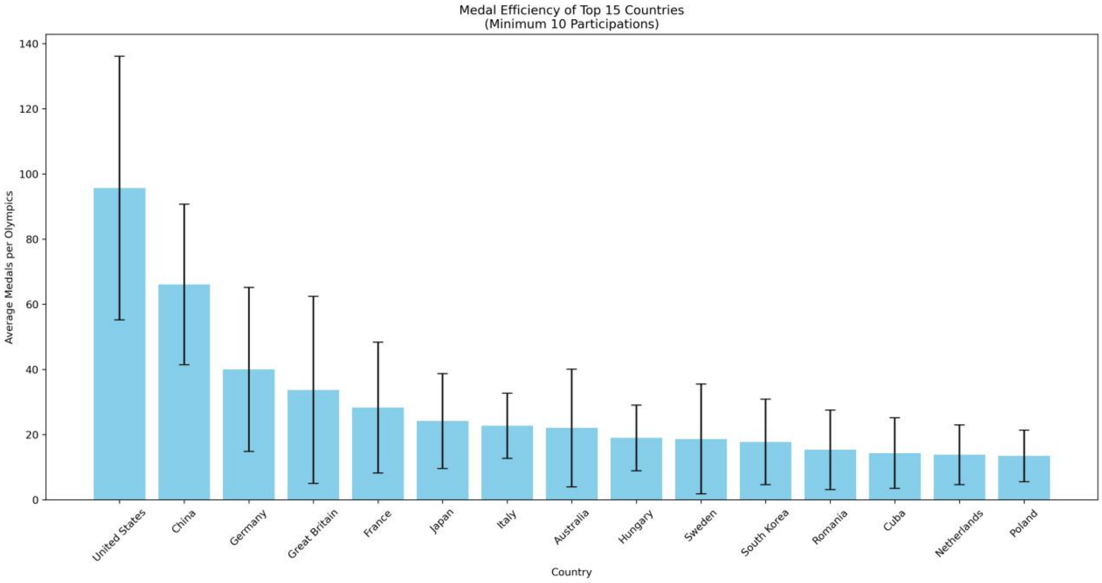
[9] Wikipedia contributors. (n.d.). Chad le Clos. In Wikipedia. Retrieved January 28, 2025, from https://en.wikipedia.org/wiki/Chad_le_Clos

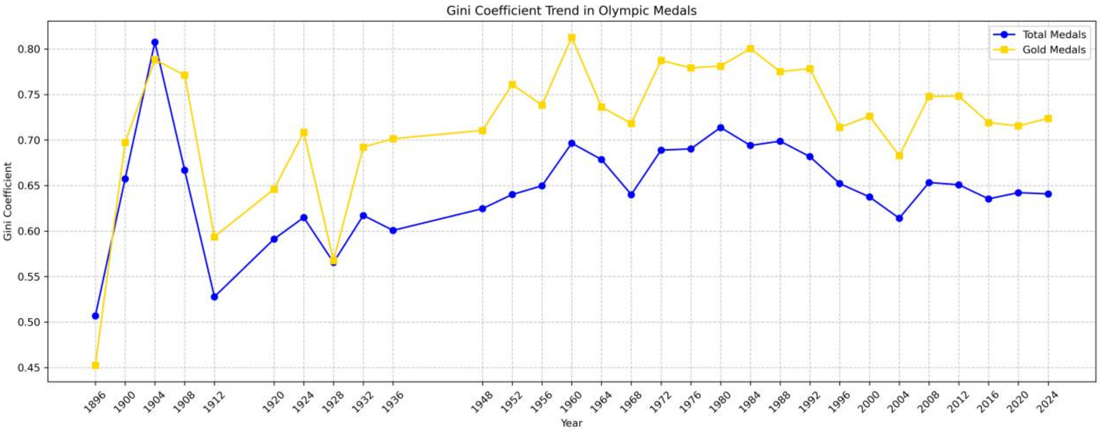
[10] Wikipedia contributors. (n.d.). Vuelta a Colombia. In Wikipedia. Retrieved January 28, 2025, from https://en.wikipedia.org/wiki/Vuelta_a_Colombia

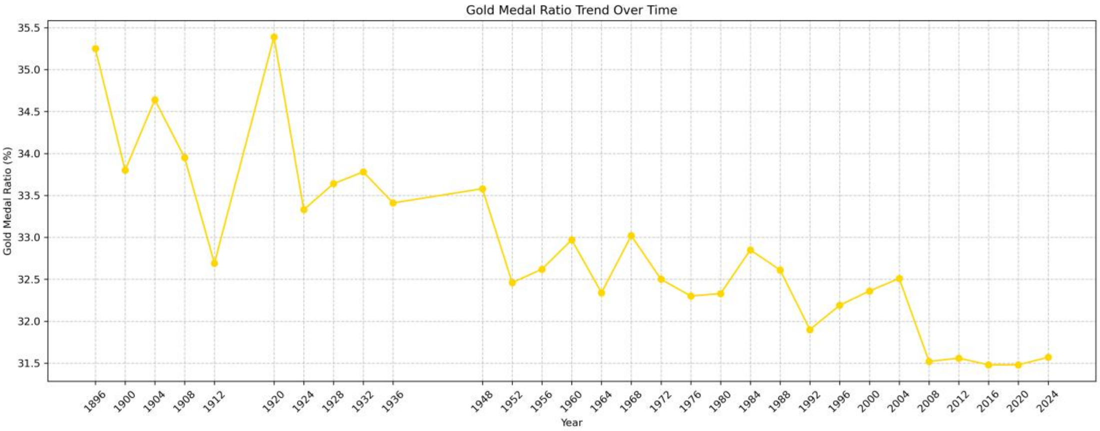
Appendix

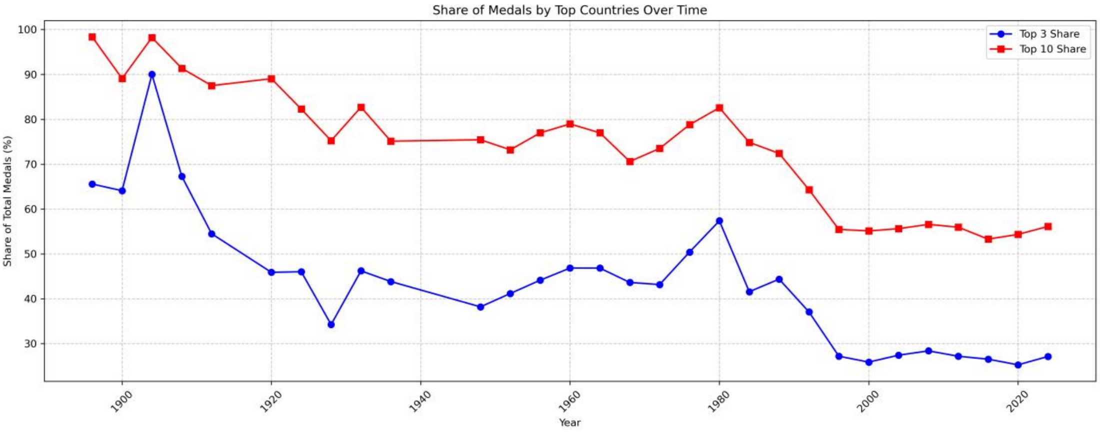

Olympic Games Medal Count Trends (1896-2024)
Gold
Silver
Bronze
Number of Medals
50
100
150
200
250
300
350
400
Year
1896
1900
1904
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
2024


Number of Participating Countries in Olympic Games (1896-2024)
Number of Countries
0
20
40
60
80
Year
1896
1900
1904
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
2024


Medal Efficiency of Top 15 Countries
(Minimum 10 Participations)
Average Medals per Olympics
0
20
40
60
80
100
120
140
Country
United States
China
Germany
Great Britain
France
Japan
Italy
Australia
Hungary
Sweden
South Korea
Romania
Cuba
Netherlands
Poland


Gini Coefficient Trend in Olympic Medals
Total Medals
Gold Medals
Gini Coefficient
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
Year
1896
1900
1904
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
2024


Gold Medal Ratio Trend Over Time
Gold Medal Ratio (%)
31.5
32.0
32.5
33.0
33.5
34.0
34.5
35.0
35.5
Year
1896
1900
1904
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
2024


Share of Medals by Top Countries Over Time
Top 3 Share
Top 10 Share
Share of Total Medals (%)
30
40
50
60
70
80
90
100
Year
1900
1920
1940
1960
1980
2000
2020