Problem Chosen	2025	Team Control Number
C	MCM/ICM	2516787
	Summary Sheet	

The Golden Guess - Predictive Modeling for Olympic Outcomes

Summary

This study explores predictive modeling for the 2028 Olympic Games by leveraging historical data from 2000 to 2024. Using statistical approaches such as Weighted Least Squares (WLS) and logistic regression, we analyze key factors influencing medal success, including historical performance trends, host nation advantages, and athlete participation. WLS modeling emphasizes the importance of recent Games, providing robust predictions for total medal counts. A logistic regression framework estimates the likelihood of nations winning their first-ever medals, incorporating predictors such as athlete profiles and proximity to prior medal-winning performances. The results suggest that the United States, China, and Japan will continue to dominate the medal standings, while nations like Saint Kitts and Nevis, Papua New Guinea, and Angola are strong candidates to achieve their first Olympic medals. The analysis also incorporates uncertainty quantification through bootstrapping, ensuring the reliability of predictions. These findings underscore the utility of data-driven insights in understanding and forecasting Olympic success, offering actionable recommendations for countries aspiring to enhance their athletic performance on the global stage.

TABLE OF CONTENTS

1. Introduction

1.1 Background

Though the legacy of the Olympic Games goes back thousands of years, its revival in 1894 with the first Olympic Congress instilled a new enthusiasm for global understanding and celebration of sports[5]. Today the Olympics captivate billions worldwide with inspirational achievements and the celebration of athletic excellence, cultural exchange, and global unity[4]. The Games embody the spirit of sportsmanship, bringing together over 200 participating teams representing sovereign states and territories. Traditionally held every four years, the Games began as a summer event, with the Winter Olympics introduced in 1924 to showcase cold-weather sports.

In recent years, the Olympics have embraced technological advancements, such as live streaming and social media, which have expanded their reach and accessibility. The addition of new sports, like skateboarding and surfing in the Tokyo 2020 Games, highlights the Olympics' adaptability to modern trends and appeal to younger audiences. Athletes compete across a wide range of disciplines, earning gold, silver, and bronze medals based on their performance. While iconic events such as the 100-meter sprint draw global attention for their intense competition, the Games also serve as a platform for smaller nations to achieve historic milestones. At the most recent Games in Paris, approximately 10,000 athletes competed, with countries like Botswana, Saint Lucia, and Albania winning medals for the first time, igniting national pride and celebrating their Olympic success[1].

While the Olympics continue to evolve and inspire global audiences, understanding the underlying factors that drive medal success remains a challenge. Patterns in medal distribution are influenced by a wide array of variables, including the number of events, host country advantages, number of returning and the strategies employed by individual nations. These dynamics raise important questions about the key contributors to Olympic success and how nations can optimize their performance on the world stage. To address these questions, this study leverages comprehensive athlete data and statistical models to uncover trends and offer insights into factors shaping Olympic outcomes.

1.2 Restatement of the Problem

While traditional predictive models like Nielsen's *Gracenote Model* take into account upcoming athlete data, countries' economics, and coach selections just before the games, our study focused on data only from historic games[3]. This research closed the gap in knowledge of predictive factors from historic data and added to the field of sports analytics by analyzing comprehensive datasets from all Summer Olympic Games from the years 2000-2024.

Our team used individual athlete data, host country data, previous medal tables, and a list of program events to determine how many gold medals a nation will win in the 2028 Olympic Games and the total number of medals a nation will win. Furthermore, based on our predictions, we derived insights into what nations would win their first medal in 2028 and quantified the number of silver and bronze medals nations will win. These insights offer actionable recommendations for countries aspiring to enhance their Olympic achievements.

Given the inherent uncertainty surrounding future events, our models accounts for variability through guiding markers that address the following objectives:

- **Predict** total medal counts per country, broken down by gold, silver, and bronze medals.
- **Estimate** the likelihood of new nations winning their first-ever Olympic medal.
- **Quantify** the degrees of uncertainty within the models to enhance reliability.
- **Investigate** patterns of interaction between athletes, host nations, and event types.
- **Identify** key statistical markers that are most indicative of an athlete's likelihood of winning a medal.

2. Preparation for Modeling

2.1 Assumptions

The premise of this problem relies on several major assumptions that must be touched upon. These assumptions allow for uniform measurements, consistent analysis, and applicable results when made. The assumptions are as follows:

Assumption 1: Past performances are indicative of future outcomes, meaning countries with consistent medal wins in previous Olympics are likely to perform similarly in future games.

Justification: Countries that consistently dominate certain sports due to well-established training programs and infrastructure.

Assumption 2: Individual sport events and athletes' performances are independent of one another, meaning one event's outcome does not directly influence another.

Justification: The result of one event does not directly influence others unless explicitly linked (relay-race) Ex:, a sprinter winning a 100m race does not directly affect outcomes in gymnastics.

Assumption 3: Host nations often experience a *"home advantage"* due to increased funding, athlete participation, and local support.

Justification: Familiarity with local climate and venues and the presence of home crowds can enhance performance.

Assumption 4: The number and type of events in each Olympics will remain relatively consistent, with only minimal additions or removals.

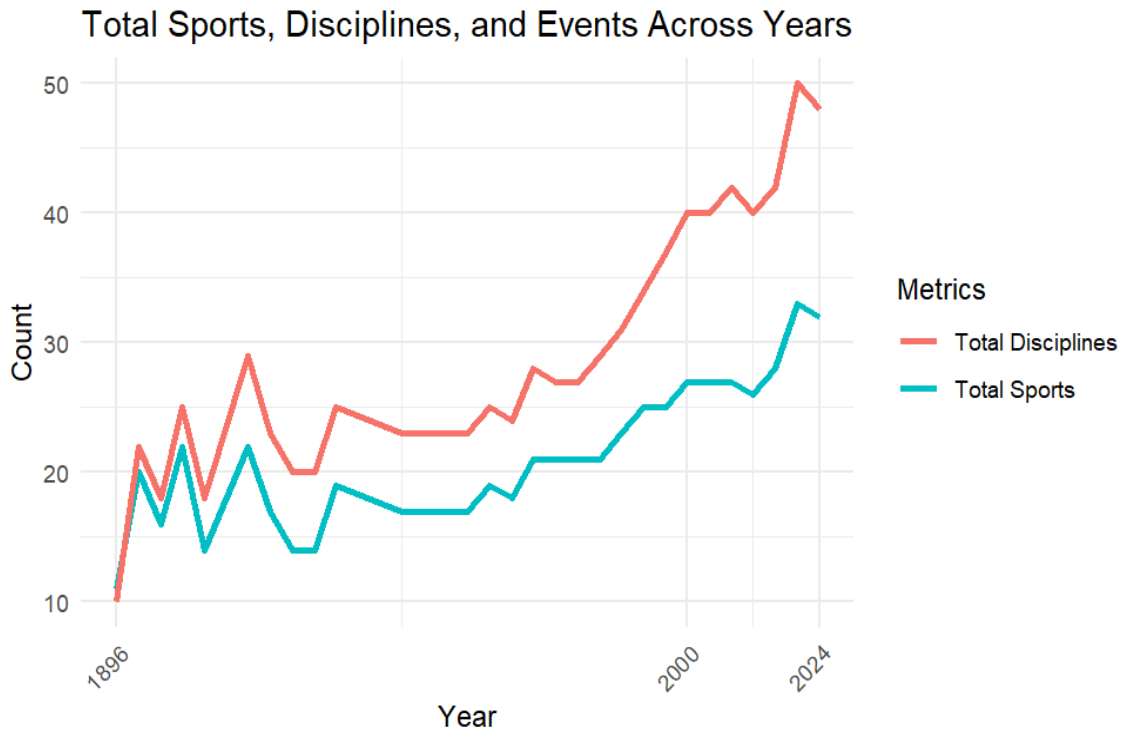
Justification: A sudden shift in number of events can create instability in comparison across years.

Assumption 5: Scoring criteria are assumed to be fair and standardized across all participating nations.

Justification: International federations oversee scoring and judging standards to maintain fairness.

Assumption 6: Model analysis and Calibration will begin from the year 2000.

Justification: The year 2000 goes far enough to collect relevant historical data whilst maintaining a stable level of shifts in events and sports.



3. Data Preprocessing

3.1 Interpreting Original Data

Starting our research by exploring the data given to us through *COMAP* Problem C Zip file (1. *athletes.csv,* 2. *hosts.csv*, 3. *medal_counts.csv*, and 4.*programs.csv*.), we discovered that historic medal count with country ranking, history of Olympic Game hosts, and athlete profiles all contained information regarding the name of the country.

***Data Source*:**	***Variable / Marker Definition:***	***Type / Example:***
Hosts	*Year:* Year of hosted Olympic Game.	*numeric*, {2000, 2004}
	Host: City and Country of hosted game.	*char*, {Athens, Greece}
Programs	*Sport:* A general category of athletic activity.	*char*, {Archery, Cycling}
	Discipline: A specialized branch within a sport. Focuses on a singular or set activity.	*char*, {Sprint, Track}
	Code: Abbreviation for Discipline.	*char*, {BOX, SWM}
	Sport Governance: Official Federation of Sport.	*char*, {FIFA, UCI}
	Years: Frequency matrix of sports / discipline included in a particular year.	*numeric*, {0 ,4, 12, NA}
Medal Counts	*Rank:* Overall placement of Olympic Team.	*integer*, {1,2,3,4,5,...}
	NOC: Name of County (according to ISO.)	*char*, {France, Denmark}
	Gold: Total gold medals count.	*integer*, {5, 10, 15, …}
	Silver: Total silver medals count.	*integer*, {4, 12, 17, …}
	Bronze: Total bronze medals count.	*integer*, {13, 17, 25, …}
	Total: Summation of gold, silver and bronze counts.	*integer*, {15, 22, 30, …}
	Year: Year of specific Summer Olympic Events.	*numeric*, {1896, 1960}
Athletes	*Name:* Name of athlete / Olympian	*string*, {Michael, Simon}
	Sex: Either identified as male or female.	*char*, {M, F}
	Team: Name of Country, or associated team.	*char*, {China, Finland}
	NOC: Name of Country (according to ISO.)	*char*, {CHN, FIN}
	Year: Year of specific Summer Olympic Events.	*numeric*, {1992, 2012}
	City: City location of Olympics.	*char*, {Barcelona, Seoul}
	Sport: General category of athletic activity.	*char*, {Judo, Hockey}
	Event: Specific competition within discipline.	*char*, {Swimming 400m}
	Medal: Whether an athlete placed in an event.	*char*, {No medal, Silver}

3.2 Data Cleaning

Host Country Data & Medal Counts Join:

We began our cleaning by parsing the 'Host' feature within the Hosts dataset. The columns of this dataset correlated with that of the Medal Counts dataset, where each dataset contained a variable describing year and country. Utilizing this key feature, we merged the two datasets along these specified columns in an attempt to consolidate the datasets that were provided. This newly created dataset, a merge of the Medal Counts and Hosts dataset, will henceforth be referred to as the Medal-Hosts dataset and accounted for the years 2000-2020.

Standardization of Name of Country Variable for Joins:

Before we could complete our join by the name of country feature (NOC) we had to convert the NOC values in the Medal-Hosts data set to the International Organization for Standardization (ISO) 3166 alpha-3 country codes that were used in the Athletes dataset. Previously, the Medal-Hosts dataset used full country names instead of ISO. Additionally, we decided that countries in both the Medal-Hosts and Athlete datasets who experienced balkanization would be dropped from our research without any data manipulation to distribute medals won by those countries to the descendant countries. This is because historic athlete data did not contain the current (then future) country names or ethnicities which prevents from determining athlete's origins in today's world. Due to the nature of our research, predicting the total medal counts for countries in the 2028 Olympics, we also dropped non-country data team data from our research. This resulted in removing countries such as Bohemia, Czechoslovakia, Australasia, etc. and non-nation teams such as the Refugee Olympic Team, and Mixed Team from our data.

Athletes Data Cleaning Pre-Join:

The Athletes dataset was an in-depth list of athletes from the 1800's Olympic Games to the most recent 2024 Games that contained: name of athlete, team, name of country, biology, the year the athlete participated in the Olympics, Event Name, and whether they won a medal in the event. Per our limitations, we began by filtering the data to only 2000-2020 year data and upheld the limitation that athletes from non-existing countries would be removed from our model. We analyzed the Athlete dataset to prepare the data for a merge with the Medal-Hosts data on the 'NOC' feature and generated useful statistics for prediction by country the athletes represented. We began by simplifying the sporting categorization… Through further analysis of

the Athletes dataset, we found that the column labelled “Sport” had in fact contained a mixture of sports classifications and discipline classifications. This motivated us to manually standardize this column to only contain disciplines, as this classification was highest in frequency. We instated three rules that would help us determine disciplines:

- Athletes cannot be listed to participate in multiple disciplines under one event, that is, a given athlete must be listed to be participating in one discipline per event.
- Keywords in a given event are used to determine the apt discipline for the given event. Ex.) An event listed as *“10km Swimming”* would fall under the discipline *“Marathon Swimming”* due to the unit *(km)* and the word “Swimming.”
- Synonymous disciplines would be merged into the IOC-recognized discipline. Ex.) “Trampolining” would be merged into the recognized discipline “Trampoline Gymnastics”

This process successfully classified each event into their proper respective disciplines, and proved to be key in removing duplicate athlete profiles due to sports having different names for the same sport.

3.3. Feature Creation

We derived the number of sports each athlete played in their career, the number of unique events they participated in, and the number of Olympic Games they attended. Additionally, because we had a record of if the athlete won a medal for a particular event and year, we created counts of gold, silver, bronze medals the athlete had won. These *'Athlete Profiles'* were organized by the nations they represented with duplicate profiles being removed from the data.

Our team created a more mathematically intense statistic for each athlete profile called the *'Athlete Score'* which became 'Average Team Rating' in the final model dataset. This was done because the Medal-Host dataset provided a ranking of countries, however, it only included countries that had previously won. This meant that the model had only athlete summary statistics to predict off of for non-ranking nations. We desired to create a more informative feature that would provide the model with information on all countries participating in the olympics. This 'Athlete Score' was calculated using the number of gold medals, silver medals, and bronze medals the athlete had won plus the 'experience' of the athlete, which was the number of Games the athlete had participated in.

$$ln(AthleteScore) = Athlete's(MedalCount) + Athlete's(Experience)$$

Due to the manipulation of athlete's scores by scaling the athlete's individual scores by the natural log, the parameters can then be interpreted as odds. Additionally, instead of incremental increases, more years and medals won by athletes have a multiplier effect on the score which is more representative of the significance of winning olympic medals and creates a larger value difference in athlete scores between athletes who have won one medal versus three.

$$Average\ Team\ Rating\ = Data\ from\ \{Olympic\ Year\},\ Country's\ Athlete\ Scores\ Mean$$

Athlete and Medal-Hosts Join

We joined the data via a match merge on 'NOC' after completing the aforementioned cleaning, allowing for the final dataset features seen below to be used in the predictive modeling.

3.4 Data Transformation

One of the challenges we encountered was how to account for the 2028 Olympics in Los Angeles without having the actual results. To address this, we decided to use the outcomes of the 2024 Summer Olympics as a proxy for estimation.

Final Variables:	***Variable Description:***
Notes: Derived from Data	
Rank:	Final placement in previous Olympic Tournaments
Participated Events:	Total number of events a given team participated in.
Year:	The Year when the Olympic was held.
Total Athletes:	The total count of athletes from a given country.
Athletes Returned:	The total count of olympians who appear in At Least 2 Olympics.
Multi Event Athletes	Number of athletes who participate in numerous events. (min 2)
Avg Team Rating:	Mean athlete score aggregated for each nation.
Discipline:	Each sport was considered as a variable capturing athlete frequency.

4. Statistical Methods

4.1 Weighted Least Squares

Weighted Least Squares (WLS) regression is an extension of Ordinary Least Squares by addressing the case of where homoscedasticity (constant variance) or errors is violated. To adjust for this, weights are introduced to the model. Our weighted function is as follows:

$$min \sum_{i=1}^{N} w_i (Y_i - \hat{Y}_i)^2$$

Typically, weights are assigned the value of the inverse variance ($1/\sigma^2$). For our purposes, we will utilize them as an adjustment with respect to year. That is, as time approaches to our target Olympic year, the weight has a stronger pull.

$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

This final result implements the combination of the transpose of the design matrix, a diagonal weight matrix, and our result vector to achieve the coefficient estimates.

4.2 Likelihood Function

To estimate the likelihood of a nation winning their first ever Olympic medal (any medal) in the 2028 Olympics, we develop a probabilistic likelihood function utilizing the historical framework.

We first assume nation independence. That is, the chances of a country winning its first medal is conditionally independent of others given our markers. Additionally, countries that have come close (4th or 5th) are more likely to win their first medal in the future games. We can model the likelihood of a Country i using as Bernoulli Random Variable:

$$L(\beta) = \prod_{i=1}^{N} P(W_i = 1|X_i)^{W_i} \times (1 - P(W_i = 1|X_i))^{1-W_i}$$

$$l(\beta) = \sum_{i=1}^{N} (W_i) log P(W_i = 1|X_i) + (1 - W_i) log(1 - P(W_i = 1|X_i))$$

- W_i is a binary indicator for win (Wi = 1) or loss (Wi = 0) in the 2028 Olympics
- X_i is a vector of predictors for a $Country_i$
- β are the coefficient vector estimates for given p-1 predictors
- $f(X_i|\,\beta)$ is the logistic function $P(W_i = 1|X_i) = f(X_i|\,\beta) = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_{i,1} - \cdots - \beta_{p-1} X_{i,p-1}}}$

Thus, by using the maximum likelihood estimator (MLE), we can compute the probabilities of a nation winning their first medal based on predictors such as number of athletes, proximity to past medal wins (ex: finishing 4th or 5th), and participation in specific events. The resulting model provides not only predictions but also measures of uncertainty, enabling us to rank nations by their likelihood of earning their first medal in the 2028 Olympics. This probabilistic approach

integrates historical data with predictive modeling, creating a robust foundation for understanding and projecting Olympic success patterns.

4.3 Uncertainty Quantification

One immediate challenge is that the 2028 Olympics have yet to occur. Thus, there is a high degree of uncertainty regarding how accurate our estimation is. We opt to use bootstrapping as a resampling-method to estimate the sampling distribution of the mean by drawing random samples with replacement from our current set of observations. The advantage is that we can quantify uncertainty without holding strong assumptions of the underlying distribution, since it assumes a non-parametric nature. By repeating sampling, we can simulate the variability and approximate for the true sampling distribution.

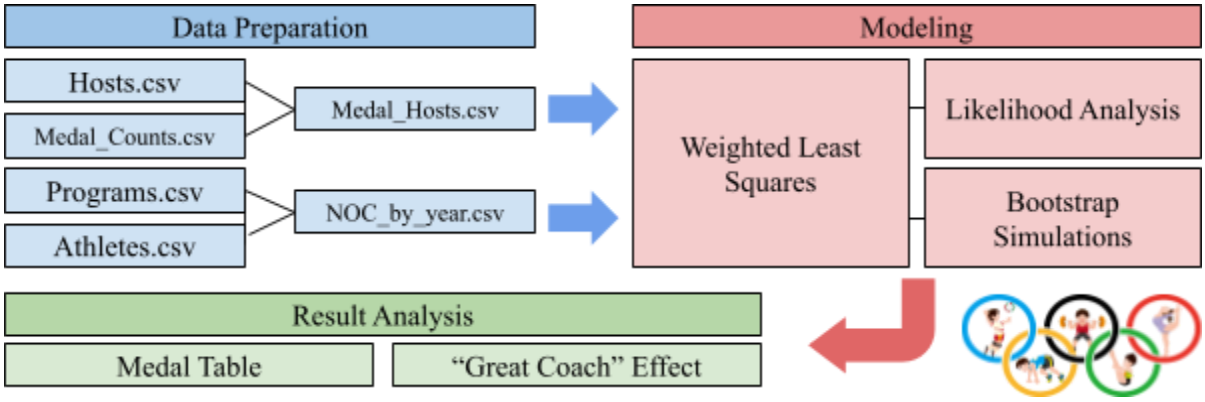
Suppose we have some data $[X_1, X_2, ..., X_n]$

The Bootstrap Estimate for each sample: $\hat{\theta}_b = f(X^*)^b$

The Bootstrap Distribution for sampling distribution follows: $\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, ..., \hat{\theta}^*_{(B)}$

Confidence Intervals for $\hat{\theta}$: $CI_{1-\alpha} = [\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}]$

With certain countries having limited historical data, such as those that have never won medals, making traditional analytical methods less reliable; bootstrapping helps address this by simulating variability in such cases. It also plays a crucial role in quantifying uncertainty when predicting future Olympic outcomes, which are influenced by evolving factors like new sports or changing rules. Additionally, bootstrapping allows for comparative analysis by ranking countries based on the robustness of their predictions, generating most reliable forecasts.



5. Analysis

5.1 Total Medal Count

Our initial approach was to develop a methodology that accounts for both country-specific and sport/event-specific factors, while also incorporating an adjustment for time. The goal was to assign greater importance to recent Olympic Games, reflecting their higher relevance to future projections. Specifically, we emphasized that results from previous Olympics closer to 2028 should carry a greater weight.

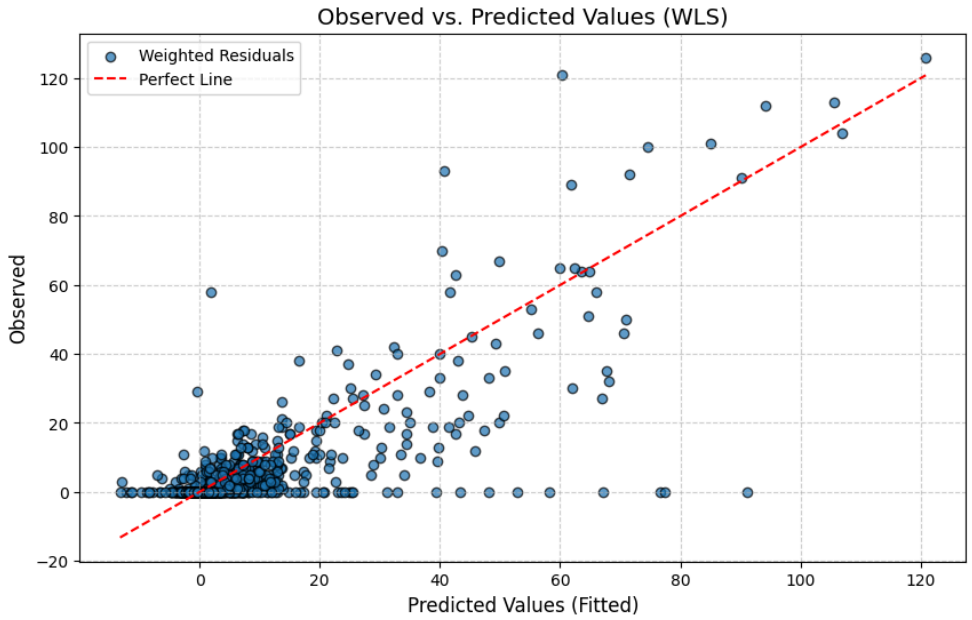
To achieve this, we tested two different weight levels with respect to year. These weights allowed us to assess how different levels of emphasis on recency impacted the model's ability to predict total medal counts. We fitted a WLS model using the finalized list of predictor variables.

For each level, we decided to test two different weights with respect to year:

1. $weight_1 = e^{-\lambda(2024-Year_i)}$
2. $weight_2 = \frac{Year_i - 2000}{max(Year_i) - min(Year_i)}$

We will first test for weight-1 fitting all parameters from our final variables list.

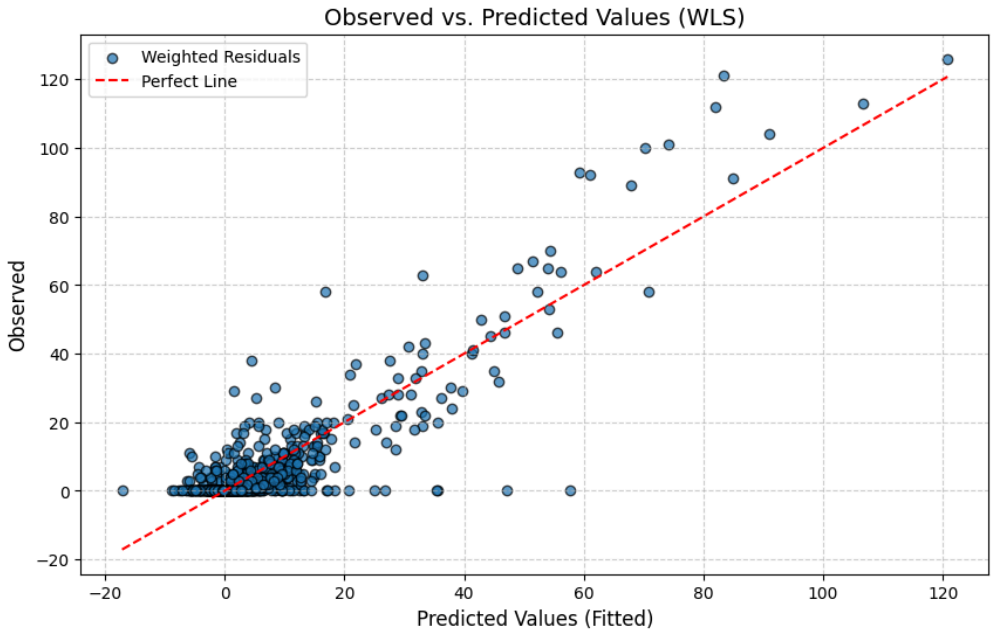
Adj R^2: .967 , F-Stat: 683.9. $_{61,1367}$ … AIC: 1.682^4



Exponential weighting in the WLS model assigns higher importance to recent Olympic data, reflecting their greater relevance for predicting 2028 outcomes. The weights decrease exponentially as the year moves further from 2028, λ controls the rate of decay. This approach emphasizes recent trends, prioritizing recent results (ex: 2020 and 2024), the model better captures current dynamics, smoothing the temporal trend and reducing the risk of overfitting.

We then explore the test for weight-2. Weights increase linearly as the year approaches 2028, giving more recent data (ex: 2020 and 2024) higher importance, while older data (ex: 2000) is down-weighted less aggressively than with exponential weights. This method provides a balanced approach, ensuring that historical trends still influence the model.

Adj R^2: .866 , F-Stat: 152.10 $_{61,1367}$ … AIC: 10^{12}

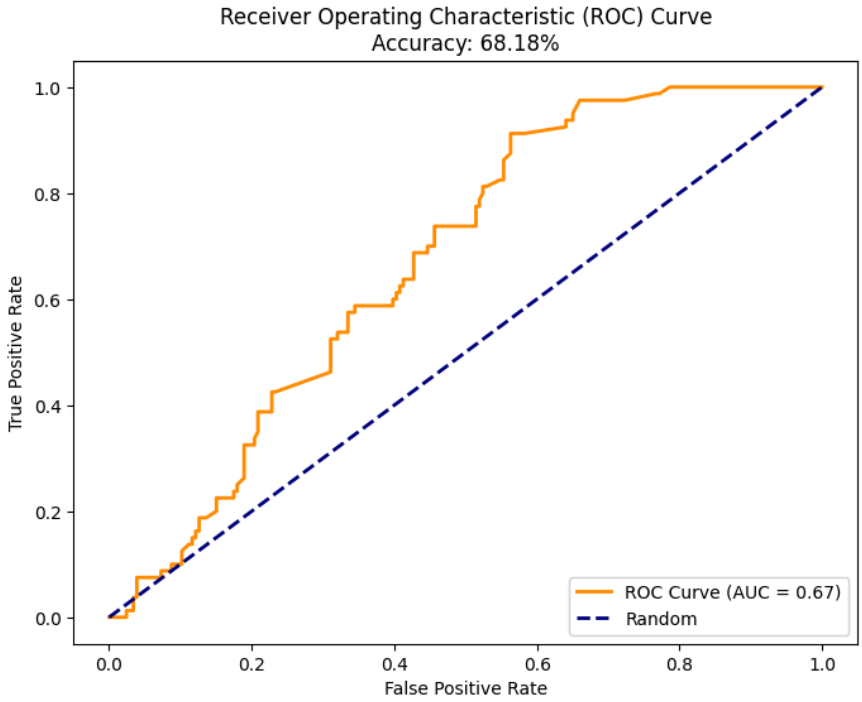


To ensure the robustness of our results, we conducted a sensitivity analysis by varying the weight schemes and observing their impact on the model's performance metrics.

1. **(Weight 1):** Provided a balanced approach, with high explanatory power and stable projections. This approach demonstrated a strong fit, suggesting that more recent Olympics carry the most predictive power for projecting total medal counts. However, the aggressive weighting risks underestimating the relevance of historical trends
2. **(Weight 2):** The lower adjusted R^2 and inflated AIC suggest that this level may not fully capture the importance of more recent Olympics, leading to a less effective model fit.

5.2 Likelihood Estimation

Our objective was to predict the likelihood of countries winning their first-ever Olympic medal using a weighted likelihood estimation approach. The Receiver Operating Characteristic (ROC) curve, displayed above, illustrates the model's performance in distinguishing between countries likely to win a medal and those not. With an Area Under the Curve (AUC) value of 0.67, the model demonstrates moderate discriminatory power. The accuracy of the predictions stands at an approximate 68%, as indicated in the figure. This suggests that while the model can reasonably predict outcomes, there is room for improvement in capturing additional predictive features or refining the weighting methodology. The diagonal line (dashed) represents a random classifier, and the orange curve above it shows the model's improvement over random chance. These findings provide insight into how likelihood estimation can support predictions for emerging medal-winning countries.

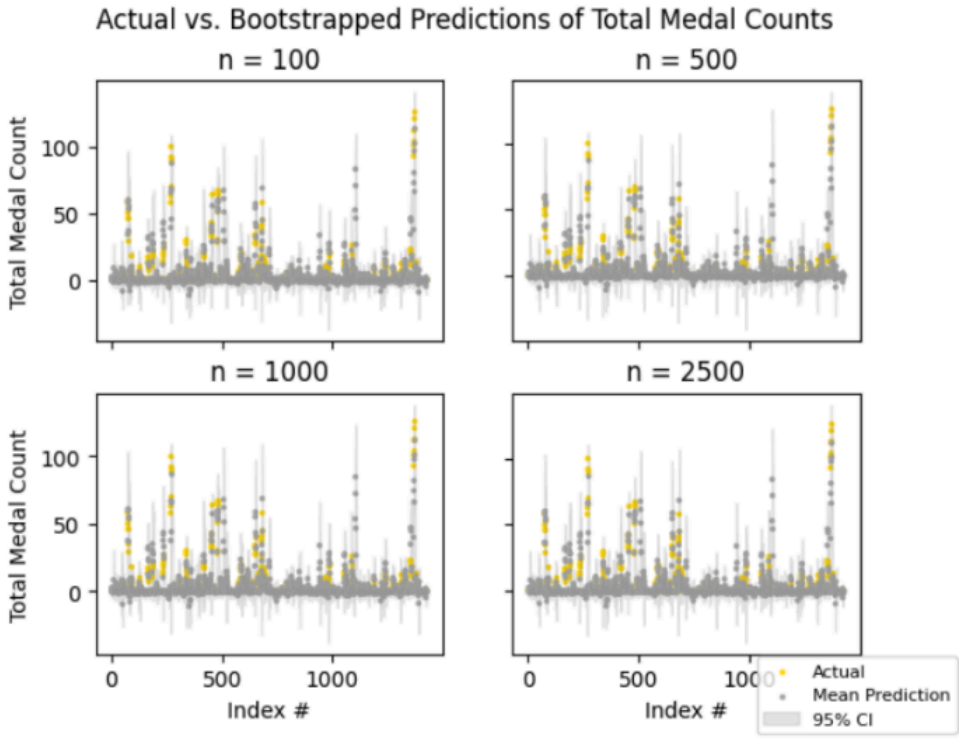


5.3 Uncertainty

Lastly, we wanted to measure how robust our model was to the future Olympics in 2028. We decided to run Bootstrap simulations for the *Total Medal Count* estimated from our WLS model.

We took the predicted estimates for each nation and then resampled the data using different sample sizes. The figure below shows the results of a bootstrap simulation assessing the

predictive performance of our WLS model in estimating total Olympic medal counts for various teams. Each subplot represents simulations with different sample sizes, N = 100, 500, 1000, and 2500 respectively. The x-axis represents the index of observations (teams), while the y-axis captures the total medal count, both actual and predicted.



The actual medal counts, represented by the gold dots, provide the basis for comparison against the model's predictions; The mean predictions are shown as gray dots. The associated uncertainty of each prediction is represented through 95% confidence intervals (CI), the light gray ranges that encompass each mean value. Across all subplots, the variability of the predictions ever so slightly decreases as n increases, showing that the model produced leaves little to be desired, per the Central Limit Theorem (CLT).

While the model generally aligns with the observed medal counts, some outliers were particularly apparent for teams with extreme medal counts (both low and high). These discrepancies may point towards unaccounted variables or unnecessary assumptions within the model that may need to be reviewed in order to reduce potential overfitting of the model.

6. Results

6.1 Medal Table Results

2024 Paris	Gold	Silver	Bronze	Total
United States	40	44	42	126
China	40	27	24	91
Japan	20	12	13	45
Australia	18	19	16	53
France	16	26	22	64
Netherlands	15	7	12	34
Great Britain	14	22	29	65
Republic of Korea	13	9	10	32
Italy	12	13	8	33

2028 Los Angeles	Gold	Silver	Bronze	Total
United States	39	42	40	121
China	37	28	25	90
Japan	19	13	13	45
Australia	18	19	18	55
France	16	26	22	64
United Kingdom	15	21	26	62
Italy	12	13	15	40
Canada	9	7	10	26
New Zealand	8	5	1	14
Ireland	6	2	5	13

Our model predicted that the United States, China, Japan, and Australia would perform marginally worse in 2028 but maintain the top four positions overall, while France, Great Britain, and Italy will maintain their medal counts. The countries that significantly decrease in performance are the Netherlands and the Republic of Korea, allowing new countries into the top ten contenders: Canada and New Zealand.

6.2 New Medalist Countries & The Great Coach Effect

The most-likely new medalists for the 2028 Olympic Games included Saint Kitts Nevis, Papua New Guinea, and the Cayman Islands,based off of Average Team Scores created from athlete profiles from these national teams. We believe that the island nations of Saint Kitts Nevis and the Cayman Islands would benefit the most from the "Great Coach Effect" due to their already high likelihood of winning a medal in the 2028 Games. Additionally, due to their geographic proximity to Brazil who has an up and coming gymnastics program, as well their proximity to the United States, it is possible that either of these nations could participate in joint coaching. In particular, we propose this idea for the Cayman Islands because they had their first Olympic gymnast in 2024 so their small team would not require as much resources or funding compared to pairing their gymnastics team with another nation. Additionally, Reagan Rutty, the Cayman Island gymnast won Miss Universe in 2024, so to allow her to be jointly coached with either Brazil's gymnast or the United States' team would be beneficial to public relations between countries[9].

Papua New Guinea's proximity to Australia, which is predicted to come in fourth place again in 2028, would be another great opportunity for joint coaching to enable a highly likely first-time medalist country to increase their chance of being a medalist. Our recommendation is

that Australia combines swimming programs with Papua New Guinea because Australia's most successful sport in the Olympics is swimming,[11] and that Papua New guinea's most successful athlete is Ryan Pini, a 100m competitor for swimming.

7. Discussion

7.1 Continuation of Research

Returning Athletes

Our team did not take into account the possibility of athletes returning to future games, instead, we opted to treat each Olympic Games as independent events in this way, though we assumed that previous performance is an indicator of future outcomes. We recommend that future studies on predictive analytics for Olympic Games conduct more research on the likelihood of athletes returning per country team and account for these athlete's previous performances.

7.2 Limitations of Research

The quality of data is always a limitation of data science and statistical models, however, there is also the limitation of unknown future variables such as number of events and participating countries, and confounding variables within the historic data of Olympic Games results. Below are the most important factors that the team considered, and believed, to heavily influence the accuracy of predictive models for Olympic Games[12]

Name of Country Inaccuracies

Our team chose to use data from the years 2000 to present because of higher consistency with event types and number of events in recent years but other reasons we were limited to these Olympic Games were the 'NOC' inaccuracies due to balkanization, countries boycotting the Olympic Games, and due to cases of exclusion based on racial justice.

As the International Olympic Committee states, the goal of the Olympic Games is to promote global unity, peace, and increased cultural understanding among youth[4]. There have been years where nations or groups of nations have boycotted the Olympics, such as the Soviet Union and East Germany, in 1984, and the 1976 Montreal Games boycotted by 28 African nations in protest of segregation policies[6]. The inverse, where countries are banned from the Olympic Games, is also true. South Africa was banned from the Olympics in 1964 and did not return until 1992[6]. Additionally, of the nations who did participate, some have experienced "Balkanization" where their borders have changed shape, split into smaller nations, or changed

their country's name. Nations like Yugoslavia, Czechoslovakia, Korea, and others have experienced this split due to nationalism, war, or both. It is impossible to redistribute the medals won by athletes in these nations according to current territories based on the data we have.

If our team took into account historic games prior to 2000, the result would be that communities from these boycotting, banned, or now-non-existent nations would lack the recognition and statistical measure of success in those historic games. We choose the year 2000 to present in order to limit this impact on our predictive model.

8. Conclusion

Through various data cleaning procedures and analysis of athlete's scores and creation of team features, our team created a Weighted Least Squares (WLS) regression model in an attempt to predict the medal distributions at the Los Angeles 2028 Olympic Games. Our results show that the United States, China, and Japan are on track to become the top three countries at the end of the Games, with Italy, Canada, New Zealand, and Ireland projected to make it into the top ten standings. In addition, Saint Kitts Nevis, Papua New Guinea, the Cayman Islands, and Angola are the most likely countries to win their first ever medals in the upcoming olympics, through our utilization of our logistic regression model. On top of this, we find that the “Great coach” effect is the most beneficial for Saint Kitts Nevis and the Cayman islands due to their geographic proximity to Brazil, a major competitor within the realm of gymnastics.

The models outlined within this paper have more than one consideration that should be kept in mind when utilizing them, however due to various constraints, we were unable to mitigate them to the point where they became irrelevant. However, even after considering the outlining shortcomings, these models have plenty of potential in accurately predicting the outcomes of the 2028 Olympic games.

9. Appendix

Repository containing written code to produce outputs and results utilized in this paper

https://github.com/Kenzo-Hoobert/MCM-ICM-2025

References

[1] Consortium for Mathematical Competition in Modeling, "Problem C: Models for Olympic Medal Tables," *COMAP,* (2025)
https://www.immchallenge.org/mcm/2025_MCM_Problem_C.pdf

[2] Geeks for Geeks, "Pandas dataframe.groupby() Method," (2024)
https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/

[3] Gracenote, "Nielseb's Gracenote Expects USA, China, Great Britain, France and Australia to Lead 2024 Paris Olympic Games Medal Table," *Nielsen News* (2024)
https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/#

[4] International Olympic Committee, "Around 5 billion people - 84 per cent of the potential global audience - followed the Olympic Games Paris 2024," *International Olympic Committee News* (2024)
https://www.olympics.com/ioc/news/around-5-billion-people-84-per-cent-of-the-potential-global-audience-followed-the-olympic-games-paris-2024

[5] Kalliopi Sakavitsi, "The History of the Olympic Games," *Olympics News* (2024)
https://www.olympics.com/en/news/the-history-of-the-olympic-games

[6] Lauren Lee, "What were the most notable boycotts in Olympic history?" *KCRA News* (2024)
https://www.kcra.com/article/boycott-protests-olympics-history/61159266#:~:text=In%201976%2C%2028%20African%20nations,segregation%20policy%20known%20as%20Apartheid.

[7] W3 Schools, "Python Sets," *Python Tutorial W3Schools,* Python Sets Tutorial
https://www.w3schools.com/python/python_sets.asp

[8] Yibi, H. (n.d.). *Lecture 14: Weighted least squares and logistic regression*. University of Chicago. Retrieved from https://www.stat.uchicago.edu/~yibi/teaching/stat224/L14.pdf

[9] Cayman Compass, "Cayman's first Olympic gymnast crowned Miss Universe Cayman Islands 2024,"*Cayman Compass* (2024)
https://www.caymancompass.com/2024/09/01/caymans-first-olympic-gymnast-crowned-miss-universe-cayman-islands-2024/

[10] International Olympic Committee, "African athletes excel at the Paris 2024 Olympics, from artistic gymnastics to fencing," *Olympics News* (2024)
https://www.olympics.com/en/news/paris-2024-africa-athletes-excel-win-medals

[11] International Olympic Committee, "Most productive Olympic sport for Australia - swimming, the gold standard", *Olympics News,* (2024)
https://www.olympics.com/en/news/australia-best-olympic-sport

[12] Olympedia, "Papua New Guinea Olympic Committee" (2024)
https://www.olympedia.org/countries/PNG