Problem C: 2028 Summer Olympics Medals Count

Ziwen Zhu, Tuyen Truong, Li Dongyuan

January 2025

Contents

1 Summary

Our paper presents am analysis of the 2028 Summer Olympics medal prediction using a neural network model, integrating both a Feedforward Neural Network (FFNN) and Long Short-Term Memory (LSTM) components. The model is designed to predict medal counts (gold, silver, bronze) for each National Olympic Committee (NOC) and year, incorporating athlete-related features and a New Medals Model to forecast countries winning their first Olympic medals. In addition, a correlation analysis identifies key relationships between specific sports and NOC medal outcomes.

The results of our model demonstrate a significant improvement in prediction accuracy compared to traditional approaches such as ordinary least-squares regression that have been done in other literature. Furthermore, the analysis highlights the persistent advantage host countries enjoy in the Olympics, with certain countries consistently dominating medal counts across multiple Games.

Despite these successes, the paper acknowledges certain limitations, such as the reliance on historical data for verification and the exclusion of certain performance factors like coaching impacts, athlete injuries, and psychological impacts. Future improvements to the model could incorporate more granular data, including athlete-level performance metrics and a broader range of external factors. With these enhancements, we aim to refine our predictions and gain deeper insights into the dynamics of Olympic success in the next 2028 Summer Games.

2 Introduction

The 1996 Summer Olympic Games in Atlanta were the largest Games held since the commencement of the modern Olympic Games in 1896. Almost 11,000 athletes from 197 countries participated in 271 events from 26 sports. About 5 million spectators and visitors came to Atlanta and approximately 3.5 billion people tuned in to watched the Olympic Games on television. While more countries than ever were represented in the "final medal count at the 1996 Summer Olympic Games (78 countries received at least one medal), some countries regularly win more medals than others. Many attempts have been made in different literature to distinguish the many probable reasons why. Population and a country's economic development (measured by GDP-gross domestic product or gross domestic product by capita) is cited most frequently as the main factors affecting a country's Olympic success. However, we only need to observe the recent performances of nations such as Cuba and India to see that other factors are involved (in 1996, Cuba won 25 medals, while India won one medal). Past research efforts to account for differences in success between nations have relied on linear modeling techniques such as ordinary least-squares (OLS) regression. In this paper, we construct several models such as two simple feed forward neural network and and one times series based model. We train each neural network model using different categories of data given.

3 Data Exploration

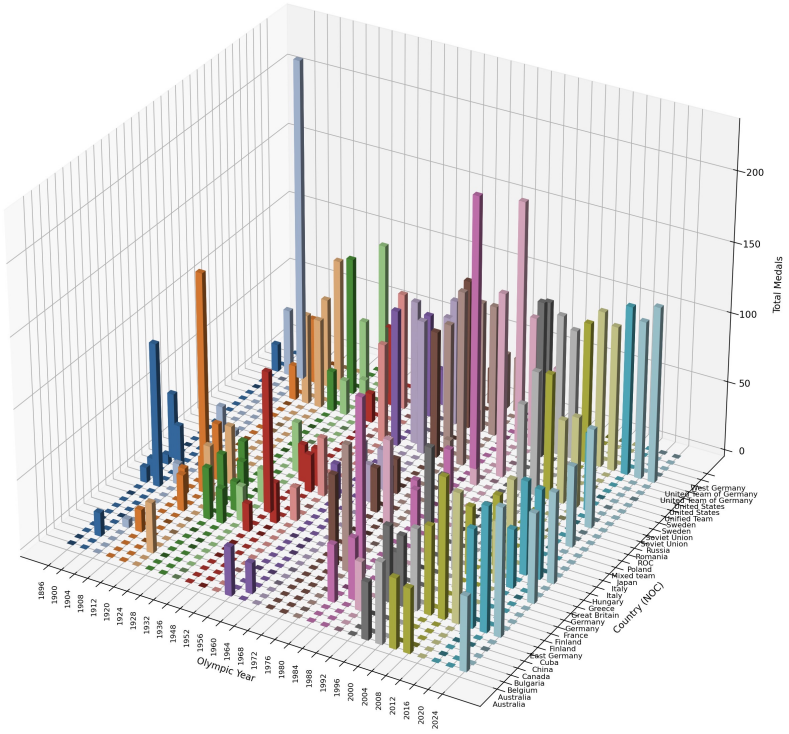

Figure 1: 3D Bar Graph showing data distribution.

We begin with an exploratory overview of the data, referencing Figures 1 through 4 for initial insights and patterns. Figure 1, a 3D bar graph of medals across multiple Olympic years and countries, shows steadily taller bars in more recent Games—a signal that participation and event offerings have likely expanded over time. Figure 2 presents the distribution of athletes by gender, with male athletes outnumbering female athletes overall, though the narrowing gap hints at more inclusive participation in recent Olympics.

Moving to Figure 3, which plots the number of medals per athlete, the histogram peaks at zero or one, indicating most competitors do not accrue multiple medals. However, the "long tail" at higher counts implies a select group of star athletes often drives a substantial portion of each country's overall medal haul. Lastly, Figure 4 focuses on the "host advantage," illustrating how hosting can produce a medal bump for some nations yet relatively modest gains for others. This variability suggests that while facilities and home crowd support may help, consistent outcomes also depend on established athlete development programs and historical strengths in specific sports.

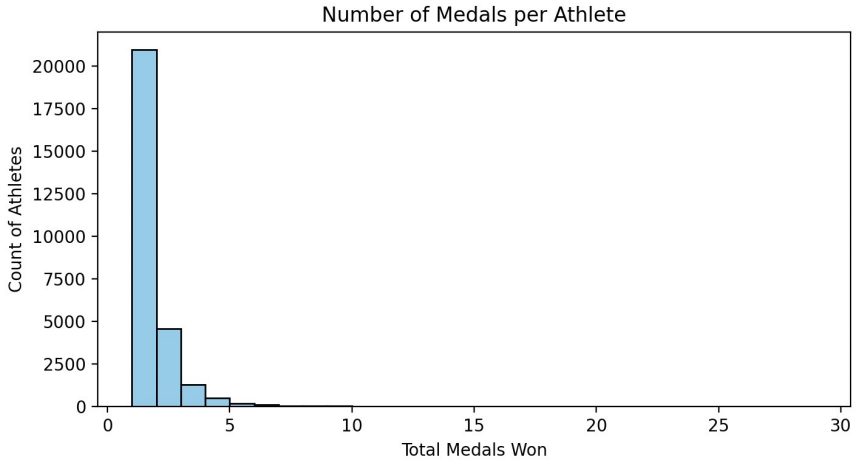

Figure 2: Number of medals per athlete over total medal wins.

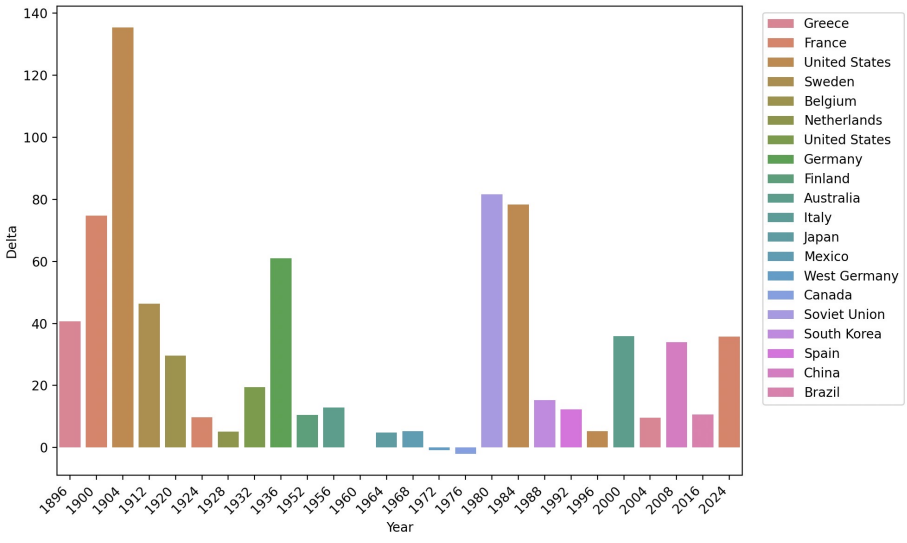

Figure 3: Advantage of host country in winning medals.

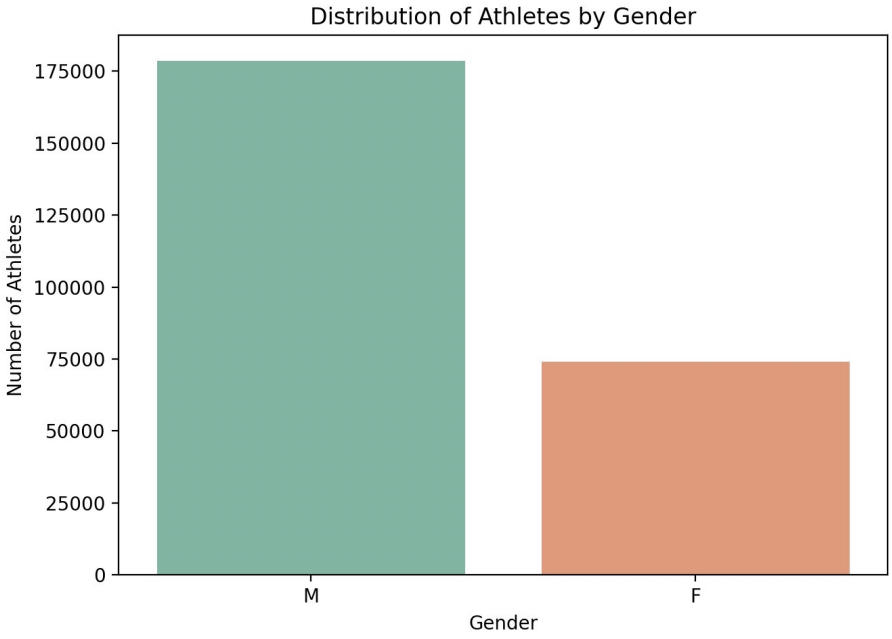

Figure 4: Gender Distribution of Athletes.

4 Model(s) Description

We begin by cleaning the data, discarding rows with problematic string patterns, removing rarely contested events, and limiting our dataset to NOCs appearing in recent Olympics or with sufficient historical entries. After merging duplicate records, we aggregate everything at the (Year, NOC) level.

Next, we generate the following features:

(i)SeedPlayer: An integer capturing whether an athlete earned a medal in their first or second Olympics.

(ii)ElitePlayer: An integer that sums total medals for an athlete, subtracts one, and adds an extra count for each gold medal.

(iii)ParticipationCount: A count of how many different Olympics an athlete has attended (only used for RetiringPlayer calculation).

(iv)RetiringPlayer: A flag set by (ParticipationCount 3) when it is positive, otherwise zero.

(v)IsHost: A binary indicator (0 or 1) identifying whether the athlete's NOC matches the host NOC for that year.

(vi) MinorEventCount: An integer measuring how many specialized or rare events the NOC entered.

(vii)TrendAthCount, TrendNumEvents, TrendPlayerList: Three slopes or ratios gauging changes over up to four previous Olympics in athlete headcount, event participation, and roster overlap.

Once these features are created, we convert the NOC labels to one-hot vectors and standardize all numeric fields. Our main predictive pipeline uses a Long Short-Term Memory (LSTM) module to forecast future athlete-related features, with its outputs then feeding a feedforward neural network (FFNN) that predicts gold, silver, and bronze medal counts. We also build a smaller FFNN variant for identifying first-time medals if an NOC did not medal in the prior Olympics, and test a model for the Great Coach Effect—though the latter shows no significant impact within our dataset. A heatmap analysis of sports versus NOCs further illustrates where certain nations concentrate their medal hauls.

5 Extrapolation Algorithms

5.1 Model 1 Simple Neural Network

Our FFNN model was able to predict medal count by taking the predicted athlete features, along with fixed features like the target year and NOC, to estimate medal counts (gold, silver, bronze) for each country. The network is trained to minimize the loss measured as Mean Absolute Error across multiple Olympic cycles. Batch normalization and drop out for first few layers are implemented, ensuring that the model is capable of generalizing to unseen data, such as the 2028 Olympics. We began by removing rows with unwanted characters (such as digits or periods in the "Team" column, or Roman numerals in the "Team"

column, these typically means the athlets is not representing a country, instead, these team name usually refers to clubs and institutes like schools). The total number of these rows are relatively small compared to the size of the dataset, so we can safely remove them and still make sure the major trend is preserved. This ensures that only valid rows of data remain. Then, the next crucial step was to have the "NOC" feature be one-hot encoded, turning it into a set of binary columns that indicate the presence of a specific country so the model can learn better. Furthermore, the features and target variables (Medals: Gold, Silver, Bronze) are standardized using StandardScaler. This ensures that all features have the same scale, which is essential for the FFNN to perform optimally. To actually "train" the model, we splitted the data chronologically, where data from years before 2016 is used for training, and data from 2016 onwards is used for testing. This mimics the real-world scenario where we train on past data and test on future data. From there, the FFNN is constructed using multiple dense layers with the Swish activation function, BatchNormalization, and Dropout regularization to prevent overfitting. The model predicts three output values: Gold, Silver, and Bronze medals for each NOC and year. The optimizer used is AdamW, which includes weight decay to help improve generalization. The loss function is mean squared error (MSE), while the model's performance is measured using mean absolute error (MAE).

5.2 Model 2 Time-Series Based Model

The first step in the extrapolation algorithm is the prediction of missing athlete-related features for countries where current data is unavailable. To address this, we utilize an LSTM network, a specialized type of recurrent neural network (RNN) designed for sequential data. The LSTM is trained on 27 years of historical Olympic data, capturing trends and temporal dependencies that inform predictions for future Olympic Games. The LSTM model is built to handle time-series self-predicted data (using historical athletes' feature to predict future athletes' feature), where the number of lag periods used in predictions can be specified.

With these features, the LSTM produces estimates for key athlete-related features (e.g., performance trends, historical improvements) for each country and year, which are subsequently fed into the FFNN. This enables the model to account for shifts in athlete performance over time, even when precise data for the target year is not available.

5.3 Model 3 Neural Network 2

To improve on our original neural network model, this model processes and prepares Olympic athlete data for predictive modeling, including cleaning, feature engineering, and scaling steps. It begins by setting the environment to allow for multi-threaded operations, enhancing the performance of subsequent data processing. Feature engineering follows, where new variables are created to represent player characteristics such as the number of medals won ("SeedPlayer"),

the count of different medals ("ElitePlayer"), and the number of years an athlete participated ("ParticipationCount"). A special feature "RetiringPlayer" is calculated based on participation count. The code also includes logic for identifying host countries based on a mapping of Olympic cities to their respective National Olympic Committees (NOCs). Any missing cities or NOCs are cleaned out ,after a confirmation that its a minor part of the set, from the dataset for clarity. Additionally, trend-related features are computed using linear regression to capture the slope of changes in player count, event count, and player overlap over the years for each country (NOC) across past four Olympic games.

The dataset is then aggregated by year and NOC, summing numeric values, including medal counts (Gold, Silver, and Bronze). One-hot encoding is applied to the NOC column to create binary features for each country, and the data is scaled using the StandardScaler for both features and target columns. The scaling models and the one-hot encoder are saved as pickle files for later use. The final processed dataset, along with the transformed features, is ready for machine learning modeling, which will be displayed in our results.

6 Verification of Solutions/Comparisons

To ensure the validity and accuracy of the processed dataset and its features, we ensured that there was no duplicate handling in the data, and we used a standard scaling technique to establish that the data is appropriately normalized for machine learning models. The scaling process was verified by inspecting the distribution of the features before and after scaling, confirming that the transformed data exhibited a mean of zero and a standard deviation of one, as expected. Note that in our verification that there was no duplicate errors in our data, the final dataset was checked for remaining duplicates using the duplicated() function, which confirmed that no redundant records persisted. This step guarantees that each athlete's medal counts are uniquely represented in the dataset, avoiding potential over-counting or data contamination.

7 Results

1. 2024 Performance vs. Model Predictions

Our model's forecasts for the 2024 Olympics align reasonably well with the actual outcomes, though a few discrepancies exist. For example: (i)The United States total was under predicted by about 6 medals (predicted 120 vs. actual 126). (ii) China was over predicted by around 4 medals (95 vs. 91). (iii) Great Britain was over predicted by about 6 medals (60 vs. 54).

Across all nations, mean absolute errors for gold, silver, and bronze medals were 0.83, 0.82, and 0.97 respectively, indicating that for each medal type, our predictions typically fall within about one medal of the actual totals.

2. Projected 2028 Medal Counts

Overall Leaders: For Los Angeles 2028, our feed forward neural-network

model continues to project the United States at the top of the medal standings (gold, silver, bronze, and total). China follows closely, reflecting its consistent performance.

Next Tier: A distinct gap separates the top two from a cluster of strong contenders—Great Britain, Australia, France, and Japan—whose gold medal counts hover around 10–20, with total medals ranging between 30 and 60.

Countries Likely to Improve: Some nations, including Australia, France, and the Netherlands, appear on an upward trajectory. Our forecasts suggest they may improve medal totals in part due to a steady rise in specific sports (for instance, cycling or swimming).

Countries Likely to Decline: Certain traditionally strong teams (e.g., Italy, Germany) show slightly lower predicted totals than in 2024, potentially reflecting generational turnover in key events or plateauing athletic pipelines.

3. "First Medals" for 2028

We define a "first medal" as any medal earned by a country that did not medal in 2024. A specialized version of our neural-network model highlights moderate probabilities for several lower-ranking NOCs to earn medals under this definition. While precise projections vary, the likelihood of one or more of these nations reaching the podium in 2028 remains noteworthy, given they exited the 2024 Games with zero medals.

4. Analysis of the "Great Coach" Effect

We attempted to identify patterns in the data that might reveal a strong coaching impact—such as sudden medal increases linked to a notable coach hire. However, due to limited data in single-sport or single-event contexts and the often subtle, multi-year nature of coaching influence, no significant or consistent patterns emerged from our current analyses. Therefore, we do not have measurable evidence of a "Great Coach" factor in our final results.

5. Heatmap Insights on Sport–NOC Synergies

The row-normalized heatmaps (covering top 30 NOCs across various time spans) indicate that many countries concentrate their medal points in particular sports. For instance: (i) Jamaica (when present among the top 30) shows a high concentration in Athletics. (ii) Italy often registers stronger results in events such as Fencing. (iii) Great Britain typically scores higher in Cycling disciplines. (iv)Japan stands out in sports like Baseball and, in some years, certain martial arts.

By highlighting each NOC's proportion of medal points across different disciplines, these heatmaps help explain why some nations excel in certain events and reinforce the sport-specific nature of medal acquisition.

6. Overall Observations and Takeaways

The 2024 predictions validated our modeling approach: on average, the error per medal category was less than one medal across the top contenders. For 2028, the United States and China are predicted to remain dominant, with a mid-tier of developed nations vying for the next positions. Countries that earned no medals in 2024 show a realistic chance of breaking through in 2028, according to our specialized model focusing on first-medal probabilities. The "Great Coach" hypothesis could not be confirmed from our data. Coaching impacts

may exist but are not clearly discernible within our aggregated Olympic-level datasets. Heat map analyses reveal that many nations' successes concentrate heavily in a handful of sports; focusing resources on these high-return disciplines can significantly influence overall medal totals. These findings combine both the 2024 performance comparisons and the 2028 projections, integrated with our heat map-based sport insights, giving a cohesive look at how NOCs may fare at the upcoming Los Angeles Olympics.

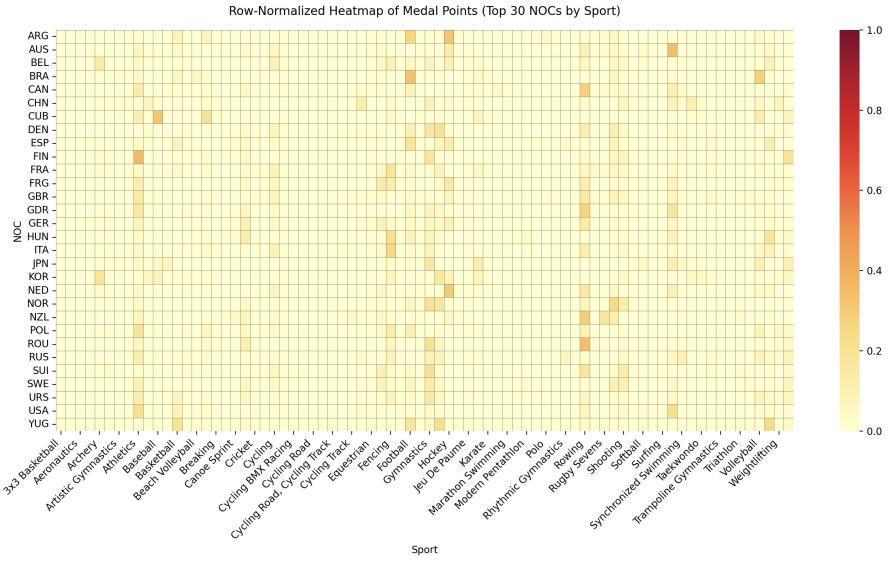

Figure 5: Historical Heatmap of medal distribution.

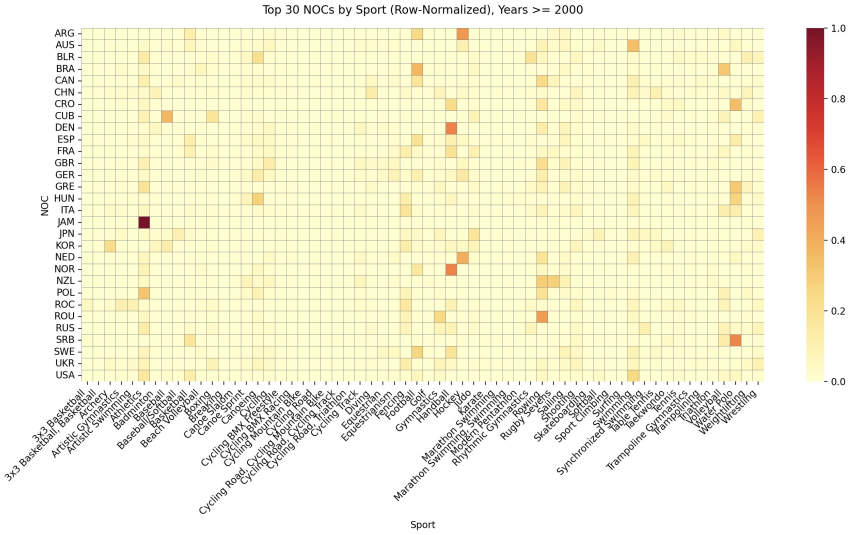

Figure 6: Heatmap for years greater than or equal to 2000.

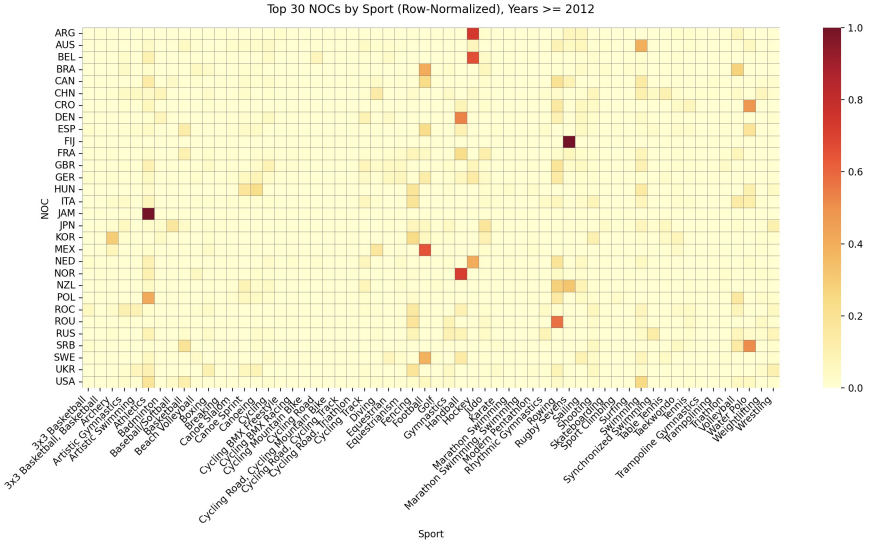

Figure 7: Heatmap for years greater than or equal to 2012.

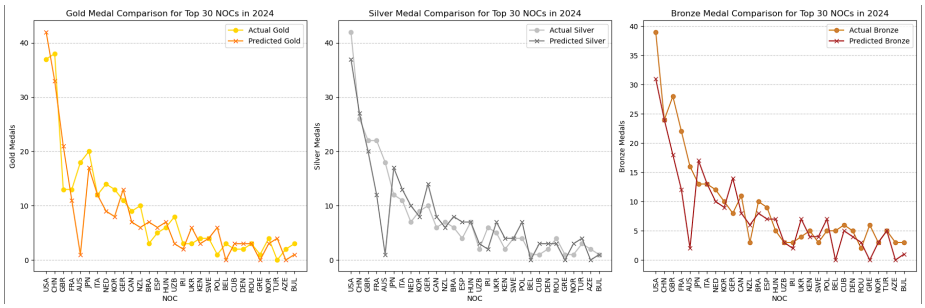

Figure 8: Graphs comparing the medal counts for each medal type for 2024.

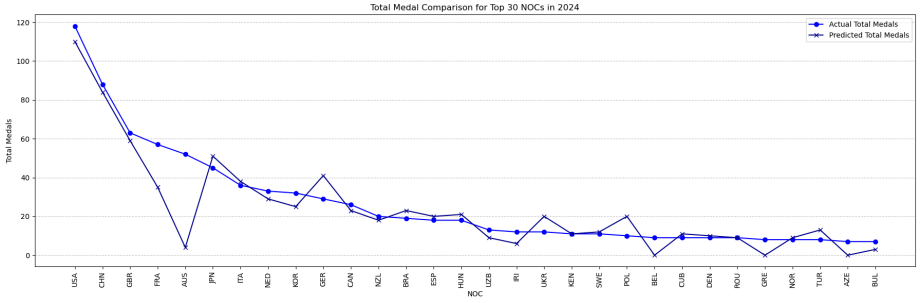

Figure 9: Trend Graph comparing the total medal counts for 2028.

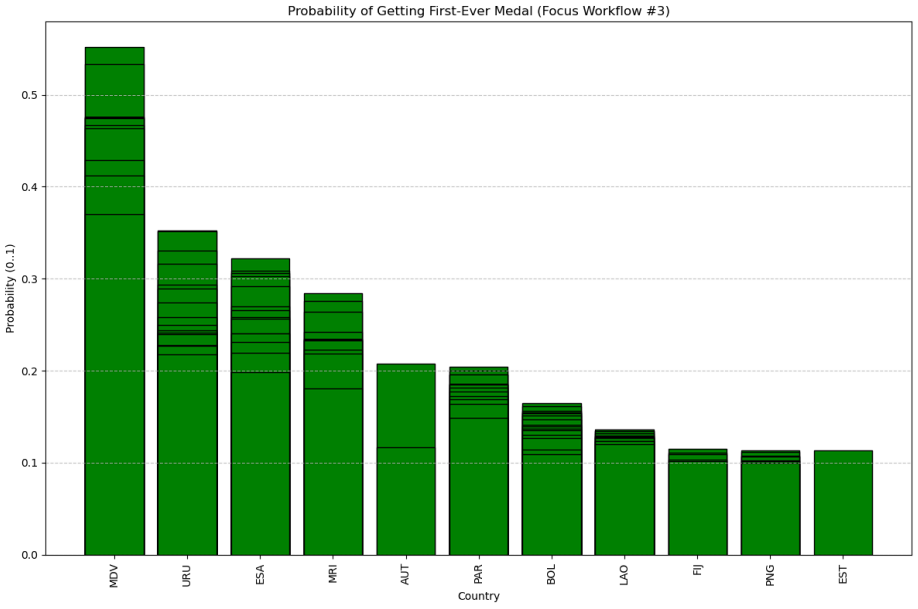

Figure 10: Probability Chart For Earning First-Ever Medal.

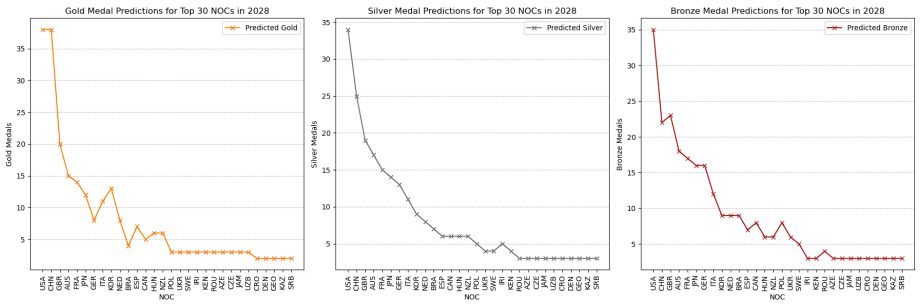

Figure 11: Graphs comparing the medal counts for each medal type for 2028.

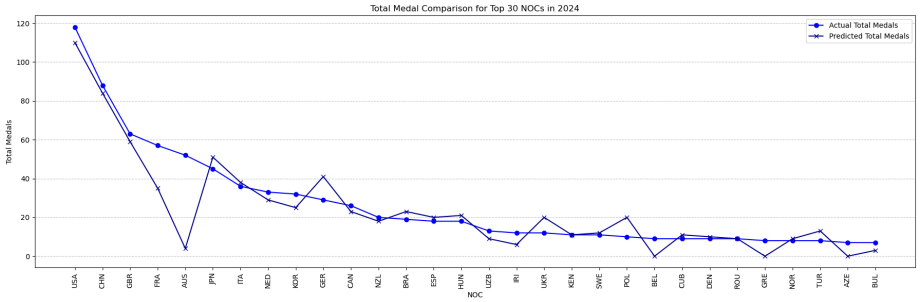

Figure 12: Graph depicting Total Medal Predictions for the Top 30 Countries.

8 Conclusion (Discussion of Limitations and Improvements)

Our model's analysis of Olympic data, focused on medal counts, athlete performance, and country trends, reveals important insights into the competitive nature of the Games and the influence of factors like home-country advantage. Our findings show that the majority of athletes win few medals, while a small group of elite athletes dominate with multiple victories. Additionally, host countries consistently perform better in the Olympics, likely due to factors such as home crowd support and familiarity with the venues and geography. Despite this advantage, global powerhouses such as the United States and the Soviet Union (in the past) have maintained strong medal counts throughout the years, indicating that, while hosting offers some benefits, sustained performance requires systemic strength in sports development and talent identification.

However, this analysis also comes with several limitations that should be acknowledged. First, the dataset primarily focuses on medal counts, which overlooks other important performance metrics such as athlete participation or the broader impact of coaches and national training programs. Although we implemented a model that includes features related to trends in athlete participation and events, it would be beneficial to extend the analysis to factors such as athlete injuries, psychological impacts, team dynamics, and the impact of the "Great Coach Effect". Additionally, the data used here is based on a limited number of Olympic Games, which may not fully capture the dynamics of global sports development over time.

Furthermore, while trends such as the host country advantage were observed, our current analysis does not account for the full spectrum of external variables that may influence Olympic outcomes, such as geopolitical events, economic factors, and changes in the structure of the Games themselves. The reliance on historical data also means that any sudden shifts in international competition or rule changes may not be fully captured by the model. Moreover, data quality and completeness are crucial factors; missing or incomplete records for certain athletes or events may have skewed some of the results, leading to potential biases in our conclusions.

To improve this analysis, future work could incorporate a more detailed exploration of athlete-level factors, including detailed performance metrics (e.g., times, distances, points), and model the "Great Coach Effect". Integrating a more robust dataset with higher granularity, such as training histories and athlete progression, would allow for more precise predictions of medal outcomes. Additionally, extending the timeframe of the study to include more recent Olympic Games would provide a more comprehensive understanding of how trends evolve, especially in response to changes in global competition and international sporting initiatives. The inclusion of advanced statistical techniques or machine learning models could further refine the predictive power of the model, allowing for better handling of complex relationships and non-linear patterns in the data.

9 References

- Edward M. Condon, Bruce L. Golden, Edward A. Wasil, Predicting the success of nations at the Summer Olympics using neural networks, Computers Operations Research, Volume 26, Issue 13, 1999, Pages 1243-1265, ISSN 0305-0548, https://doi.org/10.1016/S0305-0548(99)00003-9.
- https://digitalcommons.iwu.edu/cgi/viewcontent.cgi?article=1248context=parkplace
- Forecasting the Olympic medal distribution – A socioeconomic machine learning model - ScienceDirect swp0000.dvi
- https://www.sciencedirect.com/science/article/pii/S0169207009002088
- https://www.sciencedirect.com/science/article/pii/S0305054899000039
- https://medium.com/data-science-portfolio-ideas/building-an-impressive-data-portfolio-predicting-olympic-medal-outcomes-with-machine-learning-444b355adb26