Predictions of the 2028 Olympic Games

Team Control Number: 2528318

January 2025

Problem C
Advisor Id: 127906

Problem Chosen	2025	Team Control Number
C	MCM Summary Sheet	2528318

As the Olympic Games have risen enormously in popularity within the 21st century, every team is doing everything they can to increase their chances of winning a medal. Our goal is to create a model to give accurate predictions on the outcome of the 2028 Olympic Games hosted in Los Angeles, US.

Detailed information on the 2028 Olympics has not been released yet such as the specific contestants, and gender split among contestants. Thus we made the assumption that the athletes from 2024 will also participate in 2028 and that there will be a 50/50 split between male and female athletes.

- First, we take into account the foreseeable changes in 2028 as we engineer the data for 2028 based on the trends and patterns from the past 3 games in 2016, 2020, 2024.
- After analyzing the trends from these 3 games we were able to predict the average number of contestants per sport and average number of events per sport for 2028.
- Next, we analyzed the distribution of country representation in the Olympic Games to determine how to proportionally allocate contestants across different countries.

Once we had a complete dataset, we analyzed the relationship between various features and our target variable—the total number of medals. By calculating the Pearson correlation coefficient for each feature, we observed that there was no clear linear relationship across the data. As a result, we opted to use the Random Forest model, which is well-known for its ability to handle non-linear data effectively.

Next, we needed to determine an appropriate train/test split. We considered two options: a time-based split (before 2008 for training, after 2008 for testing), and a random 70/30 split. We ultimately chose the time-based split, as it seemed most suitable for avoiding data leakage and simulating real-world usage.

This approach ensures that our model is not tested on data it has already seen, which is ideal for predicting outcomes for the 2028 Olympics. Unfortunately, our model's performance was suboptimal because the distribution of the data before 2008 significantly differs from that after 2008. Prior to 2008, there were fewer contestants, sports, and events, leading to a mismatch between the training and testing data distributions.

We then decided to use the traditional 70/30 train-test split, which yielded excellent results with 99.9% accuracy on the test data. However, when applying the model to the predicted 2028 data, we found that the total medal predictions showed only slight variations from the 2024 outcomes to the point of almost being identical. This highlighted areas in our approach that could be improved.

Specifically, the 2028 data was primarily based on the 2024 data. To better capture potential changes, we should have averaged the data from the 2016, 2020, and 2024 Olympics before constructing the 2028 dataset. This would have introduced more variability and provided a more robust prediction model.

Table of Contents

Contents

1 Introduction

1.1 Background

Dating back to ancient Greece, the Olympic Games were held every four years, showcasing the most skilled athletes competing for medals. This tradition has not only been preserved, but has flourished in the modern era, evolving into a global competition. Initially, the Olympics were exclusively for Greek men, but today they welcome athletes of all genders and nationalities, each striving to win medals for their countries.[1]

The viewership of the Olympics has grown exponentially over the past century. In 1924, when the Paris Olympics became the first to be broadcast, approximately 650,000 spectators followed the event. By the 2024 Paris Olympics, viewership had soared to an estimated 5 billion people—roughly 84 percent of the global population.[2] As shown in Figure 1 [3], this rise in popularity has been accompanied by a growing interest across all sports categories. With such a significant increase in global attention, the stakes are higher, and each nation's team is determined to maximize their chances of success.

Given the global importance of the Olympics, analyzing historical data from past games can provide valuable insights and predictions for future outcomes. A critical aspect of this analysis involves examining the factors influencing each nation's preparation. This includes the athletes, the host location, and a variety of other key variables. Rather than focusing on individual events, national pride often hinges on the total number of medals won, making this the primary variable to predict. Since medal counts are discrete values—countries can only win whole numbers of medals (e.g., 0, 1, 2, etc.)— thus the focus will be on discrete predictions that encompass all Olympic sporting events.

Our goal is to utilize historical Olympic data, spanning from the 1896 Games to the most recent 2024 Paris Olympics, to develop accurate predictions of country rankings for the 2028 Los Angeles Olympics. Beyond generating a ranking list, we aim to analyze the key factors influencing a country's performance. This deeper analysis will provide valuable insights that nations can consider when preparing for future games.

[1] Abraham, H. M., & Young, D. C. (2025, January 21). Olympic Games. Encyclopædia Britannica. https://www.britannica.com/sports/Olympic-Games

[2] Paris Olympic Games. (n.d.-b). https://www.olympics.com/ioc/news/around-5-billion-people-84-per-cent-of-the-potential-global-audience-followed-the-olympic-games-paris-2024

[3] Voices, E. (2024, July 28). Summer Olympics: 128 years of history in 5 charts. https://earthsky.org/human-world/summer-olympics-128-years-history-5-charts/

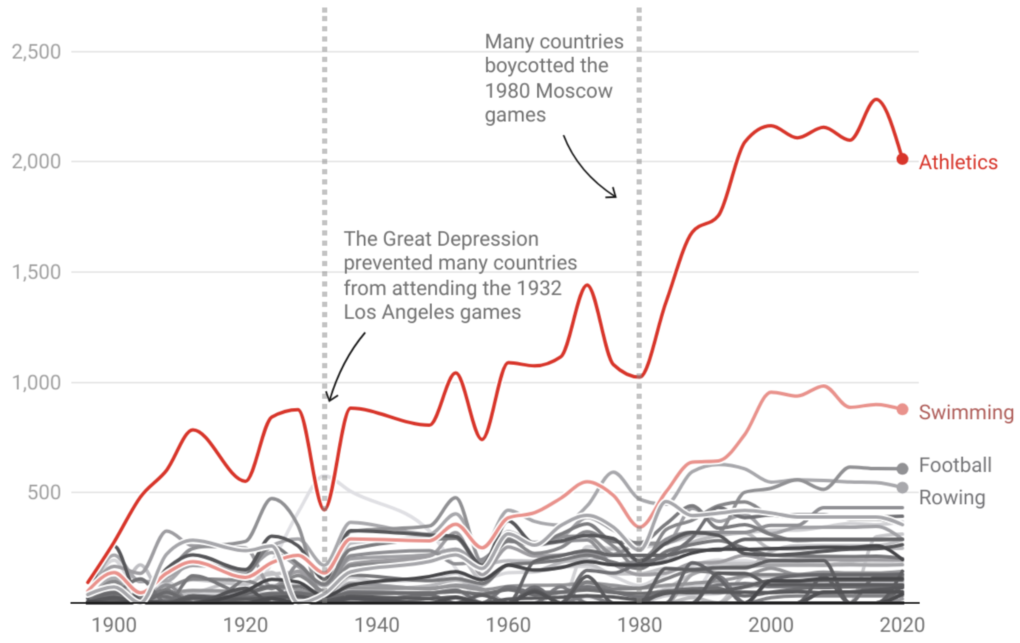

Figure 1: Number of Contestants for each sport category throughout the century

1.2 Problem restatement

Our goal is to create a comprehensive ranking of all countries participating in the 2028 Olympics, with a particular focus on both the top-performing and lowest-ranking nations. Additionally, we will conduct an in-depth analysis of the key factors driving these outcomes, offering valuable insights into the dynamics that influence Olympic success.

The problems that we need to solve are:

- What are the top 5 countries that will win the 2028 Olympics and how many medals will each win?
- What are the bottom 5 countries?
- Which country will improve the most, which will do worse?
- Is the home team advantage a factor?
- Will a country win their first medal in 2028?

1.3 Fundamental Assumptions

Since detailed information about the 2028 Olympics has not yet been released, we are unable to fully replicate the exact scenario using data from previous years. For instance, we do not know the precise number of countries that will participate or the exact number of athletes representing each nation.

To utilize the available data effectively, we must make certain assumptions about the 2028 Olympics to generate predictions. However, to ensure our analysis is as accurate as possible, we will base these assumptions on trends observed in previous Olympic Games, carefully replicating patterns and dynamics from past events.

- **Incorporating 2024 data for 2028.** We are using the latest data from 2024 to act as the foundation for 2028 data predictions, incorporating new rows to account for the announced changes to the 2028 games.
- **50/50 split between genders among the contestants.** The 2024 Olympics were the first Olympic games to adopt equal representation between male and female among the competitors. We are assuming that this policy will continue to 2028.
- **Every medal is weighted equally.** In reality, if we have a country that has 50 gold medals and another country with 50 bronze medals, these countries should not be ranked as equals but since we are ranking countries based on medals in general, these countries would be ranked evenly.

1.4 Data Dictionary

Sex	Gender of the Contestant
Team	The country the contestant is competing for
Sport	Sport category (Judo, Swimming, Badminton, etc.)
Event	Specific event under the sport (Judo Men's Heavyweight, 400m Men's Freestyle, Men's Badminton Singles)
Is_Host_Country	Whether or not the player's country is hosting the Olympics that year
Year	What year that contestant is competing in
Medals	How many medals that contestant's country has won that year

Table 1: Field and their Description

2 Data Engineering

2.1 Analysis

As previously mentioned, our goal is to predict the rankings for the 2028 Olympics in Los Angeles. To achieve this, we require relevant data to train our model and generate accurate predictions. However, since the specifics of the 2028 Olympics have not yet been confirmed, we will need to engineer the dataset ourselves, ensuring it is as realistic as possible to produce reliable results.

The 2028 Olympic Games will feature 50 sports, including new additions and returning sports such as Baseball/Softball, Lacrosse, Cricket, Flag Football, and Squash. However, since the specific contestants for these sports have not yet been confirmed, we will analyze trends from the past few Olympics to estimate the number of participants for each sport. This will help us create a realistic dataset to inform our predictions.

We will be looking through 2016, 2020, and 2024 Olympic Games to analyze the contestants and events.

To analyze the trends and identify patterns for predicting the 2028 Olympics, we will first calculate the average number of contestants per sport and the average number of events per sport for the past three Olympics. This will allow us to better understand how these variables have evolved and provide a more informed basis for making predictions for the 2028 Games.

2016	362.441176
2020	293.086957
2024	259.880000

Table 2: Average Contestants Per Sport

2016	9
2020	7.5
2024	7.18

Table 3: Average Events Per Sport

2.2 Calculation

As shown in Table 2 and Table 3, there has been a decreasing trend in the number of contestants and events per sport in recent Olympic Games. While the 2028 Olympics will feature more sports, we must account for this trend by factoring in fewer contestants and events per sport. This adjustment will help us create a more accurate representation of the 2028 Games, ensuring our predictions align with the current trajectory of the Olympics.

To do this we need to calculate the growth rate to predict the statistics for 2028.

First, we must extract the average number of contestants per sport from 2024 to 2016 and calculate the average growth rate between each game from 2024 to 2016 (a total of 3 games).

$$\text{Growth Rate (2020 - 2024)} = \frac{Average(2024) - Average(2020)}{4} \tag{1}$$

$$\text{Growth Rate (2016 - 2020)} = \frac{Average(2020) - Average(2016)}{4} \tag{2}$$

$$\text{Average Growth Rate (2016 - 2024)} = \frac{\text{Growth Rate (2016 - 2020)} + \text{Growth Rate (2020 - 2024)}}{2} \tag{3}$$

Applying equations (1), (2), (3) we are able to get the average contestant growth rate per game and the event growth rate per game.

Contestant Growth Rate Per Game	-12.82014705882353
Event Growth Rate Per Game	-0.22750000000000004
Projected Contestants for 2028	247.05985294117647
Projected Events for 2028	6.725

Table 4: Calculations for 2028

Looking back at Table 2 and Table 3, our projected calculations for 2028 follows the decreasing trend and gives us appropriate values.

We need to round the values for the predictions to whole numbers. Thus, this leaves us with:

Contestants_per_sport = 247

Events_per_sport = 7

We'll use these values to calculate how many contestants per event and how many total new contestants we'll have compared to 2024.

$$\text{contestants\_per\_event} = \frac{\text{Contestants\_per\_sport}}{\text{Events\_per\_sport}} = 35 \tag{4}$$

$$\text{total\_new\_contestants} = (\text{3 new sports coming in 2028}) \times (\text{Contestants\_per\_sport}) = 741 \tag{5}$$

Therefore, we conclude that there will be approximately 741 more contestants in 2028 compared to 2024. The next task will be to incorporating 741 new contestants into the dataset.

2.3 2028 Data Generation

Now that we have all the necessary information we will need to generate the data for LA2028.

Given the 741 additional participants in the 2028 Olympics, it would not be realistic to distribute them evenly across countries, especially considering the vast differences in representation among nations. As shown in Figure 2, countries like the United States, which has had 8,535 Olympic contestants from 1896 to 2024, have far more representation compared to smaller or less populated nations like Cabo Verde, which has only 24 contestants.

To account for this, we need to add contestants in proportion to each country's historical representation in the Olympic Games. This approach will ensure a more realistic distribution, reflecting each nation's past participation rates and adjusting for the expected number of additional athletes in the 2028 Games.

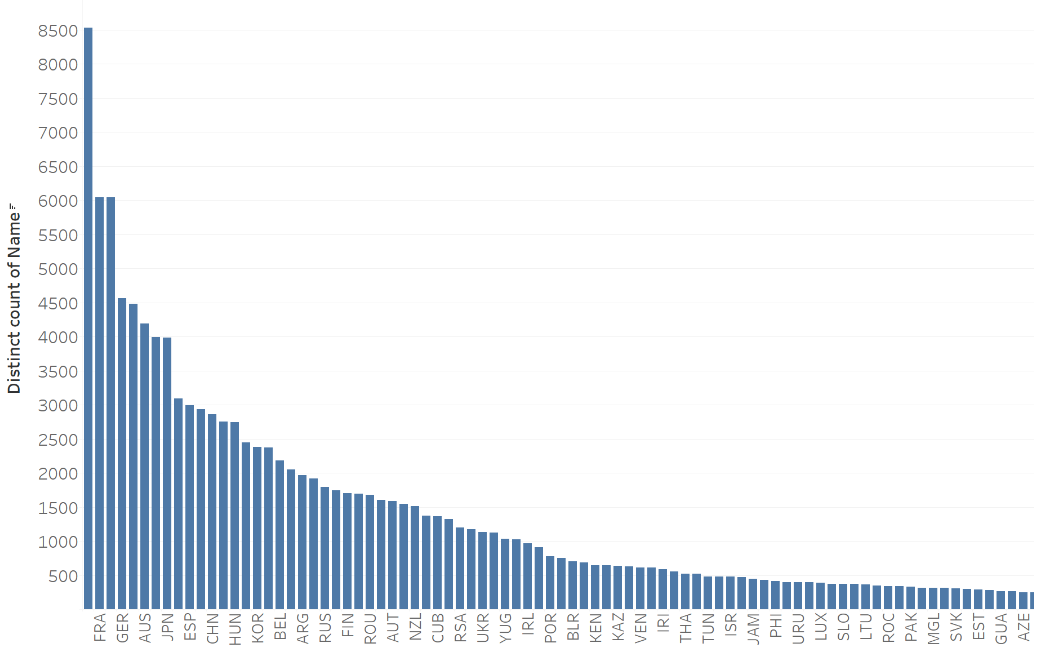

Figure 2: Total number of contestants representing each country from 1896 - 2024

We will first normalize the Country representation from 2024 as we are interested in analyzing the relative frequency of occurrences instead of the absolute count of each Country.

United States	0.065723
France	0.061644
Australia	0.049946
Germany	0.049792
Italy	0.046945

Table 5: 5 most represented countries in 2024

By assigning the majority of the new participants based on the proportional values from the 2024 Olympic Games, we can effectively incorporate the 741 new participants into the dataset. This ensures that the distribution of participants mirrors the historical patterns observed in previous Olympics.

Once this new data is added, we can proceed with the following data manipulation steps:

- **Sex Field Distribution**: We will perform a 50/50 split of the new data for the ”Sex” field, assigning half to male and half to female to maintain gender balance.
- **Host Country Indicator**: We will set the $Is\_Host\_Country$ field to 1 for Team USA, marking them as the host country for the 2028 Olympics

With these steps completed, we will have a more realistic and structured dataset ready for analysis and predictions.

3 Model Training

3.1 Model Determination

First we need to figure out what model to use before we train it. Looking at Figure 3 on the page below, we can see that the distribution is skewed with the United States holding the most Olympic medals while other less represented countries hardly have won any medals.

Next, let us analyze the relationship between the features and what we’re predicting on, which is the total number of medals to be won. We want to see if there is a linear relationship between the features and the total. Doing so will give us a good idea on what models we can/can’t use.

To do this we will use Pearson’s correlation coefficient to construct a correlation matrix, Figure 4, for visualization.

Equation for Pearson's Correlation Coefficient:

$$coeff = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

The Pearson correlation coefficient gives us a value between -1 and 1.

- 1 means we have a perfect positive linear relationship
- -1 means there's a perfect negative linear relationship
- 0 means there's no linear relationship

The closer the coefficient is to 1 or -1, the stronger the linear relationship between the two variables. A coefficient closer to 0 indicates little to no linear relationship.

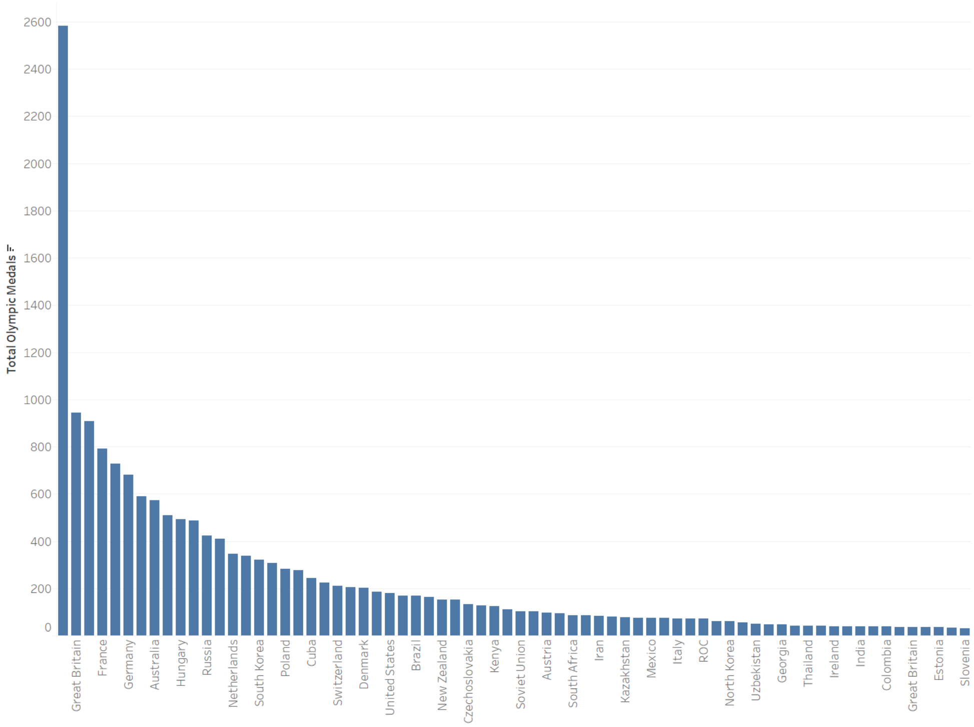

Figure 3: Total medals won by each country from 1896 - 2024

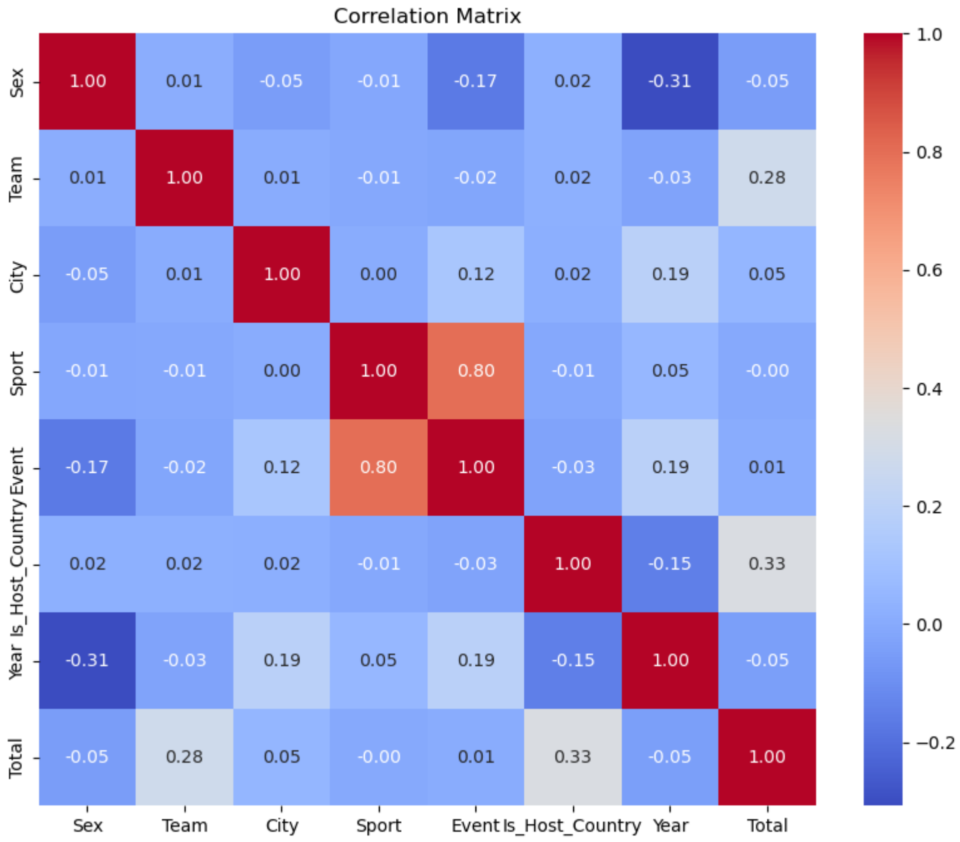

Figure 4: Correlation Matrix between features and the Total

- Looking at the correlation matrix in Figure 4, we observe that:
  - The diagonal values are all 1, which is expected, as each feature is perfectly correlated with itself.
  - The bottom row and rightmost column, most features exhibit very small correlation coefficients with the *Total* variable.
- This suggests there is no linear relationship between these features and *Total*, indicating that we need a model capable of handling non-linear data.
- For this reason, we have chosen the **Random Forest** model, which is well-suited for non-linear data. [4] A Random Forest uses **decision trees**, which are objects that divide data based on specific feature thresholds.
  - When the data is passed through these decision trees, a series of nodes are activated based on comparisons of the data with these thresholds.
  - The decision boundaries created by these trees are non-linear, allowing the model to learn trends in the dataset more effectively.

[4] Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. TEST, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

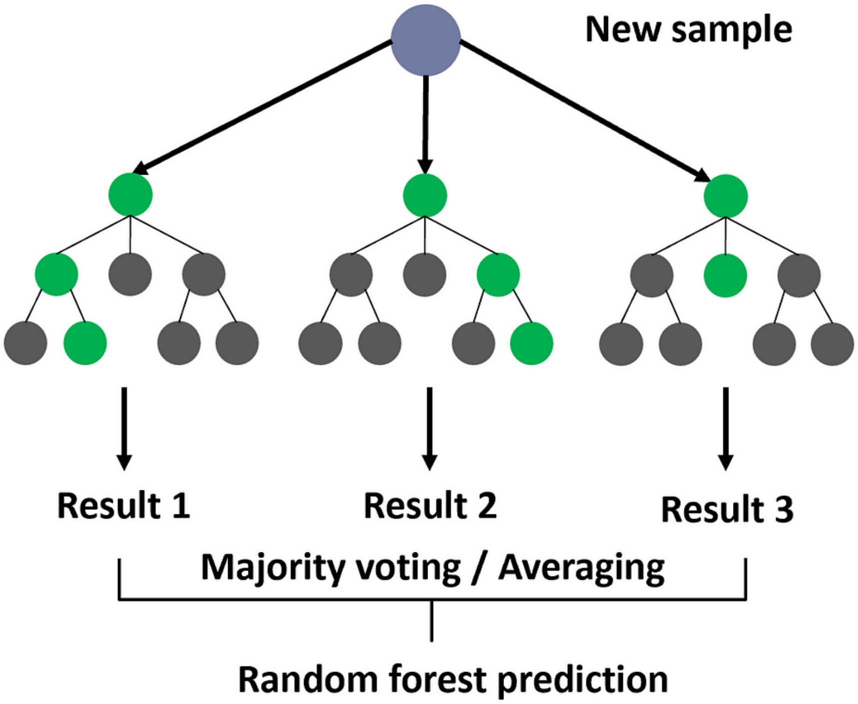

Figure 5: Simplied Random Forest Diagram

- A single decision tree, however, would likely overfit the data, especially in non-linear situations. This happens because:
  - A decision tree uses a greedy algorithm, selecting the feature threshold at each node that maximizes information gain.
  - This process often leads to a model that is too specific to the training data and doesn't generalize well to unseen data, as its decision boundaries are shaped by the specific data used for training.
- The advantage of the Random Forest is that it uses multiple decision trees, each making its own decisions based on a unique subset of features and data.
  - Figure 5 [5] provides a simplified illustration of a Random Forest.
  - By using many decision trees, each introducing random decision boundaries, the final prediction is based on averaging the outputs of all the trees.
  - This approach improves generalization and results in a more robust and accurate prediction.

[5] Yehoshua, Dr. R. (2023, July 17). Random forests. Medium. https://medium.com/@roiyeho/random-forests-98892261dc49

3.2 Train/Test Split

We reach a important crossroad on choosing how to split our data into a dataset for training and one for testing. We want to train our data based on scenarios that replicate the real world. Thus a time based split seems like the more realistic simulation.

When training our model we would train it on 1896-2008 data (70% of the total data) and the remaining 2008-2024 data will be used for testing. If we were to use a random split this means we're training our data on 1896-2024 data and also tested on 1896-2024, thus during the training phase it's likely that the model has already learned the trends relative to that year so it will perform well on testing. However when we deploy the model on 2028 data which was not used in training, it will perform a lot worse.

However, as we will see later there will also be downsides to using time-based split. For now let us train our model using this method.

3.3 Model 1

After running the data through a Random Forest Regressor, we get floating value predictions thus we round them to whole numbers.

Getting the accuracy between the testing data and the rounded predictions we have: 0.20530002558853633

Roughly 20% accuracy. With such terrible performance we can't use this model on the 2028 data. Why is it performing so poorly?

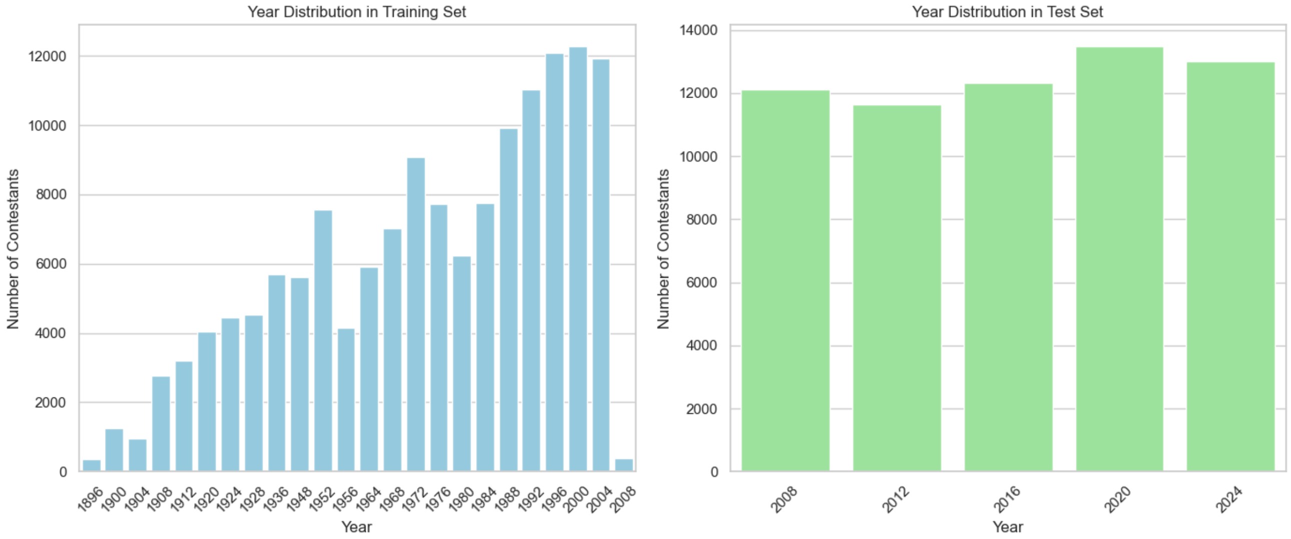

Figure 6: Contestants Per Year

Visualizing the data we can see the reason. The data is distributed differently between the training data and the testing data, thus the poor performance on the testing data.

The average number of contestants in the Olympic Games before 2008 is 6327.086956521739.

The average number of contestants in the Olympic Games after 2008 is 12580.6

Thus the testing data has nearly double the amount of contestants per game compared to the training data.

This brings us back to when we were deciding on the train/test split. Since the distribution of the data is so drastically different after 2008, our cutoff year, we can't get the model to perform properly.

3.4 Model 2

Let us choose our train/test split again, this time randomly splitting 70% for training and the remaining 30% for testing.

After training the data we make our predictions and get a accuracy of **0.999536207778915**

Thus, we are basically getting perfect predictions. This is too good to be true and thus we suspect that we are experiencing data leakage. Data leakage occurs when there is data used as input that gives the exact information on the target.

4 Interpreting the output

Now that we have a well-performing model we can implement it on the 2028 data and answer our initial research questions.

Using the model on our 2028 data we get the top 5 teams:

Team	Medal Prediction
United States	126
China	91
Great Britain	65
France	64
Australia	53

Table 6: Top 5 teams anticipated for 2028

More interestingly is the teams that perform much worse or much better than the previous year in 2024.

Team	2024 Medal Total	2028 Medal Prediction	Difference
Chinese Taipei	7	9	+2
Colombia	4	6	+2
Indonesia	3	5	+2
Croatia	7	9	+2
Azerbaijan	7	8	+1

Table 7: Difference in medal totals between 2024 and 2028

Looking at Table 7 we notice that all the teams with the highest difference experience a positive difference. Meaning that in 2028 they are predicted to experience more medal wins compared to 2024. This is likely due to more participants and events thus giving these teams more opportunities to win medals.

5 Areas for improvement

- Looking back, using 2024 as the foundation for building the 2028 predicted dataset was not ideal. This is because:
  - Testing the model on 2028 is **too similar** to 2024, yielding little to no difference in country rankings.
- As a result, many of our research questions cannot be properly answered. For example:
  - There was no country that performed worse than in 2024, as we only added contestants to the 2024 dataset to account for the increased events and sports.
  - We couldn't determine the true impact of the home-field advantage for the United States in 2028, since we obtained the same medal totals as in 2024.
- Looking at Figure 4 on page 8, we can observe that the feature `Is_Host_Country` has the highest correlation with `Total` with a coefficient at 0.33. Since the 2028 Olympics will be hosted in LA, the United States is expected to experience more medal wins. However, we ended up with the same total as 2024.

- A better approach would have been to average the data from 2016, 2020, and 2024 to create the basis for the 2028 dataset, similar to how we handled the new sports coming to 2028. This would introduce more variability compared to 2024 and provide a more realistic prediction for 2028.
- Another factor to consider is the inclusion of contestants in the data. Specifically:
  - If a country has a specific contestant who has won many medals in previous games, and they are competing in 2028, it is likely that the country will continue its trend of medal wins.
  - Conversely, if a top athlete has retired, the country's medal count may drop.
- To enhance the accuracy of our medal rankings, we should incorporate weights based on the number of gold medals won. According to the International Olympic Committee (IOC):
  - Countries are ranked first by the number of gold medals, followed by silver, and then bronze medals in case of a tie.

  Including this ranking system in our predictions will give a more detailed and accurate ranking of the countries.

6 Appendix

6.1 Extracting Datasets

```
# Load the datasets
medal_counts = pd.read_csv('2025_Problem_C_Data/summerOly_medal_counts.csv')
hosts = pd.read_csv('2025_Problem_C_Data/summerOly_hosts.csv')
programs = pd.read_csv('2025_Problem_C_Data/summerOly_programs.csv', encoding='latin1')
athletes = pd.read_csv('2025_Problem_C_Data/summerOly_athletes.csv')
```

6.2 Adding *Is_Host_Country* column

```
df_medal.to_sql(name='medal_counts', con=conn, if_exists='replace', index=False)
df_hosts.to_sql(name='hosts', con=conn, if_exists='replace', index=False)

cursor = conn.cursor()

# Add a column to indicate if the country was the host
try:
    add_column_query = "ALTER TABLE medal_counts ADD COLUMN Is_Host_Country INTEGER DEFAULT 0;"
    cursor.execute(add_column_query)
    conn.commit()
except sqlite3.OperationalError:
    # Ignore if the column already exists
    print("Column 'Is_Host_Country' already exists.")

#Query to update the Is_Host_Country column
update_query = """
UPDATE medal_counts
SET Is_Host_Country = 1
WHERE EXISTS (
    SELECT 1
    FROM hosts
    WHERE hosts.Year = medal_counts.Year
    AND TRIM(SUBSTR(hosts.Host, INSTR(hosts.Host, ',') + 2)) = medal_counts.NOC
);
"""
```

6.3 Joining Dataframes

```
df_athletes = pd.DataFrame(athletes)

conn = sqlite3.connect(db_file)

df_athletes.to_sql(name=’athletes’, con=conn, if_exists=’replace’, index=False)


join_query = ”””
SELECT * FROM athletes
JOIN medal_counts ON athletes.Year = medal_counts.Year
AND athletes.Team = medal_counts.NOC
”””
```

6.4 Determining feature columns and target

```
X = df_joined.drop([’Name’,’Medal’, ’Rank’, ’Gold’, ’Silver’, ’Bronze’, ’Total’, ’Medal_Earned’], axis=1)
y = df_joined[’Total’]
```

6.5 Correlation Matrix

```

corr_df = X
corr_df[’Total’] = y
correlation_matrix = corr_df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap=’coolwarm’, fmt=”.2f”)
plt.title(”Correlation Matrix”)
plt.show()
```

6.6 Training, Testing, and Predicting with our Model

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
sc = MinMaxScaler()
```

```
X_train_scaled = sc.fit_transform(X_train)
X_test_scaled = sc.transform(X_test)
rf = RandomForestRegressor(random_state=42, n_estimators=100)
rf.fit(X_train_scaled, y_train)
y_pred = rf.predict(X_test_scaled)
y_pred_rounded = np.round(y_pred)
accuracy_score(y_test, y_pred_rounded)
```

6.7 Determining number of contestants and events

```
recent_years = [2016, 2020, 2024]

df_recent = X[X[’Year’].isin(recent_years)]

# Group by Year and Sport, then count the number of contestants (rows)
grouped = df_recent.groupby([’Year’, ’Sport’]).size().reset_index(name=’Contestant_Count’)

# Calculate the average number of contestants per sport for each year
average_per_sport = grouped.groupby(’Year’)[’Contestant_Count’].mean().reset_index(name=’Average_Contestants_Per_Sport’)

unique_events = df_recent[[’Year’, ’Sport’, ’Event’]].drop_duplicates()
events_per_sport = unique_events.groupby([’Year’, ’Sport’]).size().reset_index(name=’Event_Count’)
average_events_per_sport = events_per_sport.groupby(’Year’)[’Event_Count’].mean().reset_index(name=’Average_Events_Per_Sport’)

#Calculate the growth rate from 2016 to 2020 Olympics contestants
growth_rate_2016_2020 = (average_per_sport.loc[average_per_sport[”Year”] == 2020, ”Average_Contestants_Per_Sport”].values[0] -
                         average_per_sport.loc[average_per_sport[”Year”] == 2016, ”Average_Contestants_Per_Sport”].values[0]) / 4

#Calculating the growth rate from 2020 to 2024 contestants
growth_rate_2020_2024 = (average_per_sport.loc[average_per_sport[”Year”] == 2024, ”Average_Contestants_Per_Sport”].values[0] -
                                   average_per_sport.loc[average_per_sport[”Year”] == 2020, ”Average_Contestants_Per_Sport”].values[0]) / 4

# Calculating the overall average growth rate
contestant_growth_rate = (growth_rate_2016_2020 + growth_rate_2020_2024) / 2

avg_2016 = average_events_per_sport.loc[average_events_per_sport[”Year”] == 2016, ”Average_Events_Per_Sport”].values[0]
avg_2020 = average_events_per_sport.loc[average_events_per_sport[”Year”] == 2020, ”Average_Events_Per_Sport”].values[0]
avg_2024 = average_events_per_sport.loc[average_events_per_sport[”Year”] == 2024, ”Average_Events_Per_Sport”].values[0]

# Growth rate from 2016 to 2020 for events
growth_rate_2016_2020 = (avg_2020 - avg_2016) / 4

# Growth rate from 2020 to 2024
growth_rate_2020_2024 = (avg_2024 - avg_2020) / 4

# Average growth rate per game
event_growth_rate = (growth_rate_2016_2020 + growth_rate_2020_2024) / 2
```

6.8 Allocating new data by distribution country representation

```
 contestants_per_event = contestants_per_sport // events_per_sport
total_new_contestants = len(new_sports) * contestants_per_sport # Total contestants in the new sports



#Create the team assignments
team_assignments = []
for team, proportion in team_counts.items():
    num_teams_for_new_sports = int(proportion * total_new_contestants)
    team_assignments.extend([team] * num_teams_for_new_sports)

index = 0  # Initialize index for team_assignments
for sport in new_sports:
    for event_num in range(events_per_sport):
        event_id = event_start + event_num  # Increment the event number for each sport
        for contestant_id in range(1, contestants_per_event + 1):

            sex = 1 if index % 2 == 0 else 0
            Is_Host = 1 if team_assignments[index] == 141 else 0
            new_rows.append({
                'Sex': sex,
                'Team': team_assignments[index],  # Assign team from shuffled list
                'Sport': sport,
                'Event': event_id,
                'Is_Host_Country': 0,
                'Year': 2028
            })
            index += 1
        # Increment event_start for next sport
        event_start = event_id + 1
```

7 References

[1] Abraham, H. M., & Young, D. C. (2025, January 21). Olympic Games. Encyclopædia Britannica. https://www.britannica.com/sports/Olympic-Games

[2] Paris Olympic Games. (n.d.-b). https://www.olympics.com/ioc/news/around-5-billion-people-84-per-cent-of-the-potential-global-audience-followed-the-olympic-games-paris-2024

[3] Voices, E. (2024, July 28). Summer Olympics: 128 years of history in 5 charts. https://earthsky.org/human-world/summer-olympics-128-years-history-5-charts/

[4] Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. TEST, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

[5] Yehoshua, Dr. R. (2023, July 17). Random forests. Medium. https://medium.com/@roiyeho/random-forests-98892261dc49