Predicting Olympic Medal Success: A Data-Driven Country-Level Analysis

Group Name: Group A

Names:
Maxim Lisiansky (206529018)
Noam Ben Moshe (318962693)
Roy Kremer (207577099)
Romi Richter (212876551)
Avraham Elbaz (209885359)

June 3, 2025

Title

Predicting Olympic Medal Success: A Data Driven Country Level Analysis

Abstract

In this project, we tried to figure out what makes some countries do better than others in the Olympics. We used data from the 1960, including things like GDP, population size, whether a country hosted the Olympics, and a country's regime status. We combined data from different sources, cleaned it up, and turned it into a format we could use to build prediction models.

After preparing the data we used four different models to predict how many medals each country might win: Poisson Regression, Linear Regression, Random Forest, and XGBoost. We looked at both country level averages and specific years to see what patterns we could find.

We found that GDP and population were the strongest predictors for winning medals in the Summer Olympics. The more a country participated, the better it did too. Hosting the Olympics and political regime type didn't really seem to matter once we included the economic data. Winter Olympics were harder to predict, but participation still made a difference there.

Overall, our best performing models were Random Forest for the Summer Games and XGBoost for the Winter Games (on country level data). Our project shows that data science can be a useful way to understand what drives success in international sports competitions, especially when we have access to long term, country level data.

Related Work

Here are the main papers we looked at that helped us understand what affects Olympic success:

- **Bernard & Busse (2004):**

  Found that both GDP and population are strong predictors of Olympic medal counts.

  Hosting the Olympics and being part of the former Soviet bloc also gave countries a boost.

  This paper helped us choose our main features (GDP, population, host, regime).

- **Rewilak (2021):**

  Showed that GDP per capita isn't always a good predictor once you control for country specific traits.

  Population and hosting were more consistent factors.

  This made us consider that GDP might not be as important as we first thought.

- **Hoffmann, Ramasamy & Ging (2002):**

  Focused on ASEAN countries and why they didn't do well in the 2000 Olympics.

  Found that low GNP and small populations were the main reasons.

  Also said that getting richer isn't enough, you need real support for sports.

  Made us think about missing variables that might not be in our dataset.

- **Xun Bian:**

  Used linear and Cobb Douglas models with Olympic data from 1988 to 2000.

  Found that GDP, population, host advantage, and socialist background all matter.

  Talked about diminishing returns—more money/population doesn't always mean more medals.

- **Winfree & Fort (Team Payroll vs. Performance):**

  Looked at North American sports leagues like the NBA and NFL.

  Found that teams with higher payrolls usually perform better, especially where salary caps are looser.

  Not directly about the Olympics, but showed how money can lead to success in sports generally.

Data

We used several datasets in this project, which we combined and cleaned to build our final dataset for modeling:

Sources:

- Athlete-level Olympic data from Kaggle's athlete_events dataset.
- GDP and population data from the World Bank.
- Hosting info and event metadata from olympic_hosts.csv.
- Regime data (communist countries) was created manually using Wikipedia.

Size and Structure:

- Athlete data: ∼69,000 rows, 15 columns (one row per athlete event year).
- GDP and population: Country level data from 1960–2020.
- Final dataset: Aggregated to country year level with features like population, GDP, medals, regime, and hosting status.

Preprocessing Steps:

- Cleaned and standardized country names (e.g., "USA" vs "United States").
- Dropped rows where GDP or population were missing.
- Aggregated medal counts per country and year.
- Created binary variables: `host` = 1 if country hosted that year; `regime` = 1 if communist that year.
- Log-transformed GDP and population to reduce skew.
- Standardized (z-score) numerical variables.
- Filtered to only include countries with complete data across required features.

Challenges:

- A lot of inconsistencies in country names (especially over time).
- Missing values for GDP/population caused some countries or years to be dropped.
- Defining the "regime" variable was tricky and required external manual sources.
- Winter Olympic data was harder to model due to smaller and more uneven participation.

Overall, we ended up with a clean dataset that we used to build regression models and analyze which features best predict Olympic medal success.

Methodology

After preparing and cleaning the dataset, we built a modeling pipeline to predict Olympic medal counts using machine learning and regression techniques. Here's how we approached the problem:

Main Idea: We wanted to see how well we could predict medal counts for each country using features like population, GDP, whether a country hosted the Olympics, and if it had a communist regime background.

Modeling Approach: We treated this as a regression problem since medal counts are numeric. We trained four different models and compared their performance:

- **Poisson Regression** – because medal counts are non negative integers.
- **Linear Regression** – a baseline to see how a simple model performs.
- **Random Forest Regressor** – to capture non linear patterns and interactions.
- **XGBoost Regressor** – for a more advanced gradient boosting approach.

Feature Choices: We included both raw and engineered features, like:

- Average GDP and population (log transformed + normalized)
- Whether the country hosted the Olympics (`host`)
- Whether it had a communist regime (`regime`)
- Number of Games attended — to capture historical Olympic involvement

Features were selected based on what prior research suggested matters most, and what improved model performance in practice.

Pipeline & Automation: Used `scikit-learn` pipelines and column transformers to combine all steps (like scaling, encoding, and model fitting) into one reproducible workflow. This made it easy to swap models or test on Summer vs. Winter Olympics.

Consistency & Reproducibility: Fixed random seeds for all models and splits. Kept train/test splits consistent across all experiments.

The methodology let us test both simple and complex models fairly while keeping everything reproducible and easy to update for future work.

Results

We tested our models separately on Summer and Winter Olympic data. Since this was a regression task, we used R^2 (coefficient of determination) as our main metric to see how well the models predicted medal counts. Here's what we found:

Summer Olympics:

- The **Random Forest Regressor** gave the best results with an R^2 of **0.17** on country level data.

- GDP and population were the most important predictors, especially after we log transformed and normalized them.
- The number of Games attended also helped improve predictions, showing that countries with more experience tend to do better.
- Surprisingly, hosting the Olympics and political regime didn't add much to the model once GDP and population were included.

Winter Olympics:

- The results were weaker overall, and most models struggled to make accurate predictions.
- The best model here was **XGBoost**, which got an R^2 of **0.685** on the country level dataset, but this performance wasn't consistent across all tests.
- The Winter Games are harder to predict, probably because fewer countries compete and success is more concentrated.

Model Comparison Summary:

Model	Olympics Type	Best R^2	Top Predictors
Random Forest	Summer	0.17	GDP, population, games attended
Poisson Regression	Summer	0.146	GDP, population
Linear Regression	Summer	0.122	Similar trends, but weaker results
XGBoost	Winter	0.685*	Games attended, GDP, population

Table 1: Model Comparison Summary

*Only on one version of the Winter dataset; not consistent across others.

Feature Importance: In most models, **GDP** and **population** had the highest weights or importances. **Host** and **regime** were often included but didn't end up being useful once other features were present.

These results support what we saw in related research: economic and demographic strength are the most reliable predictors for Olympic success, especially in the Summer Games.

Conclusion

In this project, we tried to figure out which country level factors are the best at predicting Olympic medal success. We combined data from the Olympics, World Bank, and other sources, cleaned it up, and used different regression models to test how well we could predict medal counts.

We found that the most important features were **GDP**, **population**, and **number of Games attended**. These were consistently the top predictors in both the Summer and

Winter Olympics, although the models worked better for the Summer Games. Hosting the Olympics and being a communist regime didn't make much of a difference once the economic and demographic features were included.

Random Forests performed best for the Summer Olympics, and XGBoost did okay on the Winter dataset, but results there were less stable. Overall, ensemble models did better than simple ones, but even the best models still had room for improvement.

This showed us that while money and population size definitely help, they're not the only factors. Things like investment in sports, culture, or specific athletic programs probably matter too—but we didn't have data for that in this project.

If we were to continue this research, we'd try adding more detailed features like how much each country spends on sports, the number of athletes sent per event, or even past medal trends for each sport. That might help explain more of the differences between countries.

In the end, our project helped us understand how data science can be used to explore global sports performance, and how combining multiple sources and models can lead to interesting and useful insights.