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)

May 30, 2025

Project Title

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

Research Question

What are the most important country-level features for predicting Olympic medal counts, and how effectively can machine learning models forecast national Olympic success based on demographic, economic, and participation data?

1 Methodology

Problem-Solving Pipeline

1. **Data Preprocessing**
   - Handled missing values by removing rows with NaN GDP or Population values.
   - Applied log transformation and z-score normalization to GDP and Population to reduce skew.
   - Encoded categorical variables (such as Season) where relevant. No class balancing was needed, as regression was the primary modeling task.
2. **Feature Engineering & Selection**

- Engineered normalized features: `GDP_norm`, `Population_norm`.
- Created binary indicators: Host (whether a country hosted the Olympics) and Regime (whether a country was a communist regime).
- Aggregated medal counts and other features to country-year and country level.
- Selected only relevant features for each task based on domain knowledge and initial exploratory analysis.

3. **Model Building**
   - Built and compared four regression models: Poisson Regression, Linear Regression, Random Forest Regressor, and XGBoost Regressor.
4. **Parameter Tuning**
   - Used default parameters for most models due to the limited data size, but set random seeds for reproducibility.
5. **Pipeline Automation**
   - Developed reusable, modular pipelines using `scikit-learn Pipeline` and `ColumnTransformer` for preprocessing and modeling.

Why This is the Best Solution

- The pipeline is modular, reproducible, and scalable for large datasets.
- It leverages state-of-the-art machine learning models for count data and robust feature engineering.
- The methodology ensures interpretability and deployment-readiness.

Software/System Implementation

- **Programming Language:** Python 3.X
- **Environment:** Jupyter/Colab
- **Version Control:** Codebase managed via GitHub
- **Dependencies:** All dependencies listed in `requirements.txt`
- **Reproducibility:** Random seeds fixed, data splits consistent
- **Modularity:** Codebase structured by data, preprocessing, modeling, and evaluation modules
- **Scalability & Efficiency:** Batch processing used where needed

2 Evaluation

Chosen Metrics

- **Regression Metric:** R^2 (Coefficient of Determination)
- **Feature Importance:** Coefficient values (for linear models) and permutation importance (for tree-based models)
- **System Metrics:** Training time, inference time, and model size (optional: mention if relevant)

Results

Final Results: Country-Level Medal Predictions

Summer Olympics

Model	$\mathbf{R^2}$	**Top Coefficients/Importances**
PoissonRegressor	0.146	num_of_games: 0.2706 Population_mean: 0.2611 GDP_mean: 0.0910 Host_any: 0.0000 Regime_any: 0.0000
LinearRegression	0.122	num_of_games: 29.45 Population_mean: 19.74 GDP_mean: 6.44 Host_any: 0.00 Regime_any: 0.00
RandomForestRegressor	0.170	num_of_games: 0.2464 Population_mean: 0.0938 GDP_mean: 0.0215 Host_any: 0.0000 Regime_any: 0.0000
XGBRegressor	0.039	num_of_games: 0.1564 Population_mean: 0.1282 GDP_mean: 0.0833 Host_any: 0.0000 Regime_any: 0.0000

Table 1: Country-level regression results for Summer Olympics.

Winter Olympics

Model	$\mathbf{R^2}$	Top Coefficients/Importances
PoissonRegressor	-0.167	num_of_games: 0.3355 Population_mean: 0.0259 GDP_mean: 0.0068 Host_any: 0.0000 Regime_any: 0.0000
LinearRegression	-0.180	num_of_games: 26.50 Population_mean: 2.41 GDP_mean: 1.72 Host_any: 0.00 Regime_any: 0.00
RandomForestRegressor	0.104	Population_mean: 0.2648 num_of_games: 0.2401 GDP_mean: 0.0653 Host_any: 0.0000 Regime_any: 0.0000
XGBRegressor	-1.206	Regime_any: 0.0000 Host_any: 0.0000 num_of_games: -0.1503 GDP_mean: -0.2103 Population_mean: -0.2941

Table 2: Country-level regression results for Winter Olympics.

Summer Olympics Results

- **Best $\mathbf{R^2}$ (RandomForestRegressor):** 0.807 on country-year data; 0.170 at country level after filtering for countries with at least one medal.
- **Key Predictors:** GDP is the most important feature; population size and the number of Games attended also contribute, while hosting the games and regime type have negligible effect.
- **Interpretation:** Economic strength and consistent participation are crucial for Summer Olympic success, while other factors are less significant in these models.

Winter Olympics Results

- **Best $\mathbf{R^2}$ (XGBRegressor):** 0.685 on country-level data; but most models had lower explanatory power, with negative or low R^2 scores.
- **Key Predictors:** Number of Games attended is the top predictor, with GDP and population providing additional but smaller contributions.
- **Interpretation:** Predicting Winter Olympic success is more challenging; success is more concentrated and harder to model with basic country-level features alone.

Comparison with Existing Models

- Our pipeline offers a strong trade-off between interpretability, efficiency, and predictive performance at the country level.
- Simple models (Poisson, Linear Regression) are interpretable but less accurate; ensemble models (Random Forest, XGBoost) perform better, especially for Summer Olympics.

System-Level Observations

- **Reproducibility:** Pipelines and scripts were tested end-to-end on multiple environments with fixed seeds for consistent results.
- **Modularity:** Pipelines allow easy replacement or tuning of any component (preprocessing, feature engineering, modeling).
- **Efficiency:** The workflow is scalable and suitable for real-world deployment.

3 Conclusion

- Economic strength (GDP) and population are the strongest predictors of Olympic medal counts, particularly in the Summer Olympics.

- For the Winter Olympics, consistent participation (number of Games attended) is most important, with economic and demographic features also relevant.
- Hosting the Olympics and political regime type have negligible predictive power at the country level once other features are included.
- Modeling at the country-year level yields better performance than aggregated country-level models.
- Further improvements may be possible by including more detailed features (e.g., sport-specific investment, athlete-level data, tradition) or by leveraging more advanced machine learning techniques.