Olympic Insights

Deep Learning

MSCS 612N-816

DeepLearners

Marist College

School of Computer Science and Mathematics

Submitted To:

Dr. Reza Sadeghi

February 13, 2025

Project Report of Olympic Insights

Team Name

Deep Learners

Team Members

1. Elizabeth Herrera Elizabeth.Herrera1@marist.edu (Team Head)
2. Easton Eberwein Easton.Eberwein1@marist.edu (Team Member)
3. Austin Frank Austin.Frank1@marist.edu (Team Member)
4. Luca Gristina Luca.Gristina1@marist.edu (Team Member)

Description of Team Members

1. Elizabeth graduated from Marist in 2023 with a B.S. in cybersecurity. She is now back at Marist to complete her MSCS with a concentration in AI. Elizabeth also works as a function tester for a z/OS security product at IBM.

…….

2. Easton Eberwein graduated from Marist in 2024 with a B.S. in Information Systems and Technology. He is in his final semester of Marist's 5-year M.S.I.S. graduate program to receive a degree in Information Systems Management and Business Analytics. He has also been a member of the Marist men's track and field team for 4 years, being a captain for 3.

…….

3. Austin Frank graduated from Marist in 2024 with a B.S. in Computer Science - Software Development. He is in his final semester of Marist's 5-year graduate program to receive his M.S. in Software Development. Additionally, he is a developer for the z/TPF database and business events.

…….

4. Luca graduated from Marist in 2024 with a B.S. in Computer Science - Software Development. He is currently in his final semester as a part of Marist's 5-year graduate program to receive his M.S. in Software Development.

Table of Contents

Introduction

General Description

The Olympic Games are one of the world's most popular sporting events. Held every four years, competitors worldwide come to the games to achieve personal excellence and win medals for their countries. It's always exciting to see which athletes will come away with medals, but what if we had a way to predict the outcome of Olympic events? In this project, we aim to see how accurately an AI model can predict an athlete's Olympic performance based on age, height, weight, sport, and country.

Research Question

How accurately can an AI model predict an athlete's Olympic performance (e.g. medal likelihood or placement) based on age, height, weight, sport, and country?

GitHub Repository Address

https://github.com/ElizabethHerrera12/DeepLearners

URL of your dataset

https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. This dataset includes basic bio data on the athletes as well as medal results. The dataset was collected from www.sports-reference.com in May 2018 using R code.

Name, definition, and characteristics of features

1. **ID** - Unique number for each athlete
2. **Name** - Athlete's name
3. **Sex** - M or F
4. **Age** - Integer
5. **Height** - In centimeters
6. **Weight** - In kilograms
7. **Team** - Team name
8. **NOC** - National Olympic Committee 3-letter code
9. **Games** - Year and season
10. **Year** - Integer
11. **Season** - Summer or Winter
12. **City** - Host city
13. **Sport** - Sport
14. **Event** - Event
15. **Medal** - Gold, Silver, Bronze, or NA

Review of Related Work

Machine learning and statistical models have been increasingly used to predict Olympic performance, particularly an athlete's likelihood of winning a medal. These studies take different approaches, with some focusing on country-level trends and others analyzing individual athlete characteristics. By comparing these studies, we can better understand their methodologies, results, strengths, and limitations in predicting Olympic success.

Country-level models primarily rely on economic and historical data to forecast Olympic medal counts. Studies using machine learning techniques, such as decision trees, random forests, and multiple linear regression, have identified GDP, population size, and past performance as the most significant predictors of a nation's success. These models achieve high accuracy, with some exceeding 90% in predicting medal counts for upcoming Olympics. A major advantage of these approaches is their ability to analyze long-term trends using extensive datasets spanning multiple Olympic Games. However, they fail to account for real-time adaptability, such as an athlete's current fitness level, injuries, or training improvements, limiting their applicability for predicting individual performance.

In contrast, athlete-centered models attempt to predict an individual's probability of winning a medal by incorporating personal attributes like age, height, weight, and previous results. These models, often built using classification techniques such as random forests and boosting algorithms, offer a more granular perspective on performance prediction. The Player Winning Probability Model (PWPM), for example, achieved an accuracy of 70.22%, demonstrating the potential of athlete-level predictions. However, compared to country-level models, these approaches generally have lower accuracy, possibly due to the challenge of quantifying non-physical factors such as mental resilience, injury history, or in-game decision-making.

Each approach has its strengths and weaknesses. Country-level models provide high-level insights into Olympic success, making them useful for policymakers and sports analysts, but they lack precision in predicting individual outcomes. Athlete-focused models are more tailored to personal performance but require further refinement to improve accuracy. A hybrid approach that combines macroeconomic indicators with individual athlete characteristics could enhance predictive capabilities, offering a more comprehensive model for forecasting Olympic success. Integrating real-time data, such as an athlete's recent training performance and injury reports, may further refine these predictions and provide a more dynamic and adaptable framework for future research.

Future Plan for Olympic Insights

Step 1: Data Collection

Our perspective sources for datasets:

1. Kaggle
2. Official Olympics Website

Step 2: Data Preprocessing

Categorization

Our features could be put into the following:

1. Identification Features (Not Useful for Modeling, But Needed for Tracking)
    a. ID – Unique number for each athlete
    b. Name – Athlete's name
2. Categorical Features (Require Encoding: One-Hot Encoding or Label Encoding)
    a. Sex – M or F (Binary Encoding: 0 for Male, 1 for Female)
    b. Team – Team name (One-Hot Encoding)
    c. NOC – National Olympic Committee 3-letter code (One-Hot Encoding)
    d. Season – Summer or Winter (Binary Encoding)
    e. City – Host city (One-Hot Encoding)
    f. Sport – Sport type (One-Hot Encoding)
    g. Event – Specific event (One-Hot Encoding)
    h. Medal – Gold, Silver, Bronze, or NA (Label Encoding: NA → 0, Bronze → 1, Silver → 2, Gold → 3)
3. Numerical Features (Require Normalization or Standardization)
    a. Age – Integer (Standardization)
    b. Height – In centimeters (Standardization)
    c. Weight – In kilograms (Standardization)
    d. Year – Integer (Could be used as a feature or transformed into Olympic cycles)
4. Composite Feature (Can Be Used to Derive Trends)
    a. Games – Combination of year and season (Extract year for trends or use season separately)

Standardization of Numerical Data

We will apply a standardization, most likely Z-score normalization, to our numerical features

1. Age
2. Height
3. Weight

Continuous Data

Ensure continuous data throughout our dataset and use appropriate statistical measures or predictive modeling for gaps.

Splitting Data

To test the effectiveness of our data preprocessing and model performance, we will split the dataset into training, validation, and test sets. This ensures that our model generalizes well to unseen data.

Initial Split: Training (70%) / Validation (15%) / Test (15%)

We will experiment with values to find the best balance between training size and generalization performance. Our goal is to maximize accuracy while minimizing overfitting.

Step 4: Choosing a Deep Learning Model

1. Artificial Neural Networks (ANNs)
    a. Best for tabular data like medal counts per country.
    b. Fully connected layers will capture relationships between country factors and Olympic success.
2. Recurrent Neural Networks (RNNs) / LSTMs
    a. Used if temporal dependencies are strong (e.g., predicting medals based on past Olympics).
    b. Can analyze sequential patterns in Olympic history.
3. Convolutional Neural Networks (CNNs)
    a. If image data (athlete photos, competition visuals) is included, CNNs will be used for feature extraction.

Step 5: Model Training and Evaluation

Once our dataset is preprocessed and split appropriately, we will proceed with training and evaluating our deep learning model. Our approach includes implementing optimizations such as early stopping and hyperparameter tuning to enhance model performance.

Training the Model

1. **Deep Learning Architecture:**
    a. We will likely use a Neural Network (MLP or CNN), depending on the dataset's structure.
    b. The model will consist of multiple layers with activation functions like ReLU and softmax (for classification).
    c. We will experiment with batch normalization and dropout layers to prevent overfitting.
2. **Training Process:**
    a. We will use backpropagation and gradient descent (Adam optimizer) to optimize model weights.

b. We will monitor validation loss during training to detect overfitting.
c. Batch Size & Epochs will be fine-tuned based on performance.

Hyperparameter Tuning

1. To optimize model performance, we will tune parameters such as:
    a. Learning rate (e.g., 0.001 vs. 0.0001)
    b. Batch size (e.g., 32 vs. 64)
    c. Number of layers & neurons per layer
    d. Dropout rate (e.g., 0.2 vs. 0.5)
2. Techniques used:
    a. Grid Search or Random Search to find the best hyperparameters.
    b. Bayesian Optimization for more advanced tuning.
3. Early Stopping:
    a. We will monitor the validation loss and stop training if it starts increasing, preventing overfitting.

Model Evaluation Metrics:

- Root Mean Squared Error (RMSE) – Measures prediction errors; lower RMSE indicates better accuracy.
- Mean Absolute Error (MAE) – Shows the average error in predictions.
- Confusion Matrix & F1 Score (if predicting medal categories) – Evaluates classification performance.

Comparison & Optimization:

- After initial evaluation, we will compare different models (e.g., CNNs vs. LSTMs vs. Fully Connected Networks) to determine the best-performing architecture.
- Model performance will be validated using the test set and real-world Olympic data trends.

Step 6: Insights and Visualization

Generate visualizations of medal predictions vs. actual outcomes.

Provide insights into which countries are over performing or underperforming based on socio-economic factors.

References

1. https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results
2. https://www.smithsonianmag.com/science-nature/can-statistical-model-accurately-predict-olympic-medal-counts-180949627/
3. https://nycdatascience.com/blog/student-works/olympics-medals-prediction/
4. https://www.ijfmr.com/papers/2024/2/18036.pdf
5. https://ijariie.com/AdminUploadPdf/OLYMPIC_GAME_ANALYSIS_AND_PREDICTION_USING_MACHINE_LEARNING_ijariie24076.pdf?srsltid=AfmBOopdu1a8PNutXlzheNrDUJwh_yTL6DKKoO77nzqo9gDdVGLkLyji
6. https://www.researchgate.net/profile/Nongmeikapam-Thoiba-Singh/publication/378535086_Predicting_Medal_Counts_in_Olympics_Using_Machine_Learning_Algorithms_A_Comparative_Analysis/links/65f5549d1f0aec67e29d3db3/Predicting-Medal-Counts-in-Olympics-Using-Machine-Learning-Algorithms-A-Comparative-Analysis.pdf