Olympic Medal Prediction Project (1976–2008)

Using Python, ML, and Exploratory Data Analysis

Project Overview

This project was built to understand and predict Olympic medal outcomes using machine learning. We worked with on a dataset containing medal records from the Summmer Olympic between **1976 and 2008** . The goal was to uncover trends in medal distribution and use relevant features (like sport, gender, and country) to train a model that could predict whether an athlete might win a **Gold**, **Silver**, or **Bronze** medal.

This project combines **data cleaning**, **exploratory analysis**, **feature engineering**, and **ML model building**, all while keeping the process simple, readable, and beginner-friendly.

Objectives

- Analyze how medals were distributed across different years, countries, and sports.
- Understand the participation and success patterns among male and female athletes.
- Identify which countries and athletes performed consistently well over the years.
- Create visualizations that bring these patterns to life.
- Train a machine learning model to predict the medal type based on given athlete attributes.

Dataset Details

The dataset includes individual medal records, with each row representing an athlete winning a medal in a particular year and sport. Key columns include:

- City: Where the Olympic event took place
- Year: Year of the Olympics (1976 to 2008)
- Sport and Discipline: Main category and sub-category
- Event: Specific competition
- Athlete: Name of the athlete
- Gender: Male or Female
- Country: Country name and code

- Medal: Gold, Silver, or Bronze

Step 1: Data Preparation

Before building any model, we needed to prepare the data properly. Here's what was done:

- **Encoding Fix:** While reading the file, we encountered character encoding issues. These were resolved by specifying 'latin1' encoding when loading the data using pandas.
- **Dealing with Missing Values:** Here dataset was checked for null values, weather it contains or not. Any missing entries in essential columns were either filled appropriately or dropped if necessary.
- **Column Selection:** Columns like Athlete Name, Event Gender, and Country Code were dropped, as they didn't contribute much value for prediction and could increase noise.
- **Encoding Categorical Variables:**
    - Applied **One-Hot Encoding** on Sport, City, and Gender columns.
    - Used **Label Encoding** on the Country column, since it had too many unique values (over 120+), and one-hot encoding would create too many columns.
- **Feature Scaling:** Although not every model requires it, we applied scaling to numerical columns like Year for consistency across features.
- **Target Encoding:**
    - The target variable Medal (Gold, Silver, Bronze) was **label encoded** into numerical values (0, 1, 2) so models could learn from it.
- **Train-Test Split:** The final dataset was split into a **training set and a testing set** using an 80–20 ratio to fairly evaluate model performance later.

Step 2: Exploratory Data Analysis (EDA)

Exploratory analysis was key to understanding the insights and information hidden in the data. Here are the insights we gathered:

- **Top Performing Countries:** A horizontal bar chart was used to show which countries earned the most medals between 1976 and 2008. This gave us a sense of global dominance.

- **Medals Over Time:** A line chart visualized the total number of medals awarded per Olympic year. This helped identify growth patterns and participation trends.
- **Medals by Sport:** We used treemaps and bar charts to compare the number of medals earned in each sport. Some sports (like Athletics and Swimming) dominated the medal count.
- **Gender Participation:** Pie charts and line comparisons showed how male and female athletes fared in terms of total medals and participation rate. It clearly visualized the rise of women in Olympic history.
- **Athlete Highlights:** We also created a summary of athletes who won the most medals, offering a glimpse into individual excellence across sports and years.

Step 3: Machine Learning Model Development

With the data cleaned and explored, we moved on to training several ML models to predict the medal type. The goal was to see whether we could take in an athlete's attributes (like sport, country, gender, and year) and **predict the medal type** they won.

Models Tried:

1. **Logistic Regression:**
    - As a basic baseline model, it helped us understand how well simple linear boundaries worked.
    - **Accuracy:** Around **40%** — not ideal for this multi-class task.
2. **XGBoost Classifier:**
    - A gradient boosting algorithm known for handling structured data well.
    - **Accuracy:** Improved to about **61%**.
3. **CatBoost Classifier:**
    - A model built specifically for categorical data. It gave slightly better results than XGBoost.
    - **Accuracy:** Roughly **63%**.
4. **Random Forest Classifier (Best Performer):**
    - This ensemble model consistently outperformed others.
    - **Accuracy Achieved:** Around **68%**

We also tried hyperparameter tuning on the Random Forest model, but improvements were minor — suggesting the model was already performing close to its potential on this dataset.

Summary

This project took a historical dataset and used basic tools in Python to create a predictive pipeline that included:

- **Data cleaning and transformation**
- **Exploratory data visualizations**
- **Feature engineering**
- **Model training and evaluation**

The final model, a **Random Forest Classifier**, reached about **68% accuracy**, which was a solid result considering the scope of the data and simplicity of the features.

More importantly, this project helped develop confidence in:

- Working with real-world data
- Preparing data for machine learning
- Analyzing trends using visualization
- Applying and comparing multiple ML models