Olympic Prediction

Janelle Anderson, Billy Cartwright, Kristina Castro, Matthew Price, Ricardo Robles

Project Overview and Proposal

Why do some countries outperform others in the Olympics? What factors influence olympic performance at the country and athlete level? What countries and athletes are most likely to have a stronger performance at the next olympics? This proposal seeks to answer these questions. Using the vast dataset from the 2016 olympics, we seek to determine the most influential factors that relate to medal performance of olympic athletes, and build a predictive model that can forecast athlete performance for the next summer olympics.

This analysis could provide multiple investment opportunities related to sports/olympic consumer market sizing, athlete endorsements, sports analytics, and sports betting, to name a few possibilities.

Data Set

Official dataset outlining olympic events, athletes, participating countries, and medal counts.

https://www.kaggle.com/rio2016/olympic-games

Data Cleaning and Analysis

Python will be the main language used for data cleaning and analysis. Several packages will be utilized, including Pandas, NumPy, SciKitLearn, ImbalancedLearn, and DateTime functions.

Database Storage

The data for the project will be stored in a PostgreSQL relational database.

Group Roles

Github master (square) - Janelle

Database (circle)- Ricardo

ETL/technology (X) - Kristina

Machine learning (triangle) - Matt

Final visuals/technology (X) - Billy

Machine Learning Model

The machine learning model will be created with the SciKitLearn and ImbalancedLearn packages. As the project will require over/under sampling due to class imbalances, the approach will rely on either random over/under-sampling, SMOTE, Cluster Centroid Undersampling, or SMOTEENN.

Our goal is to create multiple models to identify the model with the highest prediction score. Models under consideration are simple linear regression, logistic regression (with a binary output of "won a medal" vs. "did not win a medal"), and a decision tree/random forest model. Optimally, we can rank predictions of olympic performance at the athlete level, but also grouping them by country.

Possible Independent Variables (observations at the individual athlete level): height, weight, age, home country GDP per capita, sex (dummy variable), sport (dummy variables) Possible Dependent Variables: total medals won, total gold/silver/bronze medals won, or a "placed" variable (a binary variable outlining if the athlete won any medal vs. not)

Overall accuracy of the model will be the primary success metric. Sensitivity will likely be a priority over precision. While we would like to be able to correctly identify all of the medalists possible (higher precision), making sure the model's predicted medalists are actually winners (higher sensitivity) is a more important consideration.

There will be a lot more observations on non-medalists than medalists, so it could be beneficial to include over/under sampling for our model in order to counteract some of the sampling biases inherent in the dataset.

Dashboard

The group created a Tableau dashboard to present the findings and allow users to sort and filter data based on their required analysis.