# Identifying Olympic Athletes That Used Performance Enhancing Drugs

# Purpose


Build a model that will classify PED use to provide the International Olympic Committee with an expedient tool aiding in the identification of athlete samples to re-test.

- Athletes being retroactively stripped of their ranking and suspended from future events if found guilty of PED use.
- Provide the athletes in events with flagged PED users with an improved rank

# PROCESS

1. Exploratory Data Analysis
2. Data Preprocessing
3. Feature Engineering
4. Modeling
5. Model Evaluation
6. Model iterations (Hyper parameter and Fine tuning)
7. Deployment

# Data

The data used for this project was gathered from:

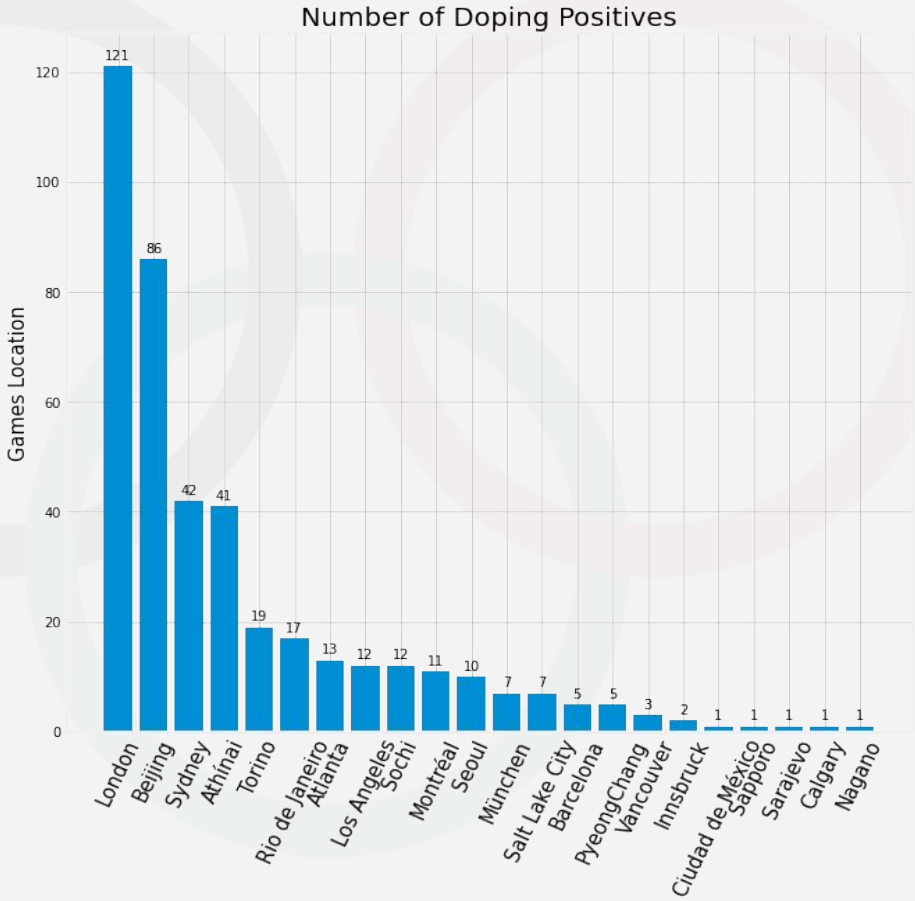
- Olympedia
- Kaggle Datasets
- World Anti-Doping Agency (WADA)
- The Doping List
- Wikipedia

*(2012-2016 data obtained)*

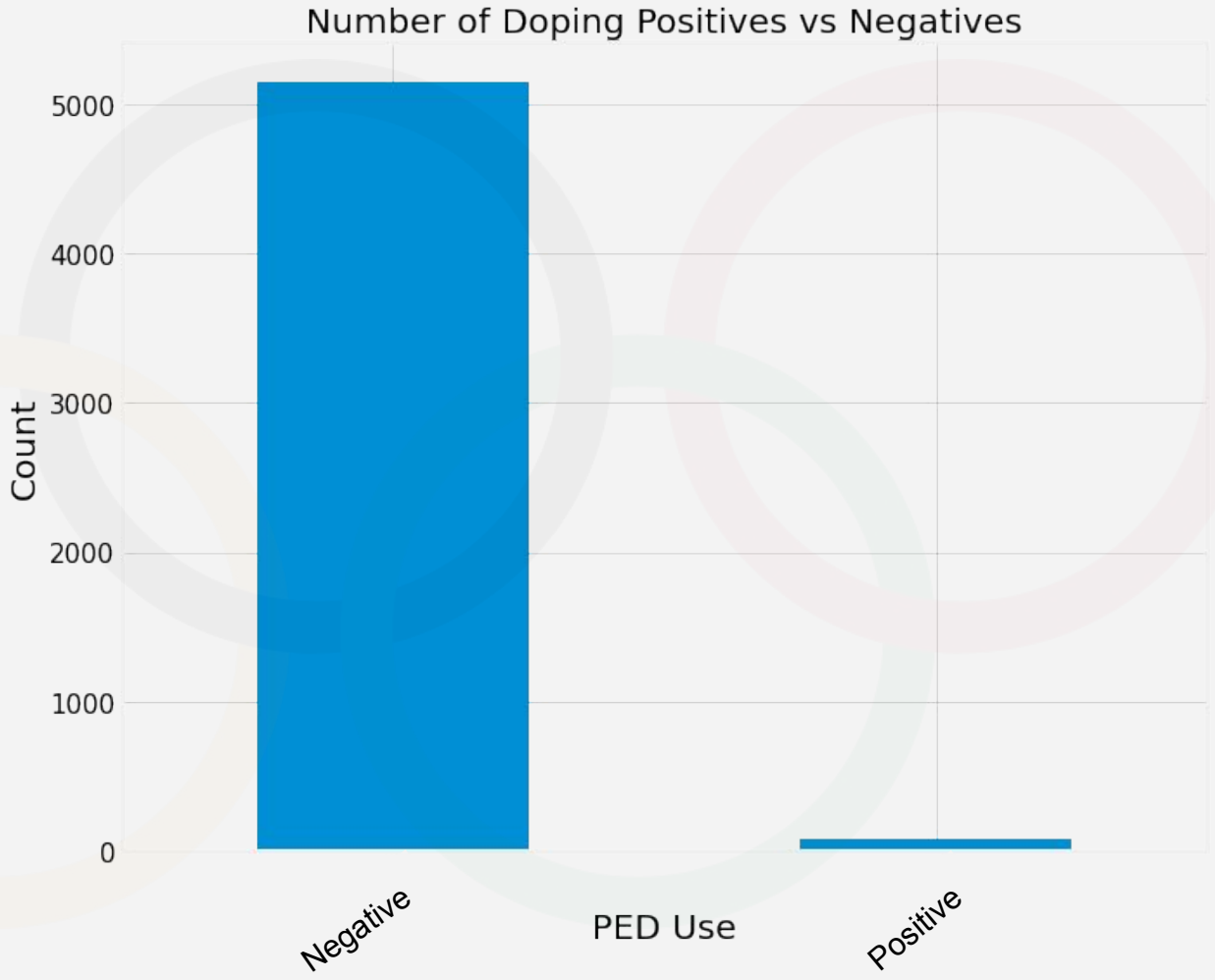
Athlete dataset from kaggle as the base dataframe and combining the Olympedia results from track and field events on athlete names. The ‘flagged’ feature added to indicate PED use and be used as the target variable.

# EDA Findings

**Doping Positives per Olympic Games**



# Limitations

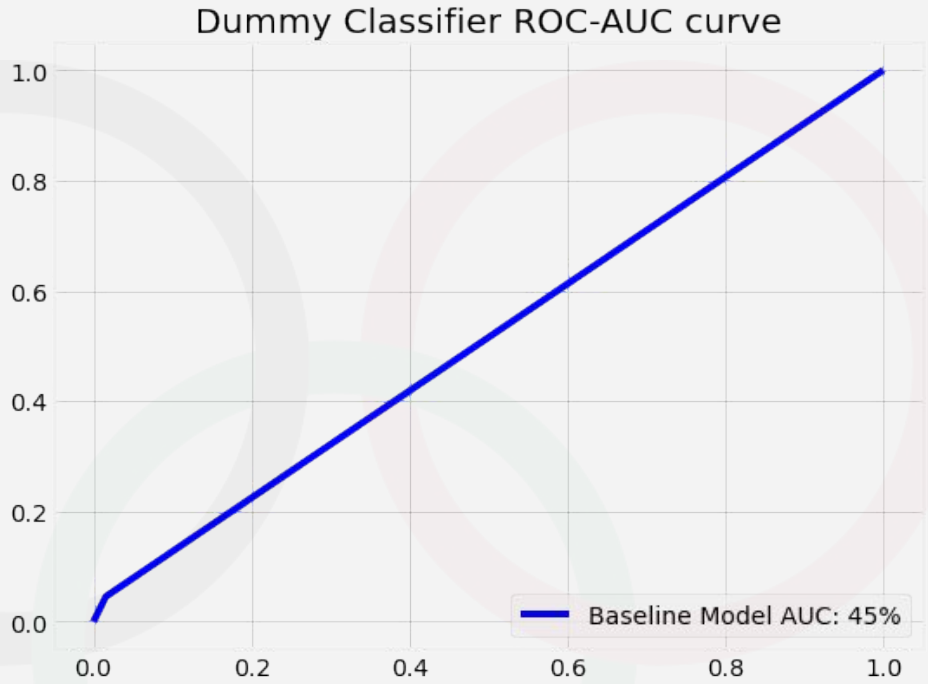


# Baseline Model

## Scikit-Learn Dummy Classifier

- Optimizing for recall to limit False Negatives

The higher the AUC score, the better the model is at distinguishing positive vs negative PED use. The goal is to have the curve as close to the top left corner of the grid as possible (largest area under the curve)




| **Doping Recall (Sensitivity)** | **5%** |
|---|---|
| Precision | .06% |
| Accuracy | 97% |

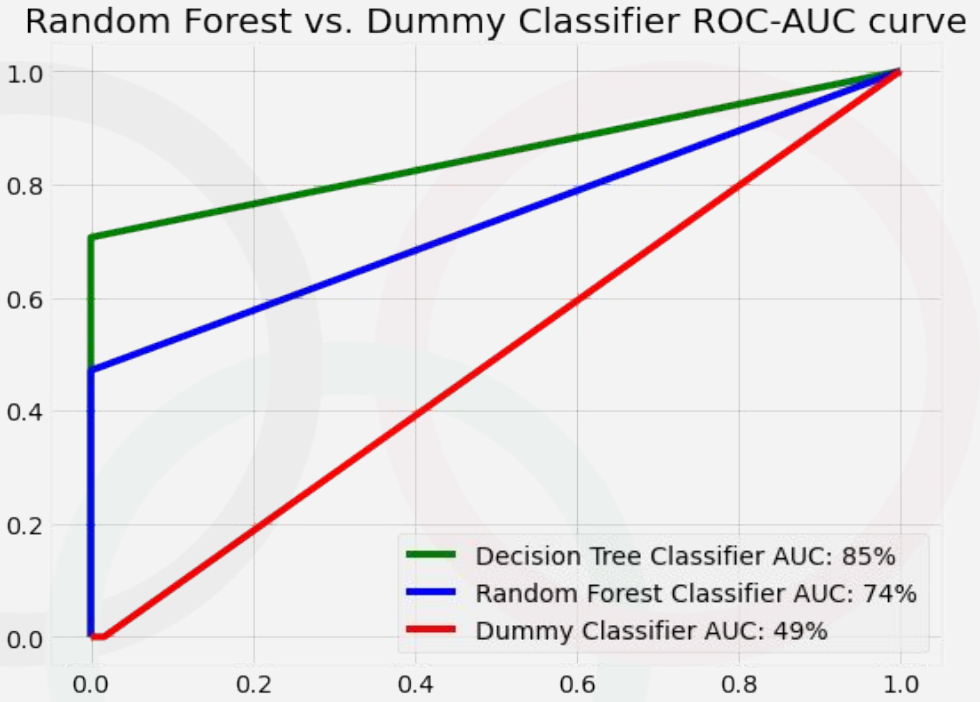
# Current Model

## Decision Tree Classifier

- Optimizing for recall to limit False Negatives

## Parameters:

- Max Depth: 7
- Min_samples_split: .13




| **Doping Recall (Sensitivity)** | **27%** |
|---|---|
| Precision | 97% |
| Accuracy | 99% |

# Next Steps

- Include event results from 2004 and 2008 Summer Olympic Games
- Include all track and field events
- Engineered feature indicating difference in event result from previous year's Olympic Games
- Neural Network classification modeling
- Model evaluation on next Olympic Games event results
- Train model to classify type of PED used

# Thank You

Jason Wong

Email:
jwong853@gmail.com

Github: jwong853