Identifying PED Use Amongst Olympic Athletes

Purpose

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

- Guilty athletes being retroactively stripped of their ranking with ban/suspension.
- Improve clean athlete rankings

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

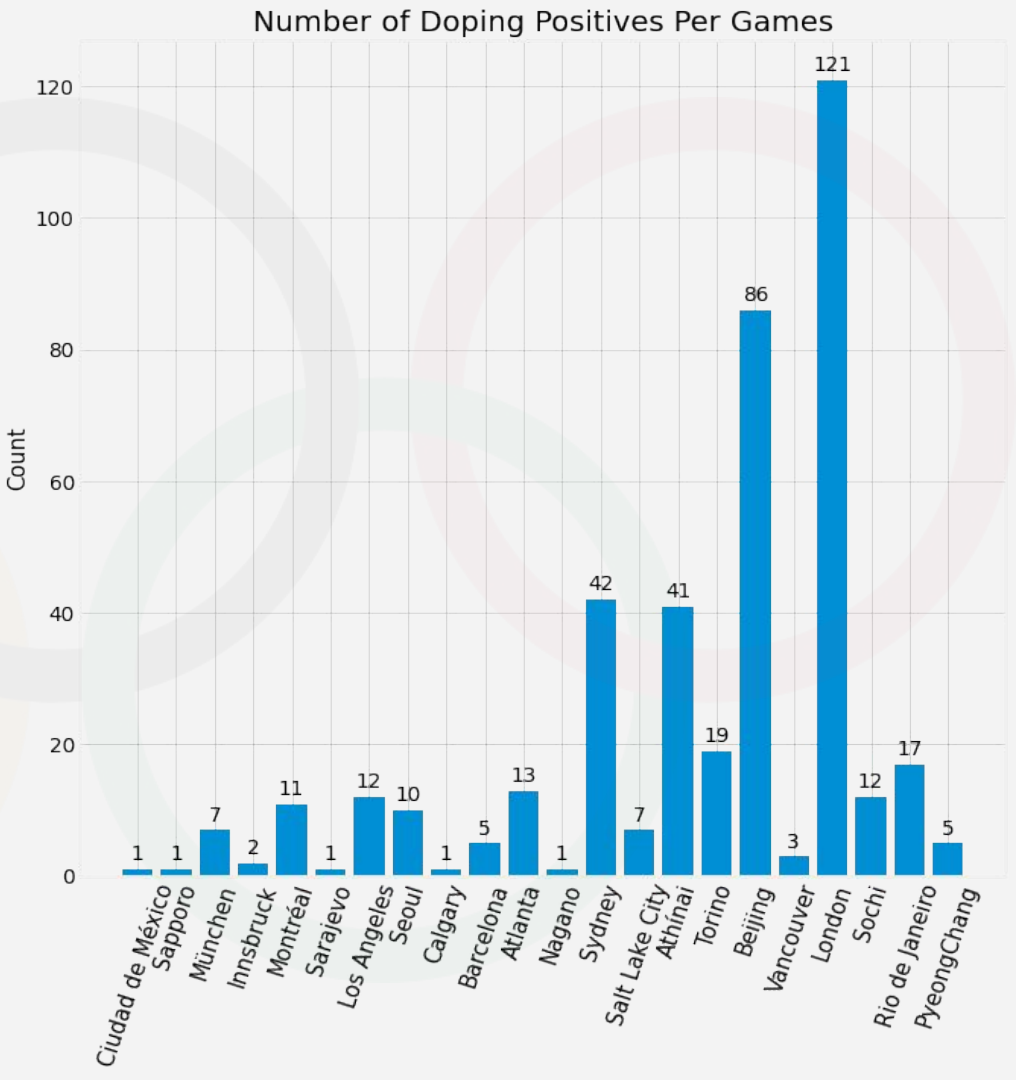
- Olympedia
- Olympic
- Kaggle Datasets
- World Anti-Doping Agency (WADA)
- Wikipedia

(2012-2016 data obtained)

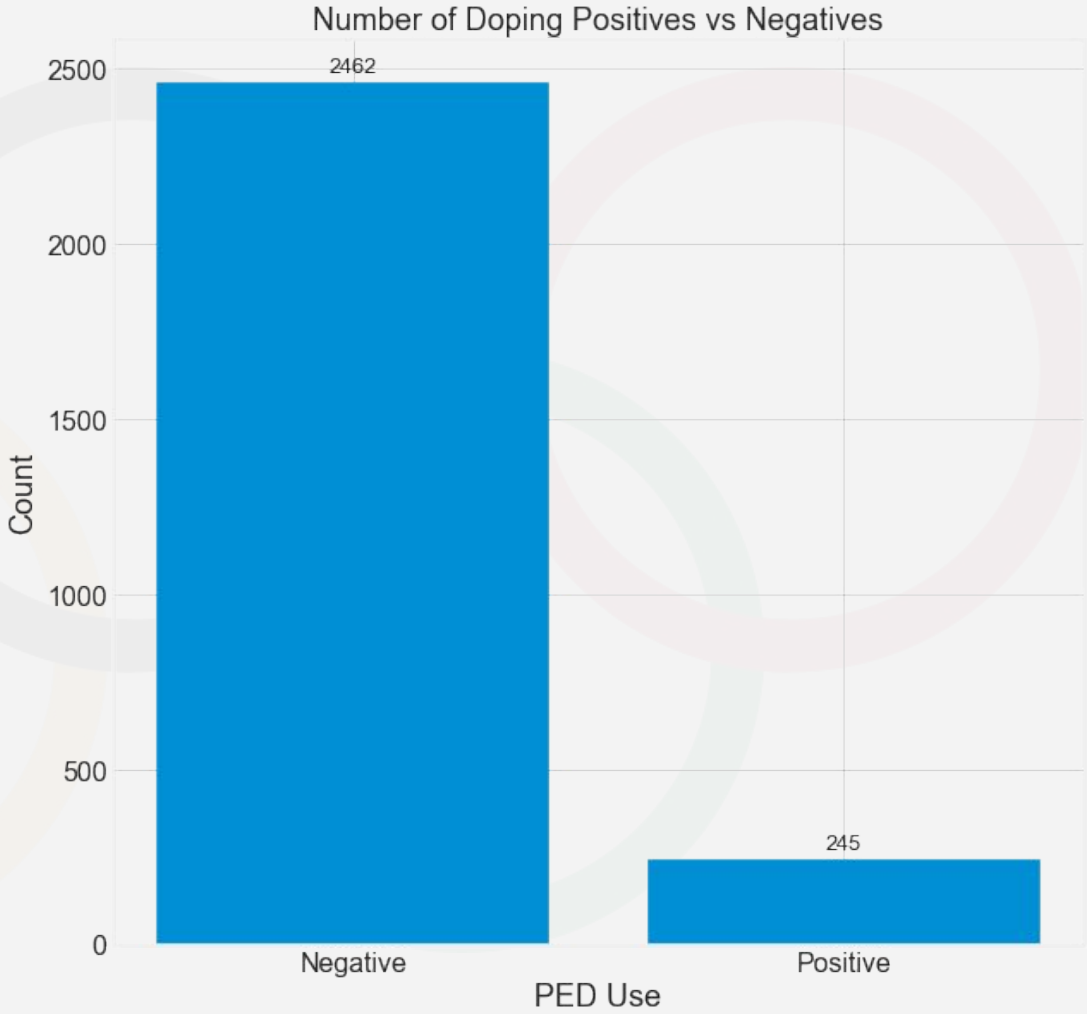
Combined multiple data sources together with added feature indicated positive or negative PED use

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.

Recall	**9%**
Precision	6%
Accuracy	81%
ROC-AUC	50.53%

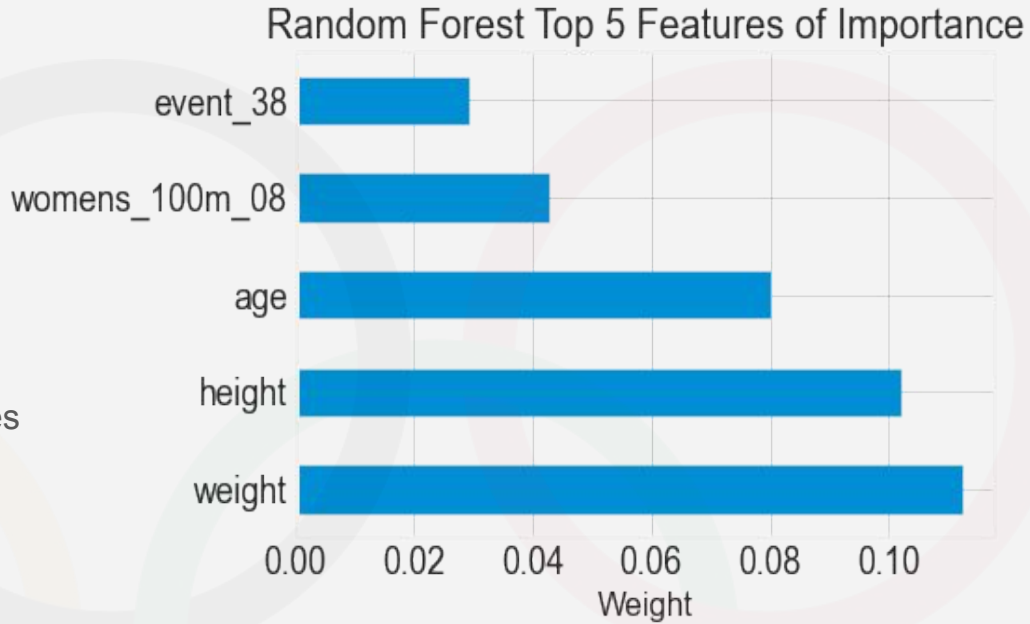
Best Model

Random Forest Classifier

- Optimizing for recall to limit False Negatives

Parameters:

- Random Undersampling the majority class



Recall	**75%**
Precision	32%
Accuracy	83%
ROC-AUC	80%

Next Steps

- Include event results from other Olympic Sports
- Improve upon class imbalance before deployment
- Create feature indicating difference in event results from previous year’s Olympic Games
- Neural Network classification modeling
- Model evaluation on next Olympic Games event results

Thank You

Jason Wong

Email:
jwong853@gmail.com

Github: jwong853