Module 03 Project Terry Stops: ‘Executive Summary’

Decision Trees and Random Forest by Alex Tyryshkin

Dataset

- Data source: https://catalog.data.gov/dataset/terry-stops
- Dataset Overview:
- Each row represents an individual Terry Stop
- The dataset includes information such as the gender and race of the officer, race and weapon type of the suspect, as well as the age group of the suspect and the time of arrest

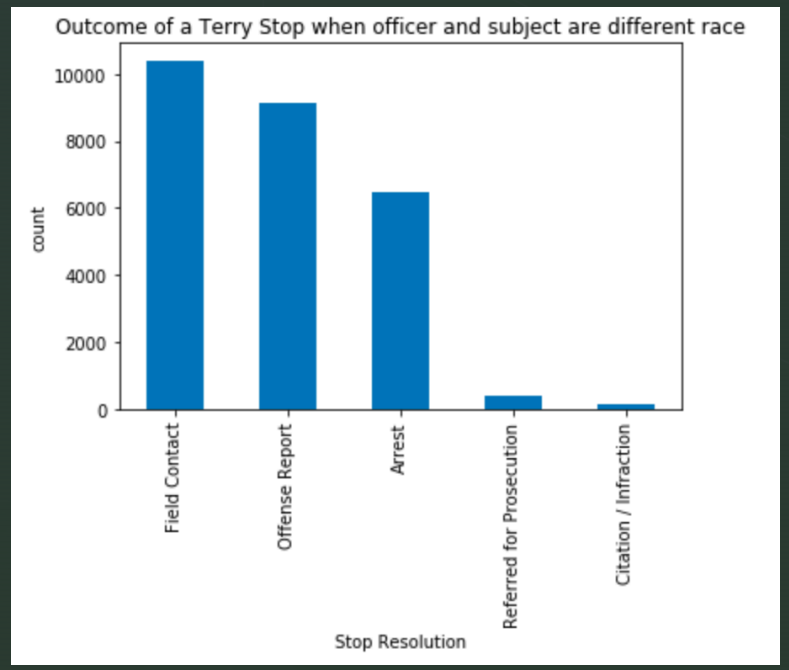
OSEMN Process: Obtaining and scrubbing

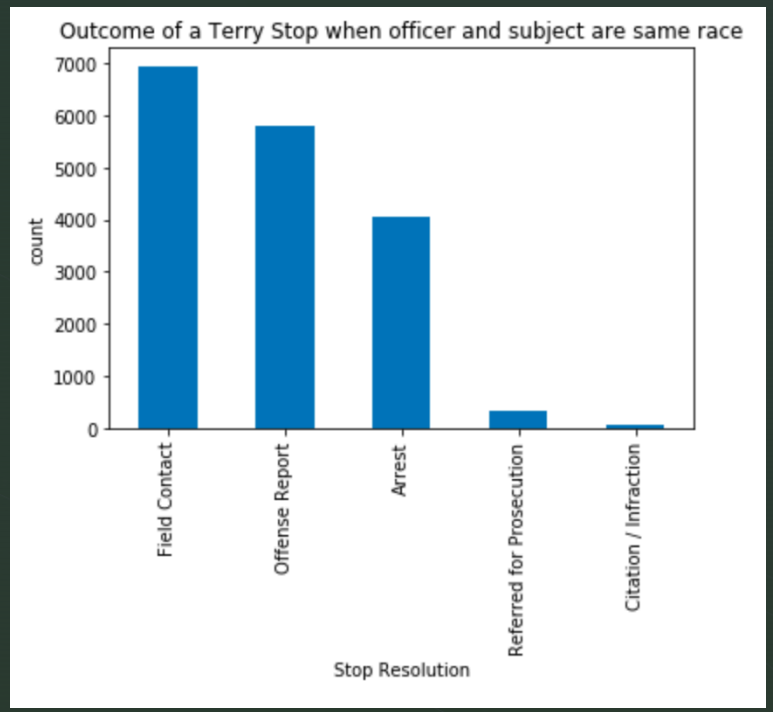
The dataset was obtained as a CSV file from the data.gov website

- Cleaning consisted of removing missing values and simplifying certain columns. For example, the column weapon_type was simplified to show only whether the suspect had a weapon, instead of a type of weapon, as the original categories were not particularly precise
- The column subject_id was simplified to show whether the subject had ID or not, instead of showing a particular ID for each subject
- Stop resolution was simplified to show whether an arrest was made or not, and missing gender values were assigned to female

OSEMN: Explore and Visualize

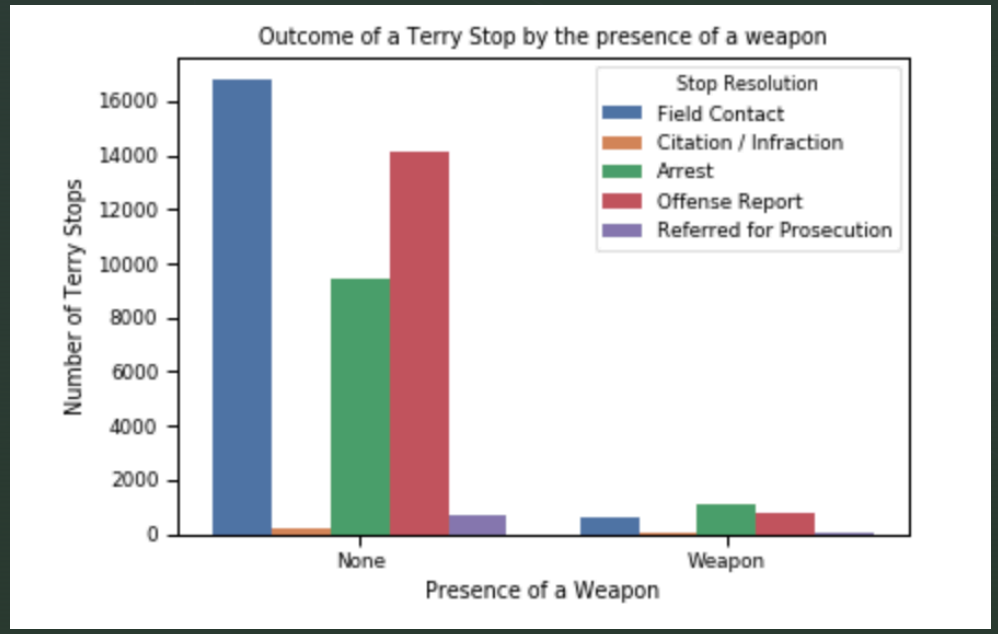
Does race of the subject and the officer have a significant impact on the outcome of a Terry Stop? It appears that the outcomes of Terry Stops are similar for both situations, however a closer inspection is needed for a more conclusive answer





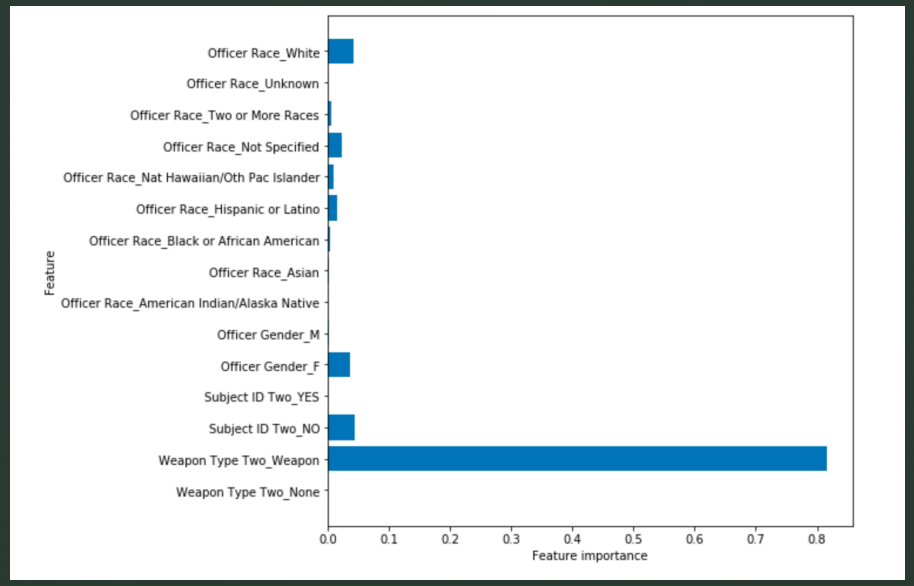
OSEMN: Explore and Visualize

- Does the presence of a weapon increase the chance of an arrest? A greater proportion of suspects carrying a weapon were arrested during a Terry Stop, in comparison to those who didn't.



OSEMN: Decision Tree

The most important feature was shown to be the type of weapon carried by the subject, which was closely followed by the race of the officer and whether the subject carried an ID.



Confusion matrix for decision tree classifier

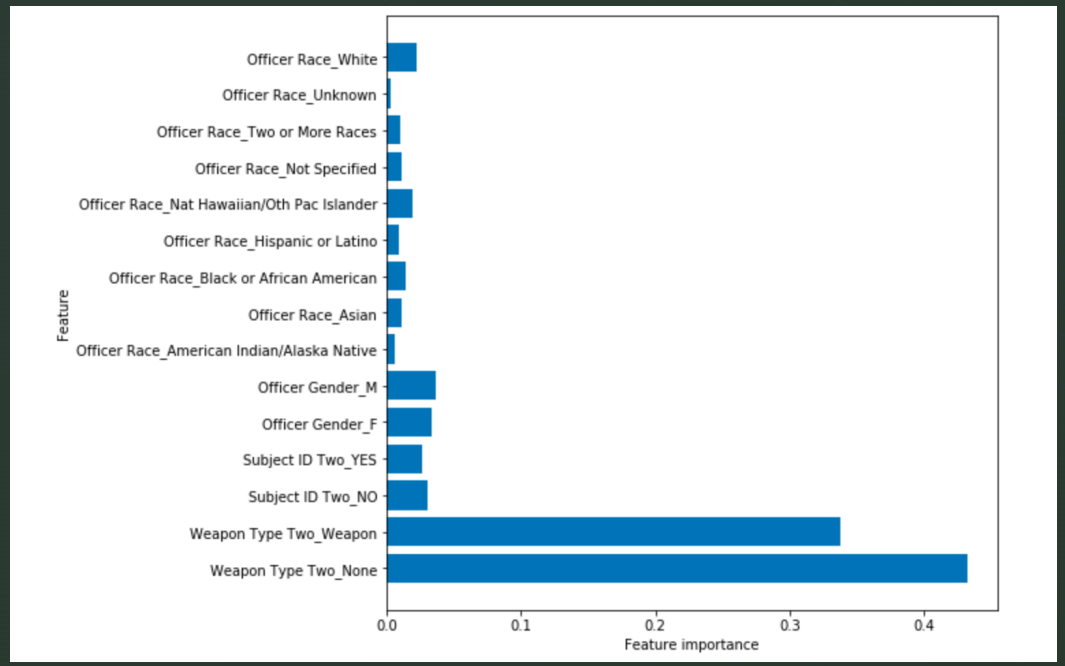
```
[[8260   18]
 [2636   11]]
              precision    recall  f1-score   support

          NO       0.76      1.00      0.86      8278
         YES       0.38      0.00      0.01      2647

    accuracy                           0.76     10925
   macro avg       0.57      0.50      0.43     10925
weighted avg       0.67      0.76      0.65     10925
```

OSEMN: Random Forest

- Weapon type, presence of ID, and officer gender were shown to be the most important features for the random forest as well
- Both training and test accuracy scores for decision tree, random forest, and bagged trees classifiers were between 0.75 and 0.76



Summary and Future work

Random Forest seemed to be the best classifier for this dataset
The presence of a weapon was linked to a greater proportion of arrests
The lack of ID and the officers' race and gender also played a role
There was no clear evidence that the difference between officer's and subject's race increased the likelihood of an arrest

Collect more data about the subject and the context of a stop:
-The subject's income and educational level
-The subject's clothing and the type of location where the stop happened
-The officer's years of service

Thank you!

Alex Tyryshkin

styrysh@gmail.com