New York Police Department Arrest Data

Proposal for Machine Learning Project

Group members:
Natasha Arokium
Jiffar Abakoyas
Bethold Owusu
Thierno Diallo

Introduction & Motivation: New York City ranks #25 in the United States list of cities with most violent crime rates[1]. Since we all reside in New York City and safety is a common concern, we thought it would be useful to understand which factors about crimes in NYC allow us to make predictions. In order to do this, we analyze the New York Police Department's citywide crime statistics for 2019. The NYPD dataset contains arrest information that contains information such as the type of offence, the age and race of the offenders, the borough where the crime occurred, and other relevant information for analysis.

Our main machine learning task will be to predict the race of the perpetrator from other descriptions of the crime. We will also predict the age group of the perpetrator from the descriptions. We think this is important to help the NYPD effectively search for the right suspects after a crime has occurred. It is very important to be accurate with this prediction and not have 'false positives' that show bias in the model towards certain races. If the race of the perpetrator can't be predicted from other features of the crime, then that will also be important information to note.

Preliminary Exploratory Analysis: To do better machine learning, we'll first do the following EDA to understand the data:

Location & Time: Brough with most offences. Number of offenses throughout the year.
Demographics: Race, age-group, and sex with most offenses.
Types of Crimes: Most common crimes. And other breakdown by type of crime.

Machine Learning Method: We will use classification to make predictions on features for race and age group. We will then use an ensemble of Logistic regression, Decision tree, and KNN to train our model to make a prediction and compare it to the stand-alone logistic regression.

---

[1] According to the USA today article "*Dangerous States: Which States have the highest rates of violent Crime and most murder*", Jan 2020.

Intended Experiments: The machine learning methods to be employed are the following:

1. Predict the race of the arrested person. The possible values are: [BLACK, WHITE HISPANIC, WHITE, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, UNKNOWN, AMERICAN INDIAN/ALASKAN NATIVE]. In this case, race is the target variable. The independent features include age, date of crime, type of crime and so on (all other features). The method employed will be multiclass logistic regression.
2. Predict the perpetrator's age group using the other columns. Similar to the above using multiclass logistic regression. Age groups are: [<18 ,25-44, 18-24, 45-64, 65+].
3. Ensemble of Logistic Regression, KNN, and Decision Tree to predict race as well as age group.
4. Dimension reduction using PCA to find the two dimensions of greatest variation. Then conduct the above two logistic regressions again to compare accuracy.
5. Use K-means clustering to see cluster the crimes in NYC into different cluster, if any. And identify what the clusters have in common.

Evaluation

We plan to evaluate using a test split of 30% and also separately a 5-fold-cv to compare. We will use categorical cross entropy for loss and stochastic gradient decent (SGD) for optimizer. We will also use Adam optimizer for comparison, in addition to SGD. We will try different learning rates (0.001, 0.1, 0.3) to see which one performs best. We will stop fitting the model when the validation loss does not decrease any further.