NYPD Crime Analysis

Overview

Topic

The main objective of this analysis to is to explore NYPD Complaint data provided from NYC Open Data using a the Random Forest classification logistic regression model.

Specifically, the analysis will examine crime descriptions, times, locations, etc. to create a binary classification model that will predict whether or not the type of crime committed is likely to be a violent crime. Furthermore, this will be supplemented by Tableau heat maps to show high crime areas per crime type. Time permitting and given the expanse of the data available, we can see if crime rates change over months/years.

Reasoning

This analysis is inspired by increased focused on policing in the United States over the past year. We hope to make unbiased observations on crime trends using the developed model.

To explore the validity of divisive political agendas, and social movements in a large American City

Data Source Description

This current data source comes from the city of New York Open Data website. The data is provided by the NYPD and contains all valid felony, misdemeanor, and violation crimes reported.

- The dataset used contains data from 2020 to 2021 (324K rows of data). It can be found here https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

- There is an alternative dataset which contains data from 2006 to 2021 (7.38M rows of data) . It can be found here: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Questions to be Answered

This analysis is intending to give the likelihood/crime type (output) most likely to occur in a certain area, at a certain time (input). Questions this model should be able to answer include:

1. Q1
2. Q2
3. Q3

Data Exploration

To predict the frequency or the level of offense depending on different factors associated with the registered complaint and to analyze the factors in the dataset and eventually leading to the crime trend.

Analysis Phase

- Using regression model to forecast crime rates in NYC.
- Benefits: Provides correlation and reasonable outputs among many variables.
- Limitations: It is assumed that the cause and effect relationship between the variables remains unchanged. The assumption may not hold an accurate estimation of the values based on the model and it may lead to misleading results.