CS109A: Final Project
Protopapas, Rader, Tanner

Predicting types of Crime

Problem Statement

While we all aspire to minimize crime, it is necessary to first understand the severity of the issue and some of the most contributing factors. It may come as a surprise, but crime tends to be ubiquitous in all areas of the United States, including Boston and Cambridge. Pinpointing the *exact* causes of crime is impossible, as it is a highly nuanced and complex issue. However, factors such as gross income, economic disparity, and government infrastructure/support are often strong indicators. For example, not only is there a global trend for less-resourced areas to resort to crime, but it is widely believed that there's an increase in crime in regions that have a large variance in citizens' income levels.

The goal of this project is to first inspect some of the contributing factors. For example, does economic disparity (e.g., property value could be one possible proxy) contribute to the amount and types of crimes committed? If so, shifting socio-economic levels is a complex, difficult, slow process to improve, but perhaps more surface-level factors could improve crime rates (e.g., installing street lights).

Additionally, the main project goal is to try to predict the type of crime that occurred. Specifically, given an incident report and its included information, try to predict the *type* of crime that was committed (e.g., larceny, robbery, grand theft auto, residential burglary). We will focus on the City of Boston, which provides great datasets for this task (described below). It is up to the discretion of the group to decide (1) which types of crimes to focus on, but there must be at least 5 different types; and (2) which crime reports will comprise the train/dev/test splits.

Data Resources

- Crime Incident Reports (August 2015 to present) -- the target data
- Property Assessment (FY2019)
- Streetlight Locations

We expect students to explore the predictive ability of property values and streetlight locations toward crime incident types. Note, the crime reports are annotated with a 'district' feature, whereas the other two files are not -- they use other geospatial labels, so students must find a reasonable approach to map the location information from all three datasets.

In addition, students are free (and encouraged) to use any other datasets that they wish.

High-level project goals

1. Determine a reasonable set of crime types to focus on, along with train/dev/splits
2. Develop an approach to use location data from all three datasets
3. Perform EDA to showcase the types of crimes, along with any correlations with property value, locations in general (e.g., maybe some neighborhoods are more prone to a certain type of crime than others), and the presence of streetlights.
4. Build a predictive model for classifying the type of crime.
5. Perform an error analysis, including details of the false positives and false negatives of your model, the common categories of issues, and analysis of features.

Note, this project should not be interpreted as being any reflection as to the CS109A staff's or Harvard's socio-economic or political views. We are not suggesting that one should try to predict crime for the sake of instituting targeted policing. We are aware of the harmful effects. This is purely an academic project whereby we want students to explore data and make statistical predictions, for the sake of illustrating learned skills that align with our course goals.