Predicting Types of Crime
Version 1.1 (October 1, 2019)

Pavlos Protopapas, Kevin Rader, and Chris Tanner

Harvard University
CS109A

1 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, it is widely believed that there's an increase in crime in regions that have a large variance in citizens' income levels.

You are tasked with inspecting factors that may be correlated and/or contributing to occurrences of Boston crimes and the *types* thereof (e.g., robbery, vandalism, etc). **You should explore datasets of your own choosing** to help answer your questions. Perform exploratory data analysis, iteratively refine your questions and chosen datasets, and produce plots to understand and communicate your findings. Toward these goals, we will start by asking you to explore two specific datasets that we provide: Street Lights Dataset and Property Values Dataset:

Let's start by looking at surface-level factors that may be correlated to crime: the presence of street lights. Are certain crimes more likely to occur where there is a dearth of street lights? Is there any correlation at all with well-lit streets and the locations of crimes?

Improving socioeconomic issues is a complex, difficult, and slow process. In attempt to understand potential correlations with inequality, we ask you to investigate economic disparity and crime. As a first approximation and proxy of economic disparity, we provide a dataset of property values. We ask you to find and explore additional datasets to estimate economic disparity.

Last, build model(s) to predict the *type* of crime that occurred. For this, you should define the set of crimes that are considered (e.g., larceny, robbery, grand theft auto, residential burglary, etc). At a minimum, pick five types of crimes.

Within the *Crime Incident Reports* dataset that we provide, you should designate reasonable sub-sections of it to serve as the training set, validations set, and test set. After training your model, provide to your model each incident report from the testing set and try to accurately predict the type of crime of that incident. Report your results and provide in-depth analysis of the performance.

What are the overall strengths and weaknesses of your models? Continue to improve your model to the best of your abilities. Last, comment on any additional ideas you have for a model that may be able to perform better, even if these ideas are not possibly by any model that we have learned about in the course – for example, explain what type of information you wish your model could use, and how do you wish it could leverage that information. This doesn't have to be explained in mathematical terms or theory, but the goal is to think about current limitations of models and how you wish to improve upon them.

2 Data Resources

- Crime Incident Reports Dataset (click here to view)
- Property Assessment Dataset (click here to view)
- Streetlight Locations Dataset (click here to view)

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.

3 High-level Project Goals

1. Describe how you will associate the location data provided in the Street Lights dataset with the location data specified in the Crime Incident Reports Dataset.
2. Describe an additional dataset that you will use, along with how it will be helpful.
3. Train and evaluate aforementioned models using the Street Lights dataset and Crime dataset.
4. Train and evaluate aforementioned models using the Property Value dataset and Crime dataset.
5. Use any combination of datasets you wish to predict the *type* of crimes that occur.
6. Use EDA, plot results, and make refinements to your models.
7. Perform error analyses, noting the best features, worst features, and any trends in misclassifications.
8. Discuss model weanesses and future ideas for improvements.

Last, this project is meant to be a chance not only to showcase your skills, but to get real, hands-on experience with data, so please take this opportunity to explore and create a project that reflects your interests and curiosity. That is, take liberty in expanding the project to how you see fit, while making use of any dataset that you find useful.

Note, this project should not be interpreted as being any reflection as to the CS109A staff's or Harvard's socioeconomic 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 in an area that is vastly affected by many confounding, real-world complex factors. The goal is for students to illustrate learned skills that align with our course goals.