STA 208 Project Proposal: Major Crime Prediction in Toronto, Canada

Matthew Chen, Mark Faynboym, Jasper Tsai

1 Introduction

Our team will pursue a big data analysis of crime data in Toronto, Canada. The overall goal is to predict different crime types using data collected by the Toronto Police Service. Our research questions are as follows:

1. How well can we predict crime types using the given data?
2. What variables are most influential in the prediction?
3. What region in Toronto has the highest occurrence of major crimes?

2 Data Description and Data Source

The data will be sourced from the Toronto Police Service (see hyperlink). The data includes 323, 296 observations collected from 2014 to 2022. The raw data includes 31 columns, including the major crime types (assault, break and enter, auto theft, robbery, and theft over $5000), time and date of the crime, location type (i.e. apartment, bars, commercial, train, etc…), and division of the police department.

3 Methods

3.1 Exploratory Data Analysis

Some preliminary ideas for EDA include creating histograms of crime types and other feature variables and creating a geographical map of the crime locations and color-coding by crime types, for example.

3.2 Feature Engineering (Unsupervised Learning)

We hypothesize that different regions have different crime types, although GPS coordinates are difficult to use in raw form. Thus, we propose using clustering to determine groups of coordinates that represent a larger region of the city. A simple approach would be to use K-means clustering, since it is fast to compute and interpretable.

3.3 Classification (Supervised Learning)

We propose several different candidate models, including multi-class logistic regression, random forest, gradient-boosted trees, and k-nearest neighbors. We will compare and evaluate the prediction performance of each model on k-fold cross validation, and we will choose a final model for the classification task based on our findings.

4 Potential Challenges

After an initial inspection, we find that the dataset may be unbalanced, particularly towards assault crimes. We will need to be careful in our evaluation metrics that are robust to this imbalance.

Additionally, as aforementioned, location is hypothesized to be important yet raw GPS coordinates may be difficult to work with. Further, the dataset includes the name of the

neighborhood, but this is also impractical to use directly as there are 140 different neighborhoods. This motivates the clustering analysis, though a potential challenge may be that clusters no longer include location information at a fine enough scale.