The University of Texas at Austin

Chicago Crime Incident Analysis
Project Proposal

INF 397 – Statistical Analysis and Learning w/ Prof. Varun Rai
Spring 2018

TEAM MEMBERS:
Milind Siddhanti (mss4376)
Nimish Kate (nk8648)
Prachi Singh (ps28755)
Sanchit Singhal (ss84657)
Shreshtha Shukla (ss83452)

Problem Statement

With a rise in the crime rate in metropolitan cities, it has become a priority for the authorities to identify crime-prone areas and take precautionary measures. To understand neighborhoods that need extra attention, it is imperative to identify zones with higher crime rates, predict recurring crimes, their frequencies, and analyze the time frames in which most crimes occur. A detailed analysis of the crimes would help the authorities understand the type of security measures that need to be put in place.

Information about the Data

The dataset reflects reported incidents of crime (except for murders where data exists for each victim) that occurred in the City of Chicago from 2001 to 2017. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. To protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

The dataset is a total of 1.85 GB. The entire data is split into 4 groups of specific time spans consisting of over 6,000,000 records.
2001-2004 - 1923517 records
2005-2007 - 1872346 records
2008-2011 - 2688712 records
2012-2017 - 1456715 records

The data has 22 columns, which will act as variables while building the models; they are listed below-

- ID - Case ID
- Case Number - Case number of the crime reported
- Date - Date of the occurrence of the crime
- Block - Area (block) of Chicago where the incident took place
- IUCR - Illinois union crime reporting code
- Primary Type - Top most classification of the crime
- Description - Description of the crime
- Location Description - Description of the location
- Arrest - Arrested or not (True or False)
- Domestic - Type of crime
- Beat - Police jurisdiction
- District - Police Jurisdiction
- Ward - Jurisdiction of the crime scene
- Community Area - Jurisdiction of the crime scene
- FBI Code - FBI code defining crime category
- X Coordinate - X coordinates of the crime scene location
- Y Coordinate - Y coordinates of the crime scene location
- Year - Year the case was filed
- Updated On - Date/Time of the update on the crime case
- Latitude - Location of the crime scene (Latitude)
- Longitude - Location of the crime scene (Longitude)
- Location - Location of the crime scene (Latitude Longitude)

Approaches Considered

- Our group will begin with various unsupervised methods to explore patterns in the dataset. Once key variables are identified, the team will use supervised learning models to classify or predict appropriate response variables.
- Potential unsupervised learning methods that might be a part of our exploratory data analysis are-
    - Dimensionality Reduction - Principal Component Analysis (PCA) to reduce the complexity of the dataset and increase the accuracy of relevant features
    - K-Means Clustering to group similar data points and develop characteristics distinct to that cluster
- In the next step, the following supervised learning methods can be used to produce predictions and inferences about the patterns found so far-
    - Linear or polynomial regression models for numeric responses
    - Classification for categorical data. Methods possible: Logistic regression (binary response such as arrested vs not arrested), Naive Bayes (this method might not be used as the assumption of independence might not hold to all the predictors), kNN (such as crime classification based on location), and decision tree (types of crime predictions).
- The occurrence of crime, as per assumption, is going to vary from block to block, so the data should be randomly picked from all the years and build according to the zones. Building individual models for different zonal areas might lead to less overfitting and lower test MSE.
- The team may also consider regularizing our model attributes to find the optimum model hyperparameters.
- The models will be using a stratified k-fold cross validation resampling method for the identification of the best model.

Key Uncertainties

- The geographical nature of the data, with multiple factors playing a role in the frequency and occurrence of crime, will abstain the use of a general model for the entire data while predicting results. A zone-wise modeling might give more accurate predictions.
- Lack of certain demographics (like unemployment, age, gender, alcoholism etc.) will leave dependent attributes from the database to be inadequate, which might lead to biased predictions (collinearity).
- Addresses have been masked to protect the privacy of crime victims, aggregating addresses to block level. This might reduce the accuracy of predictions.
- Records from the early years have missing and bad data. This may affect the analysis. The data may require preprocessing involving data cleaning and transformation. Some data generation will also be required (records with missing geospatial information but a known address might need imputation - possibly through an API).
- The data is very diverse with many variables affecting the prediction. Judging the individual effect of the predictors on the response will be difficult. The team is unsure of the number of interaction terms to add - a balance needs to be met between model complexity and accuracy.

List of references

1. (An Introduction to Statistical Learning). Retrieved March 02, 2018, from http://www-bcf.usc.edu/~gareth/ISL/

2. Boundaries - Police Beats (current) | City of Chicago | Data Portal. (n.d.). Retrieved March 02, 2018, from https://data.cityofchicago.org/d/aerh-rz74

3. C. (2017, January 28). Retrieved March 02, 2018, from https://www.kaggle.com/currie32/crimes-in-chicago/data

4. Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes | City of Chicago | Data Portal. (n.d.). Retrieved March 02, 2018, from https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

5. (Crime Type Categories). Retrieved March 02, 2018, from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html

6. (Google Maps APIs). Retrieved March 02, 2018, from https://developers.google.com/maps/