Analysis and Classification of Crimes in Chicago

Ye Zhou

September 15, 2017

1 Definition

1.1 Project Overview

Historically, Chicago saw a major rise in violent crime starting in the later 1960s [1]. More recently, the crime situation in Chicago is getting even worse. Last year, Chicago has experienced a recent spike in homicides of 762 people, an increase of 58 percent over 2015 [2]. Compared with other largest cities in United States, Chicago has a significantly higher murder rate than New York or Los Angeles [3]. There is no denying the fact that crimes have become a severe social concern for Chicago as prosperous communities in the long term. Here, we are interested to analyze the Chicago crime data through visualization tools in python, and make predictions for the category of crimes given time and location, using machine learning algorithms.

1.2 Problem Statement

This project is aimed to predict the Chicago crime type based on time and position information. The raw data from Chicago Data Portal is preprocessed to the features (year, month, weekday, hour, location description, community, latitude and longitude) and target variable (crime type), and then split into train, test and validation dataset (6:2:2). After an exploration in the visualization of the crime data, we will apply logistic regression (benchmark model) and xgboost algorithm with finely tuned parameters on the train and test dataset. The performance of two algorithms will be evaluated with log loss function and accuracy and finally reported on the validation dataset.

1.3 Metrics

This project use the log loss function and accuracy, from sklearn.metrics package, to train the model and evaluate the performance.

2 Analysis

2.1 Data Exploration

The data is imported from the Chicago Data Portal website [4], which is a collection of city data related not only to crimes, but also to education, transportation, health, and so on. The dataset has 21 columns:

- ID - Unique identifier for the record.
- Case Number - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- Date - Date when the incident occurred. this is sometimes a best estimate.
- Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- IUCR - The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description.
- Primary Type - The primary description of the IUCR code.
- Description - The secondary description of the IUCR code, a subcategory of the primary description.
- Location Description - Description of the location where the incident occurred.
- Arrest - Indicates whether an arrest was made.
- Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- Beat - Indicates the beat where the incident occurred. A beat is the smallest police geographic area each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
- District - Indicates the police district where the incident occurred.
- Ward - The ward (City Council district) where the incident occurred.
- Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas.
- FBI Code - Indicates the crime classification as outlined in the FBIs National Incident-Based Reporting System (NIBRS).
- X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

- Y Coordinate - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- Year - Year the incident occurred.
- Updated On - Date and time the record was last updated.
- Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Location - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

From the variables above, we plan to extract the year, month, weekday (from Monday to Sunday) and hour features from the column Date. And the positional information is described by a combination of Location description, Beat, District, Ward, Community Area, Latitude and Longitude. The information of IUCR and FBI Code are excluded, because these they contain the crime classification information.

The Primary Type of crimes is the target variable for prediction in our models, which has 33 categories in the 2016 crime data. To make the model more robust and efficient, here, we transform the Primary Type to major crime types (theft, battery, criminal damage, assault, deceptive practice, burglary, narcotics, robbery, motor vehicle theft), and severe crime type (homicide). Figure shows the occurrence of different crimes in Chicago from 2015 to 2016.

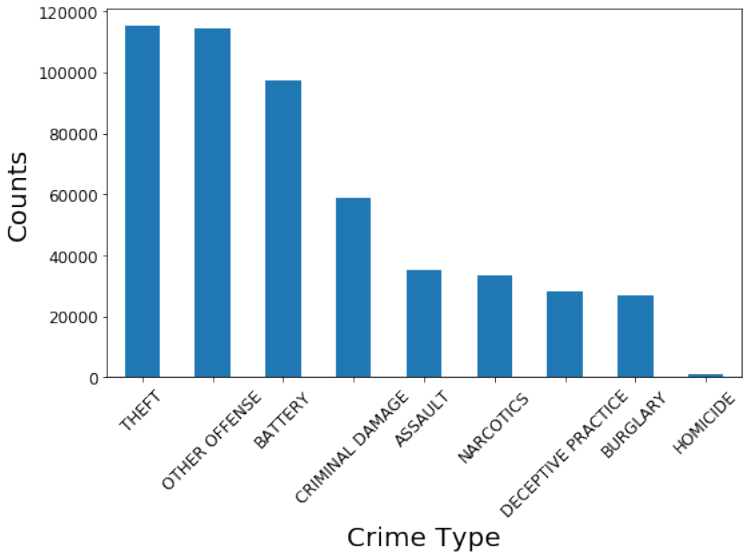

Figure 1: Total crimes in Chicago from 2015 to 2016

The whole dataset will be split into train, test set (6, 4) for the model development.

2.2 Exploratory Visualization

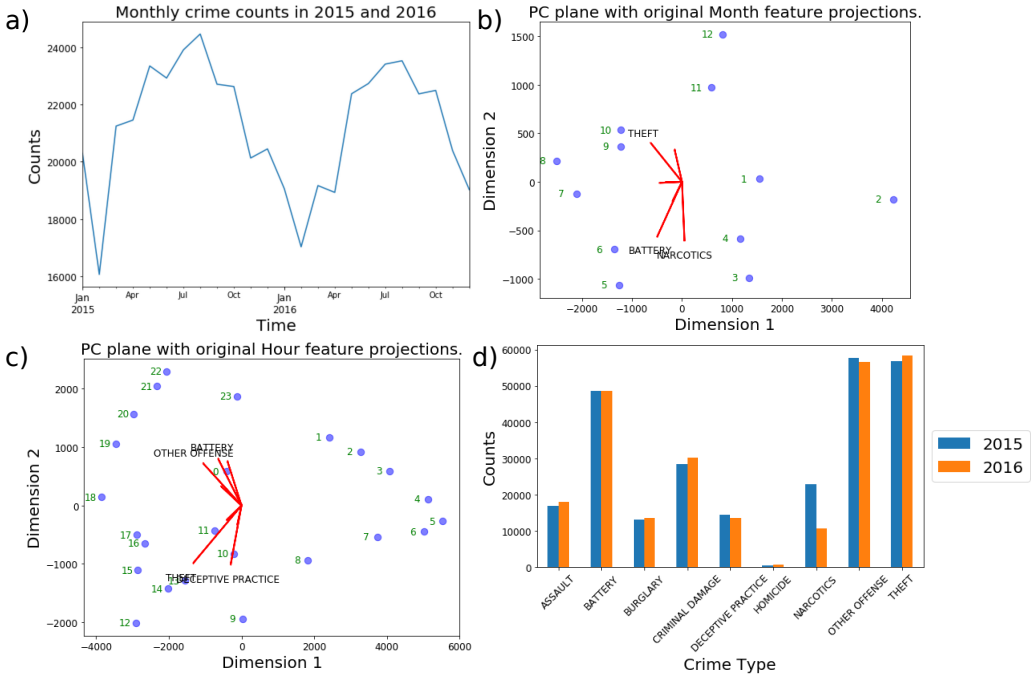

Figure 2: a) The monthly crime counts in Chicago from 2015 to 2016. b) Principal component analysis of crimes by month and type. c) Principal component analysis of crimes by hour and type. d) Comparison of crime count between 2015 and 2016.

2.3 Exploratory Visualization

2.4 Algorithms and Techniques

To go beyond linearity, the tree-based Gradient Boosting will be used with the package xgboost. This algorithm combines weak learners into one strong powerful learner. The parameters wait to be tuned are:

- The learning rate (eta)
- Min_child_weight, which defines the minimum sum of weights of all observations required in a chiral.
- Max_depth, which is the maximum depth of each weaker learner. Deep trees lead to over-fitting more easily.
- Gamma, which is defined as the minimum loss reduction required to make a split. Higher value prevents the model from over-fitting.
- Subsample, which is the fraction of observations to be randomly samples for each tree. Lower value prevents the algorithm from over-fitting.

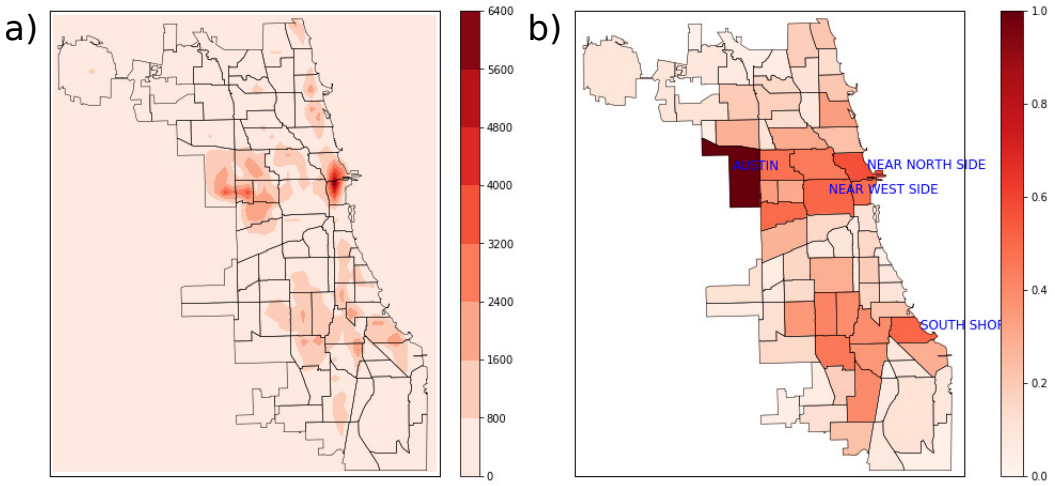

Figure 3: a) Crime density in Chicago between 2015 and 2016. b) Crimes in Chicago communities.

2.5 Benchmark Model

As a classification problem, the logistic regression model is used as a benchmark model. Logistic regression is a linear model, which could be used for the multi-class categorization. Two different regularization method, Ridge and Lasso will be tested with a range of regularization parameter values.

3 Methodology

3.1 Data Preprocessing

3.2 Implementation

3.3 Refinement

4 Results

4.1 Model Evaluation

4.2 Justification

5 Conclusion

5.1 Free Form Visualization

5.2 Reflection

5.3 Improvement

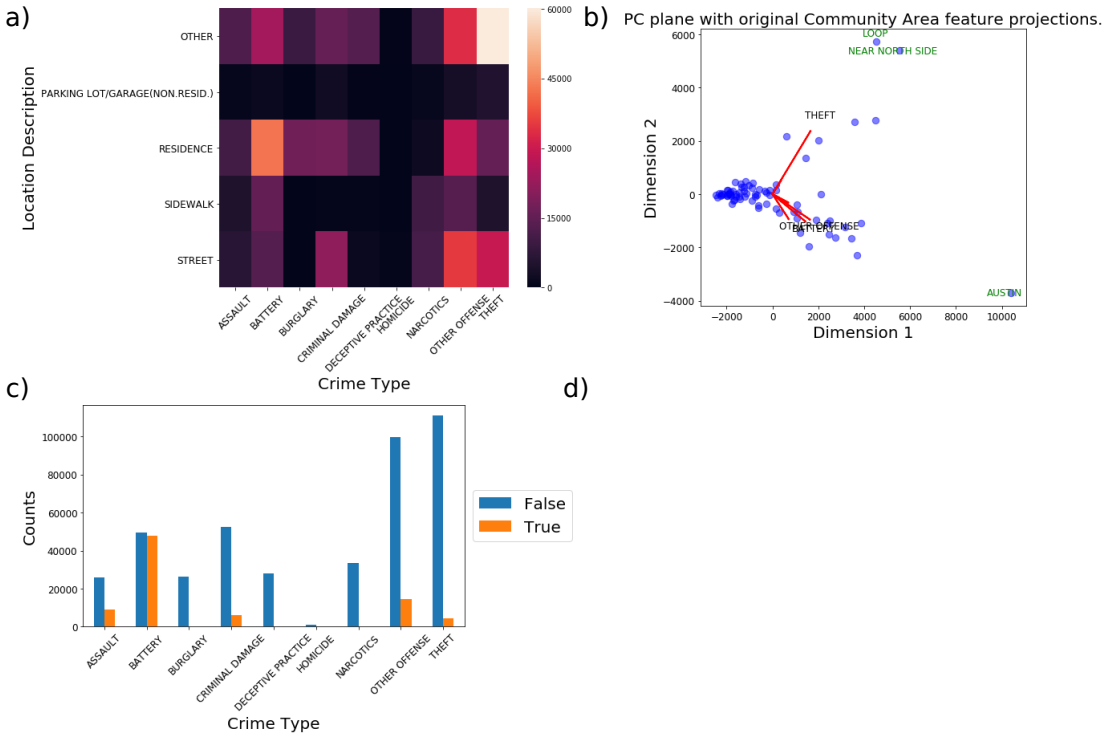

Figure 4: a) Relation between location description and crime type. b) Principal component analysis of crimes by community and type. c) Domestic related and unrelated crimes for different crime types in Chicago. d) Clusters of crimes by latitude and longitude using Gaussian Mixture Model.

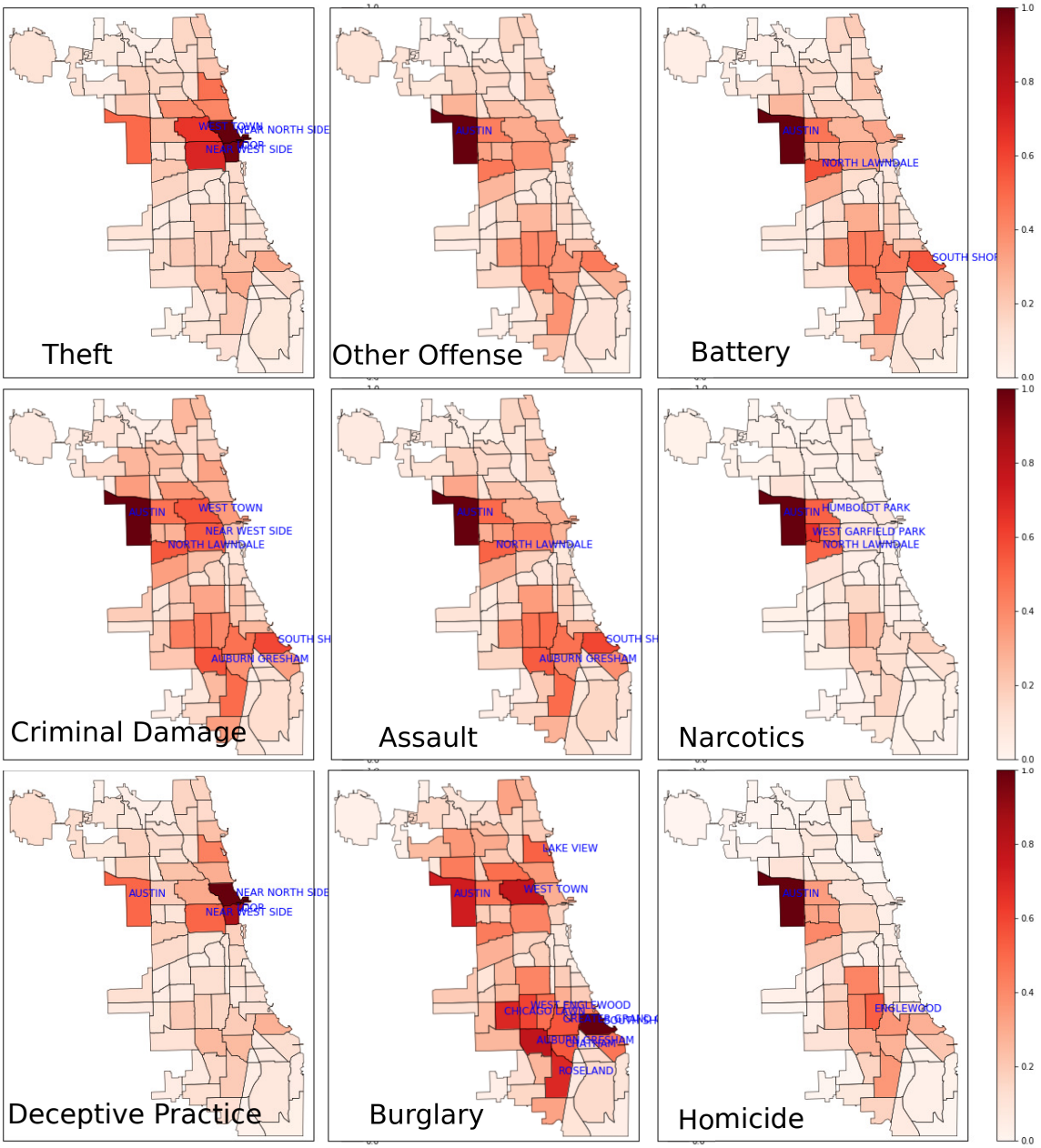

Figure 5: Standardized crime count in Chicago communities from 2015 to 2016 for different crime types.