Chicago Crime Analysis and Classification

Ye Zhou

09/07/2017

I. Introduction

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.

II. Data Description

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. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.
- 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. See the beats at https://data.cityofchicago.org/d/aerh-rz74.
- District - Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.

- Ward - The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.
- Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.
- FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.
- 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', 'Community Area', 'Latitude' and 'Longitude'.

The 'Primary Type' of crimes is the target variable to predict in our models, which has 33 categories in the 2016 crime data. The counts of each type is shown below:

Table 1 Counts of different crime types committed in 2016

Crime Type	Counts
THEFT	61530
BATTERY	50277
CRIMINAL DAMAGE	31009
ASSAULT	18734
DECEPTIVE PRACTICE	18185
OTHER OFFENSE	17211
BURGLARY	14288
NARCOTICS	13252
ROBBERY	11958
MOTOR VEHICLE THEFT	11305
CRIMINAL TRESPASS	6305
WEAPONS VIOLATION	3448
OFFENSE INVOLVING CHILDREN	2266
PUBLIC PEACE VIOLATION	1606
CRIM SEXUAL ASSAULT	1493

SEX OFFENSE	949
INTERFERENCE WITH PUBLIC OFFICER	934
PROSTITUTION	800
HOMICIDE	781
ARSON	516
LIQUOR LAW VIOLATION	227
KIDNAPPING	202
GAMBLING	189
STALKING	172
INTIMIDATION	134
OBSCENITY	53
NON-CRIMINAL	48
CONCEALED CARRY LICENSE VIOLATION	36
HUMAN TRAFFICKING	11
PUBLIC INDECENCY	10
NON - CRIMINAL	5
OTHER NARCOTIC VIOLATION	4
NON-CRIMINAL (SUBJECT SPECIFIED)	1

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).

The whole dataset will be split into train, test and validation set (6, 2, 2) for the model development.

III. Project design

1. Exploratory Visualization

We will examine the time and positional dependence for different crime types. Especially, Basemap function will be applied for the visualization of geo data. The crimes in 2016 on top of the Chicago map is shown in Figure 1.

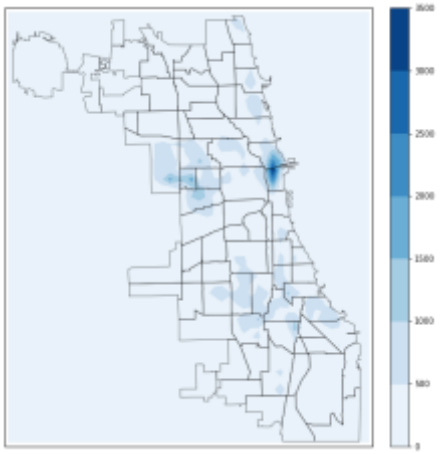

Figure 1 Crimes in Chicago in 2016

2. Benchmark Model

As a classification problem, I plan to use the logistic regression model 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. Solution Statement

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.

4. Evaluation Metrics

In order to evaluate the performance of the model, we will use the log loss function as metrics to train the model. In the end, the accuracy will also be reported in the validation set. Both are from the sklearn.metrics package.

IV. Conclusion

To sum up, 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.

Reference:

[1] https://en.wikipedia.org/wiki/Crime_in_Chicago

[2] https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/

[3] Heinzmann, David (January 1, 2003). "Chicago falls out of 1st in murders". Chicago Tribune. Retrieved August 8, 2012

[4] https://data.cityofchicago.org/