A data-mining oriented approach to public safety

Contributors

Aaron Shyuu | Max Smith | Thomas Dolan

Description

High levels of crime have plagued the city of Chicago for decades. Research has turned up little in terms of contributing factors.

The intent of this project is to mine the Chicago Crime dataset for unique patterns found amongst all crimes committed over the last four or more years, especially as they relate to standardized crime categorizations defined by the FBI and in relationship to concurrent local economic metrics reported by the CFNAI. Using these unique approaches to our dataset with KDD methods and practices, our investigation purports to supplement public safety intelligence in Chicago by answering questions like:

- How is crime correlated to economic boom and bust?
- Do patterns of property crime differ from other types of crime, and how?
- What locations are associated with what types of crime and where are the crime hotspots?
- Is there an association between violent crime and unemployment metrics?
- Can we predict whether an arrest will be made for a commited crime?

Prior Work

Chicago Crime Data set is a publicly accessible dataset on Google's BigQuery. As such, quite a few people have used this dataset either in their own analyses on Chicago Crime or for learning how to use BigQuery. Some noteworthy examples:

- https://www.kaggle.com/mkrkkinen/chicago-crime-on-christmas
- https://www.kaggle.com/alkadri/chicago-codes-crime-map-and-bigquery-in-r
- https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20
- https://cjango.wordpress.com/portfolio/chicago-crime-data-analysis-python-project/

Crime in Chicago has also been the focus of many government and academic studies and initiatives as well. A few are:

- https://www.ipr.northwestern.edu/news/2018/crime-in-chicago-research.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831577/

In order to yield a novel analysis, we will be deriving additional attributes for our crime dataset using crime categorizations from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html, which will allow us to do additional multilevel pattern analyses and more. We will also be supplementing our primary dataset with economic data from the CFNAI, shown in our datasets slide. We have not seen either approach used in other works that we are aware of.

Datasets

CHICAGO CRIME DATASET - Primary

Reported incidents of crime in the city of chicago. Each record indicates the type, description, time, location, and some nominal categorizations of the crime committed. .

Found: https://www.kaggle.com/chicago/chicago-crime

Whether downloaded: We have not downloaded the chicago crime dataset. As a publicly hosted dataset on Google's Big Query, there is no downloaded link and a reasonable assurance that it will persist in the foreseeable future. Furthermore, the dataset is very large, complicating making a copy. We plan to make a constrained copy using our ETL timeline into a pandas dataframe. As possible we will expand the size of the dataframe copy and explore other ways of copying the data as possible.

CFNAI DATASET - Supplement

An aggregate of 85 series of economic data, weighted and normalized to a mean of 0 and standard deviation of 1. Adjusted for inflation and growth trends. Predictive of recessions and other economic troughs and peaks. Can be broken down into four meaningful categories.

Found: https://www.chicagofed.org/research/data/cfnai/historical-data

Whether downloaded: Yes. We are currently storing the dataset in an excel file on our github. Will integrate into our ETL pipeline and convert to a pandas dataframe for data mining.

Additional reference sites that provide data context and will help us to derive new attributes, create association rules, and more:

- http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html
- https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Beats-current-/aerh-rz74
- https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Proposed Work

Cleaning

Both of our datasets are quite clean (as far as we can tell). However, the Chicago crime dataset contains millions of data points so we need to verify and address any noise and incomplete data. Therefore, our proposed data cleaning steps are to:

- Identify any null values and develop a strategy for filling in missing data using global constant, attribute mean, inference approach or by removing the object (eg. X coordinate, Y coordinate and location contain null values)
- Check validity and correctness of data and perform data scrubbing tasks as necessary, such as checking for and removing duplicate items.
- Review attributes for entry errors, in particular for crime type attributes such as fbi code.
- Identify noise and outliers (except when performing outlier analysis) and develop strategy for handling them, for example using clustering to detect and remove outliers.
- Remove timestamp from date; we do not expect to inspect data at that level of granularity at this time.

Proposed Work

Pre-Processing

- Converting arrest and domestic attributes into boolean values
- Normalizing other attributes for data mining. In particular this entails converting nominal attributes, in particular ‘primary type’, ‘description’, and ‘location description’ into decided upon encoded numeric values.
- Aggregating data / binning to reduce processing time and improve visualization.
- Smoothing out outliers when they affect data mining result.

Note: *CFNAI dataset has already been aggregated, each series of data weighted, and then normalized to a mean of 0 and standard deviation of 1.*

Note: *fbi code, ward, and community area are already encoded numerically. These are nominal data attributes, not quantitative, so will not be normalized to the same standards as CFNAI dateset since they are not comparable in that way.*

Proposed Work

Integration

- Our economic dataset can be joined with Chicago Crime Dataset using the month of our date attribute. How we perform this integration may vary depending on the analyses being performed. For example we may choose to do some binning of the crime dataset by month or otherwise perform an aggregation that will allow us to compare the data against economic trends. We expect this to be a large part of how our analysis is performed and to change dynamically depending on the question being asked.
- We will be deriving new attributes based on the FBI defined crime types and categorization for each crime object, especially “crimes against person”, “crimes against property”, and “crimes against society”, that can be used in multilevel pattern analyses and for identify other trends especially as they related to concurrent economic based patterns and trends. How we derive these attributes may also change dynamically depending on the question being asked, for example we may derive boolean attributes for more focused analyses, and integer sets for more general or hierarchical pattern analyses.

Proposed Work

Reduction

- Removing attributes that are not useful for our data mining questions or with incomplete data such as case number and updated on.
- Removing derivable data (e.g. year and date are redundant attributes).
- Removing ID attributes after data cleaning / pre-processing is complete.
- We plan on constraining our data to a four year period initially, which will allow us to do lower support / more general pattern analyses initially and to refine our data mining techniques. We will expand this period for more refined analyses as we are able and see fit.
- Reduce attribute by performing correlation analysis (Chi-square test for nominal data) or otherwise determining whether an attribute is valuable. For example we expect the Ward attribute, an outdated method of geographic organization, to yield un-meaningful results and will look to verify that assumption so it can be removed.

List of Possible Tools

Google BigQuery

Jupyter Notebooks

SQL

Python

Pandas / Numpy / Matplotlib / Seaborn

Tableau

Excel

D3.js

Evaluation

In general, a successful project will be evident if previously undiscovered patterns are uncovered during the KDD process. An additional milestone will be to build out a tool which can assist analysts in the KDD process.

For regression models comparing crime vs. econometric data, we plan on using standard evaluation techniques including R-squared to determine the significance of our associations.

For our prediction models, we plan on using k-Fold Cross Validation techniques to ensure that we are not overfitting our training data.