CrimeIntel

A data-mining approach to public safety

Max Smith
maxs@colorado.edu

Aaron Shyuu
aaron.shyuu@colorado.edu

Thomas Dolan
thomas.j.dolan@colorado.edu

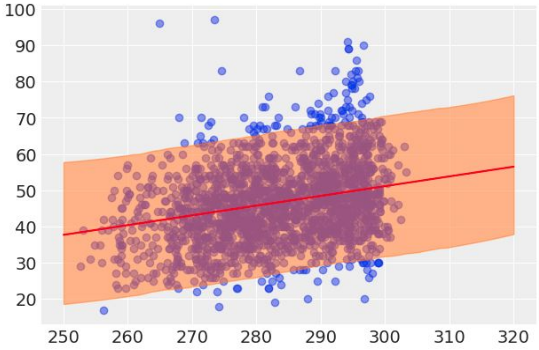
PROBLEM STATEMENT

Chicago has received national attention for its high homicide rates, especially since the rates skyrocketed in 2016 [1]. This attention has precipitated a lot of interest in generating a solution, but has also shown that the causality of the problem has many roots, which in turn means that the impact of many proposed solutions have intractable results that are difficult to predict [2]. This has been further complicated by those purporting to have a grand solution, but whose primary goal is to support their own agenda. In order to provide possible solutions, then, we have to understand the nature of crime occurring in Chicago with as much nuance, depth, and objectivity as possible. Using an objective, data driven approach is a likely means to that end. The size of Chicago and its crime data, collected by the Chicago City Police, lends itself to this approach. However, still missing is the very important community-driven data, such as surveys after an incident [2]. This means any study on crime in Chicago will be inherently incomplete. For that reason, our goal is not to solve the problem, but to lend a new perspective that adds some texture to the landscape that is our understanding of Chicago crime.

By using an unique combination of supplemental data-sets and other auxiliary information in addition to a data mining methodology, we hope to contribute to the conversation by answering smaller, more manageable questions and proffering any new, unexpected insights we might find along the way. We hope to discover whether crime is correlated to local economic trends, and if so, how? We hope to discover whether homicide is correlated or associated with other types of crime. We will be looking at broader trends as well, such as whether property crime or violent crime have similar or different patterns from other crime-type categorizations. We also hope to find what locations are associated with what types of crime in an effort to identify crime hotspots both on a per crime basis, as a crime-type aggregate, and in general. We hope to be able to delve deeper into crime type associations with more specific economic metrics as well, such as whether there is an association between violent crime and unemployment metrics. Finally, using classification methods, we will be able to predict based on crime, time, and place, whether an arrest will be made, leading us to additional insights and other data mining questions.

LITERARY SURVEY

Many analyses relating to Chicago crime have been done including those that have used our same dataset. Some of these projects exist as notebooks accessible on Kaggle's website, while some others are hosted on personal blogs or sites. Four of these in particular provided inspiration for our project, two of which utilize data mining / KDD methods. The project "Chicago Crime on Christmas" helped us to realize the benefits of integrating additional data types to yield novel insights, and the frequency of the data we would need in order to do so successfully [3]. A Python project by Cjango provided a vis heavy approach that helped us to get a better sense for the data we would be working with and what kind of questions we could be asking [4]. Two separate data mining projects using different tooling than we plan to use were especially useful for understanding how we can apply the data mining methodologies we have been learning in class to our dataset [5, 6]. In both cases we found that they used only the categories provided originally in the data set and realized that if we were able to derive more attributes using additional categorization information, we could create hierarchies within the data that will help us to yield additional insights.



Excerpt from Mikko Karkkainen's "Chicago Crime on Christmas" [3], showing a correlation between temperature and incidents of crime.

In addition to these projects, many academic and civic studies have also been done. A data driven article published by Northwestern helped give us context surrounding the problem, potential causes, and possible solutions [1]. An academic article by Julia Burdick-Will on the relation between violent crime and academic achievement in Chicago schools helped to illuminate the challenges associated with identifying causality and prescribing solutions to crime given its endemic nature and the role that data plays in this.

PROPOSED WORK

After the data collection process, we decided on the Chicago Crime dataset with the CFNAI dataset as a supplement. Before performing any Exploratory Data Analysis or data mining techniques, we need to ensure that our raw data are transformed into useful and efficient formats. Our focus for data preprocessing will be on the Chicago Crime Dataset, but the same procedures will apply to the CFNAI dataset. The Chicago Crime dataset is a huge dataset consisting of millions of rows. This means that the first step to cleaning our data will be to verify any missing/null values and address them appropriately. Through some initial filtering, it is clear that the majority of our missing values occur in the location and coordinate attributes. Due to the size of our dataset and the relatively small number of rows that contain at least a column with a missing value (< 0.1% of our dataset), we can safely remove them with minimal loss. Next, we need to check if there is any redundancy in our data and remove duplicate rows as they take up unnecessary space. We will be removing exact timestamps from dates since they overlap with other time data and we don't expect to inspect data at that level of granularity. To validate the quality of our data, we will review certain attributes that could be more prone to input errors, such as the fbi attributes and make corrections if needed. Lastly, we will check and remove noisy data with a binning or clustering approach to minimize skewing in our results.

In terms of data reduction, we will remove attributes that are not beneficial to EDA or data mining such as case number and update time. Moreover, any ID attribute columns will be dropped as they do not contribute to the data mining process. We will iteratively remove attributes to narrow down our data further until all remaining attributes are essential. Since location attributes such as beat, district and ward are similar and might yield the same correlation, we will perform a Chi-Square test (correlation analysis) to see if we can reduce from those attributes. For initial pattern and data analysis, we will constrain our data to a four year period for efficiency and faster runtime. The scope of our data will increase as we begin with more refined analysis and the implementation of our data mining techniques.

To find correlations between the Chicago Crime dataset and economic dataset, it is crucial for us to integrate them. We will join the two datasets through the date attribute. This will give us a combined table where we will be able to see the values of crime attributes and economic attributes for a given date. (represented in an object or row of the table). We may also perform other integrations and aggregations as necessary to compare crime data against economic trends. In order to find correlations and associations, we will also make some transformations which includes rearranging our data into chronological order and grouping by crime types depending on the mining knowledge we are trying to gain. We will also derive nominal attributes—crimes against person, crimes against property, and crimes against society— which will be useful for multilevel pattern analyses and their relation to economic trends. Furthermore, new attributes such as fbi code descriptions and weekend (a binary yes or no) will be created to support classification and model construction. For tactics involving a join on the CFNAI dataset, we will be reducing the data attribute to only its month value. In essence, we will derive many additional nominal attributes to aid our data mining process later.

For further data preprocessing, we will need to convert all data (with many different data types) to numbers so we can use them for our data mining techniques. Firstly, we will convert the different attribute values to numbers. For example, we will first convert the arrest attribute to binary (1s and 0s) and the description attribute to the respective encoded numeric values. When we perform this conversion, we need to be attentive to use the correct conversion. (Manhattan distance for quantitative attributes and Hamming distance for categorical attributes) Next, we will normalize and aggregate our normalized distance matrices. Although our CFNAI economic dataset is normalized and aggregated, we will need to perform this process for our Chicago Crime dataset.

While many people have utilized the Chicago Crime dataset for data analysis and mining (please see the Literary Survey above), we did not see any past studies that integrated the crime dataset to an economic dataset. The “Chicago Crime on Christmas” is an excellent prior work that integrated the Chicago Crime dataset and a weather dataset to explore crime trends with changes in weather. It inspired us to upgrade our interesting data mining questions and add the CFNAI economic dataset as a supplement to our primary Chicago Crime dataset. Unlike previous usage of the Chicago Crime dataset, we will be exploring crime data against various economic attributes and model our data in multiple dimensions with data cubes. To gain more insights, we will also use OLAP operations to aggregate, slice and zoom into our data.

DATA SET

Our primary data set is the Chicago Crime dataset, hosted by Google’s Big Query. A link to the dataset and description can be found on Kaggle, at https://www.kaggle.com/chicago/chicago-crime [9]. The data is generated and owned by the Chicago City Police. Each record represents a reported crime and includes information regarding its time, location, and various categorizations describing the crime. The dataset contains 7,213,094 entries from 2001 to 2020.

Our secondary data set is the CFNAI Historical (Real-time) data set, created and maintained by the Federal Reserve Bank of Chicago. The CFNAI metric is a monthly average of 85 series of economic data representing a wide range of economic indicators, each of which has been weighted, normalized, and adjusted for inflation. Each record represents the CFNAI metric for a given month, and includes four similar metrics each of which is derived from a subset of the 85 series whose indicators fall into one of four categories, 1) production & income, 2) employment, unemployment & hours, 3) personal consumption & housing, and 4) sales, orders & inventories. It is available for download at https://www.chicagofed.org/research/data/cfnai/historical-data.

EVALUATION METHODS

In order to evaluate the performance of our classifiers which predict economic values with our crime data, we will use metrics such as accuracy, precision and error rate. We will also perform standard evaluation techniques including R-squared, P-value, F-test and confidence intervals to observe the variance and quality of prediction. This will allow us to analyze the predictive capability of our model and make adjustments accordingly. If a model is not making predictions or classifications up to our accuracy standard, it might be in our best interest to substitute it with another model. Moreover, we can also use the holdout method and k-fold cross validation to further validate our models.

To evaluate our result for whether an arrest will be made for a commited crime, we will likely use the classification technique with a decision tree. This means that we will first use our validation test set to assess the accuracy of our classification model. Furthermore, we can access the coverage and accuracy of individual rules in our decision tree. For other correlations and associations that we want to explore between crime and economic metrics, we will perform correlation analysis and use multi level association rules to break down the underlying support and confidence.

Evidently, we will also have many visualizations to support both our EDA and data mining results. For example, plotting the root mean squared error will be helpful for us to visualize the deviation of our errors from the actual values in our crime dataset. In general, a metric that we will certainly look at is the efficiency and runtime of different methods and models.

TOOLS

The tools that will be used to explore, analyze, test and visualize our results from our Chicago Crime and CFNAI datasets will include but are not limited to: Python and its libraries including Numpy, Pandas, Matplotlib, Sklearn and Seaborn. Google BigQuery, SQL and Excel will be used to transfer our data to the python libraries. For more choices and other useful visualizations, we will use Tableau and D3.js.

MILESTONES

The most important milestones of our project from this point on will be:

1. Building an Extract-Transform-Load (ETL) pipeline
2. Exploratory Data Analysis (EDA) to determine interesting questions
3. Using data mining techniques to answer interesting questions; create report on findings
4. (Stretch goal) Development of tools for finding interesting patterns (data cube, frequent pattern analysis)

Building an Extract-Transform-Load Pipeline

Deadline: 30 October 2020

Prior to analyzing the data, it is necessary to do some initial cleaning and integration. All preprocessing will be completed in a single Jupyter Notebook, where the (remote) data is queried via SQL, converted into a Pandas DataFrame, and subsequently cleaned. Once cleaned, the DataFrame will be integrated (joined) with our econometric data set. Finally, the fully-processed DataFrame will be exported to .csv format for use in our EDA Jupyter Notebook.

Exploratory Data Analysis (EDA) to determine interesting questions

Deadline: 20 November 2020

With a cleaned and fully-integrated dataset, we can begin examining the data to find interesting trends and correlations. This process will include creating visualizations and creating summary statistics for various attributes. This process will result in fully fleshed-out questions which we will attempt to answer using data mining techniques. The final "deliverable" for this milestone will be a Jupyter Notebook including all EDA and resulting questions.

Using data mining techniques to answer interesting questions; create report on findings

Deadline: 27 November 2020

With our questions in hand, we will begin attempting to answer them using traditional data mining techniques. Some of these techniques are mentioned in the "proposed work" section above. The final deliverable for this milestone will be a Jupyter Notebook, separated by question, displaying all data mining/visualizations and answers to our questions.

Development of tools for finding interesting patterns

Deadline: 11 December 2020

Finally, given enough time, we would like to make a generalizable tool for mining the Chicago crime dataset. This will involve building an application where users can answer their own questions about the dataset using basic data-mining techniques. The tasks that the application would support include rollup and drill down on various attributes, frequent pattern analysis, and visualization.

REFERENCES

[1] Mary Pattilo. 2018. Crime in Chicago: What Does the Research Tell Us? Retrieved October 2020 from https://www.ipr.northwestern.edu/news/2018/crime-in-chicago-research.html.

[2] Matt Ford. 2017. What's Causing Chicago's Crime Spike? *The Atlantic*. Retrieved from https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/.

[3] Kaggle. 2019. Chicago Crime on Christmas. Retrieved October 2020 from https://www.kaggle.com/mkrkkinen/chicago-crime-on-christmas.

[4] Cjango. Chicago Crime Data Analysis (Python Project). Retrieved October 2020 from https://cjango.wordpress.com/portfolio/chicago-crime-data-analysis-python-project/.

[5] Kaggle. 2019. Chicago Codes - Crime Map and BigQuery in R. Retrieved October 2020 from https://www.kaggle.com/alkadri/chicago-codes-crime-map-and-bigquery-in-r.

[6] Sadaf Tafazoli. 2018. My Notes on Chicago Crime Data Analysis. Retrieved October 2020 from https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20.

[7] Julia Burdick-Will. 2013. School Violent Crime and Academic Achievement in Chicago. *Sociology of Education.* PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831577/.

[8] Chicago Police. Crime Type Categories: Definition and Description. Retrieved from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

[9] Kaggle, Chicago City Police. 2018. Chicago Crime (BigQuery Dataset). Retrieved October 2020 from https://www.kaggle.com/chicago/chicago-crime/metadata.

[10] Federal Reserve Bank of Chicago. 2020. CFNAI Historical (Real-Time) Data. Retrieved October 2020 from https://www.chicagofed.org/research/data/cfnai/historical-data.