Data Mining for Crime Analysis

Julia Holland Chen
Applied Computer Science
University of Colorado Boulder
Boulder, CO USA
julia.chen-1@colorado.edu

ABSTRACT

Although violent crime has generally decreased over the last 20 years,[1] the crime rate remains an important issue for Americans.[2] Available crime data continues to increase, as does our need to process and interpret the large amounts of raw data. Data mining techniques such as association rule mining and decision trees will be used in this study in order to better understand and predict crime incidents in each of the police districts in the City of Chicago.

1 Motivation

Determining when and where crime takes place in our cities is an important topic and relevant to everyone. This study will examine data collected from the City of Chicago. In 2019, Chicago was found to be the 31st most dangerous city in America,[3] which makes it a useful location for crime analysis. Using the dataset, this study will attempt to find patterns in location and time that may help police target their resources where they needed most and lead to a better understanding of crime patterns in the City of Chicago. The dataset spans 22 years, 2001 until September 2023, and features include the date of incident, primary type of crime, description, police district, and location where the incident occurred. This study will focus on the following areas: investigate frequent itemsets for crime incidents, use models to predict if an arrest was made given the other features in the dataset, and investigate types of crime for each police district. The results of this study may be used for strategic planning on how and when to deploy police resources for each district in the City of Chicago.

2 Related Work

Khatun et al[4] studied how data mining may be used by crime investigation agencies to discover relevant precautionary measures from prediction rates. They used several datasets, including one from the City of Chicago. The authors used decision trees, k-nearest neighbors (KNN), and random forest algorithms in their investigation. Forecasting was performed for the most frequently occurring crimes like robbery, assault, and theft and additionally prediction of arrest. While their paper focused on crime forecasting, this study will also test frequent patterns to see if additional data for crime types may be extracted along with decision trees for arrest prediction. This study will also include results on the most important features in determining arrest. Their study would be useful for police to understand when an arrest is likely, given incident features but they did not discuss what features strongly contributed to the arrest prediction. This study's goal is to produce varied data for each of Chicago's police districts to use for resource planning and evaluation of current strengths.

Jantan and Jamil[5] explored the relationship between the category of location and type of crime incidents by applying the Apriori algorithm for association rule mining. They focused on type of crime in specific locations with the goal to produce rules for a pattern visualization system to be used in crime analysis. This study will repeat association rule mining using the Apriori algorithm and additionally include classification using decision trees for analyzing the dataset. This study will incorporate the ideas from Janten and Jamil,[5] expand their work with predictive modeling, and present data for the specific case of the City of Chicago.

3 Proposed Work

3.1 Data Collection

Data was collected from Data.gov and downloaded to personal computer.

3.2 Preprocessing

Dataset cleaning will be performed by first changing each feature to correct type. Most of the features of interest have few NaN values and those objects will be removed. As one of the items of investigation is to determine crime during times of year, new features will be derived for the month, day, and time of the incident. Plots will be generated to examine the dataset for outliers and incorrect data.

3.3 Exploratory Data Analysis

Plots will be generated to investigate multiple items of interest in the dataset. Incidents per year, month, and day for the entire dataset will be examined for each police district. The most common crimes for the entire dataset as well as for each police district will be calculated. While examining the entire dataset is useful for general reporting on crime in the City of Chicago, the data for each police district will attempt to help the individual districts understand their crime patterns.

3.4 Association Rule Mining

Using the Apriori algorithm, this study will first determine frequent itemsets from a subset of the features on the entire dataset. This will investigate relationships between the type of crime, location the incidents, time of incident, and additional details. These rules will help police districts understand frequently occurring features in their crime incidents. The selection of the minimum support will be determined after test runs on smaller portions of the dataset. The goal will be to have a minimum support that gives a good number of frequent itemsets that are not so large the list would be unmanageable. Once the frequent itemsets are determined, strong association rules will be calculated. Not all strong association rules are interesting so the lift will be used to determine if the rule was truly interesting. This study will additionally calculate strong association rules for each district so districts will have information specific to them. Jantan and Jamil[5] also studied the Apriori algorithm for association rule mining, with a focus towards producing a visualization system for police department uses. This study will not explore the creation of a visualization system, but will include plots for better understanding of the results.

3.5 Classification

A decision tree algorithm will be used to predict if there was an arrest made for the incident. This algorithm was chosen as it lends itself to greater interpretability of the results. Prediction of arrest was also investigated in Khatun et al[4] on the Chicago dataset so this study will attempt to replicate, noting that our dataset includes more recent samples and differing features for input data. In addition, although we are using a prediction technique, the goal will be to produce a list of features that were most important in the predictive model. This list will be useful to police districts to understand which features tend to lead to arrests. The entire dataset will be examined and well as models for each of the police districts. The classifier performance will be evaluated by computing a confusion matrix and reviewing the values for accuracy, recall, specificity, and precision. Determining which features of crimes lead to an arrest will be informative to each district to see where they have been successful and where they may need to deploy additional resources.

3.6 District Reports

As the final step of this study, a report specific to each district will be generated to enable police districts to clearly understand crime occurring in their areas and what features best predict if an arrest will be made.

4 Data Set

The dataset was published by the City of Chicago.[6] The dataset includes approximately 7.91M rows and 22 features.

Note that features which include an exact location are shifted slightly from the true location of the incident to protect privacy, but will still fall on the same block.

4.1 Feature Information

- ID – unique identifier, nominal
- Case Number – police department records division number, nominal
- Date – date of incident, ordinal
- Block – partially redacted address of incident, nominal
- IUCR – Illinois uniform crime reporting code, nominal
- Primary Type – primary description of the IUCR code, nominal
- Description – secondary description of the IUCR code, nominal
- Location Description – description of the incident location, nominal
- Arrest – if arrest was made, Boolean
- Domestic – if incident was domestic-related, Boolean
- Beat – beat where the incident occurred, nominal
- District – police district of incident, nominal
- Ward – City Council district of incident, nominal
- Community Area – community area of incident, nominal
- FBI Code – crime classification as outlined in the FBI National Incident-Based Reporting System (NIBRS), nominal
- X Coordinate – x coordinate of incident location, continuous
- Y Coordinate – y coordinate of incident location, continuous
- Year – year of incident, ordinal
- Updated On – date and time record was last updated, ordinal
- Latitude – latitude of incident location, continuous
- Longitude – longitude of incident location, continuous
- Location – latitude and longitude of incident location formatted for maps, nominal

Derived data:

- Month – month of incident, ordinal
- Day – day of incident, ordinal
- Time – time of day, ordinal

5 Evaluation Methods

The evaluation methods will vary depending on the technique used. For association rule mining the results of this study will validate a strong association with a significant correlation between the itemsets. For the classification tasks, a confusion matrix will be calculated so accuracy, recall, specificity, and precision may be reviewed. The results will also be evaluated against the following criteria: easily understood, valid on new or test data, potentially useful, and are able to give new insights into crime patterns in the City of Chicago.

6 Tools

This study will use the Python language and Jupyter Notebook for analysis. Libraries used will include:

- Pandas, a powerful data analysis and manipulation tool[7] will be used to load and process the dataset.
- Numpy, a library for scientific computing[8] will be used to convert dataset into n-dimensional arrays for numerical computations.
- Scikit-learn is a library for predictive data analysis.[9] It will be used for decision tree and k-nearest neighbor algorithms.
- Mlxtend is a library which contains machine learning extensions.[10] It will be used for the Apriori algorithm.
- Matplotlib is a visualization library[11] and will be used for simple visualizations of the dataset.

- NetworkX is a Python package for working with complex networks.[12] It will be used for visualizing association rules.

7 Milestones

7.1 Milestones Completed

1. Data cleaning and transformation.
2. Initial algorithm tests with subset of data.
3. Review of initial tests and modifications if needed for larger run of algorithms.
4. Review of algorithm results on entire dataset.

7.2 Milestones Todo

1. Complete paper write up of final results including district reports.
2. Finalize Jupyter Notebook for project source code review.
3. Complete project presentation video.
4. Create README in GitHub repository with links to video and project paper.

8 Results So Far

8.1 Exploratory data analysis

Plotted below is the number of incidents for each year of the dataset.

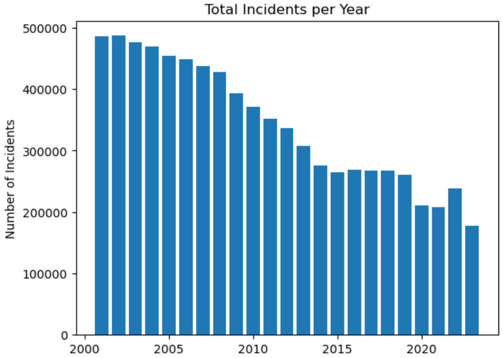

Figure 1: Total incidents reported by year

The number of incidents was steadily decreasing from 2001 with the years 2015-19 stable. 2020-2021 were lower which is understandable as that was during COVID-19 lockdowns. There is an increase in incident count in 2022 as society was resuming normal activities. The year 2023 is a lower level but this dataset only includes through September of that year. Following are plots of incidents per month and day.

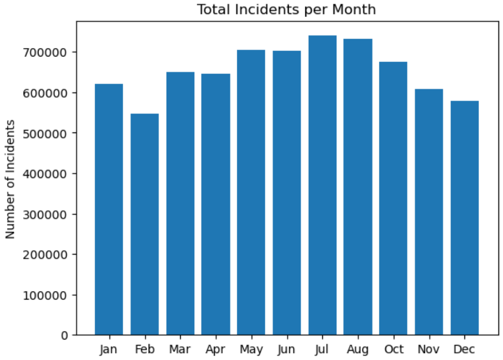

Figure 2: Total incidents reported by month

The number of incidents does not vary widely by month but we can see Jul and Aug are slightly elevated, while Dec and Feb are lower.

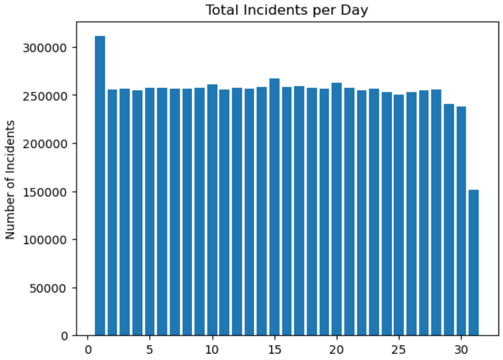

Figure 3: Total incidents reported by day

Interestingly the first of the month is significantly higher in terms of incidents. The 31st day of the month may be lower as not all months have 31 days. Next, the number of incidents reported during different times of day shown.

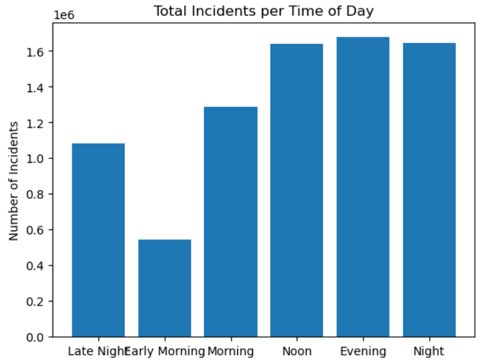

Figure 4: Total incidents per time of day

The lowest time for crime reporting is early morning (4am-8am), while noon through night (12pm-12am) is at a fairly consistent high level. The next plot is the types of crime that were most frequently reported.

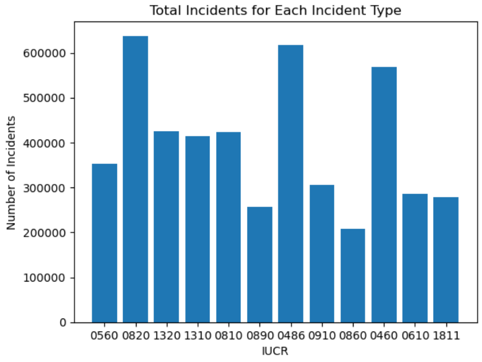

Figure 5: Total incidents reported for each type of incident (IUCR)

IUCR codes where incidents were over 200,000 were used to examine the top IUCR's for the dataset.

Top three IUCR codes are:

- 0820 (Theft – $400 and under)
- 0486 (Battery – Domestic battery simple)
- 0460 (Battery – Simple)

As this study is also concerned with data pertaining to each district, the next plot shows the total number of incidents reported for each district.

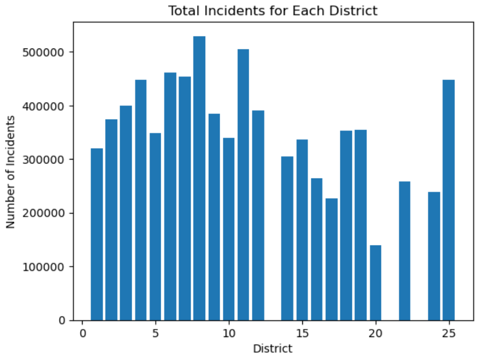

Figure 6: Total incidents for each district

Districts 8 and 11 have the highest number of incidents. District 20 has the lowest number of incidents. There are no incidents reported from districts 13, 21, and 23. According to the Chicago Police Department website, there is no information for 13, 21, or 23 districts suggesting they were closed in the past and absorbed into other districts. Most districts had similar top IUCR codes. District 18 varied more widely and is displayed below.

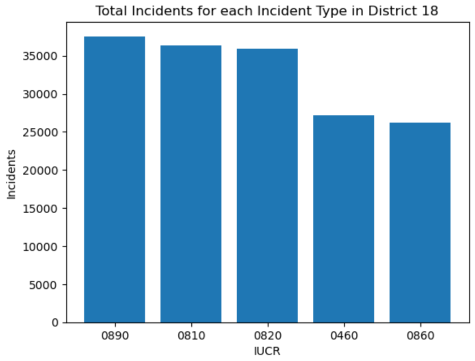

Figure 7: Total incidents reported in District 18 for each type of incident (IUCR)

Top three IUCR codes for District 18 are:

- 0890 (Theft – From building)
- 0810 (Theft – Over $500)
- 0820 (Theft – $400 and under)

Plots for each district were repeated for incidents per month, day, and time of day. They were fairly consistent with the entire dataset plots, although District 1 has noon leading with night at a lower percentage than other districts. Information for each district will be fully examined in the district reports section.

8.2 Association Rule Mining

Frequent itemsets and strong rules were calculated using the Apriori algorithm for the entire dataset using the features (primary type, location description, domestic, arrest, and time). They were then updated to remove any rules that had a lift value of one or less. The minimum support was 0.06 and minimum confidence is 0.05. Twelve rules were generated.

Table 1: Association rules for full dataset

	sup	con	lift
(Domestic)=>(Type_BATTERY)	0.10	0.56	3.1
(Type_BATTERY)=>(Domestic)	0.10	0.56	3.1
(Arrest)=>(Type_NARCOTICS)	0.10	0.36	3.8
(Type_NARCOTICS)=>(Arrest)	0.10	1.0	3.8
(Domestic)=>(Location_RESIDENCE)	0.06	0.37	2.2
(Location_RESIDENCE)=>(Domestic)	0.06	0.38	2.2
(Arrest)=>(Location_STREET)	0.07	0.27	1.0
(Location_STREET)=>(Arrest)	0.07	0.27	1.0
(Time_Night)=>(Location_STREET)	0.07	0.34	1.3
(Location_STREET)=>(Time_Night)	0.07	0.27	1.3
(Arrest)=>(Time_Night)	0.07	0.31	1.2
(Time_Night)=>(Arrest)	0.07	0.25	1.2

Examining a plot of these rules shows there are two connected groups.

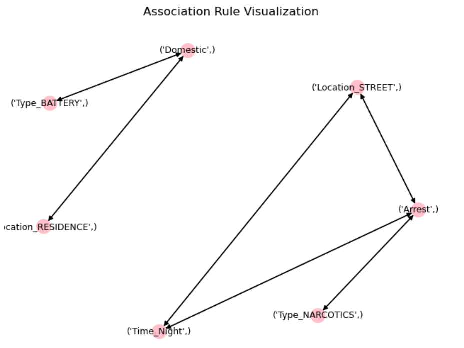

Figure 8: Association rule visualization for dataset

Strong rules were then generated for each district with a minimum support of 0.08 and minimum confidence of 0.07. Rules were removed with a lift less than or equal to one. Plots for each district will be made available in the district reports but since we examined District 18 earlier, that visualization is displayed below as an example of results.

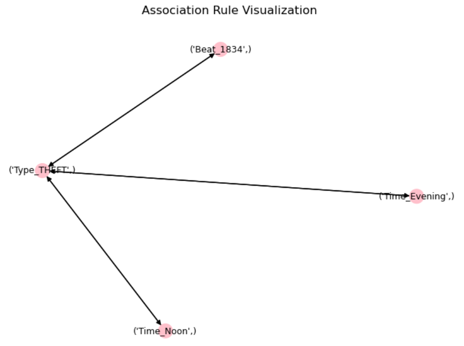

Figure 9: Association rule visualization for district 18

The top center node is Beat_1834, bottom node is Time_Noon, center left node is Type_Theft, and right node is Time_Evening. There are relationships between time of day and theft. Beat 1834 also shows a frequent occurrence with theft. This is one example of how association rule mining may benefit the police districts as police in this area can recognize the type of crime, time of day, and in this case have a more specific location to target for resources (a beat is a smaller area in a district).

8.3 Classification

A decision tree algorithm was used to predict if there was an arrest made for the incident. The dataset does not track if a conviction was successful. The entire dataset was used with the features: primary type, location description, district, domestic, day, and time. The target variable is arrest. First, the features were one-hot encoded as they are categorical variables. 70% of the dataset was used for training; the decision tree classifier used entropy as its criterion and a max depth of 15. The model accuracy was approximately 0.88 which is acceptable. Examining the values for precision, recall, and the

F1 score suggest the model was moderately successful.

Table 2: Performance measures for full dataset

	Precision	recall	F1
0	0.88	0.97	0.92
1	0.87	0.61	0.72

This study is interested in what features were most important in the classification task. For this, feature importance was calculated. Results displayed below.

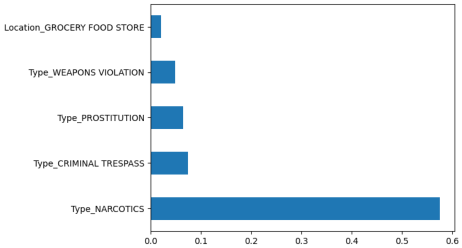

Figure 10: Feature importance

When the type of crime is narcotics, it contributes significantly in determining an arrest outcome. If we look back to our earlier association rules this is consistent with those results. This model appears acceptable but our dataset is imbalanced with respect to the arrest feature. There are approximately 26% of objects where arrest is true. In order to test the decision tree algorithm on a balanced dataset we will oversample the arrest is true objects. Random objects from the arrest false category are selected while using almost all of the arrest true objects. The same features are used, this time with an 80% training set. This model accuracy was approximately 0.79, worse than our first model. Examining the values for precision, recall, and the F1 score we can see that although the values are more balanced, this model does not offer a significantly better outcome.

Table 3: Performance measures for balanced dataset

	Precision	recall	F1
0	0.72	0.94	0.82
1	0.91	0.64	0.75

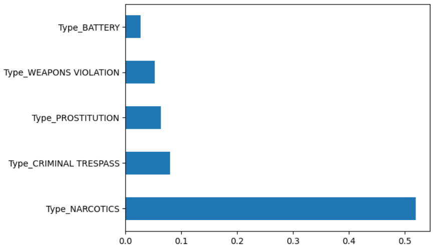

Figure 11: Feature importance for balanced dataset

The features that were the most important in the model using a balanced dataset are similar. The only difference is battery has replaced the location of grocery food store. The next step is to run the classification model for each district. The balanced model will be used in the task. Although the accuracy scores remain between .7-.8, the F1 scores are acceptable. The plots will be included in the reports but as we did before for association rules, I will include District 18 plot for comparison to full dataset. District 18 precision, recall, and F1 scores for balanced dataset below.

Table 4: Performance measures for district 18

	Precision	recall	F1
0	0.77	0.88	0.82
1	0.83	0.69	0.76

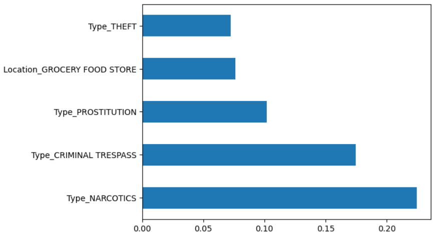

Figure 12: Feature importance for District 18 with balanced dataset

Narcotics still leads the features but it is not as dominant as it was in Fig 11. Weapons violation and battery were dropped and location grocery food store and theft are in their place. As we can see both the full dataset analysis and individual police district analysis are useful.

8.4 District Reports

Reports specific to each district will be outlined below. District 1 is included as example; final paper will include all districts.

District 1

Crime incidents are more likely in July and August, 1st of the month, and noon and evening hours. Crime incidents are least likely in February, 31st day of the month, and early morning hours. The types of crime which are the most frequently reported are theft – $400 and under, theft – from building, and theft – over $500.

Table 5: Association rules for district 1

	sup	con	lift
(Type_THEFT)=>(Time_Evening)	0.12	0.28	1.1
(Time_Evening)=>(Type_THEFT)	0.12	0.49	1.1
(Type_THEFT)=>(Time_Noon)	0.14	0.33	1.2
(Time_Noon)=>(Type_THEFT)	0.14	0.52	1.2

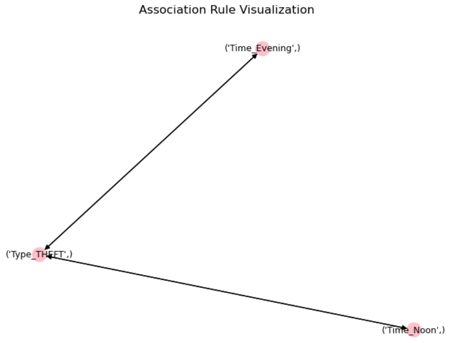

Figure 13: Association rule visualization for district 1

The association rule table and figure show that theft is occurring frequently in the evening and noon hours. In the classification model predicting arrest from incident features, an accuracy of 0.79 was calculated. Additional performance measurements are displayed in Table 6.

Table 6: Performance measures for district 1

	Precision	recall	F1
0	0.76	0.84	0.80
1	0.82	0.75	0.78

The F1 scores are significant and suggest the model was successful in predicting an arrest based on the incident features. The most important features leading to a prediction of arrest displayed in Figure 14.

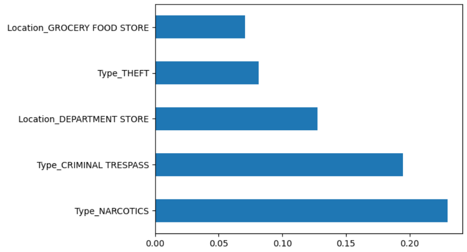

Figure 14: Feature importance for district 1

Narcotics and criminal trespass are the most likely features to determine if an arrest will be made. From the district 1 data, we can see when police should have the most resources available for response. Districts will also know which features are currently most likely to lead to an arrest.

REFERENCES

[1] Crime in the United States. https://en.wikipedia.org/wiki/Crime_in_the_United_States

[2] Domenico Montanaro 2023. Poll: Dangers for both parties on the economy, crime and transgender rights. https://www.npr.org/2023/03/29/1166486046/poll-economy-inflation-transgender-rights-republicans-democrats-biden

[3] Elisha Fieldstadt 2020. The most dangerous cities in America, ranked. https://www.cbsnews.com/pictures/the-most-dangerous-cities-in-america

[4] Most. RokeyaKhatun, SafialIslam Ayon, Md. RahatHossain, Md. JaberAlam, Data mining technique to analyseand predict crime using crime categories and arrest records. Indonesian Journal of Electrical Engineering and Computer Science Vol.22, No.2, May 2021, DOI: 10.11591/ijeecs.v22.i2.pp1052-1060.

[5] Hamidah Jantan, Aina Zalikha Mohd Jamil, Association Rule Mining Based Crime Analysis using Apriori Algorithm. International Journal of Advanced Trends in Computer Science and Engineering Vol.8, No.1.5, 2019. DOI: 10.30534/ijatcse/2019/0581.52019.

[6] Crimes - 2001 to Present. https://catalog.data.gov/dataset/crimes-2001-to-present

[7] Pandas Library. https://pandas.pydata.org

[8] NumPy Library. https://numpy.org

[9] Scikit-learn Library. https://scikit-learn.org/stable/

[10] Mlxtend Library. https://rasbt.github.io/mlxtend/

[11] Matplotlib Library. https://matplotlib.org/

[12] NetworkX Library. https://networkx.org/