Analysis of the Chicago Crimes

Danielle Líbano
Computer Science Master's Degree
Universidade do Porto
up202004103@up.pt

Hugo Valdrez
Computer Science Master's Degree
Universidade do Porto
up201704962@up.pt

Luís Pinto
Computer Science Master's Degree
Universidade do Porto
up201704025@up.pt

Vinícius Batista
Computer Science Master's Degree
Universidade do Porto
up202204007@up.pt

***Abstract*—Crime is a challenge that societies face, with its consequences affecting both individuals and communities. The ability to gain insights into the patterns and characteristics of criminal activities is important for the understanding of why it is happening and what can be done. In this report, we present an analysis of the Chicago Crimes dataset, covering the years 2017 to 2023.**

***Index Terms*—Link Analysis, Chicago Crimes, Association Rules, Machine Learning**

I. Introduction

HERE in this report, we present an analysis of the Chicago Crimes dataset, covering the years 2017 to 2023. This dataset wasn't the only one being used but that will be covered later on. This project was divived in 2 parts:

Part I focuses on exploring the dataset using techniques like data cleaning, visualization, association rules, and link analysis. We aim to identify common crime types, discover spatial clusters, and uncover significant associations between different crime situations, with a particular focus on arrests.

In Part II, we employ machine learning to develop predictive models for determining the likelihood of an arrest. We use four different machine learning algorithms to train the classifiers. Throughout the report, we emphasize the importance of feature engineering to enhance the accuracy and reliability of our predictive models by extracting relevant information from the dataset.

II. Data Preparation

A. Data Understanding

We start by loading and doing basic understanding functions with pandas, i.e., *'dtypes', 'info', 'head',* etc. With this, we start to know better our subject of interest and the challenges to come. Like the variety of domains in variables like *'Primary_Type'*, the diversity of texts in *'Description'* and so forth. Although we present the steps in a sequential order, the process often becomes cyclical due to repeated analysis, new discoveries, and assumptions.

B. Data Correction and Cleaning

After getting the basic understanding of our data, we went to the cleaning phase, first analysing how many missing values are and where they are:

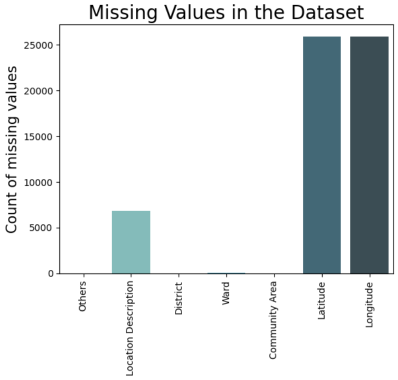


Fig. 1. Missing Values

In addition to the missing values, a set of features were corrected at this stage: District, Ward, Community and Coordinates. Afterwards, we produce a new version of the dataset more fitted to our needs along the way, with no missing values and no incorrect formats that could cause problems.

C. Data Collection

While investigating, we sought to enhance our understanding by exploring additional datasets. Cross-referencing various datasets revealed hidden insights and potential gaps that were not apparent initially. We discovered a valuable resource: https://data.cityofchicago.org/. Our primary objective was to determine whether there were any discernible disparities between socially disadvantaged areas and non-socially disadvantaged areas. Our analysis yielded precisely the insights we sought. By utilizing the Socioeconomically Disadvantaged Areas dataset, we discovered notable discrepancies in the types of crimes occurring within these regions. Specifically, while battery incidents were predominant in the Socioeconomically Disadvantaged Areas, theft overwhelmingly dominated in other districts.

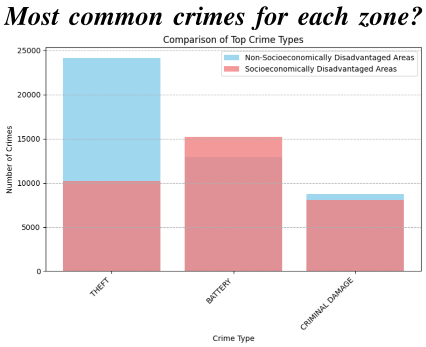


Fig. 2. Most popular crimes in each zone

D. Feature Engineering

A few features were created to help us grasp a more detailed insight over the data and add value in the models later on: *'Hour', DayOfWeek', 'periodOfDay', 'YearMonth', 'Month'* and the conversion of the existing *'Date'* to a better format.

III. Data Analysis

A. Data Visualization

Here, we formulated a series of questions and produced numerous visualizations in an attempt to answer those questions.

Due to the large number of plots generated, we will only present the most relevant ones. However, all other plots can be found in the accompanying notebook.

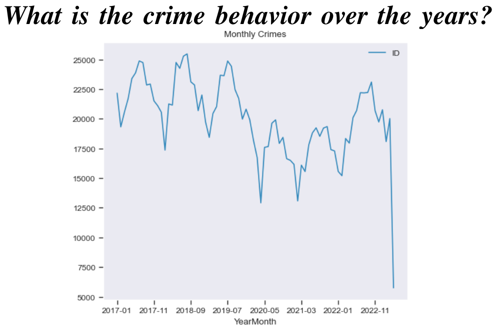


Fig. 3. Total Crimes by Year

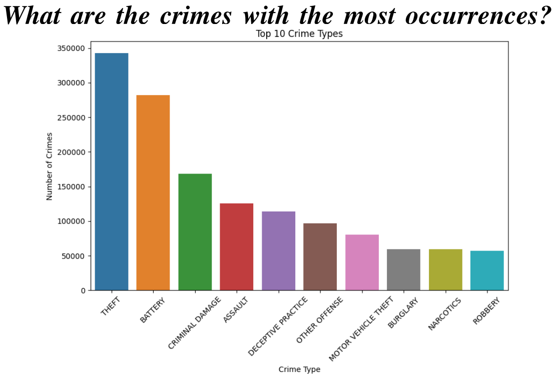


Fig. 4. Ranked Total Crimes by Type

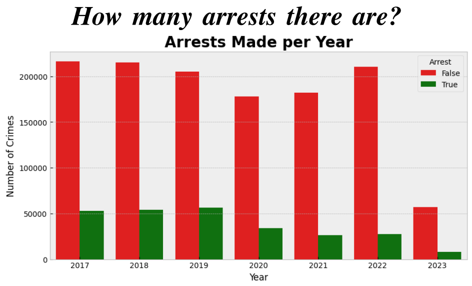


Fig. 5. Arrests by Year

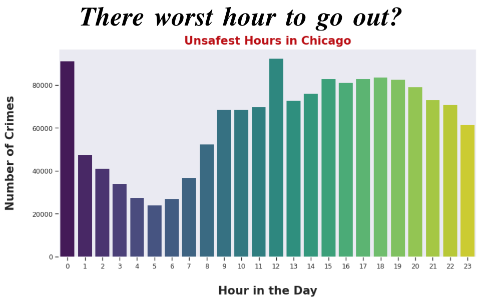


Fig. 6. Total Crimes by Hour

With these analyzes and visualizations, we started to have a better understanding of the object of study and were able to start carrying out the next studies and model creations.

B. Descriptive Modelling

We used the KMeans model to identify distinct crime clusters and differentiate zones based on crime types. Our chosen parameters for the model were 'Domestic', 'periodOfDay', 'Month', 'DayOfWeek', 'Beat', and 'Primary Type'. After applying the elbow method, we determined the optimal number of clusters and trained the model.

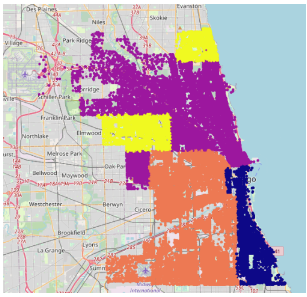


Fig. 7. Clusters from KMeans

However, the results differed from our expectations, with only four observed clusters. Due to time constraints, we couldn't explore alternative approaches thoroughly. The dominance of theft and battery crimes across all districts likely influenced the outcome. We hypothesize that excluding these common crime categories could enhance the model's efficiency in distinguishing clusters. Notably, theft was prevalent in three clusters, while battery dominated the remaining cluster. Moreover, the composition of the top crime categories varied among the three clusters that shared theft as the primary crime type.

IV. Association Rules

One of the necessary steps was the construction of association rules, here we made use of two algorithms, Apriori and FP-Growth, in both we wanted to evaluate the co-occurrence of crimes, for instance, which crimes appear related to *THEFT* or *BATTERY*? The aim is to identify relationship patterns, which can be used for treatments and prevention actions by government bodies.

Given the size of the dataset, we had to perform this analysis on a very small sample of just 2% to 4%. We carried out successive runs and even with little representation it seemed to us that the behavior was always similar.

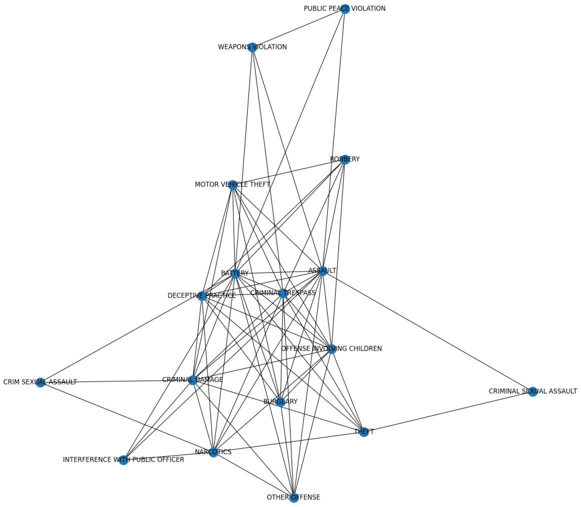


Fig. 8. Graph Network

Analyzing the results we were able to identify a great correlation between *THEFT, BATTERY, CRIMINAL DAMAGE and ASSAULT*, with lift between 0.8 and 1.1 and high support and confidence. Above this lift value are some more complex groups, for example: *ROBBERY, BATTERY, BURGLARY, CRIMINAL TRESPASS* imply *(NARCOTICS, DECEPTIVE PRACTICE, THEFT)* with support of 0.1, confidence of 0.6 and lift of 1.2.

V. Text Processing

In this report, we aimed to address challenges posed by a large and diverse dataset. Various data preprocessing techniques were employed, including cleaning, tokenization, removal of special characters, stop words, and lemmatization. Unique words were selected to reduce the data further.

To generate semantically relevant clusters, we utilized the popular pre-trained GloVe word embedding model [1]. Manual adjustments were made during the clustering process, resulting in 63 clusters for the *Description* attribute and 32 clusters for the *Location* attribute.

For instance, the Description clusters included words like *tar, black, brown, white*, and *tan* in the *synthetic* cluster. The *Location* clusters featured words like *university, grammar, college, and school* in the *college* cluster.

These clusters facilitated classification using a classifier for both *Location* and *Description*. This approach significantly reduced the data while maintaining the inherent meaning of each set of descriptions or locations.

VI. Link Analysis

The goal of this paper is to examine criminal behaviour and characteristics using a thick graph with over 1.5 million nodes and multiple edges. We focused on certain variables such as *Primary Type, Description, Location, Arrest, Domestic, Year, Month, Weekday* and *WeekMonth* to simplify the study.

A. Graph Optimization

To optimize the graph for analysis, we performed the following steps. Firstly, we grouped nodes based on their attributes. For instance, if we had 100 thefts occurring in the evening, we consolidated them into a single edge from *"Thefts"* to *"Evening"* with a weight of 100. We only retained edges with a weight above 5. This approach resulted in a graph with 141 nodes and 5676 edges. The density of the graph is approximately 0.575, indicating a high degree of interrelation between different crime elements.

B. Graph Properties

The diameter of the graph, which represents the longest shortest path between any two nodes, is 2. This suggests that there are relatively short paths connecting different crimes, i.e. the crime attributes are closely related and can influence each other.

C. Node Importance

To determine the influence and importance of nodes within the network, we used Eigenvector centrality and PageRank. The nodes with the highest eigenvector centrality and PageRank scores were *ArrestFalse* and *DomesticFalse* followed by the attributes *residential, Fourth, Night* and *Afternoon*. This implies that most crimes do not result in an arrest or involve domestic incidents. Furthermore, offences tend to occur in residential areas during darker hours, and there is an increase in crime towards the end of the month.

D. Node Connectivity

The Betweenness centrality values indicate how nodes act as bridges or intermediaries between other nodes. The nodes with the highest betweenness centrality include *cemetery, department, WEAPONS VIOLATION, explosive* and *secure*. These nodes play crucial roles in connecting different parts of the graph. For example, crimes occurring near or involving cemeteries may be linked to descriptions such as vandalism, property crimes, or criminal damage. The presence of

WEAPONS VIOLATION, explosive and *secure* in the network suggests a strong association with police activities due to their inclusion in the police force category.

E. Node Closeness

Closeness centrality measures the proximity of a node to others in the network. Nodes with higher closeness centrality scores are considered more central due to their ability to reach other nodes more quickly. In this graph, the nodes *murder* and *bridge* exhibit high closeness centrality. The presence of a bridge near a river suggests that murderers may prefer to commit crimes in such locations, as it provides an easier means to conceal a body.

F. Community Detection

Five communities were identified in the graph. One of particular interest comprises nodes such as *BATTERY, residential, sexual, abuse* and *DomesticTrue*. This suggests that a significant portion of physical abuse crimes occurs within families, as indicated by the presence of *residential* and *DomesticTrue* attributes.

VII. Predictive Modelling

First, we define the choice of models to be used based on experiences from unrelated work, willingness to compare different methodologies and personal preferences. We then have the four selected models: *Logistic Regression*, *XGBoost*, *Random Forest* and *LGBM*. Then we decided to use the variable Arrest as our variable to be predicted, this was due to the fact that the low proportion of arrests in relation to crimes called our attention, except when the crime is ’NARCOTICS’. Other forecasting ideas were related to forecasting crime types by time of day or week, but we didn’t evolve in time.

A. Feature Definition

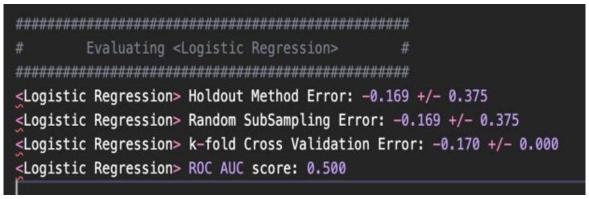
To maintain all models with the same feature set we performed a few validations with manual and automated tests, in the end we choose the features based on: Identification, to help identify the crimes: Primary Type, Description, FBI Code: Location, to pin-point areas and maintain regional behavior: Time, to maintain the relation between the different crimes:

- Identification, to help tipify the crimes: Primary Type, Description and Domestic
- Location, to pin-point areas and maintain regional behavior: Beat, Unadvantage Zone and Location Description
- Time, to maintain the relation between the different crimes: Month, DayofWeek and PeriodOfDay

B. Models

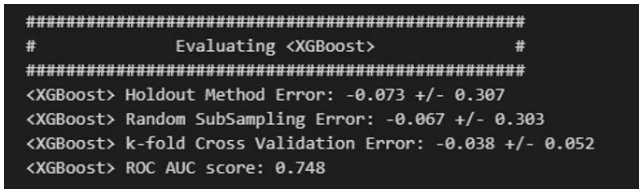
- Logistic Regression
- XGBoost
- Random Forest
- LGBM

The best model in our evaluation was LGBM with 0.748 of AUC. The worst performance was from Logistic Regression with 0.5 of AUC, wich was expected given that it was our base model.



```
################################################
#       Evaluating <Logistic Regression>       #
################################################
<Logistic Regression> Holdout Method Error: -0.169 +/- 0.375
<Logistic Regression> Random SubSampling Error: -0.169 +/- 0.375
<Logistic Regression> k-fold Cross Validation Error: -0.170 +/- 0.000
<Logistic Regression> ROC AUC score: 0.500
```

Fig. 9. Logistic Regression Results



```
################################################
#              Evaluating <XGBoost>            #
################################################
<XGBoost> Holdout Method Error: -0.073 +/- 0.307
<XGBoost> Random SubSampling Error: -0.067 +/- 0.303
<XGBoost> k-fold Cross Validation Error: -0.038 +/- 0.052
<XGBoost> ROC AUC score: 0.748
```

Fig. 10. XGBoost Results

VIII. Conclusion

Our analysis found a constant drop in the number of crimes over the years, accentuated by the pandemic period. This trend continues in 2023, even with months to come we see the general trend continuing. The most common crimes reported were theft and battery, predominantly occurring during nighttime hours, also noted in association rules. Additionally, towards the end of the month, crime rates tended to peak.

We discovered relationships between stealing, battery, criminal damage, and assault, which may be explained by their frequency as the most common crimes. Notably, because of its relationship with numerous types of crimes, a cemetery setting stood out. However, because of the high prevalence of battery and theft occurrences in these regions, residential areas accounted for the majority of recorded crimes.

Our findings emphasize the importance of considering peak crime hours when allocating resources and arranging patrols. Understanding temporal trends for different types of crime can also help in establishing focused crime prevention efforts. Seasonal fluctuations, caused by factors such as weather and vacations, should also be taken into account when putting crime prevention measures in place.

Domestic incidents require focused support systems and preventive measures to address this specific issue. Identifying crime hotspots allows for targeted policing, increased patrols, and community engagement efforts.

References

[1] Pennington, J., Socher, R., & Manning, C. D. (n.d.). GloVe: Global Vectors for Word Representation. Retrieved from nlp.stanford.edu/projects/glove/ .

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

[3] https://data.cityofchicago.org/Community-Economic-Development/Socioeconomically-Disadvantaged-Areas/2ui7-wiq8