Data Mining Project Proposal

Bach Nguyen, Matthew Mayer, Samyak Ghimire, Suyogya Poudel

September 24, 2023

1 Introduction

Our project focuses on data mining the extensive data set of reported Crimes that occurred in Chicago from 2001 to the present day. Our primary goal is to uncover correlations between various factors such as race, age, location, socioeconomic class, and criminal activities. We also aim to understand how the type of crime may correlate to those attributes. By applying advanced data analysis techniques, we aim to extract valuable insights that can inform policy decisions, resource allocation, and community interventions to enhance public safety and address the root causes of crime in the city.

2 Literature Review/References

As our project focuses on updated police records from the city of Chicago our data analysis will fall under the field of crime analysis. Crime analysis has many sub-fields as defined by the Office of Justice Programs under the Department of Justice in the United States. More specifically, our analysis of police department reported crime would fall under Tactical Crime Analysis and Strategic Crime Analysis as defined by Rachel Boba, the Director of the Crime Mapping Laboratory back in 2001. [1] Respectively they are defined as, "The study of recent criminal incidents and potential criminal activity by examining characteristics such as how, when, and where the activity has occurred to assist in problem solving by developing patterns and trends..." and "The study of crime and law enforcement information integrated with socio-demographic and spatial factors to determine long term "patterns" of activity, to assist in problem solving, as well as to research and evaluate responses and procedures." (13) As Kayla Eddy, the Training and Technical Assistance Coordinator at the Bureau of Justice Assistance National Training and Technical Assistance Center has aptly put it, "Law enforcement agencies and jurisdictions across the United States are increasingly recognizing the importance of using crime analysis to disrupt crime patterns in their communities." [2] With this increase of the presence of Crime Analysis units and agencies she has highlighted many of the desired standard products of the field including bulletins covering crime data related to crime locations and trends, top call for service (i.e. location, time of day, days of the week where crime is likely etc.), and offense maps.

With the data we are using from the Chicago PD, we should be able to properly emulate some of the standard products of the crime analysis, as well as discover interesting trends that could be modeled for the class. We look forward to working on such an important topic for our semester project.

Bibliography:

[1]: Boba, Boba 2001, 'Introductory Guide to Crime Analysis and Mapping' *The United States Department of Justice Resource Center*, November 2001, accessed 25 September 2023, https://portal.cops.usdoj.gov/resourcecenter/ric/Publications/cops-w0273-pub.pdf

[2]: Eddy, Kayla June 10 2021, Bureau of Justice National Training and Technical Assistance Center *Checklist to Assess Your Agency's Crime Analysis Capabilities* Retrieved from: https://bjatta.bja.ojp.gov/media/blog/checklist-assess-your-agencys-crime-analysis-capabilities

[3]: Bureau of Justice Statistics, U.S Department of Justice, (n.d)*All Data Analysis Tools* Retrieved from: https://bjs.ojp.gov/data/data-analysis-tools

[4]: Bureau of Justice Statistics, U.S Department of Justice, *Overview of Crime Analysis* Retrieved from:
https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/OverviewofCrimeAnalysis.pdf

[5]: Steven, Raphael, Goldman School of Public Policy, University of California Berkeley, (n.d) *Criminal Justice Data Analysis*

3 Proposed Work

3.1 Preprocessing

Before diving into the analysis, preprocessing of the data is crucial to ensure its quality and relevance. Our preprocessing steps will include:

- **Data Cleaning**: Removing any inconsistencies, duplicates, and irrelevant entries from the dataset.
- **Handling Missing Values**: Using imputation techniques to fill in missing values or removing entries with missing values, depending on the nature and amount of missing data.
- **Feature Engineering**: Creating new features that might be relevant for our analysis and removing redundant or irrelevant features.
- **Normalization and Standardization**: Ensuring that all numerical features have a similar scale to prevent any single feature from dominating the model.

3.2 Strategy of Work

Our intended strategy will be iterative and collaborative:

1. Initial exploration of the dataset to understand its structure and content.
2. Preprocessing the data as outlined above.
3. Exploratory Data Analysis (EDA) to uncover initial insights and patterns.
4. Model development, where we will experiment with various data mining techniques suitable for our dataset and objectives.
5. Evaluation of the models using the metrics defined in the "Evaluation Metrics" section.
6. Refinement of models based on evaluation results and feedback.
7. Final analysis and report writing.

3.3 Source of Data

Our primary data source is the comprehensive dataset of reported crimes in Chicago from 2001 to the present day. This dataset is maintained and updated by the Chicago Police Department. We believe that this dataset, given its extensive nature and granularity, will provide us with the necessary information to achieve our project goals.

3.4 Objective Function and Model Description

Our primary objective is to uncover correlations and patterns in the crime data. To achieve this, we will employ classification models to categorize crimes based on various attributes (e.g., race, age, location, socioeconomic class).

We aim to elucidate:

- The relationship between various demographic factors and the type of crime.
- Temporal and spatial trends in criminal activities.
- Potential predictors of specific types of crimes.

4 Evaluation Metrics

To ensure that our data mining models are both accurate and reliable, we have identified a set of evaluation metrics that are appropriate for our project. These metrics will help us gauge the performance of our models and ensure that our results are robust and meaningful.

1. **Accuracy**: This metric calculates the ratio of correctly predicted instances to the total instances in the dataset. It is given by the formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

2. **Precision**: Precision measures the number of true positive predictions among the positive predictions made. It is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

3. **Recall (Sensitivity)**: Recall calculates the number of true positive predictions among the actual positive instances. It is given by:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

4. **F1-score**: The F1-score is the harmonic mean of precision and recall, providing a balance between the two when their values diverge. It is defined as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: This metric measures the ability of the model to distinguish between the positive and negative classes. A value of 1 indicates perfect classification, while a value of 0.5 suggests that the model is no better than random guessing.

Validation Strategy: To validate our results, we will employ the following strategies:

- **Cross-Validation**: We will use k-fold cross-validation, where the data is divided into k subsets. The model is trained on k-1 of these subsets and tested on the remaining one. This process is repeated k times, with each subset serving as the test set once. This helps in reducing the variance and ensures that the model is not overfitting to a particular subset of the data.

- **Handling Outliers**: Outliers can skew the results and provide a misleading representation of the model's performance. We will use techniques like the IQR (Interquartile Range) method to detect and handle outliers in our dataset.

- **Sample Size Consideration**: Given the extensive dataset, we will ensure that our sample size is large enough to draw statistically significant conclusions. We will also ensure that the sample is representative of the entire dataset.

- **False Positives and Negatives**: By focusing on precision and recall, we aim to minimize the number of false positives and false negatives. This is crucial as misclassifying a crime instance can have significant implications.

5 Milestones

1. **Data Acquisition and Initial Exploration** (Week 5)
    - Obtain the dataset from the Chicago Police Department.
    - Conduct an initial exploration to understand the structure, size, and nature of the data.
2. **Data Preprocessing** (Week 6-7)
    - Clean the dataset by removing inconsistencies and duplicates.
    - Handle missing values through imputation or removal.
    - Normalize and standardize numerical features.
3. **Exploratory Data Analysis** (Week 8-9)
    - Visualize data distributions, correlations, and patterns.
    - Identify potential features and target variables for modeling.
4. **Model Selection and Development** (Week 9-10)
    - Experiment with various data mining techniques.
    - Develop an initial model based on findings from EDA.
5. **Model Evaluation and Refinement** (Week 10-11)
    - Evaluate the model using the defined metrics.
    - Refine and tweak the model based on evaluation results.
6. **Report Writing - Initial Draft** (Week 12)
    - Compile findings, insights, and model details into an initial draft.
    - Share the draft among team members for feedback.
7. **Finalizing Report and Presentation Preparation** (Week 13)
    - Incorporate feedback and finalize the report.
    - Begin preparing slides and materials for the final presentation.
8. **Final Presentation and Project Submission** (Week 14-15)
    - Conduct a rehearsal of the presentation.
    - Present the project findings and model to the class.
    - Submit the final report and project materials.