Concept Paper: Crime Patterns and Predictive Modelling in Chicago City

Abstract

It is important to have a better understanding of crime from a public safety and law enforcement perspective. This paper aims to examine crime trends in Chicago using data from the City of Chicago's Open Data Portal, crime reports from 2001 to 2024. This study will use Exploratory Data Analysis (EDA) and predictive modeling techniques on the data to determine the most frequent crime types, the seasonal variation in crimes and the influences on arrest rates. Also, this paper will examine the application of machine learning models to model crime occurrences. It also presents a review of the literature on the use of law enforcement resources, crime prevention, and the problems with biased predictive policing models.

Introduction

Understanding patterns of criminal activity depends on crime analysis which helps law enforcement agencies both direct their resources effectively and contribute to public policy creation. The City of Chicago will be analyzed for crime trends through this project using data from the City of Chicago's Open Data Portal. Exploratory Data Analysis (EDA) and machine learning techniques will be applied in this study to determine crime trends and examine seasonal patterns before developing a predictive model for future crime occurrences.

The dataset spanning from 2001 until 2024 enables to investigate crime events by both geographic location and time as well as category. This project explores the connections between crime type, location, and probability of arrest to provide meaningful insights about crime prevention and law enforcement tactics.

The Data

This dataset provides a structured framework for the analysis of crime, the identification of trends over time, and the construction of predictive models of future crime occurrences.

Source: City of Chicago Open Data Portal

Dataset: Crimes 2001 to 2024

Timeframe: 2001 to 2024

Number of Observations: 8,144,412 ~8.2 million crime records

Number of Variables: 22 variables

Data Collection Method: The data is collected from police reports which are compiled by the Chicago Police Department (CPD) and are updated on a regular basis.

This dataset is obtained from the City of Chicago Open Data Portal. A description of the key variables included in the dataset is provided below.

Column Name	***Description***
ID	*Unique identifier for each crime record.*
Case Number	*A unique tracking number assigned by the Chicago Police Department.*
Date	*The date when the crime was reported or estimated to have occurred.*
Block	*Partially redacted street address where the crime occurred.*
IUCR	*Illinois Uniform Crime Reporting code, categorizing the type of crime.*
Primary Type	*Broad classification of the crime (e.g., Theft, Assault, Homicide).*
Description	*More specific details about the crime.*
Location Description	*Specifies the crime scene (e.g., Street, Residence, Bank).*
Arrest	*Indicates whether an arrest was made (True/False).*
Domestic	*Whether the incident is classified as domestic-related (True/False).*
Beat	*The smallest police unit assigned to the crime location.*
District	*The police district responsible for handling the incident.*
Ward	*City Council district where the crime occurred.*
Community Area	*One of 77 designated community areas within Chicago.*

FBI Code	*Classification based on the FBI's National Incident-Based Reporting System.*
X Coordinate	*State Plane Illinois East NAD 1983 projection for mapping.*
Y Coordinate	*State Plane Illinois East NAD 1983 projection for mapping.*
Year	*The year the crime occurred.*
Updated On	*The last time the crime report was modified in the database.*
Latitude	*Approximate geographic location of the crime.*
Longitude	*Approximate geographic location of the crime.*
Location	*A formatted field for mapping and spatial analysis.*

Source: *Chicago Data Portal, About Data Section*

Data Cleaning & Preparation

To ensure the dataset is suitable for analysis, the following preprocessing steps were taken:

- Split Date and Time into separate columns for better time-series analysis.
- Handled missing values in fields such as Location Description, District, and Community Area.
- Converted categorical fields into appropriate formats for optimized memory usage.
- Created additional time-based features such as Month, Day of the Week, and Hour to analyze temporal crime patterns.

This data will be used to perform descriptive analysis, predictive modeling, and trend forecasting to better understand crime in Chicago.

Research Questions

1. What are the most frequent crime types in Chicago, and how do they correlate with location (residential vs. commercial areas)?
    - Using Primary Type and Location Description, we analyzed crime frequency across different locations.
    - Found that thefts and narcotics-related crimes are among the most common.

- Insights from this analysis could help optimize law enforcement resource allocation to high-crime areas.

2. Is there any seasonal pattern to violent crimes vs. property crimes?
    - Created Month, DayofWeek, and Hour columns to analyze crime trends over time.
    - Found that crime rates peak during certain months and hours of the day, with the theft and assault incidents increasing in summer.
    - The patterns identified in this research can be used to inform the strategic placement of police personnel at times of heightened criminal activity.

Predictive Modeling

1. Can machine learning models (XGBoost, LSTM) predict hourly/daily crime counts in the City of Chicago?
    - Predictive insights could allow law enforcement agencies to be proactive rather than reactive in crime prevention.
2. What features are most predictive of an arrest?
    - Identifying these predictors can help in policy-making for targeted crime interventions and law enforcement traning.

Discussion, Concerns, and Potential Issues

1. **Data Quality Issues**
    - Missing values in Community Area, Location Description, and Latitude/Longitude were handled by dropping missing values.
    - Data and Time columns were split and formatted correctly.

2. **Ethical Considerations**
    - Predictive policing models must account for biases in historical arrest data.
    - Ensuring data transparency to avoid misinterpretation of crime patterns.
3. **Key Findings from EDA**
    - **Crime types leading to the most arrests:** Weapons Violation, Public Indecency, and Domestic Violence.
    - **Crime types least likely to result in immediate arrests:** Theft, Burglary, and Arson.
    - **Peak crime hours:** Evening hours (6 PM – 11 PM) show the highest crime rates.
    - **Seasonality trends:** Summer months report higher violent crime rates compared to winter months.
4. **Modeling Constraints**
    - Temporal dependencies in crime forecasting need careful handling to ensure model accuracy.
    - Imbalanced crime categories (e.g., homicides are far less frequent than thefts) may affect prediction reliability.

Conclusion

The purpose of this study is both to apply EDA and predictive modeling to the crime trends of Chicago and to identify patterns in the type and location of crime as well as seasonal variation in order to gain insights that might help in crime prevention and in the way, resources are allocated by law enforcement. The final goal is to determine if crime forecasting is

possible in order to develop proactive policing strategies and to assess if it is with the aim to implement such strategies in the future.

We want to provide practical and unbiased recommendations for public safety improvements through our research, while making sure that crime data is used fairly and ethically.