Venkata Sai Charan Lysetty

OO18458

DATA-606 Capstone Project

Unveiling Insights of Chicago Crime

Introduction:

Chicago has been keeping track of its crime by the Chicago Police Department's Bureau of Records, since the early 1900s. Unfortunately, the city's crime rate, especially violent crimes, is higher than the average in the United States. In 2016, Chicago was responsible for almost half of the increase in homicides nationwide, even though the overall crime rates in the country are relatively low. The exact reasons for the high crime numbers in Chicago are not well understood.

What is the issue of interest?

Crime rates have been a big worry in the USA for a long time. For my final project, I picked a dataset that focuses on crimes in Chicago from 2001 until now. What makes this dataset interesting is that it provides real-time information about the crimes happening in the city, and it gets updated every day. I got this data from the official website of the city of Chicago, which has a lot of public datasets available.

Why is this issue important?

I chose this dataset because it has a lot of information that helps us understand and make sense of what's going on. There are many different aspects to look at, and we can discover interesting patterns and insights from it. The reason I picked Chicago is because it's known for having a high crime rate compared to other cities in Illinois. It's important to study and understand these patterns to find ways to make our communities safer.

What questions do I have in mind and would like to answer?

1. I want to figure out which types of crimes happen a lot.

2. I'm curious about what kinds of crimes are considered serious and if people get punished for them.

3. I want to see where in the city crimes happen the most, finding the hotspots.

4. By using an IUCR dataset, I'll try to combine the crime dataset to classify the criminal offenses.

5. I aim to find out which crimes are the most common throughout Chicago.

Data Context:

The dataset I have has a lot of information with 22 columns and a whopping 7.62 million rows, making it quite large (about 1.78GB in size). It covers reported crimes in Chicago from 2001 to present, excluding murders. The data comes from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. To protect victims' privacy, addresses are only given at the block level, keeping specific locations confidential.

One crucial column in the dataset is the IUCR (Illinois Uniform Crime Reporting) Codes. These are four-digit codes used by law enforcement to categorize criminal incidents. In Illinois, the State Police establish these codes, but individual agencies can add more based on their needs. The Chicago Police Department uses over 400 IUCR codes, split into "Index" (common crimes reported nationally) and "Non-Index" offenses (other types of incidents).

The Index offenses, like murder, robbery, and burglary, are part of the Federal Bureau of Investigation's Uniform Crime Reports, tracking crime trends nationwide. Non-Index offenses cover various incidents like vandalism and weapons violations. The dataset is extensive, below is the source link for the dataset.

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data

Unit of Analysis: Crimes, Community, Area, Latitude, Longitude

Variables/measures I plan to use throughout analysis:

- Date - Date on which crime occured.
- Primary Type - Primary type of crime.
- IUCR – Four-digit Illinois Uniform Crime Reporting (IUCR) codes
- Description - Short description of the type of crime
- Location description - Description of where crime occured
- District - District code where crime occured
- Community - Community area code where crime occured
- Longitude & Latitude - Exact coordinates of crime occurance
- FBI Code - numeric code indicating FBI crime categorization
- Year - Year of crime
- Arrest - Indicates whether arrest was made or not

Techniques and Planning:

I want to predict the type of crime using various classification algorithms like Logistic Regression, Decision Tree, Support Vector Machine, KNN Classifier, and Random Forest. KNN is good for real-time predictions, but all features need to be at the same scale, so I'll use standard scaling.

For optimal parameter selection, I'll use grid search. To handle overfitting and variance, I'll create an ensemble model, like Random Forest, which works well with the dataset having many categorical values.

I'll evaluate the models using metrics like confusion matrix, accuracy score, RSME value, and others like precision, recall, and F1.

Lastly, I plan to build a webpage and also show visualizations in Tableau/Power Bi, where you can input a location, and it will provide the crime rate for that area, helpful for people house-hunting in specific places.

Summary/Expected Outcomes:

The main goal is to identify specific Chicago neighborhoods with the highest reported crime numbers within a chosen time frame. The User Interface has tools like a 'date range' selector and a 'Primary Type' filter to narrow down the type of crime. Users can also check if arrests were made or if the crime was domestic. To understand where the crime occurred, there's a 'Location' filter. Clicking on options drills down to more detailed information, including a 'Word Cloud' describing the crime and a 'Heat Map' showing crime density. The Heat Map uses colors and circle sizes to indicate the severity of crime; bigger and darker circles represent higher crime numbers. From a business perspective, tackling crime issues in a specific area begins with analyzing this detailed information on the dashboard.