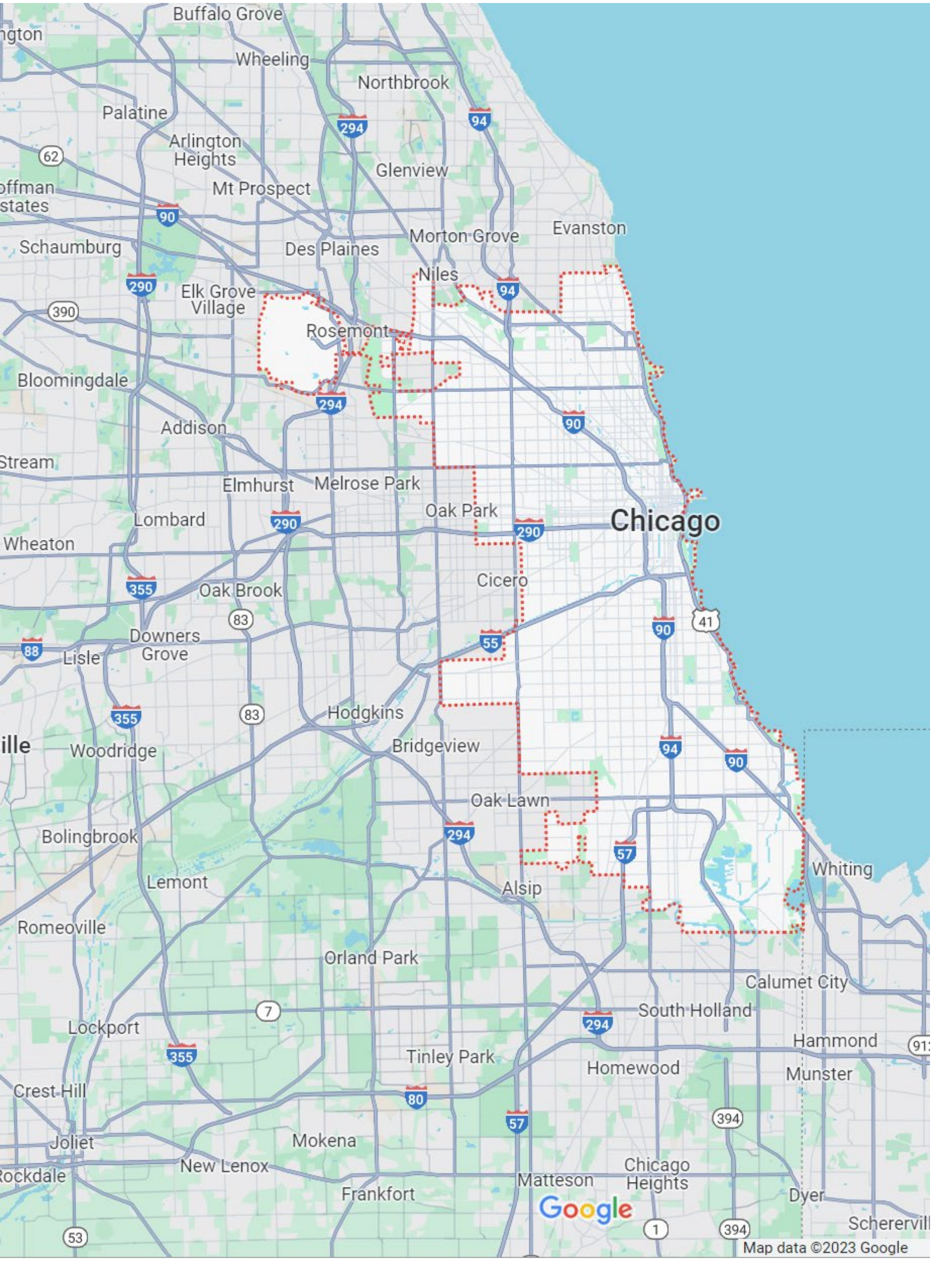


Map data ©2023 Google

DATA MINING FOR CRIME ANALYSIS

JULIA CHEN

MOTIVATION

This study examines crime incident data collected from the City of Chicago from 2001 to September 2023 in an effort to understand and predict crime patterns. The study concentrates on patterns in location, time, and type of crime that may help police target their resources where they are needed most and lead to a better understanding of crime patterns.

Questions

- What details of the crime commonly occur together?
- What factors determine if an arrest will be made?
- What are the most common characteristics of crime in Chicago?

DATA PREPARATION

The crime incident dataset was published by the City of Chicago. The dataset includes approximately 7.91M rows and 22 features. It spans 23 years, 2001 until September 2023.

Preprocessing

- Features converted to correct type in Pandas DataFrame
- Removal of rows with NaN values for features of interest (not many)
- Derive new features of month, day, and time of day
- Removal of rows with invalid district values

TOOLS

- Pandas: import and manipulation of dataset
- Mlxtend: association rule mining (Apriori)
- Scikit-learn: decision tree algorithm, metrics
- Imbalanced-learn: oversampling for classification
- Matplotlib: visualizations
- NetworkX : visualizations of association rules

TECHNIQUES APPLIED

- Exploratory Data Analysis

  Incidents per year, month, and day for the entire dataset were examined and additionally for each police district. The most common crime types in the dataset were found as well as the most common crime types for each district.

- Association Rule Mining

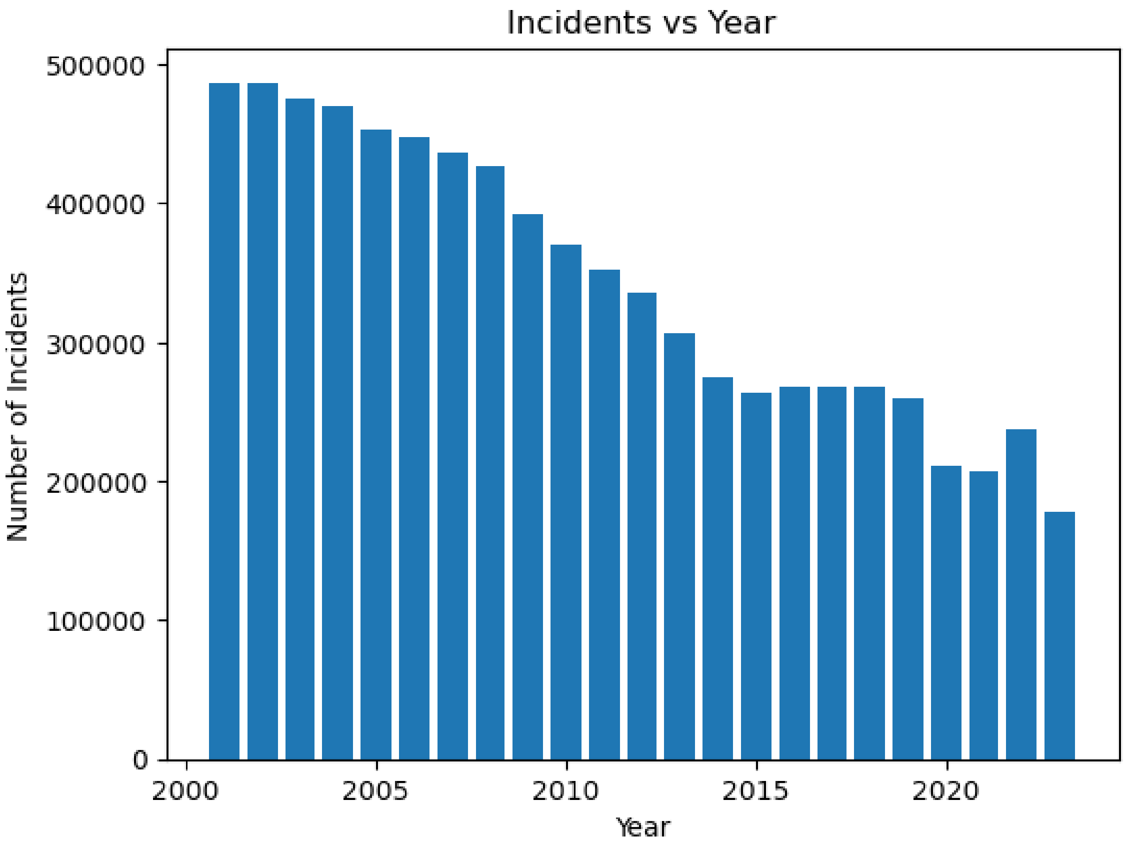
  Investigated relationships between the primary type of crime, location of the incident, if the incident was a domestic-related, whether an arrest was made, and time of incident.

- Classification

  A decision tree algorithm was used for the prediction of arrest, where the two target classes are no arrest and arrest.

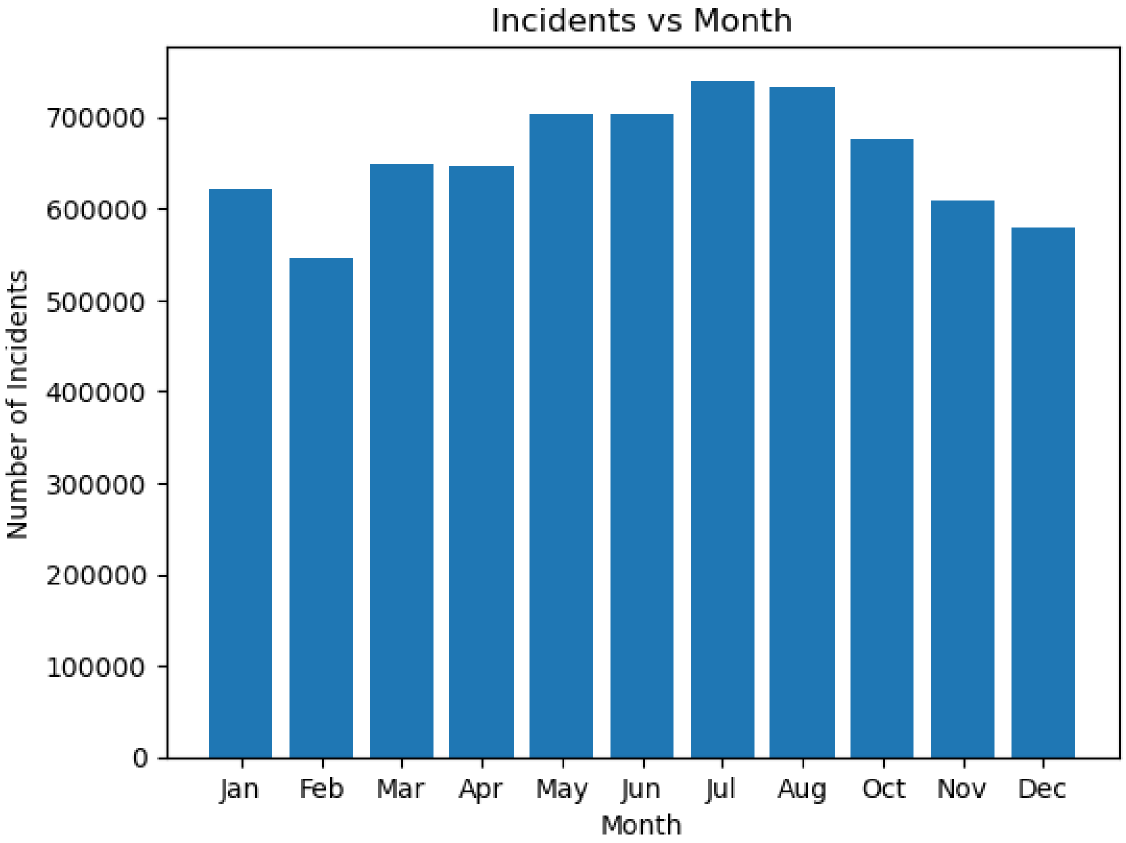
KNOWLEDGE GAINED

Crime has generally been decreasing since 2001



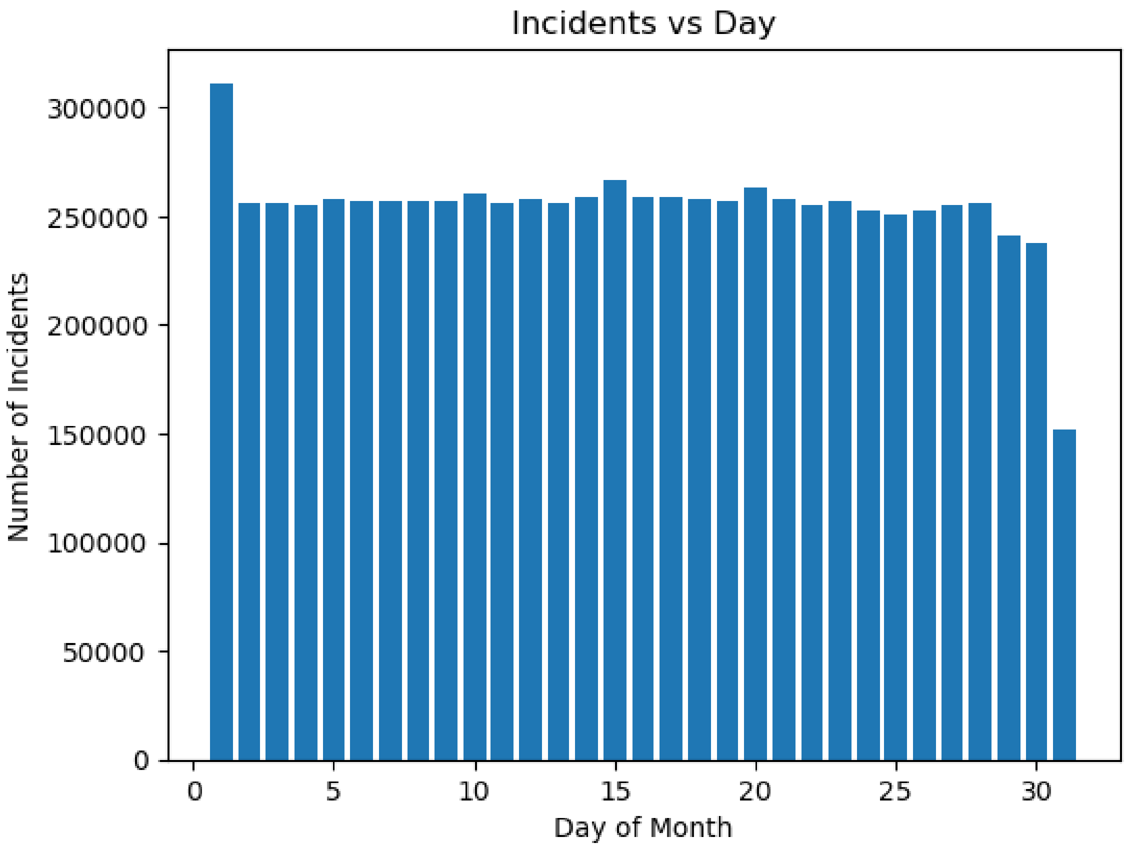
KNOWLEDGE GAINED

Most crime occurs in the summer months



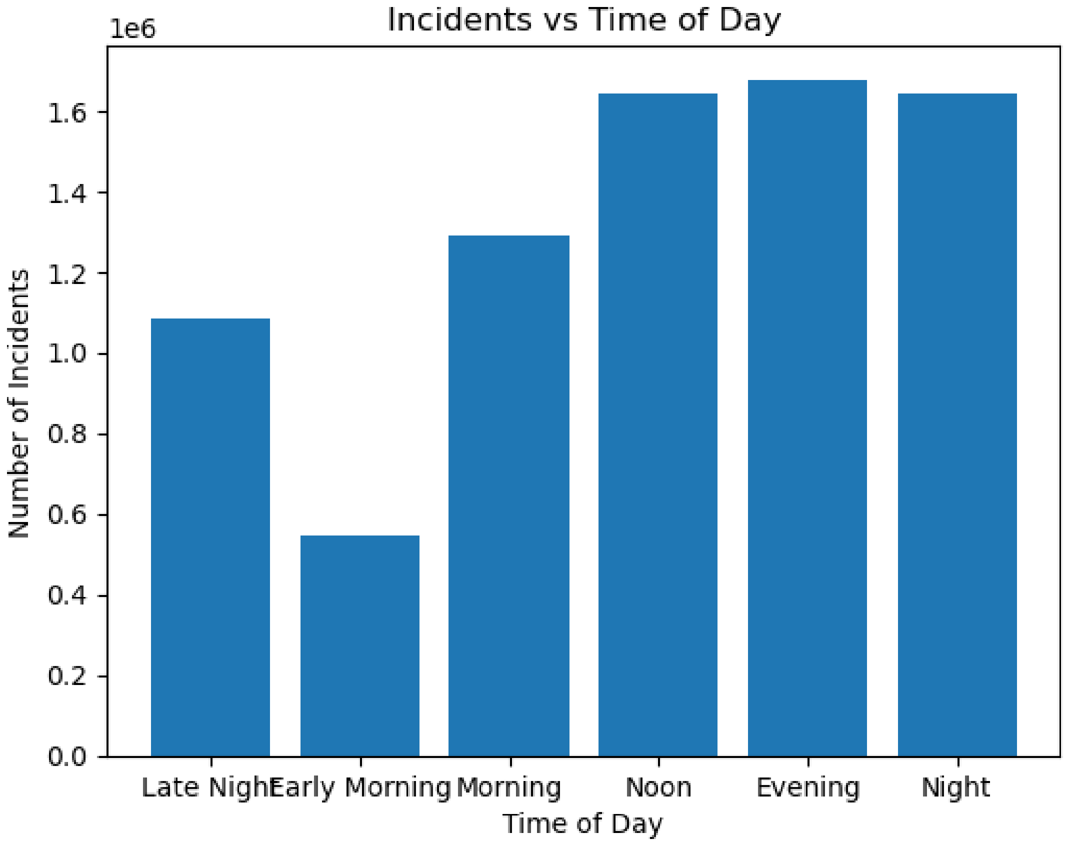
KNOWLEDGE GAINED

Most crime occurs on the first day of the month



KNOWLEDGE GAINED

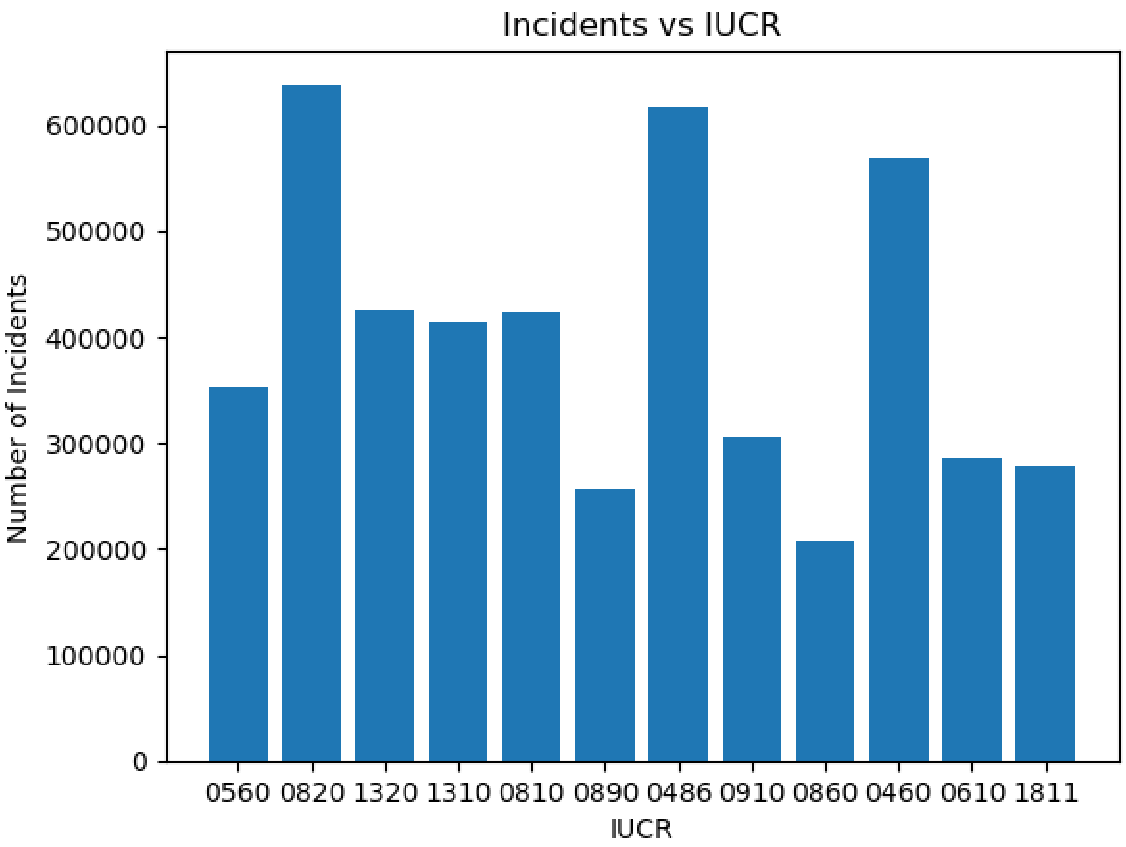
The most common time for crime is noon through midnight



KNOWLEDGE GAINED

The most common crime types are theft and battery

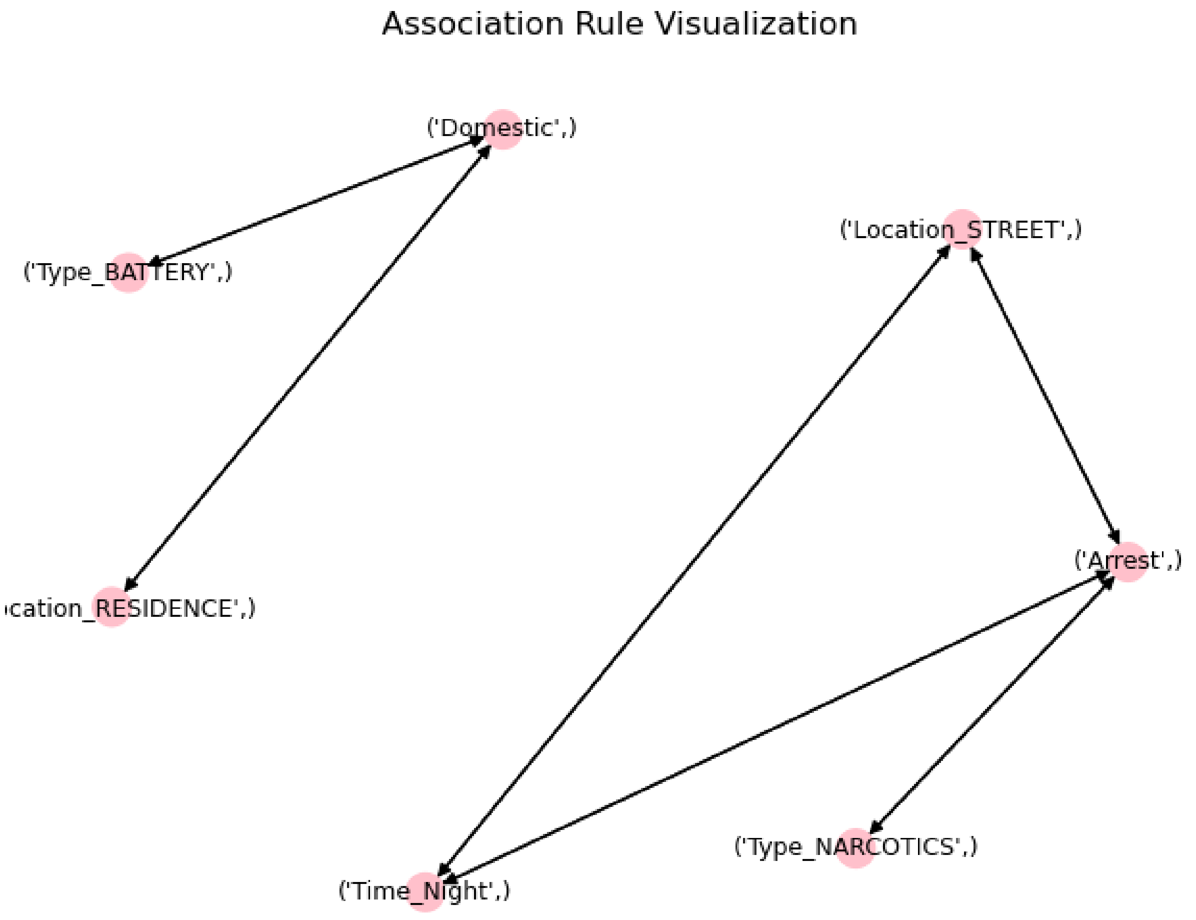
0820 (theft – $400 and under)
0486 (battery – domestic battery simple)
0460 (battery – simple)



KNOWLEDGE GAINED

Association Rules

Minimum support is 0.06 and minimum confidence is 0.05. Rules must have lift greater than one.


Association Rule Visualization

KNOWLEDGE GAINED

Classification

A decision tree algorithm was used to predict if there was an arrest made for the incident.

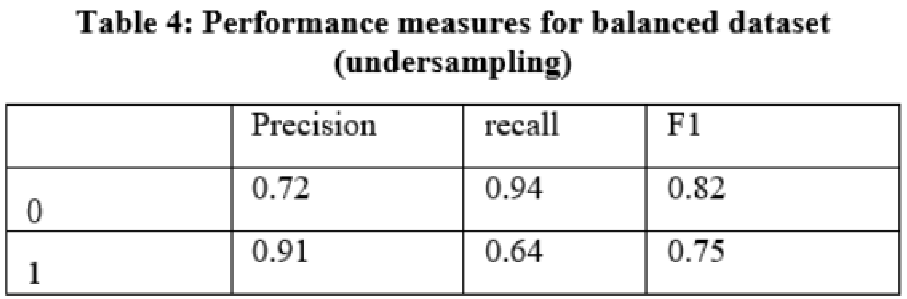
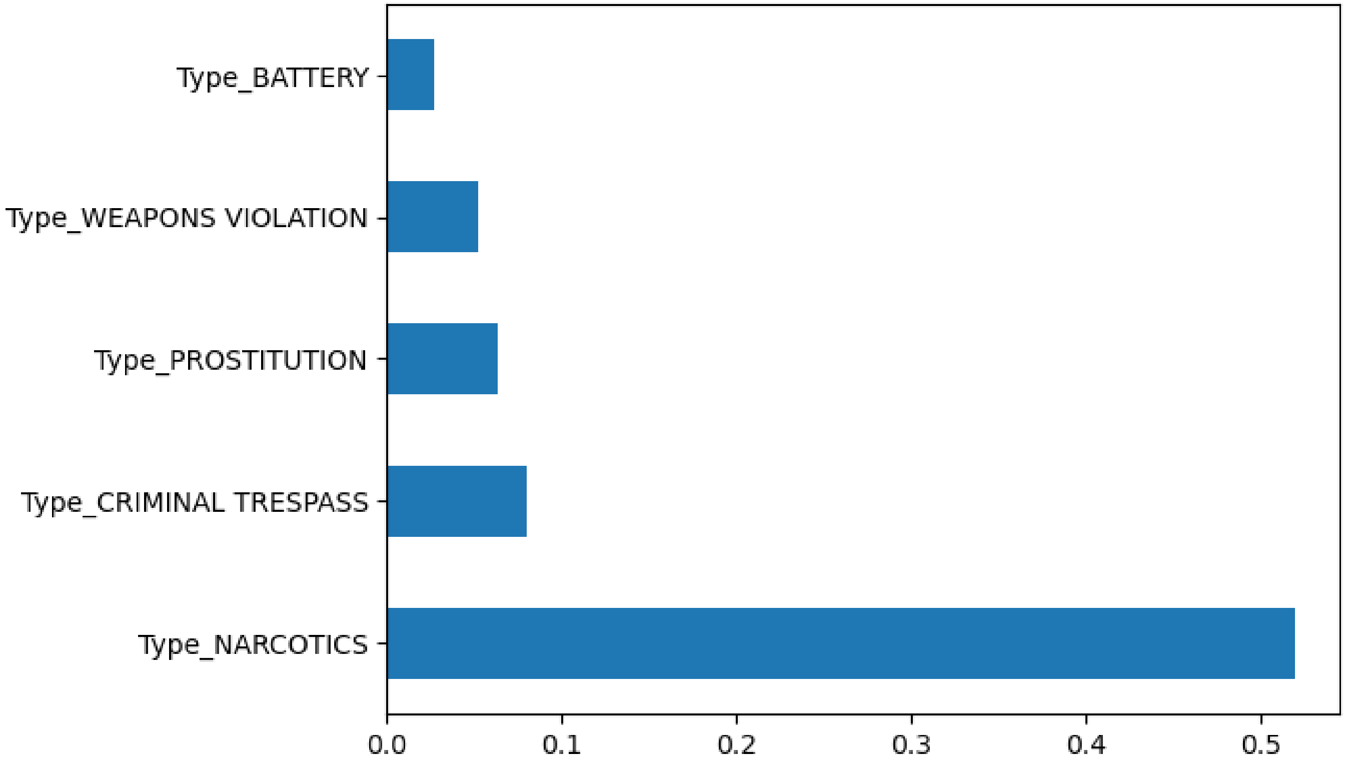


Table 4: Performance measures for balanced dataset (undersampling)

	Precision	recall	F1
0	0.72	0.94	0.82
1	0.91	0.64	0.75

Accuracy 0.79



Feature Importances

KNOWLEDGE GAINED

District Reports

Summary of exploratory data analysis, association rule mining, and classification for each district.

- The individual districts mostly diverge in the most common crime types and association rules

APPLICATIONS

- Inform police districts when and where to deploy resources for crime prevention and response
- Classification results inform what types of incidents have successfully led to an arrest
- Import data mining techniques to a visualization system for easier consumption