CRIME PREDICTION MODEL FOR CHICAGO

MIS-584 Final Report

Ani Chitransh
Department of Data Science
anichitransh@arizona.edu

Ankit Pal
Department of Data Science
apal@arizona.edu

Sarthak Haldar
Department of Data Science
sarthakhaldar@arizona.edu

Nikhil Kumar
Department of MIS
nikhil@arizona.edu

Parth Dattani
Department of MIS
parthdattani@arizona.edu

Nassim Sbai
Department of Data Science
nassimsbai@arizona.edu

Abstract

Crime is an inescapable element of our society. We examined the crime data from the Chicago data portal which contains crime occurrences for Chicago from 2001 to the present. We looked at the patterns of crime throughout time, the most common crime areas, and the hotspots for crime over time. In our experiment, we predicted crimes based on time and location using supervised machine learning approaches.

I. INTRODUCTION

In the sprawling metropolis of Chicago, ensuring the safety and security of its residents is an ongoing challenge. With a diverse range of neighborhoods and communities, allocating law enforcement resources becomes a complex task. Reducing crime rates is not only a matter of public safety but also has profound social, economic, and emotional implications for those impacted by it. The project aims to harness the power of data and advanced machine learning techniques to proactively address this problem and benefit the diverse stakeholders within the city of Chicago.

Why Are We Trying to Solve the Problem?

- Proactive Resource Allocation: With the help of developing a crime prediction model, we can assist in allocating resources more effectively, ensuring that high-risk areas receive the attention they need to deter criminal activity.
- Reducing Response Times: With accurate predictions, law enforcement can respond more rapidly to emerging crime patterns, potentially preventing criminal acts or by resolving them swiftly.
- Enhancing Community Safety: Our model can contribute to lowering crime rates, making our city a safer place for all

Who Will Benefit From It?

- Residents: A reduction in crime rates will lead to safer communities, peace of mind, and an improved quality of life.
- Law Enforcement Agencies: Police departments, emergency services, and other law enforcement agencies will benefit from more accurate and proactive resource allocation. This can lead to more effective crime prevention and resolution.
- Government: The city's government will benefit from this optimized resource allocation, increased public safety, & the potential to specifically allocate funds more efficiently, enhancing its ability to address other pressing issues.

In conclusion, the development of a predictive machine learning model for crime in Chicago represents a significant step towards a safer, more prosperous city.

II. RELATED WORK

[1] Crime pattern detection, analysis & prediction by Sunil Et Al. delves into the dynamic nature of crime patterns and their correlation with social development, emphasizing the challenge of explaining these behaviors. The research employs data analysis algorithms, utilizing the Weka Tool and R Tool, to mine and interpret crime patterns from government-collected data. The four executed algorithms include Association Mining (Apriori), Clustering (k-Means), Classification Techniques (Naive Bayes), and Correlation and regression. The developed model aims to enhance crime detection and analysis, contributing to arresting criminals and implementing measures for crime reduction.

[2] A survey on crime analysis and prediction explores the realm of predictive policing, where statistical data (numerical) is harnessed to guide decision-making in law enforcement. The study investigates various global approaches utilizing deep learning, statistical models, and algorithms to analyze historical crime data for forecasting criminal activities. It overall also emphasizes the complementary role of predictive policing alongside traditional methods, enhancing resource deployment through advanced statistical models. The paper categorizes methods into Neural Network approaches, Statistical approaches, and Spatiotemporal approaches, highlighting their efficacy and comparing them based on accuracy, crime variety, and dataset quality. Conclusions underscore the need for a combination of models, such as neural networks or data mining techniques, for more accurate predictions in crime occurrence analysis.

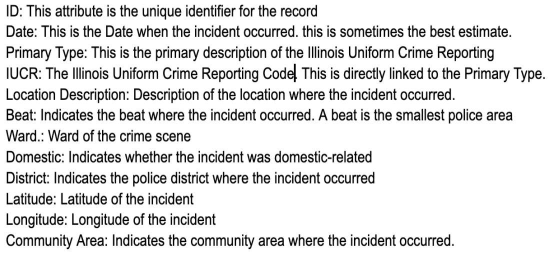
III. DATASET

[3] The Chicago crime dataset extracted from the CLEAR system offers a comprehensive insight into the city's law enforcement landscape, spanning nearly two decades from 2001 to 2019. This rich collection of information serves as a goldmine for crime analysis, providing a multifaceted view of various incidents reported within the city.

At its core, this dataset encapsulates a wide array of crucial details regarding each reported crime. The inclusion of data points such as date, type, and description of the crime provides a fundamental foundation for analysis. The temporal aspect, in particular, allows for the observation of trends, seasonal patterns, and potential shifts in criminal activity over the years. Moreover, the dataset's depth extends beyond basic details, encompassing location-specific important information. This spatial granularity enables a geospatial analysis, potentially uncovering crime hotspots, and distribution patterns across neighborhoods, and areas more prone to certain types of criminal activities.

Some characteristics of the dataset:

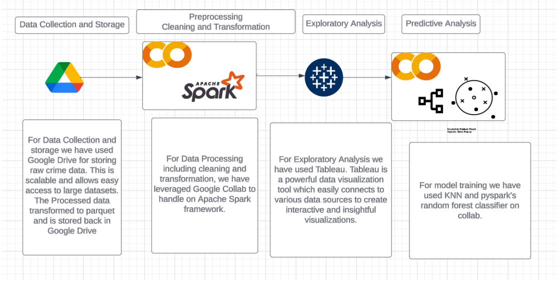
- 7,890,000 - Rows of Data; each row is a reported crime incident
- 22 Attributes which include information such as Date, Type, Description, and location of the crime.
- 11 are categorical, 9 are numerical and 2 are binary variables
- Some important attributes:



ID: This attribute is the unique identifier for the record
Date: This is the Date when the incident occurred. this is sometimes the best estimate.
Primary Type: This is the primary description of the Illinois Uniform Crime Reporting
IUCR: The Illinois Uniform Crime Reporting Code. This is directly linked to the Primary Type.
Location Description: Description of the location where the incident occurred.
Beat: Indicates the beat where the incident occurred. A beat is the smallest police area
Ward.: Ward of the crime scene
Domestic: Indicates whether the incident was domestic-related
District: Indicates the police district where the incident occurred
Latitude: Latitude of the incident
Longitude: Longitude of the incident
Community Area: Indicates the community area where the incident occurred.

The dataset includes enough information about the crime for our analysis.

IV. Methods



For Data Collection and storage we have used Google Drive for storing raw crime data. This is scalable and allows easy access to large datasets. The Processed data transformed to parquet and is stored back in Google Drive

For Data Processing including cleaning and transformation, we have leveraged Google Collab to handle on Apache Spark framework.

For Exploratory Analysis we have used Tableau. Tableau is a powerful data visualization tool which easily connects to various data sources to create interactive and insightful visualizations.

For model training we have used KNN and pyspark's random forest classifier on collab.

Fig: System Architecture

A massive dataset of 7 million crime records necessitates rapid and efficient data processing. Due to its in-memory processing capability, the Spark framework was chosen to handle data of this magnitude. We employed techniques like KNN, Random Forest, Logistic Regression, and Gradient Boost method.

The pipeline followed is discussed below:

A. Data Collection and Storage: We fetched data from the Chicago Police Department's CLEAR website and stored the compressed CSV on Google Drive for efficiency. Since the website was sluggish compared to Drive, it made sense to avoid multiple downloads. Google Collab Platform was our code collaboration space, and Drive served as our data storage. We employed PySpark to read the file, specifying a schema. While CSV files lack inherent schemas, we opted for a defined one, steering clear of potential inaccuracies in inferred schemas.

B. Data Pre-Processing:

To streamline our dataset, we discarded redundant attributes like 'Case Number,' 'FBI Code,' etc., and saved the refined data in parquet format on Google Drive. Parquet's advantages in storage and computation made it a pragmatic choice for further analysis.

Null values from the data were pruned, accounting for approximately 8% of the data. Unwanted rows were filtered out to focus on specific crime categories, and certain crime types were combined for a more balanced dataset. For instance, 'CRIM SEXUAL ASSAULT' and 'PROSTITUTION' were amalgamated into 'SEX OFFENSE.' The modified data was stored back in parquet format on Google Drive.

To enhance our predictive capabilities, we introduced three columns namely ViolentCrime, DayOfWeek, and week column. ViolentCrime is a binary flag that tells whether the reported crime was Violent.Violent/Non-Violent Classification of Crime was done using ChatGPT. These additions were deemed valuable predictors for our model. The transformed dataset was once again stored in parquet format on Google Drive.

C. Exploratory Analysis: For Exploratory Analysis, we have used Tableau. Its seamless connectivity to various data sources enables the creation of interactive and insightful visualizations that facilitate data exploration and pattern identification.

1. Arrest Rate Annual Trends

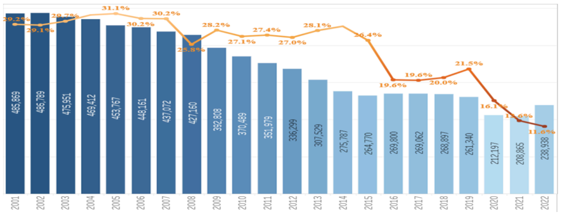


Fig: Domestic Cases Annual Trends

The chart shows yearly variation in reported crime cases with the highest number of cases registered in 2001 and the lowest number of cases in 2021, with the highest arrest rate in 2005 and a notable fall in many cases between 2009 and 2014.

2. Domestic Cases Annual Trends

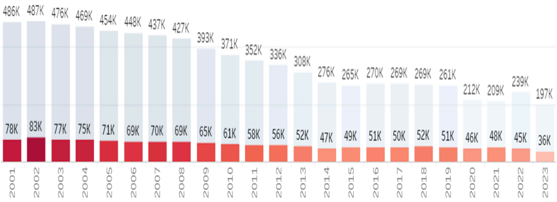


Fig: Yearly trend

The chart shows yearly variation in reported crime and domestic cases with the highest number of domestic cases registered in 2002.

3. Crime Distribution: Violent vs. Non-Violent

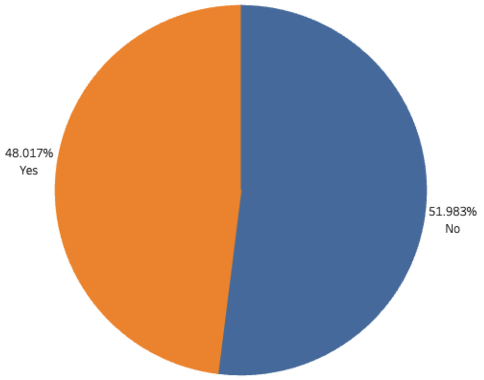


Fig: Crime Distribution

The Pie chart here shows us what percentage of the crimes in Chicago are violent crimes and what percentage are non-violent. We can see from the plot that there are more non-violent crimes than violent crimes. But there is no significant difference between the two.

4. Major Types of Non-Violent Crimes

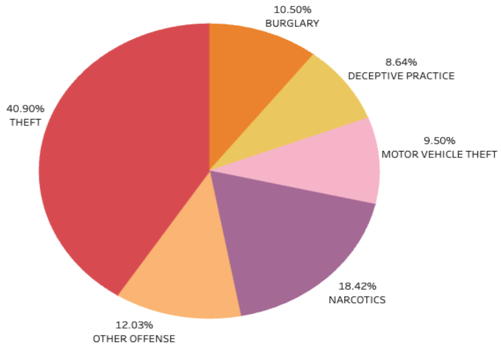


Fig: Major types of Non-Violent Crimes

This pie chart illustrates the distribution of major crimes classified as violent crimes. From the plot, it's clear that Theft is the most common Non-Violent crime.

5. Major Types of Violent Crimes

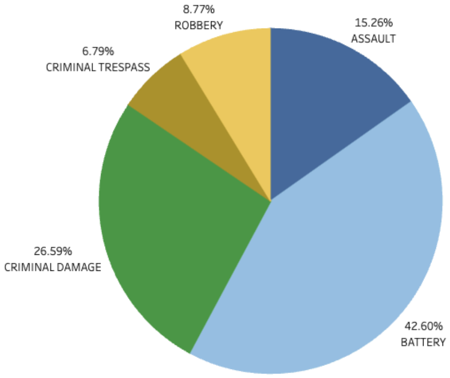


Fig: Major types of Violent Crimes

This pie chart illustrates the distribution of major crimes classified as violent crimes. Battery crime is knowingly or recklessly causing bodily harm to another person. Battery crimes account for 42% of the overall Violent Crimes.

6. Zip Code Violent Crime

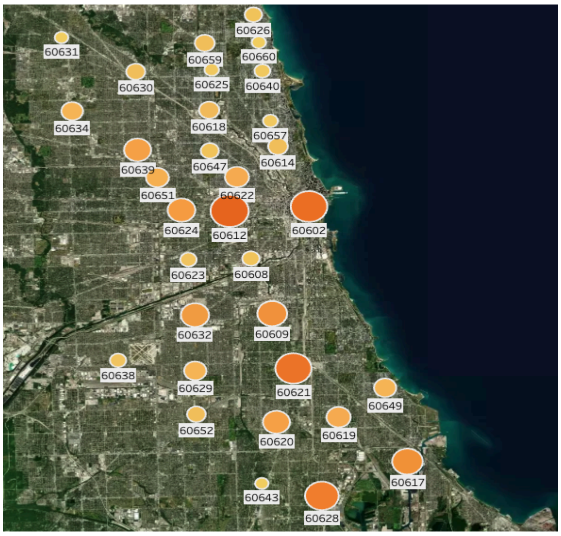


Fig: Zip Code Violent Crime

This map illustrates the prevalence of violent crimes across different ZIP codes in Chicago. Zip code 60612 of Cook County, Chicago accounts for the highest number of Violent crimes.

D. Predictive Models:

We conducted a series of operations to prepare the dataset for predictive modeling. Categorical features underwent encoding using StringIndexer and VectorIndexer, facilitating a smooth integration into machine learning algorithms.

Initially, we tried to predict the crime type from the data but that did not perform well as the data was highly imbalanced. For Example, Our model made 1,331,379 predictions for the label "THEFT," 608,987 predictions for "BATTERY," and 204,640 predictions for "NARCOTICS." The substantially higher count of "THEFT" predictions compared to the other labels suggests a potential class imbalance within the dataset. Then, we changed our approach to binary classification and trained our classifiers to predict violent crime based on location and time.

Correlation Matrix

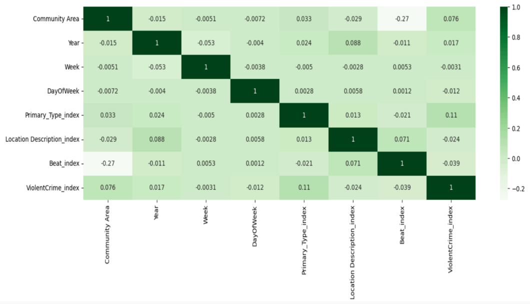


Fig: Correlation Matrix

We used PySpark to generate the correlations between features of the huge dataset. String features were transformed using Pyspark StringIndexer before running the correlations. The plotted results revealed no robust correlations with the response variable, signifying the absence of strong correlations with violent crime.

Our training approach involved data splitting based on two methods. The first was a ratio-based split, with a 90:10 training-to-test ratio. The second was a year-based split, with the 2023 crime data reserved for testing and the rest for training. Irrelevant features were removed from the dataset to enhance the model's performance. To prepare the input data for machine learning algorithms, PySpark's VectorAssembler was used to convert it into a vector format. After the model generated its predictions, PySpark's IndexToString was used to convert the numerical labels back to their original string labels for interpretation. Performance metrics such as F-1 score, precision, recall, and accuracy guided our evaluation.

Employing a Random Forest model, we initially examined the performance of time and location features separately. However, we discovered that combining both yielded superior results.

1. Random Forest Model:

Using both of the train-test split approaches ratio and year, we trained a Random Forest model using time features like DayOfWeek and Week and then location features like Beat (Smallest police geographic area), Location Description, and Community Area. Also, we trained the model on both time and location combined.

2. Logistic Regression Model:

To check the performance of the traditional models, we trained the Logistic Regression model (LR) on both time and location features combined. We used both train-test split approaches in the case of LR as well.

3. Gradient Boosting Tree Model:

To evaluate the other ensemble machine learning models's performance on the dataset, we trained the Gradient Boost Tree (GBT) model on time and location features as well. Similar to the LR model, both train-test split approaches were used.

Further refining our predictive machine learning model, we undertook hyperparameter tuning for the Gradient Boost Tree model (GBT) using 3-fold cross-validation, culminating in our final model. We used 3-fold because 5-fold was taking a lot more processing time.

V. RESULTS

We experienced low correlation features with our predicting variable. We experimented with different features to get better predictions to predict violent crime based on time and location.

Our initial approach of predicting crime type based on location and time resulted in about 32.55% F-1. Changing our approach from predicting crime type to predicting violent crime based on location and time gave us much better results. We think that this approach could be more useful in real life as well because law enforcement agencies can plan better if they know where violent crime or non-violent crime is gonna happen.

Features:

Time:

1. Week - Week of the crime date
2. Day of the Week - Day of the crime

Location:

1. Beat - Smallest police geographic area
2. Location Description - Description of the location where the incident occurred
3. Community Area - This field Indicates the community area where the incident occurred. Chicago State of USA has 77 community areas.

Table I. Performance of modes in case of train test split type Year

Model	Features	Accuracy	F1	Precision	Recall
RF	Time	0.52	0.49	0.51	**0.77**
RF	Location	0.58	0.57	0.56	0.73
LR	Time+Loc	0.53	0.51	0.52	0.71
RF	Time+Loc	0.57	0.56	0.55	0.76
GBT	Time+Loc	**0.61**	**0.61**	**0.60**	0.65

Table II. Performance of modes in case of train test split type Ratio

Model	Features	Accuracy	F1	Precision	Recall
RF	Time	0.53	0.49	0.53	**0.78**
RF	Location	0.58	0.58	0.58	0.70
LR	Time+Loc	0.54	0.53	0.54	0.68
RF	Time+Loc	0.58	0.58	0.59	0.68
GBT	Time+Loc	**0.61**	**0.61**	**0.62**	0.66

The models' performances were evaluated using evaluation metrics like F1 score, precision, recall and accuracy metrics. In both train-test split approaches, we found that time and location features combined give the best prediction results. Traditional models like Logistic Regression fell short in performance. Our models mostly had high recall, low accuracy, low precision, and low f-1 comparatively.

High Recall: The model is good at identifying a large proportion of the actual violent crimes. This means that it is sensitive to the positive class (violent crimes) and tends to minimize false negatives. The model is effective at not missing many actual violent incidents. RF with the only Time features gave the best recall score.

Low Accuracy: Accuracy is the overall correctness of the predictions and is the ratio of correct predictions to the total number of predictions. A lower accuracy suggests that the model is making a significant number of misclassifications, and it's not performing well on the overall dataset.

Low Precision: Precision is the ratio of true positives to the total predicted positives. A lower precision here indicates that the model is making a substantial number of false positive predictions, classifying non-violent instances as violent. This could lead to potential issues, especially if false positives have serious consequences.

Low F1 Score: The F1 score is the harmonic mean of precision and recall. A lower F1 score could be a result of a balance between precision and recall, and it indicates that the model is not performing well in terms of both false positives and false negatives.

The results showed that the GBT model performed the best in both train-test split approaches and gave a good balance of F-1, precision, recall, and F-1 score.

Because of Pyspark, we were able to do all the data manipulations and prediction modeling on this type of big dataset in an easy and faster way. We can easily extend our model to predict violent crime in other cities if we have enough computational resources and historical data. Pyspark can enable us to do distributed computing if we have more computing machines.

VI. CONCLUSION

The creation of a predictive machine learning model for crime in Chicago is a huge step toward making the city safer for its citizens and potential visitors. The model can predict potential future crimes with roughly 60% accuracy one week in advance, and it has the potential to help law enforcement officers reallocate resources, reduce response times, improve community safety, and maximize budget utilization. The model identified high-risk areas using historical crime data, and specifically identified locations that presented safety challenges.

The model has the potential to make Chicago a safer and more affluent city as evidenced by its capacity to proactively identify high-risk locations and contribute to speedier response times. This stands as evidence of how technology can benefit society, by contributing to a safer environment for every member of the community.

We hope that this model will serve as a foundational step for Chicago law enforcement, empowering them to make more informed decisions and strategies related to crimes, ultimately enhancing public safety.

VII. FUTURE WORKS

Many complex aspects go into predicting crime patterns; these include sociological, economic, historical, and geographic considerations. Two elements form the basis of accurate prediction: Excellent data is more crucial than a good model. Despite the size of our dataset, the features it offers are not sufficient.

1. **Enhanced Data Sources:** Integration of additional data sources, such as socioeconomic, demographic, and environmental data, to improve the accuracy and scope of the predictive model.

2. **Geospatial Analysis:** Expansion of the model to incorporate geospatial analysis, enabling more granular predictions and the identification of specific areas where proactive measures can be implemented effectively.

3. **Real-time Monitoring:** Development of capabilities for real-time monitoring and prediction, allowing law enforcement agencies to respond rapidly to emerging crime patterns.

VIII. REFERENCES

1. S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma and N. Yadav, "Crime pattern detection, analysis & prediction," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 225-230.

2. Thomas, A., & Sobhana, N. (2022). A survey on crime analysis and prediction. Materials Today: Proceedings, 58, 310-315.

3.https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2, Crimes - 2001 to Present | City of Chicago