Unveiling Insights of Chicago Crimes

Venkata Sai Charan Lysetty and Bhaavani Madabathula

Department of Data Science, University of Maryland Baltimore County

DATA 606: Capstone Project

Dr. Antonio Diana

May 2024

ABSTRACT

Security is still a major problem in metropolitan settings, which makes advanced technology that can effectively process large amounts of crime data necessary. The design of a Crime Data Information System adapted to satisfy these needs is presented in this study. To build and improve this system, we used a large dataset from Chicago. To clean and prepare the dataset for analysis, a thorough data preprocessing step is required in the first phase. We then used two analytical techniques to locate crime hotspots, identify trends, and project the likelihood of future crimes. These techniques, which each adds in a different way to our knowledge of the dynamics of crime, include statistical modeling and exploratory data analysis (EDA). We validate our results with ground truth data and compare these methods according to how well they perform in real-life situations. This research provides a framework to apply data-driven approaches in crime analysis, in addition to strengthening the strategic capacities of law enforcement.

Keywords—Data preprocessing, Exploratory Data Analysis, crime patterns, statistical modeling, data analytics, Chicago crime data.

INTRODUCTION

Improved techniques of understanding crime trends—especially where they occur—are becoming more and more crucial as worldwide crime rates continue to rise. Everywhere they operate, law enforcement organizations face difficulties with organizing and interpreting vast volumes of crime data that vary according to the location and time of incidents. All of this data needs to be rapidly analyzed, which is why having a flexible system is essential. Making sense of it all and commenting on potential criminal behavior requires the application of techniques such as data organization, which combine similar data into groups and look for patterns.

This indicates the importance of developing more effective techniques for data analysis. We will be able to keep people safe in Chicago by attaining this through improved decision-making, a deeper understanding of crime, and more efficient use of our resources. This is where methods such as data wrangling come in useful. Their collaboration can help us make this large city's crime prevention program better.

We're examining Chicago's crime trends in-depth in our study paper. The most frequent categories of crimes, along with their typical times and locations, are our main concerns. We also monitor the frequency of police arrests for certain offenses. We are doing this by utilizing advanced computer techniques such as Decision Trees, Linear Regression, and K-nearest neighbors. Additionally, we continue to investigate the most serious crimes and their connections. Furthermore, we are identifying the Chicago neighborhoods with the highest concentration of these serious offenses. We intend to develop more effective strategies to deter crime by analyzing all of these strategies

LITERATURE REVIEW

Massive datasets produced by advances in data collecting in recent years have underscored the significance of machine learning and data mining approaches in crime investigation. Holst, A., & Bjurling, B (2013). Because these technologies can detect patterns, trends, and relationships between crimes, they are being used more and more in law enforcement to help lower crime.

The correct parameter selection is critical to the effectiveness of these techniques as it allows law enforcement to more precisely classify and evaluate illegal activity. According to a study conducted in Chicago by Omonigho Edoka, N. (n.d.), machine learning techniques have been applied to classify crime episodes. This study demonstrates how crime analysis can benefit from the use of supervised learning techniques. Malathi and Santhosh Baboo (2011) suggested a better way to predict future crimes using data mining. They highlighted how important it is for law enforcement to use predictive models to be more proactive in preventing crimes before they happen.

Mookiah, Eberle, and Siraj (2015) provided a comprehensive survey of study that explored a variety of crime-related variables but concluded that factors such as age and alcohol have no significant impact on crime rate predictions. While the discussion was thorough, the study lacked a conclusive analysis. Shiju and Surya (2014) used unstructured data from blogs and websites to investigate crime prediction in India using the Naive Bayes and Decision Tree algorithms. Naïve

Bayes performed the best among them, but the study's narrow range of criminal characteristics made for inferior overall accuracy. Varvara and Sergey (2018) investigated the use of gradient boosting, logistic regression, and linear regression to forecast crime types that occur frequently. Their objective was to assess and contrast the three models' efficacy in terms of accuracy predictions.

Malathi and Santhosh Baboo (2011) worked with the MV and Apriori algorithms in particular to improve algorithms to better predict crime in India. Their efforts were directed toward resolving missing values in crime datasets and expediting the processing of crime data. Their conclusions showed that the data mining techniques used may effectively identify trends and forecast upcoming crimes in India. Formal Concept Analysis was the method used by Quist-Aphetsi (2013) to investigate the crime patterns depicted on maps. It was made easier to grasp how different crimes might be related based on where they occur and where crimes tend to occur on a map.

According to Ingilevich and Ivanov (2018), crime rates in cities can be predicted by taking into account factors such as social interactions and environmental context. They demonstrated how determining these social elements is crucial to determining the potential times and locations of crimes. These studies help us understand how crime works in cities like Chicago. They provide different ways to look at crime data and give us ideas on how to predict where crimes might happen like hotspots and also predict the arrest percentage for each crime type.

METHODOLOGY

Figure 1 outlines the research methodology used in our study, which is structured into five key stages: data gathering, data preprocessing, data transformation, data modeling and conversion, and evaluation and results.

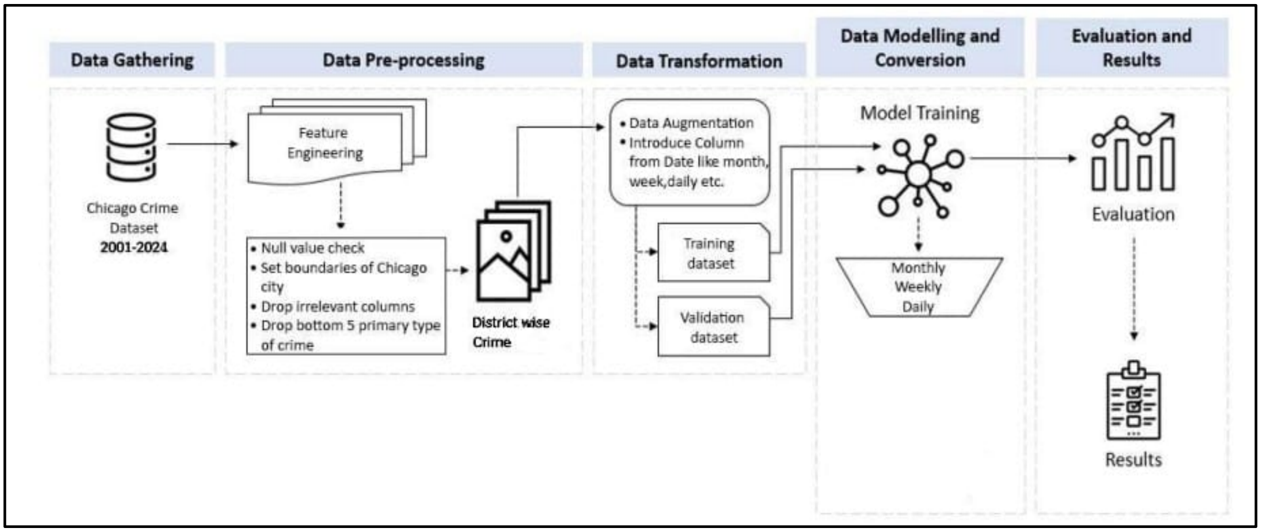

Fig 1: Research Methodology

Data Gathering

The first stage, data gathering involves downloading the Chicago crime incidents dataset from January 2001 to March 13th, 2024, from the Chicago government data portal. Here the process begins with data preparation, where features are selected excluding the target variable 'Arrest'. The dataset is then divided into training and testing sets, ensuring that 20% of the data is reserved for testing to evaluate model performance. For categorical data, a specialized pipeline involving the imputation of missing values and one-hot encoding is implemented to handle preprocessing effectively.

Data Preprocessing

The second stage, data preprocessing involves the removal of irrelevant attributes and crime instances. In the Chicago dataset, crimes were initially detected, and 30,33,173 have been removed for incorrect formatting like missing values, data, fates, etc. The dataset contains 22 attributes like ID, case number, primary type, etc. This study considers only relevant attributes and removes unnecessary columns like ID, updated on location, etc. drop duplicates method is used to check and remove the duplicate rows, and 1.6 million rows were removed. After data pre-processing, this study has 73,60,231 incidents and 16 attributes to be studied.

Data Transformation

The third stage, data transformation involves renaming the column names and creating new columns like month, dayOfWeek, dayOfMonth, and weekOfMonth from the Date of crime incident column. This helps in analyzing past crime patterns. Initially, the Latitude column in the dataset presents as an object but to check the hotspot area, this needs to be changed into float. Using the value counts method, this study checked the top categories of crime type and there are 35 different primary types of crime category. Finally, we analyzed all the primary types and from the count, we can see that a few categories like 'obscenity' have very low crime numbers.

Data Modeling

The fourth stage, data modeling, and conversion involves training of the model, conversion of the model, and model deployment. The models were trained with 80 percent of data and 20 percent of tested data that contains crime incidents from January 2001 to March 2024. For district-wise prediction for a month, week, and day, a list of unique districts is needed.

We constructed Several machine learning models including logistic regression, decision trees, and K-Nearest Neighbors (KNN). Each model is integrated into a pipeline that encompasses both preprocessing and learning algorithms. To optimize the models, a grid search technique is employed to fine-tune hyperparameters such as regularization strength in logistic regression and depth parameters in decision trees.

Model Evaluation

While analyzing the performance of the model, we evaluated using a variety of metrics including accuracy, precision, recall, F1 score, and ROC AUC score. The detailed outputs such as confusion matrices and classification reports provide insights into each model's performance, highlighting the true positive, false positive, true negative, and false negative rates. Additionally, ROC curves are plotted for both training and test data to assess the models' ability to generalize to new data.

Thus, the methodology emphasizes iterative improvement, with multiple rounds of parameter tuning conducted to enhance model accuracy. Cross-validation is utilized throughout to ensure the stability of the models and to reduce the risk of overfitting, thereby improving the reliability of predictions when applied to new datasets. This structured approach ensures the development of robust, well-tuned models capable of performing effectively in real-world applications.

RESULTS:

From Fig. 2, we can see that most crimes happen on weekends, this might be due to increased alcohol consumption, nightlife activities, parties, and social events.

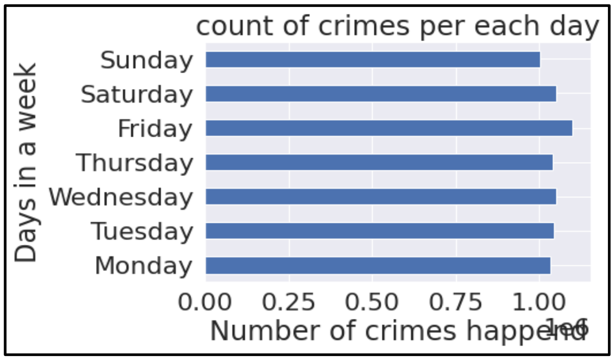

Fig 2: Number of crimes by days of the week

Similarly, from Figure 3, it is apparent that there is no significant difference between crimes occurring in specific months, but February has the lowest number of crimes occurring. One reason for this may be the number of days in that month and also similar to January, February may experience lower crime rates due to cold weather and fewer outdoor activities. However, as the month progresses, there may be slight increases in certain types of crimes as people start to venture out more.

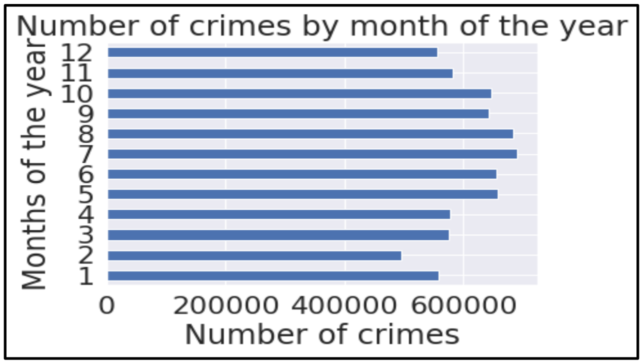

Fig 3: Number of crimes by month

Month by month, followed by Fig.4 a decline later in the year but the overall crime rate is in a downward direction. It is evident that January 2020 to 2021 December has the lowest number of crimes, and this is because of the Covid-19 pandemic period. From Exploratory data analysis, this study noticed that theft, battery, criminal damage, narcotics, and assault are the top five crime categories. Out of all districts, districts 8, 11, 6, and 7 have the maximum crime rate but according to the arrest attribute, people don't get arrested for maximum crime.

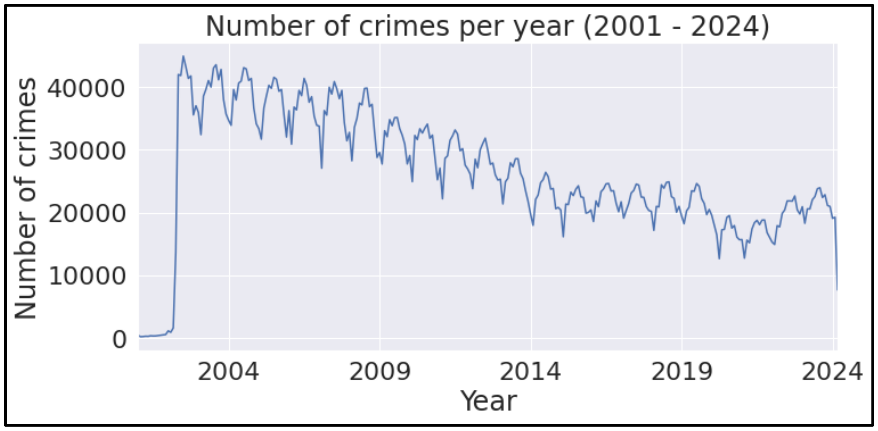

Fig 4: Number of crimes by Year

From the above Fig.5, we see that the number of arrests generally increases with the number of crimes in a certain area. We can also say that it is much more likely for both a crime and an arrest to happen in areas of Chicago like Austin (25) and the Near North Side (8) compared to Edison Park(9) or Burnside(47).

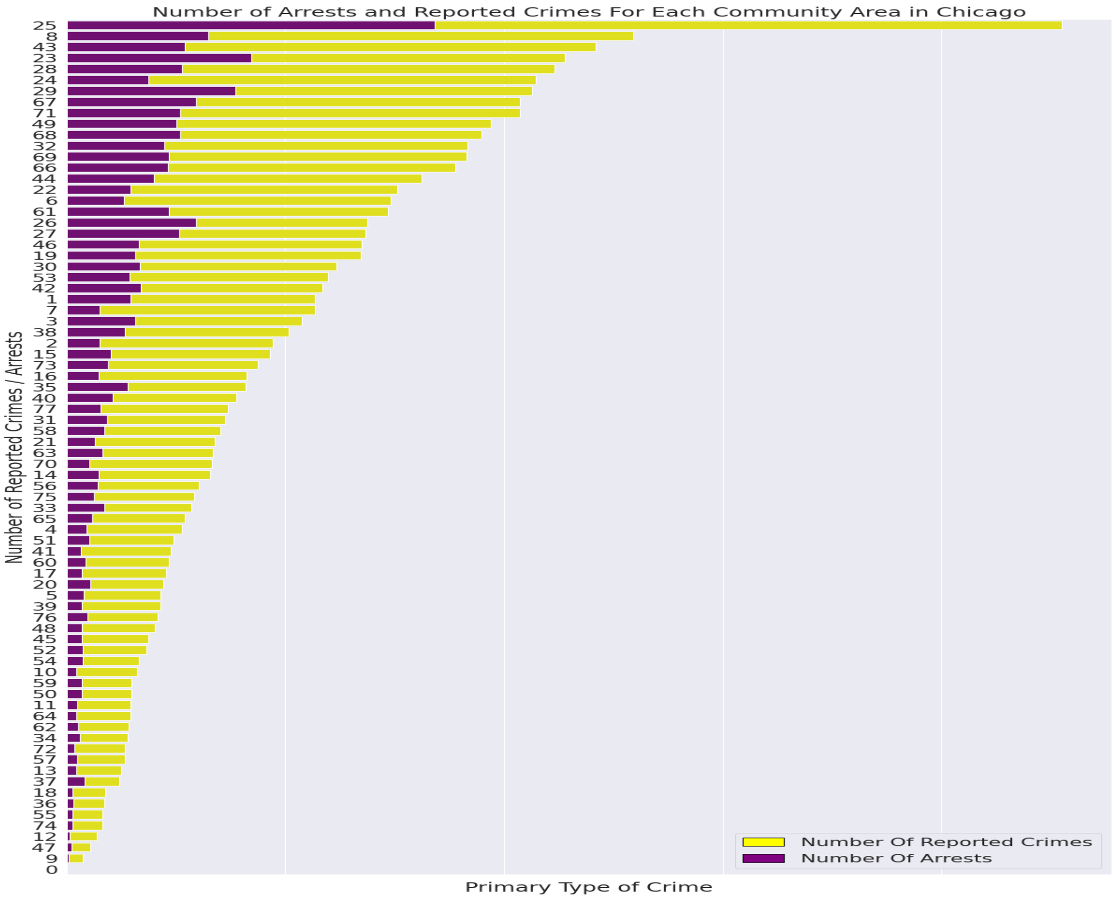

Fig 5: Number of Arrests and Reported Crimes for each Community Area

Our time series analysis offered key insights into crime trends and police activity over the years. By examining changes in crime rates and arrests, we could see how different factors, like policy changes, affected these trends. This information is very useful for policymakers and law enforcement as it helps them identify what works and what doesn't. With this knowledge, they can make better plans and strategies for reducing crime and improving police effectiveness in the future. This work provides a strong foundation for making decisions that help keep communities safer.

From the above Fig 6 and Fig 7, we can see that there is a downward trend in crimes and arrest percentage in Chicago over the last 18 years and this might be due to variations influenced by factors such as changes in law enforcement strategies, economic conditions, social dynamics, and community interventions.

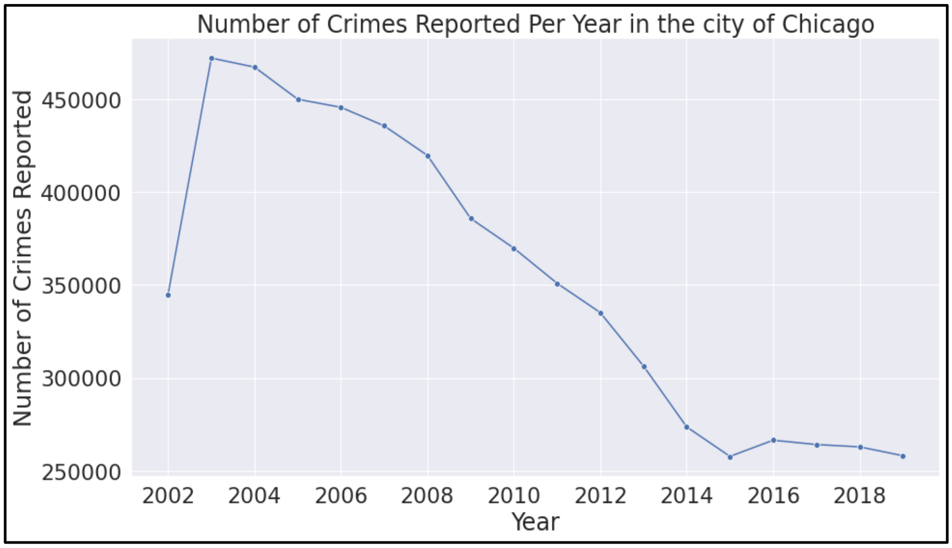

Fig 6: Number of Crimes Reported Per Year

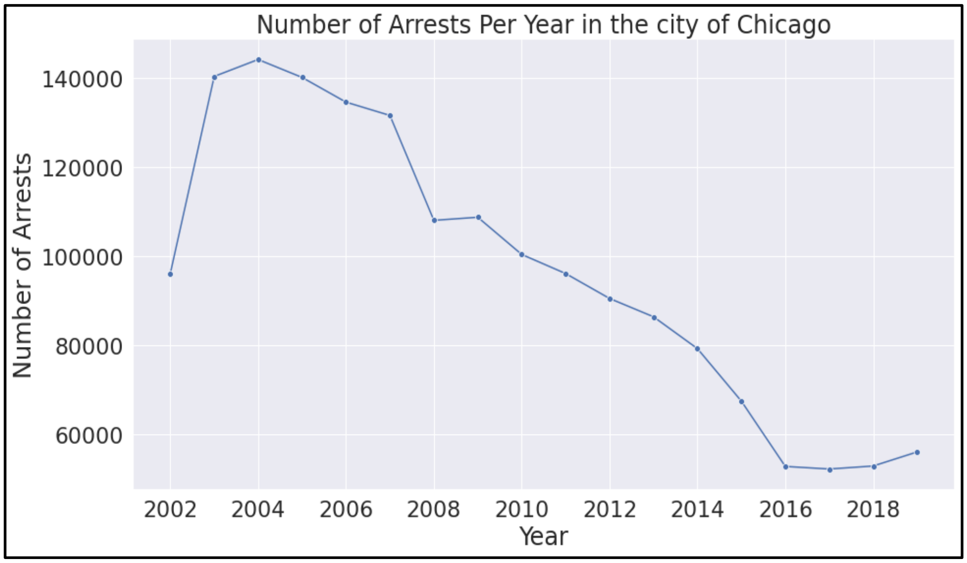

Fig 7: Number of Arrests Per Year

The above Fig 8, shows that arrest rates are notably lower in comparison to Non-Arrest. This variation highlights a significant gap between the number of crimes being reported and the number of arrests being made, suggesting potential challenges in law enforcement's capacity to apprehend suspects or other systemic issues within the criminal justice process. This observation warrants further investigation to identify underlying causes and to enhance the effectiveness of law enforcement responses.

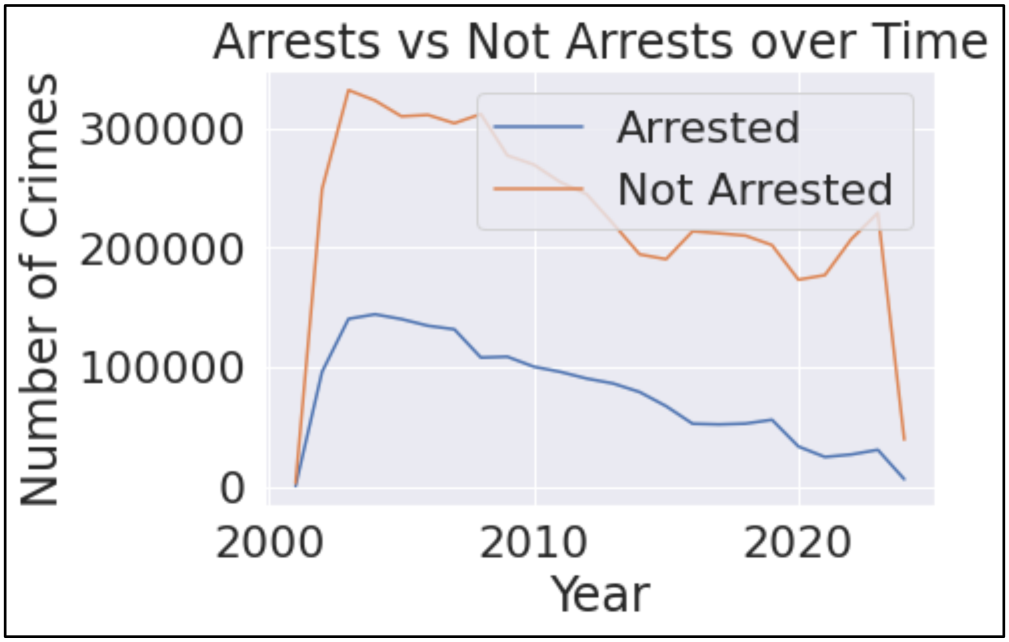

Fig 8: Arrests Vs Not Arrests Over Time

In this research, we analyzed three types of machine learning models: Logistic Regression, Decision Tree, and K-nearest neighbors (KNN). We focused on how accurately these models can predict arrests from crime data by examining their accuracy, precision, and F1 score.

ML Model	Accuracy	Precision	F1 score
Logistic Regression	88.89%	88.14%	70.04%
Decision Tree	88.56%	87.11%	69.48%
K-Nearest Neighbour	86.56%	85.16%	61.94%

Table 1: Machine Learning Model Performances for Arrest Prediction

Table 1 shows that there are far fewer arrests compared to the number of crimes reported. This suggests that there might be problems with how effectively the police are catching suspects or other issues within the criminal justice system. This issue needs more investigation to help improve police responses.

According to below, ROC for our Logistic Regression, Decision Tree, and KNN models. ROC shows us how well our models can tell the difference between true arrests and non-arrests. This helps us fine-tune the models to make them more accurate. By improving our models, we make sure they are useful in real situations, helping the police make better, fairer decisions.

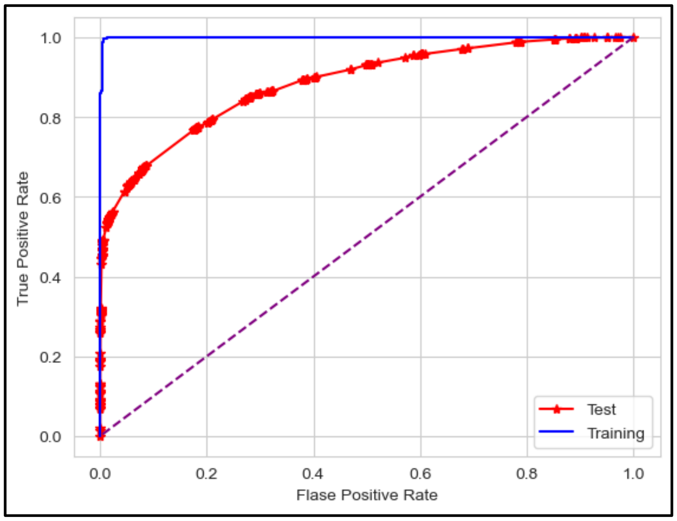

Fig 9 : ROC for Logistic Regression Classifier

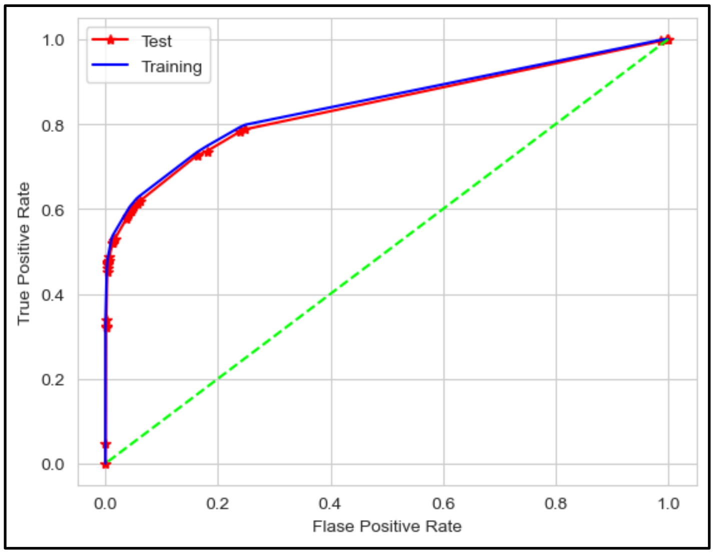

Fig 10 : ROC for Decision Tree Classifier

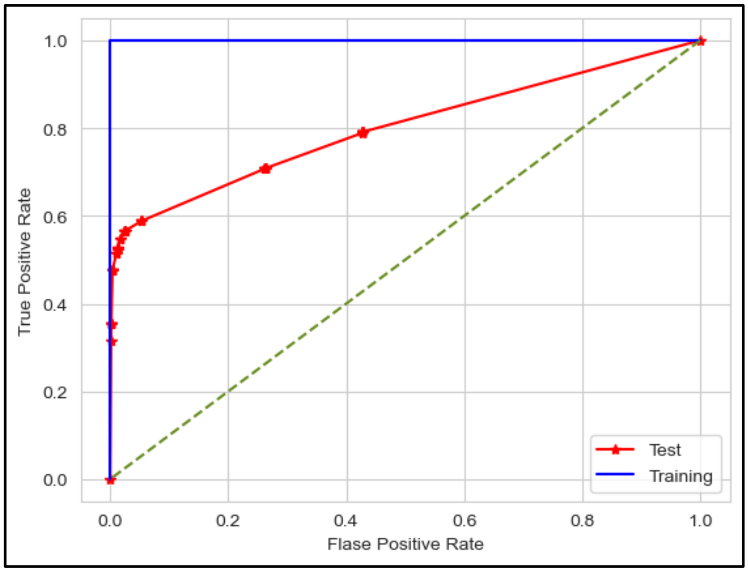

Fig 11 : ROC for K-Nearest Neighbour

MODEL TYPE	ROC CURVE
Linear Regression	89.27%
Decision Tree	85.60%
K-Nearest Neighbors	82.66%

Table 2: Evaluation of ROC Curve Across Different Machine Learning Models

CONCLUSION

In summary, our initiative has made a substantial contribution to the development of data science methods for analyzing and predicting Chicago's crime trends. By combining exploratory data analysis (EDA) with several machine learning methods, such as K-Nearest Neighbors (KNN), Decision Trees, and logistic regression, we have gained a significant understanding of the complex trends and variables affecting crime rates. Our investigation, which covers the years 2001 to 2024, has been successful in identifying crime hotspots, seasonal variations, and the effectiveness of law enforcement tactics.

Although our algorithms are rather good at pattern recognition and crime probability prediction, there are significant shortcomings, especially when it comes to predicting future crime hotspots and dynamically allocating resources based on real-time data. Our analysis of the ROC curves demonstrates our models' legitimacy even more, providing a strong basis for improving prediction accuracy and model robustness in practical applications.

To close these gaps, we will need to work together to improve our models going forward by adding sophisticated algorithms, real-time data feeds, and ongoing data refinement. We may get closer to a more proactive and adaptable strategy to crime prevention and resource allocation by encouraging interdisciplinary cooperation and utilizing emerging technologies. This will ultimately lead to safer communities and increase the effectiveness of law enforcement activities in Chicago.

ACKNOWLEDGEMENT

We are grateful to our Professor Dr. Tony Diana who took out his valuable time to review and provide feedback for our research. We would also like to thank the Chicago Data Portal for making crime datasets available.

REFERENCES:

Holst, A., & Bjurling, B. (2013). A Bayesian Parametric Statistical Anomaly Detection Method for Finding Trends and Patterns in Criminal Behavior. *European Intelligence and Security Informatics Conference*.

https://www.semanticscholar.org/paper/A-Bayesian-Parametric-Statistical-Anomaly-Detection-Holst-Bjurling/ef26da140a484f7bd9d89cb2c31dc154e629881e

***Crime Analysis and Prediction Using Data Mining*. (n.d.). ResearchGate.**

https://www.researchgate.net/publication/280722606_Crime_Analysis_and_Prediction_Using_Data_Mining

Ingilevich, V., & Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*, *136*, 472–478.

https://doi.org/10.1016/j.procs.2018.08.261

Scholarworks, S., & Pradhan, I. (n.d.). *Exploratory Data Analysis And Crime Prediction In San Francisco Recommended Citation*. Retrieved May 4, 2024, from

https://web.archive.org/web/20190427142832id_/https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1633&context=etd_projects

Mookiah, L., Eberle, W., & Siraj, A. (2015, April 6). *Survey of Crime Analysis and Prediction*.SemanticScholar. https://www.semanticscholar.org/paper/Survey-of-Crime-Analysis-and-Prediction-Mookiah-Eberle/e446c14346e1dad9e7c476c66999df9cde5bf3a8

Omonigho Edoka, N. (n.d.). *Crime Incidents Classification Using Supervised Machine Learning Techniques: Chicago MSc Research Project Data Analytics*. https://norma.ncirl.ie/4315/1/nelsonomonighoedoka.pdf

A, Malathi., & Santhosh Baboo, S. (2011). An Enhanced Algorithm to Predict a Future Crime using Data Mining. *International Journal of Computer Applications*, *21*(1), 1–6. https://doi.org/10.5120/2478-3335

G. Borowik, Wawrzyniak, Z., & Cichosz, P. (2018). *Time series analysis for crime forecasting*. 2018 26th International Conference on Systems Engineering (ICSEng). https://www.semanticscholar.org/paper/Time-series-analysis-for-crime-forecasting-Borowik-Wawrzyniak/cfa48b5b16866a1ff2dba6e7bbf89da9b0ab1f77

Yi, F., Yu, Z., Zhuang, F., Zhang, X., & Xiong, H. (2018). *An Integrated Model for Crime Prediction Using Temporal and Spatial Factors*. https://doi.org/10.1109/icdm.2018.00190

Swartz, T. (2015, January 3). *Chicago homicides down citywide in 2014 but some communities see uptick*. Chicago Tribune. https://www.chicagotribune.com/2015/01/03/chicago-homicides-down-citywide-in-2014-but-some-communities-see-uptick/

Davey, M. (2012, June 25). Rate of Killings Rises 38 Percent in Chicago in 2012 (Published 2012). *The New York Times*. https://www.nytimes.com/2012/06/26/us/rate-of-killings-rises-38-percent-in-chicago-in-12.html

Quist-Aphetsi, K. (2013). Visualization and Analysis of Geographical Crime Patterns Using Formal Concept Analysis. https://www.academia.edu/5269677/Visualization_and_Analysis_of_Geographical_Crime_Patterns_Using_Formal_Concept_Analysis