Chicago Crime Data Analysis and Predicition

Kashmira Chawan
21261109
Dublin City University
kashmira.chawan2@mail.dcu.ie

Bhargav Anant Athavale
20210278
Dublin City University
bhargav.athavale2@mail.dcu.ie

***Abstract* - Crime Analysis is the study of various crimes, their patterns, nature, frequency, to understand them, and provide counter measures against them. It allows us to distribute the police resources efficiently. It helps in identifying the criminals and take preventative measures in future to prevent such crimes. The study is also useful for creating new tactics for law enforcement agencies. As such, crime analysis involves a combination of data mining, exploratory analysis to see trends and patterns and understanding of criminal behaviour to perform effectively. This study is conducted on crime in the city of Chicago, Illinois, USA. We perform exploratory analysis on the data to get the different patterns and trends. We perform feature engineering to extract the relevant variables. Based on that, we perform three types of modelling to predict whether an arrest will occur – Logistic Regression, Decision Tree, and Naïve Bayes. The study helps to understand the nature of different crimes and take preventive countermeasures against it in the future.**

Keywords – Crime Analysis, Crime Patterns, Machine Learning, CRISP-DM, Data Mining

I. INTRODUCTION

Criminal analysis has been conducted in some form or another since the advent of human civilization. From ancient times till modern times, before the invention of computers, analysis was performed using rudimentary methods and performed on limited and available data. Since the invention of computers and the progress in information technology, it became possible to store and process data. Due to this, criminal analysis became possible to conduct on a large set of historical data. This helped us to understand the areas affected by crime, the months and seasons of increased crime and types of crimes around major cities around the world. This helped law enforcement agencies to take corrective measures to prevent such crimes.

During the last couple of years due to the advancement in the field of computing and machine learning, it became possible to see the trends, patterns in crimes and develop accurate and predictive models for the crime and allocate sufficient resources for the same. Around the world, it has helped in reducing the criminal activity significantly.

The below study has been conducted on crime data in the city of Chicago, Illinois, USA obtained from the 'Chicago Data Portal'[13] which was brought into law in December 2012. The data is present from the year 2001 onwards. We will be focusing on the data from year 2015 to 2021. Based on the historical data, we will try to understand whether an arrest will be made based on the felony. This will help to develop future models to create the right models to curtail the criminal activity. The allocation of human resources such as police officers, investigators, and material resources such as budgets, computing power can be done accordingly.

The study is organized as follows: in section 2, earlier approaches for solving the problem using various machine learning techniques and data mining methodologies is discussed. In section 3, data mining methodology is discussed to understand its implications on the study and the evolving nature of the data. In the sub sections, the processing of data is done to get the desired output. Data is explored to understand the types of crimes, location of crimes and the trends across days and across months. The techniques used to create the necessary models are discussed. In section 4, comparison is done on the models to check their accuracy and for selecting the best model. Section 5 concludes the study as well as discusses the future improvements which will help to improve the said models and better understand the dataset.

II. RELATED WORK

Several studies have been conducted in the field of criminal analysis on various datasets of several cities. The accuracy of the methodology heavily depends on the attributes from the selected data and the accuracy of the data.

In [1], two different machine learning approaches are discussed, K-nearest neighbour and boosted decision trees for predicting crime on Vancouver city crime dataset. The accuracy of the methods could be improved, fine-tuning it based on the required application. The accuracy of the prediction model was quite low and further analysis is required to improve it.

In [2], the author used WEKA(Waikato Environment for Knowledge Analysis) software to compare different data mining methods and algorithms. Simple Logistic, Logistic, Multilayer Perceptron, Naïve Bayes, Bayes Net, decision trees were compared. Decision trees showed greater accuracy among the chosen.

Four machine learning methods – Random Forest Regressor, Extra Trees Regressor, Decision Tree Regressor and Bagging Regressor are used to analyse the data for the city of Fortaleza, Brazil [3]. Bagging Regressor and Decision Tree Regressor method were considered quite accurate in predicting the crime location. Simple models were found to be more accurate than neural networks.

In [4], the authors use crime pattern analysis to predict the location and type of crime using historical data. Naïve Bayesian Classification was found to be most accurate from among the methods tested. To improve accuracy, machine learning methods were discussed as part of future work.

In [5], data mining procedure was developed to help solve crimes faster. For this purpose, Bayes theorem was used. Apriori algorithm was used to find frequent patterns of the places. Models were also created to predict the location of crime based on day. For future work, to improve accuracy more attributes such as time and more precise location state wise/ region will need to be considered.

In [6], the author makes use of k-means mining algorithm to process criminal data. It is also explained that selecting the clustering centre is important. Improvement in k-means clustering is discussed based on genetic algorithm for efficient criminal analysis.

In [7], the authors discuss outlier detection extensively. Techniques to detect outlier detection are discussed and the advantages(such as credit card fraud detection) and disadvantages of the same are discussed. Four types of outlier detection techniques are discussed namely: distance based, clustering based, density based, and depth based. Its importance in the field of data mining is explained.

In [8], the authors discuss various data mining, prediction and analytical techniques used over the years for crimes against women. The paper discusses various techniques such as Naïve bayes classification, k-means clustering, decision trees, decision trees, association rule mining used by various authors over the years.

Various data mining techniques, tools are used in the study for understanding their uses and advantages. Some of the tools are Z-Crime Tool, ID3 Algorithm, hidden link detection algorithm, Naïve Bayes classifiers and Apriori algorithm. The shortcomings of various tools are discussed. [9]

In [10], the authors discuss the usage of data mining methods for developing proactive police strategies to intercept criminal activities. It will also improve the analytical skills of the police agencies by creating new tools. It is also discussed that further work is needed to automate the data mining process.

In [11], the authors performed text mining to detect criminal activity in social networks. Using a variety of data mining and text mining procedures, models are created and tested with sample data to predict values. Different text mining algorithms are explained. It is also discussed that using advanced tools and using text mining, law enforcement agencies can track and reduce criminal activity and build systems to detect criminals.

In [12], the author discusses about Apriori algorithm and its applications in security. It also discusses that it can work better with more data, as it essentially finds new patterns to detect criminal behaviour. It also discusses the advantages over earlier manual methods of criminal investigation.

III. Methodology

Before we delve into analysis part of the study, we need to first understand the nature of criminal analysis. Unlike other domains, criminal analysis needs to be performed on latest information in this ever-changing world. The criminal activity is frequently changing based on political, social, and economic stability of the region. Therefore, all these factors need to be considered. Also, the data is constantly updating and the patterns and trends in crimes are irregular in nature. For this purpose, CRISP-DM is the best fit. It is similar to other methods in its cyclic nature, but the key difference is that if we feel the data sample taken at the start of the modelling process is insufficient then we can go back to the initial stage without breaking the cycle. Another advantage is that at the initial stage, we focus on our problem/business and can understand it in the full sense for future modelling. This helps in saving time and resources helps us to select the accurate data from the initial stage.

A. Data Pre-processing

The first step is processing of the data. For this purpose and to narrow our scope to recent data we will first select the dataset for the year 2015 to 2021. The dataset available from the portal is updated daily and continuously. For our study the dataset was taken in January 2022. Our dataset contains at the beginning a total of 1,749,922 rows and 23 columns. We check the percentage of missing values and drop the specific rows. We also drop the unnecessary columns which give similar information and not necessary for our study. We are left with a total of 1,719,211 and 13 columns

B. Outlier Check

To check the outlier present and discard them, we identified them through boxplot and removed them using Inter Quartile Range. Box plots apply the IQR approach to display data and outliers (data shape).

Q1 indicates the data's 25th percentile.
The 50th percentile of the data is represented by Q2.
The 75th percentile of the data is represented by Q3.

Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.

The interquartile range (IQR) is the distance between the first and third quartiles (Q1 and Q3): IQR = Q3 – Q1. Outliers are data points that fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR.

We discovered that the Latitude and Longitude columns contain outliers. Longitude values were mostly around 40, but there were a few outliers with values of 37. To get rid of them, we used IQR in the same way discussed above. After making a list of all outliers, we removed those rows from the data frame. The data frame had 17,19,211 rows and after dropping rows there were 17,10,193 rows.

C. Exploratory Analysis

To understand the data visually for patterns and trends, we perform exploratory analysis on it. For this purpose, we make use of matplotlib and seaborn library along with Python and Jupyter Notebook.

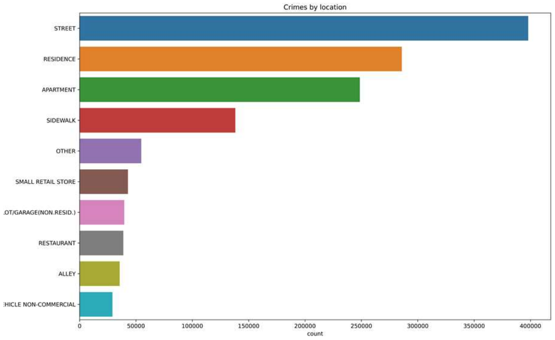


Figure 1: Top 10 crimes by location

For fig 1, we can see the top 10 crimes by location. Highest number of crimes as expected were reported in the street followed by residence.

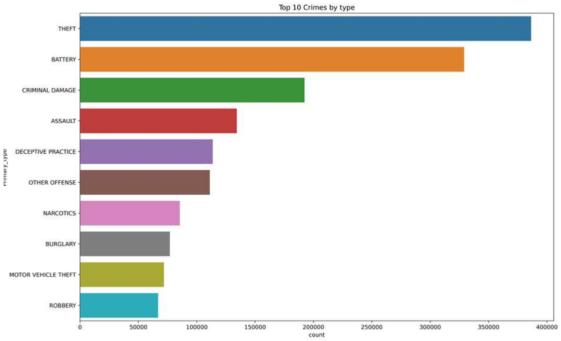


Figure 2: Top 10 crimes by type

For fig 2, we can see the top 10 crimes by type. Highest number of crimes were reported were 'Theft' followed by 'Battery' i.e., physical assault.

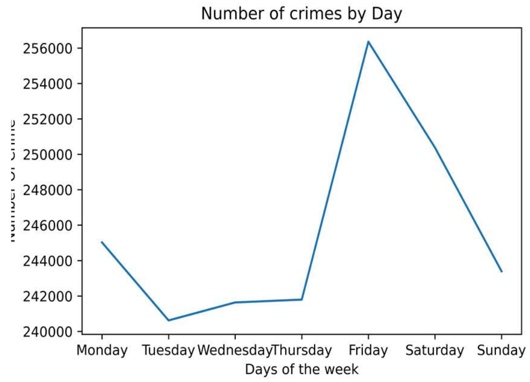


Figure 3: Crimes by days of the week

For fig 3, we see the trends of crime over days of the week. We can see that highest number of crimes take place at the beginning of the weekend- Friday & Saturday.

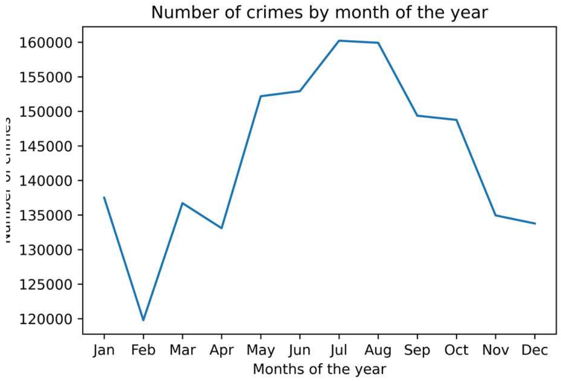


Figure 4: Crimes by month of the year

For fig 4, we see the trend of crimes over the course of months in a year. The highest number of crimes were reported at the beginning of the winter months.

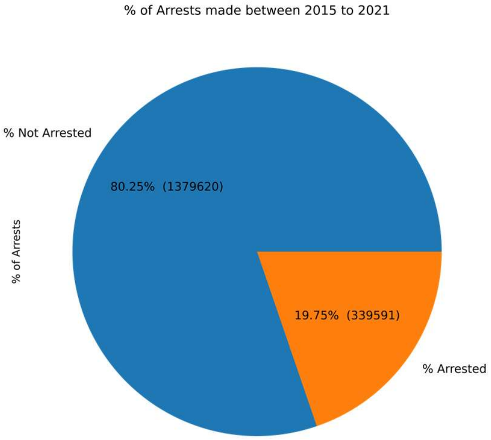


Figure 5: Percentage of arrests made between 2015-2021

For Fig 5, we have a chart explaining the number of arrests made over the given period. More advanced graphs can be used to better understand the data which are specific to the target year or target location.

D. Feature Engineering

- *Chi-square test of independence:*

Null Hypothesis: There is no relationship between variables
Alternative Hypothesis: There is a relationship between variables

Chi-square test was used to find the relationship between categorical variables

In the formula below, observed represents the actual count for each category, while anticipated represents the expected count based on the population distribution for that category.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

We took 0.05 (i.e., 5%) as the significance level and degree of freedom is 1. Degrees of freedom is calculated as (no of rows-1) * (no of columns-1). Critical value is calculated by considering the degree of freedom as 1. In p-value we are getting the cumulative distribution function to cover all the area under the distribution up to our chi-squared test with degrees of freedom equal to 1.

Dataset had three categorical columns, to check the association between categorical values we used the chi-square test of independence. In this way we can remove those variables and it will be easy to train the model faster and in an efficient way.

First, we looked at *Primary_type* and Location. If there is a relationship between them, we can drop one of them as they are not independent to impact the target variable.

1. Made a contingency table for observed data and then calculated the expected values.
2. We have chosen the p-value to be 0.05.
3. Degree of freedom was calculated to check the number of variables to be estimated in order to complete the dataset which was also required for chi-square statistic calculation.
4. Chi-Square depends upon observed and expected values, in our case the chi-square value came out to be 605323.30
5. The Chi-square value was the pretty good value to determine the relation between two variables.
6. Critical value is less than chi-square value then the result for this test statistically significant.
7. P value was also less than 0.05.
8. Considering the test results, *Primary_type* and Location are highly dependent on each other.
9. Therefore, we decided to drop one of the variables as it will not impact the target variable.

Similarly, we looked at the *Primary_type* and *Description* columns, which had a relationship, and had to remove one of them. We were able to remove the *Location* and *Description* attributes in this method.

- *One hot encoding*

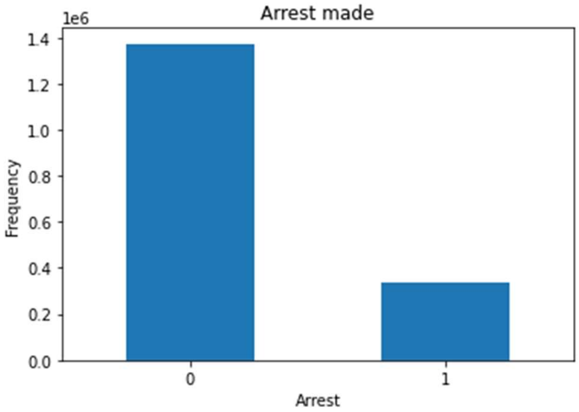
One-hot encoding was an important step in getting a dataset ready for machine learning algorithm. Categorical data is transformed into a binary vector representation using one-hot encoding. Pandas for Dummies made it simple! This creates a new column for each unique value in a column. Depending on whether the value matches the column header, the values in this column are represented as 1s or 0s.

There were 35 entries in *Primary_type,* which resulted in 46 columns in total after one hot encoding. We decided to examine the top 10 Primary-type crimes and make predictions for them. After running one hot encoding we converted the top 10 *Primary_type* into dummy variables.

In order to select the best features of modelling we plotted a correlation matrix which showed 12 variables correlating with each other. Therefore, we dropped those columns and the features that were used for modelling are: *Domestic, Beat, Ward, Community, year, month, dayOfWeek, hour, Theft, Battery, Criminal Damage, Assault, Deceptive Practice, Other Offence, Narcotics, Burglary, Motor Vehicle Theft and Robbery.*

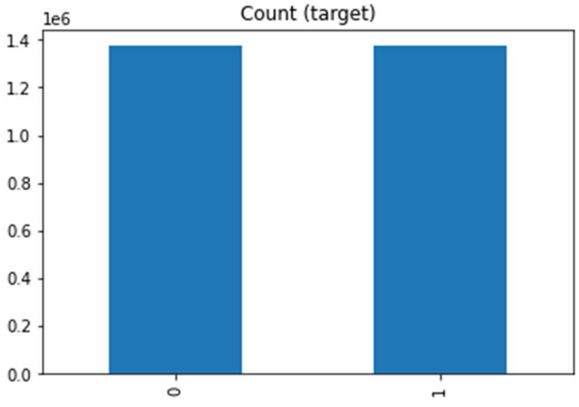
- *Handling Imbalanced dataset*

In order to understand the target variable, we plotted the bar graph which depicts that there are more of the 0 values than 1, which clearly shows the imbalance in the target variable.



A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and / or adding more examples from the minority class (over-sampling). We have performed random over sampling as we did not want the valuable information to miss out.

After performing the oversampling there are an equal number of 0's and 1's in the target variable.



E. Modelling

Any predictive model that uses the classification algorithm frequently incorporates regression techniques to construct a predictive model. Test data is recognized throughout this procedure, and classification judgments are applied to the data to determine the model's performance.

Based on our goals we used three models:

1. Logistic Regression
2. Decision Tree
3. Naive Bayes

The data is separated into two groups: xtrain, ytrain, and xtest, y test. Arrest is the target variable i.e., ‘y’ and rest are the independent variables i.e., ‘X’. Data is split into 80% training data and 20% testing data. Sklearn is used to import the algorithms model. model.fit is used to construct the model.fit. Model.predict is used to make predictions after the model has been built using the above process (xtest). The accuracy is determined by importing accuracy score from metrics - metrics.accuracy score (ytest, predicted).

IV. Evaluation

After modelling, the classification report for logistic regression was as follows:

	Precision	Recall	F1-score
0	0.81	1.00	0.89
1	0.73	0.02	0.03

Above tables depict that the model is biased towards the majority class that is False (0) - not arrest. Recall for 0 is 100% whereas for 1 is just 2%. Which means that the target variable ‘Arrest’ has imbalanced data. In order to understand the dataset, we decided to deal with an imbalanced dataset.

After learning about the imbalance data and performing over-sampling, the classification report for logistic regression is as follows:

	Precision	Recall	F1 score
0	0.84	1.00	0.92
1	1.00	0.25	0.40

Handling the imbalance dataset has given a slightly better result with respect to precision, recall and accuracy. With 100% recall for ‘0’ and around 25% for ‘1’.

Similarly, Classification report for Decision tree is:

	Precision	Recall	F1 score
0	0.80	0.95	0.92
1	0.20	0.06	0.40

For our third model, classification report is as follows:

	Precision	Recall	F1 score
0	0.90	0.78	0.84
1	0.42	0.63	0.50

Precision is defined as, out of all arrests (1) predicted; how many are right?
Logistic regression - 100%
Decision tree - 20 %
Naive Bayes- 42%

Recall is defined as, out of all arrests (1) actual values how many we got right?
Logistics regression-25%
Decision tree- 6%
Naive bayes- 63%

Considering the recall and F1-score of Naive bayes for predicting the arrest (1) we can say that naive bayes is performing well in prediction of classification.

V. Conclusion & Future Scope

Critical value is less than chi-squared value so it will be significant, and the p-value is 0.00 which is a small value. Since the p value is less than 0.05 and chi-squared value exceeds the critical value, we rejected the null hypothesis.

Handling the Imbalanced dataset has given significantly better results. Even though the accuracy is higher the model might not correctly predict the classes.

Machine Learning Algorithm	Accuracy Score
Logistic Regression	85%
Decision Tree	80%
Naive Bayes	75%

Therefore, we evaluated the model performance using precision, recall and f1-score. When compared the three model’s classification report, the naive bayes gave better prediction performance. Hence, we conclude that the naive bayes is the best suited model for our case study.

Future Scope

We discovered the benefits of dealing with imbalance data. Despite the benefits of balancing classes, these methods are not without flaws. Over-sampling can be as simple as duplicating random records from the minority class, which can lead to overfitting. We can explore how to deal with overfitting difficulties in the future, as well as which machine learning method is most suited for handling imbalance data.

The code[16] and files[17] for the study can be found at the above locations respectively.

References

[1] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.

[2] O. Llaha, "Crime Analysis and Prediction using Machine Learning," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 496-501, doi: 10.23919/MIPRO48935.2020.9245120.

[3] A. R. C. da Silva, I. C. de Paula Júnior, T. L. C. da Silva, J. A. F. de Macêdo and W. C. P. Silva, "Prediction of crime location in a brazilian city using regression techniques," 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), 2020, pp. 331-336, doi: 10.1109/ICTAI50040.2020.00059.

[4] N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 266-273, doi: 10.1109/ICAIS50930.2021.9395953.

[5] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, "Crime analysis and prediction using data mining," 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014, pp. 406-412, doi: 10.1109/CNSC.2014.6906719.

[6] T. Li, "Criminal Behavior Analysis Method Based on Data Mining Technology," 2016 International Conference on Smart City and Systems Engineering (ICSCSE), 2016, pp. 562-565, doi: 10.1109/ICSCSE.2016.0152.

[7] R. Bansal, N. Gaur and S. N. Singh, "Outlier Detection: Applications and techniques in Data Mining," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016, pp. 373-377, doi: 10.1109/CONFLUENCE.2016.7508146.

[8] B. Kaur, L. Ahuja and V. Kumar, "Crime Against Women: Analysis and Prediction Using Data Mining Techniques," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 194-196, doi: 10.1109/COMITCon.2019.8862195.

[9] C. Chauhan and S. Sehgal, "A review: Crime analysis using data mining techniques and algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 21-25, doi: 10.1109/CCAA.2017.8229823.

[10] D. Uzlov, O. Vlasov and V. Strukov, "Using Data Mining for Intelligence-Led Policing and Crime Analysis," 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), 2018, pp. 499-502, doi: 10.1109/INFOCOMMST.2018.8632122.

[11] T. Siddiqui, A. Y. A. Amer and N. A. Khan, "Criminal Activity Detection in Social Network by Text Mining: Comprehensive Analysis," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 224-229, doi: 10.1109/ISCON47742.2019.9036157.

[12] L. Bao, "Correlation Analysis of Crime Factors Based on Data Mining," 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2020, pp. 577-580, doi: 10.1109/ICICTA51737.2020.00127.

[13] Department, Chicago Police. "Crimes - 2001 to Present: City of Chicago: Data Portal." *Chicago Data Portal*, 31 Jan. 2022, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.

[14] https://github.com/IamSidharth/ChicagoCrimeAnalysisAndPrediction

[15] https://github.com/NamanJain2050/chicago-crime-predictions

[16] CA683 Assignment Code: https://github.com/Kashmira98/DADM-Assignment

[17] CA683 Assignment Files: https://drive.google.com/drive/folders/1DyKVemZmbw52hy7flsuoIva2Y0hci6bM