International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

Crime Analyses Using Data Analytics

Thanu Dayara, Manukau Institute of Technology, New Zealand
Fadi Thabtah, ASDTests, Auckland, New Zealand
Hussein Abdel-Jaber, Arab Open University, Saudi Arabia*
Susan Zeidan, Zayed University, UAE

ABSTRACT

One potential approach for crime analysis that has shown promising results is data analytics, particularly descriptive and predictive techniques. Data analytics can explore former criminal incidents seeking hidden correlations and patterns, which potentially could be used in crime prevention and resource management. The purpose of this research is to build a crime analysis model using supervised techniques to predict the arrest status of serious crimes in Chicago. This is based on specific indicators, such as timeframe, location in terms of district, community, and beat, and crime type among others. We used time series and clustering techniques to help us identify influential features. Supervised machine learning algorithms then modelled the subset of features against incidents related to battery and assaults in specific timeframes and locations to predict the arrest status response variable. The models derived from Naïve Bayes, Decision Tree, and rule induction algorithms reveal a high predictive accuracy rate at certain times in some communities within Chicago. This information can benefit the city's law enforcement to optimize resource allocation, such as patrols, new stations, and prevention strategies.

KEYWORDS

Crime Analysis, Crime Prediction, Data Analytics, Dimensionality Reduction, Machine Learning, Resource Management

INTRODUCTION

A crime is an unlawful act by an individual that is contrary to the laws of a region or a country. Crime is a significant issue, notably since the 1960s, that negatively affects the social fabric and economy of countries, and attributable to greed, poverty, and economic distress (Saleh & Khan, 2019). Crimes, such as assault, homicide, theft, battery, and those involving narcotics have grown rapidly in Chicago, USA; for example, reported robberies in 2016-2017 exceeded 16,000 cases (Saleh & Khan, 2019). Battery is a criminal charge for the illegal use of violence toward the body of another citizen resulting in violent interaction or physical harm (Castillo, 2015). According to the New Zealand Police (2020), even for small countries like New Zealand, the ratio of crimes committed in 2019 was 10% greater than that of 2018.

DOI: 10.4018/IJDWM.299014 *Corresponding Author

Crimes affect people's quality of life and the economic growth of a country. People often consider crime rates within the city as a primary factor for moving so that they can avoid areas where crime rates are high (Chalom et al., 2002). Furthermore, the Global Peace Index organised by the Institute of Economics and Peace (IEP, 2019) utilizes crime rates as an indicator when globally ranking countries. Security is a major concern worldwide (Malathi, & Babo, 2011).

The need for more police patrols burdens governments financially. High crime rates negatively affect the distribution of public funds to public services, such as emergency, education, fire, and healthcare (Ahishakiye et al., 2017).

Crime analysis is the process of differentiating and evaluating patterns related to crime past crime data. According to Ahishakiye et al. (2017), the patterns discovered when conducting crime analysis help managers allocate resources effectively and also support the police to apprehend criminals. Data analysis techniques, such as artificial intelligence (AI) and visualization, play a vital role in the process of crime analysis providing exceptional search capabilities to derive valuable results (Sharma et al., 2021; Bunker, & Thabtah, 2017).

Applying AI techniques to crime data is possible due to high incidences of crime. AI methods, such as machine learning, explore crime datasets to find concealed correlations that can be utilized by analysts to take appropriate actions related to arrests and resource management (Abbass et al., 2020). For example, Shah et al., (2021) investigated machine learning and computer vision approaches to improve crime detection rate. In addition, a recent research work by Abbas et al., (2020) revealed that predictive machine learning models can be used for predicting crimes occurring on social media such as Cyber Scam, Cyber stalking and Cyber Hacking among others.

Chicago is the third largest city in the USA; in 1974, when the population reached 3 million, 970 murders were recorded, resulting in a murder rate of 29 deaths per 100,000. This rate gradually declined in the mid-2000s, but a huge rise in violent crimes reoccurred, i.e., 749 murders in 2016; this is more than the number of murders in the largest cities of New York and Los Angeles with 334 and 294 murders, respectively (BBC, 2017). A BBC article in 2017 reported that Chicago Police were turning to big data to help predict times and locations where crimes might occur (BBC, 2017). Ahishakiye et al. (2017) stated that in a statistical sense, crime is predictable mainly because criminals tend to stay in their zone of comfort. This assumption was strongly supported with evidence that certain crimes occurred in similar timeframes and around common districts as where they were successfully committed previously. In this research, we pursue this assumption by developing a supervised learning model to predict arrest status based on Chicago crime data for three consecutive years (2017-2019). To achieve the aim, we initially explore correlations between crime type, timeframe during the day (24 hours), and locations (beats, communities, districts) to offer law enforcement analysts in the Chicago Police Department (PD) useful insights. A beat is the smallest police geographic area and usually supported with a single police car (Crimes - 2001 to present, 2020). More importantly, the new predictive model will be able to predict arrest status using features, such as types of crimes, timeframe within a day, location, etc. These features have been identified using computational intelligence methods to ensure that only highly influential features are offered to the machine learning algorithm. Therefore, models derived will improve the predictive power of arrest status, particularly for harmful crimes, such as battery and assault, within peak time zones at hot districts in Chicago.

Section 2 provides literature related to crime analysis using data anlaytics and supervised learning. The methodology is discussed in Section 3, and Data pre-processing is discussed in Section 4. In Section 5, experimentations and result analyses are conducted along with descriptive and predictive analysis; finally, we conclude in Section 6.

LITERATURE REVIEW

Iqbal et al. (2013) conducted research to predict the crime category for various US states on real crime data using Decision Tree (Quinlan, 1993) (UCI Machine Learning Repository, 2012). The authors

compared Decision Tree with Naïve Bayes (Duda & Hart, 1973) to determine the ideal model for predicting the crime category. The results in terms of classification accuracy showed 70.81% and 83.95% for models derived by Naïve Bayes, and Decision Tree, respectively.

Ahishakiye et al. (2017) investigated crime analysis on a 'Crime and Communities' dataset (UCI Machine Learning Repository, 2017) using several machine learning algorithms including Decision Tree, Multi-layered Perceptron (MLP), Naïve Bayes, and Support Vector Machine (SVM) (Quinlan, 1993; Rosenblatt, 1957; Duda & Hart, 1973; Cortes & Vapnik, 1995). After experimentations using 80% training and 20% testing subsets, the Decision Tree classification model was selected by the author for crime analyses on the dataset. The Decision Tree algorithm was able to efficiently derive the clarification models in terms of training time (time taken to build the models).

Nguyen et al. (2017) performed a crime analysis on a Portland Police Bureau (PPB) dataset obtained from two databases, the PPB (National Institute of Justice, 2017) and the American Factfinder website (American FactFinder, 2017). Several machine learning algorithms, i.e., SVM, Gradient Boosting Machines, Random Forest, and Neural Networks were used (Cortes & Vapnik, 1995; Friedman, 1999; Kam Ho, 1995; Rosenblatt, 1957). The accuracy gained by the classification models from PPB data shows 57.3%, 72.2%, and 88%, respectively for SVM, Random Forest, and Gradient Boosting Machines. The second dataset recorded 79.39%, 65.79%, 61.67%, and 74.24% for SVM, Random Forest, Gradient Boosting Machines, and Neural Networks, respectively.

Shermila et al. (2018) performed research on San Francisco Homicide data (FBI Supplementary Homicide Report 1980-2014, 2014) using kNN Classifier, Multilinear Regression, and Neural Networks algorithms to predict the criminal's age, gender, and relationship to the victim (Cover & Hart, 1967; Galton, 1989; Rosenblatt, 1957). The findings in terms of accuracy for Multilinear Regression showed 60% for 'age', kNN showed 85% for 'gender' and 48% for 'relationship' feature. The Neural Network algorithm outperformed the other classifiers deriving models with 96% and 97% for 'gender', and 'relationship', respectively. kNN was also employed by Bharati and Sarvanaguru (2018) to predict Chicago crime data along with other algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM) and Bayesian Method (Galton, 1989; Quinlan, 1993; Kam Ho, 1995; Cortes & Vapnik, 1995; Bayes, 1763). Data pre-processing, feature selection, and scaling were performed prior to data modelling. The classification accuracy rates produced by the models are 78.70%, 64.60%, 31.30%, and 78.60% for kNN, Naïve Bayes, SVM, and Decision Tree, respectively.

Saleh and Khan (2019) studied Chicago crime data for 2012–2017 using models and pre-processing methods within the Python Library. The K-means clustering technique (Lloyd, 1957) was applied after selecting the attributes: Case ID, Type, Time, Location, Year, and Arrest. Findings revealed robberies were at their peak between 2016-2017 and most of the criminals were not arrested for their crimes.

Yuki et al. (2019) extended the work of Saleh and Khan (2019) and predicted the location and type of crime that is committeds at a certain time and the location on the same dataset using Decision Tree, Random Forest, Bagging, AdaBoost and Extra Tree (Quinlan, 1993; Kam Ho, 1995; Breiman, 1994; Freund & Schapire, 2003). The investigation revealed the highest level of crime was in 2008. After experimenting with training data, the authors identified Bagging as the best model to predict Chicago's crime with 99.92% accuracy.

Sharma et al., (2021) investigated a crime dataset related to Boston Police Department to determine locations in the city that are associated with high and low crime rates. The authors utilized few variables related to time and location to conduct a descriptive analysis, and then performed feature assessment using Principal Component Analysis with classification. The reported results show slight increase in the detection rate when PCA was utilized prior to classification.

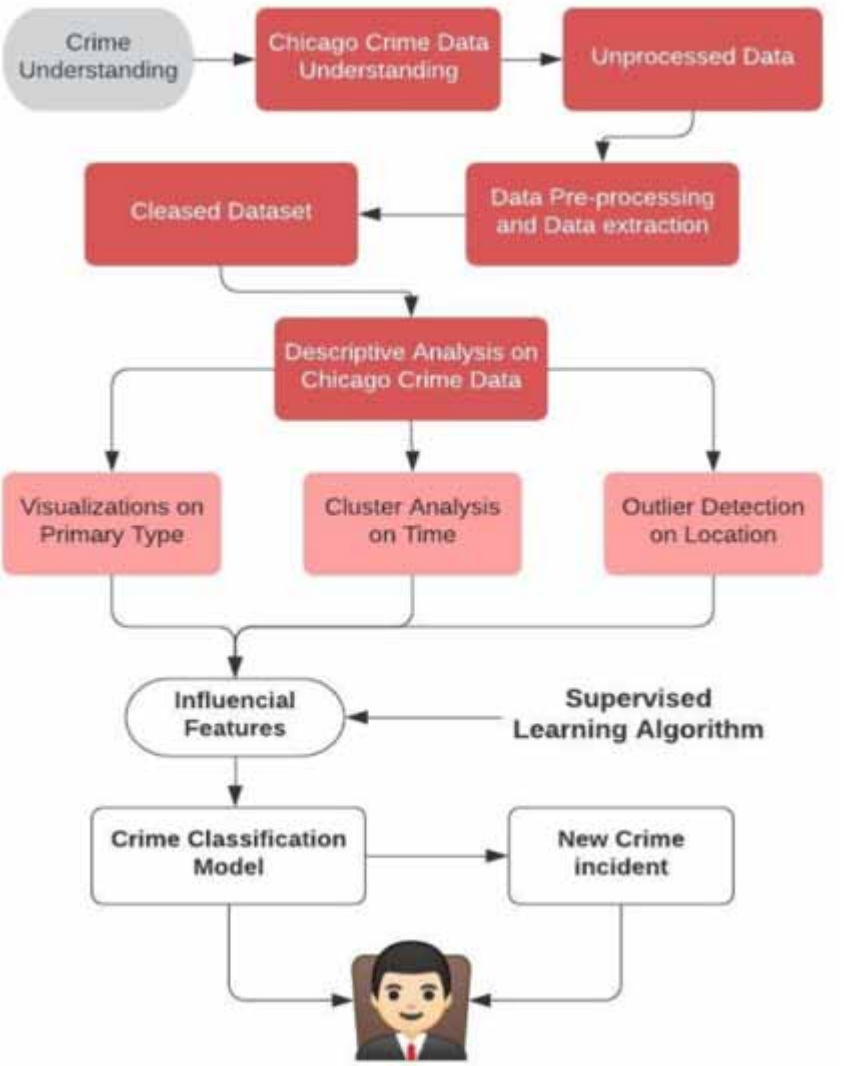
This research is distinguishable as that the developed model examines the correlations between crime type, timeframes during the day (24 hours), and locations (beats, communities, districts). It uses this knowledge to build a timeframe for 2021 to be useful for the PD to distribute resources effectively.

METHODOLOGY

Figure 1 consists of the primary phases of the developed model. Initially we obtained access and downloaded the Chicago crime dataset; data was then extracted to derive crime information for the last three years as the target training dataset for the analysis. Crime data from 2017-2019 in all districts has been considered in this research. Feature analysis and data statistic phases were conducted to better understand the features in the dataset, such as 'Primary Type', 'Time', 'District', 'Community Area' (CA), 'Beat' among others. This helped us to perform pre-processing including unwanted data removal (instances with multiple missing values totalling 9,247). The cleansed dataset was then utilized for descriptive analysis by including:

1. Visualizations based on Primary Type
2. Clustering based on Time

Figure 1. Proposed methodology



3. Outlier detection based on Location
4. Time series for the next 3 years based on incidents in the last three years
5. Combinations of A-C

We focused on influential features based on the feature selection (Section 4) and descriptive analyses conducted. To be exact, three features (Timeframe in 24 hours, per hour, and 3-hour intervals respectively, Crime Type, Location including district, CA, and beat) together with other features have been evaluated. The reason for selecting these features is to identify common crimes during any given timeframe, cluster these by 'Primary Type' and then based on information obtained, 'hot' geographical areas with certain crime types can be identified. Based on the results of these analyses, we will be able to identify serious crimes, particularly battery and assault, within a specific time of the day and locations. We can predict arrest status for these crime types by utilizing the influential features obtained.

Classification techniques including Naïve Bayes, Decision Trees, and rule induction (Duda & Hart, 1973; Quinlan, 1993; Cohen, 1995;), have been utilized to create crime prediction models to forecast arrest status in specific locations at certain times. This will be highly beneficial as allocation of resources can be optimized with respect to time, severity of the crime, and location.

Decision Tree algorithm (C4.5) employs divide and conquer strategy to learn classifiers in the format of trees. Each node in the tree represents an attribute and each branch represents an attribute value. The algorithm builds the tree using information-based metrics such as Shannon Entropy in which the attribute that has the largest information gain is always chosen. Once the tree is built, a path from the root to each leaf denotes a rule. On the other hand, rule induction approach creates a rule using separate and conquer approach. In doing so, an empty for a certain class label is formed, and the algorithm seeks for the attribute value in the training dataset that when appended into the rule increases the rule expected accuracy according to a mathematical metric. The algorithm keeps appending attributes values into the rule until the rule's expected accuracy cannot enhanced any further. When this happens, the rule gets derived, and all data observations associated with it are discarded from the training dataset. The algorithm keeps creating rules until the dataset gets empty or no more rules can be discovered. Lastly, Naïve Bayes utilizes a Bayes Theorem to predict the class of the test data in an efficient manner. The reason for selecting these classification algorithms is their applicability in other application domains besides the different learning schemes they employ to derive classification models.

The dataset used in this research is derived from the Chicago Police Department's CLEAR system (Citizen Law Enforcement Analysis and Reporting) (https://data.cityofchicago.org/public-safety/crimes-2020/qzdf-xmn8). The dataset has the crime recorded in Chicago from 2001 to the present and includes more than 7,000,000 instances and 22 features (Table 1). We cleaned noise from the dataset, such as instances with missing values and features that add no value to the learning process. Since the amount of missing data, is less than 10% of the total data we decided to remove all instances with missing values to reduce data inconsistency. Furthermore, the 'Date' attribute of the dataset is a collection of both the date and time of the incident. Therefore, to carry out graphical statistics smoothly, we create new features by separating 'Month', 'Day', 'Year', and 'Time' from the 'Date' attribute. Thus, a set of four additional features has been added to the revised dataset. More importantly, we removed 'Date' and 'Year' features to avoid duplication and 'ID', 'IUCR', 'Description', 'Location Description', 'Domestic', 'Ward' and 'Updated On' as they are insignificant.

RESULTS AND ANALYSES

Descriptive Analysis

To gain a better understanding of how the crimes occurred in Chicago we utilized Tableau, a statistical and data visualization tool used for clustering, visualization, an correlation analysis. We investigate

Table 1. Dataset characteristics

Attribute	Data Type	Description
ID	Numeric	Unique identifier for the case.
Case Number	Numeric	Records Division Number.
Date	Date and Time	Date and time of the incident.
Block	String	Address at which the incident occurred.
IUCR	Numeric	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description.
Primary Type	String	The primary description of the IUCR code.
Description	String	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	String	Description of the location where the incident occurred.
Arrest	Boolean	Arrest was made or not.
Domestic	Boolean	Indicates whether the incident was domestic related as defined by the Illinois Domestic Violence Act.
Beat	Numeric	Indicates the beat where the incident occurred*.
District	Numeric	Police district where the incident occurred.
Ward	Numeric	The ward (City Council District) where the incident occurred.
Community Area (CA)	Numeric	Indicates the community area where the incident occurred.
FBI Code	Numeric	Crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
X Coordinate	Numeric	The 'x' coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Y Coordinate	Numeric	The 'y' coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Year	Numeric	Year the incident occurred.
Updated On	Date and Time	Last updated date and time.
Latitude	Numeric	The latitude of the location where the incident occurred.
Longitude	Numeric	The longitude of the location where the incident occurred.
Location	Location	The location where the incident occurred.

*A beat is the smallest police geographic area – each beat has a dedicated police-beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.

the correlations between crime type, timeframes during the day (24 hours), and locations (beats, communities, districts), and then develop a descriptive model to provide Chicago PD with valuable insights into the highest crime type, highest crime prevailing time, hot districts, community areas (CAs), and beat per top crime types.

Figure 2 depicts crime frequency by type from 2017-2019 in Chicago. Theft-related crimes have the highest frequency at 189,161 incidents, followed by battery and criminal damage at 148,294 and 83,241, respectively. The 'Primary Type' attribute in the input dataset originally consisted of 32 possible values, out of which the top 16 crime types in terms of frequency have been plotted in

Figure 2. Total crimes against crime type 2017-2019

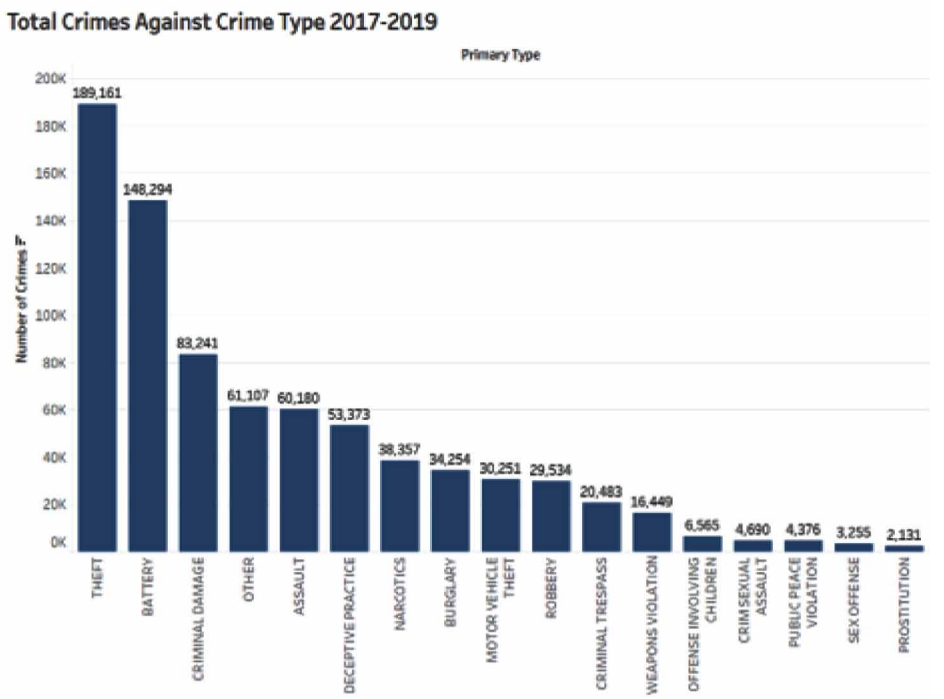


Figure 2; the remaining crimes (61,107 incidents) were integrated into one category called 'Other'. We will narrow the scope to consider harmful crimes only; arrest prediction particularly battery and assault we will see later in this research.

Figure 3 demonstrates the trend in criminal incidents in Chicago on a given day (2-hour timeslot) over three consecutive years. The trend appears to follow the same pattern throughout the years in which the timeframe between 8-12 am is at its peak. This information is highly useful, and it pinpoints that despite the efforts made by the Chicago PD in terms of resource allocation, alarmingly, the crime frequency pattern with regards to timeframes per hour has not changed for three consecutive years. To further investigate the pattern seen in Figure 3, we investigate the sort of crime that has taken place on an hourly basis broken down by year. We discover that each year from 1.00 am until 3.00 am, and from 8.00 am until 10.00 am, the rates of incidents are similar. More importantly, the highest number of crimes were reported as occurring at 12:00 midday in any given month over the last three years. The lowest number of crimes is registered between 1 pm and 11 pm. It is also feasible to group the hours to distinguish which types of crimes are more prominent in peak hours. We have been able to recognise the trends and the level of crime over the corresponding three years.

According to descriptive analyses conducted, so far against the historical data considered, there is still a significant likelihood of the same trend for crimes occurring at 12:00 midday in the future, as depicted in Figure 4. This leads to studying crimes by specific location, such as district, CA, or beat, to find the 'hot' locations per crime type and time. By gaining insights on these three attributes, Chicago PD will be able to optimize the resources allocated, such as new stations, patrols, and community service among others, and more importantly, the type of resources needed per specific communities and when.

We isolate the top five crime types for 23 districts broken down by year. Theft incidents are the highest in districts #1, #6, and #18 having 7,014, 3,583, and 6,546 incidents, respectively between 8.00 am and 12.00 midday in any given year. Battery is the highest crime type in district #8 with 3,927

Figure 3. Crime type 2017-2019

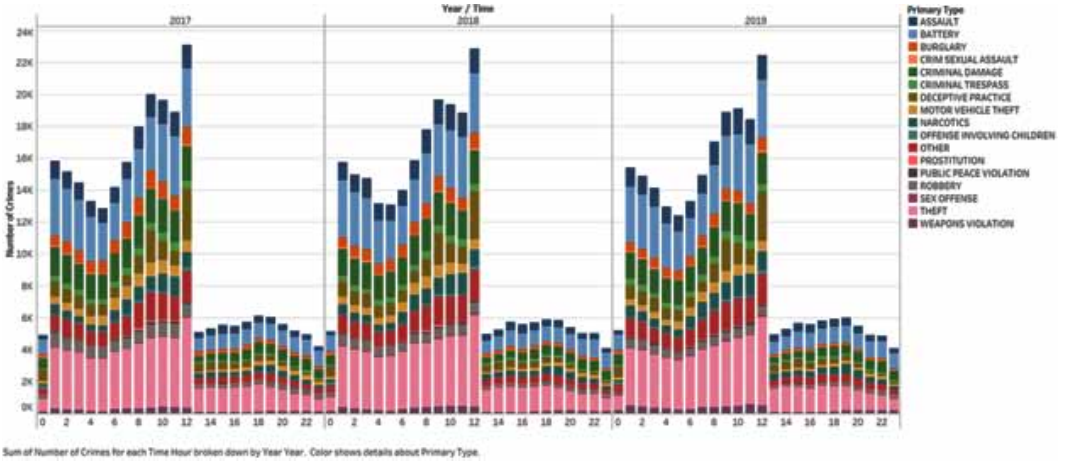
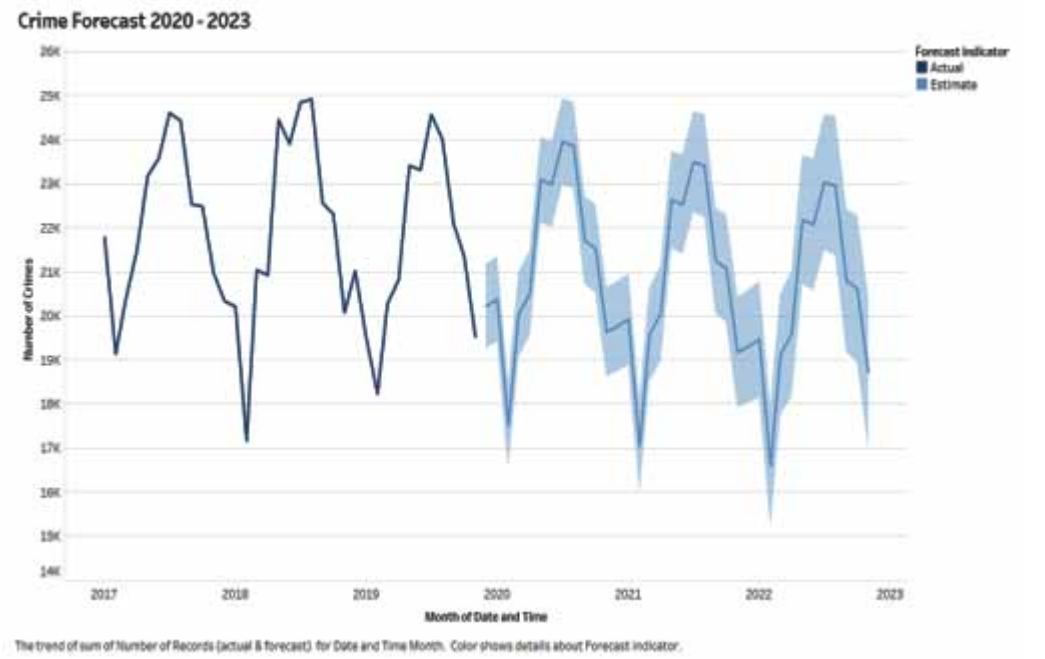


Figure 4. Crime forecast 2020-2023



incidents between 8.00 am and 12.00. These four districts reported 16,342 crimes for 2017, 16,744 for 2018, and 16,396 for 2019. Districts #1 and #18 rank first and second in the past three years in terms of crime frequency with a high likelihood that it will be repeated in 2020. This information is useful for the Chicago PD as it can concentrate additional resources on these two districts. Apparently, those two districts fall under Area Central covering Central, the Loop, and the Near North Side of Chicago and standing out as outliers exceeding all other districts' crime rates (Halat et al., 2015). The near North Side district has the highest number of skyscrapers, and the largest population with high per capita income. It is the oldest part of Chicago, and thus, it a potential spot for criminal activities.

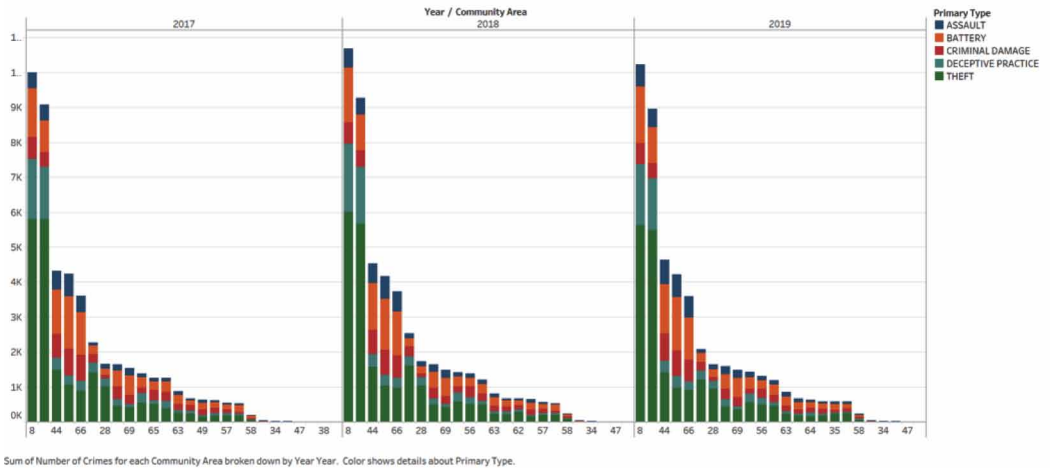
Some famous locations, such as the Gold Coast, Navy Pier, and Magnificent Mile are in this area. The shortage of economic opportunities for youth in the central and Near North Side districts have contributed to a rise in crime (Vélez, & Richardson, 2012). Chicago has high youth unemployment levels, so it's hard to make a living with limited opportunities. Hence, if the government is keen to provide viable opportunities for young citizens, the crime rate will be significantly reduced (Vélez, & Richardson, 2012).

District 31 stands out showcasing the least crime with just one per year indicating that it is safe. This district includes Norridge and Harwood Heights, which are surrounded by the city. The populations of these two villages were 8,612 and 14,572, respectively according to the 2010 census. Based on United States census data, Harwood Heights is an Illinois community comprising 32.5% Polish Americans. The estimated median household income of Harwood Heights in 2017 was $55,881 whereas for other Illinois communities it was $62,992; this region is still expanding and developing (Charles, 2013) (Chalom et al., 2002).

We condensed the number of crimes to ascertain smaller communities that might be hotspots for crimes by splitting districts into CAs, i.e., 77 in Chicago. Figure 5 reflects the top five crimes in CAs by year. It is obvious that Central Side's CAs with larger population groups, i.e., #8 and #32, have more crimes than all other CAs. Serious crimes, such as battery and assault are concentrated in the Auburn Gresham, Chatham, and Chicago Lawn communities, peaking between 8.00 am-12.00 midday in any given year. Specifically, CAs #8 and #32 come under districts #18 and #1, respectively. Furthermore, Chatham and Auburn Gresham fall within district #6, while Chicago Lawn falls within district #8. The total number of incidents between 8.00 am and 12.00 pm involving battery and assault are [1051, 918], [1066, 944] and [1070, 1017], respectively for Auburn Gresham, Chatham, Chicago Lawn CAs between 2017 and 2019. These three areas are classified as dangerous neighbourhoods in Chicago. This pattern of crimes indicates an alarming situation for the PD requiring increased patrols in these specific CAs especially at peak times. Placing more police community support officers in these communities provides a reassuring presence and is a promising crime prevention strategy.

We further investigate smaller locations within CAs to find specific beats related to crime types and timeframes. The focus is on investigating the Chatham, Chicago Lawn, and Auburn Gresham CAs to identify more specific beats. Table 2 reflects the frequency of top crime types by beat between 8.00 am and 12.00 pm and for three consecutive years (2017-2019) along with arrest rates. Battery-related crimes have the highest frequency in beat #631 with 175, 163, and 167 incidents in 2017,

Figure 5. Top crimes by location (CA) 2017-2019



2018, and 2019, respectively; this is followed by beat #612 with 154, 127, and 172 incidents in 2017, 2018, and 2019, respectively.

It's obvious from Table 2 that there is a need to maximize police patrols in beats, such as #631, #632, and #612 where there is a huge gap between rate of crimes and rate of arrests. The average arrest rate from 8.00 am to 12.00 pm. for 2017-2019 is 16%. Therefore, instead of one patrol car assigned per beat, the Chicago PD could allocate more patrol cars between 8:00 am and 12.00 pm. Between 8.00 am and 12.00 pm beats, such as #631, #632, and #612 would also benefit from a greater police presence.

Predictive Analysis

We expanded the analysis to forecast the arrest status class using supervised machine learning algorithms for serious crimes including battery and assault. In particular, we employed three classification algorithms: probabilistic-Naïve Bayes, Rule Induction-Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and Decision Tree- C4.5 (Duda & Hart, 1973; Cohen, 1995; Quinlan, 1993) on features identified using feature selection methods. We created a new feature 'timeframe' which includes three possible values:

1. Midnight to 6.00 am
2. 6.00 am to 12.00 pm
3. 12.00 pm to midnight

To select the relevant features, we used Leave One Out Cross Validation (LOOCV), Information Gain (IG), and Correlation Feature Set (CFS) methods (Brown, 2000; Quinlan, 1993; Hall, 1999) with Naïve Bayes as a base classifier. The results of the feature selection process are shown in Table 3. Block and Beat features were common and often ranked high by the considered feature selection methods. Hence, we retained these two features in the final subset of features, however, we will use only one of them during the building of the classification models as they both represent location.

Supervised learning techniques have been applied against the subset of features retained, i.e., Beat/Block, Primary Type, Timeframe, Date, and Arrest Status - class label, and the results are shown in Table 4a. We consider just one feature to represent the locations, and run the algorithms against two subsets of features: one with location as Block and one with location as Beat. The supervised learning algorithms maintain acceptable classification models with slightly over 80% predictive accuracy except for the Naïve Bayes algorithm, which produced a model slightly less than 80% when Beat was replaced with Block. Overall, the classification accuracy of the models derived by the different machine learning algorithms were consistent and stable. However, we suspected

Table 2. Top crime types by beat in peak hours

	Year/ Beat								
	2017			2018			2019		
Battery	175	154	94	163	127	111	167	172	92
Theft	87	99	144	98	121	152	74	91	113
Criminal Damage	76	70	69	53	64	45	84	66	63
Assault	59	81	52	59	74	53	80	59	59
Deceptive Practice	41	30	51	30	34	53	31	34	46
Total Crimes per beat	438	434	410	403	420	414	436	422	373

Table 3. Feature selection methods and ranking

LOOCV	IG	CFS
Block	Block	Block
Beat	Beat	Primary Type
Community Area	Date	Time Frame
District	Community Area	District
Primary Type	District	Date
Date	Primary Type	Beat
Time Frame	Time Frame	Community Area

Table 4a. Misleading model performance without balancing data

		Features Used	Classification Accuracy
Naïve Bayes	Block	Timeframe, Primary Type, Date	79.36%
	Beat		80.22%
RIPPER	Block	Timeframe, Primary Type, Date	---
	Beat		80.22%
C4.5	Block	Timeframe, Primary Type, Date	80.22%
	Beat		80.22%

that the reason behind similar classification accuracies derived by the algorithms is due to a data imbalance issue. Therefore, we further analysed the confusion matrix results and noticed that Naïve Bayes, RIPPER, and C4.5 have high misclassification rates for the true positive instances. In other words, all machine learning algorithms failed to classify incidents which had 'Arrest Status = Yes' due the original training dataset having over 80% incidents with 'Arrest Status = No'. Therefore, the classification accuracies derived by the considered algorithms are misleading and sampling the data before the learning phase becomes essential. For example, the Naïve Bayes algorithm produced models that had no true positive rates, misclassifying 41,236 that were supposed to be 'Arrest Status = Yes' into 'Arrest Status = No', which is misleading. So, all results obtained have been ignored and we applied the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) before the training phase to balance the class in the dataset.

Table 4b depicts the performance of the machine learning algorithms for predictive Arrest class for battery and assault crimes after applying the SMOTE data sampling method (Chawla et al., 2002). The results of the models derived by Naïve Bayes, RIPPER, and C4.5 were low, but the C4.5

Table 4b. Model performance after using SMOTE with location beat

	Accuracy	Precision	Recall	F-measure
Naïve Bayes	59.78%	0.598	0.598	0.598
RIPPER	51.89%	0.526	0.519	0.494
C4.5	63.50%	0.635	0.635	0.635

algorithm showed superiority, i.e., 63.50% predictive power. The classification models derived have low predictive accuracy, precision, recall, and harmonic mean rates. These results, if limited, show that predicting the arrest status for just two crime types using the Date, Timeframe, Primary Type, and Block/Beat subset of features is not highly influential, at least on the Chicago crime dataset. The results show that more features are needed to improve arrest status especially in high areas for crimes, such as battery and assault, for example in beats #612, #631, and #632. Beat #612 belongs to CA #71; beats #631 and #632 belong to CA #44; these three beats have 1,723, 1,812, and 1,308 battery and assault incidents, respectively. More importantly, the question of when these beats have a low or high prediction of arrest status becomes crucial.

After delving into the statistics of the models derived by the best performing algorithm in terms of accuracy after data sampling, i.e., C4.5, we noticed that the number of incidents on beats #612, #631, and #632 without arrest were 1,327, 1,441 and 1,032, respectively. Therefore, we managed to output the classifier with predicted class for the test data, i.e., test data contains 100 instances (See Table 5). The results revealed that: (1) Beat #632 had the most misclassified instances; (2) Totals of 87, 87, and 85, and 79, 74, and 83 instances on beats #612, #631, and #632 contributed to accuracy and resulted in no arrests, respectively. There should be an increase in resources, especially in beat #632, to improve the number of arrests and prevent serious crimes that may result in physical harm.

A possible reason for the low arrest rate is that limited resources are used to patrol those communities with high battery and incident crimes. Not having enough features to distinguish between cases of arrest and others without arrest in the dataset is another potential factor. More importantly, balancing the dataset using SMOTE data sampling has not had a significant impact on the performance of the machine learning algorithms.

CONCLUSION

Law enforcement continues to maintain records of criminal incidents that are examined using advanced intelligent techniques to understand causes and effects and then develop resource allocation and prevention strategies. This paper investigated crime analysis using a real dataset to understand why arrest status is low at specific times and in various communities in the city of Chicago for offenses, such as battery and assault. To achieve the aim, in-depth clustering and visualization against the last three years' incidents (2017-2019) has been conducted to understand the correlations between timeframes, locations, and crime types. The knowledge learnt was then employed to produce a dataset with influential features that could be processed by machine learning techniques to produce classification models.

Findings based on feature selection pinpointed that features, such as Beat and Block are impactful. Moreover, 12:00 noon was identified as the hour when the most crimes were reported with 23,091, 22,848, 22,443 for 2017, 2018, and 2019, respectively. In terms of specific locations with a high level of crime, districts #1 and #18, CAs #44, #66, and #71, and Beats #631, #632, and #612 were classified with the top five crime types of theft, battery, criminal damage, assault, and deceptive practice.

A time series analysis was also conducted to show the crime trend will reoccur in 2020–2022. More importantly, the predictive analysis using the machine learning algorithms, i.e., Naïve Bayes, RIPPER, and C4.5, produced models from the subset of features retained with around 80.22% classification accuracy for the arrest status. The performance of these models is biased toward the majority class in the dataset because of the data imbalance issue, so these models are dropped. We balanced the dataset using SMOTE with ten-fold cross validation to produce more realistic classification models. The results obtained after data sampling by the machine learning algorithms were surprisingly low with C4.5 being superior with classification accuracy of 63.5%.

There is a huge gap between the rate of crime and the arrest status resulting in imbalanced data. However, this research demonstrates a successful approach to building a machine learning model to forecast arrest status using a subset of the Chicago crime data related to battery and assault. This

research is beneficial for Chicago PD to re-examine and model a service-oriented system linked with effective crime prevention and management. The department needs to allocate resources more efficiently to increase the possibility of crime prevention and thus reduce crimes in specific timeframes in recognised beats, districts, and communities as indicated earlier.

FUNDING AGENCY

The publisher has waived the Open Access Processing fee for this article.

REFERENCES

Abbass, Z., Ali, Z., Ali, M., Akbar, B., & Saleem, A. (2020) A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning. *2020 IEEE 14th International Conference on Semantic Computing (ICSC),* 363-368, doi:10.1109/ICSC.2020.00073

Ahishakiye, Omulo, Taremwa, & Niyonzima. (2017). Crime Prediction Using Decision Tree (J48) Classiðcation Algorithm. *International Journal of Computer and Information Technology*, *6*(3), 188–195.

American FactFinder. (2017). Retrieved from https://factfinder.census.gov/ faces/ nav/ jsf/ pages/index.xhtml

BBC. (2017). *Chicago Battling Violence with Crime Predicting Tech*. Retrieved from BBC News: https://www.bbc.com/news/av/technology-39748345/chicago-battling-violence-with-crime-predicting-tech

Bharati & Sarvanaguru. (2018). Crime prediction and analysis using machine learning. *International Research Journal of Engineering and Technology*, *5*(9), 1037–1042.

Breiman. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Brown. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108–132.

Bunker & Thabtah. (2017). A machine learning framework for sport result prediction. *Science Direct: Applied Computing and Informatics*, 27-33.

Castillo, M. L. (2015). Not so simple: How simple assault and battery became distorted in the context of crimes involving moral turpitude. *Washburn Law Journal*, 55.

Chalom, Vanderschueran, & Vezina. (2002). Urban safety and good governance: The role of the police. United Nations Centre for Human Settlements (UNCHS – HABITAT), International Centre for the Prevention of Crime (ICPC).

Charles, S. L. (2013). Understanding the determinants of single-family residential redevelopment in the Inner-ring suburbs of Chicago. *Urban Studies (Edinburgh, Scotland)*, *50*(8), 1505–1522. doi:10.1177/0042098012465908

Chawla, B., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th international conference on machine learning, ICML* (pp. 115–123). Morgan Kaufmann. doi:10.1016/B978-1-55860-377-6.50023-2

Cortes, C., & Vapnik, V. (1995). Support vector machines. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. doi:10.1109/TIT.1967.1053964

Crime at a Glance. (2020). Retrieved from New Zealand Police: https://www.police.govt.nz/sites/default/files/publications/crime-at-a-glance-jan2020.pdf

Crimes - 2001 to present. (2020). Retrieved from Chicago Data Portal: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Duda & Hart. (1973). Naive Bayes. *Pattern Classification and Scene Analysis, 3*.

FBI Supplementary Homicide Report 1980-2014. (2014). Retrieved from https://www.ojjdp.gov/ojstatbb/ezashr/

Galton. (1989). Kinship and correlation. *Statist. Sci., 4*(2), 81-86.

Halat, Saberi, M-Frei, A-Frei, & Mahmassani. (2015). Impact of crime statistics on travel mode choice: Case study of the city of Chicago, Illinois. *Transportation Research Record: Journal of the Transportation Research Board*, *2537*(1), 81–87.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* Academic Press.

Ho, K. (1995). Random forest - Document analysis and recognition. *Proceedings of the Third International Conference*, 1, 278-282.

IEP. (2019). Retrieved from Institute for Economics and Peace: IEP http://economicsandpeace.org/

Iqbal, Azmi, & Mustapha, Panahy, & Khanahmadliravi. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*, 4219–4225.

Kang,, H-W., & K., H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PLoS One*, *12*, 1–19.

Lloyd. (1957). K-means clustering - Least squares quantization in PCM. *IEEE Trans Inf Theory, 28*, 129–137.

Malathi & Babo. (2011). Algorithmic crime prediction model based on the analysis of crime clusters. *Global Journal of Computer Science and Technology, 11*(11).

National Institute of Justice. (2017). Retrieved from https://www.nij.gov/ funding/ Pages/fy16-crime-forecasting-challenge.aspx#data

Nguyen, Hatua, & Sung. (2017). Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology*, *8*(2), 141–146.

Quinlan. (1993). *C4.5: Programs for machine learning.* Morgan Kaufmann.

Rosenblatt. (1957). Multi-layered Perceptron & Artificial Neural Networks. In *The Perceptron, a perceiving and recognizing automaton Project Para.* Cornell Aeronautical Laboratory.

Saleh & Khan. (2019). Crime data analysis in Python using K-means clustering. *International Journal for Research in Applied Science and Engineering Technology*, *7*(4), 151–155.

Schapire, F. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933–969.

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention. *Vis. Comput. Ind. Biomed.* 10.1186/s42492-021-00075-z

Sharma, H. K., Choudhury, T., & Kandwal, A. (2021). Machine learning based analytical approach for geographical analysis and prediction of Boston City crime using geospatial dataset. *GeoJournal*. 10.1007/s10708-021-10485-4

Shermila, B., & Santiago. (2018). Crime data analysis and prediction of perpetrator identity using machine learning approach. *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, IEEE Conference Record: 42666*, 107-114.

Tableau Software, LLC. (2020). Retrieved from https://www.tableau.com/about

UCI Machine Learning Repository. (2012). Retrieved from Available from: http://archive.ics.uci.edu/ml/datasets.html

UCI Machine Learning Repository. (2017). Retrieved from Available from: http://archive.ics.uci.edu/ml/datasets.html

Vélez & Richardson. (2012). The political economy of neighbourhood homicide in Chicago: The role of bank investment. *British Journal of Criminology*, *52*(3), 490–513.

Yuki, Sakib, Zalam, Habibullah, & Das. (2019). *Predicting crime using time and location data*. Academic Press.