Early Risk Assessment and Linkage Analysis of Violent Crime Cases

Chamundeshwari Kottur
MSDS Data Science Practicum
Anderson College of Business and Computing
Regis University, Denver, CO, USA
ckottur@regis.edu

Abstract

Violent crime is one challenge that public safety institutions are constantly faced with, especially when investigations take a long time to be resolved. When government agencies begin to release huge amounts of crime data through open data programs, it often isn't that the data doesn't get used, but it's big, complex and doesn't have a solid framework for rigorous and repeatable analysis. As a result there's a lot of untapped info on how investigations actually go and how cases get cleared.

Through this practicum, i suggest a combined data science system of risk assessment at an early stage and violent crime cases analysis through linkage by use of the Chicago crime dataset. The project resolves two operational issues that are of interest to investigators and other public-safety organizations who are interested in finding at-risk cases that may turn into cold cases and detect past incidents which might be behaviorally, spatially and temporally associated with a new report(target case).

More than twenty years of publicly available is used in the current prototype which records Chicago crime incidents, assigns violent-crime categories, fills in missing latitude and Distinct at the district level, and filters analysis to a crime that involved force or threat of force. In the case of cold-case, risk assessment proposal makes the task a supervised proposal. classification issue in which a case is termed as a cold case when it has not been arrested and the case is more than a year old. Logistic regression, random forest, and hist-gradient boosting models are considered, and their performance is compared across ROC-AUC, PR-AUC, precision, and recall values.

through this practicum i intend to present a usable, and ethically conscious, analytical pipeline which provides insight into how supervised learning can be responsibly utilized in an approachable way to the understanding of crime data involving the population. The project will contribute to the existing academic knowledge of clearance trends and deliver actionable insights in terms of evidence-based discussion of the investigative issues without making or automating decisions.

I. Introduction

Violent crime is quite an elaborate issue to the public communities, the police and the general authorities. In addition to the short-term safety issues, the investigations may cause suspicions and inefficiency among the investigators and an irreparable social damage in the long-term when conducted on a long hanging basis. The growing need of people to experience openness has led to the publication of crime data on publicly accessible forums by a number of the federal and city departments. Despite the fact that these datasets are highly useful, the datasets are usually dirty and high-dimensional and thus hard to handle when not in the technical field over a long period.

Violent crime is associated with lengthy investigation periods, elevated levels of uncertainty, and social damages. Agencies have to choose what new incidents require intensive and immediate attention and what old incidents can be of significance to a new case pattern. The decisions are usually made when there are resource limitations, incomplete information and when decisions have to be made fast. Data science is not going to take the place of investigators, but may assist them identify those indicators that are hard to identify consistently at scale.

While working in data science, we are usually shown only bits and pieces, e.g., performing a quick exploratory plot, building a simple predictive model, or a dashboard. But practical considerations, especially those as weighty as the problem of public safety demand that the whole pipeline: of cleaning and

justifying the data, engineering features, establishing the results that we want, and that we are morally reading results. In this practicum, I will be taking the whole supervised learning process through a real life data with real stakes to the wider society.

The concept behind this practicum is the notion that historical incident data may be converted to operational support in two ways that are complementary to each other. To start with, historical variables can be utilized to determine the likelihood that a reported case would live to see another meaningful day instead of without being resolved. Second, previous incidents may be compared with one of the target case to determine possible linkable events by using common behavioral traits, local concentration, and time proximity.The practicum demonstrates a preliminary manifestation of this concept.

The relevance of this practicum is that most projects in the public-sector analytics field often terminate with descriptive dashboards. Instead, this work infers the next step over description to decision support: it approximates risks of cases, exposes associated cases, and displays the context of the investigation in a form, which can be checked and refuted by humans. The more general academic importance is shown in proving how structured tabular feature as well as text-based behavioral cues, geo-spatial distance, and time-related similarity can be combined into a single reproducible analytic process.

II. Problem Statement

Although there is an abundance of data on crimes in the public, we lack a systematic and machine learning oriented analysis, that unravels the violent crime clearance results immediately after the occurrence. Most research only considers simple statistics or mere graphs and overlooks the time, place and circumstances interaction to determine whether a case would be closed or not.

Crime data sets are also sloppy: they have strange data structures, gaps, after-the-fact added variables, and can cause havoc to a model if not handled correctly. A simple analysis is not sufficient, and we require a data-science workflow that includes all aspects finishing with the acquisition of the data all the way through the cleaning process, modeling and analysis. We can afford not to do that and people may not come out with the right things or overlook important patterns.

The gap that I am addressing as part of this practicum is the unavailability of reproducible, leakage-free supervised learning architecture, which can forecast the probability of an instance of violent crime being cleared based solely on the information we have upon the initial reporting of the incident. Doing that correctly requires doing all the steps of the data-science pipeline: loading in the data, cleaning up the data, creating good features, labeling the results, creating the model itself, validation checking its performance, and ensuring we understand what is taking place.

This is one way of getting it right in the society. A better understanding of why certain cases are closed more quickly than others can assist law-enforcement chat to discuss the actual drawbacks, spend the resources efficiently, and hold the populace to blame. It should be an ideal match with the cutting-edge tricks of data-science and cannot be resolved through a simple making of charts or the application of rule based rule sets.

This is a challenging issue due to four reasons. First, it runs on noisy data which is large and geographic omissions, changing crime names and extensive periods of time. Second, class imbalance should be the norm since cold cases are not evenly distributed between the offense categories and years. Third, it is multidimensional; crimes are similar in a number of fields determine the relationship between two incidents. Fourth, settings of operation demand and transparency, since the explanation of such a score may not be believed or utilized responsibly.

III. Related Work

Crime analytics and criminal investigation research has greatly advanced alongside the growth of crime datasets of large scale and the development of new and improved methods of data science. The past studies on criminology were based on the theoretical explanations of criminal behavior, whereas the

contemporary ones involve the applications of statistics, machine learning, and geographic information systems to determine crime trends and assist with investigation.

Investigative psychology is one of the constructions, established by [1], to study the criminal behavior. Investigative psychology is concerned with the way offenders behave and seeks to establish relationships among characteristics of crime. Based on this model, criminals tend to exhibit criminal consistency in their behaviors, and this has the potential to assist law enforcers in connecting cases that could have been done by the same person. This idea is the theoretical foundation of present-day crime linkage analysis.

[2] were the first to perform empirical research on crime linkage when they studied how crime incidences were related to the modus operandi characteristics. Their study showed that behavioral patterns including entry methods, choice of location, and mode of crime can be employed to statistically determine the relationship between two crimes. The participants used regression models and ROC analysis to determine the usefulness of crime connection using behavioral attributes. This publication formed a significant methodological basis of the study of crime linkage by computers.

The studies based on behavioral profiling have also led to the comprehension of the working mechanisms of offenders in relation to various crimes.[3] explain how behavioral profiling can be used during a criminal investigation and the significance of the analysis of tendencies in the description of crimes, the selection of victims, and moving. In their work, the authors emphasize the importance of integrating behavioral analysis with data-based approaches to aid in decision-making during the investigation.

Spatial analysis is very important in the analysis of crime patterns. According to research on environmental criminology, crimes do not occur randomly on geographic space but rather follow regular patterns in violation of the surroundings of urban life and human activity. According to [4], environmental criminology deals with the relationship between offenders, targets and places. The method emphasizes the effects of geographical characteristics like areas, routes to transportation, and locations of activities on crime. On the same note, [5] specify the ability of geographic information systems (GIS) to visualize crime trends and the formation of hotspots, and facilitate law enforcement planning.

As open government data has become more accessible, more and more scholars have used governmental information on crime in their analysis and prediction studies. Among the aspects addressed by Caplan, Rosenblat, and Boyd [6], open data initiatives and their role in enhancing transparency and facilitating approaches to the criminal justice research based on data are mentioned. The openness of crime data enables statisticians to investigate long-term trends in crimes and create the analytical model of predicting and preventing crimes. On the same note, the National Incident-Based Reporting System (NIBRS) offers more detailed crime reporting standards that allow a more profound study of the crime cases and offender behavior [7].

Recent studies have also examined predictive policing and machine learning methods of analyzing crimes.[8] have carried out randomized controlled trials that prove that predictive policing models have the ability to predict crime hotspots with the help of historical crime. These strategies employ statistical learning techniques to examine the spatial and temporal trend of criminal occurrences. These models give useful information to law enforcers as it can help them to distribute resources in a more effective way.

The crime data has also been subjected to text mining methods with crime descriptions and reports being the most analyzed.[9] coined the name frequency-inverse document frequency (TF-IDF) method that has been a popular method of designating textual information in numerical terms. TF-IDF allows comparing textual documents in terms of the significance of terms in a document as compared to a group of documents. This may apply in crime analysis in order to detect similarities in the crime descriptions.

The most recent advances in natural language processing have brought about word embedding algorithms, including Word2Vec to the concept by [10]. Such techniques encode semantic relays among words by coding them in the form of a vector space, which allows complex text analysis. These methods allow one to enhance the analysis of the crime reporting and behavior description in sets of investigation.

Research that specifically concentrated on Chicago crime data has also helped in comprehending the crime dynamics in the city. [11], the authors analyzed the crime trends in Chicago with methods of analysis of data mining. Their analysis revealed space and time crime patterns in the city and how a

visualization and interpretation of any given crime trend can be made through analytical approaches. The paper demonstrates the importance of large scale city crime data in building insights about criminal behavior using data.

Current study forms the continuation of these earlier works by incorporating the spatial analysis, modelling of behavioural similarity, and clustering technique to study patterns of violent crime. The suggested framework helps to define possible connections between crimes and offer a multifaceted analytical strategy of the crime data exploration by using geography, time-based characteristics as well as textual descriptions of crime.

IV. Methodology

The methodology follows a 6-stage structure where the data is prepared, violent-crime subsetting, cold-case risk modeling, exploratory clustering and target-case linkage analysis. The whole end-to-end workflow is written in Python such that every transformation produce the final tables and plots as well as map outputs, which are based on the original Chicago crime file.

Data Collection

I retrieved some publicly available crime data out of open data portals of the chicago city crime data portal using public API. I concentrated on the grave cases such as homicides, robberies, sexual assaults and kidnappings. These data span several years and include time and place of each crime, as well as information regarding each crime. All anonymous, hence no problem with research.

Data Cleaning and Preprocessing

During the preparation phase, the names of columns are normalized and category names are standardized. The cold-case modeling implemented, has a substituted old label CRIM SEXUAL ASSAULT with Criminal Sexual Assault and has no latitude and imputes missing data by dropping records that still have not obtained critical values and use the values of longitude and latitude with the median of districts fields. Violent-crime category has such categories as assault, sex offense, stalking, kidnapping, robbery, battery, homicide, arson, human trafficking, criminal trespass and crimes against children.

Features have primary type, description, location description, domestic indicator, beat, district,latitude, longitude, days of age of cases. On sorting, a time-aware split is then produced, age of cases and the eighty percent of the cases of the training to go to the oldest and the remaining to go to the training then testing.

cold case definition

To identify the clearance risk I chose to apply a 12 months rule where; the requirements are not caught within a year of crime, then the delayed or not resolved mark is made.

In cold-case task, the response variable cold-case refers to 1 representing arrest as false and if case age is more than 365 days, it is 0 otherwise. The prototype calculates the age of cases with the use of distinct differences in the highest observed case date and every incident date.

Supervised Learning and Baseline Modeling

I plan to conduct a regularized logistic regression initially to verify whether the intake characteristics are indeed predictors of clearance. I will use precision-recall curves and other measures which are good with a single class that is significantly less common.

Fields that are categorical, The pipeline used processes the field casts values to strings and imputes UNKNOWN on demand, and can employ one-hot encoding. Meanwhile, numerical characteristics include median. imputed and standardized.

link-ability risk i formulated a formula using behavior similarity, spacial proximity and temporal consistency to calculate a linkability score to find a potential pattern between criminal cases.

Ethical Considerations

I ensured that I never attempted to determine who is guilty and their behavior. There are no fields which are entered into the model and my results are introduced as scholarly knowledge, not guidelines. discuss the boundaries and prejudices in order to make everything straight.

V. Dataset Description

The data employed in this project is obtained in this portion of the Chicago Crime Data Portal which offers publicly accessible records of reported crimes in the city of Chicago. The data consists of a number of years of crime statistics gathered by the Chicago Police Department.

The data set consists of different attributes that characterize every crime incidence. The use of the following variables in this study will include:

- ID – Unique identifier for each crime record
- Date – Timestamp indicating when the crime occurred
- Primary Type – Main classification of the crime (e.g., assault, robbery, battery)
- Description – Detailed description of the offense
- Location Description – Type of location where the crime occurred
- Latitude and Longitude – Geographic coordinates of the crime location
- Beat and District – Police jurisdiction identifiers
- Arrest – Indicator of whether an arrest was made
- Domestic – Indicator of whether the crime was domestic in nature

To present the data according to the aim of the current study, it was narrowed down to include the categories of violent crime, which included assault, robbery, battery, homicide, and crimes involving children. The crimes were further filtered to target the recent records within a selected time range to guarantee relevance in terms of time.

Various preprocessing procedures were done before analyzing. Missing geographic coordinates were estimated by means of mining district medians, and still provided the possibility to conduct spatial analysis. To ensure uniformity in dataset, duplicate and incomplete records were eliminated.

The processed and filtered dataset had thousands of crime events that could be subjected to clustering and linkage analysis.

VI. System Architecture

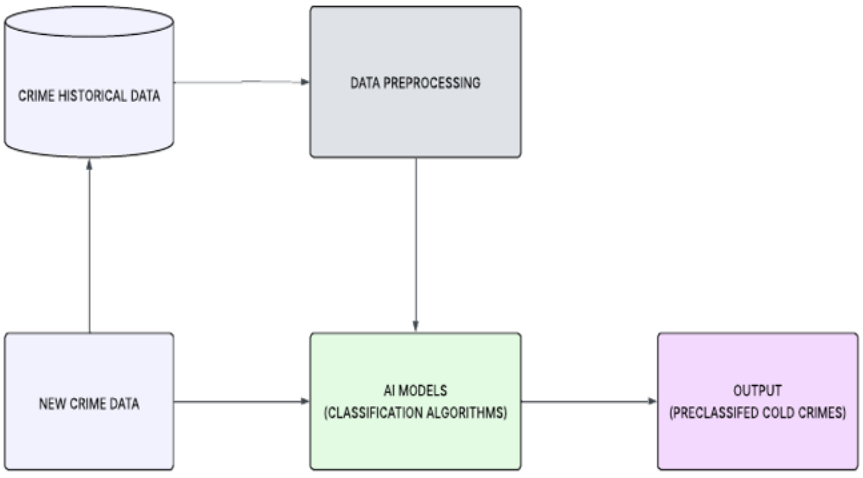

Fig. 1. flow diagram

The offered system aims at analyzing violent crime data by applying a set of data preprocessing, feature engineering, clustering method and linkage analysis. The general architecture is a multi-stage pipeline involving the processing of the raw crime records and generation of analytical data to give indications of the crime connections.

The initial architecture process is the acquisition and pre-processing of data. The records of crime that take place in the past have been acquired using the Chicago open data portal that provides organized datasets of crime data like date, location, nature of the crime, and arrest status. A Python environment loads the data into it with data analysis libraries. On the preprocessing, in missing cases of geographic coordinates, a median imputation at the district level is used. Data quality is ensured by deleting rows whose records are not shared with other records in the same table.

The second one creates feature engineering and transformation. The dataset is extracted to give spatial features which include latitude, longitude, district, and beat. The year field, month field, day of week field and hour are some of the temporal aspects based on the crime date field. Behavioral features are produced through the use of text descriptions of crimes and locations to produce behavioral signatures. These characteristics allow the system to obtain various dimensions of criminality.

The third step does data exploration and data visualization. Statistical menus and formulas of distribution are created to make the sense of crime patterns within locations and types. Relationships between crimes types and location descriptions are analyzed with the help of heatmaps and bar charts. Visualizations can be used to determine the patterns of crime that are predominant and possible hotspots.

The fourth phase uses machine learning methods. One-hot encoding is used to transform the numerical and categorical features into machine-readable numbers and feature scaling is used to normalize them. The K-Means clustering algorithm uses this to categorize crimes together sharing similar features in space and behavioral features. A cluster size of the optimum is calculated based on the elbow method.

The fifth phase puts in place the crime linkage analogy. Based on a historical case, a linkage score is calculated according to a target case. The score incorporates three factors, the similarity of behaviors, measured with the TF-IDF and cosine similarity, spatial proximity measured with the Haversine distance formula and time consistency measured with circular time distance. A combination of these features is done through weighted aggregation to give a resulting linkage score.

Lastly, the system produces images and interactivity maps. Geographic representations show clusters and possible connected cases on the forms of scatter diagrams and interactive maps. These products give an intuitive test to the investigators on spatial relationships among crimes.

VII. Outcomes

The main focus of the project is to reveal the idea of how data science methods can be applied to help the crime analysis in terms of establishing patterns within the violent crime statistics and recognizing possible correlations between the incidents. In the proposed framework, machine learning models with the use of spatial analysis and behavioral similarity are integrated to analyze historical crime data and find potentially related cases.

A. Performance of Machine Learning Model

In order to measure the performance of the predictive modeling, a set of fielded machine learning algorithms were laid upon the filtered violent crime data. The models were then trained to scan our dataset based on spatial or categorical variables i.e. district, beat, latitude, longitude and crime type. Three models were compared: the Logistic Regression, the Random Forest and the Histogram Gradient Boosting.

These models were evaluated on other standard evaluation measure such as accuracy, precision, and recall. The findings of these models are presented in table below.

The model that was found to be most accurate and using the highest amount of recall, was the Histogram Gradient Boosting model, showing that it is indeed effective in determining patterns within the dataset. RF also proved to be highly effective, especially in the recall aspect, which is relevant in the crime analysis situation where it is desirable to identify as many meaningful cases as possible. Logistic Regression was used to carry out a strong baseline model that was found to be somewhat less accurate than the ensemble-based approaches.

Model	PR-AUC	Precision	Recall
LOGISTIC REGRESSION	0.7794	0.7579	0.8574
RANDOM FOREST	0.7933	0.7463	0.9344
HISTOGRAM GRADIENT BOOSTING	0.7961	0.7417	0.9801

TABLE I
MACHINE LEARNING MODEL PERFORMANCE

The findings indicate that ensemble based learning is more appropriate in generalizing complex relations in a crime dataset than linear technologies.

B. Results of the Linkability Analysis

Along with predictive modeling, a crime linkage analysis structure was also adopted in this project and it aimed at determining historical crimes, which could be connected to a target crime. The linkage analysis represents a combination of three main components:

Crime description and location description based on the use of TF-IDF and cosine similarity to extract behavioral similarity.

Spatial proximity is measured based on Haversine distance of geographic coordinates.

The temporal consistency and the distinction between the time of occurrence of the incidents.

The weighted aggregate is used to form a single score which is the linkage score. Crimes, which are more linked are regarded as more likely to relate to the target case.

The system was able to find the 10 most linkable cases in historical records in the data. These instances showed high degree of similarity in the behavioral features, geographical location and time pattern towards the target case. The comparison of these cases in terms of geographic mapping showed that the majority of associated offenses happened within close-by districts, which proved the hypothesis that criminals tend to be in well-known spatial locations.

The outcomes of the analysis of linkages prove that the use of behavioral, spatial, and time characteristics types in identifying criminals of potential relationships is efficient. The linkage scores generated offer investigators with a line of incident rank that might be important to investigate more.

Also, a K-Means algorithm to cluster the data showed the spatial groupings of crimes in the dataset. The clusters point to higher crime rate geographic hotspots where violent crimes are more prevalent and therefore give good understanding of how the geographic distribution of crime is in the city.

On the whole, the findings suggest that combining machine learning methods with spatial and behavioral analysis may contribute greatly to the exploration of crime data and provide an investigative process. The suggested system evidences the potential help of data-based techniques in detecting crime trends and possible case interconnections in large city crime records.

REFERENCES

[1] D. V. Canter and D. Youngs, *Investigative psychology: Offender profiling and the analysis of criminal action*. John Wiley & Sons, 2009.

[2] C. Bennell and D. V. Canter, "Linking commercial burglaries by modus operandi: Tests using regression and roc analysis," *Science & Justice*, vol. 42, no. 3, pp. 153–164, 2002.

[3] C. R. Bartol and A. M. Bartol, *Criminal & behavioral profiling*. Sage, 2012.

[4] M. A. Andresen, "The place of environmental criminology within criminological thought," in *Classics in environmental criminology*, Routledge, 2010, pp. 21–44.

[5] S. Chainey and J. Ratcliffe, *GIS and crime mapping*. John Wiley & Sons, 2013.

[6] R. Caplan, A. Rosenblat, and D. Boyd, "Open data, the criminal justice system, and the police data initiative," *Data and civil rights: A new era of policing and justice. Washington, DC: Data*, pp. 1–13, 2015.

[7] L. A. Addington, "National incident-based reporting system (nibrs)," *The encyclopedia of research methods in criminology and criminal justice*, vol. 1, pp. 88–91, 2021.

[8] G. O. Mohler et al., "Randomized controlled field trials of predictive policing," *Journal of the American statistical association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] A. Garima, A. Alaiad, et al., "Crime analysis in chicago city," in *2019 10th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2019, pp. 166–172.

[12] P. J. Cook and A. Mancik, "The sixty-year trajectory of homicide clearance rates: Toward a better understanding of the great decline," *Annual Review of Criminology*, vol. 7, no. 1, pp. 59–83, 2024.