Making Arrests

A Machine Learning Approach to Predicting On-Site Arrests for Crimes in the City of Chicago

Author:
Jiaqing Zhao
jz2902@columbia.edu.

Faculty Advisor:
Michael D. Parrott
mp3675@columbia.edu

Master'sThesis
Quantitative Methods in the Social Sciences
Columbia University
December 30th, 2018

ABSTRACT

Researches have suggested that decreasing crime rates will lead to a multitude of economical and social advantages. However, existing literatures which attempted to achieve predict crime using machine learning techniques lack the granularity and accuracy to produce real world impacts, granting these researches no advantages over traditional predictive crime prevention techniques. This thesis is the initial attempt on taking down the path of predicting arrests instead of crimes themselves since previous literatures have proven a deterrence effect of high arrest rates can decrease crime rates. We hypothesize the main variables that can determine arrests are 1) distance to closest police department and 2) the traffic of surrounding areas. After modeling arrest using crime data provided by the City of Chicago, we obtained an AUC-ROC score of 0.92 from a tuned random forest model, showing that arrests can be predicted accurately and therefore it is possible for law enforcements to maximize chances of arrests using this line of research.

TABLE OF CONTENTS

I. BACKGROUND

a. CRIMES, PREVENTIONS AND MAKING ARRESTS

Crimes and Their Impacts

We study Crimes because they bring too many harms to the world. Not only crimes directly harm the victims involved in each of the bugler, robbery even murder at a micro level, but also crimes act as an active part in the economic down-cycle (Mehlum et al., 2005) at a macro level. There are many reasons contribute to the economic decline brought by high crime rates.

First of all, fighting crimes is expensive, as case studies on Latin American cities show that the cost can sometimes be above 7% of GDP (Londono and Guerrero, 2000.) In the US, the cost burdens caused by criminal activities are also significant. Case studies have shown that medical care for assault victims alone is estimated to be $4.3 billion per year (Robert J. Shapiro and Kevin A. Hassett, 2012.)

Second of all, apart from economic burdens on taxpayers and governments, human resources become scarce as people would avoid moving into or traveling to a city with high crime rates (Arulanandam et al., 2014,) damaging both tourism and other industries. Furthermore, crimes can break a community's sense of unity by forcing people to habitually avoid certain places, creating disorders and chaos (Ahishakiye et al., 2017.) These problems will induce poverty, a critical factor that brings more

individuals into criminal activities thus potentially creating a vicious cycle (Currie, 1997; Fajnzylber et al., 2002.)

Therefore, attempting to lowering crime rates seem to be a crucial task for any crime-heavy cities or nations as reducing crime rates can reverse economic burdens. A case study done by Shapiro and Hassett indicated that, a goal of savings which range from “$6 million per year in Seattle” to “$12 million per year in Boston and Milwaukee", to “$42 million per year in Philadelphia and $59 million for Chicago” can be achieved “if their rates of violent crime declined by either 10 percent or 25 percent.” These costs are directly tied to “lower expenditures on law enforcement and the justice system,” as well as the additional tax revenues that each city could expect to collect from those who otherwise would have been “victims or perpetrators of those crimes.” These numbers didn’t even include benefits like lowered out-of-pocket medical costs for those potentially saved victims of crimes (Robert J. Shapiro and Kevin A. Hassett, 2012.)

Shapiro and Hassett also pointed out that the most significant economic benefits can arise from housing values that would otherwise be affected by high crime rates. In their analysis, they found that on average, a reduction of one homicide in a zip code for a given year raise the housing values by 1.5%. Estimations showed that with a 10 percent reduction in homicides, Jacksonville would increase $600 Million. Other cities like Philadelphia (as well as the surrounding suburbs) and Boston can raise $3.2 billion to $4.4 billion (Robert J. Shapiro and Kevin A. Hassett, 2012.) These budgetary savings and benefits can be crucial in today’s tight fiscal and economic environment.

What is more critical, such benefits can be easily projected beyond a single beneficiary. In a report by Sarah Ford from the Australian Crime Prevention Council, she described “Benefit Diffusion” which refers to a distribution of benefits beyond specific place, individuals or time frames that were initially targeted given with proper crime prevention policies (Ratcliffe & Makkai 2004.) Such benefits usually are not only monetary gains but also psychological effects that bring people satisfaction and happiness. Sarah pointed out that crime prevention policies have been shown to be beneficial in reducing not only crimes but also fear of crime (Cozens et al. 2005.) Employing crime prevention policies brings lower fear of crime, increased property values (in line with Shapiro and Hassett’s report) and in turn, higher quality of life (Schneider & Kitchen 2002.) It would also be reasonable for one to expect that through crime prevention, orders can be restored in communities thus bring back the union of people that once disrupted by criminal activities.

In short, we study crimes because we want to be better at reducing the wrongs crimes bring. The goal can be attempted through designing, comparing and employing various crime prevention policies which will not be discussed with detail in this thesis. With the rise of computation power and data science approaches, such crime prevention policies can be assisted with a data-centered predictive approach aiming to predict crimes before they happen.

Predictive Nature of Crimes and Crime Prevention

On a broader level, crimes are predictable. Researchers have found predictable patterns and traits among criminal activities. For example, criminals tend to operate in their comfort zone. That is, as the RAND report points out, “they tend to commit the type of crimes that they have committed successfully in the past, generally close to the same time and location”(L. Perry et al., 2010.) Many major theories of criminal behavior, such as routine activity theory, rational choice theory, and crime pattern theory support such view (Clarke and Felson, 2004.) Merging these theories into what the RAND researchers call a “blended theory” gives us a detailed picture of how they depict criminal behaviors:

1. *Criminals and victims follow common life patterns; overlaps in those patterns indicate an increased likelihood of crime.*
2. *Geographic and temporal features influence the where and when of those patterns.*
3. *As they move within those patterns, criminals make “rational” decisions about whether to commit crimes, taking into account such factors as the area, the target’s suitability, and the risk of getting caught.*

— L. Perry et al., 2010

The primary interpretation of the blended theory is to recognize the fact that crimes form in clusters. Criminals’ behaviors are indeed related to certain characteristics of their environment in which they grew and lived. Factors like exposure to specific peer networks and neighborhood influence contribute to criminals’ actions

(Weinberg, 1954.) This means that criminal activities can be closely related to social-economic variables such as education (Ehrlich, 1975), ethnicity (Best and Braithwaite, 1990), income level (Patterson, 1991) and unemployment (Freeman, 1999), which make crimes very likely to be clustered: a statement confirmed both empirically in our daily life and academically (Weisburd and Lorraine, 1993).

These theories and findings, deeply rooted in the fields of criminology, sociology, psychology, and economics, are usually used to predict the following four categories, according to RAND's report:

1. *Methods for predicting crimes*: These are approaches applied to forecast places and times with an increased risk of crime.
2. *Methods for predicting offenders*: These approaches identify individuals at risk of offending in the future.
3. *Methods for predicting perpetrators' identities*: These techniques are used to create profiles that accurately match likely offenders with specific past crimes.
4. *Methods for predicting victims of crimes*: Similar to those methods that focus on offenders, crime locations, and times of heightened risk, these approaches are used.

These four categories, however, aren't the type of predictions one would expect from a modern machine learning approach. They are either generalized to a group of people or region (like 1, 3 and 4) or too specific on certain individuals (like 2) which is almost like "overfitting" in a machine learning perspective. Generalizations like this are

useful and effective on a higher level for policy implementations, but they lack the granularity to deal with specific situations. For example, there can be a discrepancy in crime patterns between smaller and larger geographic levels (streets vs. communities) as studies show that sometimes streets with strong crime concentrations can be in a commonly perceived good neighborhood while the seemingly bad neighborhoods can be free of crime (Weisburd and Lorraine, 1993.) Opposing to the machine learning approach of aiming accuracy, such predictions should, in fact, avoid accuracy which is listed as the number-one pitfall of predictive policies in the RAND report since accuracy is often a trade-off to location precision in these applications.

Past Works on Crime Prevention using Machine Learning

Since the bloom of machine learning and data science, many researchers attempted to marry crime prevention with machine learning. The trend in this line of research is to use enrichment data acquired through all sorts of technical means (internet, mobile networks, etc.) to predict future crime density for a given region.

So far, a relatively successful and popular approach to predict crimes is using Twitter data such as studies done by Wang et al. (2012) and Gerber (2014). They used sentiment analysis/topic grouping on twitter contents and tried to predict crime levels of given regions. Gerber obtained an AUC score of 0.71 and Wang also obtained a model performance better than random. In another study Crime Prediction Using Twitter Sentiment and Weather (2015), Chen, Cho, and Yang took the idea a little bit further by taking into considering both sentiments of tweets and weather. They

hypothesized that more negativity in text sentiments, as well as more extreme weathers, are correlated with more theft. They used twitter data, weather data and crime data from the city of Chicago during a specific time period and trained a logistic regression model and obtained a testing AUC of .67 (Chen, Cho, and Young Jang, 2015) which has no improvements over Wang et al. ‘s result.

Besides Twitter data, Bogomolov et al. used data from the mobile network which can tell the number of people in each cell (a particular geographic division of the London Metropolitan Area) as well as their gender, age and status (residents, workers, visitors, etc.) in London. They used PCA to reduced feature space and fed the data into a random forest model. They achieved an AUC score of 0.64 (Bogomolov et al., 2014) which still isn’t very impressive as well.

These studies shed lights on how to approach crime prediction by combining data mining methods and modern network datasets. However, they suffer from the following problems:

1. Enriched datasets (like mobile network) seem to be a little bit of a stretch. Although most of the times people understand why did they choose certain enrichment dataset, there still aren’t much explanations and theoretical background.
2. Predicting crimes on a regional level is not granular enough to make meaningful machine learning predictions. As machine learning methods are undoubtedly competent in these prediction tasks, they probably aren’t more effective than existing predictive policy strategies.

3. They lack model performance that can give actionable insights. Namely, their ROC-AUC scores are so low such that their models cannot be employed by law enforcements and researchers.
4. They lack explanatory analysis on what are their models telling us. They usually keep the models as a black-box.

Predicting Arrests: Why is it Better?

Aiming to marry crime prevention and machine learning further, this thesis will focus on the prediction of another variable: on-site arrests instead of predicting crimes themselves. The first and foremost reason is that increasing arrest rates can also reduce crime rates.

Many studies confirm a negative correlation between crime rates and arrest rates. For example, in his study, Levitt attributed the negative correlation to deterrence effect and incapacitation effect. These effects can be in short described as the fear factor that prevents criminals to engaging in more criminal activities due to afraid of penalties, and deterrence effect seems to be much more important as an influence empirically than incantation effect (Levitt 1998.) “If the deterrence effects found in this paper are correct,” states Levitt, “then one would expect to see reductions in crime.” Moreover, since deterrence effect is stronger than incapacitation, the reduction in crime would come with only small increases in prison populations in the states where a “three strikes” type of policy has been implemented. A real world example would be that,

after passing a “three strike” law, California saw a 7.2% decline in violent crimes in the first year after the policy was implemented.

Furthermore, making arrests is the second best outcome in the crime-fighting scene while the best scenario is preventing crimes before they happen. If a crime has occurred, knowing the best strategy to make arrests is much better than letting criminals on the loose. Please note that for the sake of simplicity we are giving simple assumptions like most of the time police arrest the actual criminals. We are also only looking at on-site arrests, since other types of successful arrests can be due to completely different and specific reasons (e.g. solving murders.)

Predicting arrests can also make our lives easier when building the predictive model. For example, arrest or not is a boolean variable present in our dataset that is feasible to predict on while “a crime happened or not” isn’t something that can be predicted directly. The prediction outcomes are also granular enough to have an impact on every single incident of crime while not losing the power to apply such models to crimes or different attributes.

b.WHY CHICAGO? AN OVERVIEW

Challenged by crimes like homicide, armed robbery, gang violence, and aggravated battery for the last 50 years, The Chicago Police Department's Bureau of Records has been tracking Crimes in Chicago since the beginning of the 20th century, and the numbers have not been promising. The city’s numbers, regardless if it is violent crime rate, murder rates, rape rates, or assault rates, are much higher than national

averages (City-data.com, 2018.) By 2010, Chicago's homicide rate had surpassed that of Los Angeles (16.02 per 100,000 population) and was more than twice that of New York City (7 per 100,000 population)

Chicago's homicide rate had surpassed that of Los Angeles by 2010 (16.02 per 100,000) and was more than twice that of New York City (7.0 per 100,000) in the same year (Gorner, 2014.) By 2016, Chicago recorded more homicides and shooting victims than New York City and Los Angeles combined (Gorner, Nickeas and Malagon, 2016.) In the same year, Chicago was responsible for nearly half of 2016's murders in all US (Sanburn, 2016.) As pointed out by the TIMES article, Chicago Is Responsible for Almost Half of the Increase in U.S. Homicides, Chicago is the only major U.S. city to report an increase in murders in both 2015 and 2016 during which the nation's crime rates remain near historic lows (Sanburn, 2016.)

The reason behind the high crime rates is often thought to be the pervasive infestation of street-gangs. Estimations showed that there are over 100,000 active gang members from almost 60 factions (ABC News, 2012.) Gangs were responsible for 61% of the homicides in Chicago in 2011 (Home.chicagopolice.org, 2011.) That been said, the situation in Chicago is still very complicated as researchers have failed to capture the reasons behind the recent murder rate spike in 2016 (Ford, 2017.) Because there haven't been enough time-series data on the different social aspects that can present a clearer picture of motions of criminal behavior in Chicago. "What researchers don't know is vast," as Ford points out, "A breakdown in police-community relations after the McDonald shooting could have played a role, but without regular public surveys on the issue, there are no reliable data to prove the effect."

Thus, Chicago is one of the perfect samples to study crimes. It has a high-volume of crime data of different variety, complicated reasons and interactions behind crimes and a dire situation of which the city's law enforcement wants a solution.

c. WHY A MACHINE LEARNING APPROACH?

In this thesis, I will be approaching this topic from a machine learning approach as opposing to a social science approach. A social science approach is when researchers choose their hypothesis and variables carefully based on previous literature and utilize interpretable models like regressions. This thesis will be more of a machine learning project for the following reasons:

1. We have abundant data on single crime incidents to perform techniques like data-mining/machine learning since the City of Chicago has recorded an ample amount of Crime data. Adopting data-mining/machine learning techniques means that the models are usually hard to interpret.
2. Machine Learning methods can provide the granularity of accuracy needed to understand single incidents.
3. If we want to approach crime prevention with a social science perspective, it would take a much longer time and more effort to uncover the complicated crime situation in Chicago. The most challenging part may be finding and collecting a multitude of social data (public, surveys, etc,) making the project unrealistic to finish.

4. Predicting chances of arrests for a single crime incident has not been done before, meaning that many of my choices on which variables to include to the model will be solely decided by myself, either using variable selection techniques or simple assumptions. Social science research can't justify these choices without presenting proper theoretical supports. I also do not want my model design to be constrained by past studies as well because I want to be explorative.

d. HYPOTHESIS/PREDICTIVE GOALS

Aiming to combine crime prevention and machine learning, this thesis will look at factors that predict higher chances of on-site arrests in the hope of generating an accurate model that can predict whether an **on-site arrest** (will be referred to as on-site arrest or simply arrest) will be made. More specifically, I hypothesis that whether or not an on-site arrest can be made is related to:

1. The distance of the incident to the nearest police department
2. Traffic

The reason behind this is the simple assumption of how fast can the police arrive at the crime scene. Ideally, the nearer the police departments to the incidents and the faster the traffic, the higher chance polices will make arrests before criminals escape. This is the kind of assumption that cannot be backed up by any preceding

researches and I will not officially include the direction of relationships in my hypothesis as my models will not have interpretable coefficients.

My models will have other control variables which will be presented in coming sections.

II. DATA/METHODOLOGY

a. DATASETS

City of Chicago Crimes - 2001 to present

This dataset is a log of all the crimes reported (except murders where data exists for each victim) in the City of Chicago from 2001 to the present. On the City of Chicago website, the records were updated on December 23rd (same day as I am writing this) and the whole dataset gets updated on a daily basis. I downloaded the dataset on September 12th. The origin of this dataset is from Chicago PD's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Each row is a crime incident with records like dates, time, type of crime, location and etc. For a sample of this dataset please see Appendix or download it from: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

The copy of dataset I currently use has 6,695,947 rows and 22 columns. Essential columns are Date, Primary Type, Description, Location Description, Domestic,

Community Area, Arrest, Year, Latitude and Longitude. For a complete data dictionary please see Appendix.

This is the main dataset I will be using in this thesis. All other dataset will serve as enrichment datasets that get matched to records in this main dataset.

City of Chicago Police Departments

This dataset contains the locations of all Chicago Police Departments, published officially by the City of Chicago. There are 23 police stations in this dataset and there is only one station per district. For a sample row please download from https://data.cityofchicago.org/Public-Safety/Police-Stations/z8bn-74gv.

This dataset contains 15 columns and the most essential columns are the longitude and latitude columns which we can use to identify their exact locations. Unfortunately, this dataset does not include anything about attributes of these police stations other than location and telephone number. This way we won't be able to know information like number of officers per police station or number of police cars per station.

Chicago Traffic Tracker - Historical Congestion Estimates by Region - 2013-2018

Released officially by the city of Chicago, this dataset contains the historical estimated congestion for the 29 traffic regions from January 2013 to May 2018. This dataset contains rows of standardized speed for a given region, of a given time and

date. For a sample row please see download from https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/emtn-qqdi.

The standardized speed calculation was not explained by the City of Chicago. From their descriptions of the columns we can think of it as some sort of "standardized speed" that standardizes congestion level with speed limits such that roads with different speed limits can be compared with each other. The downside of this dataset is that it only covers time period of 2013 to 2018, which is a smaller time interval than that of our main dataset.

Demographics in Community Areas

The demographics data of Race by Poverty level of given community areas in Chicago. This is obtained by downloading and cleaning the 2010-2016 decennial census. This data has been difficult to locate but "Chicago Data Guy" (http://robparal.blogspot.com/) parsed the Chicago portion out of decennial census and some of the American Community Survey of 2006-2010. I want to show my appreciation to whoever is running this blog that saved me a huge amount of time.

Please see a sample of this data in the Appendix, or download from http://robparal.blogspot.com/.

b. MAIN PREDICTORS

Distance to the Nearest Police Station (*min_pd*, continuous)

We hypothesized that distance to nearest police stations affects how fast can the police officers arrive at crime scenes. Knowing the location of each crime, and the location of each police department, distance to the nearest police station is calculated through the following method that is rudimentary in calculating any distance between two points:

$$d_{jmin} = min\{d_{j1}, d_{j2}....d_{ji}\} \ \ where\, d_{ji} = \sqrt{(longitude_j - longitude_i)^2 + (latitude_j - latitude_i)^2}$$

Where j is a given crime location and i is a given police station, which in this case is from 1 to 22. After the min_distances are calculated, they are normalized and centered to 0.

Traffic (*avg_speed*, continuous)

Traffic data was obtained from Historical Congestion Estimates dataset. Ideally, we should have traffic information by road, time and day for the entire time span of 2001 to 2018. Unfortunately, not only our data is generalized to a region level (each region in Chicago contains about 2 - 3 community areas.) Thus I made a bold assumption that for any given day and time, the traffic situation of a given area is similar throughout years. This is a very weak statement that comes from empirical experiences: although sometimes one may experience traffic anomalies due to events like constructions, most of the times one would expect similar travel time for a familiar route at a routined time of day. Thus, the way I am using this dataset is to average

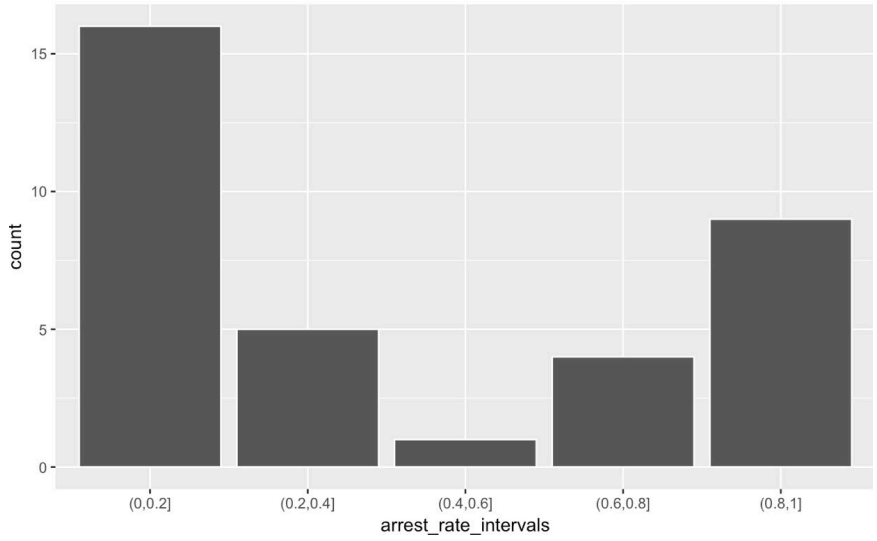
speeds throughout the years of a given region after grouping by date and time, and join this information to the main dataset on date and time. For example, we assume that the traffic speed for the surrounding region of an incident happened on 2010/1/1 at 12:00 p.m. is actually the average speed of that region for all 1/1 12:00 p.m.s across 2013 to 2018.

c. IMPORTANT CONTROLS:

Primary Crime Type (*Primary.Type*, Categorical)

Different types of crimes have dramatically different arrest rates. From Graph 1 we can see that the distribution of arrest rates for different types of crime is clearly bimodal. In Graph 1 the axis shows different brackets of arrest rates, from 0 ~ 20% to 80% ~ 100%. The y axis is the count of crime types that falls into each category. Looking at the crimes where solve rate is almost 100% (See Output 1,) it is understandable why would they reach 100% arrest rate. For example, sometimes police have planned strikes on narcotics and prostitution for a long time so they may only attempt arrests when they are very confident. Some of these crimes require a more stationary location, like gambling, making it easier to make on-site arrests. These crime types, since they already have a perfect arrest, is not worth predicting.

Graph 1. Histogram of How Many Crimes Types Fit into Each Interval of Arrest Rates



An anecdotal fact is that there is only one incident recorded as domestic violence. Most of the domestic violences are recorded as other types (like “ASSAULT”) with a “true” under column “Domestic.”

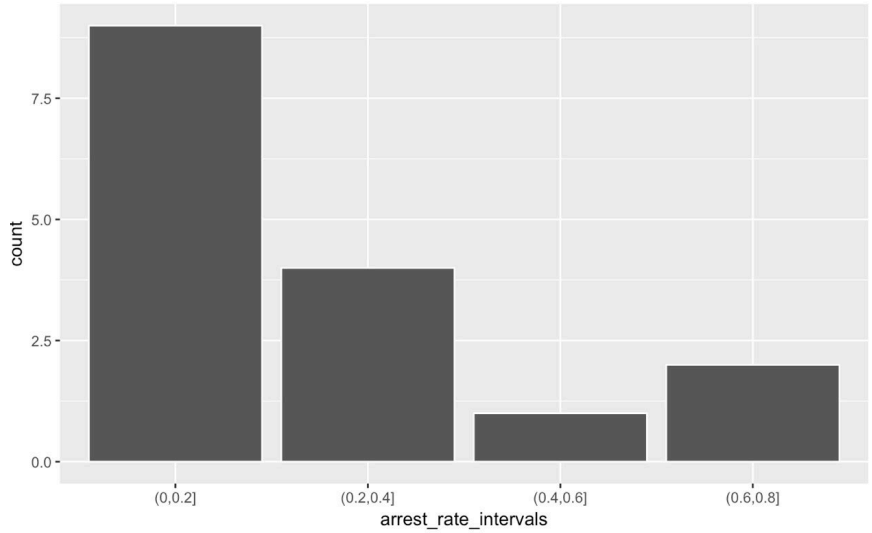
Output 1. List of all Crime Types with an Arrest Rate Larger Than 99%

type	arrest_rate
DOMESTIC VIOLENCE	1.00000000
PROSTITUTION	0.99606888
NARCOTICS	0.99405238
PUBLIC INDECENCY	0.99371069
GAMBLING	0.99277175
LIQUOR LAW VIOLATION	0.99107398

After taking out crimes with extremely high arrest rates and those with less than 1000 occurrences, we are left with the following crime types: "BURGLARY", "MOTOR

VEHICLE THEFT", "ROBBERY", "KIDNAPPING", "THEFT", "ARSON", "CRIM SEXUAL ASSULT”,

Graph 2. Adjusted Histogram of How Many Crimes Types Fit into Each Interval of Arrest Rate



"STALKING", "DECEPTIVE PRACTICE", "INTIMIDATION", "OFFENSE INVOLVING CHILDREN", "BATTERY", "ASSAULT", "SEX OFFENSE", "HOMICIDE","CRIMINAL TRESPASS”, and "WEAPONS VIOLATION.” Our target crime types include 7 of the 10 most frequently happening crimes (Robbery, Deceptive Practice, Motor Vehicle Theft, Burglary, Assault, Batter and Theft.) Given that The other 3 include Narcotics which has a 99% arrest rate, accounting for this seven should be meaningful enough. The adjusted histogram is no longer a bimodal distribution (See Graph 2.) The selected crime type accounts four about 70% of the total crimes. We can see from graph 2 that the arrest cases are highly imbalanced.

Domestic (*Domestic*, Boolean)

As previous evidence suggest that it may be easier to make on-site arrests with domestic cases yet this is not always true. Domestic incidents have an arrest rate of 19.8% while non-domestic incidents have 29% arrest rate. Several reasons behind this may be that the non-domestic incidents have a much higher count than domestic ones (5,818,975 non-domestic incidents Vs 876,972 domestic incidences.)

When we look closer we find some interactions between Domestic and crime types. While some crime types have no domestic incidences at all, some crime types show less arrest rates with domestic incidents. Crime types like “offense involving children” have a higher arrest rate for domestic incidents (22% vs 20%.) This is probably because domestic can mean different things under different crime types. A domestic incident involving children may be domestic violence by a parent or family member while a domestic theft may simply be that some outsider attempted stealing inside someone’s house. Therefore the interactions between this variable and crime types can be very complicated. I will include this variable hoping our model can capture some of its interactions with other variables.

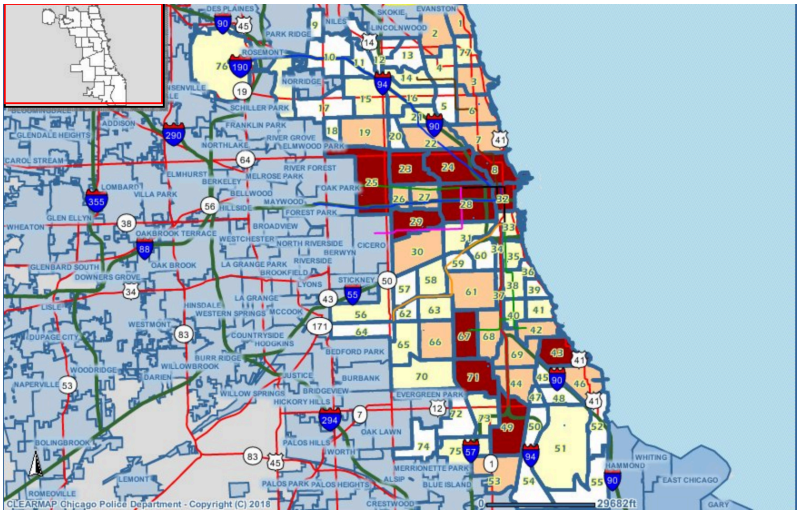
Regional Information

Regional information include community area, as well as their related demographics information: percent poverty, percent black, percent white, percent hispanic and percent asian.

The 8 most crime heavy community areas in our dataset includes community area 25, 8, 43, 23, 24, 67, 28 and 29, which can be found in the red regions from the crime heat map of last year (see graph below.) It is obvious that the crimes form clusters in upper Chicago and south-west Chicago. This variable is included because we want to see if arrest rates can be determined somewhat by crime rates. We also want to see that besides the obvious different of crime rates, if other traits of community areas contribute to higher arrest rates.

As for the demographics variables, I want to investigate if certain racial groups or income groups are more prone to be arrested.

Graph 3. Heat Map of Crime Heavy Community Areas For the Past 365 Days



d. DATA WRANGGLING

First I took away all crime types that are either too rare or have extremely high arrest rates. The remaining crime types mostly have arrest rates around 10 - 20% and are the leading types of crimes which are also the ones we care about since they cover the most prevalent violent crimes mentioned earlier.

Then I filtered out incidents that only happened in summer because I want to control for the time series nature of this data. Also, in summer crimes tend to spike (with the exception of robbery and auto theft.)

The next step is to use SMOTE method to treat imbalanced class problem. Before treatment the ratio of arrested to not-arrested was about 1:6. After treating the dataset with SMOTE, we over-sampled our success case and down-sampled our not-success case to achieve a class ratio of 1:1. After SMOTE we have 900k rows in our dataset[1] .

Later I join the enrichment data to our main dataset firstly by calculating the distance to the closest police station for each dataset. This process took a very long time given that I had limited computing power[2]. A whole iteration of calculations through the whole dataset took about 4.5 hours to complete. Then I joined the traffic information to my main dataset. Because the traffic information is by regions, not community areas, I manually parsed the region names to their corresponding community areas. One region usually covers 2 to 3 community areas. But there are

---

[1] Raw out-of-box model has an ROC-AUC score of ~ 0.75 before SMOTE using just down sampling.

[2] 3 hours to iterate through the entire 900k sample.

exceptions when the records' names were messed up. Also sometimes the regions don't overlap exactly through examining recorded boundaries in longitudes and latitudes, despite the official assurance from the City of Chicago that they should. The result is that some of the community areas cannot be matched to a region that has traffic information.

The last step was joining the community area demographics to the main dataset which was pretty straight forward. After some numeric re-codings and null removals my dataset is ready to be fed into a random forest model.

e. THE RANDOM FOREST MODEL

Random Forest Models

Initially introduced by Leo Breiman in his 2001 paper *Random Forests, r*andom Forest is a type of ensemble decision tree method.

A decision tree can be considered as a series of if-else statements that split the conditions into different subgroups and see if each subgroup at the maximum depth can be more or less homogeneous of one predictive class. This is done by firstly select a feature and find a cut point where leads to the greatest possible reduction in RSS. Next, the process will be repeated, aiming for the best predictor and best cut point in order future splits can minimize the RSS within each of the resulting regions even further. The generally used criteria for assessing cut point is the Gini index, which is a metric that measure impurity, defined by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Whereas pmk is the proportion of kth class present in the mth subgroup. Thus, for any split made, the Gini index is calculated for both of the new subgroups created and the split that makes the least average Gini index will be chosen as the best split. A small Gini index indicates the subgroup is more homogenous while a large Gini index means the classes are equally mixed in a subgroup.

Decision trees alone suffer from overfitting and high variance because they can be sensitive to specific data on which they are trained and the trees can be quite different if they were trained on different training-sets. Thus, ensemble methods like bagging was created. Such methods average multiple trees that were trained on different subsets of the data (bootstrapping). This way we are not so concerned about a single tree overfitting the data as that tree will be outvoted by other trees.

One one hand, Algorithms like decision trees and bagging utilize greedy algorithms such that they choose one variable from all available variables to split on in order to minimize errors. This will result in decision trees to have structural similarities because they went down the same path choosing from the same features during tree growing. In such cases the prediction from trees will be very similar and averaging with a majority vote will lose its advantage.

On the other hand, Random Forests improves bagging methods further to solve the correlated trees problem. Random Forest chooses a random subset of parameters for the tree to grow during each split instead of all variables(hence the name, RANDOM

forest.) This way the trees are much less likely to be correlated with each other and combining their predictions would be more useful than averaging inherently similar sub-models like bagging.

Benefits of a Random Forest Model For This Thesis

First of all, a random forest model is extremely suitable for my work given that most of my variables are categorical. Random forest supports both categorical and continuous values in an elegant way, comparing to models like Logistic Regression where you are eventually creating numerous dummy variables that acts like continuous variables.

Second of all, with this many categorical values, it will be hard to account for the numerous interactions with Logistic Regressions while tree based methods are analogous to a huge pile of interactions themselves. Thus there is a good chance that a Random Forest model will capture some of the complex interactions like the street vs community example (Weisburd and Lorraine, 1993) presented in earlier sections while a Logistic Regression won't. Random Forest models capture the local features of data like all tree based models but aren't prone to overfitting which makes them a good choice for this thesis.

Third of all, it runs much faster than other "black box" algorithms such as SVM. With the computation power I can get access to, running SVM and neural nets on my

data would cause my machine to freeze. Random Forest models takes sometime to run, but they always finish within an acceptable period of time[3].

Lastly, because I don't have much computation power, Random Forest alleviated the cross-validation problem and enables me to use a little number of folds. This is because Random Forest itself randomly samples the entire dataset for each iteration (bootstrapping). This behavior is comparable to performing cross validations already.

III. RESULTS/INSIGHTS

a. MODEL SELECTION

Logistic Regression Vs Random Forest

I first assessed some performance differences between out-of-the-box Random Forest and Logistic Regression, using 3-fold cross-validation. The number was chosen based on the computation power I have. Also Random Forest models are already not prone to overfitting so it is not necessary to choose a large fold value for cross-validation

The out-of-box Logistic Regression model yielded an average 0.64 ROC-AUC score which is not good at all. Although it can be argued that with more tuning we can get better performances out of a Logistic Regression model. But in this case, the

[3] 10 ~ 30 mins for one model, depend on the parameters.

different levels of interactions are the most important, and they cannot be fully addressed by a regression model. Probably by purposefully adding certain interaction terms we can improve the performance. However, by then the coefficients can also be too messy to interpret, giving us no advantages to use interpretable models like a Logistic Regression.

On the other hand, our random forest model produced an average ROC-AUC score of 0.9 out of the box[4], which is an astoundingly good score comparing to the Logistic Regression model.

Other models were either not considered due to simplicity (KNN, Naive Bayes,) or gave up due to computational difficulties (SVM.)

b. PARAMETER TUNING/OTHER ACCURACY BOOSTINGS

To improve my model I firstly used cross-validation to tune the model over a parameter grid. The original out-of-the-box model used 400 trees. In my parameter tuning I aimed to test: a) the optimal number of trees and b) the optimal number of parameters used at each split.

The reason for testing the optimal number of trees is because usually increase the number of trees in the forest result in higher prediction accuracy and less variance of the overall model. The reason is that with more trees, the model can capture more details in the dataset to make better predictions. The original *n_estimators* = 400 (the parameter that controls number of trees in scikit-learn) from the out-of-box model was

[4] I set the n_estimators to 400 for computation reason which is not exactly out-of-the-box

chosen such that the random forest model could be finish on-time within a reasonable amount of time. I want to see, with a more invested tuning, what is the number of trees from numbers below 400 to numbers way above 400 that can give me the best performance, given the usual display of diminish of returns when n_estimator gets larger. Thus I tested n_estimators in the set of 200, 400, 600, 800 and 1000. Any number larger than 1000 is too much of a computational burden that I had to avoid running.

As for max_features (the parameter that controls the number of variables considered while making each split), considering more features should increase the chance of finding a better split. Usually, the recommended default value for random forests in a classification problem is the square root of the total number of predictors. Since we have 12 variables and very complicated interactions, using $\sqrt{12}$ which is less than 4 parameters can be too small a number of parameters that can be taken into consideration while making each split, especially when our computation power limits the number of trees we can grow. Thus I wanted to see if a slightly higher max_features can improve the performance. Without turning the model into a bagging model, I chose to test the following values: default, 4, and 6.

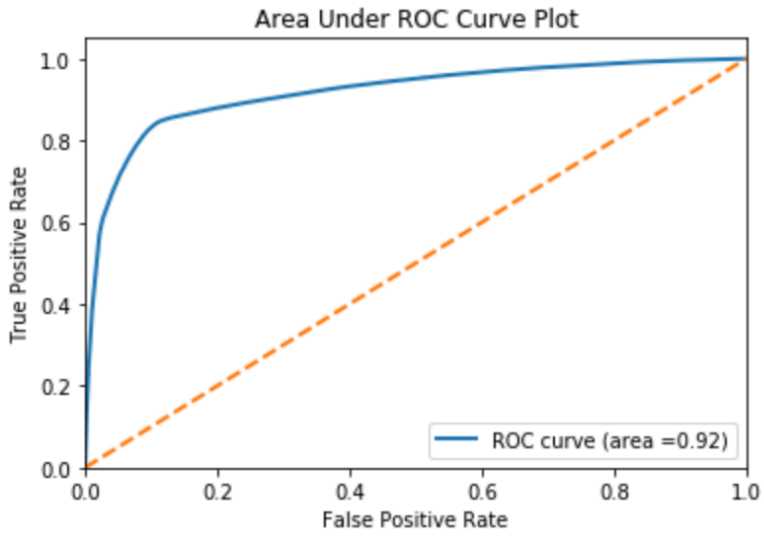
With a 3-fold cross validation, the total model that needed to be run was 30. The entire grid search took me over 8 hours to finish[5]. I tried running a more complicated grid on Amazon Sagemaker but it kept breaking regardless of my settings.

After done running the Grid Search, the best parameters appears to be the maximum value from both parameter categories, which are n_estimators = 1000 and

[5] 3.6 ghz dual core intel i5, 16G ram.

max_features = 6. I obtained an ROC-AUC score of 0.92, which is not a big improvement from 0.9 (See graph below).

Graph 4. ROC-AUC Curve For Final Model After Grid Search

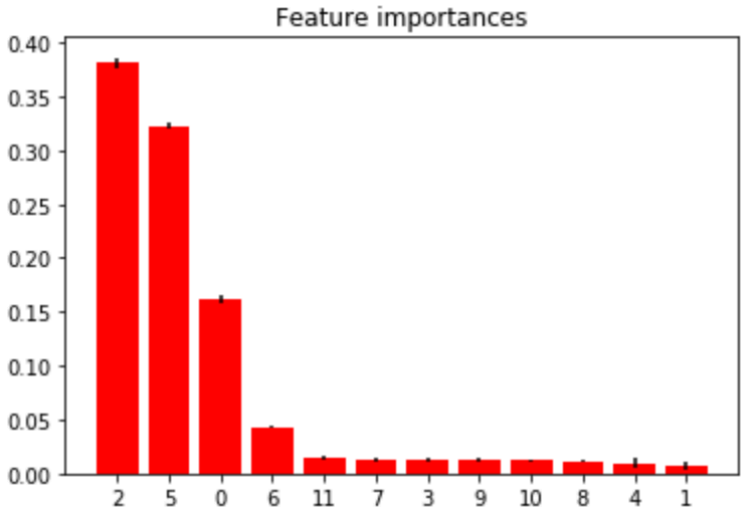


c. VARIABLE IMPORTANCE

A variable importance plot shows that our main predictors are the most important variables in the model (See Graph below.) Feature 2 is min_pd and feature 5 is avg_speed. Following the two is feature 0 which is Primary.Type in our dataset. From the graph we can see that the feature importance of other features are incomparable with our main variables. Especially for features that form the tail of this plot, they are almost negligible. The feature importance in scikit-learn calculates the Gini importance, or mean decreased impurity. Mean decreased impurity is the average of a variable's

total decrease in the node's Gini, or impurity, weighted by each individual decision tree generated.

Graph 5. Feature Importance Plot



Thus, a higher value in Mean Decreased Gini means that the variable is more important such that without that variable, the trees will have a harder time finding better splits. Note that this does not reflect a causal relationship: our main features are important, but not necessarily important in a way that they have cause arrests directly or that they determine the chance of getting arrests. Instead, the importance plot says that our main features provide purer split points overall, which can mean many things, like our main features have a lot of interactions with other features.

It is safe to say that besides min_pd, avg_speed, and Primary.Type, they rest are not essential in building this random forest model, which include all regional variables

like community areas as well as their demographics. This could mean that they are quite irrelevant to classifying arrests.

d. MODEL IMPLICATIONS USING SIMULATED DATA

Given that we achieved a very accurate model, we can try extract more meanings from our model by using a group of simulated data to find out what our model thinks about the arrested situations. We generated 577,805 data points by producing all possible combinations of values in our dataset and fed this simulated data to our model[6]. The benefit of such simulated data is no matter how we split and compare the categories the rest will always be controlled.

Firstly, from simulated data we can see that the average and median distance to the nearest police station is closer for the arrest cases than the non-arrest cases. The average and median min_pd for non-arrest cases are 9.53 and 3.69 respectively while those for the arrest cases are 9.34 and 3.53 respectively.

The average speed of surrounding areas for arrest cases are also faster than the non-arrest cases both in terms of average and median. The average and median for the non-arrest cases is 22.468 and 22.938 respectively while those for the arrest cases are 22.247 and 22.88 respectively.

If we split the two continuous features by their medium and look into the intersections between average speed, distance to nearest police department and arrest rates, we can find that most of the times for each individual crime type, the

---

[6] Except regional variables. Same community areas are always associated with the same demographic variables

highest arrest rate happens when both average speed and distance are at their lower 50% and the opposite when they are both at their upper 50%. However, if we look at the relationship between one of the main features and arrest, grouped by the other main features, we find such a pattern only exist in the upper 50%. For example, in cases where distance to the closest police department is larger than the 50th percentile, arrests have a faster traffic comparing to non-arrests. In the group where distance to the closest police department is smaller than the 50th percentile, this different is not so large and arrests even seem to have a lower traffic comparing to non-arrests. Such relationship is also true when we look at the relationship between distance and arrest, grouped by traffic.

On one hand, things get more interesting when we look at Community Areas X Primary Type interactions. For each primary type, the arrest rate differs dramatically for each community area. For example, Battery has a 60% arrest rate in some community areas but only 31% in some other community areas. Burglary has an arrest rate from 3% to 30%. There seem to have no obvious pattern as to which community areas stand out and which don't. High arrest rate areas and low arrest rate areas seem to be inconsistent for different crime types.

On the other hand, no obvious trends can be spotted with ethnicity or poverty information. Sometimes a race arrest more for certain types of crimes sometimes they are arrested less. Sometimes the arrested group are poorer for some crime types and sometimes it is the opposite. This is in-line with the variable importance plot earlier which is that the demographic variables do not contribute to cleaner cuts, and their function is probably not generalizable.

IV. FINAL REMARKS

a. SUMMARY

In short, our model reached a very high ROC-AUC performance score of 0.92 after tuning parameters, showing that one can predict arrests with high accuracy. Our main predictors, min_pd and avg_speed contributed greatly to tree growing, added the overall node purity to the structure, and appear to the most important variables in our model. Combining this result and the simulated data result we can confidently say that generally the easier access to police force, the higher the chance of making arrest although things can get more complicated with interactions between all variables. What is suspected is that min_pd and avg_speed serve as the premises to identifying arrests that takes into consider other variables like crime type, thus it is possible success trees have them at a shallower depth of decision trees.

The fact that variables like crime types and location seem to have low variable importance and no distinguishable pattern is telling us that given our results, making arrest is not as trivial as “arrests are more likely to be made in locations and demographic segments where crimes happen more.” These variables can be crucial in deciding the accuracy of predictions at deeper nodes. The simulated data analysis showed that there are many interactions that cannot be explained by just looking at the variables, which is what we expected for a Random Forest Model.

Some actionable insights based on our results can be that the law enforcement can experiment with police forces arrangement and adjust the police force accordingly based on traffic conditions.

This line of research can probably be best utilized by law enforcements to optimize police resources when multiple incidents happen. For example, if 3 crimes have been reported in, which one has a higher chance to make arrest and which one don't? If the law enforcement can figure out these prediction they can maximize the chances of making arrest and waste less time and effort.

b.LIMITATIONS/FUTURE STEPS

As an initial attempt to predict arrests, this thesis has many limitations. They include but are not limited to the following:

a) Limited computation power. The best parameters in our thesis is only the best parameters that I can afford. There are still other parameters that I have not explored with yet. With more computation power, models with better prediction accuracy could be made.
b) Simplified complex assumption. There are many times where I simplified assumptions like "distance and speed determines access to police force," and there are also times where I just ignored reality like I didn't investigate if on-site arrests are actually accurate in capturing responsible personals.
c) Low granularity data on traffic. Our traffic data is less than ideal: it is averaged traffic information within a small time span of regions as large as 2 - 3

community areas. If we could have more detailed traffic data in a long time span, we could probably get more accuracy with our model and do other analysis like investigating extreme traffic situations.

d) Interactions too complex to extract actionable insights. Complicate interactions were expected and welcomed while building the Random Forest model in that they give better prediction accuracies. However, they are still not helpful while trying to create actionable insights. After all, one of the goals of this thesis is to try to let machine learning algorithms produce more real world impacts in our particular subject matter.

e) Did not account for the time series nature of the data. This data we have on crimes is time series data. Because I did not take the class on time series, I just controlled for summer season and used a model that isn't susceptible to time series problems. However, it is necessary to take into consider all seasons and do more detailed investigations such that we can really see a model that is interpretable.

V. REFERENCES

Ahishakiye, E., Taremwa, D., Opiyo, E. and Niyonzima, I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. International Journal of Computer and Information Technology, 6(3).

Arulanandam, R., Tony Roy Savarimuthu, B. and A. Purvis, M. (2014). Extracting Crime Information from Online Newspaper Articles. AWC 2014.

Best, J. and Braithwaite, J. (1990). Crime, Shame and Reintegration. Social Forces, 69(1), p.318.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F. and Pentland, A. (2014). Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. ACM International Conference on Multimodal Interaction.

Chen, X., Cho, Y. and Young Jang, S. (2015). Crime Prediction Using Twitter Sentiment and Weather.

City-data.com. (2018). Crime in Chicago, Illinois (IL): murders, rapes, robberies, assaults, burglaries, thefts, auto thefts, arson, law enforcement employees, police officers, crime map. [online] Available at: http://www.city-data.com/crime/crime-Chicago-Illinois.html [Accessed 23 Dec. 2018].

Clarke, R. and Felson, M. (2004). Routine activity and rational choice. New Brunswick, N.J.: Transaction.

Cozens, P., Saville, G. and Hillier, D. (2005). Crime prevention through environmental design (CPTED): a review and modern bibliography. Property Management, 23(5), pp.328-356.

CURRIE, E. (1997). Market, Crime and Community. Theoretical Criminology, 1(2), pp. 147-172.

Ehrlich, I. (1975). On the Relation between Education and Crime. Education, Income, and Human Behavior.

Fajnzylber, P., Lederman, D. and Loayza, N. (2002). What causes violent crime?. European Economic Review, 46(7), pp.1323-1357.

Ford, M. (2017). What's Actually Causing Chicago’s Homicide Spike?. [online] The Atlantic. Available at: https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/ [Accessed 23 Dec. 2018].

Ford, S. (2014). Benefits of Crime Prevention. [online] The Australian Crime Prevention Council. Available at: http://www.acpc.org.au/images/documents/Benefits-of-Crime-Prevention.pdf [Accessed 22 Dec. 2018].

Freeman, R. (1999). The Economics of Crime. Handbook of Labor Economics, 3.

Gaviria Trujillo, A., Guerrero, R. and Londoño de la Cuesta, J. (2000). Asalto al desarrollo. Washington, D.C: Banco Interamericano de Desarrollo.

Gerber, M. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, pp.115-125.

Gorner, J. (2014). Chicago violence continues to outpace NYC, LA. [online] Chicagotribune.com. Available at: https://www.chicagotribune.com/news/ct-chicago-police-shootings-homicide-met-20140701-story.html [Accessed 23 Dec. 2018].

Gorner, J., Nickeas, P. and Malagon, E. (2016). August most violent month in Chicago in nearly 20 years. [online] Chicagotribune.com. Available at: https://www.chicagotribune.com/news/local/breaking/ct-august-most-violent-shootings-chicago-20160829-story.html [Accessed 23 Dec. 2018].

Home.chicagopolice.org. (2011). Chicago Murder Analysis. [online] Available at: http://home.chicagopolice.org/wp-content/uploads/2014/12/2011-Murder-Report.pdf [Accessed 23 Dec. 2018].

J. Shapiro, R. and A. Hassett, K. (2012). The Economic Benefits of Reducing Violent Crime - Center for American Progress. [online] Center for American Progress. Available at: https://www.americanprogress.org/issues/economy/reports/2012/06/19/11755/the-economic-benefits-of-reducing-violent-crime/ [Accessed 22 Dec. 2018].

L. Perry, W., McInnis, B., C. Price, C., C. Smith, S. and S. Hollywood, J. (2010). Predictive POLICING The Role of Crime Forecasting in Law Enforcement Operations. RAND.

LEVITT, S. (1998). WHY DO INCREASED ARREST RATES APPEAR TO REDUCE CRIME: DETERRENCE, INCAPACITATION, OR MEASUREMENT ERROR?. Economic Inquiry, 36(3), pp.353-372.

Lisa Cole, N. (2018). Why Does Crime Spike in Summer?. [online] ThoughtCo. Available at: https://www.thoughtco.com/why-does-crime-spike-in-summer-3026089 [Accessed 24 Dec. 2018].

Mehlum, H., Moene, K. and Torvik, R. (2005). Crime induced poverty traps. Journal of Development Economics, 77(2), pp.325-340.

News, A. (2012). Chicago Gang Violence: By The Numbers. [online] ABC News. Available at: https://abcnews.go.com/Nightline/fullpage/chicago-gang-violence-numbers-17509042 [Accessed 23 Dec. 2018].

PATTERSON, E. (1991). POVERTY, INCOME INEQUALITY, AND COMMUNITY CRIME RATES. Criminology, 29(4), pp.755-776.

Ratcliffe, J. and Makkai, T. (2004). Diffusion of Benefits: Evaluating a Policing Operation. Australian Institute of Criminology.

SANBURN, J. (2016). Chicago Is Responsible for Almost Half of the Increase in U.S. Homicides. [online] Time. Available at: http://time.com/4497814/chicago-murder-rate-u-s-crime/ [Accessed 23 Dec. 2018].

Schneider, R. and Kitchen, T. (2002). Planning for crime prevention. London: Routledge.
Wang, X., S. Gerber, M. and E. Brown, D. (2012). Automatic Crime Prediction using Events Extracted from Twitter Posts.

Weinberg, S. (1954). Theories of Criminality and Problems of Prediction. The Journal of Criminal Law, Criminology, and Police Science, 45(4), p.412.

Weisburd, D. and Lorraine, G. (1993). Defining the street-level drug market. Drugs and crime : Evaluating public policy initiatives.

VI. APPENDIX

Appendix 1. Sample data from City of Chicago Crimes - 2001 to present

	ID	Case.Number	Date	Block	IUCR	Primary.Type	Description
1	10000092	HY189866	03/18/2015 07:44:00 PM	047XX W OHIO ST	041A	BATTERY	AGGRAVATED: HANDGUN
2	10000094	HY190059	03/18/2015 11:00:00 PM	066XX S MARSHFIELD AVE	4625	OTHER OFFENSE	PAROLE VIOLATION
3	10000095	HY190052	03/18/2015 10:45:00 PM	044XX S LAKE PARK AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
4	10000096	HY190054	03/18/2015 10:30:00 PM	051XX S MICHIGAN AVE	0460	BATTERY	SIMPLE
5	10000097	HY189976	03/18/2015 09:00:00 PM	047XX W ADAMS ST	031A	ROBBERY	ARMED: HANDGUN
6	10000098	HY190032	03/18/2015 10:00:00 PM	049XX S DREXEL BLVD	0460	BATTERY	SIMPLE
7	10000099	HY190047	03/18/2015 11:00:00 PM	070XX S MORGAN ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE
8	10000100	HY189988	03/18/2015 09:35:00 PM	042XX S PRAIRIE AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
9	10000101	HY190020	03/18/2015 10:09:00 PM	036XX S WOLCOTT AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS
10	10000104	HY189964	03/18/2015 09:25:00 PM	097XX S PRAIRIE AVE	0460	BATTERY	SIMPLE
11	10000105	HY189984	03/18/2015 09:30:00 PM	130XX S DR MARTIN LUTHER KING JR DR	1320	CRIMINAL DAMAGE	TO VEHICLE
12	10000108	HY189719	03/15/2015 04:10:00 PM	078XX S VINCENNES AVE	2825	OTHER OFFENSE	HARASSMENT BY TELEPHONE
13	10000109	HY189966	03/18/2015 09:14:00 PM	086XX S EXCHANGE AVE	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN
14	10000110	HY190056	03/18/2015 10:50:00 PM	014XX S ASHLAND AVE	0460	BATTERY	SIMPLE
15	10000111	HY190019	03/18/2015 10:31:00 PM	051XX W CHICAGO AVE	0860	THEFT	RETAIL THEFT
16	10000112	HY189725	03/18/2015 12:55:00 PM	077XX S KINGSTON AVE	0610	BURGLARY	FORCIBLE ENTRY
17	10000114	HY190071	03/18/2015 08:00:00 PM	024XX W NORTH AVE	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
18	10000115	HY190036	03/18/2015 09:00:00 PM	069XX S LOOMIS BLVD	0890	THEFT	FROM BUILDING
19	10000116	HY190063	03/18/2015 10:56:00 PM	105XX S LAFAYETTE AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT
20	10000117	HY190068	03/18/2015 10:45:00 PM	087XX S KIMBARK AVE	0890	THEFT	FROM BUILDING
21	10000118	HY190031	03/18/2015 10:00:00 PM	075XX S STONY ISLAND AVE	0860	THEFT	RETAIL THEFT
22	10581023	HZ329792	09/01/2014 08:00:00 AM	0000X E LAKE ST	1140	DECEPTIVE PRACTICE	EMBEZZLEMENT
23	10000119	HY190072	03/18/2015 11:55:00 PM	054XX N KENMORE AVE	0320	ROBBERY	STRONGARM - NO WEAPON
24	10000120	HY190073	03/18/2015 11:40:00 PM	078XX S CREGIER AVE	0430	BATTERY	AGGRAVATED: OTHER DANG WEAPON
25	10000123	HY189969	03/18/2015 09:44:00 PM	0000X N CENTRAL AVE	141A	WEAPONS VIOLATION	UNLAWFUL USE HANDGUN
26	10000124	HY190060	03/18/2015 11:30:00 PM	024XX S BELL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
27	10000126	HY190035	03/18/2015 10:45:00 PM	003XX E 59TH ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE
28	10000127	HY190027	03/18/2015 10:33:00 PM	091XX S RACINE AVE	0560	ASSAULT	SIMPLE
29	10000129	HY190096	03/19/2015 01:20:00 AM	078XX S EMERALD AVE	0460	BATTERY	SIMPLE

Location.Description	Arrest	Domestic	Beat	District	Ward	Community.Area	FBI.Code	X.Coordinate	Y.Coordinate	Year
STREET	false	false	1111	11	28	25	04B	1144606	1903566	2015
STREET	true	false	725	7	15	67	26	1166468	1860715	2015
APARTMENT	false	true	222	2	4	39	08B	1185075	1875622	2015
APARTMENT	false	false	225	2	3	40	08B	1178033	1870804	2015
SIDEWALK	false	false	1113	11	28	25	03	1144920	1898709	2015
APARTMENT	false	false	223	2	4	39	08B	1183018	1872537	2015
APARTMENT	false	true	733	7	17	68	08B	1170859	1858210	2015
APARTMENT	false	true	213	2	3	38	08B	1178746	1876914	2015
STREET	true	false	912	9	11	59	18	1164279	1880656	2015
RESIDENCE PORCH/HALLWAY	false	false	511	5	6	49	08B	1179637	1840444	2015
PARKING LOT/GARAGE(NON.RESID.)	false	false	533	5	9	54	14	1180907	1818839	2015
CTA GARAGE / OTHER PROPERTY	false	true	623	6	17	69	26	1175130	1853144	2015
DRIVEWAY - RESIDENTIAL	true	false	423	4	10	46	15	1197309	1848290	2015
SIDEWALK	false	false	1233	12	2	28	08B	1165950	1893388	2015
GAS STATION	true	false	1531	15	37	25	06	1141741	1904839	2015
APARTMENT	false	false	421	4	7	43	05	1194535	1854110	2015
OTHER	false	false	1423	14	1	24	07	1159959	1910569	2015
GROCERY FOOD STORE	false	false	734	7	17	67	06	1168192	1858832	2015
ALLEY	true	false	512	5	34	49	24	1177790	1835106	2015
BAR OR TAVERN	false	false	412	4	8	48	06	1186312	1847473	2015
GROCERY FOOD STORE	true	false	411	4	5	43	06	1188123	1855151	2015
OTHER	true	false	111	1	42	32	12	*NA*	*NA*	2014
OTHER	false	false	2023	20	48	77	03	1168270	1936260	2015
APARTMENT	false	false	414	4	8	43	04B	1189564	1853354	2015
VEHICLE NON-COMMERCIAL	true	false	1513	15	29	25	15	1139051	1899715	2015
APARTMENT	true	true	1034	10	25	31	08B	1161687	1887883	2015
APARTMENT	true	true	232	2	20	40	08B	1179208	1865959	2015
RESIDENCE	false	false	2222	22	21	73	08A	1169928	1844242	2015
APARTMENT	false	false	621	6	17	71	08B	1172661	1852870	2015

Updated.On	Latitude	Longitude	Location
02/10/2018 03:50:01 PM	41.89140	-87.74438	(41.891398861, -87.744384567)
02/10/2018 03:50:01 PM	41.77337	-87.66532	(41.773371528, -87.665319468)
02/10/2018 03:50:01 PM	41.81386	-87.59664	(41.81386068, -87.596642837)
02/10/2018 03:50:01 PM	41.80080	-87.62262	(41.800802415, -87.622619343)
02/10/2018 03:50:01 PM	41.87806	-87.74335	(41.878064761, -87.743354013)
02/10/2018 03:50:01 PM	41.80544	-87.60428	(41.805443345, -87.604283976)
02/10/2018 03:50:01 PM	41.76640	-87.64930	(41.766402779, -87.649296123)
02/10/2018 03:50:01 PM	41.81755	-87.61982	(41.817552577, -87.619818523)
02/10/2018 03:50:01 PM	41.82814	-87.67278	(41.828138428, -87.672782106)
02/10/2018 03:50:01 PM	41.71745	-87.61766	(41.71745472, -87.617663257)
02/10/2018 03:50:01 PM	41.65814	-87.61367	(41.658138493, -87.613672862)
02/10/2018 03:50:01 PM	41.75241	-87.63379	(41.752406801, -87.633792381)
02/10/2018 03:50:01 PM	41.73856	-87.55268	(41.738563465, -87.552678593)
02/10/2018 03:50:01 PM	41.86304	-87.66629	(41.86304084, -87.666288555)
02/10/2018 03:50:01 PM	41.89495	-87.75487	(41.894945606, -87.754874977)
02/10/2018 03:50:01 PM	41.75460	-87.56265	(41.754602618, -87.562650741)
02/10/2018 03:50:01 PM	41.91031	-87.68781	(41.910312648, -87.687806494)
02/10/2018 03:50:01 PM	41.76817	-87.65905	(41.768167414, -87.659053795)
02/10/2018 03:50:01 PM	41.70285	-87.62459	(41.70284845, -87.624588931)
02/10/2018 03:50:01 PM	41.73659	-87.59299	(41.736588206, -87.59299436)
02/10/2018 03:50:01 PM	41.75761	-87.58612	(41.757614433, -87.586115266)
03/01/2018 03:52:35 PM	*NA*	*NA*	
02/10/2018 03:50:01 PM	41.98063	-87.65653	(41.980634157, -87.656529996)
02/10/2018 03:50:01 PM	41.75265	-87.58089	(41.752648838, -87.580891872)
02/10/2018 03:50:01 PM	41.88093	-87.76488	(41.880934043, -87.764879438)
02/10/2018 03:50:01 PM	41.84802	-87.68209	(41.848024395, -87.682090877)
02/10/2018 03:50:01 PM	41.78748	-87.61846	(41.787480544, -87.618458018)
02/10/2018 03:50:01 PM	41.72809	-87.65311	(41.72809296, -87.653113317)
02/10/2018 03:50:01 PM	41.75171	-87.64285	(41.751709641, -87.642848244)

Appendix 2. Data Dictionary for Crimes Data

This dataset contains 22 columns:

ID - Unique identifier for the record.

Case Number - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.

Date - Date when the incident occurred. This is sometimes the best estimate.

Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.

IUCR - The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.

Primary Type - The primary description of the IUCR code.

Description - The secondary description of the IUCR code, a subcategory of the primary description.

Location Description - Description of the location where the incident occurred. *Arrest* - Indicates whether an arrest was made.

Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

Beat - Indicates the beat 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. See the beats at https://data.cityofchicago.org/d/aerh-rz74.

District - Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.

Ward - The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.

Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.

FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

Y Coordinate - They coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

Year - Year the incident occurred.

Updated On - Date and time the record was last updated.

Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

Location - The location where the incident occurred in a format that allows for the creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.