Chicago Crime Recommendation System

(From an Alternate Perspective)

Jarrod Johnson

University of Houston

Data Science II (COSC4337)

March 5, 2021

Source

Data Provided by: Chicago Police Department

Contact Dataset Owner

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Licensing and Attribution

License See Terms of Use

Source Link https://portal.chicagopolice.org/portal/page/portal/ClearPath

Overview

Date Created

September 30, 2011

Data Provided by

Chicago Police Department

Dataset Owner

Cocadmin

Rows

7.28M

Columns

22

Each row is a

Reported Crime

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. Should you have questions about this dataset, you may contact the Research & Development Division of the Chicago Police Department at PSITAdministration@ChicagoPolice.org. Disclaimer: These crimes may be based upon preliminary information supplied to the Police Department by the reporting parties that have not been verified. The preliminary crime classifications may be changed at a later date based upon additional investigation and there is always the possibility of mechanical or human error. Therefore, the Chicago Police Department does not guarantee (either expressed or implied) the accuracy, completeness, timeliness, or correct sequencing of the information and the information should not be used for comparison purposes over time. The Chicago Police Department will not be responsible for any error or omission, or for the use of, or the results obtained from the use of this information. All data visualizations on maps should be considered approximate and attempts to derive specific addresses are strictly prohibited. The Chicago Police Department is not responsible for the content of any off-site pages that are referenced by or that reference this web page other than an official City of Chicago or Chicago Police Department web page. The user specifically acknowledges that the Chicago Police Department is not responsible for any defamatory, offensive, misleading, or illegal conduct of other users, links, or third parties and that the risk of injury from the foregoing rests entirely with the user. The unauthorized use of the words "Chicago Police Department," "Chicago Police," or any colorable imitation of these words or the unauthorized use of the Chicago Police Department logo is unlawful. This web page does not, in any way, authorize such use. Data are updated daily.

To access a list of Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) codes, go to http://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

Column description *Data Types*

Column Name	***Description***	***Type***
ID	*Unique identifier for the record.*	*Number*
Case Number	*The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.*	*Plain Text*
Date	*Date when the incident occurred. this is sometimes a best estimate.*	*Date & Time*
Block	*The partially redacted address where the incident occurred, placing it on the same block as the actual address* .	*Plain Text*
IUCR	*The Illinois Uniform 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.*	*Plain Text*
Primary Type	*The primary description of the IUCR code.*	*Plain Text*
Description	*The secondary description of the IUCR code, a subcategory of the primary description.*	*Plain Text*
Location Description	*Description of the location where the incident occurred.*	*Plain Text*
Arrest	*Indicates whether an arrest was made.*	*Checkbox*
Domestic	*Indicates whether the incident was domestic related as defined by the Illinois Domestic Violence Act.*	*Checkbox*
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.*	*Plain Text*
District	*Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.*	*Plain Text*
Ward	*The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.*	*Number*
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.*	*Plain Text*
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.*	*Plain Text*
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.*	*Number*
Y Coordinate	*The y 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.*	*Number*
Year	*Year the incident occurred.*	*Number*
Updated On	*Date and time the record was last updated*	*Date & Time*
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.*	*Number*
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.*	*Number*
Location	*The location where the incident occurred in a format that allows for 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.*	*Location*

Novel idea

AI Alphabet Inc. is planning to become the world's first industrial supervillain. After recently accessing the power of data science and statistical predictions, they plan to use the Chicago crimes dataset to realize their vision of supervillainy. Within this dataset they've identified a very special feature column; One that can hopefully empower their evil AI hench-bot (powered by Jupyter Notebook) to reveal the best course of action for them to commit crimes in Chicago and avoid immediate capture. The feature column is titled "Arrest" and is conveniently binary. This column reflects true or false for whether an arrest has been made sometime (1 or more week) after a crime occurred.

Data/business understanding

This information will hopefully allow them to predict different combinations of which, where, and possibly when different crimes happen that do not have arrests related to them. They would like to arrange the data so they can have the ability to:
Choose a crime type given a place (and time), or
Choose a place (and time) given a crime type,
that will give them the best opportunity to escape arrest,
(at least immediately (Arrest = 0/F)).

Feature Engineering *Explanation of the steps followed to clean and reduce the dimensionality of the data*

Under unmentionable duress I've been forced to assist in working with this dataset to achieve these goals. Since the company is primarily interested in features that are related to crime type and place, I will work towards reducing the dataset around those types of features. Notebook rows 1-10 involve getting a set size workable for my machine and suitable for model testing. After noticing variations in feature content between different sets, I decided to use the largest set since it has consistent feature shape and data and covers the most time. Even though this set is too large for my machine, it should be a manageable file size after feature reduction.

Since my machine can't do much correlation work with a file this size, I decided to use my eyes and general judgement to cull features that will bog my machine down and that should be highly correlated with the features I intend to keep, or generally

useless for the companies' goals. Below is al list of features with their associated unique entry count. This list alone took 20 mins to produce!

Exploratory Data Analysis *Data visualization*

0	ID	7283796
1	Case Number	7283335
2	Date	2966372
3	Block	61445
4	IUCR	402
5	Primary Type	36
6	Description	534
7	Location Description	214
8	Arrest	2
9	Domestic	2
10	Beat	304
11	District	24
12	Ward	50
13	Community Area	78
14	FBI Code	26
15	X Coordinate	78792
16	Y Coordinate	130005
17	Year	21
18	Updated On	4001
19	Latitude	875197
20	Longitude	874633
21	Location	876411

Data/business understanding

I chose from the above features as follows:

Primary Type - Is the feature that will allow the company to select a specific type of crime.

Location Description, Beat, District, Ward, Community Area - Are the features that will allow the company to select where to commit the crime.

Arrest - Is the target variable the company is interested in with regards to the type and place of crimes committed.

Domestic - Is a variable that will allow me to remove many rows of data that are not important to the company. Domestic indicates a crime that happened between two or more people that know

each other well, and often time takes place inside a residence while the victim and offender are living together. This situation is not valid for the company and therefore all Domestic crimes must not be included in the data.

Year - Is a variable that will allow me to sort the data and partition it into sections that will allow me to produce train, validate, and test sets for modeling. Setting a period of a year will allow me to use future data as a baseline ground truth for the predictions in a previous year.

Updated On – Is a feature that may have purpose later with respect to determining which crimes are ignored over time more than others (or something similar).

Feature Engineering *Explanation of the steps followed to clean and reduce the dimensionality of the data*

With the dataset now sorted and manageable for my machine, I'd like to look at my group of "where" features and make sure they are in fact well correlated so I may choose just one. In order to do that I needed to transform "Location Description" into a nominal numeric form. I decided to use the Label Encoder function since it normalizes the transformed data. I accomplished this by truncating the data within "Location Description" leaving only the first word of data to describe the feature column. This will better enable the nominal transformation.

Exploratory Data Analysis *Data visualization*

	Location Description	Beat	District	Ward	Community Area
Location Description	1.000000	-0.003799	-0.004223	-0.008905	-0.004963
Beat	-0.003799	1.000000	0.999998	0.999724	-0.998922
District	-0.004223	0.999998	1.000000	0.999765	-0.999005
Ward	-0.008905	0.999724	0.999765	1.000000	-0.999516
Community Area	-0.004963	-0.998922	-0.999005	-0.999516	1.000000

Observing the 10 pairs correlations indicates a strong relation between all the attributes except for "Location Description" having relatively no correlations (although curiously having double the relative correlation to "Ward" as to the others). I expected "Location Description"

not to have strong correlations due to the Label Encoder function's normalizing the conversions. I probably just should have excluded it altogether.

"Community Area" produced a surprise by having an inverse relation to the other features (suspected trivial cause by *inverse numbering pattern).

"Beat", "District", and "Ward" outcomes were as predicted because they are likely subsets of themselves in some manner.

To Do:

"Primary Type" needs some special attention. Since it is categorical, I'll need to convert it to numeric. I plan to group the crime types based on their real-life severity (felony, misdemeanor, etc.), then list them numerically in an ordinal fashion.

More:

Exploratory Data Analysis *Data visualization*

Feature Engineering *Explanation of the steps followed to clean and reduce the dimensionality of the data*

Look into "Beat" (strange spread of numbers); look for outliers; get the baseline correlation with the other "places", then alter "Beat" (maybe normalize) and see if the correlation changes.

Add notes to Notebook directly for future review.