Analysis of Crimes in Chicago

Project Final Report

1ST DECEMBER'2017

UNIVERSITY OF ILLINOIS CHICAGO
ASHIMA HORRA [UIN;654087182]

Table of Contents

Analysis of Crimes in Chicago

1. Introduction

1.1. Objective

We study the crimes that have taken place in Chicago over the years and develop prediction models to help gauge the risk associated with the specific location in Chicago. The project is essentially divided into two sections: Prediction of Crime Index and the Prediction of Arrests in Chicago. Given the location and time information, the objective of the former is to predict the crime index associated with those inputs. The objective of the latter is to predict whether a crime at the select location and time, given the police district, will result in an arrest or not.

1.2. Motivation

Public safely is one of the most important concerns all around the world. Over the past few year, the crime rate in Chicago has become a topic of alarming concern and calls for dedicated measures. According to a 1960's science fiction author Philip K. Dick, "In fifty years or so our society will have the ability to predict certain crimes before they happen." While predicting exact crimes is still out of our scope, we do have the capabilities to predict an estimate of the risk associated with a crime in the next hour at any street in Chicago.

1.3. Future Scope

In this project, we predict the index of the crime (Index Crime v/s Non-Index Crime) based on time and space information, along with whether a crime associated with those inputs will result in an arrest. However, in the future, we can predict the exact crime: Robbery, theft, Assault, etc. with the exact probability of the crime happening.

2. Data Source

2.1. Raw Data

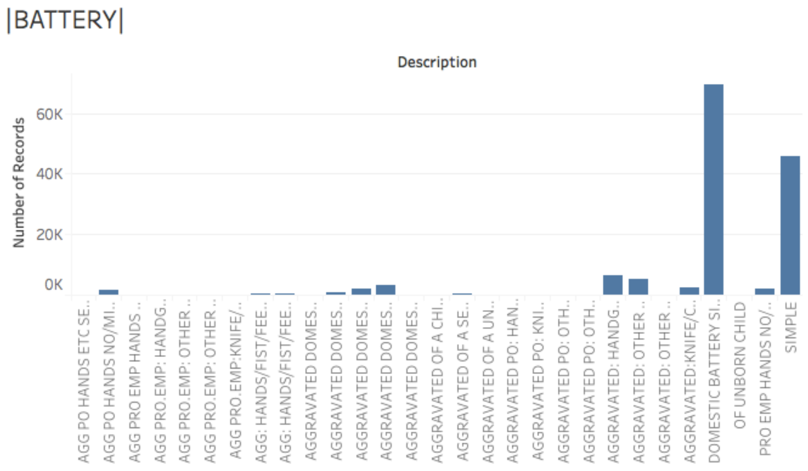
The data has been picked up from https://data.cityofchicago.org/. The original data set reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. This project only includes a part of the dataset (say from 2014 to present), to ensure that the data volume in manageable.

P.S: Since the volume of the 3 years data is huge, i.e. ~700,000 rows, data for only 2 years has been considered for analysis. This makes up ~343,000 rows. To further decrease the volume of the data a random sample of 50% of this data has been considered to build models ~171,000 rows

Name	Description	Type
ID	Unique Identifier for the record	Integer
Case Number	Chicago police department RD number	Character
Date	Date and time when the incident occurred	Character
Block	Block level address of the place where the incident occurred	character
IUCR	The Illinois uniform Crime Reporting code	Integer
Primary Type	The primary description of the IUCR code	character
Description	The secondary description of the IUCR code (subcategory of the primary description)	character
Location Description	Description of the location where the incident occurred	character
Arrest	Indicates whether an arrest was made for the crime	character
Domestic	Indicates whether the incident was domestic- as defined by the Illinois Domestic Violation Act	character
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area	integer
District	Indicates the police district where the incident occurred	character
Ward	The ward (City Council district) where the incident occurred.	integer
Community Area	Indicates the community area where the incident occurred	integer
FBI Code	Indicates the crime classification as outlined in FBI's National Incident-Based Reporting System	integer
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD	integer
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD	integer
Year	Year the incident occurred.	integer
Updated On	Date and time the record was last updated.	character
Latitude	The latitude of the location where the incident occurred.	Double
Longitude	The longitude of the location where the incident occurred.	Double
Location	Latitude and longitude combined	character

2.2. Data Preparation

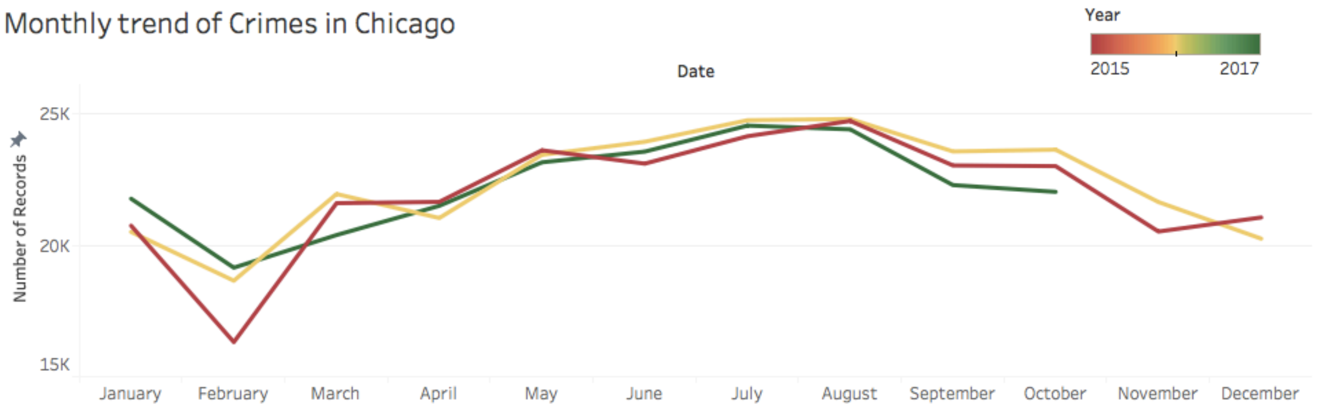
The data set consists of crimes with their Primary type and Description. The primary type gives the type of the crime and description gives a description about how sever the crime is. All crimes in this data set have been categorized into INDEX and NON-INDEX crimes based on the categorization of crime types given by the City of Chicago Municipal Code: MCC.

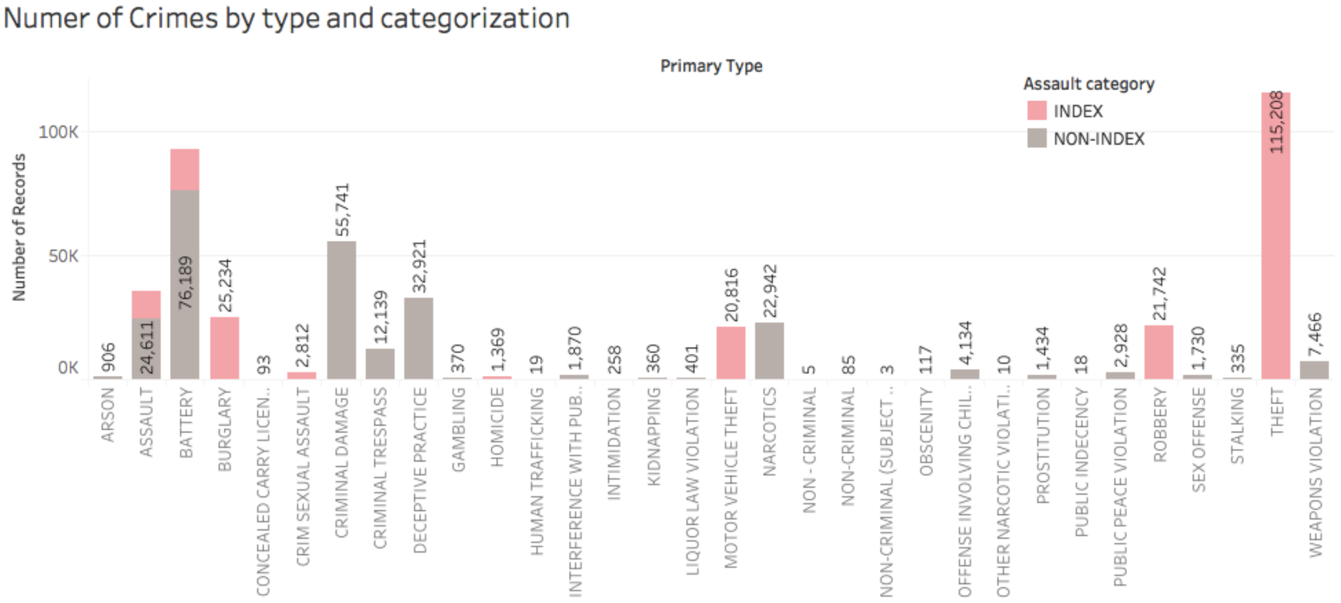


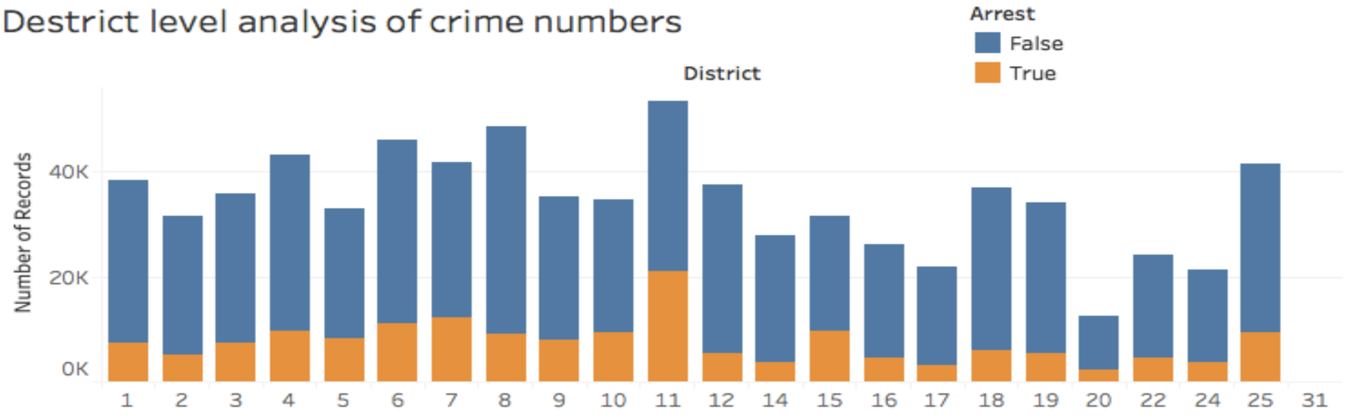
For example, in case of BATTERY, the aggravated battery crimes are considered to be INDEX and simple batteries are considered to be NON-INDEX crimes.

3. Exploratory Data Analysis

Data exploratory analysis was performed to check for any anomalies, missing values, or outliers in the data and to better understand the data and its elements. In the case of this data, 5% of the rows were found to have missing values. Since the percentage of missing values is low, these rows were omitted from the dataset.


Monthly trend of Crimes in Chicago


Numer of Crimes by type and categorization


Destrict level analysis of crime numbers

4. Statistical Models

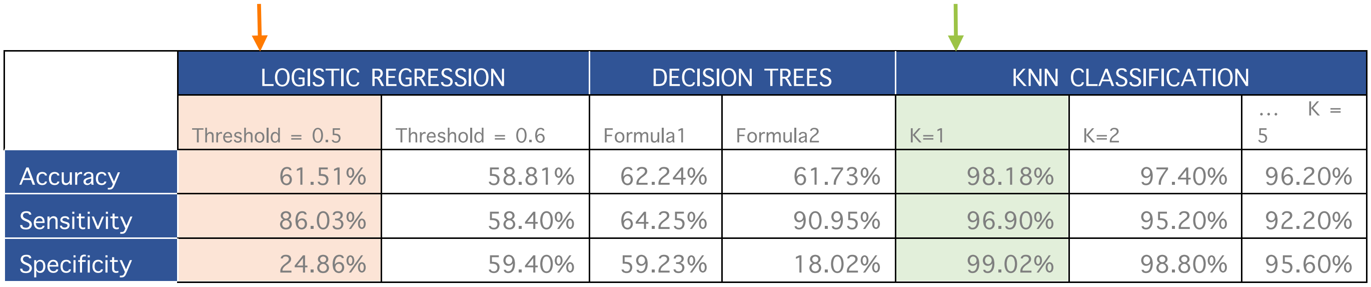
4.1. Prediction of Crime Index

Different models have been tried in order to get a good performing model. Since in our problem, we want higher correct positive and negative predictions, we wouldn't only be looking at the accuracy. The measures that have been used to measure the performance of the model are:

- Accuracy: Overall accuracy of the model. How many records have been predicted correctly
- Sensitivity: Number of INDEX crimes(positives) predicted correctly
- Specificity: Number of NON-INDEX crimes(negatives) predicted correctly

Benchmark Model

Final Model



	LOGISTIC REGRESSION		DECISION TREES		KNN CLASSIFICATION		
	Threshold = 0.5	Threshold = 0.6	Formula1	Formula2	K=1	K=2	… K = 5
Accuracy	61.51%	58.81%	62.24%	61.73%	98.18%	97.40%	96.20%
Sensitivity	86.03%	58.40%	64.25%	90.95%	96.90%	95.20%	92.20%
Specificity	24.86%	59.40%	59.23%	18.02%	99.02%	98.80%	95.60%

The performance of three types of models has been compared: Logistic Regression, Decision Tress and K-Nearest Neighbor. Above are the Accuracy, Sensitivity and Specificity of the three models. Logistic regression with threshold = 0.5 has been considered as the benchmark model.

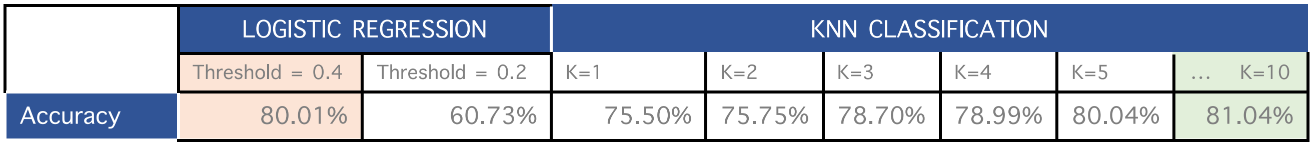
K-Nearest Neighbor with K=1 has been selected as the final model as it gives the best results. Also, since our model predictions are highly dependent on neighborhood, the KNN model seems to be the most suitable fit.

4.2. Prediction of Arrests

The performance of Logistic regression and KNN models has been compared. Only Accuracy has been considered as the measure of performance since in predicting arrests, accuracy is what is most important to us.

Benchmark Model

Final Model



	LOGISTIC REGRESSION		KNN CLASSIFICATION					
	Threshold = 0.4	Threshold = 0.2	K=1	K=2	K=3	K=4	K=5	… K=10
Accuracy	80.01%	60.73%	75.50%	75.75%	78.70%	78.99%	80.04%	81.04%

KNN Nearest neighbor with K=10 has been selected as the final model as it gives the best accuracy. While the logistic regression model with threshold = 0.04 has comparable accuracy, it obtains this accuracy by predicting "No Arrest" for all cases, thus not being a suitable model.

5. Inferences

5.1. Prediction of Crime

Intuitively, we can say that crime rates are highly dependent on neighborhood; Hence, prediction of crime index is highly dependent on neighborhood. On this ground, we can assess and draw interpretations about the models that we have tested.

5.1.1. Logistic Regression

Logistic regression works best when the decision boundary between different classes is linear. On our data, logistic regression performs very poorly. Hence, we can say that the decision boundary between index and non-index crimes is highly non-linear.

5.1.2. Decision Trees

Decision trees is a method that works on dividing the space into different regions based on the predictor variables. It works well for both linear and non-linear decision boundaries. But for our data, it doesn't give satisfactory performance.

5.1.3. K-Nearest-Neighbor

If what we established above about the boundary being non-linear and our intuition are correct, KNN should work best as it performs very well for non-linear decision boundaries when we are dealing with lower dimension data and the model learns based on neighborhood. For our data, KNN does perform the best and gives satisfactory accuracy, sensitivity and specificity.

To explore and assert our intuition further, we run KNN based on **only latitude and longitude** information. This will give us the predictions of crime index based on only space (latitude and longitude) information. We can see the following results.

	Accuracy	Sensitivity	Specificity
KNN; K=1	100%	100%	100%

Based on the latitude and longitude information, locations can be categorized as INDEX-RISK zones (Locations with prediction of index crimes) and NON-INDEX RISK zone (Locations with predictions of Non-index crimes) with an **accuracy of 100%**

5.2. Prediction of Arrests

Looking at the arrest data, we observe that there is imbalance in the data. The percentage of crimes resulting in arrests are only 20% of the entire data set of crimes.

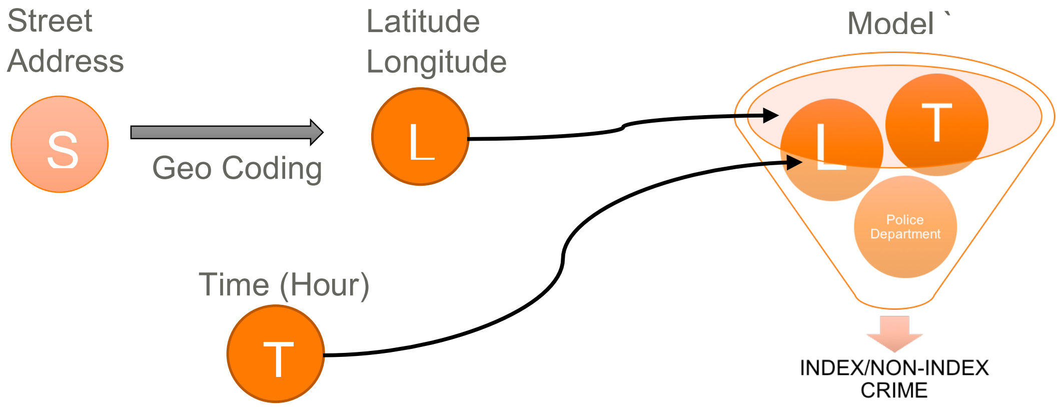
5.2.1. Logistic regression

The logistic regression model predicts very low probabilities for arrests and using a threshold of 0.4 results in all cases being predicted as negative. Decreasing the threshold value further, decreases the accuracy of the model.

5.2.2. K-Nearest-Neighbor

For the prediction of arrest, KNN with K=10 works the best. It gives us an accuracy of ~80%. Considering that we have used a 50% random sample of 2 years of **IMBALANCED** data (too make the data volume manageable) this accuracy is decently sufficient, but of course ha scope for improvement.

6. Process Flow



7. Conclusion

We have successfully predicted the crime index based on time and space information of locations in the city of Chicago, with an accuracy of ~98%. We have predicted if the location (latitude and longitude) in Chicago, is an INDEX RISK ZONE or a NON-INDEX RISK zone with an accuracy of 100%. We have also predicted whether a crime will result in an arrest, based on where the crime took place and which police Community was in-charge.