Project Report

on

“Crime Analysis and Prediction in Chicago using Data Mining”

By

#Team 4

Shivani Ghatge

Vinaya Chinti

INTRODUCTION

Crime is prevalent in society and precautions need to be taken to avoid it. Modern data science techniques can be used as precautionary measures to avoid the crime. The Chicago Police Department has records and documentation that have been gathered over the years. We analyze the crime datasets to find trends in the data and build a model to predict the crime type. The process of building a model involves following steps.

1) Analyzing the data to find trends in the data and decide on the dependent variable in our case the dependent variable is the Primary Type column.
2) After the data is analyzed we need to do data preprocessing to find out which columns have null values. After the data is cleaned, we need to extract the important features from the data that will be used for training the model.
3) After feature extraction is done, we need to divide the data into training and test data sets. The training data is used for building the model whereas the testing data is used for testing the model and finding the accuracy of the model.

SYSTEM DESIGN AND ARCHITECTURE

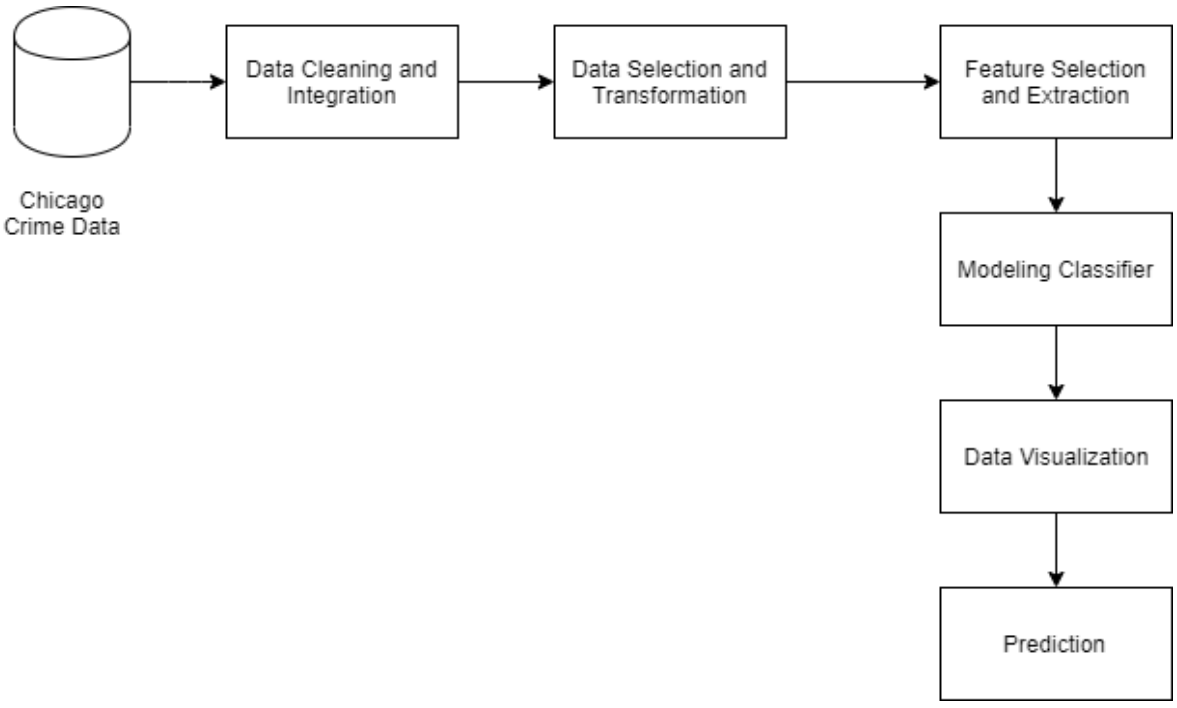


Figure 1: System Architecture

DATA CAPTURE

The data is taken from the Chicago Data portal. It has data from year 2001 to present year(2019)
Source: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

ID	Case Num	Date	Block	IUCR	Primary Ty	Descriptio	Location D	Arrest	Domestic	Beat	District	Ward	Communit	FBI Code	X Coordin	Y Coordin	Year	Updated On	Latitude	Longitude	Location
1.1E+07	HZ250496	5/3/2016 23:40	013XX S S	486	BATTERY	DOMESTI(	APARTME	TRUE	TRUE	1022	10	24	29	08B	1154907	1893681	2016	5/10/2016 15:56	41.8641	-87.7068	(41.864073157, -87.706818608)
1.1E+07	HZ250409	5/3/2016 21:40	061XX S D	486	BATTERY	DOMESTI(	RESIDENC	FALSE	TRUE	313	3	20	42	08B	1183066	1864330	2016	5/10/2016 15:56	41.7829	-87.6044	(41.782921527, -87.60436317)
1.1E+07	HZ250503	5/3/2016 23:31	053XX W (	470	PUBLIC PE	RECKLESS	STREET	FALSE	FALSE	1524	15	37	25	24	1140789	1904819	2016	5/10/2016 15:56	41.8949	-87.7584	(41.894908283, -87.758371958)
1.1E+07	HZ250424	5/3/2016 22:10	049XX W I	460	BATTERY	SIMPLE	SIDEWALK	FALSE	FALSE	1532	15	28	25	08B	1143223	1901475	2016	5/10/2016 15:56	41.8857	-87.7495	(41.885686845, -87.749515983)
1.1E+07	HZ250455	5/3/2016 22:00	003XX N L	820	THEFT	$500 AND	RESIDENC	FALSE	TRUE	1523	15	28	25	6	1139890	1901675	2016	5/10/2016 15:56	41.8863	-87.7618	(41.886297242, -87.761750709)
1.1E+07	HZ250447	5/3/2016 22:35	082XX S M	041A	BATTERY	AGGRAVA	STREET	FALSE	FALSE	631	6	8	44	04B	1183336	1850642	2016	5/10/2016 15:56	41.7454	-87.6038	(41.745354023, -87.603798903)
1.1E+07	HZ250489	5/3/2016 22:30	027XX S S	460	BATTERY	SIMPLE	CHA HALL	FALSE	FALSE	133	1	3	35	08B	1176730	1886544	2016	5/10/2016 15:56	41.844	-87.6269	(41.844023772, -87.626923253)
1.1E+07	HZ250514	5/3/2016 21:30	002XX E 4	460	BATTERY	SIMPLE	RESIDENC	FALSE	FALSE	215	2	3	38	08B	1178514	1874573	2016	5/10/2016 15:56	41.8111	-87.6207	(41.811133958, -87.62074077)
1.1E+07	HZ250523	5/3/2016 16:00	014XX W I	460	BATTERY	SIMPLE	SIDEWALK	FALSE	FALSE	2432	24	40	1	08B	1165696	1942616	2016	5/10/2016 15:56	41.9981	-87.6658	(41.99813061, -87.665814038)
1.1E+07	HZ250667	5/3/2016 22:30	069XX S A	486	BATTERY	DOMESTI(	STREET	FALSE	TRUE	735	7	17	67	08B	1166876	1858796	2016	5/10/2016 15:56	41.7681	-87.6639	(41.768096835, -87.663878589)
1.1E+07	HZ250469	5/3/2016 21:44	074XX S S	143A	WEAPONS	UNLAWFU	VEHICLE N	TRUE	FALSE	334	3	7	43	15	1195696	1856719	2016	5/10/2016 15:56	41.7617	-87.5583	(41.761733286, -87.558309979)
1.1E+07	HZ250541	5/3/2016 23:11	006XX N V	486	BATTERY	DOMESTI(	SIDEWALK	TRUE	TRUE	1834	18	42	8	08B	1176630	1904401	2016	5/10/2016 15:56	41.893	-87.6268	(41.893026751, -87.626750829)
1.1E+07	HZ250415	5/3/2016 17:30	011XX W J	890	THEFT	FROM BU	OTHER	FALSE	FALSE	1232	12	2	28	6	1168776	1898793	2016	5/10/2016 15:56	41.8778	-87.6558	(41.877811861, -87.655758012)
1.1E+07	HZ250513	5/3/2016 9:00	028XX S D	820	THEFT	$500 AND	STREET	FALSE	FALSE	133	1	4	35	6	1179375	1886199	2016	5/10/2016 15:56	41.843	-87.6172	(41.843016958, -87.61722727)

Figure 2: Snapshot of Data

DATA DESCRIPTION

The data has 7012575 rows and 22 attributes. The description of each of the attributes is given below:

1) ID: Unique identifier for the record.
2) Case Number: The Chicago Police RD Number(Record Division Number) which is unique to the incident.
3) Date: Date when the incident occurred.
4) Block: The partially redacted address where the incident occurred, placing it on the same block as the actual address.
5) IUCR: The Illinois Uniform Crime Reporting code. It is directly linked to the primary type and description.
6) Primary Type: The primary description of IUCR code.
7) Description: The secondary description of the IUCR code, a subcategory of the primary description.
8) Location Description: Description of the location where the incident occurred.
9) Domestic: Whether the incident was domestic related as defined by the Illinois Domestic Violence Act.
10) Beat: Indicated the beat where the incident occurred. A beat is the smallest police geographic area.
11) District: Indicates the district where the incident took place.
12) Ward: The ward(City Council District) where the incident occurred.
13) Community Area: Indicated the community area where the incident occurred. Chicago has 77 community areas.
14) FBI code: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System.
15) X-coordinate: The X-coordinate of the location where the incident occurred.
16) Y-coordinate: The Y-coordinate of the location where the incident occurred.

17) Year: Year the incident occurred.
18) Updated On: Date and time the record was last updated.
19) Latitude: The latitude of the location where the incident occurred.
20) Longitude: The longitude of the location where the incident occurred.
21) Location: The location where the incident occurred. It is a combination of latitude and longitude.

Out of the 22 attributes, 11 are location attributes, 3 are time attributes, 3 for IUCR code, 3 attributes are nominal which are used for unique identification and 2 are categorical attributes.

DATA CLEANING AND FEATURE EXTRACTION

There are missing attributes for the ward, community area, X coordinate, Y coordinate, Latitude, Longitude and Location attributes. ID column was removed as it is just an identification number and will not be helpful for prediction. Case number and IUCR columns were removed as well as they represent the case identification and crime reporting code respectively. X-coordinate, Y-Coordinate, Latitude, Longitude and location are removed as there are other attributes that represent the location where the incident occurred like the ward, community and considering them would result in redundancy of attributes. The missing values for ward and community were replace by the mode of respective columns. Mode represents a variable which occurs most frequently.

DATA TRANSFORMATION

Domestic and Arrest attributes have Boolean values that is True or false. We have converted these values to 1 and 0. 1 represents true and 0 represents false. The date attribute represented a sting value, so it was converted to datetime data type. Further, the date attribute was subdivided into month and day columns. We have used Principle Component Analysis on the data with components as 5 and standard scalar is also applied on the data .Primary type is the independent variable which we will be used for predicting. It has 31 different crime types. To reduce the number of the categories used for prediction we have reduced the categories of primary type from 31 categories to 10 categories. Following table shows the grouped variables and their categories.

Primary Type	New Primary Type
CRIM SEXUAL ASSAULT, PROSTITUTION, SEX OFFENSE	SEX
MOTOR VEHICLE THEFT	MVT
GAMBLING, INTERFERENCE WITH PUBLIC OFFICER, INTIMIDATION, LIQUOR LAW VIOLATION, OBSCENITY, NON-CRIMINAL, PUBLIC PEACE VIOLATION, PUBLIC INDECENCY, STALKING, NON-CRIMINAL(SUBJECT SPECIFIED)	NONVIO
CRIMINAL DAMAGE	DAMAGE
CRMINAL TRESPASS	TRESPASS

NARCOTICS, OTHER NARCOTIC VIOLATION	DRUG
DECEPTIVE PRACTICE	FRAUD
OTHER OFFENCE, ARSON, BATTERY, DOMESTIC VIOLENCE, HOMICIDE, RITUALISM	OTHER
KIDNAPPING, WEAPON VIOLATION, OFFENCE INVOLVING CHILDREN	VIO
THEFT, ROBBERY, BURGLARY	THEFT

Table 1: Categories for Primary Type

	Arrest	Domestic	Beat	District	Ward	Community Area	Year	Month	Day	Primary Type
0	0	0	1822	18.0	27.0	8.0	2013	12	10	SEX
1	0	0	422	4.0	7.0	46.0	2019	07	15	SEX
2	1	0	2313	19.0	46.0	3.0	2002	09	08	DRUG
3	0	0	624	6.0	6.0	44.0	2015	11	02	OTHER
4	0	1	833	8.0	42.0	25.0	2001	02	14	OTHER

Figure 3: Data frame after Preprocessing and Transformation

As it can be seen from the image above that the final data frame which will be used for model building contains Arrest, Domestic, Beat, Ward, Community Area, Month, Day and Primary Type attributes. From the attributes mentioned above Primary Type is the dependent variable.

Exploratory Analysis on Dataset

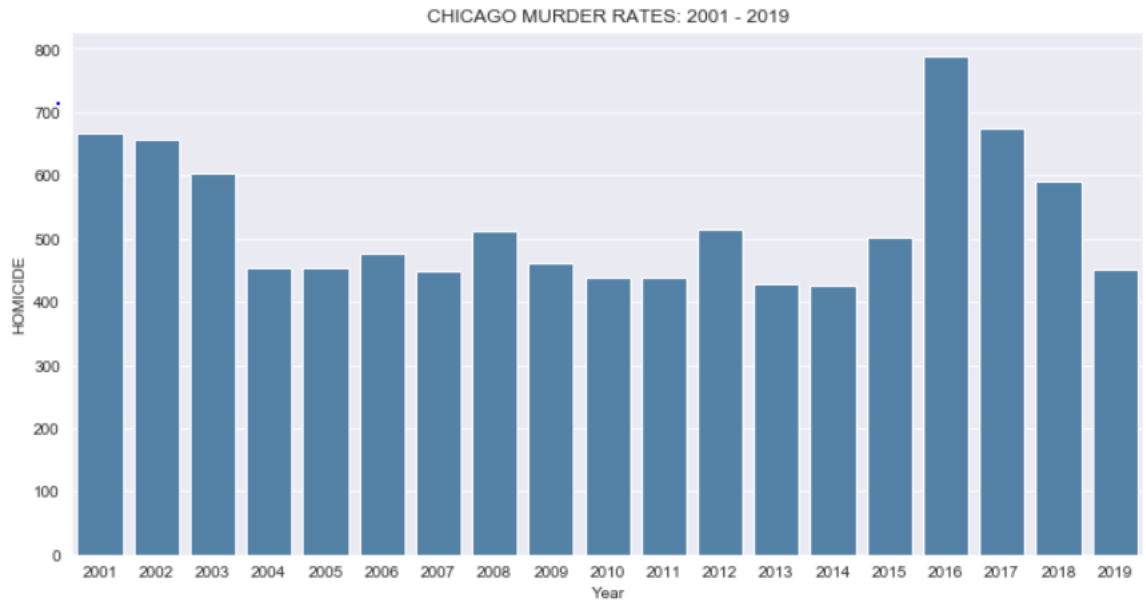


Figure 4: Year Wise Crime Rate

From the graph above 2016 has the highest number of crimes.

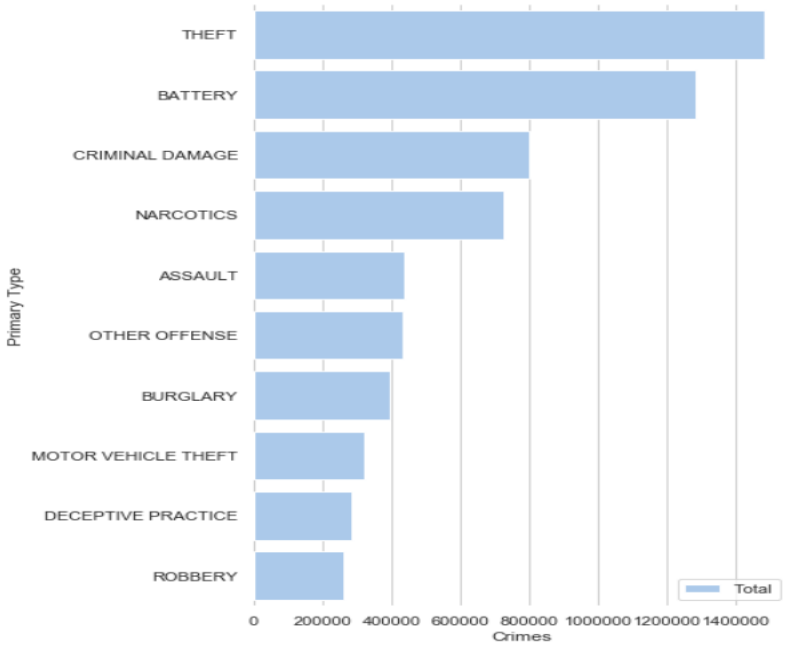
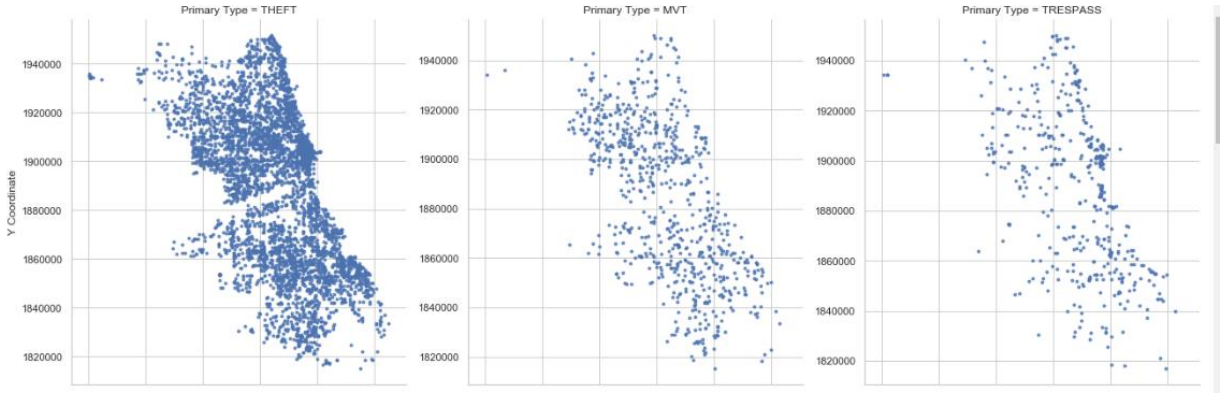


Figure 5: Number of Crimes from each category

From the figure above it can be seen that theft is the most frequent type of crime after Battery and criminal damage.



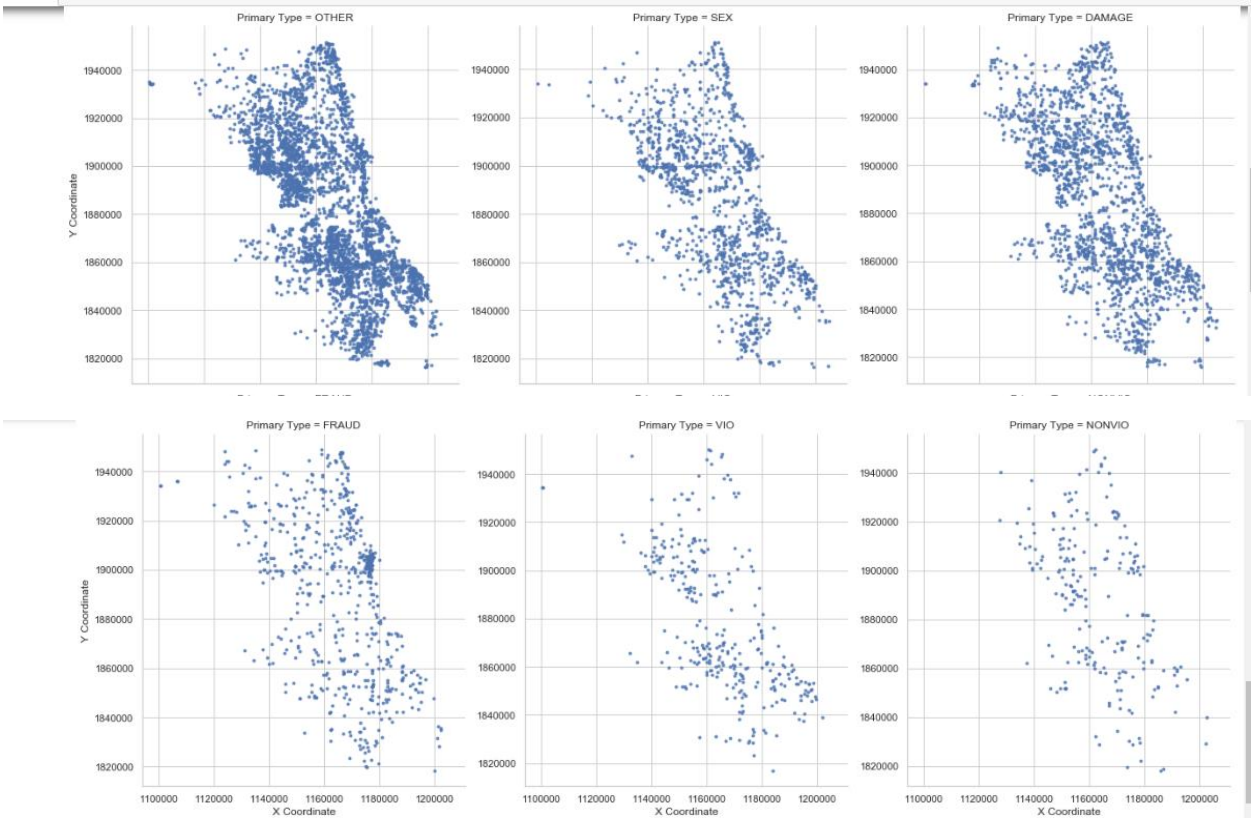


Figure 6: Scatter Plot for the Primary Type

Figure above plots the X and Y co-ordinates and shows that theft is the most common crime type followed by other crimes. This makes the dataset biased towards the 'THEFT' and 'OTHER' primary type.

DATA MINING

Decision Tree

The decision tree has flowchart like structure with root node, internal nodes, and leaf nodes. The root node is a node with no incoming links and zero or more outgoing links. Internal nodes have precisely one incoming link and two or more outgoing links. Leaf and terminal nodes each of which has precisely one incoming link and no outgoing link. Every leaf node in the decision tree is associated with a class label in our dataset, it will be the primary data type. The non-terminal nodes, which include the root and the internal nodes, contain test conditions that are typically defined using a single attribute. Each possible outcome of the attribute test condition is associated with exactly one child of this node. Given a decision tree, classifying a test instance is straight forward. Starting from root node we apply its attribute test condition and follow the appropriate branch based on the outcome of the test .

These are the hyper parameters which we got after hyper parameter tuning and the model is built using these Hyper parameters

Best hyperparameters:
{'max_depth': 7, 'max_features': 0.4, 'min_samples_leaf': 0.02}

Random Forest

Random forest consists of large number of individual decision trees that function as an ensemble. Each decision tree gives a class prediction and the class with the most votes becomes the prediction for our data. The individual decision trees should have the least correlation between them. If the trees are correlated, then they protect each other for their individual errors. The chance of high accuracy is more when the trees have least correlation among them.

These are the hyper parameters which we got after hyper parameter tuning and the model is built using these Hyper parameters

Best hyperparameters:
{'criterion': 'entropy', 'min_samples_leaf': 3, 'min_samples_split': 7, 'n_estimators': 200, 'n_jobs': -1, 'random_state': 123}

Naïve Bayesian Classifier

Naïve Bayesian classifier is a supervised learning algorithm, which is effective and widely used. It is a statistical model that predicts class membership probabilities based on Bayes' theorem. It assumes the independent effect between attribute values.

This formula is used for naïve Bayes classification:

$$P(H|X) = P(X|H)\,P(H)/\,P(X)$$

Where,

P (H|X) –Posterior Probability
P (X|H) - Likelihood
P (H) – Class Prior Probability
P (X)- Predictor Prior Probability

K-Nearest Neighbor

K Nearest Neighbor is a simple supervised learning algorithm that can be used for both classification and regression. The drawback of this algorithm is it become slow as the data grows.

Best Hyper Parameters:
{'algorithm': 'auto', 'leaf_size': 1, 'n_jobs': -1, 'n_neighbors': 10, 'weights': 'uniform'}

Support Vector Machine
A support vector machine is a discriminative classification model that learns linear or nonlinear decision boundaries in the attribute space to separate the classes. It offers strong regularization capabilities i.e. it can control the complexity of the model in order to ensure good generalization performance.

Best Hyper Parameters:
{'algorithm': 'auto', 'leaf_size': 1, 'n_jobs': -1, 'n_neighbors': 10, 'weights': 'uniform'}

METRICS AND PERFORMANCE EVALUATION

Following Performance metrics will be used in order to evaluate the performance of the models

1. Accuracy: Accuracy is the fraction of predictions that our model is predicted correctly.

   Accuracy= (True Positives+ True Negatives)/Total

2. Sensitivity: It is also known as Recall and it measures the proportion of actual positives that got predicted as positives.

   Sensitivity = (True Positive)/ (True Positive + False Negative)

3. Specificity: It is the proportion of actual negatives that got predicted as negatives.

   Specificity = (True Negative)/ (True Negative + False Positive)

4. Precision: Precision defines how precise the model is out of those predicted positive.
   Precision= (True Positive)/Total Predicted Positive

5. F1-Score: Precision and Recall are combined as a harmonic mean to find the F1-Score.

6. Prevalence: It is the fraction of data that is positive. It is an important factor which can be used to see if the data used for training is balanced or unbalanced.

SOFTWARE AND HARDWARE DISCUSSION

Software:
Jupyter Notebook
Python 3.7.4

Hardware:
16 GB RAM
Core i7 Processor
64-bit Windows Operating System

Results

We have randomly sampled the data and worked on a data frame containing 16000 rows. The data was split into training and test sets. 80% was used for training and 20% was used for testing. Models were build using the 80% training data and tested on the 20% test data. Test train split and 5-fold cross validation is used for validations. The test data has 5000 rows. Test data was cleaned and given to the models built. The accuracies that we obtained on the train test split data, Cross validation and test data are summarized in the table below:

	Decision Tree	**Random Forest**	**KNN**	**Naïve Bayes**	**SVM**
Accuracy	46.15	45.375	40.09	23.84	46.18
Cross Validation Accuracy	47.06	45.5	28.74	32.93	
Accuracy on test data	46.73	45.30	29.16	31.74	24.54

From table above we can see that decision tree performed best in all the three cases as compared to the other models. We have observed that our data is skewed it has more samples of "THEFT" and "OTHER" primary type. So, when we see the classification report of models, we see that the samples for theft and other primary type have highest accuracy of getting predicted correctly. The decision tree performs well on skewed data which is the reason it performed well on our data set.

CONCLUSION AND FUTURE WORK

We have used the Chicago crime dataset to predict the type of crime based on the attributes chosen after data cleaning and data transformation. We built 5 models on the data namely, decision tree, Random forest, KNN, Naïve Bayes and SVM. From the models-built decision tree performed well on test as well as validation data set.
In the future, we can use neural networks to predict the crime type and evaluate its performance. We can apply the models built, on crime data available from cities other than Chicago. We can use a distributed file system for the data processing.