University of New Haven

Data Mining

CSCI-6401-01

Final Report

CRIME ANALYSIS USING DATA MINING TECHNIQUES

Contributors:

Akkanapally NaveenKrishna - nakka2@unh.newhaven.edu

Aravind sudulaguntla - asudu1@unh.newhaven.edu

Submitted to:

Dr. Shivanjali Khare

Abstract:

Crimes are increasing day by day with an inclined trend. The goal of this research is to calculate, based on the city's crime statistics from the past, how likely it is that a specific address would be the scene of a criminal event in the near future. In order to predict crime hotspots. To predict the crime hot spots using Three of those algorithms like Gradient Boosting Classifier,Logistic Regression,K Neighbors Classifier.

The data used for the prediction was collected from Kaggle. Performance of different algorithms were then compared to find one that best suits the data set. Furthermore, the model was evaluated using test data and the comparison of different evaluation metrics like accuracy in this experimental learning the greatest accuracy is obtained with catboost algorithms with scores of test 1.0 and 1.0.

Introduction:

According to data released by the FBI, violent crime rise by an estimated 1% in 2021 compared with the previous year. However, the number of murders increased by more than 4%. For predicting of the hotspots, we took Chicago Crime dataset which have 1048517 rows17 columns in which the attributes are 'Date', 'IUCR', 'Primary Type', 'Location Descript ion', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Latitude', 'Longitude', 'Location' .

The dataset collected in the Kaggle,the purpose is about to predict the potential customer. We split the historical data set from the competition into two groups– train set, testset.The train dataset is going to be input for the training model.This test data set will be used to check the correctness for the generations of predictions for the test data. There can be many flaws in the train set because of which training of the model on the basis of such data is not possible or can lead to an incorrect output.To predict the crime hot spots using Three of those algorithms like Gradient Boosting Classifier,Logistic Regression,K Neighbors Classifier.

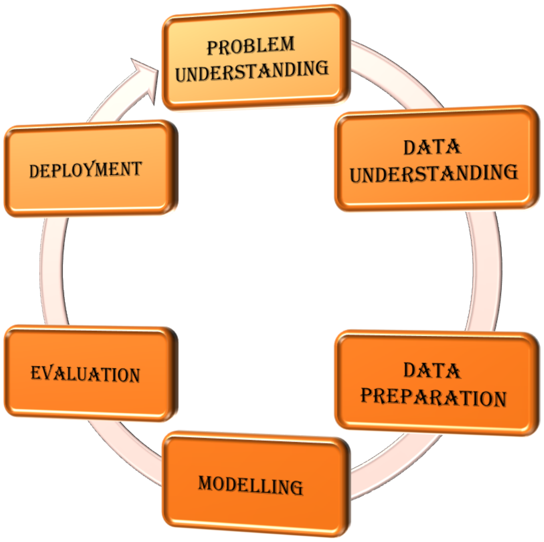
Related work :

Paper	Author	Learning Outcome
Z - CRIME: A Data Mining Tool for the Detection of Suspicious Criminal Activities Based on Decision Tree	Mugdha Sharma	Using ID3 to predict the suspicious emails with accuracy 0.75
Prediction of Future possible offender's network and role of offender's	Sushant Bharti and Ashutosh Mishra	Getting the confidence score between the offenders network and offenders
Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology	Mohammed MahmoodAli, Khaja Moizuddin Mohammed	Using , Association rule mining get the confidence of messages which is suspicious

A Multivariate Time Series Clustering Approach for Crime Trends Prediction	B. Chandra, Manish Gupta	Using time series algorthims solved with accuracy of 0.86
A Survey of Cyber Crime in India	Vinit Kumar Gunjan, Amit Kumar	Using Data mining technique to improvise surveillance

Proposed methodology:

Approach for prediction of the potential customer proposed in this paper is composed of several steps :

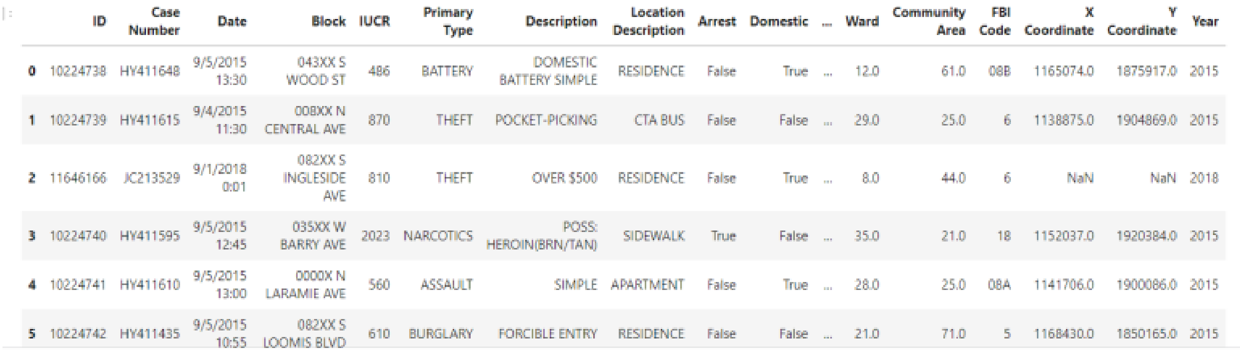


Problem Understanding:
In this first stage of the methodology we understand what the business wants to solve and how it is impactful for the society. We determine the business question and objective: In this stage we are solving the most common problem face by all people in the society fear of live, how and where to live in the society. Using previous data we are resolving this problem. Based on the city's crime statistics from the past, how likely it is that a specific address would be the scene of a criminal event in the near future. In order to predict crime hotspots

Data Understanding:

Data Collection:

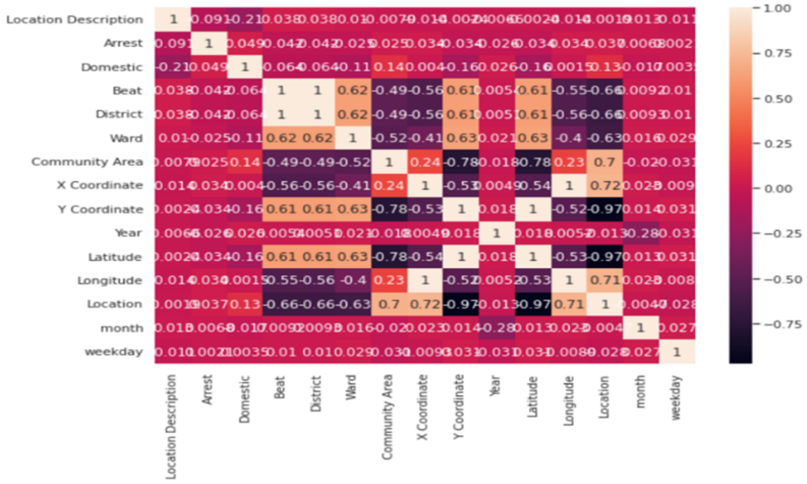
Data is collected from kaggle organized by HU Berlin The following attributes were captured from the dataset. After raw data has been collected and stored to local database, The primary Raw data contains 1048517 rows samples with 17 features attributes like



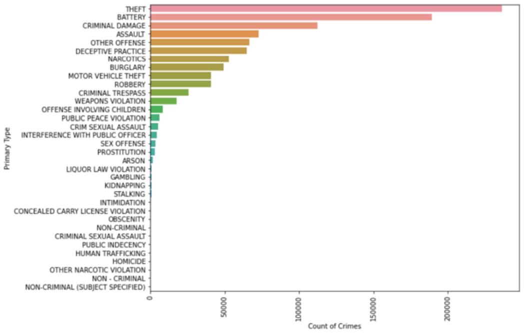
	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year
0	10224738	HY411648	9/5/2015 13:30	043XX S WOOD ST	486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	True	...	12.0	61.0	08B	1165074.0	1875917.0	2015
1	10224739	HY411615	9/4/2015 11:30	008XX N CENTRAL AVE	870	THEFT	POCKET-PICKING	CTA BUS	False	False	...	29.0	25.0	6	1138875.0	1904869.0	2015
2	11646166	JC213529	9/1/2018 0:01	082XX S INGLESIDE AVE	810	THEFT	OVER $500	RESIDENCE	False	True	...	8.0	44.0	6	NaN	NaN	2018
3	10224740	HY411595	9/5/2015 12:45	035XX W BARRY AVE	2023	NARCOTICS	POSS: HEROIN(BRN/TAN)	SIDEWALK	True	False	...	35.0	21.0	18	1152037.0	1920384.0	2015
4	10224741	HY411610	9/5/2015 13:00	0000X N LARAMIE AVE	560	ASSAULT	SIMPLE	APARTMENT	False	True	...	28.0	25.0	08A	1141706.0	1900086.0	2015
5	10224742	HY411435	9/5/2015 10:55	082XX S LOOMIS BLVD	610	BURGLARY	FORCIBLE ENTRY	RESIDENCE	False	False	...	21.0	71.0	5	1168430.0	1850165.0	2015

Exploratory Data Analysis:

Checking Correlation between the Attributes:

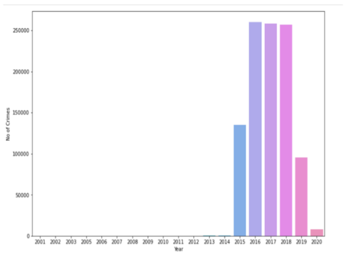


From the above image it will clearly visible that the attributes are clearly correlated with each other.

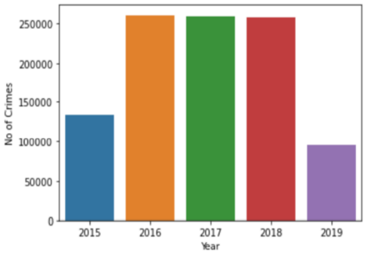


From the above graph it was clearly visible that theft cases are more in this data set.

Showing Crimes Vs Year

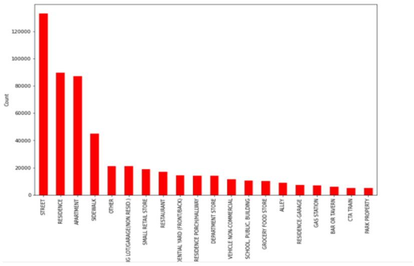


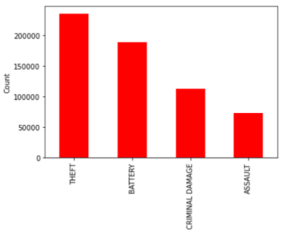
Showing clear view major years vs crime



From above image it was clearly shows that from the year 2015 to 2019 the crimes are increaing rapidly.

Types vs count

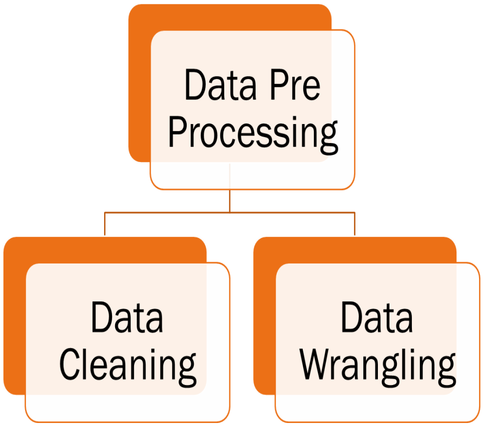




From above image it was clearly shows that from the year Theft have many cases then any other crimes.

Data Prepartion:

Data prepration further divided into two categories in that Data pre processing and Data Wrangling.



In this stage data wents into two forms first though Data cleaning and second through data wrangling

Data cleaning:

Missing values and outliers are frequently encountered while collecting data. The presence of missing values reduces the data available to be analyzed, compromising the statistical power of the study, and eventually the reliability of its results. In addition, it causes a significant bias in the results and degrades the efficiency of the data. Outliers significantly affect the process of estimating statistics (*e.g.*, the average and standard deviation of a sample), resulting in overestimated or underestimated values.

Mising values; I used the following methods to replace missing values in the attribute like Simple Imputer. Simple Imputer is a scikit-learn class. It is one of the most useful techniques to handle the missing data in the predictive model dataset. It replaces the Nan values with a specified placeholder. It is implemented by the use of the Simple Imputer () method which takes the arguments like missing values, fill value and strategy By using Simple imputer we fill the data with mean,median and mode for the suitable variable.

Outliers: outliners is the heart for feature Engineering because this can influence the distribution and effect the model. There are different techniques to detect and handle outliers. To detect the outliers we used Boxplot, Inter quartile range. Treating outliers is the most important aspect in feature engineering. We used Fissurization, Arbitrary Outlier Capper and rectify using Inter quartile range.

Data Wrangling:

In this data wrangling method we solved using Encoding, Transformation and Train Test Spilt.

Encoding: Categorical data refers to variables that are made up of label values, for example, a "color" variable could have the values "red", "blue, and "green". Think of values like different categories that sometimes have a natural ordering to them.

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.

Transformation:Data transformation is the process of converting data from one format to another. The most common data transformations are converting raw data into a clean and usable form, converting data types, removing duplicate data, and enriching the data to benefit an organization. During the process of data transformation, an analyst will determine the structure, perform data mapping, extract the data from the

original source, execute the transformation, and finally store the data in an appropriate database. We used minmax scaler to change the data

Train Test Spilt: Train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data. In this Section Split the data set into two pieces — a training set and a testing set. This consists of random sampling without replacement about 80 percent of the rows (you can vary this) and putting them into your training set. The remaining 20 percent is put into your test set.

Modelling:

We used 3 types of algorithms for this problem are Gradient Boosting Classifier,Logistic Regression,K Neighbors Classifier.

Logistic Regression:

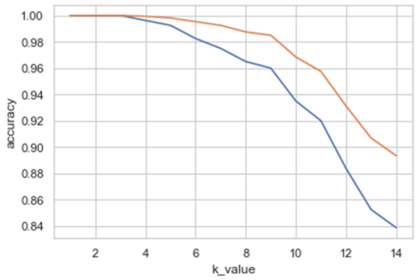
Using Default parameters:

class sklearn.linear_model.LogisticRegression(*penalty='l2'*, ***, *dual=False*, *tol=0.0001*, *C=1.0*, *fit_intercept=True*, *intercept_scaling=1*, *class_weight=None*, *random_state=None*, *solver='lbfgs'*, *max_iter=100*, *multi_class='auto'*, *verbose=0*, *warm_start=False*, *n_jobs=None*, *l1_ratio=None*)
The accuracy of the train data is 0.41375
The accuracy of the test data is 0.385

K Nearst Neighbors Classifier.

class sklearn.neighbors.KNeighborsClassifier(*n_neighbors=5*, ***, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None)*
The accuracy of the train data is 0.998125
The accuracy of the test data is 0.9925

We optimised N neighbours using for loop and graph shown as below.



The accuracy of the train data is 1.0

The accuracy of the test data is 1.0

Gradient Boosting Classifier:

class sklearn.ensemble.GradientBoostingClassifier(***, *loss='log_loss'*, *learning_rate=0.1*, *n_estimators=100*, *subsample=1.0*, *criterion='friedman_mse'*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0.0*

The accuracy of the train data is 0.9246875
The accuracy of the test data is 0.90375

From above w e can tell that KNN is the best algorithm to resolve this issue.

Results:

Algorthim	Train	Test
KNN Algorthim	1.0	1.0
Gradient Boosting Machine	0.92	0.92
Logistic Regression	0.41	0.38

The model that performed best was KNN, With an accuracy of 100%. The precision and recall are all weighted as 56%. This implies that the model can predict a particular area's crime rate with 100% accuracy. Precision is the fraction of relevant instances, while recall is retrieved instances among all instances. In other words, precision is the number of correctly predicted positive values divided by the total number of predicted positive values. The recall is the number of correctly predicted positive values divided by the number of all relevant samples.

Conclusion:

Many factors affect the crime rate of a particular area. The crime rate of a particular area can be calculated by analyzing the city's crime statistics, which can be analyzed using data mining techniques. As we have seen, technologies like data mining and machine learning can be used to find crime hotspots and predict the crime rate of a particular area by using machine learning models. Machine learning models can be used to predict the probability of a crime occurring in a particular area. This undertaking is relevant to society because it can be used to predict the crime rate of a particular area and take necessary precautions to prevent crimes. Using KNN algorithm we solved the research question is to calculate, based on the city's crime statistics from the past, how likely it is that a specific address would be the scene of a criminal event in the near future. In order to predict crime hotspots.

Future works:
To build an extra parameter model using Deep learning with help of Neural Networks.

References:

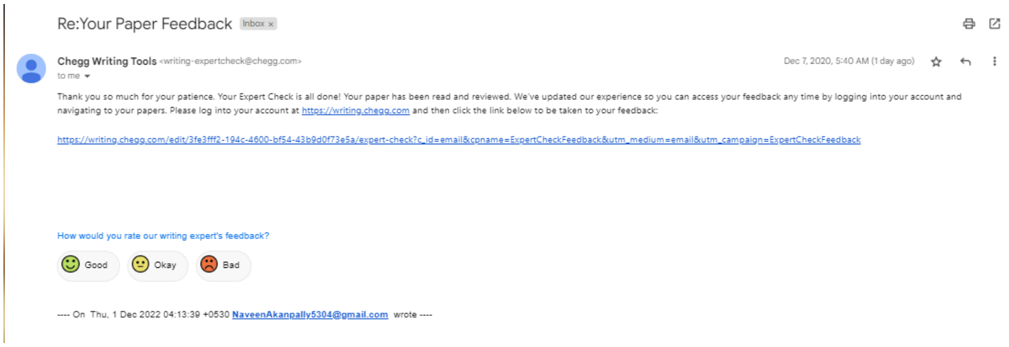
1. Z - CRIME: A Data Mining Tool for the Detection of Suspicious Criminal Activities Based on Decision Tree by Mugdha Sharma
2. Prediction of Future possible offender's network and role of offender's by Sushant Bharti and Ashutosh

Mishra

3. **Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology by Mohammed MahmoodAli, Khaja Moizuddin Mohammed**
4. **A Multivariate Time Series Clustering Approach for Crime Trends Prediction by B. Chandra, Manish Gupta**
5. **A Survey of Cyber Crime in India by Vinit Kumar Gunjan, Amit Kumar**

GitHub Link https://github.com/Naveen0970/DataMiningPhase5-.git

Exper Check Report:



Re:Your Paper Feedback

Chegg Writing Tools <writing-expertcheck@chegg.com>
to me

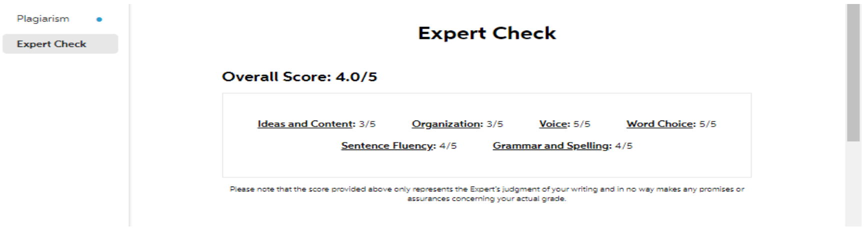
Dec 7, 2020, 5:40 AM (1 day ago)

Thank you so much for your patience. Your Expert Check is all done! Your paper has been read and reviewed. We've updated our experience so you can access your feedback any time by logging into your account and navigating to your papers. Please log into your account at https://writing.chegg.com and then click the link below to be taken to your feedback:

https://writing.chegg.com/edit/3fe3fff2-194c-4600-bf54-43b9d0f73e5a/expert-check?c_id=email&cpname=ExpertCheckFeedback&utm_medium=email&utm_campaign=ExpertCheckFeedback

How would you rate our writing expert's feedback?

---- On Thu, 1 Dec 2022 04:13:39 +0530 NaveenAkanpally5304@gmail.com wrote ----



Expert Check

Overall Score: 4.0/5

Ideas and Content: 3/5 Organization: 3/5 Voice: 5/5 Word Choice: 5/5
Sentence Fluency: 4/5 Grammar and Spelling: 4/5

Please note that the score provided above only represents the Expert's judgment of your writing and in no way makes any promises or assurances concerning your actual grade.