Chicago Crime Data Analysis

Abstract

Crime is an ever-pervasive part of society and while our police forces work tirelessly to reduce the crime rate, there is only so much they can do when working solely off their intuition and training. This is where data analytics plays a major role. Crime is rarely random, and there are vast crime archives that can provide insight into the patterns in which crimes are committed. In addition to this, there may be a multitude of factors that affect the crime rate, and the nature of the crime, which is not immediately obvious, but the identification of which could aid the police in the prediction and thus, the prevention of criminal activities. In this paper, we aim to identify such features that influence, and thus can be used to predict, the occurrence of crimes in the City of Chicago.

Keywords: Chicago, Crime, Data Analysis, Machine Learning, Supervised Learning

1. Introduction

Crime is an inextricable element of our society. We hear about them every day, and some of us have been involved in at least one of them at some point in our lives. Being cautious and improving safety is no longer a simple command. To act more wisely against this problem, we need to leverage current technology and data science methodologies. The police department has many records that have been accumulated through time and can be used as a valuable source of data for data analytics jobs. Applying analytical tasks to these data yields useful knowledge that can be utilized to improve society's safety and reduce crime rates. In our project, we use the “Crimes in Chicago” dataset, which contains incidences of crimes in Chicago from 2012 to 2017. The dataset contains 350000 records and 30 features. We initially perform Exploratory Data Analysis to get a bird’s eye view of the data to look at crime patterns throughout time, areas of most offenses, and crime hotspots. We also use machine learning approaches such as Logistic Regression, KNN, SVM, Ada Boost, Random Forest, and K-Means to predict crimes based on time, location, and other features in our project.

2. Related work

Sushant Bharti et al. proposed a hidden link algorithm to detect hidden links of the networks of co-offenders which show the possible future crime partner and different network beyond the real network. This paper also analyzes the centrality of nodes. This analysis describes the importance of nodes of the network. This is used to discover the strongest person,power of the person and role of the person in the network. This paper gave future approaches i.e. predictive approach in crime analysis which helps in stopping the crime before it occurs and also analyze the network of Co-offenders in India and predict the possible future network of offenders.

Shiju Sathyadevan et al. proposed Apriori algorithm to identify the trends and patterns in crime. This algorithm is also used to determine association rules highlighting general trends in the database. This paper has also proposed the naive Bayes algorithm to create the model by training crime data. After testing, the result showed that Naive Bayes algorithm gave 90%accuracy.

Prashant K. Khobragade et al. proposed Forensic Tool Kit 4.0 which provides remote data investigation and visualization analysis. In remote data, investigation includes to analyze process information, service information, driver information, network device, network information. This tool generates the file and analyzes the data. This tool is also used to analyze the victim system where the attack is occurring. With the help of crime investigation, the physical and logical memory data are analyzed.

K. ZakirHussain et al. used data mining techniques for analysis of criminal behavior. This paper proposed a criminal investigation analysis tool (CIA). This tool was used within the law enforcement community to help solve violent crimes. It was based on a review of evidence from the crime scene and from witnesses and victims. The analysis was done from both an investigative and a behavioral perspective. It provided insight into the unknown offender as well as investigative suggestions and strategies for interviews and trial. Mugdha Sharma et al. proposed advanced ID3 algorithm for presenting importance-attribute significance on the attributes which have less values but higher importance, rather than the attributes with more values and lower importance as well as solve the classification defect to choose attributions with more values. The analysis of the experimental data shows that the advanced ID3 algorithm gets more reasonable and more effective classification rules. In this Z-crime tool was also proposed to analyze the criminal activities through e-mail communication

3. Methodology

As our target variable is the “primary type of crime” which is a categorical variable, we need to use the algorithms which are able to predict categorical variables. Hence, we decided to use the Logistic Regression as our baseline model because it is easier to implement, interpret, and very efficient to train. It also can easily extend to multiple classes which is a good choice for our data with more than 20 different classes. As for algorithm model we decided to use KNN algorithm model because it uses nearest neighbors to predict the type of crime based on the given input features and also reduces the problem of overfitting, reduces variance and also improves the accuracy of the prediction. In the continuation of the project we will consider implementation of the other models such as SVM, RandomForest and Clustering techniques. You can see the theoretical models explanation below.

3.1 Logistic regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where

there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

Algorithm 2 Lasso Logistic Regression

Input: $D = (x^{(i)}, y^{(i)})$
repeat
  (per epoch)
  for $j = 1$ **to** n **do**
    $\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$
  end for
until convergence

Figure1. Lasso Logistic Regression

3.2 KNN

The k-nearest neighbor (KNN) rule is one of the most basic judgment processes for classification. categorizes a sample based on its closest neighbor's category. To categorize a test pattern, nearest neighbor based classifiers employ some or all of the patterns available in the training set.

To determine the K-closest neighbors, the K nearest neighbor algorithm uses the lowest distance between the query instance and the training examples. The data for the KNN algorithm is made up of a number of multivariate qualities that will be utilized to classify the data. Load the information. Set K to the number of neighbors you want. Calculate the distance between the query example and the current example from the data for each example in the data.

To an ordered collection, add the example's distance and index. Sort the distances from smallest to greatest (in ascending order) in the ordered collection of distances and indices. From the sorted collection, select the first K elements. Get the labels for the K entries you've chosen. Return the mean of the K labels if regression is true. Return the mode of the K labels if classification is true.

3.3 SVM

A support vector machine (SVM) is a supervised machine learning model that uses classification

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to have access to a dataset of at most a couple of thousands of tagged samples. As you can see in the picture Maximum-margin hyperplane and margins for a SVM trained with samples from two classes. Samples on the margin are called support vectors.

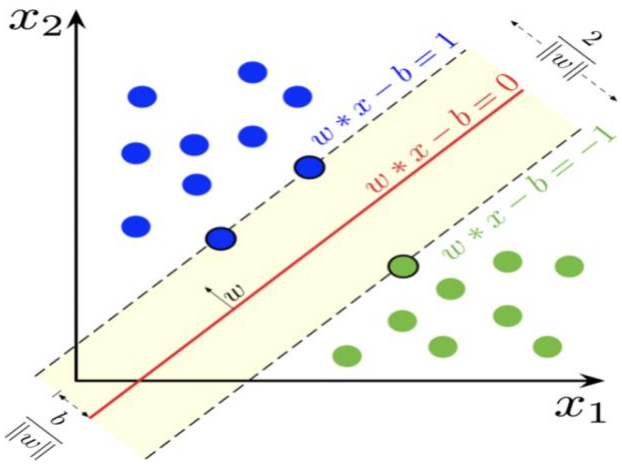


Figure2. SVM Hyperplane

3.4 Random Forest

A random forest (RF) of is a collection of tree predictors grown as follows:

The bootstrap phase: select randomly a subset of the learning dataset – a training set for growing the tree. The remaining samples in the learning dataset form a so-called out-of-bag (OOB) set and are used to estimate the RF's goodness-of-fit. The growing phase: grow the tree by splitting the training dataset at each node according to the value of one from a randomly selected subset of variables (the best split) using classification and regression tree (CART)
method.
Each tree is grown to the largest extent possible. There is no pruning. The bootstrap and the growing phases require an input of random quantities. It is assumed that these quantities are independent between trees and identically distributed. Consequently, each tree can be viewed as sampled independently from the ensemble of all tree predictors for a given learning set.

For prediction, an instance is run through each tree in a forest down to a terminal node which assigns it a class. Predictions supplied by the trees undergo a voting process: the forest returns a class with the maximum number of votes. Draws are resolved through a random selection. To present our feature contribution procedure in the following section, we need a probabilistic interpretation of the forest prediction process.
Denote by C =
{C1, C2,...,CK} the set of classes and by ΔK the set

$$\Delta K = (p1,\ldots,pK) : K\ k=1\ pk = 1 \text{ and } pk \geq 0.$$

An element of ΔK can be interpreted as a probability distribution over C. Let ek be an element of ΔK with 1 at position k – a probability distribution concentrated at class Ck. If a tree t predicts that an instance i belongs to a class Ck then we write Yˆi,t = ek. This provides a mapping from predictions of

$$\hat{Y}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{Y}_{i,t},$$

a tree to the set ΔK of probability measures on C. Let,

where T is the overall number of trees in the forest. Then Yˆi ∈ ΔK and the prediction of the random forest for the instance i coincides with a class Ck for which the k-th coordinate of Yˆi is maximal.

3.5 K-means clustering

K-means clustering is a method for grouping data points together in such a way that differences between data points in the same group are minimized. We can take n data points and partition them into k clusters

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

using these methods. The goal of the algorithm is to reduce the following function to the smallest possible value:

In this equation x is a vector that represents a specific crime instance in the data collection. The "centroid" is i. In a given cluster, it is the average point. The entire technique is based on minimizing the distance between all data points (crimes) and the cluster's associated centroid. To make all distances positive, we square the Euclidean distance between these two sites, which is a popular statistical procedure. All cluster assignments are contained in the set S. As a result, Si contains every point in the ith cluster. To get a total "distance" of all points from the centers of their associated clusters, we add all the elements in a given cluster, Si, and then add all the other clusters. We then try to reduce this by determining the best cluster assignment, S.

4. Experimental Results

Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. It contains 350000 records and 30 variables.

4.1.1. Data Description

Features that include:

- 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 a best estimate.
- Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- IUCR - The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description.
- 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.
- District - Indicates the police district where the incident occurred.

- Ward - The ward (City Council district) where the incident occurred.

- Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas.

- FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).

- 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 - 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.

- 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 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.

Features groups (Independent variable)	Features
Geographical features	1-Block=beat, district, ward=community area=X coordinate and Y coordinate=latitude=longitude=location
Time features	Year, Month, Date
Arrest feature	Arrest type
Domestic feature	Domestic Type4

Table1. SVM Dataset features

4.1.2. Exploratory Data Analysis

Before machine learning and other modeling techniques were implemented, we wanted to get a bird's eye view of the data and distribution of the target variable. We choose the type of crime "Primary Type" as our target variable. We have visualized using bar chart having different primary types of crimes. The chart provides information regarding the amount of crimes happened in Chicago with respect to each primary type. We observed that the theft, battery, criminal damage, narcotics and assault were the top five types of crimes.

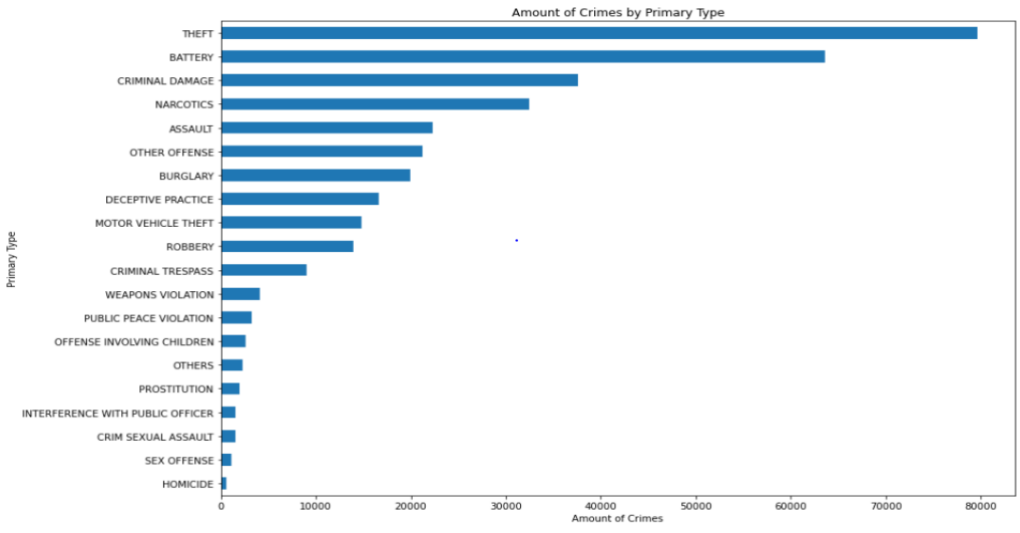


Figure3. Amount of Crimes by Primary type

We also grouped the less occured crime type into the category of others to reduce the imbalance in the data and the target class amount. Then we encoded the different types of crimes into categorical variables. We then observe how the number of crimes and arrests vary throughout the years. We see that in the year 2012, crimes were at peak. A major reason could be due to the drug abuse that was

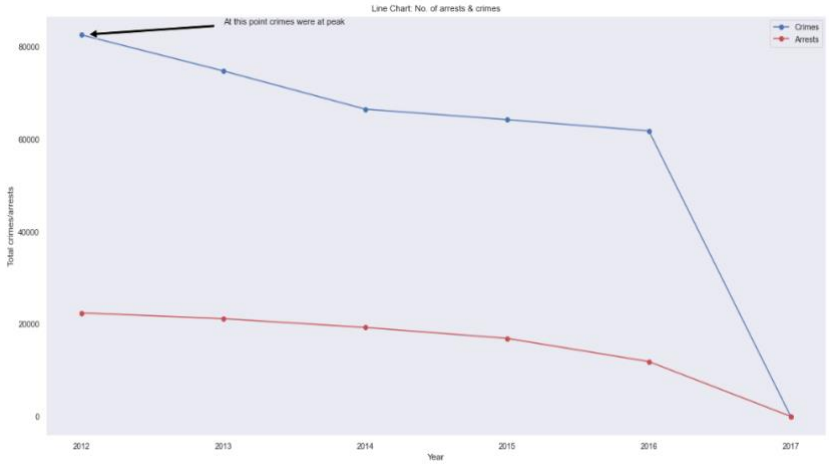


Figure4. Line Chart of Arrest and Crime

prevalent in Chicago around 2012. We see that as the years go by, the number of crimes and arrests decrease - implying that Chicago's crime department is doing pretty well at handling crimes recently. The total number of crimes occured in every particular year.

```
total_crimes_in_every_year
```

	Year	Total Crimes
0	2012	82602
1	2013	74800
2	2014	66523
3	2015	64258
4	2016	61808
5	2017	9

The total number of arrests are 91567. The total number of arrests per each year is listed below:

```
Total No.of arrests are 91567
```

	Year	Arrest
0	2012	22404
1	2013	21155
2	2014	19253
3	2015	16894
4	2016	11861

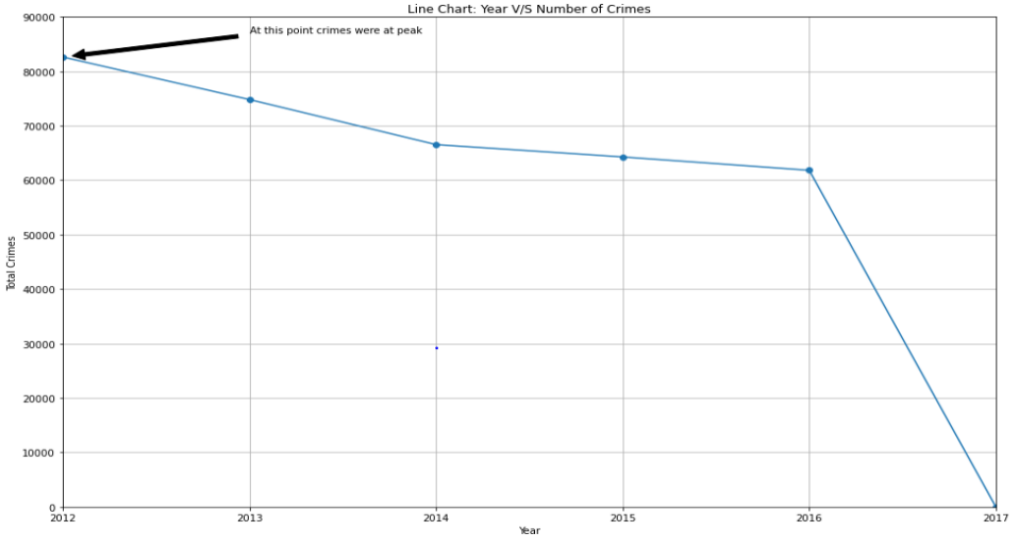


Figure5. Number of crimes per year

4.2. Feature Engineering

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. To do so, we use statistical methods to evaluate the relationship between each input variable and the target variable. These methods can be fast and effective, although the choice of statistical measure depends on the data type of both the input and output variables.

Based on our statistical and content analysis, we found out that some of our independent variables have similar meaning and also have correlation resulting in multicollinearity. The below graph represents the correlation between the variables.

Based on our analysis, we can categorize some independent variables into the same groups based on their similar meaning and correlation. As you can see in the table 1 the “Block”, “beat”, “district”, “district”, “community area”, “X-coordinate ”, “Y-coordinate”, “latitude”, “longitude”, “location” have similar meaning and are highly correlated. We used all these independent variables in our models and saw the accuracy in both baseline and main model did not significantly change, so we found out using all of them in our model just undermined the statistical significance of our independent variables. Then, we decided to select those features which are statistically significant.

We then perform ANOVA f-test for feature selection of numerical variables. Analysis of variance is a statistical method used to check the means of two or more groups that are significantly different from each other. Herewith ANOVA, we will check the variance of features to determine how much it is impacting the response variable. If the variance is low, it implies there is no impact of this feature on response and vice-versa. Only features with scores greater than 2 are considered as features of importance.

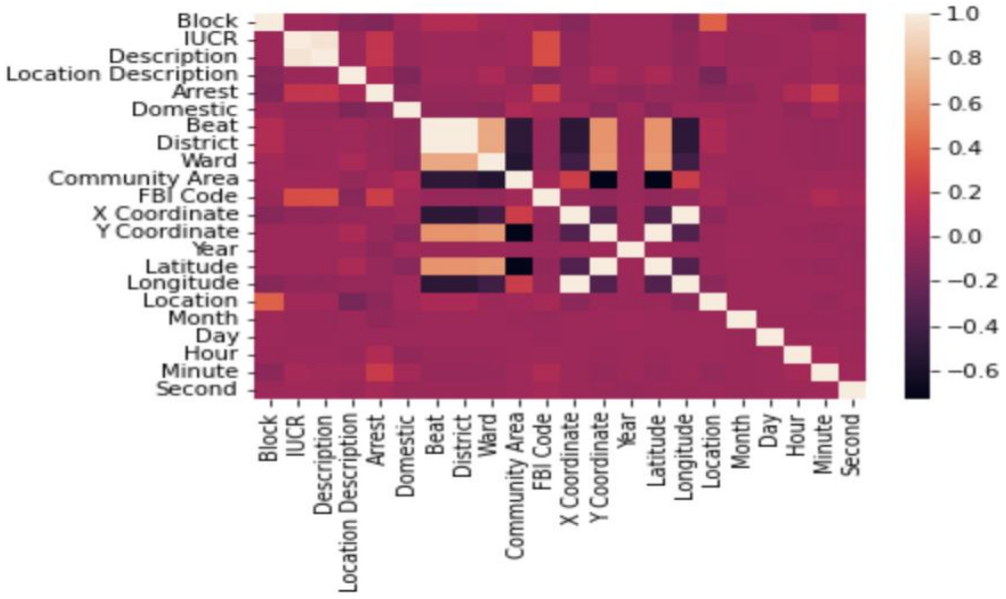


Figure 6. Correlation between features

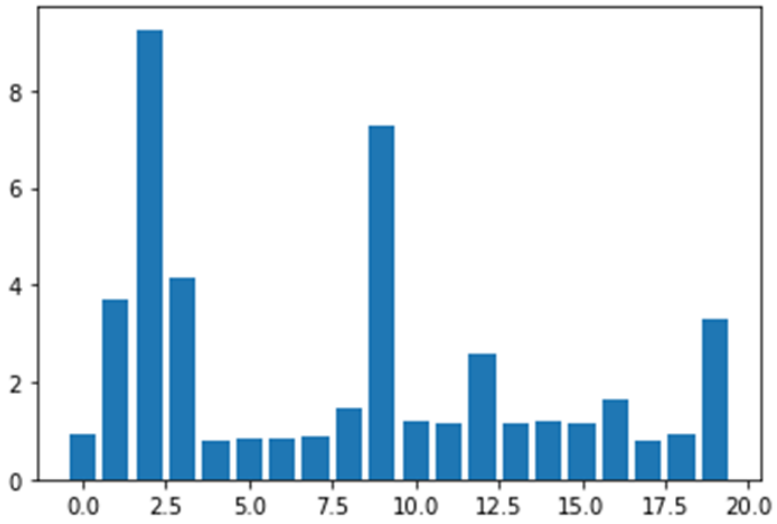


Figure 7. Features' Scores

```
# At Current Point, the attributes is select manually based on Feature Selection Part.
Features = ["Arrest","Domestic","Year","Month","Day","Hour","Longitude","Latitude","FBI Code","Location Description"]
print('Full Features: ', Features)
```

```
Full Features:  ['Arrest', 'Domestic', 'Year', 'Month', 'Day', 'Hour', 'Longitude', 'Latitude', 'FBI Code', 'Location
Description']
```

Finally, as seen in the above figure we selected these features as our independent variables. These have low correlation with together and are statistically significant. Besides, these features: "Primary type", "IUCR", and "Description" have the same meaning.

4.3. Models' Results

4.3.1 Logistic Regression

```
# Create Model with configuration
lr_model = LogisticRegression(random_state = 0, max_iter = 1000)

# Model Training
lr_model.fit(X=x1_scaled,
             y=x2)

# Prediction
result_lr = lr_model.predict(y1_scaled)
```

The result obtained from the Logistic regression had an accuracy of 0.802 for the test data. The precision, recall and f1 score for the model were 78.3%, 80.2%, and 80.7% respectively. The result obtained from the Logistic regression had an accuracy of 0.804 for the train data. The precision, recall and f1 score for the model were 78.3%, 80.4%, and 80.4% respectively.

```
========== Logistic Regression Results ==========
Accuracy     :   0.803
Recall       :   0.803
Precision    :   0.783
F1 Score     :   0.803
```

4.3.2 KNN (K-Nearest Neighbors)

```
# Model Training
knn_model.fit(X=x1,y=x2)

# Prediction
result_knn = knn_model.predict(y[Features])
```

The result obtained from the k nearest neighbors had an accuracy of 0.817 i.e for the test data. The model overall performance in classifying crimes was 81.7%. The precision, recall and f1 score for the model were 80.1%, 81.7%, and 81.7% respectively for the test data. The Accuracy for train data is 0.89 and The precision, recall and f1 score for the model were 89.9%, 89.6%, and 89.9% respectively for the train data. This result shows that the performance of the model was significantly higher in predicting the crime type.

```
========== K-Nearest Neighbors Results ==========
Accuracy     :   0.817
Recall       :   0.817
Precision    :   0.802
F1 Score     :   0.817
```

Figure 7. Knn and Logistic regression performance comparison in train data

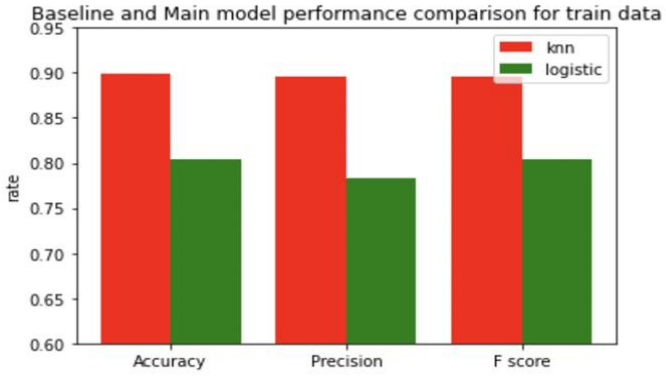


Figure D shows that our main model, the KNN model has higher performance rather than our Baseline model which is Logistic regression for the train data.

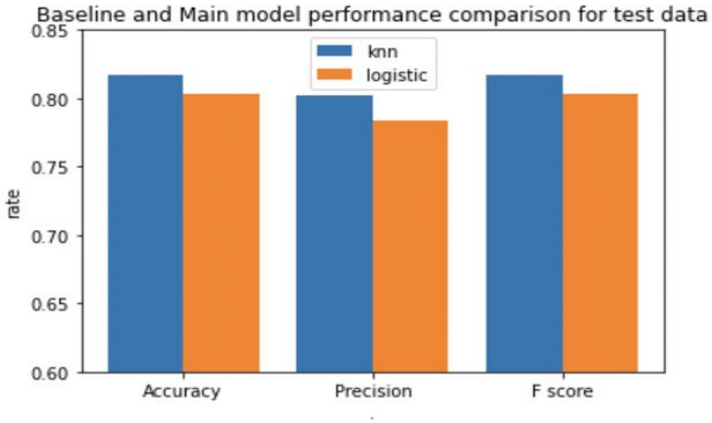


Figure 8. Knn and Logistic regression performance comparison in tet data

Figure E shows that our main model, the KNN model has higher performance rather than our Baseline model which is Logistic regression. The KNN model is better than the baseline model in predicting the primary type based on the input features.

4.3.3. SVM

```
========== SVM Results ==========
Accuracy    :  0.8865
Recall      :  0.8865
Precision   :  0.877
F1 Score    :  0.8865
```

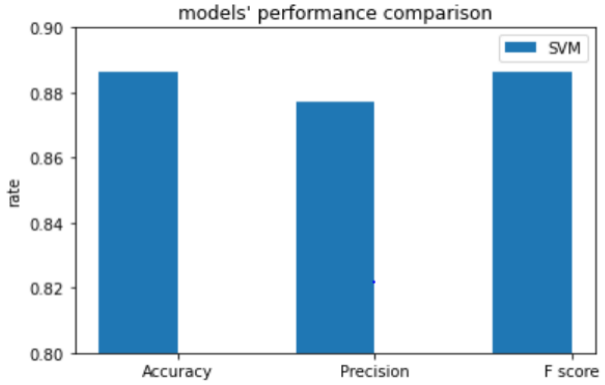


Figure 9. SVM Performance Measures

4.3.4. Random Forests

```
#Random Forest Classifier
clf=RandomForestClassifier(n_estimators=100)

#Model Training
clf.fit(X1,X2)

#Prediction
result_rf=clf.predict(Y1)
```

```
========== Random Forest Results ==
Accuracy     :  0.9385
Recall       :  0.9385
Precision    :  0.9335006118561461
F1 Score     :  0.9385
```

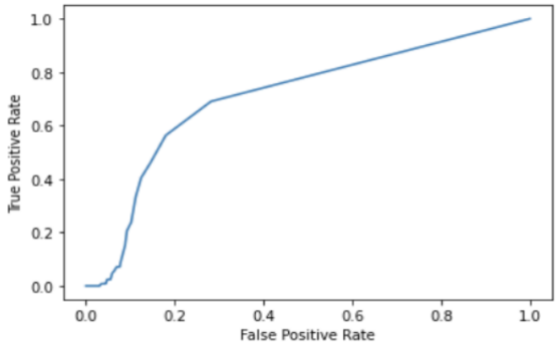


Figure 10.. Random Forests Performance Measures

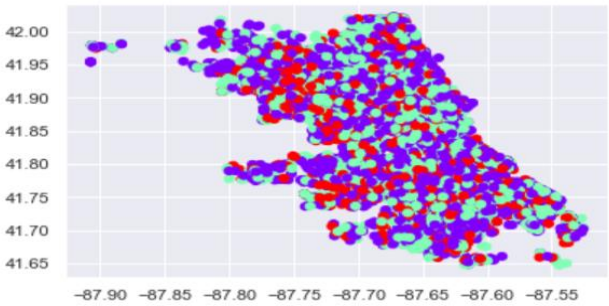
4.3.5. K-Means Clustering

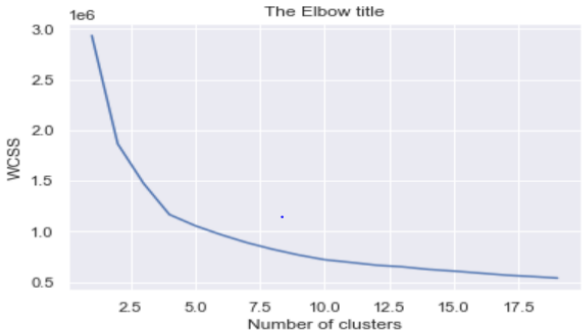
We want closer data values to correspond to more similar offenses because we're using the Euclidean distance in each dimension. This is accomplished because a small change in the date or time serialization value results in a small change in the date or time. And, because the state of Illinois uses codes to classify crimes, comparable IUCR code numbers correspond to similar offenses. For example, first and second degree murder are neighboring IUCR codes, which means that When the data, they

take on values that are next to each other. As a result, we may utilize this kind of serialization to preserve distance, allowing related items to be valued more closely.

Figure 11. K Means clustering results

As a side note, it was discovered that the city of Chicago had missing items in the IUCR code data set during the research and implementation of K-means clustering. Several IUCR codes used in the





Chicago crime data collection did not have comparable entries in their data set. I opted to believe that they are valid IUCR codes that were simply not cataloged in their IUCR code database. As a result, I hard-coded a few IUCR entries for crime codes that would fit the Euclidean distance. If IUCR codes 120 and 122 are legitimate and serialized as 3 and 4, respectively, I would serialize 121 as 4 and shift 122 to 5 in this case.We have summarized our experimental results for our models into tables.

4.3.6 Chicago's neighborhoods' crime prediction examples

	Arrest	Domestic	Year	Month	Day	Hour	Longitude	Latitude	FBI Code	Location Description	Block	Primary_type	prediction
1276629	False	False	2016	5	1	5	-87.623770	41.736313	13	5	0000X E 87TH ST	12	12
733137	True	False	2014	2	3	10	-87.625105	41.736277	16	17	0000X W 87TH ST	8	16
708368	False	False	2013	12	26	4	-87.713921	41.784256	0	0	036XX W 60TH ST	0	0
958467	False	False	2014	11	7	11	-87.727335	41.944301	1	14	034XX N PULASKI RD	10	10
580951	False	False	2013	7	24	23	-87.644803	41.780364	4	2	062XX S HALSTED ST	4	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1154600	False	False	2015	11	1	14	-87.753843	41.876672	9	52	051XX W JACKSON BLVD	5	5
439457	False	True	2013	2	10	8	-87.774158	41.909270	6	5	059XX W NORTH AVE	5	5
781543	False	False	2014	4	12	23	-87.611873	41.691874	0	18	004XX E 111TH PL	0	0
1369828	False	False	2016	9	21	23	-87.666900	41.922224	4	0	022XX N DOMINICK ST	4	4
559639	False	False	2013	6	30	22	-87.674446	41.940533	4	0	018XX W MELROSE ST	4	4

Figure 12. Crimes prediction examples for specific neighborhoods

As you can see in the above table, for instance the 051XX Jackson BLVD's crime type is 5 ("Battery") which is correctly predicted by our Random forest model. In a sample of ten data points there is just one example that was incorrectly predicted. The west 87TH ST crime's type is predicted as 16 ("Weapon Violation") but its correct type is 8 (Deceptive Practice").

5. Conclusion

With the help of machine learning technology, it has become easy to find relations and patterns among various data. The work in this project mainly revolves around predicting the type of crime and crime per capita which may happen in future. In this project, we have executed different machine learning algorithms using a real world dataset inorder to predict the crime in chicago. In our case, we experienced low correlation features with our predicting variable. We experimented with different features in order to get better predictions such as week/ month/ year to predict the crime type based on time and using additional features such as location description, arrest. The results were better however, not significant enough. As per the below results we concluded that Random Forest has the better accuracy with 0.9465 as compared to other models. Our model predicted the primary crime type with better accuracy.

Below table shows the summary of our models with accuracy, recall, precision & F1-score value.

Model	Accuracy	Recall	Precision	F1
Logistic	0.803	0.803	0.783	0.803
K-NN	0.817	0.817	0.802	0.817
Ada Boost	0.697	0.697	0.581	0.697
SVM	0.8865	0.8865	0.877	0.8865
Random Forest	0.9465	0.9465	0.945	0.9465

In this project when it comes to graph based analysis, we only showed the results of different models because it is the easiest and fastest to compute. But we can use other complex and heavy datasets as mentioned above to improve our results even further. As some factors were only considered, accuracy may vary with large datasets. For better results in prediction, more crime attributes are to be added for places instead of limited attributes. As time is an important factor in crime, it is very essential to predict not only the crime prone regions but also the proper time. In future, we would like to try to create predictions on where or when the crime will be committed. Also we can work on which areas of the city have evolved over the time span.

6. References

[2] SushantBharti, Ashutosh Mishra. “Prediction of Future possible offender’s network and role of offender’s”,Fifth International Conference on Advances in Computing and Communications, 2015.

[3] ShijuSathyadevan, Devan M.S and Surya Gangadharan.S.“Crime Analysis and Prediction Using Data Mining”,First International Conference on Networks & Soft Computing, 2014.

[4] Prashant K. Khobragade and Latesh G. Malik.“Data Generation and Analysis for Digital Forensic Application using Data mining”, Fourth International Conference on Communication Systems and Network Technologies, 2014.

[5] Amit Kumar Manjhvar (Asst.prof) and Nidhi Tomar (Research Scholar) “An Improved Optimized Clustering Technique For Crime Detection “2016 IEEE Symposium on Colossal Data Analysis and Networking (CDAN)

[6] Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes | City of Chicago | Data Portal. url: https : / / data . cityofchicago . org/Public-Safety/Chicago-Police-Department-Illinois-UniformCrime-R/c7ck-438e/data.

[7] Crimes - 2001 to present | City of Chicago | Data Portal. url: https://data. cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzpq8t2.

[8]https://ieeexplore.ieee.org/abstract/document/6642461?casa_token=nASzOzWM4YwAAAAA:-JNWuZklMT-VougLm7C-m1F9Vw9cWOEEqHar4arvgfI4sjZSyh4WT12rtsbAiYsJgIYl7b5A6Q

[9]https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/#:~:text=A%20support%20vector%20machine%20(SVM,able%20to%20categorize%20new%20text.

[10] http://journalstd.com/gallery/36-sep2021.pdf