CMPE-255 Project Report

Chicago Crime Classification

- ***Link of Dataset:-***

  Chicago Crime Classification Dataset

- ***GitHub Link of Implementation:-***

  Chicago Crime Classification Implementation

- ***Team Members:-***

1. ***Parshv Patel***
2. ***Priyansh Patel***
3. ***Het Brahmbhatt***
4. ***Monil Sakhidas***

Section 1: Introduction

Motivation:-

- The motivation behind taking this topic for the project is that in today's world every conscious person wants to live in a place that is safe and neighborhood. However we all know that crime in some manner persists in our society. While we have no control on what happens around us, we may take a few measures to assist the government and police in their efforts to maintain control. The Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system has made the 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 available to the general public. As a result, inspired by the facts mentioned above, we decided to process and evaluate the data presented in order to identify crime patterns over time and make an attempt to predict future crimes.

Objective:-

- The main objective of our project is to predict whether a given area in the city will be a crime hotspot at a given time of day with an acceptable rate of accuracy. Our project aims to draw the knowledge of criminal background from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Also the objective is to perform Exploratory Data Analysis on the dataset and draw some conclusions from the dataset. Also we will perform training using different classification algorithms and will evaluate the model on the testing dataset. So first we are going to perform exploratory data analysis to draw patterns in crime and then we will build a prediction model to predict the type of crime that can take place in Chicago in future.

Approach:-

- The approach of our project started with loading the dataset and then checking if there are any null values in the dataset or not. Since the dataset is very large (shape = (7312666, 22)) we dropped the null values from the dataset. After that we checked if there were any duplicates in the dataset or not but there weren't any duplicate values in the dataset. Then after we performed the Exploratory Data Analysis and plotted some graphs to visualize the dataset and to explain the relationship between different columns. Then we chose the target variable as "Arrest" and splitted the dataset into training, testing and validation sets.
- After that we calculated the accuracy score for our majority class baseline. We calculated the accuracy of the mode of the target variable using the mode. Then we

wrangled the dataset in terms of splitting the Date column into Month, Day, Hour and Weekday.
- We replaced False and True in the Arrest column with zeros and ones respectively. Also we dropped the 'ID' Feature as it doesn't contain any important information for the analysis.
- Finally, we started to train the dataset with different models and then we evaluated the models based on accuracy score and plotted different performance metrics such as confusion matrix, classification report and ROC curve.

Literature Review:-

- A crime is an act that disrupts society's peace, such as a violation of human life or a governmental or a private organization/property, committed by a person or a group of individuals. This report summarizes crime predictions that have been conducted in Chicago city, using various Machine Learning methods applied to actual datasets to forecast and interpret crime data based on specific parameters.
- In recent years, computers have played an increasingly important role in virtually every area, including locating and tracking offenders. Since offenders may have a variety of traits and criminal careers, they can have unique assets.

Section 2 : System Design and Implementation

Algorithms considered / selected:-

- For the purpose of training and testing we have considered many different algorithms. Since this is a classification problem we have implemented many different classification algorithms.
- We have implemented Logistic Regression, XGBoost, Random Forest Classifier, Decision Tree Classifier, Naive Bayes Classifier and Support Vector Machine.
- Here we have considered Arrest as the target column since this categorical feature doesn't have high cardinality. Since the target column here is categorical we have used many different classification algorithms to classify the type of crime.
- All the algorithms implemented gave different performance on the dataset. The best accuracy was given by XGBoost and Random Forest Classifier. Both of them gave us an accuracy of 89% on both test and validation dataset.

Technologies and Tools used:-

- We used many different python libraries to perform Exploratory Data Analysis and training. We used pandas, sklearn, numpy, matplotlib and seaborn libraries.

- Also, we used a ‘folium’ library to visualize the number of criminal incidents based upon latitude and longitude on a Map. Also we plotted the community index, arrest and the location description of each criminal incident on a Map using this “folium” library.
- Also we used a datetime library to convert the given date in the dataset to Month, Day, Hour and Weekday.
- Then we used a categorical encoders library to convert the categorical variables into numerical variables with different techniques.
- We used a sklearn library to split the dataset into train, validation and test sets. Also we used sklearn library to load the algorithms for training, to evaluate the model and also to plot some performance metrics such as confusion matrix, classification report and ROC curve.
- Then we used the “make_pipeline” library of sklearn to construct a pipeline consisting of categorical encoder, simple imputer and the classifier/estimator used.

Section 3: Experiments / Proof of Concept Evaluation

Datasets used:-

- The Data Set is publicly available through the city of Chicago’s website. The information presented in this data set is quite comprehensive, including information about the date and time of the crime, location of the crime, type of crime, etc. For the purposes of this project, we will focus on the time of the crime and the type of crime.
- The type of crime is given a standardized set of codes called the Illinois Uniform Crime Reporting (IUCR) codes. Thus, each IUCR corresponds to a specific type of crime. The list of crime codes and corresponding crimes can also be found through the city of Chicago’s website.
- This dataset contains information regarding the criminal incidents that took place in Chicago from 2001 till present. The dataset is taken from Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
- The dataset has a mix of boolean, int, float and object features. The dataset is too large in shape (7312666, 22). Hence there are 7312666 rows and 22 columns in the dataset.
- There are some null values in the dataset but since the dataset is very large we dropped the null values. The resulting shape of the dataset after dropping null values is (6628274, 22).
- There aren’t any duplicate values in the dataset. Also we checked the value counts of each column of the dataset. Also we splitted the date column into time_hour and month columns.

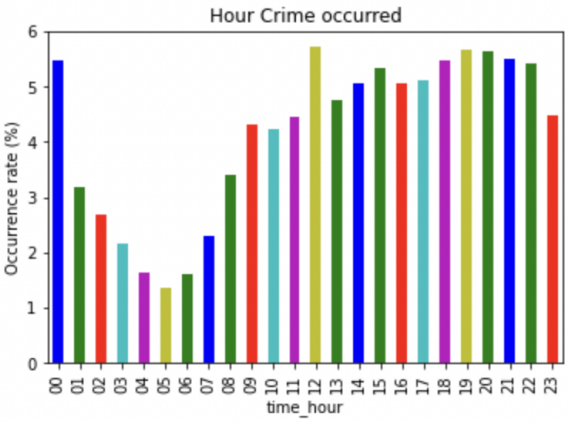
Methodology followed:-

- We splitted the dataset into X and y where y is the target feature i.e. Arrest column and X has rest all features except the target and ID column. The shape of X is (6628274, 22) and y is (6628274,).

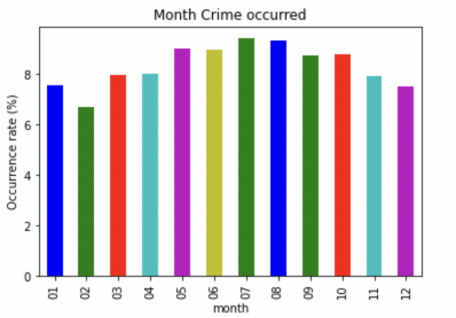
- Then we splitted the dataset into train-validation and test dataset i.e. X_trainval, X_test, y_trainval and y_test using sklearn's train_test_split where the input will be X and y keeping the train_size as 0.8 and test_size as 0.2. The shape of X_trainval is (5302619, 22), X_test is (1325655, 22), y_trainval is (5302619,) and y_test is (1325655,).
- Then we splitted the train-validation dataset i.e. X_trainval and y_trainval into train and validation dataset i.e. X_train, X_val, y_train and y_val using test_size as 0.2. The shape of X_train is (4242095, 22), y_train is (4242095,), X_val is (1060524, 22) and y_val is (1060524,).
- Then we calculated the accuracy score for the majority class baseline. In that case we calculated the accuracy of the mode of the target using the mode and then plotted the confusion matrix, classification report and ROC curve of the same.
- Next, we wrangled the X_train, X_val and X_test by splitting the "Date" column into "Month", "Day", "Hour" and "Weekday". Then we dropped the "Date" and "Updated On" columns from X. And then we started to train and evaluate the dataset using different classification algorithms.

Graphs:-

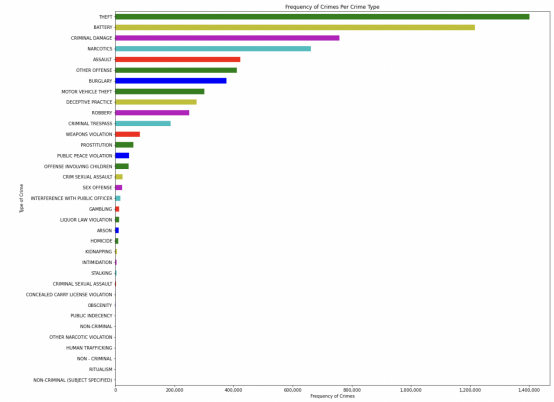
- After performing pre-processing of data we plotted the graphs to visualize the data better.
- First of all we plotted a graph i.e. a bar plot of Hour Crime Occurred which shows the percentage of occurrence of crime at a particular hour.



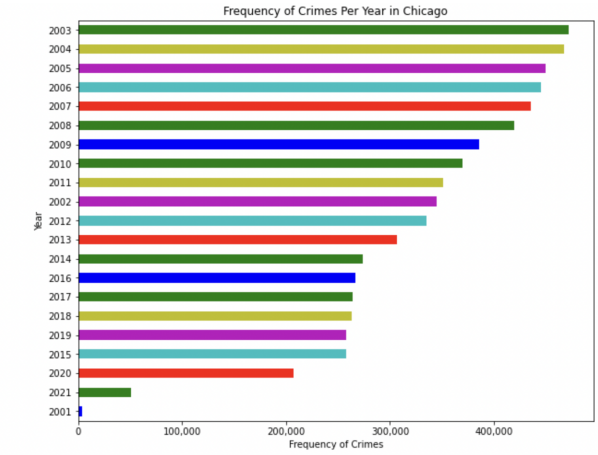
- Then we plotted a graph i.e. a bar plot of Month Crime Occurred which shows the percentage of occurrence of crime at a particular month.



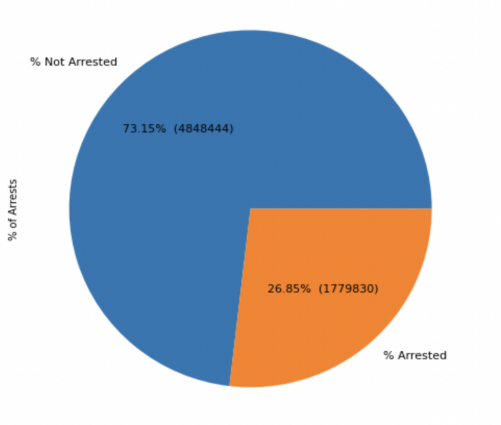
- Then we plotted a graph i.e. a horizontal bar plot of Frequency of Crimes Per Crime Type which shows the Frequency of Crimes for each particular crime type.



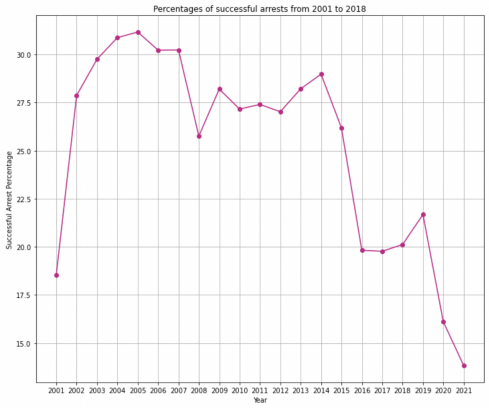
- Then we plotted a graph i.e. a horizontal bar plot of Frequency of Crimes per year in Chicago which shows the number of crimes that took place at a particular year.



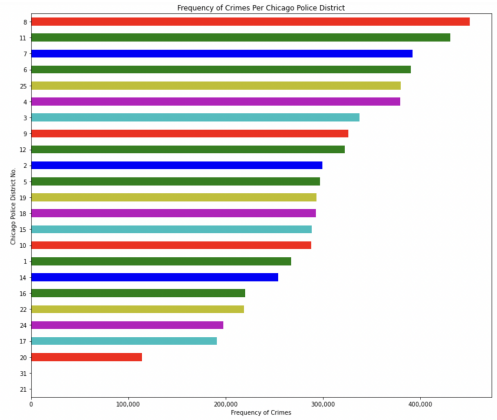
- After that we plotted a pie chart to visualize the percentage of arrests. For this we converted the Arrest values into Percentages.



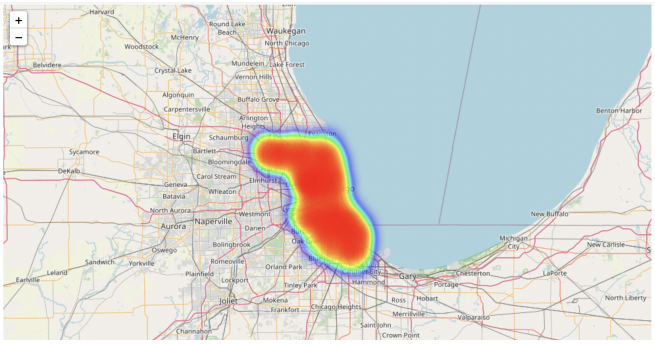
- After that we plotted a graph i.e. a line plot of Percentage of successful arrests from 2001 to 2021 which shows the successful arrest percentage for each year from 2001 to 2021.



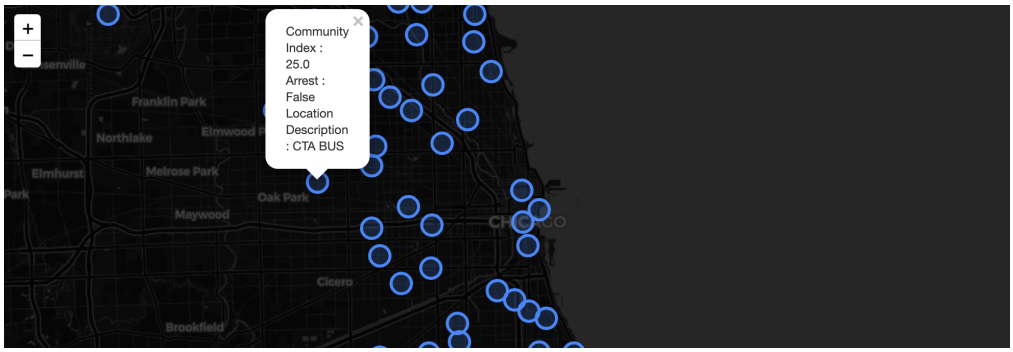
- Then we plotted a graph i.e. a horizontal bar plot of Frequency of Crimes per Chicago Police District which shows the number of crimes registered at a particular Chicago Police District No.



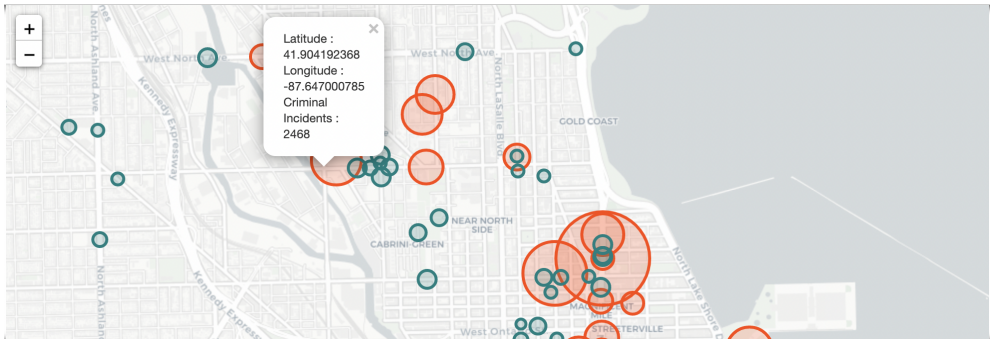
- Then we created a heatmap of crimes in Chicago March 2018 using a Folium library. We passed the mean of Latitude and mean of Longitude as the location parameter in the Folium map.



- Then we plotted a Map of Chicago city using Folium Library which shows the Community Index, Arrest and Location Description for each crime that took place in the city.



- Then we plotted a Map of Chicago city using the Folium library which shows the Latitude and Longitude at which the crime took place and Number of criminal incidents that took place at that location. Latitude and Longitude are the coordinates of the location at which the crime took place. The red circles in the Map are the places where there are a high number of criminal incidents i.e. the places which are highly prone to the criminal incidents. Whereas the green circles in the Map are the places where there are lower numbers of criminal incidents compared to the ones with red circles.



Analysis of Results:-

- We used multiple algorithms for the purpose of training and evaluating the dataset. We used different classification algorithms namely Logistic Regression, XGBoost, Random Forest Classifier, Decision Tree Classifier, Naive Bayes Classifier and Support Vector Machine.
- All the algorithms performed differently on the dataset but the best performance was given by XGBoost and Random Forest Classifier. Both of them gave us the accuracy of 89% in classifying the crime.
- We also used different performance metrics such as confusion matrix, classification report and ROC curve to evaluate the model.

Section 4: Discussion & Conclusions

Decisions made:-

- Initially we decided to perform Exploratory Data Analysis to understand the dataset better and we also decided to plot some graphs to visualize the dataset better.
- Then we decided to use multiple classification algorithms to train the dataset so that we can compare the performance of all the algorithms using several performance metrics.

Difficulties faced:-

- The difficulties we faced during the project was that since the dataset was very large it was time consuming to load the dataset and perform preprocessing on it.
- Also several algorithms such as Support Vector Machine, XGBoost and Random Forest Classifier took much time for training and performing predictions.

Things that worked well:-

- Since the problem definition was clearly understandable it was easy for us to perform the preprocessing and training on the dataset.
- Since there are already available libraries of python to plot the graphs and maps it was easy for us to visualize the dataset better.
- Also since all of us had better understanding of data mining and machine learning techniques it was easy for us to work on the project starting from Exploratory Data Analysis till model training and evaluating.

Things that didn’t work well:-

- Some algorithms took much time for training and predicting such as K-Nearest Neighbors so we weren't able to implement it.
- Also we tried out hyperparameter tuning using GridSearchCV and RandomizedSearchCV but they too were taking so long so we weren't able to implement that too.

Future Work:-

- For the future work we can do feature importances using different methods. We can also perform clustering on the dataset by making clusters of crimes that occurred during the same year, month, day or time. We can also try Deep Learning Neural Networks - Multilayer Perceptron to perform classification and evaluation on the dataset.

Conclusion:-

- It has become much easier to find relationships and patterns within different datasets using machine learning technology. The work in this project is primarily focused on predicting the type of crime that may occur based on the location where it happened.
- We developed a model using the principle of machine learning and a dataset that had undergone data cleaning and transformation. The model gave us an 89% accuracy rate in predicting the form of crime using XGBoost and Random Forest Classifier. The interpretation of a dataset is aided by data visualization.
- We have plotted different graphs such as bar, pie, line and folium maps which helped us in understanding the Chicago Crime Dataset in a better way.

Section 5: Project Plan / Task Distribution

Who was assigned to what task:-

- We are a team of 4 people. Parshv was assigned the Exploratory Data Analysis and Model Training / Fitting part. Het was assigned the data preprocessing part. Monil was assigned the data wrangling part. Priyansh was assigned to calculate the accuracy score for the majority class baseline part.

Who ended up doing what task:-

- Everyone in the team ended up doing the task that was assigned to them. But Parshv helped out in all the tasks. So, Parshv did the Exploratory Data Analysis and Model Training / Fitting part. Parshv and Het both did the data preprocessing part. Parshv and Monil both did the data wrangling part. Parshv and Priyansh both did the calculation of accuracy score for the majority class baseline part.