Unsupervised Learning on Chicago Crime Dataset

Name:	**Yashika Patil**
Registration No./Roll No.:	19339
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	January 19, 2022
Date of Submission:	April 24, 2022

1 Introduction

The Chicago Police department keeps an up-to-date record of crimes taking place from 2001 till date. This data can be used for predictive analysis to gain a better understanding of the crimes taking place in Chicago. The dataset is available on Chicago Data Portal which has about 7 million data points and 22 columns [1]. In this project, 100k data points with 22 categorical features namely ID, Case number, Date, Block, IUCR, Description, Primary type, Location description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location were considered.

So far, the Chicago crime data has been used for supervised learning to predict crime, arrest and data analysis. To the best of our knowledge, no previous attempt has been made to perform unsupervised learning on the crime data. This project is an attempt to identify natural clusters in the data and features contributing to the formation of clusters. The dataset being huge and as only few features showed correlations, preprocessing and feature selection was an important task here to reduce the complexity of the data. As the data is categorical in nature, a major difficulty was to find an algorithm that takes categorical variables as input. For this, the alternative of k means: K-Modes was explored[2]. The challenge here was to specify the number of clusters(k). By data exploration and intuition, we assumed the number of clusters as 31 which is the number of unique categories in the feature 'Primary Type'. Another model explored was Spectral clustering which doesn't require the number of clusters to be specified by the user. The Normalised Mutual Information (NMI) score for both these models were calculated. Based on this score, Spectral Clustering was seen to perform better on our dataset.

2 Methods

GitHub link: https://github.com/yashikapatil27/Unsupervised-Learning-on-Chicago-Crime-Dataset

2.1 Exploratory data analysis (EDA)

The features 'Latitude' and 'Longitude' were dropped first as they were redundant features, and were already included in the 'Location' feature. This reduced the number of columns to 21. 'Date' was converted to pandas datetime format and the index was set to be the date. Then the dataset was visualized by plotting various graphs to see the trends and patterns present[3].

(i) Heatmap: It shows the correlation between all the features present.
(ii) Number of crimes per month: It can be inferred that the number of crimes are decreasing with time and the city is becoming safer.

(iii) Number of crimes by type: There are a total of 31 primary types in the data out of which 'Theft' is the most occurring crime in the city.
(iv) Primary type Vs. Year: The plots show trends of each primary type with passing years.
(v) Number of crimes by day: 'Friday' records the highest number of crimes that. Hence it can be considered as the most unsafe day of the week in the city.
(vi) Number of crimes by month: August' and September' records the highest number of crimes that occur.
(vii) Number of crimes by location: Most of the crimes occur in the streets of Chicago which is followed by residence.

2.2 Data Preprocessing

The index that was previously set data for EDA was reset. Features like 'ID', 'Beat', 'District', 'Ward', 'Community Area', 'Year' were converted from float to object data type. The dataset had no null or missing values. The dataset consists of categorical variables. Many models require numerical values as input. So, One Hot Encoding is performed. One hot encoding using get dummies takes the categorical data and converts it into binary variables with a finite set of label values.

2.3 Feature Selection

Entropy is the unpredictability of the data. High entropy indicates that the data is scattered while low entropy means that nearly all data is the same. We compute entropy for all the features and then drop the ones whose entropy is too high. This results in selecting IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, Year, Updated On, as the most relevant features.

2.4 Clustering

2.4.1 K-Modes Clustering

This method is chosen as it clusters categorical data and deals efficiently with large amount of data. It generates clusters on the basis of dissimilarity. The points with less dissimilarity are clustered together. For cluster analysis, the value of k is set to 31 and the algorithm is run for 100,000 data points, 5 iterations, after which 31 clusters are obtained. The column that associates each data point to the corresponding cluster label is then added to the dataframe. The obtained clusters are then visualized through pie charts obtained for every Primary Type.

2.4.2 Spectral Clustering

This method takes similarity matrix as input to generate clusters. First dataset is splitted into train and test. The algorithm is run for 25,000 data points. The dataframe containing one hot encoded variables is passed to find matrix of cosine similarity between the features which is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.The spectral clustering model is then trained to predict the clusters. The number of clusters obtained is 8, with labels from 0 to 7. A column containing these labels is added to the dataframe which associates each datapoint to its corresponding cluster label. The clusters are then visualized through the pie charts obtained for 'Theft', 'Battery', 'Criminal Damage', 'Narcotics', 'Assault', 'Other Offense', 'Burglary' and 'Motor Vehicle Theft'.

3 Experimental Analysis

The evaluation criteria for our model is Normalised Mutual Information (NMI). It tells us the reduction in entropy of predicted cluster labels when actual cluster labels are specified. NMI score is normalized. So, it can be used to compare performance of different clustering models which form different numbers of clusters [4]. On calculating the NMI of predicted clusters and comparing it with all the features,

the feature “Primary Type” gave the highest NMI score. So, we considered “Primary Type” as the actual cluster label to compare NMI scores of the two models we used: K-modes and Spectral Clustering.

The NMI score for K-Modes clustering was 0.5766669383756399. The NMI score for Spectral clustering was 0.70319837876003. It can be concluded that for our dataset Spectral clustering performs better than K-Modes clustering.

4 Discussions

Biggest limitation of the project was the huge dataset with a high number of categorical and numerical features and feature variables which made it difficult to implement various models and algorithms. The models which finally were usable had high space and time complexity, thus the dataset was engineered and the results were extrapolated. The project was unique in itself as it required hands-on skills in data cleaning, manipulation and feature engineering more than model selection and improving accuracy. The future scope of this project looks good as it would help us learn more about clustering of categorical variables which is yet to be explored and discovered unlike supervised learning methods. This work can be extended as we can explore and self implement more algorithms and also, the current proposed model can be replicated for other datasets.

References

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[2] https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

[3] https://medium.com/@namanjain2050/hands-on-machine-learning-with-chicago-crime-data-3657b713d62c

[4] https://towardsdatascience.com/evaluation-metrics-for-clustering-models-5dde821dd6cd