Analysis of Crime in Denver

Richard Terrile
University of Colorado Boulder
Boulder, United States
rite5632@colorado.edu

Sai Divya Sivani Pragadaraju
University of Colorado Boulder
sapr6102@colorado.edu

ACM Reference Format:
Richard Terrile and Sai Divya Sivani Pragadaraju. 2022. Analysis of Crime in Denver. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION/MOTIVATION

This project titled " Analysis of Crime in Denver " is an attempt to understand and analyze the nature of crimes that took place in and around the city of Denver. The rise in the number of crimes with every passing day is one of the major concerns among the citizens. It is often believed that crime victims often suffer from psychological and social injuries for a long time; sometimes forever even after their physical injuries have been healed. We believe that working on this dataset will give vital insights into various types of crimes, locations and timings which in turn can be used to take preventive measures and support better law enforcement by the government to protect common people of the county from being victims of heinous crimes.

For this project we would like make use of various Data Mining techniques to answer the following questions :
1) What types of crimes are most common in Denver? 2) What is the most common time period(be it time as well as month) for a specific type of crime to be committed?
3) Which locality in Denver has the highest crime rate ?
4) How has the nature/type of crime evolved over the past five years? Like are there any types of crime that are no longer committed?

2 LITERATURE SURVEY

The following two research papers use historical Denver Crime data to analyse their own unique questions.

2.1 "Optimizing Denver Patrol Routes"

The Denver Metro area contain many different policing zones and the distribution of crime is not equal throughout all zones. Optimizing patrol routes is done to achieve more efficient and effective policing. The motivation behind this idea is that patrolling, "in particular foot patrolling," has been suggested to reduce crimes and disorders in "hotspots" and at "hot times." This project by University of Colorado Denver analyzed historical crime data to identify locations that would benefit the most from patrol routes. Then by applying graph theory, they modeled each location as a vertex in a graph and determined the optimized patrol routes by solving the classical "Traveling Salesman Problem."

To model crimes on a graph, K-means clustering was used. This algorithm was decided upon because it provides geometric centroids which are simple to plot as vertices on a graph.

After the centroids were found, Miller-Tucker-Zemlin formulation was used to find the shortest path traveling through all points. In other words, the formulation was used to solve a "Traveling Salesman Problem" revolving around these centroids. The group used geometric distance between points, but notes that if they had more time, a more effective strategy would have taken the more accurate driving/walking times from Google Maps.

In the future, the team notes that this idea could be applied to construct precincts in better locations and to optimize capacity at each station. Further work would involve programs to automatically generate patrol routes as well as splitting day and night crimes.

More details are available here http://math.ucdenver.edu/~sborgwardt/wiki/index.php/Optimizing_the_patrolling_route.[1]

2.2 "A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining"

The goal of this project was to predict and highlight trends related to crime occurrence to better support law enforcement agencies and to help create effective crime prevention measures and policies.

The methods used were to first apply statistical analysis as well as data visualization methods. Followed by implementing a variety of classification algorithms, some of which are Random Forest, Decision Tree, AdaBoost Classifier, K-Neighbors Classifiers, and more. Ensemble Models were used to classify 15 different classes of crimes with the outcomes captured by train-test split, and k-fold cross-validation.

To evaluate the performance "flawlessly," the author utilized "precision, recall, F1-score, Mean Squared Error, ROC curve, and paired-T-test." Aside from AdaBoost Classifier, many of the algorithms reached adequate accuracy, however Ensemble Model 4 had the best results for every evaluation basis.

The study proposes that its results could be most useful in raising awareness of crime occurrence hotspots and to aid security agencies in predicting violent crime locations within a particular time.

More information as well as the research paper are available from https://www.researchgate.net/publication/338500247_A_Comparative_

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Study_on_Crime_in_Denver_City_Based_on_Machine_Learning_and_Data_Mining.[2]

3 PROPOSED WORK

The data set contains many values such as “incident_id” and "offense_id” which will not be used when mining the data. These unnecessary attributes will be removed for more efficient analysis. A few entries are written with errors or improper encoding which prevents easy import into Python. These entries will need to be assessed and cleaned as well.

Dates are logged in several attributes such as “FIRST_OCCURRENCE_DATE,” “LAST_OCCURRENCE_DATE,” and “REPORTED_DATE.” For general classification we will only use one of these date attributes, namely the “FIRST_OCCURRENCE_DATE” attribute. Date and time are also stored together as one attribute in the form of a string. In order to get better results from mining, these will be split into two separate attributes for “Date” and “Time.”

3.1 Algorithms

We will look at frequent item sets within our data. The goal is to find frequent combinations of attributes to reveal correlations which could be useful in predicting crimes if the given item set occurs periodically. This method was not explored in the prior work using this data.

- Apriori Algorithm

Classification techniques will be utilized to gain information regarding various categories of crimes.

- Decision Tree
- Random Forest
- K-Neighbors
- Binary Classification ('is_traffic' attribute)

We will utilize clustering algorithms to find various patterns within the data. The focus is clustering on location and time. With those clusters we will look at crime "hotspots" and "hot times," when and where, crime is most frequent. Clustering will also be used to analyze the types of crimes most prevalent in each cluster. The literature review showed usage of clustering algorithms, however they only found location hotspots. We plan to explore clusters around time as well as analyzing the entries present within each cluster.

- K-means clustering algorithm
- Gaussian Mixture Model algorithm
- Density-based clustering (DBSCAN algorithm)

4 DATA SET

For this project we are making use of The “ Denver crime dataset” which contains criminal offenses recorded in Denver county from the year 2017 to 2022. This dataset is taken from the official government website of Denver and was populated from the National Incident Based Reporting system, which incorporates records of all victims of person crimes and all crimes within an occurrence. Furthermore, This dataset has 376813 instances and 20 attributes such as Offense Code , Occurrence date (first and last occurrence) , Address ,Location (Latitude and Longitude or X and Y) , Victim Count etc along with 16 types/classes of offenses.

The link to this dataset is as follows :
https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime?ref=hackernoon.com.

5 EVALUATION METHODS

5.1 Data Pre-Processing

Data Cleaning and Data Reduction:
In this dataset we have many missing values, particularly in two attributes in the Denver crime dataset: incident_address and last_occurance_date. We plan to clean them. The dataset also contains many values such as ‘incident_id” and “offense_id” which we will not use. These along with other unnecessary attribute types will be removed for easier analysis. We also plan to employ the dimensionality reduction procedure by utilizing attribute subset selection.

Data Integration:
For this analysis we strongly believe that reported date and Time is crucial. In this dataset, Date and time are stored together as a string.so,We are planning to convert crime reported date attribute to 4 new attributes such as year, month, day, and hour.

Data Conversion:
Further we would like to convert our object type data into categorical data so that they can be easily fed into our models for better analysis.

Data Normalization:
We would also like to use min-max normalization technique so that all the data falls in the range of 0-1 which would be easier to process.

Evaluation metrics:
While making use of Apriori algorithm to mine frequent itemsets for boolean association rules; for instance which two types of crimes have occurred at the same time, we would like use **support** and **confidence** as evaluation measures. Support tells us the frequency of association while confidence tells us about the strength of association rules.
For algorithms like decision tree we plan to calculate the entropy, information gain and check accuracy and MSE to decide which decision tree is perfect for our dataset.
The k- nearest neighbors, which is a supervised algorithm used for both regression and classification problems would help us in classifying and segregating similar kinds of crimes. To evalaute it we would like to check accuracy and generate a confusion matrix for better insights.
Further, we would also like to use clustering algorithms and look at similarity scores between clusters for similar locations and similar crime types using Cosine Similarity, and Euclidean Distance.

6 TOOLS

For this project we will code in PYTHON and we will be making use of some of the prominent libraries such as Numpy, Pandas, Matplot, Geoplot, scipy, tensorflow which are believed to be quite efficient

in yielding better results for our analysis.
we will be coding in environments such as Jupyter Notebook, coding.csel.io, Google Collab Notebook

7 MILESTONES

7.1 Milestone 1: Processing and Setup

Expected Completion: July 21

- Set up collaborative coding environment and upload data
- Clean data set (fix encoding, remove dimensions)
- Process data (split up date and time attribute)
- Preliminary data queries for better understanding

7.2 Milestone 2: Frequent Item Sets

Expected Completion: July 26

- 1: Run Apriori Algorithm on data
- 2: Note results and check if they make sense
- 3: Note interesting findings that could be looked into further with other algorithms

7.3 Milestone 3: Classification

Expected Completion: July 31

- 1: Run algorithms on data set
- 2: Analyse results
- 3: Note interesting patterns

7.4 Milestone 4: Clustering

Expected Completion: August 5

- 1: Run algorithms and collect results
- 2: Look at results
- 3: Note interesting features

7.5 Milestone 5: Analysis

Expected Completion: August 8

- Evaluate Apriori
- Evaluate Classifications
- Evaluate Clustering
- Combine information for analysis and begin formulating results

REFERENCES

[1] Dongdong Lu. 2019. *Optimizing Denver Patrol Routes.* http://math.ucdenver.edu/~sborgwardt/wiki/index.php/Optimizing_the_patrolling_route

[2] Md Ratul. 2020. A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining. (01 2020).