Crime Patrol: Working With Crime Data

Amit Rokade
ar5618@rit.edu

Tejas Arya
ta2763@rit.edu

Anmol Jaising
akj9051@rit.edu

ABSTRACT

Today, the crime rate is more than ever across the country. To ensure the safety of each citizen, the policing agencies strive to respond to the complaints as soon as a crime is reported. The details of each complaint are filled, noted, and preserved. Increasing crime rates have been a problem. Traditional crime prevention methods have proven to be useful. With the help and application of data mining tools, insights could be generated that help law enforcement agencies even further reduce, if not completely solve the problem of increasing crimes. Applications of analysis on crime data can be in crime forensics, prevention and future forecasting. All of these applications in tandem will bring around more awareness and knowledge among civilians and policing forces, thereby acting as a means to reduce crime. Through this project, a complete data cleaning and mining analysis on crime data is presented along with results found from mining the data. Evaluation metrics such as precision, recall and f-measure are used to present the effectiveness of various classification algorithms. Challenges faced are discussed along with future work to be done in the area of a continuous data-flow environment to incorporate continuous integration and continuous delivery of useful insights from unclean data.

Keywords- **Crime**, **Data**, **Cleaning**, **Algorithms**, **Data Mining**, **Data Preparation**, **Data Cleaning**.

1. INTRODUCTION

It happens frequently that by the time a complaint has been filed, another crime has already been committed. This jeopardizes the safety of civilians because of lack of prior knowledge of where the crime occurs. With the help of information backed by statistics and data mining, preventive measures could be applied by the police that might help decrease the crime rate. Other application of data mining on crime data would be in forensics analysis to figure out the exact time of occurrence of crime or age of the victim or suspect. To apply statistical studies on crime data, the data collected needs to be in a specific and clean format. Having a data cleaning and preparation pipeline and the complaint records all incorporated in one place will allow systematic access to it. Further, this enables better data analysis component in a dynamic data flow environment

In this project, crime datasets are taken from New York City (NYC) and Los Angeles (LA), where each record is complaint data for a crime.

Crime Patrol is intended to provide a data cleaning and preparation pipeline for crime data collected for NYC [5] and LA [4] and provide insights into that help police force to take preventive measures. This would in turn outline the importance of cleaned and prepared crime data.

Raw and cleaned data is visualized to show the impact of data cleaning.

Data visualization can also be used to provide better insights into the crime trend of various cities which can be further used to mitigate crime rates and ensure well being of civilians and properties.

This project is split into 5 parts, namely: Data collection, Data assembly, Data cleaning, Data mining, and Data visualization.

Data sets are described in section 2 followed. Motivations for this project are discussed in Section 3. Section 4 discusses Methodology and Implementation. Section 5 provides insights into data visualization. Results and challenges faced are discussed in section 6. Future work is covered in section 7. Finally, the project is concluded in section 8.

2. DATASET DESCRIPTION

2.1 New York police department Complaint data

NYPD complaint data [5] gives a complete overview on each crime reported between 2006 and 2017. It contains some important attributes such as the complaint type, place where it occurred, date, suspect information, victim information, geo-location etc. This dataset contains 7 million complaints with 35 attributes.

2.2 Los Angeles police department crime data

LAPD Crime Data [4] contains the crimes records that occurred in Los Angeles between 2010 and 2019 [4]. One of the key reasons this data was chosen was it was transcribed from paper documents into a database and thus has added noise during the process of converting the data. Default values for some attributes were used for geo-location during this paper to database transformation where the field entry was blank in the original format. This dataset contains 2 million crime incidents with 28 attributes.

3. MOTIVATION

Victims of any type of crime are prone to face Post Traumatic Stress Disorder in social situations. With the increase in population the crime rate increases as well. The goal for

this project is to provide a standard procedure to convert raw crime data to a high quality data that can be used to provide better insights into crime trends of specified cities. The aim is to aid in prevention of crime using insights from crimes that have already occurred. This in turn would enable a safer environment for civilians and provide them a sense of security.

4. METHODOLOGY AND IMPLEMENTATION

4.1 Data Collection

Data has been collected from two different sources in the Comma Separated Value (CSV) format.

It was found that both data sets had similar features, although their names were inconsistent.

Some similar attributes with different names were as follows:

1. **LAPD**: Date_Rptd, TIME_OCC, AREA, AREA_NAME, Crm_Cd, Crm_Cd_Desc

2. **NYPD**: attributes like RPT_DT, CMPLNT_FR_TM, BORO_NM, KY_CD, OFNS_DESC

Before merging the dataset, some amount of preprocessing is performed to find relevant attributes and making the attribute names consistent.

Merging the two datasets, based on the common attributes, we get a dataset containing:

· **DATE_REPORTED** refers to the date at which the complaint was reported.

· **DATE_OCCURRED** indicates the date when the crime took place.

· **TIME_OCCURRED** refers to the time of occurrence of the crime.

· **PATROL_DIVISION** gives the code of the patrol borough where the complaint occurred.

· **AREA_NAME** feeds us with the name of the patrolling borough where the crime/complaint occurred.

· **CRIME_CODE** indicates the crime code.

· **CRIME_DESCRIPTION** describes the crime based on the crime code.

· **VICTIM_AGE** refers to the age group of the victim.

· **VICTIM_SEX** refers to the sex of the victim.

· **VICTIM_RACE** refers to the victim's race.

· **PREMISE** refers to type of dwelling or public location where the crime occurred

· **CITY** is the city where the crime occurred.

· **LAT** is the latitude where the complaint/crime took place.

· **LON** is the longitude where the complaint/crime took place.

4.2 Data Cleaning and Preparation

Several data cleaning and preparation techniques have been applied. Some of them are: Replace null values with mode, Handling outliers using Tukey IQR Method, De-duplication.

Moreover, the Dates and Times in both datasets were inconsistent with each other as well as there were inconsistencies within the datasets for same attribute. To make the values within the merged dataset consistent dates were converted to MM/DD/YYYY format and time was converted to hh:mm:ss format.

For LA Crime Data, VICTIM_AGE were in intervals. However, for NY Crime Data, VICTIM_AGE values were absolute and some innvalid values were present. For example, some records where victim's age was 0 were found. This is clearly not possible and such cases were handled using Tukey IQR Outlier detection method. Unknown values were replaced by mode. VICTIM_AGE was then converted to a categorical variable where victims' ages were converted to bins.

Also, percentages of missing or empty values is obtained for each attribute. The attributes which have more than 50% missing or empty values are discarded.

CRIME_CODE and PATROL_DIVISION were inherently in float format. This were converted to string to ensure they are not interpreted as decimal values.

VICTIM_SEX contained defaults as well as unknown values such as 'D' and 'E' respectively. Both these values were replaced by the mode of the attribute.

VICTIM_RACE was entirely mapped from letters to words using the column description from the dataset sources .

For data cleaning, an effective python library, pandas [2] is used.

4.3 Data Storage

After the cleaning had been performed, data was stored on MySQL Workbench [1]. Data mining was performed on WEKA [6] by pulling in the data from MySQL Workbench [1].

Following is a screenshot of data:

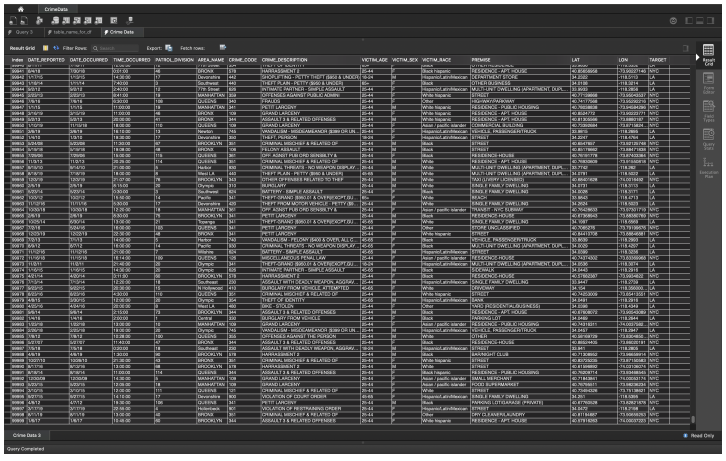

Figure 1: **Data stored in MySQL Workbench**

4.4 Data Mining

WEKA [6] tool is used for data mining component of the project. Three classification algorithms were used to clas-

sify the targets. The classification algorithms used were J48 Decision Tree, BayesNet and NaiveBayes.

Two problem statements were taken for the data mining component which are:

1. To predict VICTIM_AGE in case of a crime. This can be used by officials in analysis where no information about the victim could be found

2. To predict the AREA_NAME to find better insights of where a crime might occur based on given factors and thus come up with preventive measures for high crime areas.

For each classifier, we split the dataset on a 70:30 ratio to create a training and test set.

The J48 classifier is based on the decision tree model, the training model learns the relations between the attributes and the target class. The trained model is used to predict the target variable for test data.

The BayesNet classifier is used to classify a target variable based on the conditional probability between the attributes and the class variable.

The Naive Bayes algorithm, like BayesNet classifier, classifies the class variable assuming an independence assumption between the attributes.

5. DATA VISUALIZATION

To graphically compare the raw data and cleaned data, box-plots, histograms, bubble plots, scatter-plots as well as a size-based chart were used.

Heat Map was generated for data, visualizing locations with respective crime description markers that occurred at the locations.

WEKA [6] and Tableau[3] were used for visualization. Heat Maps were generated using Google API.

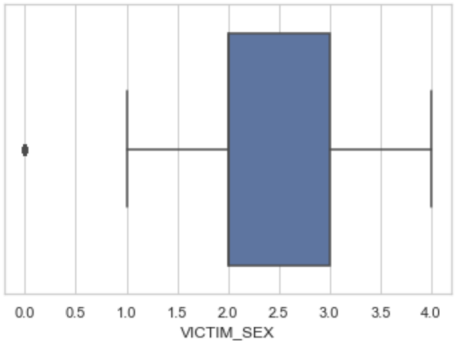

Figure 2: **VICTIM_SEX box-plot for raw data**

Box-plot of VICTIM_SEX in raw data shows outliers for ages.

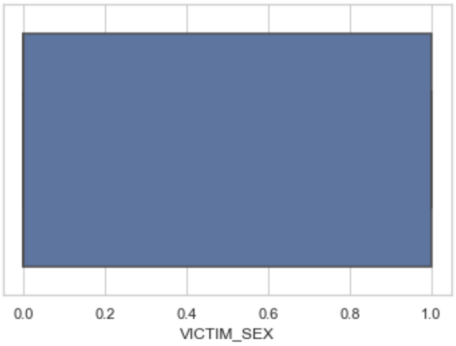

Figure 3: **VICTIM_SEX box-plot for clean data**

After outlier removal, VICTIM_SEX can be seen to be evenly distributed

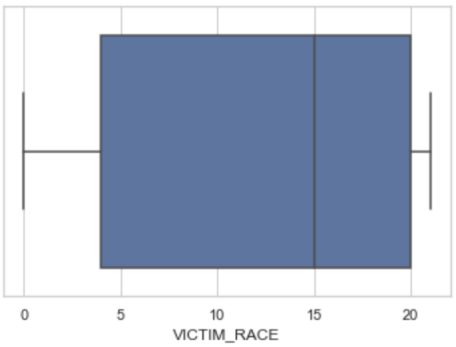

Figure 4: **VICTIM_RACE box-plot for clean data**

VICTIM_RACE can be seen to have no outliers after cleaning data.

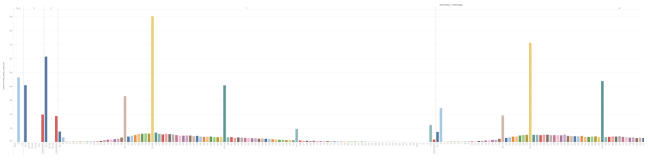

Figure 5: **VICTIM_AGE vs VICTIM_SEX for raw data**

Each color is a different age group or age for five sex groups. Since the format of representing age is different in LA (absolute) and NY (interval + absolute) , the charts automatically get laid side by side showing similar highs and lows.

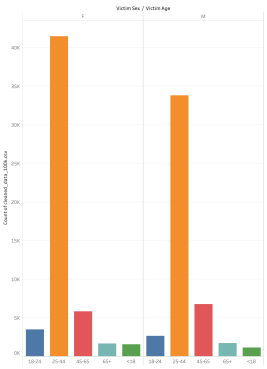

Figure 6: **VICTIM_AGE vs VICTIM_SEX for clean data**

After cleaning the data, sex groups were reduced to M and F. Absolute + interval values for age were converted to intervals for age groups. The charts look much cleaner and understandable after cleaning and it can be seen that most people belong to the age group 25-44.

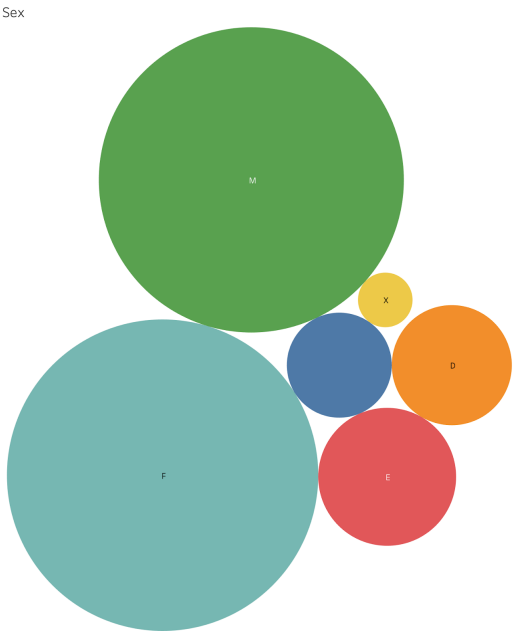

Figure 7: **Bubble cloud depicting VICTIM_SEX in raw Data**

Bubble cloud depicting different sex with their respective numbers represented by the size of bubbles.

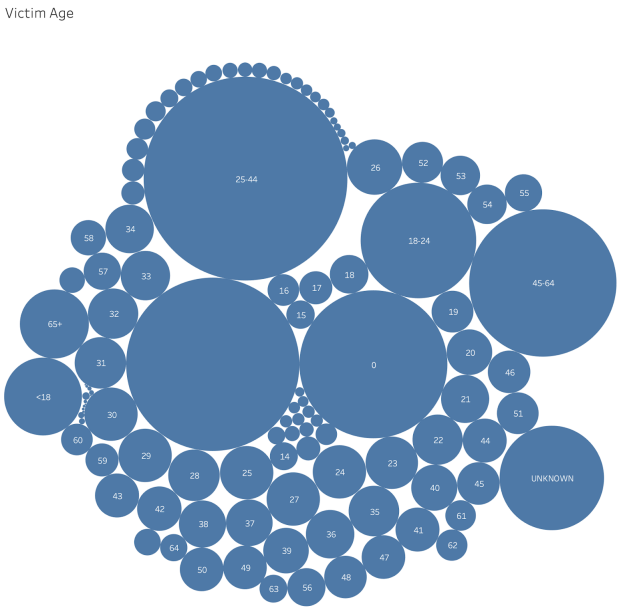

Figure 8: **Bubble cloud depicting VICTIM_AGE in raw Data**

Bubble cloud depicting VICTIM_AGE clusters before cleaning data. Size of bubble represents the number of specific cluster in the group.

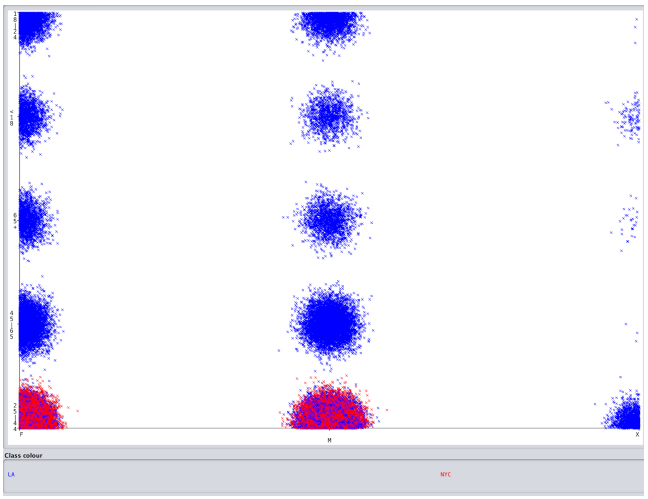

Figure 9: **Scatterplot for VICTIM_AGE vs VICTIM_SEX**

Clusters are laid out vertically with Y axis showing the age ranges [18-24 , <18 , 65+ , 45-65, 25-44]. X axis separates the sex Male (M) and Female (F). A third dimension is also used for visualization which separates LA and NYC data with color.

Figure 10: **TIME_OCCURRED vs AREA_NAME for raw Data**

Inconsistency is seen when TIME_OCCURRED is mapped against AREA_NAME for the raw data which makes it very difficult to infer any information from this data. This is due to the inconsistency in time recording formats.

Figure 11: **TIME_OCCURRED vs AREA_NAME for clean Data**

The visualization is understandable now , color shows the frequency and it can be seen that the most number of complaints were received for Brooklyn Area with 900 crimes occurring around noon.

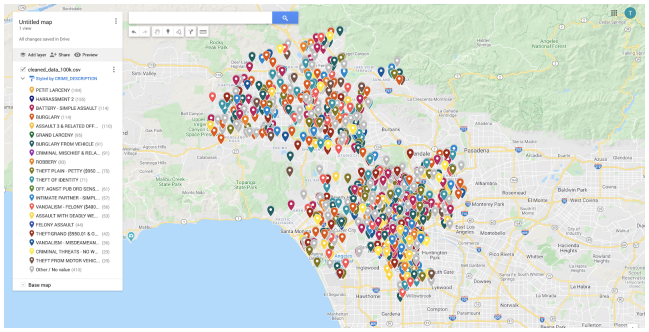

Figure 12: **Location of crimes reports in LA**

Locations in LA are marked. Location markers are color-coded with their respective crime descriptions in the legend.

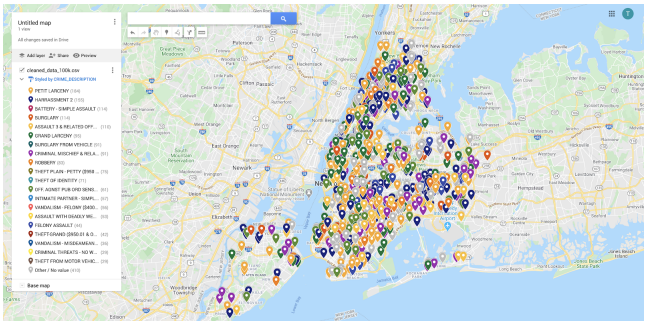

Figure 13: **Locations of crime occurring in New York** newline
Locations in NY are marked. Location markers are color-coded with their respective crime descriptions in the legend.

6. RESULTS AND CHALLENGES

6.1 Results

The results of classification are as follows:

Data	Algorithm	Target	Accuracy	Precision	Recall	F-Measure
100K_Cleaned	J48	AGE	**75.81%**	0.758	0.997	0.861
100K_Cleaned	BayesNet	AGE	62.40%	0.814	0.663	0.73
100K_Cleaned	NaiveBayes	AGE	60.74%	0.819	0.662	0.732
100K_Cleaned	J48	AREA_NAME	**76.82%**	0.836	0.886	0.860
100K_Cleaned	BayesNet	AREA_NAME	64.98%	0.643	0.565	0.601
100K_Cleaned	NaiveBayes	AREA_NAME	**73.21%**	0.744	0.721	0.73

Table 1: **Cleaned Data Result**

It was observed that J48 Decision Tree performed the best in comparison to BayesNet and Naive Bayes classification algorithms.

6.2 Challenges

1. One of the key challenges was to make modifications to the changes such that it can be merged. This was handled by applying basic pre-processing such that data could be merged on similar attribute names.

2. Another challenge was to bring consistency to dates and times due to being in different formats.

3. Another important thing was class imbalance. The instances for each city were highly imbalanced with LAPD crime data containing 2 million records and another containing 7 million records. This was solved by creating oversampled and undersampled data sets by applying class balancing techniques.

4. Lastly, handling such vast amount of data took a lot of time for training. Several times, the training crashed due to not enough heap memory being available. This was solved by reducing the size of data to 100,000 with equal number of records from LAPD crime [4] data and NYPD crime data [5].

7. FUTURE WORK

Future work would involve building a data cleaning and preparation framework for crime data that handles inconsistencies in features involving ages, dates, and times data. This framework would also incorporate outlier handling techniques along with class balancing solutions to maintain the veracity of model generation.

Along with this, another task would be to scale the data cleaning, preparation, and data mining to handle vast amount of data. This might involve use of better system configurations.

8. CONCLUSION

3 algorithms J48 Decision Tree, Bayes Net and Naive Bayes were used to classify most targeted victim age with J48 showing a 74% accuracy amongst 5 classes. Similarly, these algorithms were also used to find out the most targeted area. Along with data cleaning, algorithm performance was enchanced due to balancing the target classes which tackled the dataset imbalance problem and thus mitigated the bias the dataset had. Not all attributes were chosen to feed into the model to avoid making the model learn on non - related attributes. For example, the attributes latitude and longitude were not fed into the machine learning models which could lead to over fitting as models could directly learn the area based on location. Clustering was done along with visualization to infer important information like, Brooklyn was found to be the most dangerous area at around 12 o 'clock with most number of cases. Finally, the analysis done on the data can be used to help provide security in an efficient manner at a particular place and/or at a particular time.

9. REFERENCES

[1] *MySQL*. `https://www.mysql.com/`.

[2] *pandas*. `https://pandas.pydata.org/`.

[3] *Tableau*. `https://www.tableau.com/`.

[4] L. A. P. D. 2020. Crime data from 2010 to 2019: Los angeles - open data portal., June 2020. https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z.

[5] N. Y. P. D. 2020. Nypd complaint data historic: Nyc open data., May 2020. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.

[6] E. Frank, M. A. Hall, and I. H. Witten, 2016. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition, 2016.