MCM Practicum

Name: Amshumann Singh
Student ID: 18210185
Email address: amshumann.singh9@mail.dcu.ie
Programme :MSc in Computing (Data Analytics)
Practicum Title: Machine Learning Approach to Crime Prediction and Identification of Hotspots
Supervisor: Dr. Andrew McCarren

Disclaimer:
A report submitted to Dublin City University, School of Computing.
I understand that the University regards breaches of academic integrity and plagiarism as grave and serious.

I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken fromthe work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy (available at:http://www.dcu.ie/registry/examinations/index.shtml)

Name: Amshumann Singh Date: 17-August-2020

Machine Learning Approach to Crime Prediction and Identification of Hotspots

Amshumann Singh
Student Number: 18210185
Mail: amshumann.singh9@mail.dcu.ie
Sup: Dr. Andrew McCarren
17/08/2020

***Abstract*— Crime prediction is becoming a fast-growing area of research in the field of data science due to increase in the availability of crime data and adoption of data driven approaches by police departments all around the world. This practicum aims to find spatial and temporal hotspots along with predicting type of crime using association rule mining and machine learning classification models by using Denver crime data along with GDELT data which is a database for news events from all over the world. The model is able to identify criminal hotspots and provides a basic structure for high level crime prediction models. The results from this practicum can help the police force improve their response time to incidents, allows them to be better prepared while handling situations and help recognize areas of risk.**

Keywords— crime prediction, hotspots, association rule mining, crime classification

I. INTRODUCTION

Crime is a persistent problem all over the world that affects quality of life and economic growth [1,4,5]. Crime negatively affects the image of the associated localities resulting in residents relocating and further discourages business growth. It also burdens the taxpayers due to increased need for the police, judiciary and prison systems [5].

By studying how world economies and societies are structured, it is reasonable to assume that crime will persist and most likely increase in the near future. Another observation is that crime occurs more frequently in urban areas rather than rural. This is due to higher population density and larger economic disparity in urban localities. Crime statistics also have a correlation with factors such as unemployment and income [2]. This makes it essential to study incidents, identify solutions to reduce and prevent crime in urban surroundings.

Crime has been analyzed from a behavioral, psychological and sociological perspective since the beginning of the 20th century. The studies have focused on the relation of crime behavior with their environment and actions [1]. Based on data available, credible theories such as 'centrography', 'journey to crime' and 'routine activity theory' have also been established [8]. But now, with the growth of data and information available over the last decade, there are new opportunities and possibilities to advance the fields of crime analysis [3,14]. Nowadays, information and data are critical tools in law enforcement [10]. As computer-based systems are increasingly used in tracking crimes, data analysts have started helping uniformed officers in solving them [13].

Data driven approaches to crime-prevention in recent years has made it easier to observe large-scale trends and patterns. But specific crime patterns are still not clearly visible by these approaches. Thus, crime analytics is still fundamentally a difficult manual job for analysts [3]. Another problem to take into consideration is that the patterns and trends observed in small geographic clusters may not align with large-scale trends or assumptions such as the 'broken windows' theory [1]. Due to the popularity of the idea of solving crime based on the criminal and his motives being central to police work over decades, methods based on geographical profiling have not been expanded on or utilized frequently by law enforcement [1,8]. Even data driven approaches initially just compared crimes being committed to past crimes to observe patterns [3]. But now with the advent of machine learning algorithms, crime and its patterns are more easily observable through multiple metrics [4].

Machine learning is the science of automating the learning process for a model or a system without explicitly programming it. The concept of machine learning has been applied to the concepts of self-driving cars, web searches and pattern recognition [4]. Various machine learning models can now be used for criminal analysis and decision-making [10]. This allows for identification of factors influencing crime. These factors can then be correlated with crimes being committed in either an area of a city or a whole country. Using machine learning and data mining methods also saves time in an otherwise laborious process. Time being a crucial factor in law enforcement. With the support of large amount of data available and machine learning algorithms, the concept of predictive policing in law enforcement is starting to take shape [3].

This practicum attempts to obtain a model which can identify crime hotspots and predict crime to a level of accuracy which can be expanded upon to obtain a high-level prediction model.

Section 2 of this report covers the literature. Section 3 expands on the methodology of the project and section 4 describes the results. Section 5 covers the conclusion and future work of the practicum.

II. RELATED WORK

A. Datasets

The exploration of crime analysis has been carried out from quite a few viewpoints using a few different types of datasets. Some researchers used a London-based crime dataset along with a dataset of geo-localized information such as transportation, households and London borough profiles as well as a behavioral dataset computed from mobile network activity in the city [1]. Some papers used a crime dataset based solely on homicides in Brazilian cities along with 10 urban indicators such as child labor, elderly population and illiteracy [2]. Another approach went with crime in the city of Vancouver along with neighborhood data [4].

A few papers used the Denver crime dataset to identify types of crime occurring in different areas and also to identify patterns and correlation between various factors. These are then used to either predict future crime or to identify hotspots where particular types of crime are occurring. This dataset records the reported offences in the city of Denver and contains information regarding the location, time and type of crime [17,18,19].

A very popular dataset that has been used in crime analysis using machine learning and data mining approaches is the 'Crime and Communities' dataset attained from the UCI machine learning repository. This dataset is prepared from the socio-economic data of the 1990 census, law enforcement data of the 1990 US LEMAS survey and the 1995 FBI UCR crime data. It has a total of 128 attributes with each instance having data from a different state [5,9,10,11].

One paper used crime data in Taiwan along with spatial-temporal data to split city maps into grids, as they were applying a "broken windows" approach to observe crime patterns [7]. One paper used metrics such as victim characteristics such as gender, age and jobs as well as characteristics of the crime site such as location, ownership etc. in order to create geographical profiles for investigating crime [8]. Another paper used spatial-temporal features by applying Google Places API to theft data for Taoyuan city, Taiwan for grid-based crime prediction and to observe crime displacement in the city [12]. Some researchers worked with the police department of a Northeastern US city to collect data and generate aggregated data for different types of crimes along with spatial-temporal information related to the crime data. They applied different data mining and ensemble learning techniques on the same [14,15].

B. Crime Prediction and Hotspots

Previous research in this field has been generally tried to accomplish one of the following objectives:

- Predict crime on a given level such as in a city
- Identify patterns in crimes being committed
- Forecast criminal hotspots by methods such as geographical profiling
- Identify patterns in crime displacement and forecasting the same

Some of the research has been based on a single type of crime or have focused their research on a type of criminal such as serial offenders. Different types of machine learning and data mining algorithms have been used in criminal analysis. Both supervised and unsupervised learning methods have been used but there is clearly more research in supervised algorithms. This is obvious as most crime data available is historic data which gives us clear outputs and thus supervised algorithms have been effective in crime prediction. The models attempting to determine crime patterns require unsupervised learning models to be utilized. There is extensive research in different parts of the field but there is still room for improvement.

One paper used mobile network activity, demographic data and criminal records to identify crime hotspots in the city of London. They divided the city of London into grids and then classified each cell in the grid as either a crime hotspot or not. They used algorithms such as logistic regression, support vector machines, neural networks, decision trees and implementation of ensemble of tree classifiers. They compared the algorithms using metrics such as accuracy, F1 score and AUC score. They also compared performance by comparing features of the combined dataset to the individual datasets. The combined dataset was obtained by using the feature ranking and feature subset selection approach [1].

Another paper focused on finding criminal hotspots in terms of both space and time. They conducted a statistical analysis on two crime datasets from Denver and Los Angeles. They then used the Apriori algorithm to find frequent hotspot patterns. The patterns were in the form of the location, day and time. Then it uses Decision Tree and Naïve Bayes classifiers to predict potential crime types. It also combined Denver's dataset with demographics data to determine factors for crime [2].

One paper attempted to used machine learning algorithms to identify specific criminal patterns to order to recognize a person or group which continuously commit crimes. The model attempts to recognize the *modus operandi* of the criminal by going through the database to recognize growing patterns in the crimes. They devised an algorithm for detecting patterns in crime by using coefficients to define an incidence of crime. They used a set of coefficients to capture the common characteristics of all patterns, patterns specific coefficients which grow to capture the M.O. of a crime and dynamism in the patterns is captured by a similarity coefficient [3].

Another paper attempted to create a prediction model which can accurately predict crime by using 2 approaches. The first approach defines all variables as binary and the second approach defines all variables as numeric. They used k-nearest neighbours and boosted decision trees to predict crime in the city of Vancouver. The boosted decision tree algorithm AdaBoost was used which combines several weak learners to obtain a classifier which is much stronger [4].

One paper used different classification algorithms to compare and contrast the approaches to crime prediction in order to determine the best approach for their model. They used variables such as the state, the median income, education and employment statistics along with criminal statistics to determine the type of machine learning algorithm model that can be used to predict crime [5].

Another paper used a data-driven approach based on the "broken windows" theory which states that failure to respond to low level criminal activity in a location will lead to more serious crime. They designed a model which predicts incidence of drug-related crime in a given month based on crime such as theft, assault, intimidation occurring in the previous month. They determined features by using types of crime and spatial-temporal patterns for each grid on the map, set drug crime in the next month as the dependent variable and ran deep learning and random forest algorithms to predict crime hotspots [7].

One paper uses crime theories and geographical profiling to identify the next possible location of offenses in a serial crime. They use decay functions to create the geographical profile for each factor and use probability distribution functions to see which site will be the location of the next crime. Then they combine all the geographical profiles by

weighing the effect functions for each factor which is adaptively adjusted using Bayesian learning [8].

Researchers also use regression algorithms on two different datasets for a comparative study of violent crime patterns. They only observe crimes such as murder, rape, robbery and aggravated assault and trained data on local level to test it on national statistics. They used regression algorithms such as additive regression and decision stump in order to compare violent crime patterns in two datasets, one from the Crime and Communities dataset and one from Mississippi's crime statistics. An interesting observation in their approach was the use of additive regression, which enhances performance of a regression base classifier by fitting the model using the predictions of the previous iteration [9].

One paper used compared classification algorithms on crime datasets to identify approaches that work best for crime prediction. They also determined rules for classification using the model that worked best and also predicted attributes that contributed most to the crime that occurred [10]. Another paper similarly attempted to predict the type of crime occurring in a location using the state in which the crime occurred among the features for classification. It used socio-economic and spatial data as the features and compared the Naïve-Bayes and Decision Tree approaches to predict the type of crime occurring [11].

One paper used a grid-based prediction model to establish a range of spatial-temporal features which were used to predict crime. They attempted to use machine learning in order to reinforce traditional policing methods for improving crime prevention by designing two types of models, spatial-temporal and empirical. They combined features from both to obtain the final model which they applied in a grid of Taoyuan city on a map [12].

One research used clustering algorithms to detect crime patterns and also used semi-supervised learning for knowledge discovery and improved accuracy for predicting the patterns. They used k-means clustering to predict crime hotspots in a North-eastern city in the US. The features used in the analysis were determined by interacting with domain experts and also by running an attribute importance algorithm. Different types of crime had different sets of attributes becoming important and future crime patterns were detected by using observations given by a detective on the small clusters formed by the model [13].

One research team initially use aggregated data counts and additional features in an ensemble of data mining techniques to perform crime forecasting. Then they use classification methods to determine best models for predicting crime hotspots along with observing identifying emerging hotspots. They also identify the approach which provides the most stable outcomes [14].

Another paper also attempted to find crime hotspots but by using a combination of spatial data mining methods such as point mapping, kernel density estimation (KDE) and Spatial and Temporal Analysis of Crime (STAC). The method they use actually attempts to also capture the difference in patterns of classes in the dataset, providing additional variables used in the forecasting [16].

One research team used non-linear regression models such as random forest regressors on homicide crime data along with urban metrics from a city in Brazil in order to predict crime and used the same regressors to rank the urban indicators in order of influence over crime in the city [19].

III. METHODOLOGY

The objective of this project is to ascertain crime hotspots and also to predict type of crime occurring in the city and county of Denver by using variables such as the location (neighborhood), date and time and the tone of the news cycle on that particular day. Thus, the approach here is to use these variables to first identify frequent patterns by using association rule mining which will determine the hotspots of criminal activity and then we use these variables in a classification model to try and predict the type of crime occurring at a particular instance of crime.

In this section we discuss what datasets are being used in this practicum how they have been prepared, how the data has been analyzed and observed to identify crucial information and how the models have been built.

A. Datasets

Two datasets have been used in this project, namely the Denver crime dataset and the GDELT dataset. The Denver crime dataset is the main dataset which is used in the modeling while data from the GDELT dataset is integrated in order to provide more information about the news environment in Denver.

Denver Crime Dataset-

The Denver crime dataset includes all incidents of criminal offenses reported to the police in the city and county of Denver in the past five years. The data ranges from the 2nd of January 2014 to the 15th of November 2019. The data format is based on the National Incident Based Reporting System (NIBRS). This contains information such as first occurrence date, type of crime, geographical location of crime and the neighborhood in which the crime was committed. The dataset has 19 columns and 528587 instances [20]. The dataset has been taken from Kaggle, which updates the data from the Denver open data catalog. A table of key attributes is Table 1.

Attribute	Data Type	Values
Offense_Category_Id	Nominal	13 categories such as Aggravated-assault, Larceny, Murder, Drug-alcohol etc.
First_Occurence_Date	Date and Time	9/30/2019 9:00:00 PM
Neighborhood_Id	Nominal	78 neighborhoods such as Montbello, Stapleton etc.

Table 1. Denver key attributes

GDELT Dataset-

The basic idea behind GDELT is to provide a global database of society. GDELT monitors the broadcast, print and web news from nearly every corner of the world, identifies the parties involved, locations, emotions and sources, creating a free and open platform for news and events for the entire world [21].

In this project, the GDELT data helps gauge the news environment for the city of Denver. This allows us to

understand what type of incidents were occurring in Denver around the time a particular instance of crime took place in the city. It also gives an idea of the mindset of the people in the city at the time.

The GDELT dataset has been obtained using a data wrangler in R as the GDELT platform provides large amounts of data but only with date as a filter. The final GDELT file for Denver has more than 2 million rows and 58 columns [22]. Below is the key attributes table for the GDELT file in Table 2.

Attribute	Data Type	Values
NumArticles	Numerical	Number of articles on event
ActionGeo _ADM1Code	Nominal	Indicates part of world where action takes place
AvgTone	Float	Average tone of articles being written

Table 2. GDELT key attributes

B. Data Preprocessing

The main steps in preparation of data for modeling on the two datasets were:

Data Cleaning-

To obtain the GDELT data related to Denver in the time period required, a data wrangler written in R is used which passes through GDELT files for each date, extracts rows related to Denver and exports them to a smaller set of csv files. It then combines the filtered set of csv files to obtain one file fulfilling the requirements of the filter in place. In the GDELT dataset, there were no Nan, missing or inconsistent values in the dates, action code, number of articles or average tone for any event.

There were a few Nan values in some of the latitude and longitude locational values of incidents in the Denver dataset. As these variables aren't directly used the final dataset, these incidents do not need to be dropped from the dataset. There were also some incidents that took place well beyond Denver's boundary, so they were removed from the dataset. There were no other inconsistencies in the data for the attributes to be used.

Data Reduction-

In both the datasets, we selected attributes that were of use to us and removed the rest. In the Denver dataset, we retained only three columns to use for the modeling. From the GDELT dataset we retained three columns which were then transformed and integrated into the Denver crime dataset.

Another way we reduced the data was by using the 'IS_CRIME' column to remove the traffic accidents in the datasets as traffic accidents were reported as incidents in the dataset.

Data Transformation-

The data transformation on the Denver dataset was to break down the 'FIRST_OCCURRENCE_DATE' data into features such as year, month, day, day of the week and hour. We also minimized the 'OFFENSE_CATEGORY_ID' values from 14 different values to 7. Crimes such as aggravated assault, murder and arson which risk fatal harm are categorized as dangerous crimes, incidents of some form of theft such as larceny or robbery were combined into one category and vehicle related crimes were combined into one category. The 'HOUR' feature and the 'TONE' feature was categorized into 6 intervals for association rule mining part of the modeling.

In the GDELT dataset, we multiplied the number of articles and the average tone for each event, then we calculated the average tone for events on a particular day, which became our variable representing the news cycle in the city of Denver on that day.

Data Integration-

We integrated the GDELT and Denver datasets using the date as the key to obtain the final dataset that was used for the modeling. The date attribute that was used as the key was present in the GDELT dataset as 'SQLDATE' so we converted the 'FIRST_OCCURRENCE_DATE' by using datetime functions to obtain a variable in the form that was present in the GDELT dataset.

Table 3 gives the final set of variables that were used in the model.

Attribute	Number of Distinct values
Offense_Category_Id	7
Neighborhood_Id	78
Year	6 (2014 to 2019)
Month	12
Day	31
Day_of_Week	7
Hour	24 (converted to 6)
Tone	Float (converted to 6)

Table 3. Description of final dataset

C. Data Analysis

By conducting an initial statistical analysis of the datasets, we tried to understand the data and realize the features that shall be used in the modeling. Data has been visualized to better understand the data and each of them cover how criminal incidences were spread according to each aspect of a variable.

Figure 1 shows how crime is distributed in the city of Denver in terms of the different types of crimes that are occurring in the city. As we can clearly see the most frequent type of crime occurring in Denver (except for 'all-other-crimes) are public disorder, larceny and thefts from motor vehicles. Crimes such as robbery, arson and murder are the least frequent types of crime in the city.

Figure 2 shows the hourly distribution of crime in Denver. Crime is significantly less in late night to early morning and peaks during the early evening from 4 P.M. to 6 P.M. and late night at 10 P.M.

Figure 3 shows the overall trend of crime throughout the time period of the data. There are a few peaks and troughs and the overall mean is around 185 cases per day in the city of Denver.

Figure 4 shows neighborhoods with the highest number of cases in the city. Neighborhoods such as Five Points are clearly the most dangerous with well over 20,000 cases recorded in the neighborhood, with at least 5000 more cases than the second most dangerous neighborhood which is CBD.

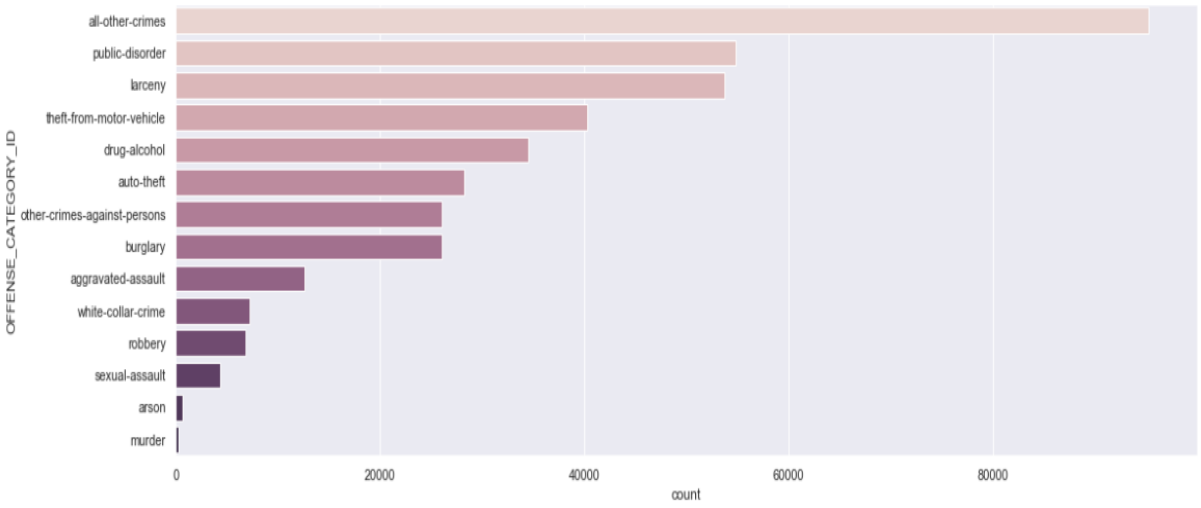

Figure 1. Number of incidents based on crime types

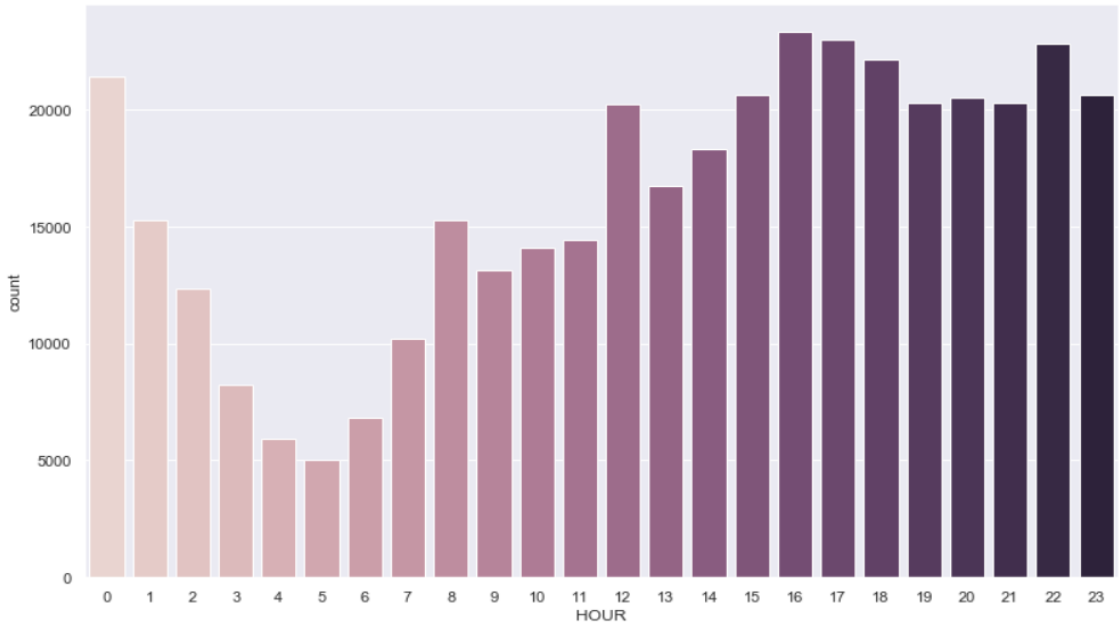

Figure 2. Number of incidents for every hour

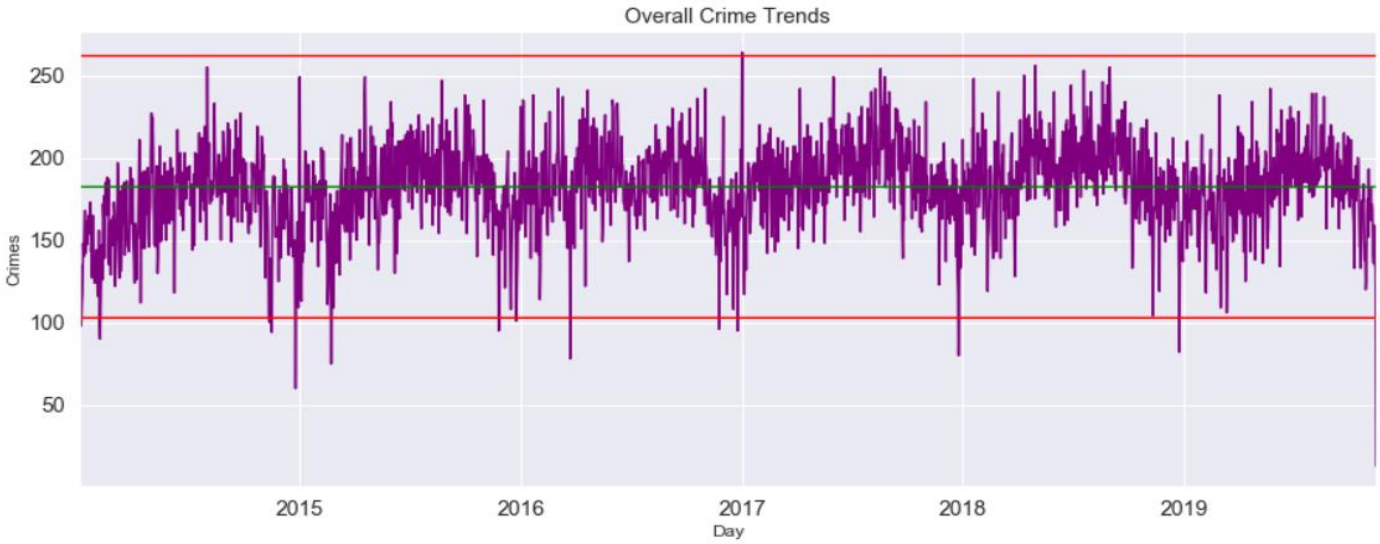

Figure 7. Overall Crime Trends

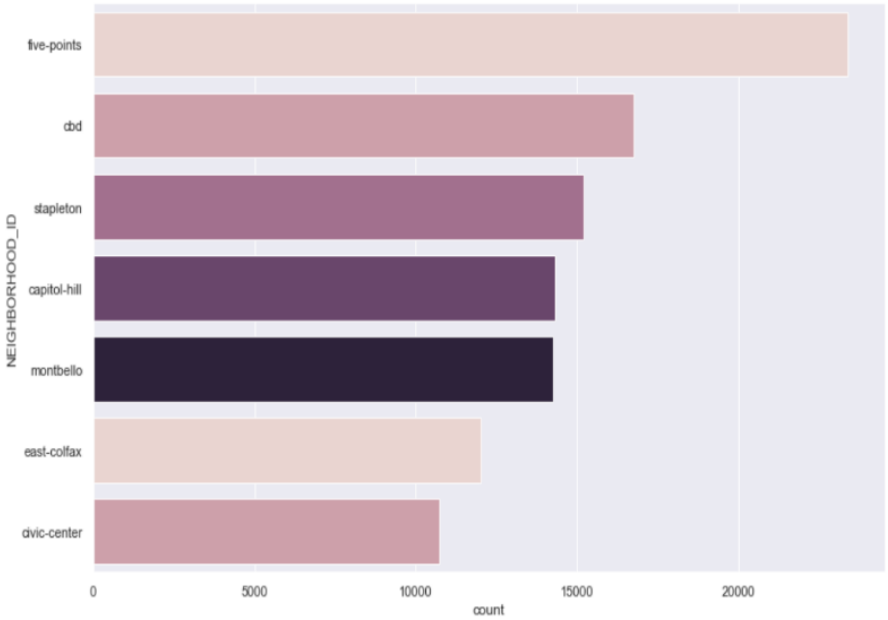

Figure 6. Most dangerous neighborhoods

D. Data Modeling

The final dataset was modeled in two different ways. In the first model, the 'TONE' and 'HOUR' values were converted to categorical intervals in order to execute the FP-growth algorithm which gives the most frequent patterns in the data. This helps identify hotspots in terms of location (neighborhood), time, day and tone of the news cycle. In the second model, the 'NEIGHBORHOOD_ID' variable is converted to dummy variables so that classifiers such as logistic regression, random forest and decision trees can be used in order to predict the type of crime occurring in a particular incident.

FP-Growth Algorithm-

FP-growth is currently one of the fastest approaches for association rule mining. The basic logic behind the FP-growth is that it starts eliminating items from transactions which do not fulfill the minimum support requirements individually. It then starts removing the least frequent item of the remaining items from the transaction set. This process is repeated with the second least frequent item with minimum support to obtain a reduced dataset. This process is known as the recursive elimination method which gives us the final frequent itemset based on the requirements [23].

This practicum uses the FP-growth algorithm to identify hotspots based on the location, time, day and tonality in the city of Denver. We used the mlxtend library in Python to implement the algorithm [24]. This association rule mining method allowed us to come up with a list of crime hotspots along with related time and news events features. Multiple experiments were conducted with different support values in order to find the optimal choice. The final value chosen was 0.0005, which would be around 187 absolute frequencies i.e. incidents of crime occurring at that particular instance.

Classification Algorithms-

Different types of classification algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machines and Neural Networks have been used in order to identify the method that works best in terms of predicting the type of crime occurring in a particular incident recorded.

The classification models are applied on two datasets, one which contains the incidents that have been categorized as 'other-crimes' and one in which incidents classified as 'other-crimes' have been removed. This has been done to observe how the metrics of the models change when it attempts to classify crimes that have been more specifically classified.

The classification models have all been constructed using the Scikit-Learn, an open source library which provides data mining and machine learning tools for Python.

- Logistic Regression: Logistic regression, in its basic form uses a logistic function to classify a binary variable. It can also be used to classify multinomial variables. Here, the solver we have used for logistic regression is the 'lbfgs' solver because it can handle multinomial dependent variables better than other solvers [25].
- Decision Tree: Decision trees classifiers attempt to create a model that predicts the target value using simple decision rules that it obtains from parsing the data features. The decision tree here uses the 'gini' criterion, which forms the decision tree on the basis of the Gini Impurity value for each node. Gini impurity measures the probability of incorrect classification of an instance of the random variable [26].
- Random Forest: Random forest classifier is a meta estimator which uses sub-samples of data to fit a number of trees and estimates the value by averaging the results. We have used 150 estimators (sub-trees) in order to classify the incidents in this case [27].
- Support Vector Machine: Support vector machines are supervised machine learning algorithms which classifies data points by forming a hyperplane which best distinguishes the data. Here, the kernel being

used is ‘poly’ which creates polynomial hyperplanes for the decision boundary [28].

- Neural Networks: Neural network is a supervised algorithm which learns a function by passing the data through multiple layers of nodes and modifying the weights on those layers through correction by error over multiple iterations of the training data. Here we use the ‘relu’ activation and the ‘adam’ solver [29].

IV. Results and Evaluation

In this section, we evaluate and summarize key results that we obtained from the association rule mining and classification models.

A. Frequent Crime Hotspots

The first goal of this project was to find crime hotspots in the city of Denver. We have extracted the most frequent patterns which gives us a reasonable idea of the spatial and temporal conditions in which crime occurs more frequently. We obtained 43 values, 10 of which are in Table 4.

The most frequent pattern observed is that a criminal incident takes place in Five points, late in the evening on Friday on a negative newsday. Five points is clearly the most dangerous neighborhood looking at the itemsets with Union Station, Stapleton and CBD the next most dangerous areas. An interesting observation is that after midnight on a Sunday is a frequent itemset in both Union Station and Five Points, even though the support value of the individual items compared is comparatively low..

Support	Itemsets	Length
0.000999651	'Friday', 'T5', 'five-points', 'Negative'	4
0.000933007	'five-points', 'T6', 'Saturday', 'Negative'	4
0.000831709	'Friday', 'five-points', 'T6', 'Negative'	4
0.000794389	'Monday', 'T5', 'five-points','Slightly_Negative'	4
0.000786392	'Sunday', 'T5', 'five-points', 'Slightly_Negative'	4
0.000775729	'T5', 'five-points', 'Saturday', 'Negative'	4
0.000765066	'T1', 'Sunday', 'Slightly_Negative', 'union-station'	4
0.000733077	'T5', 'Thursday', 'Negative', 'five-points'	4
0.000733077	'Sunday', 'T1', 'five-points', 'Slightly_Negative'	4
0.000719749	'T5', 'five-points', 'Wednesday', 'Negative'	4

Table 4. 10 most frequent itemsets

B. Predicting Type of Crime

The second objective of the project is to predict the type of crime occurring at a particular instance. The accuracy metric of all the models is given in Table 5. This table includes accuracy metrics for both the datasets that have been used in prediction, one with the ‘other-crimes’ data and one without.

Classification Algorithm	Accuracy (with other-crimes)	Accuracy (without other-crimes)
Logistic Regression	0.3473	0.3630
Decision Trees	0.2918	0.3251
Random Forest	0.3397	0.3765
Support Vector Machines	0.1868	0.2497
Neural Networks	0.3661	0.3831

Table 5. Accuracy metric for classification algorithms

Neural networks are the most accurate model for both datasets in this scenario. And looking at the classification report, it’s observed that the neural network predicts other crime most accurately along with strong F1-scores in a few categories. It can also predict theft, vehicle related theft and drug and alcohol crimes reasonably well. But it is very poor in predicting crimes with smaller number of instances. It does significantly better in predicting all types of crimes when we remove the all other crimes category and data. The classification reports of all the models are available in Appendix B.

Random forest is the most even in terms of accuracy of prediction for each category. It’s the best predictor for smaller support values such as dangerous and white-collar crimes and it has the best F1-score of all the models. It’s the best predictor across all categories.

Logistic regression is only able to predict a few categories of values such as other crimes and theft, which have large support values. It has zero accuracy for small support values such as dangerous crimes and even public disorder.

Decision trees exhibit similar prediction patterns as the random forest model but is not as accurate as the logistic regression model. But its recall, precision and F1-scores are better than the regression model This indicates that decision trees actually work better than logistic regression in terms of predicting across all categories.

Support vector machine is the poorest performer of the lot. The only category that it is able to predict to some extent is vehicle related crimes. It is a very poor performer in all other categories. It is only able to define a hyperplane for one category to some extent and is ineffective for this purpose.

V. Conclusion and Future Work

This project delves into the crime dataset for the city of Denver and attempts to identify hotspots and type of crime occurring in the city. The FP-growth algorithm helps identifying the location and time at which crime would be likeliest to occur in the city, which can help in police monitoring and faster response teams in the area. Using the classification algorithm, we can try to guess the type of crime occurring which allows the police and the government to take action and better deal with crime. The neural network model worked best for this purpose but if the objective is to have reasonable probability over the different categories of crime, random forest is more effective.

As the accuracy metric indicates, it's quite clear that any of the classification models cannot actually be applied in a real-world scenario. There is clear scope for improvement in utilizing the data to make improved predictions. This can be done by the following ways:

- Adding different types data as variables such as social and population demographics, housing data etc. to capture the environment of the city of Denver better. This would definitely help in identifying dangerous locations better and predicting and preventing crime in the city.
- Using social media, we can identify the density of people in different parts of the city at different times. This can help identifying crimes in areas, for example crimes such as pickpocketing would occur in higher density areas and crimes such as robbery would occur in sparsely populated areas.
- One future expansion that can done is to improve the association rule mining in such a way that trends in the crime hotspot mapping can also be observed. This can be done by adding variables for the year and month features to the model.
- Another way the model can be improved is by using the tonality feature over a period of days rather than using just one day as a feature. We can use a time series approach to give more dynamism to the tonality feature. This can help as the effect of news events is dynamic in terms of time, which may lend greater weight to how news can have an impact of crime.

Crime has been growing all around the world over the past few decades. As technology has advanced and more data has become available, data-driven approaches to predict crime can really help in preventing and deploying the police force more effectively. Adaptation of data-driven approaches would help the police force and make cities safer and more habitable.

VI. REFERENCES

[1] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, Alex Pentland, "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data", 16th International Conference on Multimodal Interaction, November 2014, Trento, Italy.

[2] Tahani Almanie, Rsha Mirza, Elizabeth Lor,"Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots", International Journal of Data Mining & Knowledge Management Process (IJDKP), July 2015, Boulder,USA.

[3] Tong Wang, Cynthia Rudin, Daniel Wagner, Rich Sevieri, "Learning to Detect Patterns of Crime", Machine Learning and Knowledge Discovery in Databases, September 2013, Cambridge, USA.

[4] Suhong Kim, Param Joshi, Parminder Singh Kalsi, Pooya Taheri, "Crime Analysis Through Machine Learning", 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Novmber 2018, British Columbia, Canada.

[5] Emmanuel Ahishakiye, Elisha Opiyo Omulo, Danison Taremwa, Ivan Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm", International Journal of Computer and Information Technology, Volume 06 –Issue 03, May 2017, Nairobi, Kenya.

[6] Maria R. D'Orsogna, Matjaž Perc, "Statistical physics of crime: A review", Physics of Life Reviews, March 2015, Los Angeles, USA.

[7] Ying-Lung Lin, Tenge-Yang Chen, Liang-Chih Yu, "Using Machine Learning to Assist Crime Prevention", 2017 6th IIAI International Congress on Advanced Applied Informatics, July 2017, Hamamatsu, Japan.

[8] Renjie Liao, Xueyao Wang, Lun Li, Zengchang Qin, "A novel serial crime prediction model based on Bayesian learning theory", 2010 International Conference on Machine Learning and Cybernetics, July 2010, Qingdao, China.

[9] Lawrence McClendon, Natarajan Meghanathan, "Using Machine Learning Algorithms to Analyze Crime data", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015, Jackson, USA.

[10] Umair Saeed, Muhammad Sarim, Amna Usmani, Aniqa Mukhtar, Abdul Basit Shaikh, Sheikh Kashif Raffat, "Application of Machine learning Algorithms in Crime Classification and Classification Rule Mining", Research Journal of Recent Sciences, Vol. 4(3), March 2015, Karachi, Pakistan.

[11] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy, Nasim Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction", Indian Journal of Science and Technology, March 2013, Selangor, Malaysia.

[12] Ying-Lung Lin, Meng-Feng Yen, Liang-Chih Yu, "Grid-Based Crime Prediction Using Geographical Features", ISPRS International Journal of Geo-Information, July 2018, Tainan City, Taiwan.

[13] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining", 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, December 2006, Hong Kong, China.

[14] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding, "Crime Forecasting Using Data Mining Techniques", 2011 IEEE 11th International Conference on Data Mining Workshops, December 2011, Vancouver, Canada.

[15] Jacky Yu, Wei Ding, Ping Chen, Melissa Morabito, "Crime Forecasting Using Spatio-Temporal Pattern with Ensemble Learning", Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 2014, Boston, USA.

[16] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, Melissa Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach", Applied Intelligence, Volume 39, December 2012, Boston, USA.

[17] Md. Aminur Rab Ratul, "A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining", Cornell University Machine Learning, Jan 2020, Ottawa, Canada.

[18] Mrinalini Jangra, Shaveta Kalsi, "Naïve Bayes Approach for the Crime Prediction in Data Mining" , International Journal of Computer Applications, May 2019, Jalandhar, India.

[19] Luiz G.A. Alves, Haroldo V. Ribeiro, Francisco A. Rodrigues, "Crime prediction through urban metrics and statistical learning", Physica A: Statistical Mechanics and its Applications, September 2018, Sao Paulo, Brazil.

[20] 'Denver Crime Data', Available: https://www.kaggle.com/paultimothymooney/denver-crime-data#crime.zip

[21] 'GDELT Project', Available: https://www.gdeltproject.org/

[22] 'GDELT Data Wrangler', Available: https://nbviewer.jupyter.org/github/JamesPHoughton/Published_Blog_Scripts/blob/master/GDELT%20Wrangler%20-%20Clean.ipynb

[23] Christian Borgelt, "An Implementation of the FP-growth Algorithm", 2005 Open Source Data Mining, Aug 2005, Chicago, USA.

[24] 'Frequent Itemsets via the FP-Growth Algorithm', https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/

[25] 'SkLearn Logistic Regression', https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[26] 'SkLearn Decision Tree', https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[27] 'SkLearn Random Forest', https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[28] 'SkLearn SVM', https://scikit-learn.org/stable/modules/svm.html

[29] 'SkLearn Logistic Regression', https://scikit-learn.org/stable/modules/neural_networks_supervised.html

APPENDIX A

This appendix contains the complete set of values for the crime hotspots as obtained by the FP-Growth algorithm:

support	itemsets	length
0.001	frozenset({'Friday', 'T5', 'five-points', 'Negative'})	4
0.000933	frozenset({'five-points', 'T6', 'Saturday', 'Negative'})	4
0.000832	frozenset({'Friday', 'five-points', 'T6', 'Negative'})	4
0.000794	frozenset({'Monday', 'T5', 'five-points', 'Slightly_Negative'})	4
0.000786	frozenset({'Sunday', 'T5', 'five-points', 'Slightly_Negative'})	4
0.000776	frozenset({'T5', 'five-points', 'Saturday', 'Negative'})	4
0.000765	frozenset({'T1', 'Sunday', 'Slightly_Negative', 'union-station'})	4
0.000733	frozenset({'T5', 'Thursday', 'Negative', 'five-points'})	4
0.000733	frozenset({'Sunday', 'T1', 'five-points', 'Slightly_Negative'})	4
0.00072	frozenset({'T5', 'five-points', 'Wednesday', 'Negative'})	4
0.00072	frozenset({'Tuesday', 'T5', 'five-points', 'Negative'})	4
0.000717	frozenset({'T5', 'five-points', 'Wednesday', 'Slightly_Negative'})	4
0.000709	frozenset({'Thursday', 'T6', 'Negative', 'five-points'})	4
0.000704	frozenset({'T1', 'five-points', 'Saturday', 'Negative'})	4
0.000666	frozenset({'Friday', 'T5', 'stapleton', 'Negative'})	4
0.00065	frozenset({'T1', 'Saturday', 'Negative', 'union-station'})	4
0.00065	frozenset({'Tuesday', 'T5', 'five-points', 'Slightly_Negative'})	4
0.000645	frozenset({'Monday', 'T5', 'five-points', 'Negative'})	4
0.000626	frozenset({'stapleton', 'Thursday', 'T4', 'Negative'})	4
0.00061	frozenset({'stapleton', 'T5', 'Thursday', 'Negative'})	4
0.000608	frozenset({'Friday', 'cbd', 'T5', 'Negative'})	4
0.000602	frozenset({'Friday', 'five-points', 'T4', 'Negative'})	4
0.000597	frozenset({'T6', 'five-points', 'Wednesday', 'Negative'})	4
0.000594	frozenset({'cbd', 'T5', 'Thursday', 'Negative'})	4
0.000594	frozenset({'Sunday', 'Slightly_Negative', 'five-points', 'T6'})	4
0.000592	frozenset({'Thursday', 'T4', 'five-points', 'Negative'})	4
0.000578	frozenset({'Monday', 'Slightly_Negative', 'five-points', 'T6'})	4
0.000576	frozenset({'cbd', 'T5', 'Saturday', 'Negative'})	4
0.000562	frozenset({'Sunday', 'T5', 'five-points', 'Negative'})	4
0.000557	frozenset({'T5', 'Slightly_Negative', 'five-points', 'Saturday'})	4
0.000549	frozenset({'T5', 'stapleton', 'Wednesday', 'Negative'})	4
0.000549	frozenset({'Slightly_Negative', 'five-points', 'T6', 'Saturday'})	4
0.000536	frozenset({'Tuesday', 'montbello', 'T4', 'Negative'})	4
0.000536	frozenset({'Friday', 'stapleton', 'T4', 'Negative'})	4
0.000528	frozenset({'Friday', 'T5', 'five-points', 'Slightly_Negative'})	4
0.000525	frozenset({'T1', 'Sunday', 'Negative', 'union-station'})	4
0.000525	frozenset({'Monday', 'five-points', 'T6', 'Negative'})	4
0.000517	frozenset({'T6', 'Slightly_Negative', 'five-points', 'Wednesday'})	4
0.000512	frozenset({'five-points', 'T4', 'Saturday', 'Negative'})	4
0.000509	frozenset({'Friday', 'Slightly_Negative', 'five-points', 'T6'})	4
0.000506	frozenset({'Friday', 'T1', 'five-points', 'Negative'})	4
0.000504	frozenset({'cbd', 'T4', 'Wednesday', 'Negative'})	4
0.000501	frozenset({'cbd', 'Thursday', 'T4', 'Negative'})	4

APPENDIX B

This appendix contains the classification reports of all the classification models in the project in order of their accuracy. The reports for both the datasets are together:

1. Neural Networks

	dangerous-crimes	drug-alcohol	other-crimes	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.001823	0.213411	0.493207	0.065288	0.368348	0.284877	0.006604	0.366132	0.204794	0.316299
Recall	0.3	0.333249	0.382632	0.263467	0.355758	0.339998	0.125	0.366132	0.300015	0.34002
F-score	0.000914	0.156966	0.693668	0.037261	0.38186	0.245135	0.003392	0.366132	0.217028	0.366132
Support	3282	8384	29517	13258	20717	16856	1769	0.366132	93783	93783

Neural Network with other crimes

	dangerous-crimes	drug-alcohol	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.027162	0.349823	0.196337	0.488024	0.395936	0.005672	0.383114	0.243826	0.348166
Recall	0.196787	0.389186	0.317243	0.397399	0.37968	0.178571	0.383114	0.309811	0.358571
F-score	0.014588	0.317691	0.142159	0.632192	0.413645	0.002882	0.383114	0.253859	0.383114
Support	3359	8247	13499	20788	16768	1735	0.383114	64396	64396

Neural Network with other crimes

2. Random Forest

	dangerous-crimes	drug-alcohol	other-crimes	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.058912	0.276873	0.450228	0.194368	0.34495	0.292614	0.056386	0.339784	0.23919	0.326133
Recall	0.101783	0.316328	0.39807	0.233628	0.338632	0.302146	0.103286	0.339784	0.256268	0.321398
F-score	0.041452	0.246169	0.518116	0.166404	0.351508	0.283666	0.038778	0.339784	0.235156	0.339784
Support	3305	8287	29532	13335	21092	16530	1702	0.339784	93783	93783

Random Forest with other crimes

	dangerous-crimes	drug-alcohol	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.074675	0.38896	0.304458	0.450439	0.383947	0.057906	0.376561	0.276731	0.364539
Recall	0.136719	0.416908	0.335195	0.408158	0.365273	0.126459	0.376561	0.298119	0.361075
F-score	0.051365	0.364524	0.278885	0.502492	0.404633	0.037551	0.376561	0.273242	0.376561
Support	3407	8293	13346	20868	16751	1731	0.376561	64396	64396

Random Forest without other crimes

3. Logistic Regression

	dangerous-crimes	drug-alcohol	other-crimes	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0	0.037248	0.487815	0	0.328742	0.235311	0	0.347302	0.155588	0.272474
Recall	0	0.269355	0.358895	0	0.332721	0.311098	0	0.347302	0.181724	0.267336
F-score	0	0.020007	0.761276	0	0.324856	0.189216	0	0.347302	0.185051	0.347302
Support	3259	8347	29331	13073	21194	16859	1720	0.347302	93783	93783

Logistic Regression with other crimes

	dangerous-crimes	drug-alcohol	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0	0.292194	0.133412	0.476008	0.381765	0	0.363066	0.213897	0.319244
Recall	0	0.332579	0.300507	0.388199	0.343661	0	0.363066	0.227491	0.32014
F-score	0	0.260556	0.085738	0.615154	0.429374	0	0.363066	0.231804	0.363066
Support	3375	8455	13133	21037	16559	1837	0.363066	64396	64396

Logistic Regression without other crimes

4. Decision Tree

	dangerous-crimes	drug-alcohol	other-crimes	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.072272	0.25814	0.39996	0.191465	0.291174	0.260875	0.054688	0.291812	0.218368	0.290429
Recall	0.069089	0.239781	0.386867	0.195342	0.302706	0.273379	0.058786	0.291812	0.217993	0.29003
F-score	0.075762	0.279543	0.413972	0.187739	0.280489	0.249465	0.051124	0.291812	0.219728	0.291812
Support	3313	8310	29345	13359	20860	16816	1780	0.291812	93783	93783

Decision Tree with other crimes

	dangerous-crimes	drug-alcohol	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.095612	0.34635	0.298177	0.386466	0.329132	0.074824	0.325067	0.255093	0.325092
Recall	0.09387	0.335541	0.294277	0.390577	0.333088	0.079843	0.325067	0.254533	0.325303
F-score	0.097419	0.357878	0.302182	0.382439	0.325268	0.070398	0.325067	0.255931	0.325067
Support	3254	8352	13290	21070	16697	1733	0.325067	64396	64396

Decision Tree without other crimes

5. Support Vector Machine

	dangerous-crimes	drug-alcohol	other-crimes	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.001821	0	0.093734	0.000901	0.003502	0.301677	0.034639	0.186814	0.062325	0.085423
Recall	0.044118	0	0.334378	0.166667	0.224242	0.179854	0.061135	0.186814	0.144342	0.213508
F-score	0.00093	0	0.054507	0.000452	0.001765	0.934979	0.024166	0.186814	0.145257	0.186814
Support	3227	8289	29354	13279	20963	16933	1738	0.186814	93783	93783

SVM with other crimes

	dangerous-crimes	drug-alcohol	public-disorder	theft	vehicle-related-theft	white-collar-crime	accuracy	macro avg	weighted avg
Precision	0.00242	0	0.041095	0.125469	0.396127	0.028583	0.249674	0.098949	0.152486
Recall	0.061538	0	0.233877	0.299355	0.258617	0.022146	0.249674	0.145922	0.216719
F-score	0.001234	0	0.022527	0.079367	0.845903	0.040295	0.249674	0.164888	0.249674
Support	3241	8390	13362	21054	16587	1762	0.249674	64396	64396

SVM without other crimes