Predicting Safety Levels of Neighborhoods in Austin

Sharayu Shenoy [*] [1] Shweta Mane [*] [1]

1. Introduction

Today, safety has become a huge concern and curbing crime has become a necessity. One would want to know if an area safe enough especially when a person is new to that place. Travelling alone is a risk and knowing ”Can I go out in the evening?” is what inspired us to build a solution that can answer these questions for us. Our proposed idea will analyze historical crime data (currently restricted to Austin) and predict what is the crime severity in a neighbourhood or location is during a particular time of the day. This will enable people determine their safety levels and decide whether or not to go to a certain place at certain time. This solution will benefit not just people who live in that neighborhood/area/city but also people who are visiting from outside.

Existing solutions in this space aim at providing safety by attempting to reduce crimes. Research done till date focuses on predicting crimes in a certain area in order to help law-enforcement to determine where to tighten the security. There are few websites which only show that certain area is more crime-prone than the other which can be useful when people are dealing with real estate problems. Safety is always spoken in context of reducing crime. We couldn’t find any application which will simply tell user how safe the area is at any given time.

While it is the duty of the law-enforcement to ensure our safety, we have not yet been able to define a fool-proof system that prevents crimes and guarantees citizen safety. Guaranteeing 100 percent safety seems far fetched, therefore we propose a solution that will help its users take charge of their own safety and make well informed decisions. We propose to develop a solution which predicts crime severity levels by taking into consideration the geographic location and its crime history refined by the time of day, type of crime and exact location type(Garage, parking lot etc.). This differs from existing solutions in two aspects. First, in terms of the target audience and second, in terms of parameters considered to analyze crime and the output. We intend to consider ’time of day’ and ’location type’ as important attributes which will make predictions more precise. We also intend to add different weights to different types of crime according to severity of the crime. Our model will predict safety based on the inputs and weights. We feel providing users with such a solution may not reduce the threats on the street but it will reduce the number of victims.

2. Related Work

Crime Prediction

The current research in the ”Crime Prediction” space mostly caters to the methods that can be used to predict the highest offence crime in that area based on latitude and longitude. ’Prediction of crime occurrence from multi-modal data using deep learning’ (Kang & Kang, 2017) uses Deep Neural Networks to predict crime occurrences. Comprehensive Comparative analysis of methods for crime rate prediction (Vaidya et al., 2018) highlights all the different algorithms that maybe useful in predicting crime rates. While these studies focus on the achieving the problem using historical data, ’Once upon a crime: towards crime prediction from demographics and mobile data’ (Bogomolov et al., 2014) uses a more novel approach of using mobile network activity. Another similar study ’A Study on Public Safety Prediction using Satellite Imagery and Open Data’ (Najjar, 2017) explores new ways to achieve better prediction of crime. Our proposed project aims at defining the severity of a crime in a particular area narrowed down based on the time of the day and location type. We will consider x.y coordinates instead of latitude and longitude in order to get more precise geographic location. Most of the related work speaks to regulatory bodies and authorities while we intend to speak to the general public and help them in their planning process.

Crime and Safety Analysis

Research paper ’Safety in urban neighborhoods: A comparison of physical characteristics and informal territorial control in high and low crime neighborhoods’(Greenberg et al., 1982) shows how some areas manage to maintain low level of criminal activity despite their physical proximity

[*]Equal contribution [1]School of Information, UT Austin, Austin, Texas. Correspondence to: Sharayu Shenoy <sharayushenoy@utexas.edu>, Shweta Mane <smane@utexas.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

and social similarity to high crime areas. A book 'Crime analysis with crime mapping'(Santos, 2016) documents the processes, data and demonstration of how results are used within police organization. Another similar book 'CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0)'(Wortley & Townsley, 2016) talks about environmental criminology and its effects on crime patterns. We aim to go one step further and use the crime analysis to predict safety level in the particular area.

Aggregate Data Regression

Poisson-based regression analysis of aggregate crime rates (Osgood, 2000)
Regression analysis of aggregate continuous data (Moineddin & Urquia, 2014)
Ratio variables in aggregate data analysis: their uses, problems, and alternatives (Bollen & Ward, 1979)
Above articles talk about aggregate data regression techniques at very broad level. We intend to use these methods specifically for crime data analysis and safety determination.

3. Methods

3.1. Dataset Description and Features

Crime records dataset was obtained from the official city of Austin open data portal. Dataset contains approximately 225,000 records and has several features but we decided to select only those features which we thought were the most correlated with crime Each record has following information-

- Occurred Date Time
- Location Type
- Zip code
- X co-ordinate
- Y co-ordinate
- Offence Type

3.2. Solutions

We decided to solve the problem using 2 approaches.

3.2.1. Method 1 : Supervised Classification

Method 1 of solving the problem uses classification algorithms to determine what is the most probable threat that one may encounter as a given location and what is the severity of that threat.

Design Decisions: We decided to use Random Forest Classifier for this method as it was giving the best accuracy score as compared to other classification algorithms. Other algorithms that we used were:

- Decision Tree
- Ridge Classifier
- K-nearest Neighbor
- Neural Network

Implementation Set-up:
Data Pre-processing: Firstly, we extracted year, month, day and hour from 'Occurred Date Time' column then, we calculated severity based on offence type column of the dataset. Since we had 225,000 records, we dropped rows which had null values. To further clean the data, we factorized all the categorical variables including variables of number type like zip code which have a categorical meaning.

Data Augmentation: The dataset contained only the offence type and our intention was to determine the severity of these offences. We conducted research to determine the offence severity as per the state of Texas and Federal government. This resultant severity type is considered to be the output array which will be used for classification.

Table 1. Severity Levels

Offence Type	Severity
Theft	1
Auto Theft	2
Robbery	3
Burglary	4
Aggravated Assault	5
Rape	6
Murder	7

Data Split: In order to avoid over fitting or under fitting, we decided to include validation of model before running it on test data. The data is split into train, validation and test set in 8:1:1 ratio.

Figure 1 shows the flow diagram for this method.

3.2.2. Method 2: Unsupervised Clustering

Method 1 gives the severity of the highest probable threat but with Method 2 we try to solve this and provide the user with a holistic severity rating. Method 2 utilizes unsupervised learning to cluster the dataset and for each cluster based on the all the data-points available in the cluster, severity of the entire cluster is determined based on individual crime severity.

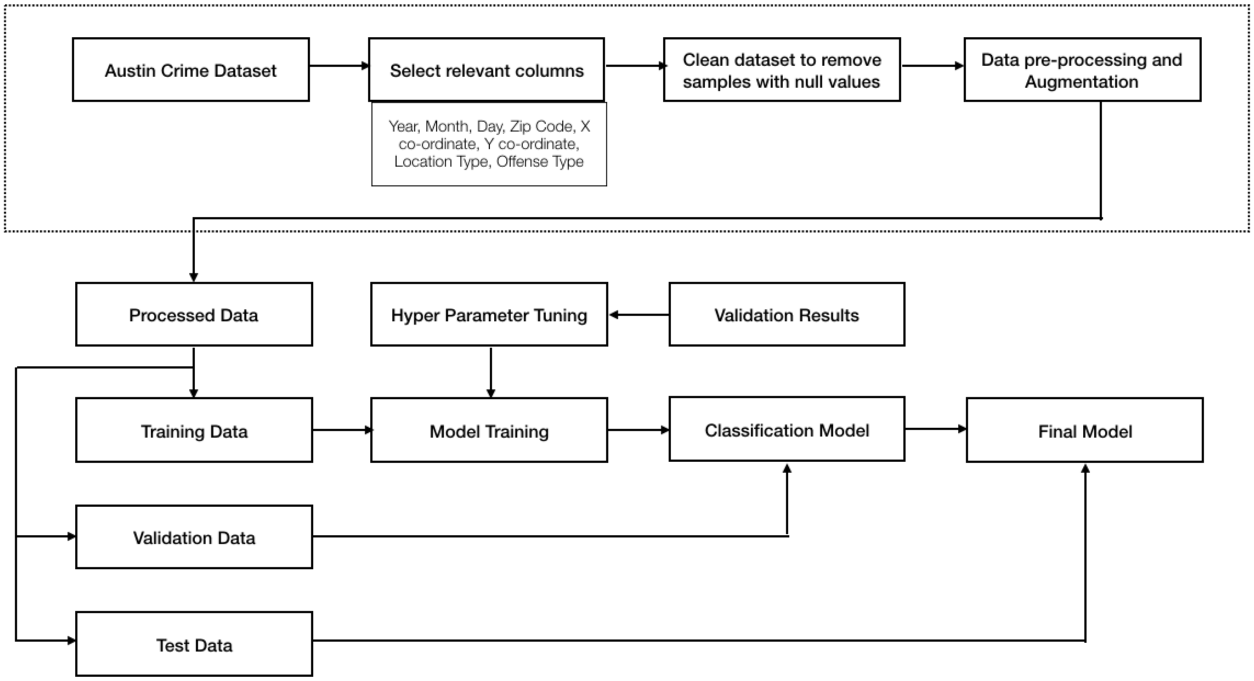

Figure 1. Machine Learning Method 1

Design Decisions: We decided to use k-means method of clustering to partition the data into relevant groups. k-mean clustering was compared to hierarchical clustering using ward and linkage. However, since the data is not of hierarchical nature, we decided to use k-means.

Implementation Set-up:
Data Pre-processing and Augmentation: Data cleanup and modification for this method is same as Method 1. However, instead of separating the offence type as an output array, it is included in the input array.

Data Split: In order to avoid over fitting or under fitting, we decided to include validation of model before running it on test data. The data is split into train, validation and test set in 8:1:1 ratio.

4. Experimental Design

4.1. Experiment 1: Using all features in dataset

Main purpose: We extracted only those features from the dataset that we think are relevant. The aim of this experiment is to evaluate the model without any additional processing.

Dataset: Original Crime dataset with data filtered for years 2013 to 2018 (45067 data samples)

Baseline: Since we do not have any prior work which solves the same problem, this experiment will act as a baseline against which all further experiments will be measured.

4.2. Experiment 2: Using PCA

Main purpose: To perform PCA to group such features together to avoid effect of correlation between variables. Using 5 features.

Dataset: Original Crime dataset with data filtered for years 2013 to 2018 (45067 data samples)

Baseline: Comparable to Experiment 1 or any other better performing experiment.

4.3. Experiment 3: Truncating the data set based on 'Year' feature

Main purpose: To observe the effect of reducing dataset on accuracy, we included data for 3 years, then for 2 years and finally for just 1 year.

Dataset: Original Crime dataset with data filtered for year 2017 (1042 data samples)

Baseline: Comparable to Experiment 1 or any other better performing experiment.

4.4. Experiment 4: Using One Hot Encoding

Main purpose: To observe the effect of enumerating the categorical variable using different method, we decided to use One Hot Encoding and then run models on the modified data.

Dataset: One hot encoded version of the Original Crime dataset with data filtered for years 2017 (1042 data samples)

Baseline: Comparable to Experiment 1 or any other better performing experiment.

4.5. Experiment 5: K-means clustering (Unsupervised Learning)

Main purpose: To determine the severity using clustering by feature or seat of features

Dataset: Original Crime dataset with data filtered for year 2013 to 2018 (45067 data samples)

Baseline: The results of this experiment will be compared to best performing experiment for *Method 1*

4.6. Experiment 6: K-means clustering followed by logistic regression

Main purpose: To improve the crime severity of clusters by performing Logistic Regression by using clusters as a feature for this algorithm.

Dataset: Original Crime dataset with data filtered for year 2013 to 2018 (45067 data samples)

Baseline: The results of this experiment will be compared to best performing experiment for *Method 1*

4.7. Experiment 7: K-means clustering followed by Ridge Classifier

Main purpose: To improve the crime severity of clusters by performing Ridge Classification by using clusters as a feature for this algorithm.

Dataset: Original Crime dataset with data filtered for year 2013 to 2018 (45067 data samples)

Baseline: The results of this experiment will be compared to results of *Experiment 5*

Evaluation Metrics:
For all the experiments, we will calculate 'Accuracy','Precision', 'Recall' and 'F1-Score' for this experiment. For the crime severity measurement, it is important to have high 'Accuracy' and 'Recall' than having high precision. Hence these two will be selected as evaluation metrics.

5. Experimental Results

5.1. Experiment 1: Using all features in dataset

We got accuracy of 73.36 using important features (mentioned in section 3.1)

Table 2. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
73.91	0.66	0.74	0.66

5.2. Experiment 2: Using PCA

Surprisingly accuracy and Recall reduced a bit after performing PCA which indicates that initial feature set was more appropriate than the reduced one.

Table 3. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
71.92	0.63	0.72	0.64

5.3. Experiment 3: Truncating the data set based on 'Year' feature

After considering data for just one year, accuracy reduced than that of the experiment 1 but the difference was very less. This might have happened because we still have enough data points even though it is for one year.

Table 4. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
73.36	0.63	0.73	0.65

5.4. Experiment 4: Using One Hot Encoding

Accuracy and Recall remained pretty much same but the computation time increased

Table 5. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
73.21	0.62	0.73	0.64

Observations and next steps for Method 1:
Classification models studied above gave fairly good accuracy. However, this method gives us severity of the highest probable threat at a particular location. More in-depth and accurate analysis can be obtained by computing a holistic

severity rating. We believe this can be achieved with unsupervised learning, by clustering the dataset. For each cluster, severity of the entire cluster is determined based on individual crime severity in that cluster. To explore this option, we added method 2 to our 'research methods' and conducted experiment 5 6.

5.5. Experiment 5: Using K means

We performed K-means clustering of the data and calculated crime severity. The Accuracy of this method was the best as compared to all the other experiments.

Table 6. Evaluation Metrics

Accuracy
75.95

5.6. Experiment 6: Using K means followed by Logistic regression

The accuracy of this method was 0.7364. We expected the results to be much more accurate.

Table 7. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
73.64	0.54	0.74	0.63

5.7. Experiment 7: Using K means followed by Ridge Classifier

The accuracy of this method was 0.8704. The results after using a classifier after clustering the data were very promising.

Table 8. Evaluation Metrics

Accuracy	Precision	Recall	F1-Score
87.04	0.76	0.87	0.81

Observations and next steps for Method 2:
As expected, the accuracy of this method was much better than supervised method. Given the short time span in which this experiment was conducted we couldn't tune the hyper parameters properly. We plan on optimizing this solution and hope to achieve higher accuracy.

6. Conclusion

From the experiments conducted, we can see that Unsupervised approach using K-means clustering worked better than Random Forest classifier. Also, including a classifier after clustering the data provided an even higher accuracy and better performance metrics. Considering current and future work, we plan to include other types of offences like DUI in order to increase the usability of the solution. We will be able to develop a model which clusters Austin area into segments based on a relevant feature or combination of features and provides overall crime severity of the location. At any given location and time, user will be able to gauge the crime status and take precautionary measures accordingly.

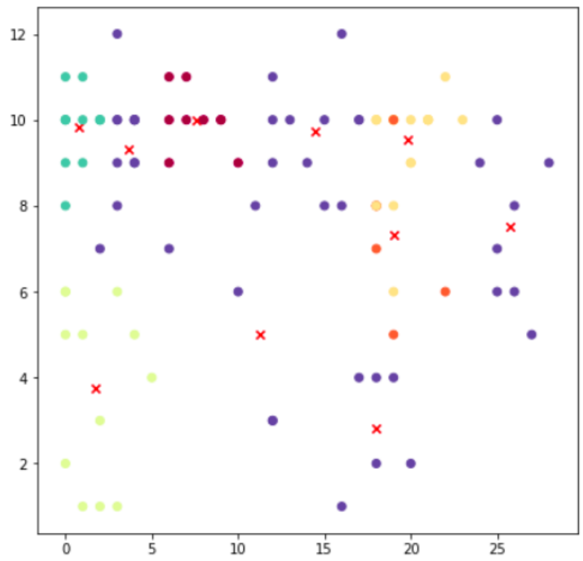

Figure 2. Plot of clustered data for 100 samples

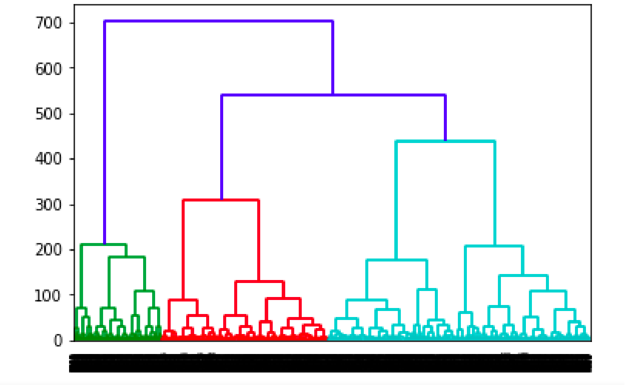

Figure 3. Plot of hierarchical data for entire dataset

References

Predicting-crime-in-toronto. URL https://github.com/7cb15/Predicting-Crime-in-Toronto. [Online; accessed 14-November-2018].

Bogomolov, Andrey, Lepri, Bruno, Staiano, Jacopo, Oliver, Nuria, Pianesi, Fabio, and Pentland, Alex. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 427–434. ACM, 2014.

Bollen, Kenneth A and Ward, Sally. Ratio variables in aggregate data analysis: their uses, problems, and alternatives. *Sociological Methods & Research*, 7(4):431–450, 1979.

Greenberg, Stephanie W, Rohe, William M, and Williams, Jay R. Safety in urban neighborhoods: A comparison of physical characteristics and informal territorial control in high and low crime neighborhoods. *Population and Environment*, 5(3):141–165, 1982.

Handbook, UCR. Uniform crime reporting (ucr) summary reporting. *Retrieved August*, 27:2009, 2004.

Kang, Hyeon-Woo and Kang, Hang-Bong. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, 12(4):e0176244, 2017.

Moineddin, Rahim and Urquia, Marcelo Luis. Regression analysis of aggregate continuous data. *Epidemiology*, 25 (6):929–930, 2014.

Najjar, Al-ameen. A study on public safety prediction using satellite imagery and open data. 2017.

Osgood, D Wayne. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*, 16(1):21–43, 2000.

Santos, Rachel Boba. *Crime analysis with crime mapping*. Sage publications, 2016.

Vaidya, Omkar, Mitra, Sayak, Kumbhar, Raj, Chavan, Suraj, and Patil, Rohini. Comprehensive comparative analysis of methods for crime rate prediction. 2018.

Wortley, Richard and Townsley, Michael. *Environmental criminology and crime analysis*, volume 18. Taylor & Francis, 2016.