Crime Prediction Using Machine Learning

Riya Rahul Shah
California State University, Sacramento
shahriya1995@gmail.com

Abstract

The objective of this project is to tackle a vital issue in the society - Crimes. Analyzing and examining of crimes happening in the world will give us a Broadview in understanding the crime regions and can be used to take necessary precautions to mitigate the crime rates. Identifying Crime patterns will allow us to tackle problems with unique approaches in specific crime category regions and improve more security measures in society. Current studies show the reason of increase in crime rates is more in areas that are economically backward. In few decades' property crime will be a target. Physical hardware like tv sets, mobiles continue to be a target for thefts.

The following approach involves predicting crimes classifying, pattern detection and visualization with effective tools and technologies. Use of past crime data trends helps us to correlate factors which might help understanding the future scope of crimes.

Keywords: crime prediction, classification, SVM, KNN machine learning, analysis

1. Introduction

Vancouver is most populated city in Canada. It is most ethnically diverse cities in Canada. Crime is one of the biggest and dominating problem in our society and its prevention is an important task [1]. Even though Vancouver known to be the safest city it is observed that vehicle break-ins and many more thefts is still a problem.

There has been tremendous increase in machine learning algorithms that have made crime prediction feasible based on past data. The aim of this project is to perform analysis and prediction of crimes in states using machine learning models. It focuses on creating a model that can help to detect the number of crimes by its type in a particular state. In this project various machine learning models like K-NN, boosted decision trees will be used to predict crimes. Area Wise geographical analysis can be done to understand the pattern of crimes. Various visualization techniques and plots are used which can help law enforcement agencies to detect and predict crimes with higher accuracy. This will indirectly help reduce the rates of crimes and can help to improve securities in such required areas.

The following paper describes in brief the research work done previously and followed with details of datasets used along with data preprocessing. In next section data analysis and models are implemented with the prediction results. At the end conclusion is mentioned.

2. Related Work

In [1], crime prediction is done on Chicago data set in which various machine learning models are used. Comparison of models like KNN, Naïve Bayes, SVM is done this paper. It is seen that prediction varies depending upon the dataset and features that have been selected. The prediction accuracy found in [1] is 78% for KNN, 64% for GaussianNB, 31% for SVC.

Auto regressive integrated Moving average models was used in [2] to make machine learning algorithms to forecast crime trends in urban areas. One of the major problems in crimes is detecting and analyzing the pattern of crimes. Understanding datasets is also an important concept in this case. We surely want to accurately predict so that we don't waste our resources due to false signals.

In paper [3], Algorithms like KNN and neural networks are developed, tested and crime prediction is done on San Francisco.

It is observed that many machine learning models are implemented on datasets of different cities having unique features, so predictions are different in all cases. Classification models have been implemented on various other application like prediction of weather, in banking and finances also in security [3].

Most of the research in crime prediction is finding the location of crimes and doing analysis based on proposed area-specific models using geographical data. Based on the review and studying previous work, KNN classification and decision tree models is shown to be giving high accuracy so we choose to use the same to predict crimes in Vancouver city.

3. Data Collection

The dataset used is Crime dataset of the city of Vancouver available on Kaggle [4]. The dataset consists of crimes in Crime in Vancouver from 2003 to 2017 which consists of 530,652. It consists of features like type, year, month, day, hour, location, latitude, longitude and many more.

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y	Latitude	Longitude
0	Other Theft	2003	5	12	16.0	15.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
1	Other Theft	2003	5	7	15.0	20.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
2	Other Theft	2003	4	23	16.0	40.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
3	Other Theft	2003	4	20	11.0	15.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
4	Other Theft	2003	4	12	17.0	45.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
5	Other Theft	2003	3	26	20.0	45.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
6	Break and Enter Residential/Other	2003	3	10	12.0	0.0	63XX WILTSHIRE ST	Kerrisdale	489325.58	5452817.95	49.228051	-123.146610
7	Mischief	2003	6	28	4.0	13.0	40XX W 19TH AVE	Dunbar-Southlands	485903.09	5455883.77	49.255559	-123.193725
8	Other Theft	2003	2	16	9.0	2.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
9	Break and Enter Residential/Other	2003	7	9	18.0	15.0	18XX E 3RD AVE	Grandview-Woodland	495078.19	5457221.38	49.267734	-123.067654

Fig. 1. Dataset

4. Data Preprocessing

Initially we need to preprocess data by removing all null values and removing all columns that are unnecessary. Following fig shows the count null values analyzed while preprocessing.

TYPE	0
YEAR	0
MONTH	0
DAY	0
HOUR	54362
MINUTE	54362
HUNDRED_BLOCK	13
NEIGHBOURHOOD	56624
X	0
Y	0
Latitude	0
Longitude	0

(a)

TYPE	0
YEAR	0
MONTH	0
DAY	0
HOUR	0
MINUTE	0
HUNDRED_BLOCK	0
NEIGHBOURHOOD	0
X	0
Y	0
Latitude	0
Longitude	0

(b)

Fig. 2 - a) original having null values b) dataset preprocessed

4.1. Architecture

The proposed work is divided into 4 parts:

1. Data preprocessing
2. Data Analysis
3. Data Modelling
4. Evaluation of performance

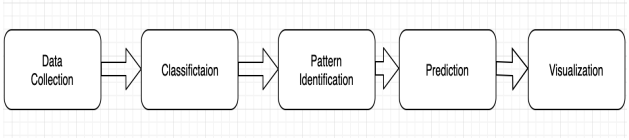

Fig. 3- Architecture

4.2. Prepare Data for Training and Testing

After data cleaning, we will use some preprocessing techniques for numerical and categorical data like normalization and one hot encoding.
Model sampling for train and split:

- Training dataset consists 70% or 80% data
- Testing dataset consist of 30% or 20% data

Once data is ready it can be trained using machine learning models.

5. Data Analysis

Following are some of the data analysis done to understand the dataset and problems we are actually having Figure 3 shows the number of incidents that have happened previously based on the type of crime.

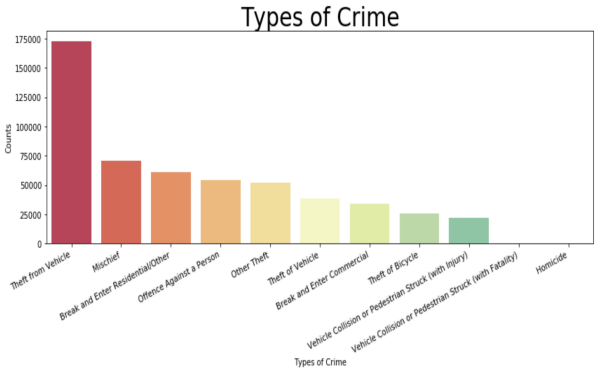

Fig. 4. Number of crimes

It is observed from figure 4 that most crimes are Theft from Vehicle.

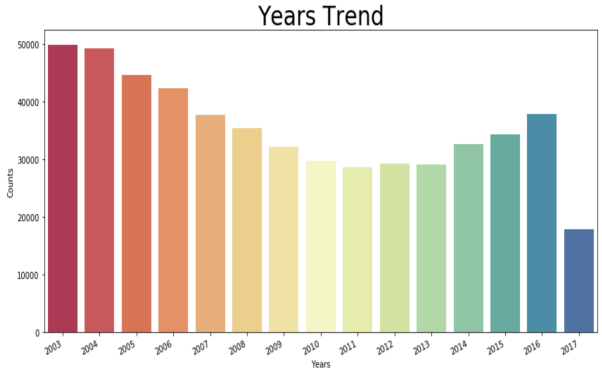

Fig.5. Crime trends per year

Figure shows trends of crimes each year It is seen that in year 2003 crime rate was highest in Vancouver city whereas the rate is quite decreased till year 2017. So it is a good indication that if we study more we can reduce the rate more.

6. Machine learning models

In this section various classification model's comparison is done to understand which model works best for our crime prediction.

One approach used to implement crime prediction is, Categorical Variables are encoded and then used for training in model. Crime type is our output (target). Since we are going to classify types of crimes, we are going to implement following machine learning models

A. K-Nearest Neighbor

This is one the simplest model, its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.
Using features Day, Date, Year of the crime using knn it is found to be 40% accuracy. Adding extra features and trying to improve accuracy is still under construction.

B. Logistic Regression

Logistic regression is one of the regression models where the variables dependent is either binary is categorical. It cannot handle continuous data.
Implementation under construction.

C. Decision Trees

Decision trees is a tree shaped graph which includes outcomes, utilities which helps in making decision.
Implementation under construction.

D. Support Vector Machine

In svm we find a hyperplane which separate two or more classes. It takes maximum time for processing.
Implementation under construction.

E. Bayesian Methods

The implementation is based on Naïve Bayes algorithm which constructs models as classifiers and is represented as vectors of all values of features.
Implementation under construction.

7. Data Visualization

Under construction
After implementing all scikit learn models. I will use matplotlib library to some data visualization and analysis of crime.

8. Results

The accuracy will be calculated using scikit learn function known as score_accuracy. We will import the metrics and then calculate the F1 score, accuracy, recall, precision of each models.

Model	F1 score	Accuracy	Recall	Precision
K-Nearest Neighbor				
Logistic Regression				
Decision Trees				
Support Vector Machine				
Bayesian Methods				

9. Conclusion

Under construction

10. References

[1] Alkesh Bharati, Dr Sarvanaguru RA.K ,"Crime Prediction and Analysis Using Machine Learning" in *International Research Journal of Engineering and Technology (IRJET)* ,Volume: 05 Issue: 09 | Sep 2018

[2] E. Cesario, C. Catlett, and D. Talia, "Forecasting crimes using autoregressive models*," IEEE 14th Intl. Conf. on Dependable, Auton. And Secure Comput.*, Auckland, New Zealand, Aug. 2016.

[3] M. V. Barnadas, Machine learning applied to crime prediction, Thesis, *Universitat Politècnica de Catalunya, Barcelona, Spain*, Sep. 2016.

[4] Kaggle Inc , Data of crimes in Vancouver (Canada) from 2003 to 2017
https://www.kaggle.com/wosaku/crime-in-vancouver

[5] Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. *In Electronics, Communication and Aerospace Technology (ICECA)*, 2017 International conference of (Vol. 1, pp. 225- 230). IEEE.