Crime Type Prediction For Smart Policing using Machine Learning

Yash Gupta
Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA
(Dated: April 28, 2021)

The paper aims to predict the crime type in a locality based on latitude, longitude, time of day(morning, afternoon, evening, night), and other relevant features using the Los Angeles crime dataset. Finally, a safest path algorithm is designed to find the safest path between a source and destination similar to shortest fastest path algorithm in Google maps. Machine learning models such as K-nearest neighbors (K-NN), Logistic Regression, convolutional neural networks (CNN), and Random Forests Classification are capable to predict crimes efficiently. The following strategy entails using effective tools and technologies to predict crimes, classify patterns, and visualize data. The use of historical crime data trends allows us to correlate factors that may aid in understanding the scope of future crimes.

Keywords: Crime Classification, Smart Policing, K-NN, Random Forests, data mining, Machine Learning, Deep Learning, Safe route, analysis

I. INTRODUCTION

[3,4]Nowadays with the increasing population, continuously increasing rich-poor gap, and increased demand for resources the rate of criminal activities has increased significantly over the past few years and is predicted to be increasing in the coming years if stringent decisions are not enforced. There is an urgent need for the governments to implement intelligent policing systems to help control such criminal activities. Analyzing and examining crimes that occur around the world can provide us with a broad perspective on crime regions, which can be used to take necessary precautions to reduce crime rates. Identifying crime patterns can enable us to address issues using unique approaches in specific crime category regions, as well as improve societal security. Machine learning algorithms have experienced tremendous growth, allowing for crime prediction based on historical data. The goal of this project is to use machine learning models to analyze and predict crimes in states. It focuses on developing a model that can assist in detecting high/moderate/low crime zones. [3,4]Machine learning models such as K-nearest neighbors (K-NN), Logistic Regression, convolutional neural networks (CNN), and Random Forest Classification will be used to predict type(severity) of crime zones in this project. To understand the pattern of crimes, a detailed geographical analysis can be performed. Various visualization techniques and plots will be used to aid law enforcement agencies in better detecting and predicting crimes. This will indirectly aid in the reduction of crime rates and the enhancement of security in such critical areas.

This paper is organized into seven sections, these are as follows: section 3 explains data exploration(dataset). Section 4 explains the methodology, section 5 presents the model results, section 6 discusses the results and the findings, and section 7 concludes the complete paper.

II. LITERATURE REVIEW

The motivation behind the study[1] is the recent surge in the crime rate. The work suggests a methodology that may enable law enforcement to embrace innovative preventive technological measures that can accurately forecast crimes based on weather attributes, and allocate necessary resources based on crime type and location. Instead of using traditional features, it uses weather, and time information to identify crime attractors, crime generators, and the surrounding population in the city for crime anticipation models. Feature selection methods(best subset selection, forward stepwise selection, and backward stepwise selection) were employed to determine if only demographic attributes play an essential role, or if the weather-related elements have any impact on the major and minor crime events happening in New York City(Dataset: New York Police Department Dataset 2018). Examining findings from all three feature selection methods, it was found that the most important attributes are nighttime, Sunday, and cloud coverage. The work performed multiple machine learning classification models(AutoMLP, Decision Tree, Logistic Regression, Random Forest, Neural net, and SVM as traditional algorithms and then finally deep learning model) to check the confusion matrix and observe various performance metrics(Area Under Curve, Accuracy, and Cohen's Kappa). Compared to other machine learning models, the decision tree algorithm is highly recommended as it is easy to interpret and also has less computational complexity than deep learning. Statistical crime analysis identified that weather-related attributes play a marginal role in the prediction of a crime even though they seemed relevant based on feature selection methods. In my opinion, the results can be further improved by combining population density with current characteristics based on location and studying whether such a combination plays an important role in crime prediction.

The focus of this paper[2] is to find spatial and temporal criminal hotspots. For Denver, CO, and Los Angeles,

CA, it analyzes two different real-world crime data sets and provides a comparison through statistical analysis between them. The paper's motivation and uniqueness are to take into account three main elements of crimes data, which are the type of crime, the occurrence time and the crime location. Most other works merely predict the location of the crime with no information about its occurrence. The initial dataset had to undergo many data pre-processing operations. The attribute subset selection technique was used for dimensionality reduction. The military time system was used for data integration to standardize date and time in the dataset. Data features were mapped into 4-hour intervals for data transformation and discretization (6 new types). Then, to find frequent crime patterns in both cities, the Apriori algorithm was implemented (using an open source tool). Decision Tree and Naïve Bayesian classifiers were subsequently applied to help predict future crimes within a particular time in a specific location. It achieved 51% prediction accuracy in Denver and 54% prediction accuracy in Los Angeles. In my opinion, it is possible to improve feature selection. Studying the additional relationship between the average income of the neighborhood and the rate of crime may improve the reliability of the model. The results[2] of the suggested work can be very helpful in preventing crime by raising awareness among people to avoid visiting vulnerable places at odd times.

III. DATASET

The dataset[7] being used for the analysis is taken from public crime dataset repository of Los Angeles[5]. It has 2,82,226 rows of data with 14 columns(features). It stores fields such as date reported, date occurred, time occurred, area, area name, crime number, crime number description, status, status description, location, latitude, longitude, etc.

Date.Rptd	DR.NO	DATE.OCC	TIME.OCC	AREA	AREA.NAME	RD	Crm.Cd	CrmCd.Desc	Status	Status.Desc	LOCATION	Cross.Street	Location.1
03/20/2013	132007717	03/20/2013	2015	20	Olympic	2004	997	TRAFFIC DR #	UNK	Unknown	OXFORD	OAKWOOD	(34.0776, -118.308)
03/10/2013	130608787	03/10/2013	445	6	Hollywood	635	997	TRAFFIC DR #	UNK	Unknown	ODIN ST	CAHUENGA BL	(34.1113, -118.3336)
12/18/2013	131820260	12/18/2013	745	18	Southeast	1839	997	TRAFFIC DR #	UNK	Unknown	105TH ST	CROESUS AV	(33.9406, -118.2338)
10/18/2013	131817514	10/18/2013	1730	18	Southeast	1827	997	TRAFFIC DR #	UNK	Unknown	101ST ST	JUNIPER ST	(33.9449, -118.2332)
05/26/2013	130510483	05/25/2013	2000	5	Harbor	507	440	THEFT PLAIN - PETTY (UNDER $400)	UNK	Unknown	1300 W SEPULVEDA BL	NaN	(33.8135, -118.2992)

FIG. 1. Dataset Exploration

Before performing further analysis, following steps have to be taken(data cleaning):

1. Remove rows with any null values.
2. Drop irrelevant columns, and crime types(crime types that cannot be possibly controlled/prevented, for example traffic accidents) from the dataset.
3. Segregating the time of the day into morning, afternoon, evening, night.
4. Labeling the dataset as per the frequency of crime occurrence into 'low', 'Moderate', and 'High'. The threshold for the frequency count was set taking care of balancing these three categories in the dataset.

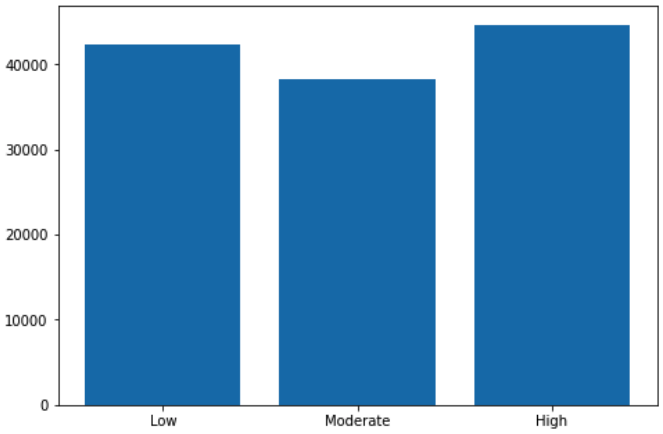

FIG. 2. Severity of crimes reported

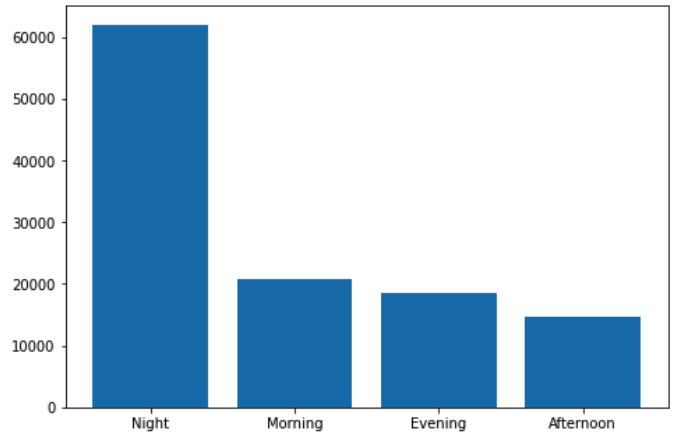

FIG. 3. Number of crimes reported at different times

Figure 3, and Figure 4 represents the number of crimes reported at different times and frequency of severity of crimes reported.

IV. METHODOLOGY

This section compares various classification models to determine which model works best for our crime prediction. [6]Categorical Variables are encoded and then used for model training is one approach used to implement crime prediction. Our output is crime type severity(target). Because we are going to classify different types of crimes, we will use the machine learning models listed below.

A. Logistic Regression

Logistic regression is one of the regression models in which the dependent variable is either binary or categorical. [6]It is incapable of dealing with continuous data. In our study, logistic regression has been implemented to predict the severity of the crime in a particular area at a particular time.

B. K-Nearest Neighbor

[8]This is the most basic model; its goal is to predict the classification of a new sample point by using a database in which the data points are separated into several classes. In our study, K-Nearest Neighbor has been implemented by assigning a value of 30 to the nearest neighbor factor(as after that the error rate is mostly decreasing, refer to figure 4). This value was determined experimentally by employing the elbow method of determining the optimal number of k.

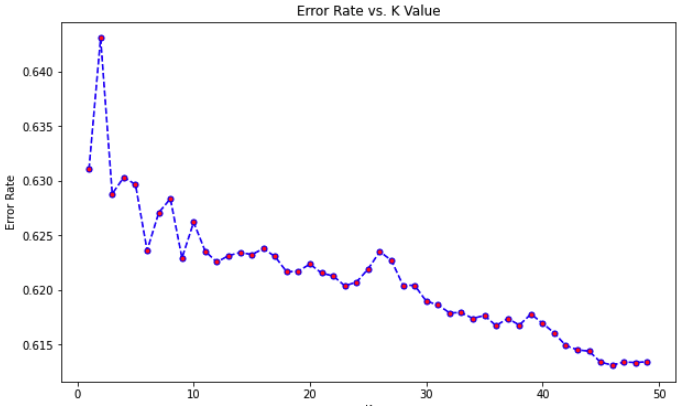

FIG. 4. Error Rate vs K

C. Convolution Neural Networks

[9]Convolutional Neural Networks, or CNNs, are a type of neural network that uses multiple layers to learn intelligent deep patterns from a given dataset, with each convolutional layer generally considering convolution, pooling, and activation operations. This neural network looks for local patterns of feature detection detected by the filters used in the convolution operation. In our study, four(4) dense neural network layers have been included(first with 128, second with 64, third with 32, and fourth with 3 neurons). The first three(3) layers have 'relu' as the activation function, and the last layer has 'softmax' as the activation function.

The architecture of the convolutional neural network(CNN) for multi-class classification has been depicted in figure 5.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	3712
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 32)	2080
dense_3 (Dense)	(None, 3)	99

Total params: 14,147
Trainable params: 14,147
Non-trainable params: 0

FIG. 5. Convolution Neural Network for multi-class classification

D. Random Forest Classification

[10]Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that works by training a large number of decision trees and then outputting the class with the most votes. In our study, sklearn Random Forest Classifier is being used to predict the severity of the crime(low,moderate,high) in a particular area at a particular time(morning, afternoon, evening, night).

E. Safest Path Between Source and Destination

The study introduces the concept of the "safest path" between source and destination. Google Maps, one of the most well-known navigation platforms, provides the shortest and fastest (lowest traffic) path between the source and the destination. In our study, the safest path is determined by giving preference (high priority or weight) to a route with a greater number of public places. For each alternate route, the experiment counts the number of public places (assuming that a route with more public places contributes to more public, light, and safety) and determines the safest route(one with most public).

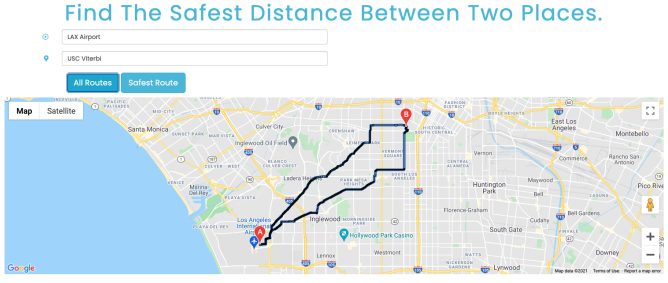

FIG. 6. All Available Paths Between Source and Destination

Figure 6 represents google maps displaying all possible paths between source and the destination, Figure 7 represents the safest path between the two points.

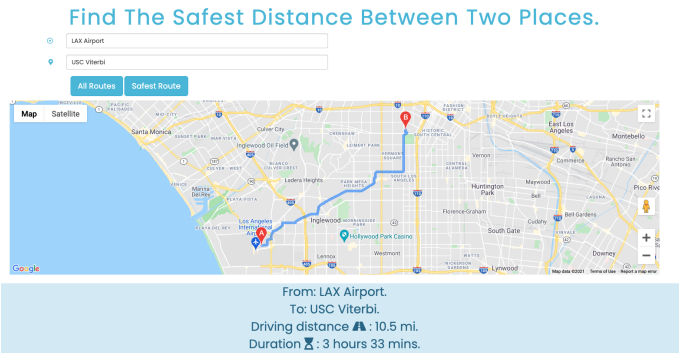

FIG. 7. Safest Available Path Between Source and Destination

V. RESULTS

The accuracy has been calculated with the scikit learn function score_accuracy. We will import the metrics and then compute the F1 score, accuracy, recall, and precision for each model. Table 1 represents the results obtained.

Model Name	Precision	Recall	F1-Score	Accuracy(%)
Logistic Regression	0.38	0.39	0.35	39
K-Nearest Neighbors	0.37	0.37	0.37	37
Convolutional Neural Networks	0.39	0.36	0.19	36
Random Forest Classification	0.70	0.70	0.70	0.70

TABLE I. Results

From the above results, it is clear that random forest performs the best of all the other models. The explanation for its superior output is that the random forests classification methodology considers the relationship between variables (if any), as well as the pattern between consecutive data records, whereas other models are less versatile and robust. The lack of more features to train the models, as well as a lot of sparsity in the dataset, is the second major explanation for the low accuracy in all of the above models. More features, such as weather, income, criminal, and victim age, may be gathered in pieces from various datasets and then mapped into one, but the task is extremely time consuming, and there is a high likelihood that all of these feature data will be incompatible(not coherent).

VI. DISCUSSION

This section discusses the possibility of biases in the dataset, the limitations and shortcomings of the proposed model, i.e. environments where it may fail to predict accurately, resulting in incorrect predictions. There is a high likelihood that a location is safe, but due to a higher number of crimes reported (complaints registered), that location was labeled as dangerous/high crime zone. Whereas a true high-crime zone area may be labeled as moderate or safe due to a low number of reported or registered cases. Another thing to keep in mind is that our assumption for the safest path (the path with the most public places is safe) may be incorrect. High/severe crimes are very likely to occur on the busiest roads. Furthermore, the model may result in unintentional discrimination against certain groups. For example, the model may develop a bias against members of a particular community and label areas where members of that community predominate as "unsafe." The safest path algorithm will always try to avoid those areas, which may have an impact on the businesses that operate there. As a result, such bias can have a negative impact on an entire community. As a result, many other factors must be considered before implementing this solution in a real-world scenario. This paper proposes a hypothetical model as the foundation for future advancement and development scope. Another enhancement that can be implemented is to map each location with the residents' income, education level, and other similar important factors that may lead our study to determine the type and severity of crime occurring in that specific location.

VII. CONCLUSIONS

The Los Angeles Crime and Collision dataset is analyzed using K-Nearest Neighbors, Logistic Regression, and Convolutional Neural Networks in the study. The dataset was pre-processed. Using the exact time of the crime, the crime data was labeled as 'Morning,' 'Afternoon,' 'Evening,' or 'Night' during this process. In addition, each crime zone was classified as 'Low,' 'Moderate,' or 'High,' based on the severity and frequency of crimes. It was discovered that the majority of crimes occurred at night and in high-crime areas. The next step was to one-hot encode the input features and output labels (0,1,2 classes depicting low, moderate, high crime zone areas). Finally, four different models (KNN, Logistic regression, and CNN, random forest classification) were implemented, and training accuracy, testing accuracy, F1 score, Recall, and Precision were calculated. It was discovered that Random Forests Classifier had an accuracy of 70 percent(approximately), whereas Logistic Regression, K-Nearest Neighbor(KNN) and Convolutional Neural Networks (CNN) had a much lower accuracy of approximately 40 percent. The explanation for the low accuracy is that the existing Los Angeles dataset lacks more features such as income, weather, age, and so on, and the dataset in its current form is very sparse, resulting in lower accuracy. In addition, using the Google Maps APIs, a safest path algorithm was implemented to determine the safest route between the source and destination. It is based on the assumption that a route with more public places is safer. It overrides Google Maps' basic functionality, which provides the shortest and fastest

path between the source and the destination. Towards end, the paper discusses about he possible shortcomings and challenges that may lead to failure of proposed solution and model. This paper proposes hypothetical models based on the given available dataset as a good strong foundation for future advancement and development scope.

ACKNOWLEDGMENTS

I express my deep gratitude to Prof. Marcin Jaroslaw Abram, Department of Physics and Astronomy for his valuable guidance and suggestions throughout the study.

I would like to extend my sincere thanks to my Teaching Assistant, Ms. Ninareh Mehrabi for her time to time suggestions to complete the study. I am also thankful to Ms. Supriya Devalla, and Mr. Pratik Singhavi, course producers and graders for providing me the necessary guidance and healthy learning environment to carry out my project work.

VIII. REFERENCES

[1] L. Elluri, V. Mandalapu and N. Roy, "Developing Machine Learning Based Predictive Models for Smart Policing," 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Washington, DC, USA, 2019, pp. 198-204, doi: 10.1109/SMARTCOMP.2019.00053.

[2] Almanie, T., Mirza, R. and Lor, E., 2015. Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050.

[3] 9. Perry, W. L. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. Rand Corporation.

[4] Ranson, M. (2014). Crime, weather, and climate change. Journal of environmental economics and management, 67(3), 274-302

[5] https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

[6] Section 4.1, Architecture, http://athena.ecs.csus.edu/ shahr/progress_report.pdf

[7] https://data.world/losangeles/lapd-crime-and-collision

[8] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[9] https://iopscience.iop.org/article/10.1088/1742-6596/1447/1/012021/pdf

[10] https://en.wikipedia.org/wiki/Random_forest