Predicting Crime Using Time and Location Data

Name: Md. Abdullah Al Shafi

***Abstract*—To be better prepared to respond to criminal activity, it is important to understand patterns in crime. In our project, we analyze crime data from the city of San Francisco, drawn from a publicly available dataset. At the outset, the task is to predict which category of crime is most likely to occur given a time and place in San Francisco. To overcome the limitations imposed by our limited set of features, we enrich our data by adding information from the United States Census to it and also attempt to make this classification task more meaningful by merging multiple classes into larger classes.**

I. Introduction

Criminal activities are regular occurrence all over the world. Governments spend huge amount of time to use technology to tackle criminal activities. Crime Analysis, a sub branch of criminology, studies the behavioral pattern of criminal activities and tries to identify the indicators of such events. Machine learning deals with data and by analyzing data machine learning helps to predict. Machine learning can learn and analyze previous crime datasets and can predict hotspots for the crime based on time. This technique is known as supervised classification.

In this project, we will use a dataset from San-Francisco OpenData which contains the reported criminal activities in the neighborhoods of the city San Francisco for a duration of 12 years. We will use different classification techniques to find hotspots of criminal activities based on the time of day. So, the aim of this project is to use machine learning techniques to classify a criminal incident by type, depending on its occurrence at a given time and location.

II. Motivation

Criminal and sociology scholars are studying the pattern of the criminal activity and its relation with the region. Researchers have shown that many criminal activities are happening in a specific region. This is called hotspot. Machine learning can be used to identify hotspots by data driven approach.

So, our motivation is to explore the different machine learning techniques to analyze the crime patterns so that this will help law enforcement agency to conduct their operations. If law enforcement agencies have a prior assumption of the class of the crime, it would give them tactical advantages and help resolve cases faster.

III. Background

Criminal activities are common around the word. So, researchers have done many works on this topic. Researches have been studying the relation between criminal activities and socio-economic variables like unemployment, income level, level of education etc.

Combining two datasets - 1990 US Law Enforcement Management and Administrative Statistics (LEMAS) and crime data 1995 FBI Uniform Crime Reporting (UCR) and applying classification techniques like Decision Tree and Naive Bayesian algorithm, 83.95% accuracy have been achieved when asked to predict a crime category for different states of USA [1]. Sadhana and Sangareddy [2] have used twitter data and sentiment analysis to predict crime in real time. They also used this data to map the concentration of crime occurrences and find large scale hotspots.

In preparation for this project, we came across various articles tackling crime prediction in various cities. Previous works in other cities, predicted the types of crimes occurring in the city based on the assumption that a crime has occurred [4]. We read that the best classifiers for this type of datasets tended to be tree-based methods, which allows for the decision-based classifications.

Additionally, we found other papers that corroborated the random forest model. They were able to achieve best accuracy across 39 different categories by using random forest and k-nearest neighbor [5]. The authors believed the random forest would be the best since the data was highly noisy. In other paper, Chandrasekar et. al. were able to achieve high accuracy Gradient Boosted trees and Support Vector Machines [4].

The authors propose an approach for detecting and identifying Crime using data-mining and machine learning techniques [6]. For crime detection, they have used k-means clustering, which iteratively generates two crime-clusters that are based on similar crime attributes. For crime identification and prediction, they have analyzed the data using K-Nearest Neighbor classification, by looking at the past crimes and finding similar ones that match the current crime based on number of nearest neighbors matched. To enhance k-means results, they have used GMAPI which embeds Google maps through Netbeans. Their model gives an accuracy of 93.62% and 93.99%, respectively, in the formation of two crime clusters using selected crime attributes.

The authors [7] have used the same finite set of features from the Communities and Crime Dataset for all the algorithm. For comparing the result five metrics are used to evaluate the effectiveness and efficiency of the algorithms. Among all the algorithms Linear Regression yields the best result. The concept of data mining and machine learning are used for identifying and analyzing patterns of crime [8]. For prediction, decision tree concept is used which is similar to a graph where internal node represents test on an attribute and each branch represents outcome of the test.Here using this algorithm the accuracy obtained is more than 90%.

Clifton Phua, Damminda Alahakoon, and Vincent Lee [9] proposes an innovative fraud detection method built upon existing fraud detection research to deal with data mining problem of skewed data distributions [9]. They used three algorithms together on the same skewed data, within the context of classification analysis, their strengths can be combined and their weaknesses reduced. Using these algorithm the accuracy obtained is more than 90%.

IV. Description of Dataset

Source: The experiment is conducted on a specific dataset. The dataset is provided by SF Opendata from SFPD Crime Incident Reporting System[3].

It provides information on crime incidents that occurred in San Francisco for the period of 1/1/2003 to 5/13/2015. The dataset is in a csv file containing 878049 rows.

A. Features

There are different attributes of the the dataset.The attributes are:

- Date
- Category
- Crime description
- Day
- PDistrict
- Address
- Latitude
- Longitude

B. Dataset Preprocessing

For preprocessing the dataset, we have used Tableur software and Python library *Scikit-learn (sklearn)*.

1. Some attributes in the csv files contains string values and others are numeric values. In order to use this dataset in machine learning models, the text features need to be converted into a numeric values. Python library *numpy* is used to contain both features and label of the dataset after converting them into numeric values.

2. Attributes with string data type are “Date”, “Category”, “Pdistrict” columns. By using python, we assigned numeric values for these features.

3. Datetime attribute is also a string data type, however this is converted into four different attributes and obtained from it: “Hour”, ”Date”, ”Month” and “Year”.

4. There is a specific pattern in occurrence of crime during different part of a day, as shown in previous analysis of different features. A new feature can be extracted by dividing a day in few different parts rather than considering 24 hours.

- Early morning: 1AM-7AM (1)
- Late morning: 8AM-1PM (2)
- Afternoon: 2PM-7PM (3)
- Night 8PM-12AM (4)

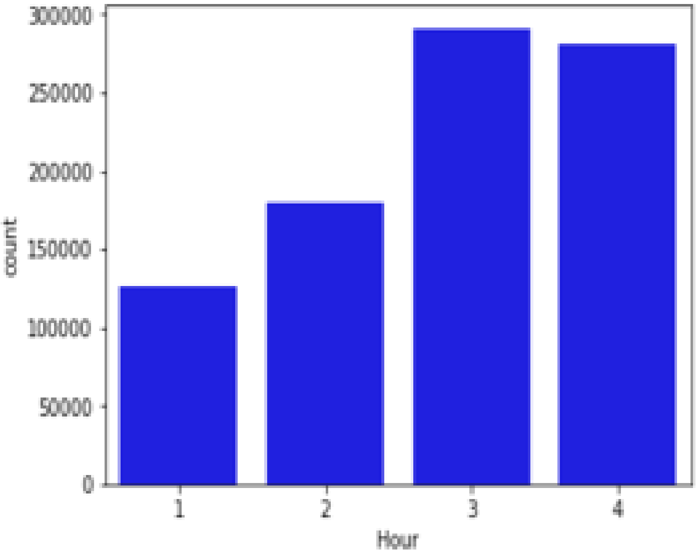

Fig. 1. Crime occurring in 4 different time range

In Fig. 1, there represents the number of crime occurring in different hour in a day. Hour is converted into 4 different parts.

V. Data Visualization

There are 39 types of crime in the San Francisco Crime Dataset. Top 5 categoies are:

- LARCENY/THEFT
- OTHER OFFENSES
- NON-CRIMINAL
- ASSAULT
- DRUG/NARCOTIC

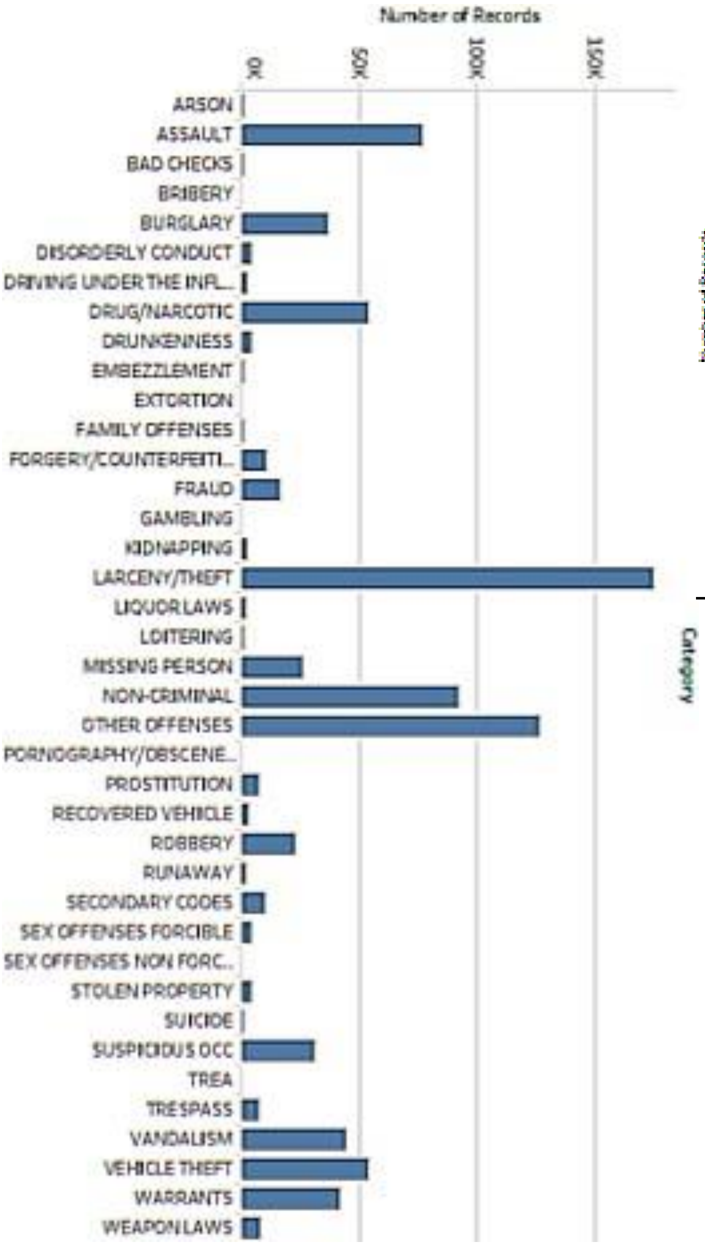

Fig. 2. Frequency of crime categories.

Fig. 2 shows there are few crimes that occur very frequently, and some crimes are really rare. Larceny/Theft is the most common and Trespassing(TREA) is the least common crime. There are assault, drug/Narcotic and vehicle theft are in high positions. The number of other offenses are in the position behind Larceny/Theft.

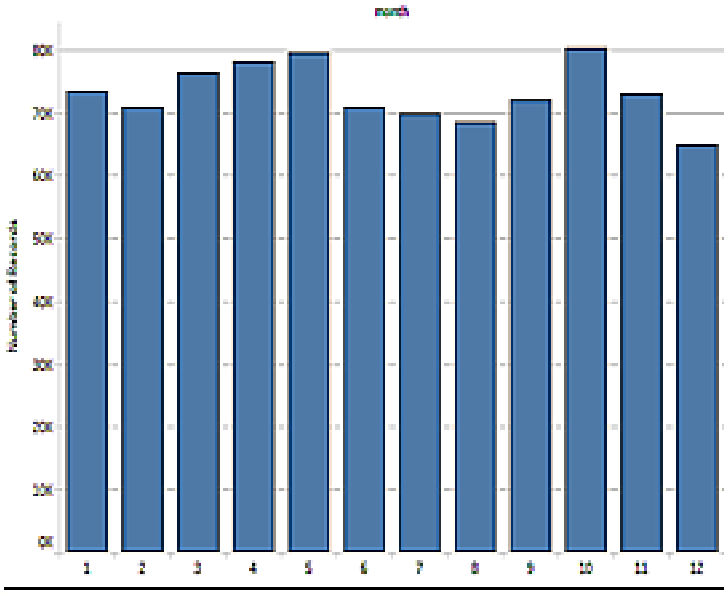

Fig. 3. Crimes occurring in different months.

Looking at the plots of fig. 3 reported criminal activities occurring throughout the year. Months are in x-axis and number of crime records in y-axis. January to December is represented by 1 to 12. It is apparent that more crimes occur in October and August is the month of less crime.

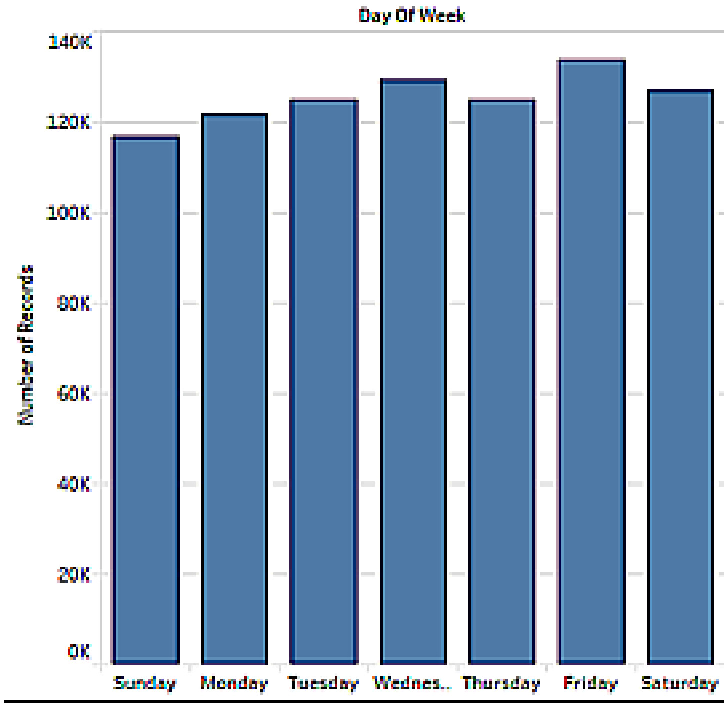

Fig. 4. Crimes occurring in different days of the week.

In fig. 4, most crimes occur on Friday, least crimes occur on Sunday. Crime rate almost gradually increases from

Monday to Thursday.

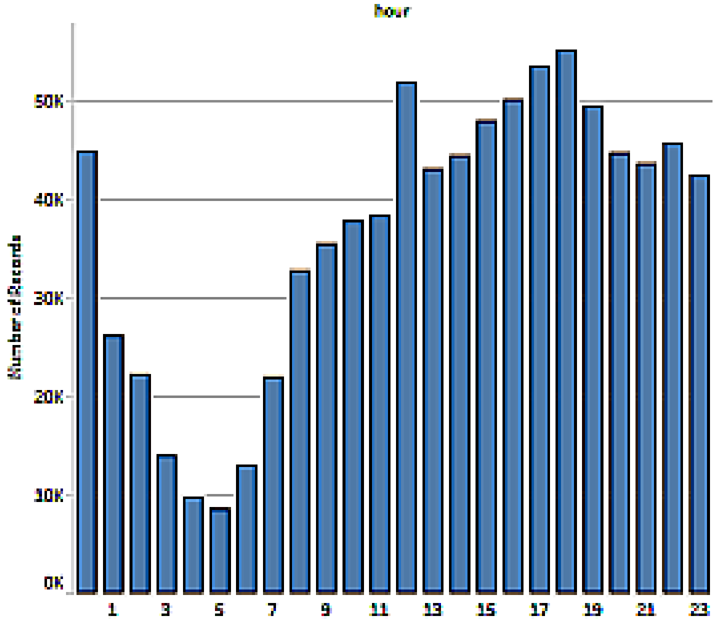

Fig. 5. Criminal activities occurring in different hour of day (in 24 hours format).

In fig. 5, most crimes occur during afternoon-evening. There is an increase of criminal activities at 6 PM and 8 PM. At 5 AM, the number of crime occurrence is least.

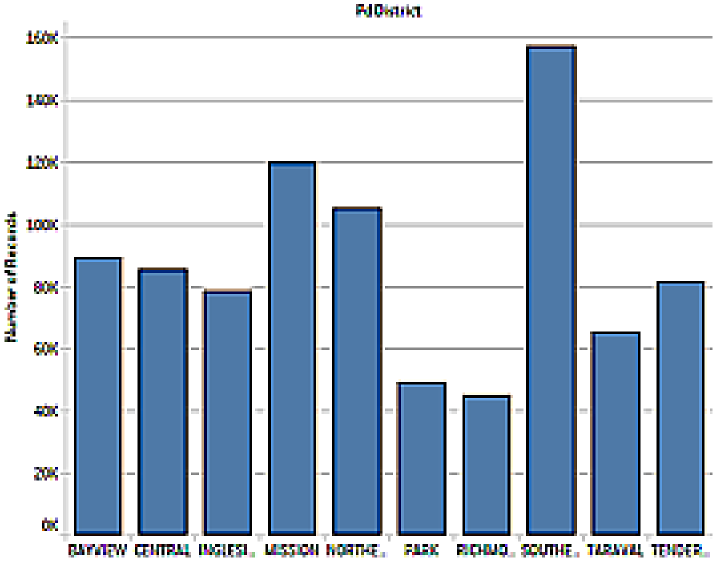

Fig. 6. Crimes occurring in different police district.

Fig. 6 presents the graph of crimes in different police districts. Among the ten police districts, criminal activities in the southern district is higher than any other district and Richmond is lower.

VI. Methodology

- **Cross Validation:** Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
- **Accuracy:** Accuracy measures how many predictions are matched exactly with the actual or true label of the testing dataset and returns the percentage of correct results.
- **Log-loss:** Log-loss is used to measure performance of classifiers by penalizing false classifications. Therefore, smaller value of log loss means the classifier is more accurate. The log loss value of a perfect classifier is 0. The classifier assigns a probability of predicting each class rather than picking the most probable class.
- **Confusion Matrix:** Confusion matrix returns a table layout that helps to visualize the performance of an algorithm rather than producing a numerical value that indicates the goodness of the algorithm.
  Confusion matrix is an X × Y matrix where X dimension indicates the true classes and Y dimension indicates the predicted classes. The value of X and Y is equal, and they indicate the number of classes in a dataset.

VII. Algorithm

Supervised classification models are applied on San Francisco Crime Dataset to predict the category of a crime incident. In the following part, performance of different classification models are discussed.

- **Logistic Regression:** Logistic regression is mainly used when dependent variable is binary. Logistic regression can work on both binary and multiclass problems. Logistic regression uses linear boundaries to classify data into different categories. Logistic regression can work on both binary and multiclass problems. For multiclass dataset, one vs the rest scheme is used. In this method, logistic regression trains separate binary classifiers for each class. Meaning, each class is classified against all other classes, by assuming that all other classes is one category. We use this algorithm in our project.

  sklearn.linear_model.LogisticRegression class is used for this model. The parameter multi_class is set to ovr which provides one vs the rest scheme, as multi class model is needed.

  TABLE I
  Logistic Regression Result

Algorithm	Accuracy	Log-loss
Logistic regression	20.054%	2.662

  The accuracy of logistic regression classifier is 20.054% with a better log-loss 2.662.

- **Decision Tree Classifier:** Decision tree datasets build a tree on the dataset. At each step, the tree is split into several trees until it reaches the result. We will use decision tree classifier in our project.

Sklearn.tree module provides Decision Tree Classifier class.

TABLE II
DECISION TREE CLASSIFIER RESULT

Algorithm	Function	Accuracy	Log-loss
Decision Tree	gini	17.760%	24.506
Decision Tree	entropy	17.741%	24.449

The accuracy we get using entropy is 17.741%, while maintaining a log loss score of 24.449. Gini performs better with 17.760% accuracy.

- **K-Nearest Neighbors:** K-Nearest Neighbors method is used in both supervised and unsupervised learning. We used this algorithm in our project. K indicates the number of neighbors voting to classify a data point. The distance can be measured with various metrics.

  sklearn.neighbors module provides supervised nearest neighbors classification models using k nearest neighbors. Among different parameters of the class, n_neighbors indicates the value of k, metric indicates the metric used to measure the distance of neighbors.

TABLE III
K-NEAREST NEIGHBORS CLASSIFIER RESULT

n neighbors	Accuracy	Log-loss
5	16.901%	19.177
10	18.488%	14.017
20	19.622%	9.359
25	20.042%	8.137
50	20.848%	5.435
100	21.071%	3.998

Log Loss improves as neighbor numbers are increased. Here, the features give the best accuracy and log loss for 100 number of neighbor.

- **Gaussian Naive Bayes:** Gaussian Naive Bayes is a supervised classifier that uses naive assumption that there is no dependency between two features. This classifier is implemented by applying Bayesian Theorem.

  Sklearn.naive_bayes provides GaussianNB class.

TABLE IV
GAUSSIAN NAIVE BAYES CLASSIFIER RESULT

Algorithm	Accuracy	Log-loss
Gaussian Naive Bayes	20.369%	2.636

Although this classifier gives poor accuracy, Log loss measurement is relatively better in this model.

VIII. REDUCING CLASSES

A dataset set is imbalanced when the classification categories are not represented approximately equally. As seen in Fig. 3 the crime classes are imbalanced with highest frequency of 174900 and lowest frequency of 6. That's why, we need to reduce the classes which has lower frequency to become the dataset balanced.
Only top 4 crime categories are used in attempt to reduce the number of classes. All other crime classes are labeled by "other crimes".

- OTHER OFFENSES (0)
- ASSAULT (1)
- DRUG/NARCOTIC (2)
- LARCENY/THEFT (3)
- VEHICLE THEFT (4)

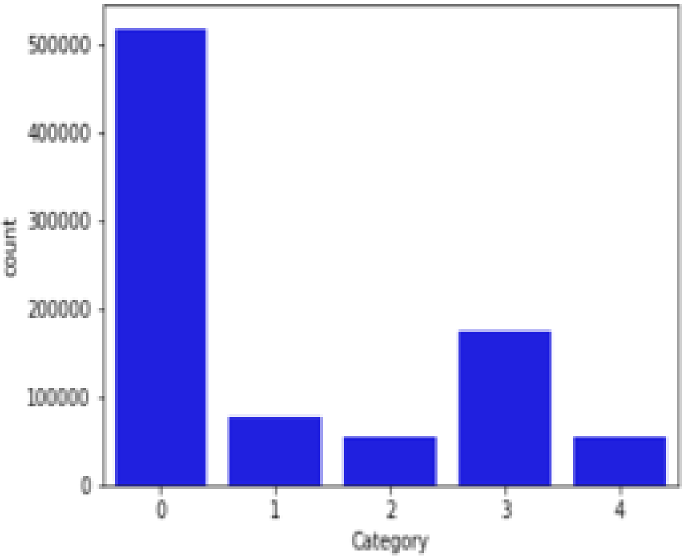

Fig. 7. Crimes frequencies for reducing classes

After reducing the classes, there are 5 categories of crimes. Fig. 7 presents the graph of crime frequencies for 5 classes. Among all classes other crimes are higher and all other classes are close to each other.

IX. RESULT

The result after reducing the classes is shown in the below table for all algorithms.

TABLE V
CLASSIFICATION RESULT FOR REDUCED CLASSES

Algorithm	Training set	Test set	Log-loss
Logistic regression	59.097%	58.950%	1.172
Decision Tree	59.812%	59.527%	1.103
K-Nearest Neighbors	59.192%	58.972%	1.187
Gaussian Naive Bayes	27.752%	27.665%	1.650

From the table V we get the accuracy and log-loss for every algorithm. The decision tree gives the best accuracy (59.527%) and best log-loss value (1.103).

X. Evaluation

- **Logistic Regression:**

TABLE VI
LOGISTIC REGRESSION RESULT FOR DIFFERENT CLASS WEIGHT

class weight	Training set	Test set	Log-loss
none	59.097%	58.950%	1.172
balanced	55.323%	55.203%	1.370

From Table VI gives accuracy and log-loss result for different class_weight. For none class_weight, this classifier gives better accuracy and log-loss than balanced class_weight.

- **Decision Tree Classifier:**

TABLE VII
DECISION TREE CLASSIFIER RESULT FOR DIFFERENT MAX DEPTH WEIGHT

max depth	Function	Training set	Test set	Log-loss
3	gini	59.097%	58.953%	1.143
5	gini	59.145%	58.999%	1.127
7	gini	59.347%	59.171%	1.110
9	gini	59.812%	59.527%	1.103
10	gini	60.111%	59.599%	1.119
3	entropy	59.132%	58.986%	1.145
5	entropy	59.156%	59.018%	1.127
7	entropy	59.218%	59.074%	1.111
9	entropy	59.522%	59.263%	1.109
10	entropy	59.798%	58.459%	1.127

Log-loss performance gets better with increasing max_depth for both kernels - gini and entropy. As well as the accuracy gets better. When max_depth=10, log-loss started increasing. That's why the better result is produced for max_depth=9. In our problem, we get better accuracy for gini rather than entropy.

- **K-Nearest Neighbors:**

TABLE VIII
K-NEAREST NEIGHBORS RESULT FOR DIFFERENT K VALUE WEIGHT

K	Training set	Test set	Log-loss
5	65.449%	55.039%	6.319
10	61.670%	57.012%	3.992
20	60.039%	58.216%	2.411
25	59.774%	58.445%	2.058
50	59.311%	58.906%	1.387
100	59.192%	58.972%	1.187

The test set accuracy is increasing as the value of k is increasing. However, the training set accuracy is decreasing. The log-loss result is improving for k. The best accuracy is 58.972% and log-loss is 1.187 for the value of k is 100.

XI. Confusion Matrix

Confusion matrix is used to visualize the results obtained from different classifier. The classes are indicated with the following integers:

- OTHERS OFFENSES (0)
- ASSAULT (1)
- DRUG/NARCOTIC (2)
- LARCENY/THEFT (3)
- VEHICLE THEFT (4)

True label indicates the true value of a given class, and predicted label indicates the prediction made of that class. True and predicted results are a match on coordinates such as (0,0), (1,1), (2,2) etc.

TABLE IX
CONFUSION MATRIX FOR LOGISTIC REGRESSION

		Predicted Label				
		0	1	2	3	4
Actual label	0	155282	0	0	0	8
	1	23212	0	0	0	0
	2	16215	0	0	0	0
	3	52624	0	0	0	1
	4	16073	0	0	0	0

TABLE X
CONFUSION MATRIX FOR DECISION TREE CLASSIFIER

		Predicted Label				
		0	1	2	3	4
Actual label	0	149858	22	615	4642	153
	1	22620	4	58	511	19
	2	15397	2	657	153	6
	3	46370	5	61	6083	106
	4	15341	2	10	519	201

TABLE XI
CONFUSION MATRIX FOR K-NEAREST NEIGHBORS

		Predicted Label				
		0	1	2	3	4
Actual label	0	154725	0	13	552	0
	1	23153	0	1	58	0
	2	16168	0	17	30	0
	3	52023	0	2	600	0
	4	16026	0	0	47	0

TABLE XII
CONFUSION MATRIX FOR GAUSSIAN NAIVE BAYES

		Predicted Label				
		0	1	2	3	4
Actual label	0	59835	0	95438	0	17
	1	9235	0	13975	0	2
	2	3178	0	13037	0	0
	3	16748	0	35874	0	3
	4	7709	0	8361	0	3

The confusion matrix shows that even though accuracy and log-loss function give good result after reducing classes, most classes are being predicted to be in class 0 (others) as that is the largest class. It also shows that, logistic regression are predicting 0 for most data points. As class 0 is quite larger than the rest selected top crimes, the dataset still remains imbalanced.

XII. Future Work

In our research, although we were successful in applying our algorithms to the dataset, we achieved a medium level accuracy. We aim to improve this accuracy rate by experimenting with our features. The dataset may require further cleaning so it fits our algorithm better. Using oversampling and undersampling on the dataset may help us to gain the desired level of accuracy.

XIII. Conclusion

We experimented with different type of algorithms to see which yields the most accurate result for. We were able to apply four different algorithms with moderate performance. Although the accuracy rate isn't very high we are satisfied with the result as the purpose of this research is to observe how different algorithms performs.

References

[1] Iqbal R. Murad, M.A.A. Mustapha, A. Panahy P.H.S, and Khanahmadliravi N., "An experimental study classification of algorithms for crime prediction," Indian Journal of Science and Technology , vol. 6(3), pp. 4219–4225, 2013.

[2] Sadhana C.S, ”Survey on Predicting Crime Using Twitter Sentiment and Weather”, Data israce, 2015.

[3] ”San Francisco Crime Classification.” San Francisco Crime Classification — Kaggle . N.p., n.d. Web. 16 Apr. 2017.

[4] Addarsh Chandrasekar, Abhilash Sunder Raj, and Poorna Kumar, ”Crime Prediction and Classifcation Using Machine Learning”, 2015.

[5]Christian Tabedzki, Amruthesh Thirumalaiswamy, and Paul van Vliet, ”Yo Home to Bel-Air: Predicting Crime on The Streets of Philadelphia”, 2018.

[6]D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in India using data mining techniques," Ai and Society, vol. 30, no. 1, pp. 117–127, Jan 2014.

[7] Mcclendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," Machine Learning and Applications: An International Journal, vol. 2, no. 1, pp. 1–12, 2015.

[8] Shiju Sathyadevan, Devan M.S,Surya Gangadharan, ”Crime Analysis and Prediction Using Data Mining”, International Journal of Computer Applications, 21(1):1–6,August 2014.

[9] Clifton Phua, Damminda Alahakoon, and Vincent Lee, ”Minority Report in Fraud Detection: Classification of Skewed Data”.