Hindawi
Complexity
Volume 2022, Article ID 4830411, 13 pages
https://doi.org/10.1155/2022/4830411

Research Article

Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques

Muzammil Khan,[1] **Azmat Ali**,[2] **and Yasser Alharbi**[3]

[1]*Department of Computer & Software Technology, University of Swat, Swat, Pakistan*
[2]*School of Computer Science, Wuhan University, Wuhan, China*
[3]*College of Computer Science, University of Hail, Hail, Saudi Arabia*

Correspondence should be addressed to Muzammil Khan; muzammilkhan86@gmail.com

Received 3 August 2021; Accepted 25 January 2022; Published 25 February 2022

Academic Editor: Gonzalo Farias

The crime is difficult to predict; it is random and possibly can occur anywhere at any time, which is a challenging issue for any society. The study proposes a crime prediction model by analyzing and comparing three known prediction classification algorithms: Naive Bayes, Random Forest, and Gradient Boosting Decision Tree. The model analyzes the top ten crimes to make predictions about different categories, which account for 97% of the incidents. These two significant crime classes, that is, violent and nonviolent, are created by merging multiple smaller classes of crimes. Exploratory data analysis (EDA) is performed to identify the patterns and understand the trends of crimes using a crime dataset. The accuracies of Naive Bayes, Random Forest, and Gradient Boosting Decision Tree techniques are 65.82%, 63.43%, and 98.5%, respectively, and the proposed model is further evaluated for precision and recall matrices. The results show that the Gradient Boosting Decision Tree prediction model is better than the other two techniques for predicting crime, based on historical data from a city. The analysis and prediction model can help the security agencies utilize the resources efficiently, anticipate the crime at a specific time, and serve society well.

1. Introduction

Data mining is the knowledge discovery process used to collect and analyze a large dataset and summarize it with helpful information. It is critical in different fields of science to serve analytical purposes and plays an essential role in human life and fields such as education, business, medicine, health, and science. Data mining is an attractive process of discovering a valid, understandable, helpful pattern and valuable information in large amounts of data [1]. The main goal of data mining is to find out fascinating and concealed knowledge in the data and summarize it in a significant form [2–4]. Similarly, the results should be in the form that conveys the inside information effectively [5–7]. Therefore, classification techniques are among the most important and commonly used techniques in data mining, and supervised class prediction techniques allow nominal class labels for predictions [8].

San Francisco is one of the largest cities in the United States of America. Therefore, it is vital to understand the pattern of crimes to ensure the safety of the citizens. San Francisco Crime Classification is an open-source dataset available for an online competition administrated by Kaggle Inc. The main task in the dataset is to predict the crime category based on a given set of geographical and time-based variables. The limited and constrained police resources prove insufficient to handle the city's law and order issues. Therefore, it is vital to study and understand the distribution of different types of crimes in the city based on the occurrence time and the location for security agencies to channelize resources efficiently. Naive Bayes, Random Forest, and Gradient Boosting Decision Tree are used for prediction and classification of crimes into two types of violent and nonviolent crimes.

In this paper, the main goal is to propose a prediction model that predicts crime based on past criminal records.

The proposed model contains three techniques and performs evaluation through accuracy, precision, and recall evaluation matrices. The data is descriptively analyzed and statistical crime distribution over space and time is visualized to help attain potential patterns. The features are extracted from the original dataset, and the classification is performed using Naive Bayes, Random Forest, and Gradient Boosting Decision Tree techniques. The experimental results show that the Gradient Boosting Decision Tree prediction model is better than the other two techniques for predicting crime, based on historical data from a city. The analysis and prediction model can help the security agencies utilize the resources efficiently, anticipate the crime at a specific time, and serve society well. Conclusions of the study and future directions for further research are presented in the last section of the paper.

2. Related Work

Data mining has been frequently used in crime prediction models for the last couple of years, considering different features. Yehya used variables such as longitude (X), latitude (Y), address, day of week, date (YYYY-mm-dd: hh : MM : ss), district, resolution, and category to analyze and predict San Francisco crime data. The study used different techniques and principal component analysis to classify the accuracy and avoid overfitting. He also used four different classifiers: K-NN, XGB Decision Tree, Bayesian, and Random Forest, applied them to the task, and obtained the log-loss of 2.39031 by the Random Forest classifier [9]. Wenbin Zhu et al. conducted an experiment for the classification of crime based on the San Francisco dataset. According to their explanation, it was mentioned that crime classification helps police to keep the city safe. They predicted crime categories based on time and location. They used Naive Bayes, logistic regression, and the Random Forest as baseline classifiers with best prediction results [10]. Umair Saeed et al. experimented with data mining techniques to identify and predict crimes and compared the experiment results of Naive Bayes and Decision Tree classifiers. They observed that the Naive Bayes classifier performed better and accurately predicted crime prediction [11]. Somayeh Shojaee et al. conducted an experimental study for crime prediction using supervised classification learners. They used two different feature selection methods executed on real crime datasets. They compared these two methods based on AUC (i.e., Area Under the Curve) values. They found that Naive Bayes, K-Nearest Neighbor (KNN), and Neural Networks are better classifiers against Decision Tree (J48) and Support Vector Machine (SVM). The Chi-square feature selection technique is used in their experiment for the performance measurement of the classifiers. The investigation is conducted in a RapidMiner environment to enhance the quality of crime mining [12]. Junbo et al. predicted crime categories from 2003 to 2015 surrounding San Francisco city based on a dataset derived from SFPD Crime Incident Reporting System. They investigated Naive Bayes, K-NN, and Gradient Tree Boosting classification models and analyzed their advantages and disadvantages on that prediction task. According to their results, Naive Bayes did not perform as a perfect model for that task because some features did not represent the count or frequency. On the other hand, K-Nearest Neighbor improved the prediction result to a large extent. Gradient Tree Boosting performed as the best model in their experiment, but it was slightly slow. Gradient Tree Boosting model generated a score of 2.39383 and was ranked 93 among 878 teams [9]. R. Iqbal et al. (2013) conducted an experimental study for the classification algorithms. They experimented with the prediction of crime categories for the different states of USA. They compared Naive Bayes with the Decision Tree classifier for crime prediction. Naive Bayes achieved 70.81% accuracy and the Decision Tree classifier achieved 83.95% accuracy, which shows that the Decision Tree classifier performs better for the crime classification problems [13].

3. San Francisco Dataset

The study uses a dataset from Kaggle to build up the model [2]. The dataset (training set/data) has different attributes, each having a different connection. The training dataset contains the incidents taken from Kaggle on San Francisco crimes. The data ranges from January 2003 to May 2015. The dataset contains almost 12 years of criminal reports from San Francisco. The dataset has classified categories of all crimes, which contain different crime types. The training set consists of 878049 observations and the testing set consists of 884263 observations. The dataset is used to check the accuracy of the classification techniques with new unclassified data. The training set consists of nine variables as shown in Table 1.

The study arbitrarily mixes the original training dataset and divides it into a training dataset and testing dataset with 80% and 20% sizes, respectively.

3.1. Exploratory Data Analysis. A simple script is run and explores several unique categories of crimes in the dataset, and 39 different crime categories are identified. The figure also shows the distribution of crime and change in the type of crime since 2003. For example, from the below plot, larceny/theft is the most common type of crime. Further, there appears to be a skewness in the type of crimes. For example, there have been 174,900 incidents of larceny/theft, whereas there have been only 6 of TREA since 2003.

From Figure 1, it is found that the top 10 crimes are larceny/theft, other offenses, noncriminal, assault, drug/narcotic, vehicle theft, vandalism, warrants, burglary, and suspicious OCC, accounting for 83.5% of the whole records statistically [10]. It is reasonable to suggest allocating more police resources to deal with these crimes as they are more likely to occur.

Figure 2 indicates that the lower overall density of sex offenses compared to the other categories of crime is expected, as there are fewer crimes of this category in the data. The overall structure here indicates the aggregate with the most prominent hot spot in the north area centered in the

Table 1: Selected features for analysis.

Attributes	Descriptions
Dates	Date is the timestamp of the moment when the crimes occurred
Category	Category shows the crime category
Description	Description shows the short description of the crime
DayOfWeek	DayOfWeek shows the day on which the crime occurred
PdDistricts	PdDistricts shows the district of the city where the crime was committed
Resolution	Resolution shows a short description of the crime resolution
Address	Address shows the address of the crime where it was located.
X	X shows the latitude of the crime position
Y	Y shows the longitude of the crime position

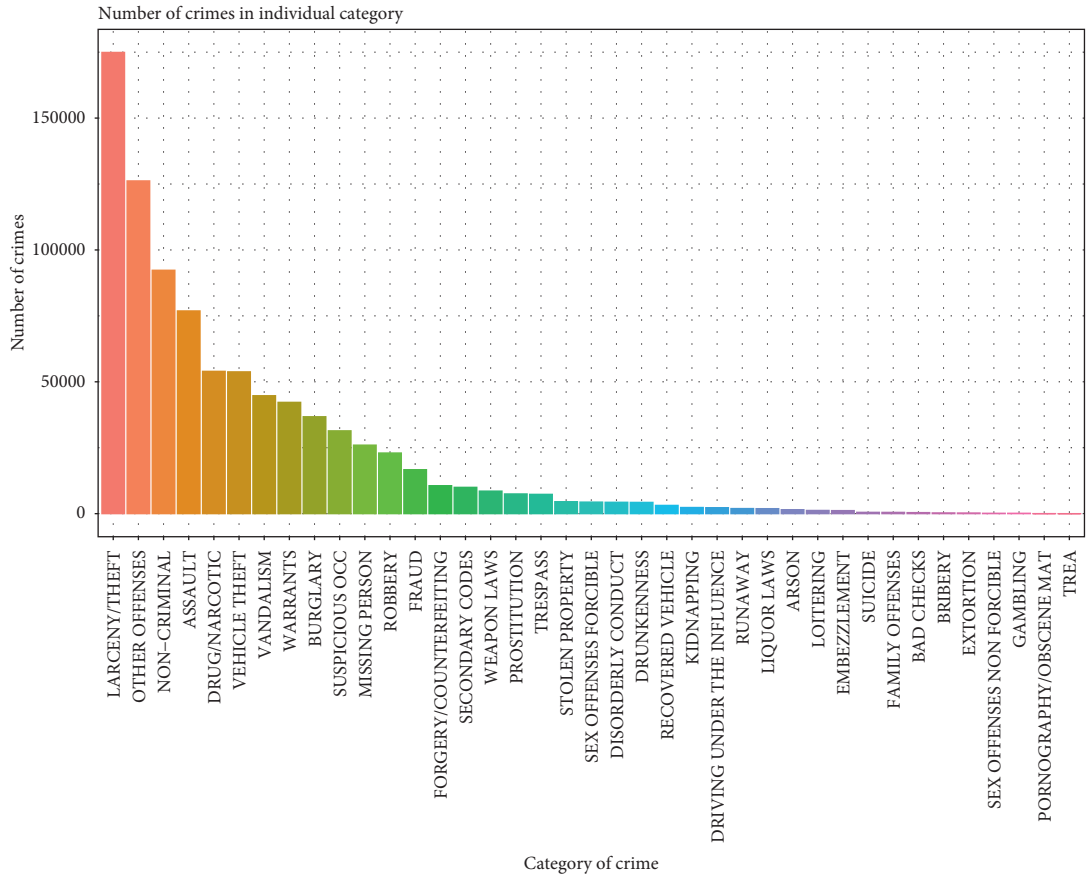

Figure 1: Number of crimes in individual category.

Tenderloin neighborhood area. Larceny/theft, other offenses, noncriminal, and assaults seem to be more concentrated on the map. However, four crimes seem to cover a larger area: vehicle theft, vandalism, burglary, and suspicious OCC. At the same time, other crimes come into a smaller area with larger density crime. It is interesting to explore further other columns of the dataset to help us extract useful features. What are the distributions for day of week, hour, month, and even year for the record of the crimes? We visualize how their occurrences alter with year, month, day of week, and hour for the ten most occurring crimes.

Figure 3 shows interesting figures and results based on years. This map reveals the increase or decrease in the top ten crimes in different years in San Francisco from 2003 to 2015.

Figure 4 shows monthly reports of the top ten crimes in San Francisco, revealing the expansion and reduction of crime month-wise. However, the interesting point is that all crimes (top 10) are increased after three months and also decreased after three months, which reveals that the top ten crimes in the San Francisco area based on seasonal pattern increased in the 3rd month (March) with same pattern in Spring, decreased in the 6th month (June) with the same pattern in Summer, and increased again in September, Autumn.

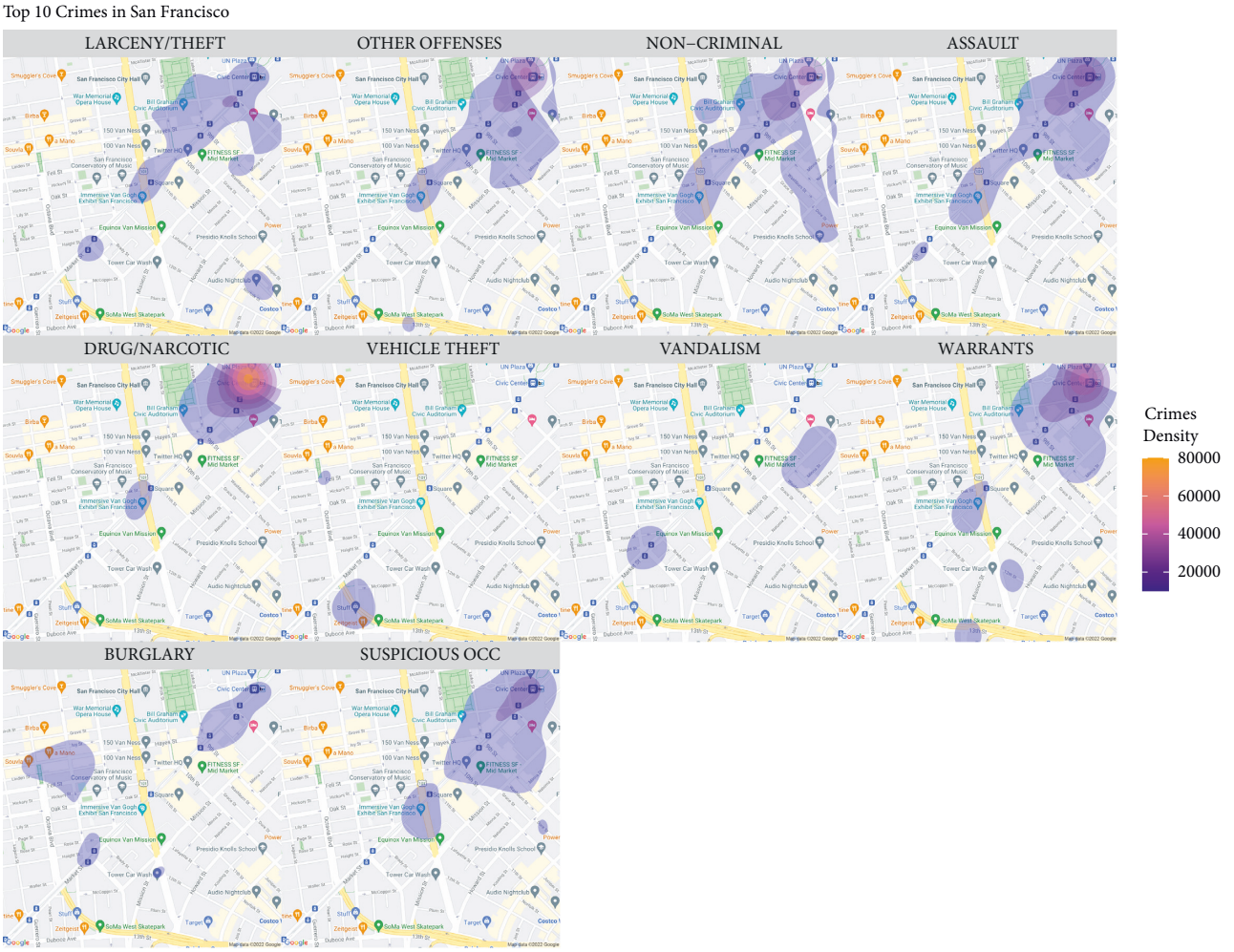

Figure 2: Top 10 crimes' density in San Francisco.

Figure 5 shows the top ten crimes' ratio (increase or decrease) for days of the week. The crime is more concentrated in northern areas on Friday, Saturday, and Wednesday. Larceny crime, vehicle crime, and vandalism crime increased on Friday and Saturday with the same pattern, while the rate of suspicious OCC crime occurred and increased on Friday and Wednesday. Burglary crime increased on Friday, and assault crime increased on Saturday and Sunday. Drug/narcotics and warrants crime occurred and increased on Wednesday. All these crimes indicate the ratio and occurrence of crime in San Francisco based on days (days of weeks).

Figure 6 shows the aggregate of the crime and the crime rate in each hour. In this graph, the results suggest that all the top ten crimes decreased between 3 : 00 AM and 6 : 00 AM but reached their second peak at midnight and the first peak around 5 : 00 PM to 6 : 00 PM. So, when police resources are limited, our suggestion is to allocate more police from noon to midnight.

There are seasonal patterns in data, where although the total crime counts were different, the normalized values followed similar trends. When normalized by mean and standard deviations, seasonal patterns in a month appear. Similar patterns emerge for hours also. Different lines represent crimes for different categories (top 10 only) in Figures 7 and 8, respectively.

3.2. Variable Selection. The variable "Category" is the dependent variable for prediction. The variables "Resolution" and "Description" are irrelevant for the analysis because of their nature and were dropped from the dataset during preprocessing steps. The remaining variables are considered the independent variables, used for predicting the dependent variable.

3.3. Variable Transformation. Few variables are transformed to enrich the features of the dataset:

(1) The "Date" variable is divided into four separate variables: year of the incident (2003–2015), month and place of the incident (1–12), day of the incident (1–31), and the hour of the day when the incident happened (0–23).

(2) The variables DayOfWeek and PdDistrict are indexed and replaced by numbers (i.e., DayOfWeek: 1,2, . . ., 7, and PdDistrict: 1, 2, . . ., 10)

Figure 3: Year-wise top 10 crimes.

4. Prediction Model

The prediction model is based on Naive Bayes, Random Forest, and Gradient Boosting Decision Tree prediction techniques, briefly discussed below.

4.1. Naive Bayes. Naive Bayes is based on the Bayesian theorem, and it is a conditional probabilities method that calculates the probability by counting frequent values [14].

Naive Bayes is summarized as follows:

(1) A simple classification process classifier

(2) Best suited for historical data and prediction

(3) Classification technique analysis of the relationship between each attribute and the class instance

(4) A supervised learning method that can solve categorical and probabilistic problems

(5) A popular classification technique in text categorization [14].

This Naive Bayes classifier was introduced in 1995 [14]. It is known with different names in the community of data mining and machine learning, such as simple bases and independence Bayes [15]. The Naive Bayes classifier is commonly used in many applications like sentiment classifications and in different ensemble prediction models [16–18].

Using the Naive Bayes classifier, two types of quantities need to be calculated from the dataset, that is, Class Probabilities and Conditional Probabilities.

The method of the Bayesian classifier is given in the following equation:

$$P\left(\frac{C}{X}\right) = P\left(\frac{X}{C}\right)\frac{P(C)}{P(X)}. \quad (1)$$

Here, $P(C\text{-}X)$ is a maximum posterior hypothesis, $P(C)$ is prior, $P(X)$ is evidence, and $P(X\text{-}C)$ is the likelihood of the hypothesis [8].

4.2. Random Forests. Leo Breiman and Ahele Culter developed the Random Forest algorithm. In 1995, Tin Km Ho (Bell Labs) used for the first time the term Random Decision Tree. Ensemble learning method, Random Forests, or Random Decision Forest is a very famous classification and regression method. It is building numbers of the classifier on the training dataset which makes good predictions. This technique is also

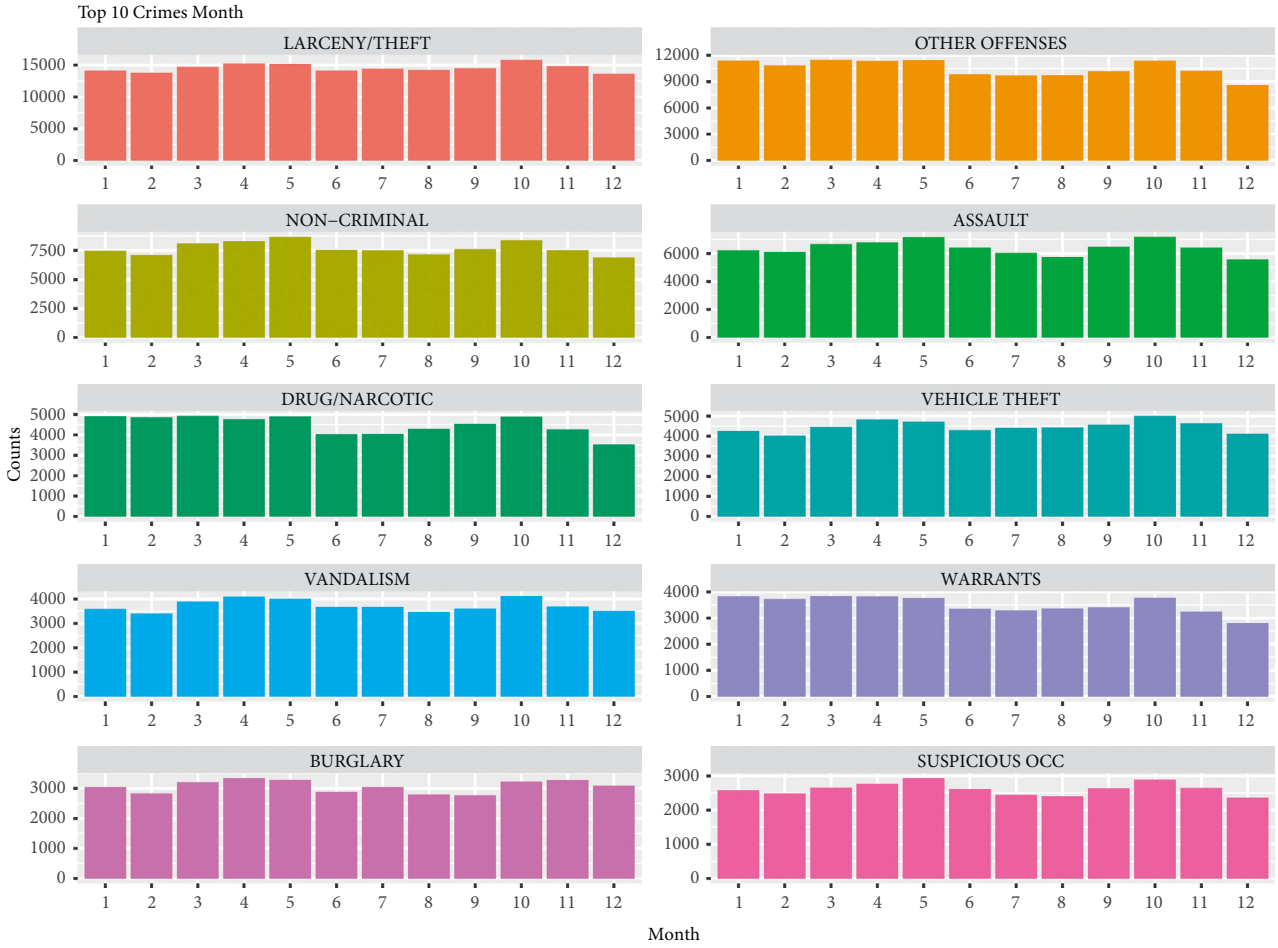

Figure 4: Month-wise top 10 crimes.

used for the predictions of handwriting character, digital pattern recognition, semantic analysis, language feature extraction, and hybrid models [19–22]. In this technique, every tree depends on randomly selected values sampled and independently corresponding distribution for every single tree around it. The numbers of trees increase in the forest general error for the forests converges as become to the limit for the forest's trees. The generalization error of the classifier depends on the correlation and individual strength between the trees of the forest. Each node in the Random Forest is split and randomly selected; the features yield an error rate that is better as compared with AdaBoost.

Definition. Random Decision Forests or Random Forest is a technique consisting of a tree-structured classifier $h(x, k)$, $k = 1,\ldots,$ where k represents independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

Correlation and Strength. In Random Decision Trees or RF, the generalization error can be obtained in terms of two parameters: how the single classifier measures the value accurately and the dependence between them [23].

Random Decision Forests correct for Decision Trees' habit of overfitting to their training set, and a Random Forest produces a large number of decision trees. For data including categorical variables with a different number of levels, Random Forests are biased in favor of those attributes with more levels. Categorical variables also increase the computational complexity to create trees [24].

4.3. Gradient Boosting Tree. Gradient Boosting Tree is a machine learning technique for classification and regression problems. This technique makes a prediction model that uses typically Decision Trees in the form of an ensemble of the weak prediction model. In this technique, the models are built in the same way as in other boosting models. It constructs the model in a stage-wise way as other boosting methods do, and it generalizes it by allowing optimization of an arbitrary differentiable loss function. The idea of gradient boosting originated in the observation by Leo Breiman where boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed by Jerome H. Friedman simultaneously with the more general functional gradient boosting perspective of Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Freeman.

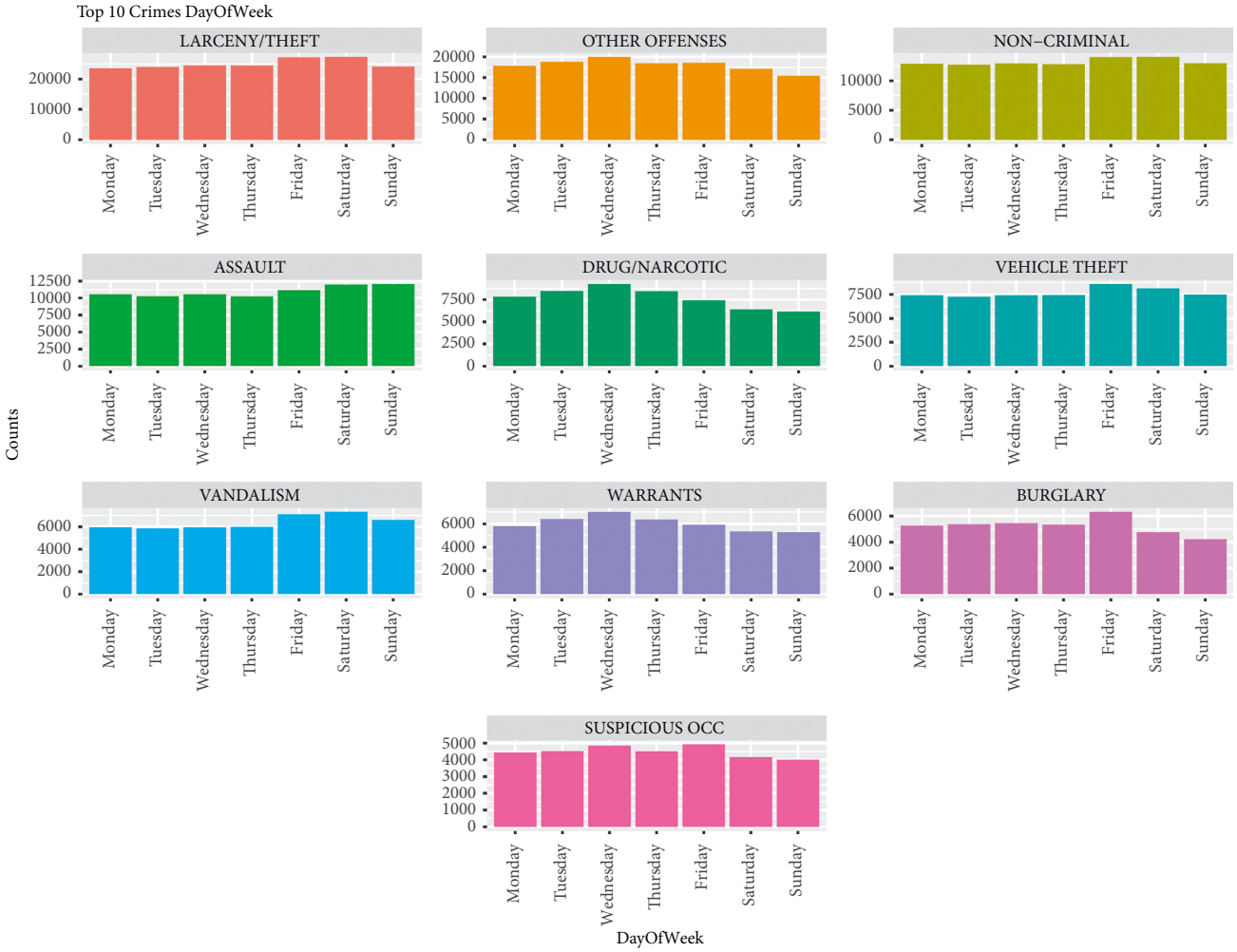

Figure 5: Top 10 crimes per day of week.

4.4. Performance Evaluation Metrics. The proposed prediction models are evaluated on the accuracy, precision, and recall, and ROC and Lift are the performance metrics for estimating the classification models [25]. Therefore, it is imperative to compare the accuracy using an alternative method, precision and recall; because of a two-class problem, the performance of a classifier is presented using the "confusion matrix" in Table 2.

The following are standardized equations for computing accuracy, sensitivity/recall, specificity, and precision.

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{FP} + \text{TN} + \text{FN}.$$

$$\text{Sensitivity} = \text{recall} = \text{TP}/t = \text{TP}/(\text{TP} + \text{FN}).$$

$$\text{Specificity} = \text{TN}/n = \text{TN}/(\text{TN} + \text{FP}).$$

$$\text{Precision} = \text{TP}/p = \text{TP}/(\text{TP} + \text{FP}).$$

TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative in the confusion matrix presented in Table 2. Precision in this context refers to the actual percentage of crime predicted by the classification model, which translates into the returns on the cost of categories. On the other hand, recall measures the percentage of crime identified and needed to be targeted. Thus, at last, specificity measures how good a test is at avoiding false alarms.

5. Experiment Results and Performance Evaluation

All three models were trained and presented in the previous section with different setting parameters and feature selections. The data exploration section observes that both the time-related features and geographic features are important. For analysis, all the three models are trained and tested, that is, the training dataset with 878,049 records from Kaggle, and they are divided into two parts in the ratio of 80 : 20 for all the models. Thus, 80% of the dataset were used to train the model, whereas 20% were used to test the model. The subsections discuss the performance and results.

6. Naive Bayes

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. In Table 3, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table

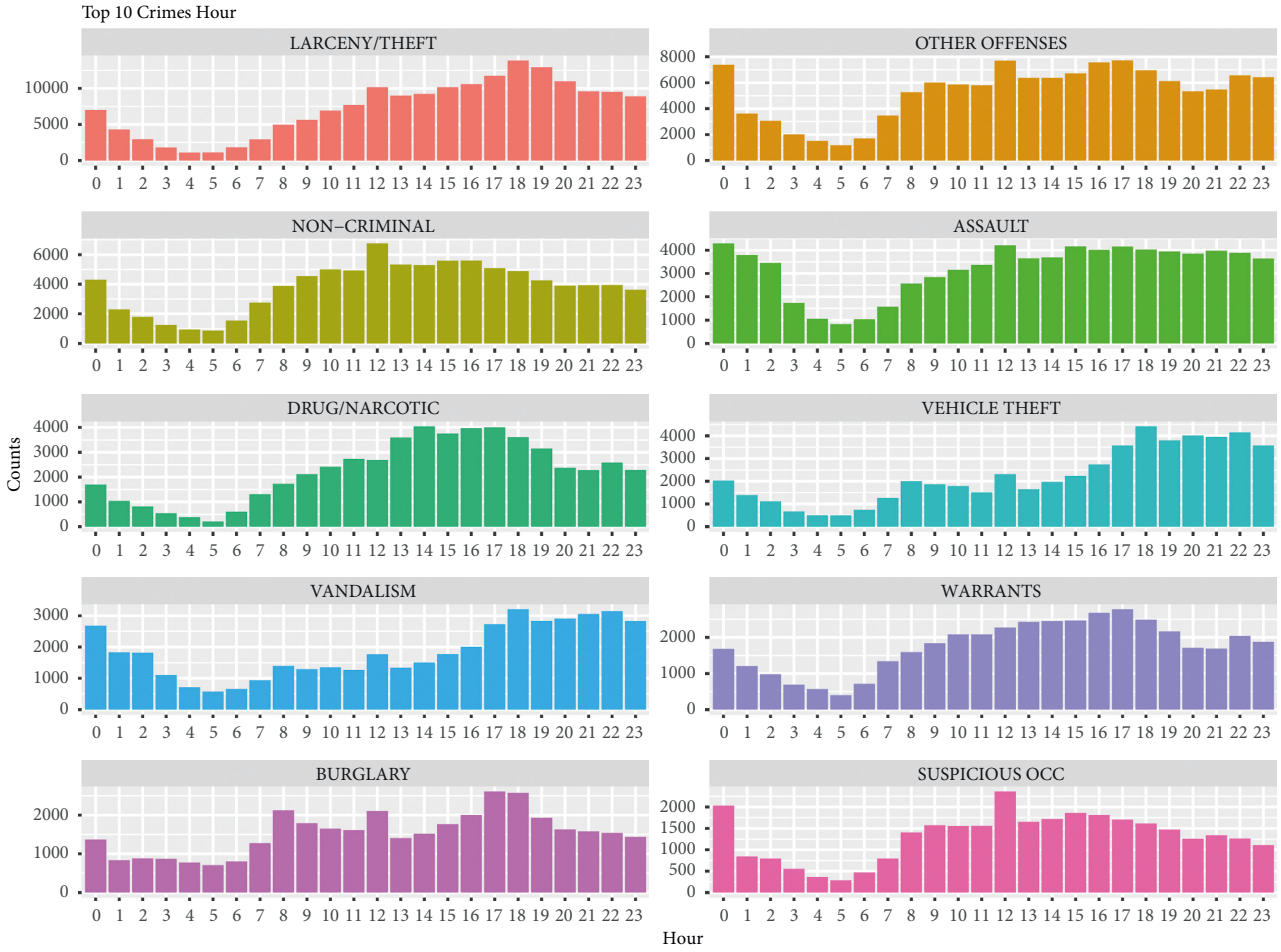

Figure 6: Top 10 crimes per hour.

classifies nonviolent crime and violent crime classes using the Naive Bayes algorithm for the training set. For each class, the result of a confusion matrix is discussed below:

There are 345,082 items classified into the nonviolent crime class.

(1) In the nonviolent crime class, the correctly classified items are 226,209.

(2) In the violent crime class, the wrongly classified items are 118,873.

There are 357,357 items classified into the violent crime class.

(1) In the violent crime class, the correctly classified items are 236,107.

(2) In the nonviolent crime class, the wrongly classified items are 121,250.

In Table 4, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table classifies nonviolent crime and violent crime classes using the Naive Bayes algorithm for the testing set. For each class, the result of a confusion matrix is discussed below.

There are 86,399 items classified into the nonviolent crime class.

(1) In the nonviolent crime class, the correctly classified items are 55,282.

(2) In the violent crime class, the wrongly classified items are 31,117.

89,211 items are classified into the violent crime class.

(1) n the violent crime class, the correctly classified items are 57,693.

(2) In the nonviolent crime class, the wrongly classified items are 31,518.

6.1. Random Forest. Random Forest technique is an ensemble learning method for classification, regression, and other tasks, operated by constructing a multitude of Decision Trees at training time and outputting the class, that is, the mode of the classes (classification) or means prediction (regression) of the individual trees. Random Decision Forests correct for Decision Trees' habit of overfitting to their training set. In this experiment, Random Forest was selected as a technique to estimate the predictors (Table 5).

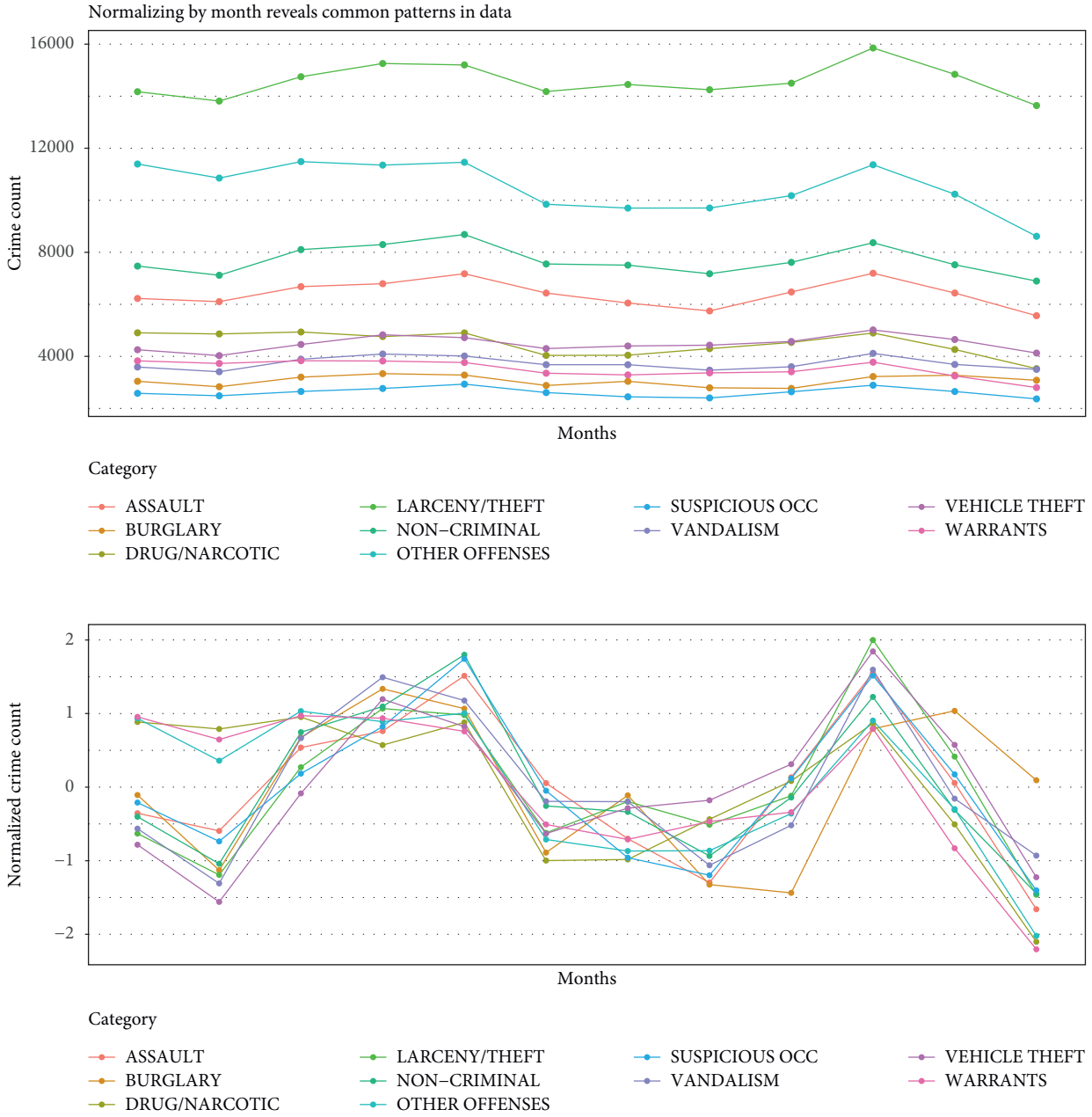

Figure 7: Normalizing by month reveals common pattern in data.

In Table 6, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table classifies nonviolent crime and violent crime classes using the Random Forest algorithm for the training set. For each class, the result of a confusion matrix is discussed below.

There are 349,230 items classified into the nonviolent crime class.

(1) In the nonviolent crime class, the correctly classified items are 280,840.

(2) In the violent crime class, the wrongly classified items are 68,390.

353,209 items are classified into the violent crime class.

(1) In the violent crime class, the correctly classified items are 287,017.

(2) In the nonviolent crime class, the wrongly classified items are 66,192.

In Table 7, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table classifies nonviolent crime and violent crime classes using the Random Forest algorithm for the testing set. For each class, the result of a confusion matrix is discussed below.

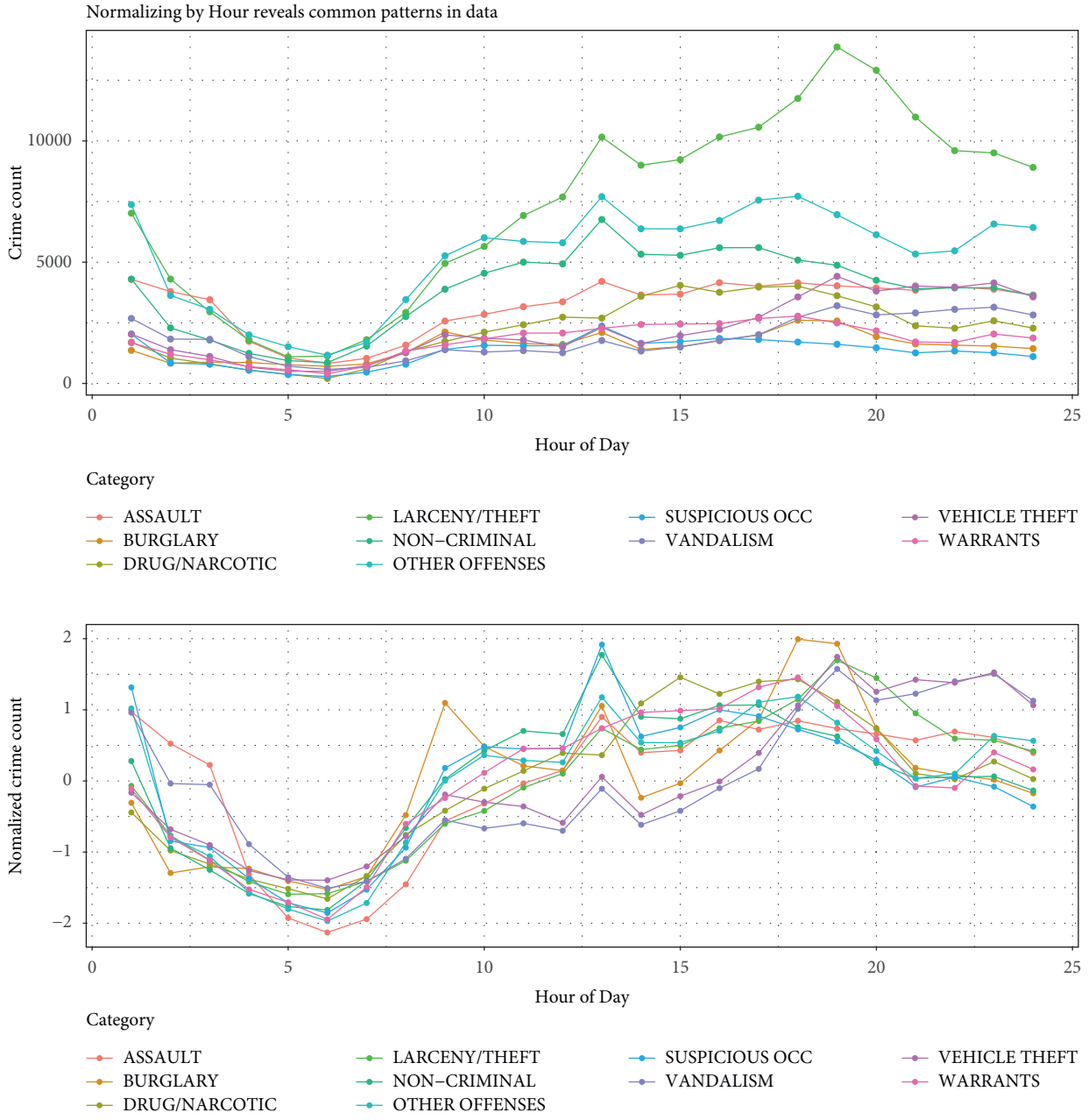

Figure 8: Normalizing by hour reveals common pattern in data.

Table 2: Confusion matrix.

		Yes	No
Actual class	Yes	True positive (TP)	False negative (FN)
	No	False positive (FP)	True negative (TN)

Table 3: Confusion matrix results of Naive Bayes on training data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	226,209	118,873
Violent crime	121,250	236,107

Table 4: Confusion matrix results of Naive Bayes on testing data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	55,282	31,117
Violent crime	31,518	57,693

Table 5: Accuracy, incorrectly classified instances, recall, and precision for Naive Bayes on training and testing data.

Method	Accuracy (correctly classified instances)	Incorrectly classified instances	Recall	Precision
Naive Bayes (training data)	65.82%	34.18%	65.55%	65.10%
Naive Bayes (testing data)	64.33%	35.67%	64.67%	63.8 8%

Table 6: Confusion matrix results of Random Forest on training data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	280,840	68,390
Violent crime	66,192	287,017

Table 7: Confusion matrix results of Random Forest on testing data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	254,779	31,766
Violent crime	32,448	56,617

There are 86,545 items classified into the nonviolent crime class.

(1) In the nonviolent crime class, the correctly classified items are 54,779.

(2) In the violent crime class, the wrongly classified items are 31,776.

89,065 items are classified into the violent crime class.

(1) In the violent crime class, the correctly classified items are 56,617.

(2) In the nonviolent crime class, the wrongly classified items are 32,448.

6.2. Gradient Boosting Trees. Gradient Boosting Decision Trees is a robust machine learning technique used in predictive modeling due to its high prediction accuracy compared to other modeling techniques. Gradient Boosting Decision Trees produces a prediction model in the form of an ensemble of weak prediction models, that is, Decision Trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes it by optimizing an arbitrary differentiable loss function (Table 8).

In Table 9, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table classifies nonviolent crime and violent crime classes using the Gradient Boosting Decision Trees algorithm for the training set. For each class, the result of a confusion matrix is discussed below.

There are 351,145 items classified into the nonviolent crime class.

Table 8: Accuracy, incorrectly classified instances, recall, and precision for Random Forest on training and testing data.

Method	Accuracy (correctly classified instances)	Incorrectly classified instances	Recall	Precision
Random Forest (training data)	80.84%	19.16%	80.41%	80.93%
Random Forest (testing data)	63.43%	36.57%	63.29%	62.80%

Table 9: Confusion matrix results of Gradient Boosting Decision Trees on training data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	347,260	3,885
Violent crime	0	351,294

(1) In the nonviolent crime class, the correctly classified items are 347,260.

(2) In the violent crime class, the wrongly classified items are 3,885.

351,294 items are classified into the violent crime class.

(1) In the violent crime class, the correctly classified items are 351,294.

(2) In the nonviolent crime class, the wrongly classified items are 0.

In Table 10, each column holds the reference (or actual) data and within each row is the prediction. The diagonal represents instances where our observation correctly predicted the class of the item. The table classifies nonviolent crime and violent crime classes using the Gradient Boosting Decision Trees algorithm for the testing set. For each class, the result of a confusion matrix is discussed below.

There are 86,569 items classified into the nonviolent crime class.

(1) In the nonviolent crime class, the correctly classified items are 86,569.

(2) In the violent crime class, the wrongly classified items are 0.

Table 10: Confusion matrix results of Gradient Boosting Decision Trees on testing data.

Predictions/references	Nonviolent crime	Violent crime
Nonviolent crime	86,569	0
Violent crime	430	88,611

Table 11: Accuracy, incorrectly classified instances, recall, and precision for Gradient Boosting Decision Trees on training and testing data.

Method	Accuracy (correctly classified instances)	Incorrectly classified instances	Recall	Precision
Gradient Boosting Decision Trees (training data)	99.44%	0.66%	98.89%	100%
Gradient Boosting Decision Trees (testing data)	99.75%	0.25%	100%	99.50%

89,041 items are classified into the violent crime class.

(1) In the violent crime class, the correctly classified items are 88,611.

(2) In the nonviolent crime class, the wrongly classified items are 430.

Tables 5, 8, and 11 present the accuracies of Naive Bayes, Random Forest, and Gradient Boosting Decision Tree techniques, respectively, and it is shown that the Gradient Boosting Decision Trees technique has better results.

7. Conclusions and Future Directions

The study presents exploratory data analysis using a prediction model based on classification techniques and compares the results of San Francisco crime data. The Naive Bayes, Random Forest, and Gradient Boosting Decision Tree are used for predicting the crime category attribute labeled "violent" and "nonviolent." The techniques are implemented in R languages, and the experimental results for all three algorithms manifest that Gradient Boosting Decision Tree performed better than Naive Bayes and Random Forest for the crime classification. The Gradient Boosting Decision Tree achieved 98.5%, 96.96%, and 100% for accuracy, precision, and recall, respectively. Law enforcement agencies can take great advantage of using machine learning algorithms like Gradient Boosting Decision Tree to fight crime effectively, channelize the resources efficiently, anticipate the crime up to some extent, and serve society. The proposed prediction models can be implemented to any dataset or crime data for predictions and resource management.

In the future, the same models using more advanced classification algorithms can be applied to the crime dataset and evaluate their prediction performance to discover trends and improve the subject knowledge. To design a comprehensive framework for the prediction that helps law enforcement agencies manage the resources in a specific area quickly, it is believed that higher accuracy can be achieved when employing more feature engineering in the address field. A more temporal analysis can be performed to determine the number and intensity of criminal activities using time series analysis, a mix of temporal and spatial analysis, which can help allocate resources more efficiently and effectively.

Data Availability

The dataset used in this research is available in the UCI repository online.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. M. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide*, Elsevier Science, Amsterdam, Netherlands, 1998.

[2] B. Kochar and R. Singh Chhillar, "An effective data warehousing system for rfid using novel data cleaning, data transformation and loading techniques," *The International Arab Journal of Information Technology*, vol. 9, no. 3, pp. 208–216, 2012.

[3] P. Santhi and V. Murali Bhaskaran, "Performance of clustering algorithms in healthcare database," *International Journal for Advances in Computer Science*, vol. 2, no. 1, pp. 26–31, 2010.

[4] A H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E M. Al-Shawakfa, "A comparison study between data mining tools over some classification methods," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, pp. 18–26, 2011.

[5] M. Khan, S. S. Khan, and M. D. Awan, "Comparative exploration of features for data mining results by legend navigation interactive technique," *International Journal of Database Theory and Application*, vol. 9, no. 9, pp. 49–58, 2016.

[6] M. Khan, S. S. Khan, K. Ullah, and G. Ullah, "Evaluating interactive visualization techniques on small touch screen devices," *International Journal of Grid and Distributed Computing (IJGDC)*, vol. 12, no. 02, pp. 31–48, 2019.

[7] M. Khan, A. Shah, and I. Ahmad, "Framework for interactive data mining results visualization on mobile devices,"

International Journal of Database Theory and Application, vol. 7, no. 4, pp. 23–36, 2014.

[8] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2011.

[9] Y. Abouelnaga, "San francisco crime classification," arXiv preprint arXiv:1607.03626, 2016.

[10] T. A. Shen, W. Wang, and S. Chyou, *San francisco Crime Classification*, University of California, California, USA, 2015.

[11] U. Saeed, M. Sarim, A. Usmani, A. Mukhtar, S. Abdul Basit, and S. Kashif Riffat, "Application of machine learning algorithms in crime classification and classification rule mining," *Research Journal of Recent Sciences*, vol. 4, no. 3, pp. 106–114, 2015.

[12] S. Shojaee, A. Mustapha, Fatimah Sidi, and M. A. Jabar, "A study on classification learning algorithms to predict crime status," *International Journal of Digital Content Technology and its Applications*, vol. 7, no. 9, p. 361, 2013.

[13] R. Iqbal, M. A. Azmi Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," *Indian Journal of Science and Technology*, vol. 6, no. 3, pp. 4219–4225, 2013.

[14] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, vol. 18, no. 60, pp. 1–8, 2006.

[15] K. M. Leung, *Naive Bayesian Classifier*, Polytechnic University Department of Computer Science/Finance and Risk Engineering, Hong Kong, China, 2007.

[16] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[17] A. Onan, "Ensemble of classifiers and term weighting schemes for sentiment analysis in Turkish," *Scientific Research Communications*, vol. 1, no. 1, 2021.

[18] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, no. 1–16, 2016.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, 2017.

[21] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[22] M. A. Toçoğlu and A. Onan, *International Conference on Intelligent and Fuzzy Systems*, pp. 1693–1700, Springer, Berlin, Germany, 2020.

[23] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[24] Xi Xia, W. Zhao, X. Rui et al., "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," in *Proceedings of the 2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, November 2015.

[25] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 69–78, ACM, Seattle WA USA, Auguest 2004.