Received May 13, 2020, accepted June 17, 2020, date of publication July 17, 2020, date of current version August 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009969

Exploration of the Hidden Influential Factors on Crime Activities: A Big Data Approach

JIANMING ZHOU[1], ZHENG LI[2], JACK J. MA[2], AND FEIFENG JIANG[3]

[1]China Unicom Guangzhou Branch, Guangzhou 510335, China
[2]Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong
[3]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong

Corresponding author: Feifeng Jiang (jiangfeifeng0302@outlook.com)

ABSTRACT Crime activities have long been a great concern of all the countries. Analysis of crime data has been a key part yet a considerable challenge for discovering crime patterns and reducing crimes. In recent year, along with the development of data collection and data mining techniques, lots of big data-related studies have been conducted to analyze the crime data. Studying the numerical influential factors is one important yet challenging problem, especially for those indirect features. Though a number of studies have been conducted to analyze the influential factors of crime activities, most of them have some limitations in the era of ''big data''. Some adopted the linear statistical methods, of which the basic assumption is opposite to the non-linear real world. Some limited their studied factors within one or two aspects. Some overlooked the importance of ranking the influence of factors. To fill these research gaps, this paper proposes a big data approach to analyze the influential factors on the crime activities, and experimented it on New York City. More than 1515 different factors ranging from demographic, housing, education, economy, social, and city planning were considered and analyzed. The proposed framework combines non-linear machine learning algorithms and geographical information system (GIS) to study the spatial determinants of crimes. Recursive feature elimination (RFE) is used to select the optimum feature set. Performance of gradient boost decision tree (GBDT), logistic regression (LR), support vector machine (SVM), artificial neural network (ANN) and random forest (RF) are compared to generate the optimum model. Important impact factors were then investigated using GBDT and GIS. The experimental results demonstrate that the combined GBDT and GIS model can find out the most important factors of crime rate with high efficiency and accuracy.

INDEX TERMS Big data techniques, feature analysis, felony assault, gradient boost decision tree, machine learning, recursive feature elimination.

I. INTRODUCTION

Crime has long been a ubiquitous social problem in human society. It will not only damage the individual rights and interests but also threaten the security of a country or even the world. Since the 9/11 attacks in 2001, growing concern about crime has been a serious problem in many countries [1]–[3]. Efficient crime prevention methods are necessarily required. However, in front of the growing volumes of crime data, how to accurately and efficiently analyze the data and extract helpful information for crime prevention has been a major challenge facing all law-enforcement and intelligence-gathering organizations [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Eunil Park.

In academia, lots of data related studies have been conducted on crime analysis. Majority of the existing literatures are exploring three kinds of problems. First is the prediction of crime activities. To better support police and governments, researchers implemented different statistical methods or time series models to forecast the crime rates and trends. For example, Catlett *et al.* [6] presented a predictive approach based on auto-regressive models to automatically forecast crime trends in different high rick regions. Ingilevich and Ivanov [7] adopted three different approaches, including liner regression, logistic regression and gradient boosting, to forecast the number of crimes in different areas of the city of Saint-Petersburg. Rummens *et al.* [8] applied an ensemble model based on logistic regression and made bi-weekly predictions and temporally disaggregated monthly predictions.

The second kind of problems that attract scholars is on the detection of crime hot spots. By identifying the hot spots, governors would be able to arrange their managements in a more strategical way. In this aspect, scholars developed different kinds clustering methods to identify the spatial and temporal patterns of crime hot spots. For example, Catlett *et al.* [6] using spatial DBSCAN based clustering techniques to identify high-risk crime regions. Mohler [9] combined several years of data and many different crime types to develop hotspot maps by extending point clustering models of crime. Bsoul *et al.* [10] utilized the affinity propagation clustering algorithm to detect the crime patterns.

The third group of data driven studies were exploring the cause effect behind the crime activities. These literatures intended to disclose more hidden influential factors on crime, and therefore provided more practical suggestions for crime prevention. For example, Khan *et al.* [11] investigated the socio-economic determinants of crime. Their study found that higher unemployment would diminish the rate of return of legal activities, and was more likely to increase the return of illegal activities. Lochner and Moretti [12] investigated the effect of education on crime. Using Census and FBI data, they found that schooling could significantly reduce the probability of incarceration and arrest. Immergluck and Smith [13] studied the impact of single-family mortgage foreclosures on neighborhood crime by analyzing the data of foreclosures, neighborhood characteristics and crime. They found that higher foreclosure levels would contribute to higher levels of violent crime.

This paper is also focusing on the third problem. Although many studies have been conducted to explore the influential factors, some limitations commonly exist in the existing literatures. The first lies in their methods for analysis. Most existing studies analyzed the crime data using statistical methods, such as ordinary-least-squares (OLS) linear regressions [14] and Poisson-based regressions [15]. These methods assume the relationship between crime and possible influential factors are linear, which is opposite to the non-linearity of the real-world situations. Therefore, the performance of these methods can be limited.

Secondly, when investigating the influential features of crime, most existing literature considered limited amount of factors, or focused on only one kind of factors such as economy, education, and population [16]–[19]. This, obviously, limits their discoveries in the era of big data. Other factors like place of interest, house vacancy rate, and city infrastructure-related factors have not been well studied. These factors can also affect citizen activities and therefore indirectly influence the crime rate.

Thirdly, the studied influential factors are not properly ranked. Majority of the non-linear machine models are referred as ''black-box'' algorithms, and cannot reveal the variable importance, while linear models such as logistic regression and lasso regression have the problem of multicollinearity, which may mitigate or even omit the weights of some important variables. The calculated results may therefore be misleading [20]. This may lead a wrong direction to the governments on which aspect to emphasize during policy making.

To overcome the limitations, this paper proposes a big data approach to analyze the influential factors on the crime activities. The experimented district was classified into 2168 areas, and each area was modeled as a sample case. More than 1515 different factors ranging from demographic, housing, education, economy, and social of each sample were considered and analyzed. A non-linear machine learning algorithm namely gradient boost decision tree (GBDT) is integrated to learn the numerical relationships between crime rates and the influential factors. A recursive feature elimination framework was implemented to support feature selection and feature ranking. The performance of the proposed framework is validated by comparing with other commonly seen machine learning methods, such as logistic regression (LR), support vector machine (SVM), artificial neural network (ANN) and random forest (RF). Based on the most important factors of crime rate given by the integrated framework using GBDT and RFE, GIS analysis is then utilized to analyze the influence of factors in different census tract regions. Practical suggestions are proposed for local law enforcement to control the growing crime rate.

The remaining of the paper is organized as follows: Section II presents the methodology framework, Section III shows the case study in New York City, Section IV analyzes and discusses the most important features and Section V concludes the work.

II. METHODOLOGY FRAMEWORK

The proposed methodology framework is shown in Figure 1. It mainly consists of three parts: 1) data collection and preprocessing, 2) selection of models and features, and 3) feature analysis. Details of the framework are introduced as follows.

A. DATA COLLECTION AND PREPROCESSING

Big data analysis is the process of discovering patterns in large datasets. Therefore, a well-collected and preprocessed dataset is the basis. The proposed methodology tends to identify important influential factors among different kinds of feature data, including crime data, demographic data, economic data, housing data, social data, and educational data. However, these data are collected from different sources and are difficult to join directly. To achieve this, this study proposes GIS for data fusion. It can join different datasets together as long as they can be described using geographical information such as latitude/longitude and city/county [21]. In the experiment conducted in Section III, the datasets were indexed using census tract, a geographic region defined for the purpose of taking a census. Details will be introduced later.

Furthermore, the collected raw datasets normally have some flaws, such as redundant data, missing value, and high correlational data. Redundant data should be removed to improve modeling efficiency and reduce the computation

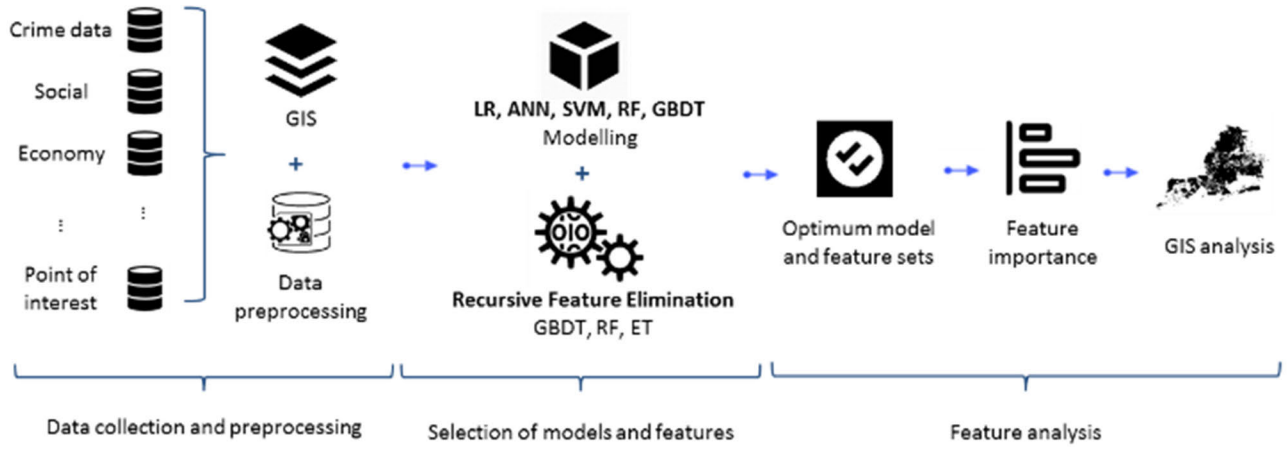

FIGURE 1. Methodology framework.

cost. Missing values occur when no data value is stored for the variable in an observation. They are usually filled with mean/median value, zero value, or simply be deleted when too much are missing [22]. High correlational data means the feature has similar meanings to the study target or other features. These features will interfere in the analysis of influential factors, especially for linear based models. The coefficient of some important features will be assigned using very small value upon the existence of highly correlated features. This is one crucial problem that most existing studies have not well addressed. This study will not only implement the Pearson correlation coefficient to remove part of highly correlated features, but also use ensemble tree-based models to mitigate the multicollinearity problem.

B. SELECTION OF CLASSIFICATION MODELS AND FEATURES

1) GRADIENT BOOST DECISION TREE (GBDT)

In this paper, gradient boost decision tree (GBDT) is adopted to model the relationship between crime rate and influential factors. It is a powerful machine learning technique developed based on CART decision tree. It has the advantages of high accuracy, fast training, and small memory footprint [23]. Due to the strong abilities of GBDT, it has been implemented in many domains, such as advertising systems, sales prediction, medical data analysis and image classification [23]–[25].

GBDT produces a prediction model for classification in the form of an ensemble of CART decision trees using gradient boosting techniques. It builds one tree at an iteration to fit the residual of the trees that precede it [23]. For given training data $X = \{x_i\}_{i=1}^{N}$, their labels $Y = \{y_i\}_{i=1}^{N}$, the classification model F(x) and the differentiable loss function $\mathcal{L}(y_i, F(x_i))$, the goal of GBDT is to choose a classification function that minimizes the aggregation of the loss function, which can be expressed as Equation (1) [23].

$$F^* = \underset{F}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{1}$$

One important ability of GBDT is calculating the variable importance, which is also used in this paper. The calculation of relative influence in GBDT can be shown as Equation (2) and (3).

$$I_k^2 = \frac{1}{M} \sum_{m=1}^{M} I_k^2(h_m) \tag{2}$$

$$I_k^2(h_m) = \sum_{t=1}^{L-1} \Delta Loss_t^2 \cdot I(t = k) \tag{3}$$

where I_k is the relative influence of a variable k in the GBDT model, M is the number of iterations or the number of trees, h_m represents the mth tree, and $\Delta Loss_t^2$ is the decrease of the square loss after the node t split.

Besides GBDT, our methodology also proposes to try four other commonly seen machine learning methods, including logistic regression (LR), support vector machine (SVM), artificial neural network (ANN) and random forest (RF). Modeling performances of these models will be experimented and compared in the case study.

2) RECURSIVE FEATURE ELIMINATION (RFE)

In this paper, more than one thousand features are collected. However, not all the collected features are important to the crime rate. Some of them may contain a lot of noise and therefore become redundant. Those features will not only slow down the calculation process but also lead to overfitting and lower the model performance. As a result, a proper feature selection method is required to remove those features.

This study proposes the recursive feature elimination (RFE) strategy to select features. It is a recursive process that uses the remaining features to fit a base model and removes the weakest features until the specified number of features is reached [26]. Features are ranked by the coefficients or feature importance of the base model. By recursively eliminating a feature or a small number of features per loop, RFE can gradually eliminate dependencies and collinearity that may exist in the model. The pseudo code of the RFE framework is shown in Algorithm 1.

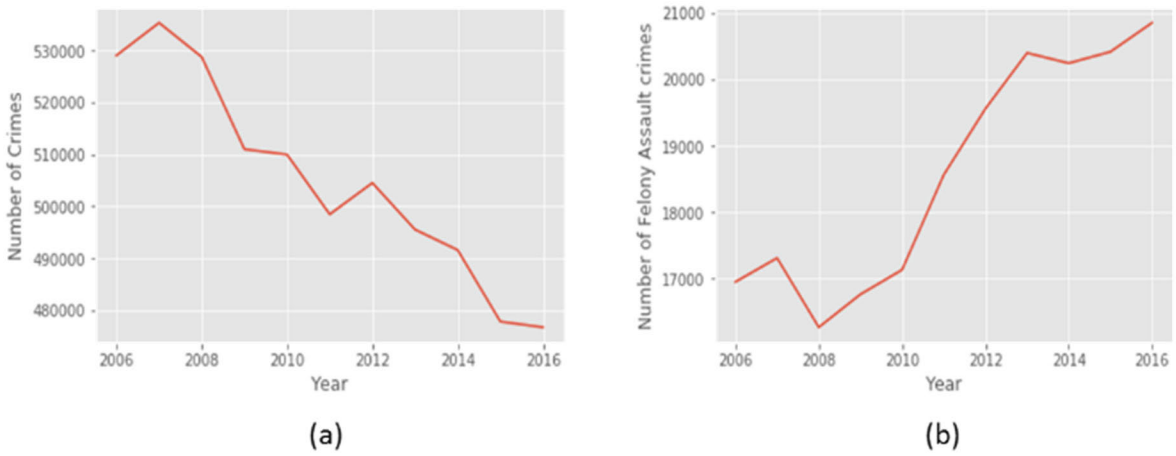

FIGURE 2. Trends of (a) the total number of crimes and (b) the number of felony assault crimes.

Algorithm 1 RFE Algorithm

Inputs:
  Training set T
  Set of p features $F = \{f_1, \ldots, f_p\}$
  Ranking method M(T, F)
  Number of features to select n
  Number of features to remove at each iteration s
Outputs: Final ranking R
Code:
  While p>n
    Rank set F using M(T, F)
    $F^* \leftarrow$ last s ranked feature in F
    p $\leftarrow$ p − size of F^*
    F $\leftarrow$ F − F^*

The number of features to keep and the number of features to remove per iteration are two important parameters of RFE. However, it is often not known how many features are valid in advance. Therefore, this parameter needs to be tuned in practice. Cross-validation and grid search may help get a stable result more efficient.

As mentioned above, RFE requires a base model to help calculate the feature importance and remove the irrelevant and redundant features. In this paper, random forest (RF), gradient boost decision tree (GBDT) and extremely randomized trees (ET) are selected as the candidates of the base model for RFE due to their capability in ranking variable importance.

The way that GBDT calculates the feature importance has been introduced in Section II-B-1. Random forest (RF) calculates the variable importance using the out-of-bag (OOB) data of each bootstrap sample [27]. Basically, the algorithm will develop lots of decision trees to model the variables and the target. It will perform random permutation of a variable's value in the OOB data and check the number of votes for correct class. By comparing the difference and calculate the average of this value over all trees in the forest is the raw importance score for a variable.

Extremely randomized trees (ET) is also a tree-based ensemble learning method proposed by Geurts *et al.* [28]. On the one hand, ET is very similar to RF. It randomly selects samples for each sub-set and features for each tree. On the other hand, while RF builds a tree with the best classification attributes at each node, ET selects the cut-point at random. It removes the need for the optimization of the discretization thresholds. Therefore, compared to RF, ET has a clear advantage in terms of computing times and ease of implementation [28].

C. FEATURE ANALYSIS

After comparing the performance of the algorithms mentioned above, an optimal model and an optimal feature set can be obtained. Using the optimal model, feature importance can be calculated and the most important factors can be identified. Then GIS can be applied to detect the spatial relationships between the crime rate and the influential factors. Based on the associations, possible reasons behind are analyzed and practical suggestions are proposed.

Overall, our methodology framework adopts non-linear machine learning algorithms to investigate the relationship between crime rate and influential factors. The framework was able to integrate data from different dimensions and filter out the most important factors. In the following part of the paper, a case study is conducted to validate the effectiveness of the proposed framework.

III. CASE STUDY

A. DATA COLLECTION

1) TARGET DATA

This paper selected New York City (NYC) to conduct the case study. Since the 1990s, crime volume in NYC has continuously reduced under the unremitting efforts of New York City Police Department (NYPD). In 2017, NYC was ranked as one of the metropolitans with the lowest crime rate in America. However, as shown in Figure 2, though the total number of crimes decreases, the number of felony assault crimes

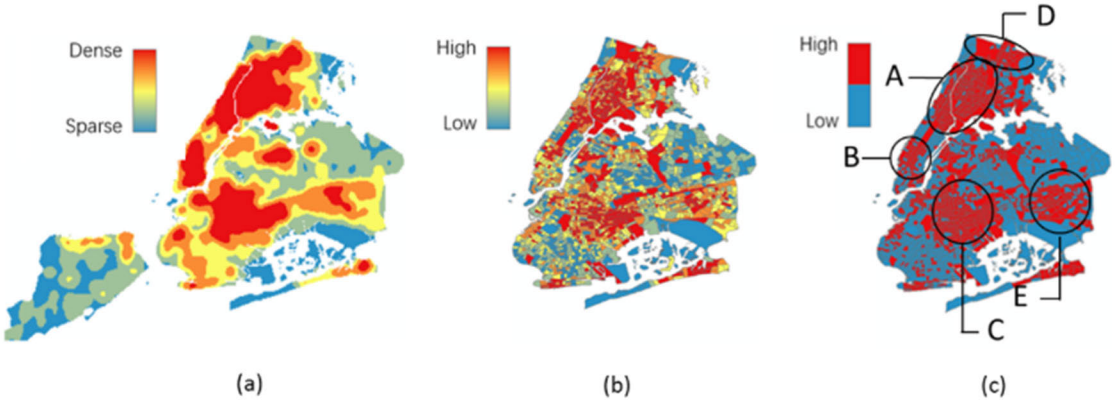

FIGURE 3. The distribution of (a) the kernel density of felony assault, (b) the crime rate of felony assault, and (c) the binarized crime rate of felony assault.

shows an increasing trend from 2006 to 2016. Felony assault means the victims suffer a physical injury of varying severity caused by either a deadly weapon or dangerous instrument. It, obviously, will seriously threaten the safety of NYC citizens. Therefore, this paper selects this kind of crime as the target to numerically investigate its prediction model and influential factors.

The felony assault data of NYC 2016 are collected from NYC Open Data. NYC is composed of five boroughs, including Manhattan, Brooklyn, Queens, Bronx and Staten Island. Distribution of the kernel density of felony assault crimes is shown in Figure 3 (a). It can be seen that in some areas, such as Manhattan, Bronx, and Brooklyn, the felony assault crimes are very dense. However, some districts have denser or more crimes simply because these places have more people. In order to eliminate the interference from the population, we choose crime rate as the target of the experiment. It is calculated by dividing the number of total felony assault crimes by the population of the census tract. This study uses census tract (CT) as the geographical unit because most demographic and economic features were collected at this unit according to United States Census Bureau. The calculation is conducted with the help of GIS, which will be introduced later. In total, there are 2168 census tracts in NYC. However, after an inspection of the data, we find that the CTs in the Staten Island borough either involve a small number of crimes or a small number of citizens. Their crime rate, thus, is either too high or too low. These CTs are very likely to be outliers and interfere the modeling. Therefore, CTs in the Staten Island are excluded in this paper. After this, we obtained the crime rate of 2057 census tracts in NYC. Distribution of the crime rate is presented in Figure 3 (b). It can be observed that the crime rate varies from CT to CT. To simplify the problem and also filter the outliers, this study divides the census tracts into two groups. One is that their crime rates are higher than the median of all census tracts. They are marked as high rate census tracts (high rate CTs), and are the positive samples in the experiment. The other is that their crime rates are lower than the median. They are the low rate CTs, and are marked as the negative samples. In total, the experiment gets 1050 positive samples and 1007 negative samples. Distribution of the high rate CTs and the low rate CTs is shown in Figure 3 (c). It can be seen that the high rate CTs mainly concentrate at area A-E.

2) FEATURE DATA

To study the possible reasons behind the increasing number of felony assault crimes, various kinds of features are investigated. Six datasets including demographic, housing, education, economic, social and place of interest are collected from United States Census Bureau and NYC Open Data. Detailed features of the six datasets are presented in Table 1. The demographic group contains 268 features, such as age, gender, race, and household owner. The housing group describes the vacancy rate, house structure, house price, etc. It includes 286 features. The education group describes the education level of different age, gender, race, etc. 384 features are included. The economic group consists of 274 features, involving employment rate, income, insurance, poverty rate, etc. The social group covers the information of family type, family relationship, marriage, etc. It includes 290 features. Place of interest involves 13 features, including recreational facility places, commercial places, public safety places, etc. In total, 1515 features are collected.

B. DATA PREPROCESSING

After the data are collected, preprocessing is conducted. First, after an inspection of the data, we found that some of the cases are describing the same felony assault at the same time and the same location. These are duplicated cases and should be removed. They account for 2.4% of the crime data.

Then data fusion is conducted using GIS. The crime data is firstly imported into ArcMap, a GIS software platform. Locations of crime are shown on the NYC map based on the latitude and longitude. Then the 2010 Census Tracts map collected from NYC Open Data is added. The location of felony assault is connected to the Census Tracts Map using spatial connection. The number of felony assault crimes within each census tract is calculated. Crime rate of each census tract is

TABLE 1. Datasets description.

1st level aspect	2nd level category	features	1st level aspect	2nd level category	features
Demographic	Population by sex and age	60	**Education**	Attainment population by sex and age	158
	Population by race	108		Attainment population by race	166
	Household	100		Poverty rate by Attainment	24
	Total	268		Median earnings by Attainment	36
Housing	Rent price	36		Total	384
	Value	20	**Social**	Ancestry	56
	Structure	66		Disability status	16
	Monthly owner costs	66		Computer and internet use	6
	Occupancy	18		Educational attainment	20
	Mortgage status	6		Fertility	14
	House Heating Fuel	20		Grandparents	18
	Year structure built	22		Households	32
	Year householder moved into unit	14		Language spoken at home	24
	Vehicles available	10		Marital status	24
	Facilities	8		Place of birth	14
	Total	286		Relationship	14
Economic	Income	126		Residence 1 year ago	16
	Class of worker	10		School enrollment	12
	Commuting to work	16		U.S. citizenship status	6
	Employment status	34		Year of entry	14
	Health insurance coverage	48		Veteran status	4
	Occupation	12		Total	290
	Industry	28	**Place of Interest**	Different places (recreational, safety, ...)	13
	Total	274		Total	13
Total features: 1515					

given by dividing the number of felony assault crimes by the population of each census tract. At the same time, the borough census tract code is provided. Based on the code, the label data would be able to join the collected 1515 features.

After the data fusion, missing values are handled. In the dataset, 14 features have missing values. The number of missing cases on these 14 features are all 193, accounting for 9% of the data. Since most of the missing values are non-available values at places with few inhabitants, zero value is adopted to fill them.

Fourth is the high correlational data. A number of features are highly correlated with other features in the prepared 1515 features. For example, the feature ''percent of adults 25 years and over with bachelor's degree or higher'' and the feature ''percent of adults with bachelor's degree or higher'' has a correlation value of 0.9995. The existence of a highly correlated feature can interfere with the importance value of the other one. This is called the multi-collinearity problem [21]. Though GBDT adopted in this paper can effectively mitigate the multi-collinearity between features, these high correlational data will slow down the computation speed and increase the complexity of the feature analysis. To address the problem, Pearson correlation coefficient is deployed. It is based on co-variance and can give information about the magnitude of the association. In this experiment, the correlation coefficients of 16,510 pairs of features are higher than 0.9, accounting for 0.72% of all the pairs. For each high correlational pair, the feature that has a relatively lower correlation with our label is deleted. We also examined the correlations between the features and the label. The highest correlation value is 0.47, so no more features were excluded. In sum, 1045 features are removed in this step, and there remained 470 features.

Furthermore, to mitigate the influence of different value range and speed up the modeling, z-score transformation is conducted to normalize the data. Calculation of the z-score transformation is shown in Equation (4).

$$X_{\text{transform}} = \frac{X - \mu}{\sigma} \tag{4}$$

TABLE 2. Model comparison.

Performance	Feature selection method			Average	Using the original 1515 features	Improvements
	GBDT	RF	ET			
LR	0.673	0.679	0.691	0.681	0.635	7.24%
SVM	0.730	0.742	0.753	0.742	0.674	10.09%
ANN	0.755	0.769	0.761	0.762	0.693	9.96%
RF	0.785	0.791	0.796	0.791	0.769	2.86%
GBDT	0.805	0.799	0.809	0.804	0.766	4.96%

where μ represents the mean value, and σ represents the standard deviation.

C. FEATURE SELECTION AND MODEL OPTIMIZATION

The collected raw data contains 1515 features. After the data preprocessing, 470 features are remained. Still, not all of them are important features. As mentioned in Section II-B-2, some features might increase the computation cost and lower the model performance. Therefore, they need to be found out and removed.

Recursive feature elimination (RFE) is adopted to select the features. The number of selected features is required to be pre-set for RFE. To find the optimal number of selected features, we take extremely randomized tree (ET) as the base model of RFE and combine it with GBDT. The number of removed features per loop is set as 10. The result of parameter optimization is presented in Figure 4. It can be seen that when the number of selected features is 300, GBDT has the best performance with the accuracy of 0.809. Therefore, 300 features are selected from 470 for further treatments.

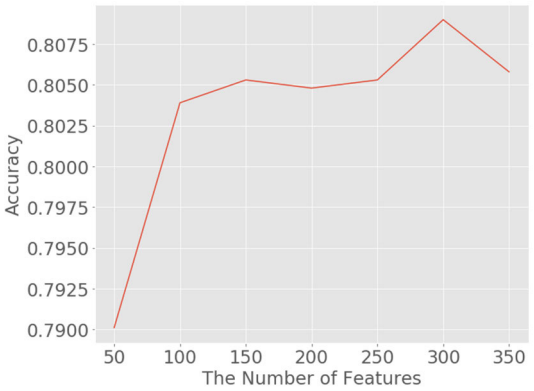

FIGURE 4. Parameters optimization for RFE.

At the same time, the parameters of GBDT need to be tuned to achieve better performance. Max depth D_{max}, number of trees N_T and learning rate l are three important parameters. Larger D_{max} generally improves the performance of individual trees but decreases the generalization of the model and increases the chance of overfitting. Larger N_T could give better performance but make the code slower. Learning rate l will influence the convergence speed of the model. The selection of l is usually associated with N_T, because smaller l may require more trees to converge, while larger l may not need much trees to find an optimal value. For simplicity, we set l as the default value of 0.1. N_T and D_{max} are optimized using grid search. Figure 5 shows the optimization results. It can be observed that when D_{max} is 5 and N_T is 800, GBDT has the highest accuracy of 0.809. Default parameters for GBDT is $N_T = 100$, and $D_{max} = 3$. The accuracy for this group of parameters is 0.775. Therefore, it can be calculated that the process of parameter optimization increased the model performance 4.39%.

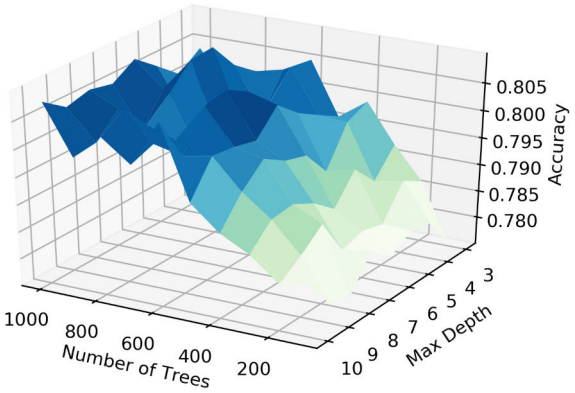

FIGURE 5. Parameter optimization for GBDT.

Moreover, to prove that GBDT is a reasonable choice for our experiment, its performance is compared with four other models, including logistic regression (LR), support vector machine (SVM), artificial neural network (ANN) and random forest (RF). The performance of three feature selection base models of RFE is also compared. They are GBDT, RF and extremely randomized trees (ET). Also, the performance of using the original 1515 features are also calculated for comparison. The results are shown in Table 2.

It can be seen that compared with the models without feature selection techniques, the performance for different algorithms improved 2.86%-10.09%, especially for SVM and ANN. RF and GBDT improved the least, but this reflects that these two algorithms have a better tolerance on noise. Especially for the RF algorithm, it has the highest accuracy

when using the original 1515 features. In addition, compared to other models, GBDT has the highest average accuracy after RFE. When ET is chosen as the base model of RFE, GBDT has the highest accuracy of 0.809. This proves that it is reasonable to combine GBDT with ET for feature selection.

Besides, six kinds of features are considered as the possible influential factors of crime rate in this study. To explore the influence of each kind of feature on the model performance, a contrast experiment is conducted. The performance of GBDT model when different types of features are excluded is shown in Table 3. It can be seen that when the economic-related features are dropped, the model has the lowest score of accuracy. This means economic has a higher impact on the model. However, the impact is also associated with the number of features and the correlation among features. For example, social features and demographic features have a higher correlation with each other. Therefore, when either of these two categories is dropped, the model accuracy does not reduce too much. To further analyze the relationship between the six types of features and the crime rate, impact of individual features on crime rate needs to be examined.

TABLE 3. Impact of each feature type on model performance.

Dropped feature group	Model performance
Place of interest	0.788
Social	0.799
Economic	0.786
Education	0.792
Housing	0.789
Demographic	0.797
None	0.809

IV. FEATURE ANALYSIS AND DISCUSSION

A. FEATURE IMPORTANCE

After the optimum feature set and model are obtained, feature importance is calculated using GBDT to uncover the most important features on the crime rate of felony assault in NYC. One important reason that this study employs GBDT to calculate the variable importance is that it can help mitigate the multicollinearity problem. The tree structure and ensemble methods in GBDT help reflect the variable importance more objectively. Table 4 lists the top 10 most important features among the selected 300.

It can be seen from Table 4 that some features have similar meanings. For example, the first and the fourth feature both refer to single women or marital situation. The second and the third feature both represent the statistics of Black or African American. The fifth, sixth and eighth feature all represent the poverty degree or economic level of the census tract. The seventh and the last feature describe the education level. The ninth feature belongs to the aspect of place of interest.

TABLE 4. Feature importance.

Rank	Feature	Importance
1	Percent of females 15 years and over who never married (Percent_NM)	0.0187
2	Percent of Black or African American (Percent_BA)	0.0174
3	Estimated population of Black or African American (Pop_BA)	0.0153
4	Percent of families with female householder or no husband present (Percent_NH)	0.0119
5	Percent of people with food stamp/supplemental nutrition assistance program benefits in the past 12 months (Percent_SNAP)	0.0118
6	Percent of families whose income is less than $10,000 in the past 12 months (Percent_less than $10,000)	0.0117
7	Percent of adults 25 years and over with bachelor's degree or higher (Percent_BD)	0.0099
8	Percent of families whose income is below the poverty level in the past 12 months (Percent_PL)	0.0098
9	Number of recreational facilities (Num_RF)	0.0094
10	Percent of Black males with high school diploma (Percent_HS)	0.0085

Therefore, for a clearer feature analysis, we divide the top 10 features into five aspects, including:

a) Marital: Percent_NM, Percent_NH
b) Black or African American: Percent_BA, Pop_BA
c) Economic: Percent_SNAP, Percent_less than $10,000, and Percent_PL
d) Education: Percent_BD, Percent_HS
e) Point of interest: Num_RF

B. MARITAL

The first and the fourth feature both describe the marital situation of single women. To find out the relationship between these two factors and the crime rate, the distribution of them are visualized in the NYC map using GIS. Since these two features show a similar distribution in the map, we pick the first feature as an example for illustration. Its distribution is presented in Figure 6. It can be seen that the Percent_NM is higher in area A, B and C. Compared Figure 6 with Figure 3 (c), it can be concluded that the high percent of single women might be one of the reasons behind the high crime rate in area A, B and C. We also compare the average percent of females 15 years and over who never married, with who has ever married in high/low rate CTs. Figure 7 shows

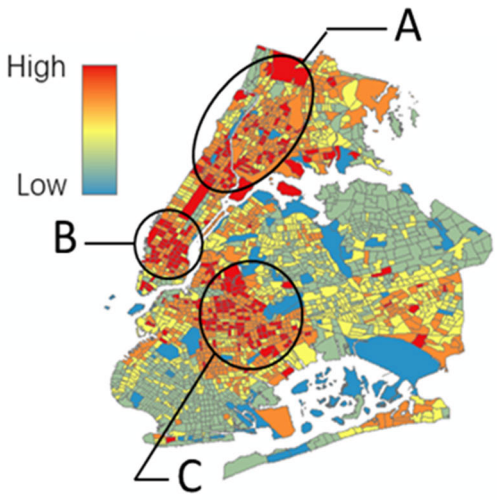

FIGURE 6. The distribution of the Percent_NM.

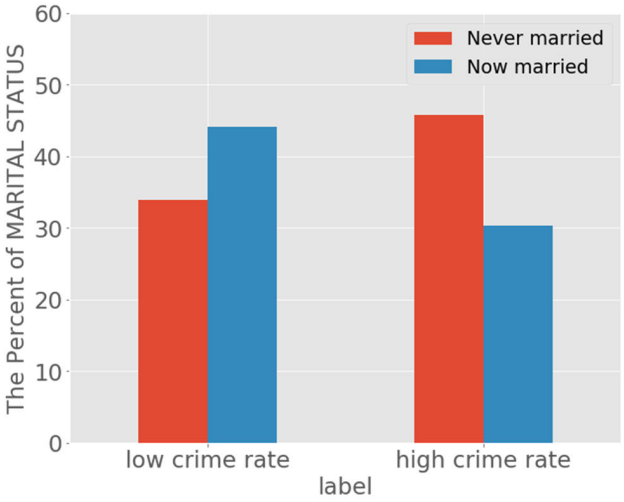

FIGURE 7. Average Percent_NM and who has ever married in low/high rate CTs.

the results. It can be seen that in high rate CTs, the average Percent_NM is much higher than that in low rate CTs.

The reason behind the high crime rate in census tracts with high single female rate might lie in several aspects. On the one hand, single women without companies usually have less strength to resist. Therefore, they are more likely to be targeted by criminals. On the other hand, higher percent of single female reflects the lower rate of marriage to some extent. Previous studies have proved that marriage has a negative correlation with crime activities [29], [30]. Married people are much more mature, responsible, easier to calm down, and are less likely to be criminals. In addition, for rapists, more single women means more potential targets. Therefore, the rape crime, which is one important kind of felony assault crimes, might be higher in CTs with high single female rate.

Therefore, to lower the crime rate of felony assault, government should study more policies to protect single women, especially in area A, B, and C. Single women should also be more careful and ask for a company when under potential risks of being attacked.

C. BLACK OR AFRICAN AMERICAN

The percent and the population of Black or African American are two important factors of crime rate as shown in Table 4. Distribution of the Percent_BA is presented in Figure 8. It shows that in area C, D and E, Black or African American accounts for a higher percent. Compared Figure 8 with Figure 3 (c), it can be observed that the high percent of Black or African American in area C, D and E might be one of the surface causes of the high crime rate in these areas. We also compute the average percent of other kinds of races in high/low rate CTs. Figure 9 gives the results. It shows that in high rate CTs, the Percent_BA is almost three times as that in low rate CTs.

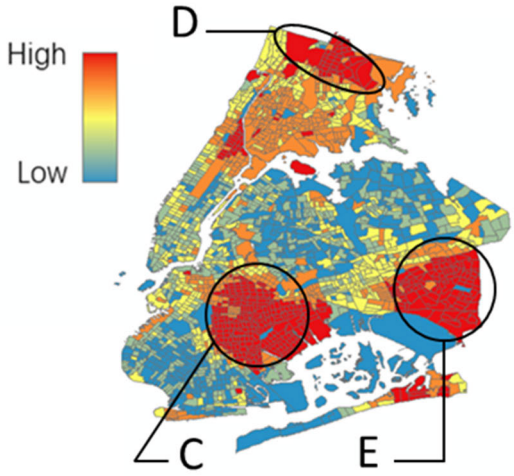

FIGURE 8. The distribution of the Percent_BA.

Relationship between race and crime has long been a sensitive topic of public controversy and scholarly debate in the United States [31]. Crime statistics have shown that the rate at which Black or African American both commit and are the victim of homicide is about six to eight times higher than that of white American [32]. The incarceration rate of Blacks is more than three times higher than their representation in the general population [32]. However, race is not the reason for crime activities. A number of previous studies have attempted to explain why some racial groups appear to be over-represented in official crime statistics [33]. Socioeconomic factors, such as social stereotype, single-parent families, lower educational attainment, low personal income, lower social class position and crime-ridden neighborhoods, have been generated to be the cause effects behind the high crime rate of Black or African American [33]–[35]. Among these factors, economic and education level have been further

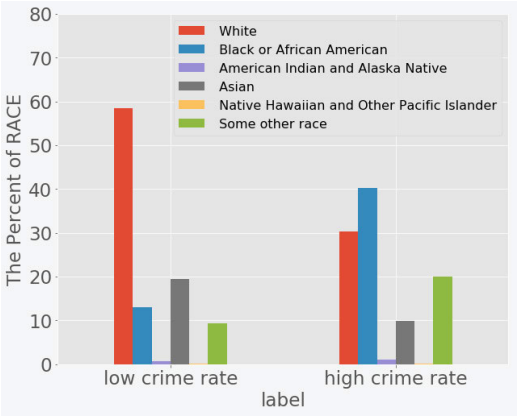

FIGURE 9. Average percent of all kinds of races in low/high rate CTs.

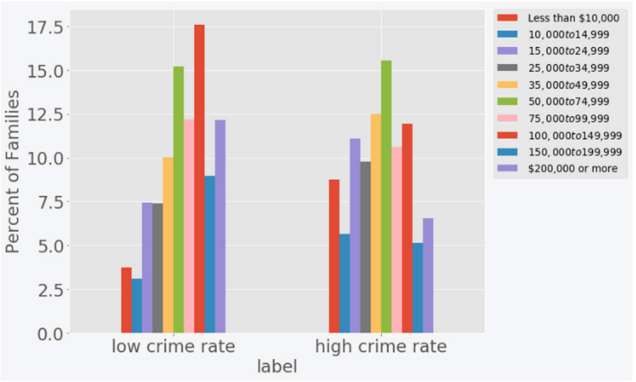

FIGURE 11. Average percent of families at different annual average income levels in low/high rate CTs.

concluded to be the major reasons [36], [37]. Relationship between the economic and education level and the crime rate will be explained later.

D. ECONOMIC

The fifth, sixth and eighth most important factors of crime rate uncovered in Table 4 all reflect the economic level of the CTs. Higher percent of these features means lower economic level. Figure 10 presents the distribution of the Percent_SNAP. It shows that the percent is higher in area A and C. Compared Figure 10 with Figure 3 (c), it can be concluded that the high crime rate in these two areas might be related to their lower economic level. We also calculate the average number of families at different annual average income levels in low/high rate CTs. Figure 11 presents the results. It can be seen that the average percent of families whose annual average income is around $50,000 to $74,999 is similar in high rate CTs and low rate CTs. However, in high rate CTs, the average percent of families whose income is lower than $50,000 is much higher than that in low rate CTs, while the average percentage of families whose income is higher than $74,999 is much lower than that in low rate CTs. This means census tracts with more lower-income households have higher crime rates.

Economy is the basis of society. Low economic level will lead to high unemployment rate and high poverty rate. These will further lead to a high level of stress and mental illness which in turn causes individuals to adopt the criminal behavior [11], [20]. Therefore, in order to decrease the crime rate of felony assault, NYC government should give priority to the economic development in poorer census tracts in area A and C. For example, use tax incentives to attract more firms, improve living conditions in these tracts through water, sanitation and solid waste management, encourage innovation and entrepreneurship, etc.

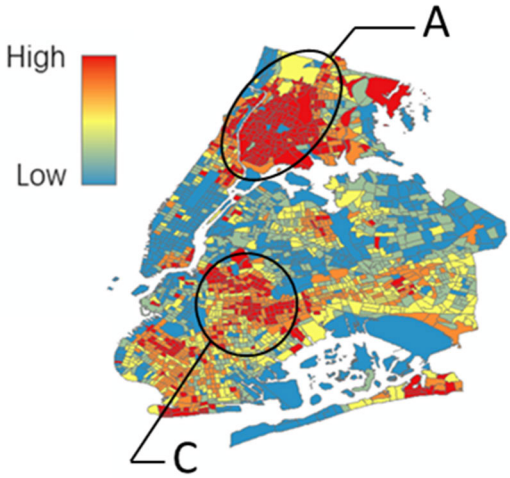

FIGURE 10. The distribution of the Percent_SNAP.

E. EDUCATION

Table 4 shows that Percent_BD and Percent_HS are two important factors of crime rate in NYC. These two factors both reflect the education level. To uncover the relationship between the education level and the crime rate, distribution of the Percent_BD is presented in Figure 12. It can be observed that the percent is lower in area A, C and E. Compared Figure 12 with Figure 3 (c), it can be concluded that the lower percent of adults with higher education level could be one of the reasons of the high crime rate in area A, C and E. In addition, we calculate the average percent of adults 25 years and over with bachelor's degree or higher and less than high school graduate in high/low rate CTs. Figure 13 presents the results. It can be seen that the percent of adults with higher education level in high rate CTs is lower than that in low rate CTs. The percent of adults less than high school graduate, on the other hand, is higher in high rate CTs. This suggests that there is a significant negative relationship between the crime rate and the education level.

Possible reasons behind this negative association are that people with higher education level are more likely to solve problems in a rational and legal way instead of in an extreme way that will hurt others. Also, they have a better

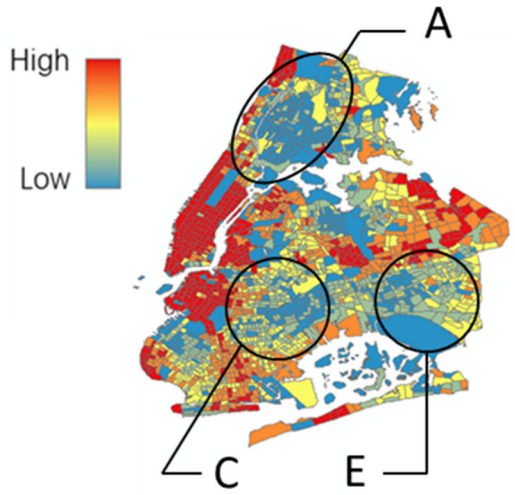

FIGURE 12. The distribution of the Percent_BD.

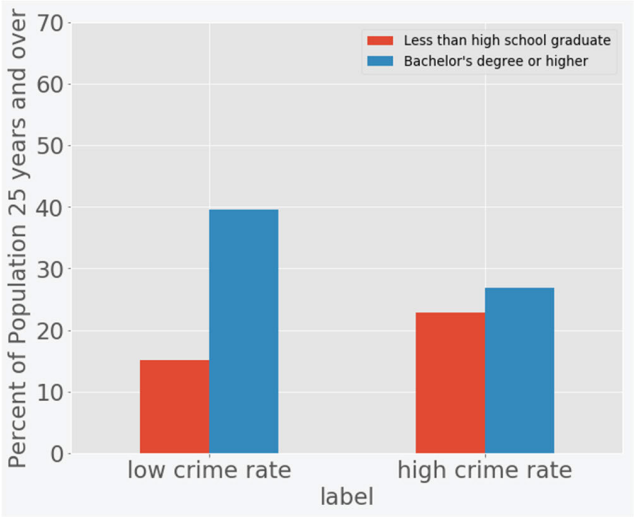

FIGURE 13. Average percent of adults 25 years and over with bachelor's degree or higher and less than high school graduate in high/low rate CTs.

understanding of the harm of crimes. They know what is wrong and what is right to do. In addition, well-educated people are more likely to have higher income, and therefore the opportunity costs to commit crimes for them are higher. This also reduces their crime rates [11], [12].

Therefore, to reduce the crime rate and guarantee the social security, education level in area A, C and E should be improved. For short-term, government could provide more job training and vocational educations for people with low education level in these areas. For long-term improvement, government could set up more community schools in these tracts and introduce policies to encourage people to take up advance studies.

F. PLACE OF INTEREST

Num_RF ranks the ninth most important factors of crime rate in Table 4. To explore the relationship between it and the crime rate, its distribution is presented in Figure 14. It can be seen that there is a larger number of recreational facilities in area A. Compared Figure 14 with Figure 3 (c), it can be inferred that the high crime rate of felony assault in area A might be associated with the large Num_RF in the area. We also calculate the average number of recreational facilities in low/high rate CTs. Results are shown in Figure 15. It can be observed that the average Num_RF in high rate CTs is twice as much as that in low rate CTs. Combined Figure 14 and Figure 15, it can be concluded that the Num_RF might have a positive correlation with the crime rate.

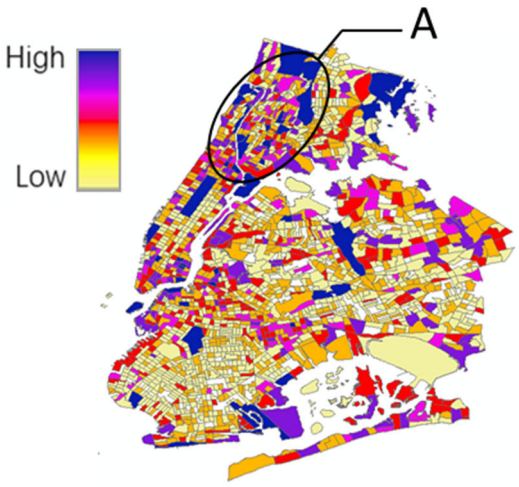

FIGURE 14. The distribution of the Num_RF.

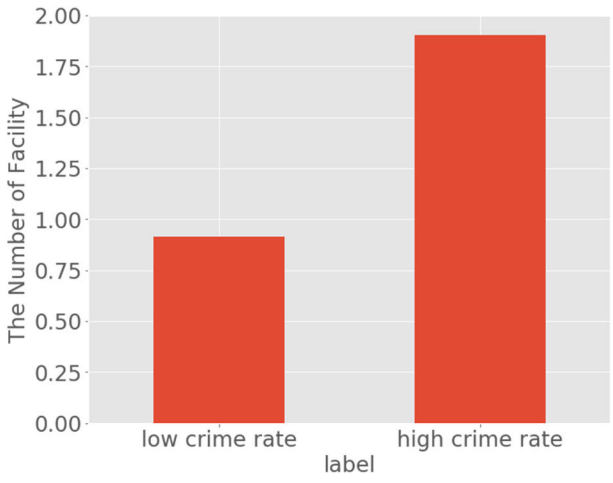

FIGURE 15. Average Num_RF in low/high rate CTs.

Recreaitional facility refers to a building or place for leisure activities, such as parks, swimming pools, tennis courts, amusement rides and golf courses. These places are usually designed for people to hang out and get relaxed, and therefore they could also be a desired place for criminals to seek targets. Several studies have investigated the role of recreational facilities, such as parks, as crime inhibitors or

generators [38]–[40]. Their results showed that neighborhood parks were associated with increased levels of crime.

Thus, to reduce crime rate, security measurements should be strengthened around places with more recreational facilities. More surveillance cameras, field and walkway lightings should be installed and police officers should pay more attention to these places.

V. CONCLUSION

To conclude, this paper proposes a ''big data'' methodology framework for analyzing factors of crime rate. An integrated model combining gradient boost decision tree (GBDT), recursive feature elimination (RFE) and geographical information system (GIS) is used to filter, rank and analyze the important features. Effectiveness of the methodology is validated by a case study in New York City (NYC). Crime rate of felony assault at census-tract level is studied and multiple possible influential features are analyzed. Our results show that:

1) The Recursive Feature Elimination framework can effectively increase the machine learning performance 2.86%-10.09% for big data analysis.
2) Parameter optimization is an important step for implementing machine learning algorithms. In our experiment, this help increase 4.39% of the model performance.
3) Compared with Neural Networks and SVM, tree-based ensemble algorithms, such as RF and GBDT, have a higher tolerance on noise data during the learning process.
4) The integrated model based on RFE-GBDT yields the highest modeling accuracy in this study. Surpass other well-known machine learning algorithms like SVM, ANN and RF.
5) Features in the economic group has the largest group effect on the crime rates of felony assault in NYC.
6) Five kinds of features are found to be the most important factors of crime rate of felony assault. They are marital, Black or African American, economic, education and place of interest.
7) For areas concentrated with high crime rate census tracts, possible reasons are explained and practical suggestions are proposed for local government based on the influential features.
8) Our methodology framework has a strong ability of generalization. It not only shows great performance in the case study in this paper, but also can be applied to other cities/countries to study other types of crimes.

Based on this paper, the future study could focus on (1) field-testing the relationships between the crime rate and the uncovered features to explore more practical suggestions for the local government, (2) applying our model at city/country scale to investigate more macro level influential features.

REFERENCES

[1] R. Krishnamurthy and J. S. Kumar, ''Survey of data mining techniques on crime data analysis,'' *Int. J. Data Mining Techn. Appl.*, vol. 1, no. 2, pp. 47–49, Dec. 2012, doi: 10.20894/IJDMTA.102.001.002.006.

[2] J. R. Faria, L. M. Ogura, and A. Sachsida, ''Crime in a planned city: The case of Brasília,'' *Cities*, vol. 32, pp. 80–87, Jun. 2013, doi: 10.1016/j.cities.2013.03.002.

[3] D. E. Brown, ''The regional crime analysis program (ReCAP): A framework for mining data to catch criminals,'' in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 3, Oct. 1998, pp. 2848–2853, doi: 10.1109/ICSMC.1998.725094.

[4] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, ''Crime data mining: A general framework and some examples,'' *Computer*, vol. 37, no. 4, pp. 50–56, Apr. 2004, doi: 10.1109/MC.2004.1297301.

[5] W. V. Ackerman and A. T. Murray, ''Assessing spatial patterns of crime in Lima, Ohio,'' *Cities*, vol. 21, no. 5, pp. 423–437, Oct. 2004, doi: 10.1016/j.cities.2004.07.008.

[6] C. Catlett, E. Cesario, D. Talia, and A. Vinci, ''Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments,'' *Pervasive Mobile Comput.*, vol. 53, pp. 62–74, Feb. 2019, doi: 10.1016/j.pmcj.2019.01.003.

[7] V. Ingilevich and S. Ivanov, ''Crime rate prediction in the urban environment using social factors,'' *Procedia Comput. Sci.*, vol. 136, pp. 472–478, Sep. 2018, doi: 10.1016/j.procs.2018.08.261.

[8] A. Rummens, W. Hardyns, and L. Pauwels, ''The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context,'' *Appl. Geography*, vol. 86, pp. 255–261, Sep. 2017, doi: 10.1016/j.apgeog.2017.06.011.

[9] G. Mohler, ''Marked point process hotspot maps for homicide and gun crime prediction in Chicago,'' *Int. J. Forecasting*, vol. 30, no. 3, pp. 491–497, Jul. 2014, doi: 10.1016/j.ijforecast.2014.01.004.

[10] Q. Bsoul, J. Salim, and L. Q. Zakaria, ''An intelligent document clustering approach to detect crime patterns,'' *Procedia Technol.*, vol. 11, pp. 1181–1187, Jan. 2013, doi: 10.1016/j.protcy.2013.12.311.

[11] N. Khan, J. Ahmed, M. Nawaz, and K. Zaman, ''The socio-economic determinants of crime in Pakistan: New evidence on an old debate,'' *Arab Econ. Bus. J.*, vol. 10, no. 2, pp. 73–81, Oct. 2015, doi: 10.1016/j.aebj.2015.01.001.

[12] L. Lochner and E. Moretti, ''The effect of education on crime: Evidence from prison inmates, arrests, and self-reports,'' *Amer. Econ. Rev.*, vol. 94, no. 1, pp. 155–189, Feb. 2004, doi: 10.1257/000282804322970751.

[13] D. Immergluck and G. Smith, ''The impact of single-family mortgage foreclosures on neighborhood crime,'' *Housing Stud.*, vol. 21, no. 6, pp. 851–866, Nov. 2006, doi: 10.1080/02673030600917743.

[14] L. G. A. Alves, R. S. Mendes, E. K. Lenzi, and H. V. Ribeiro, ''Scale-adjusted metrics for predicting the evolution of urban indicators and quantifying the performance of cities,'' *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0134862, doi: 10.1371/journal.pone.0134862.

[15] D. W. Osgood, ''Poisson-based regression analysis of aggregate crime rates,'' *J. Quant. Criminol.*, vol. 16, no. 1, pp. 21–43, 2000, doi: 10.1023/A:1007521427059.

[16] T. C. Pratt and F. T. Cullen, ''Assessing macro-level predictors and theories of crime: A meta-analysis,'' *Crime Justice*, vol. 32, pp. 373–450, Jan. 2005, doi: 10.1086/655357.

[17] C. J. Baier and B. R. E. Wright, '''If you love me, keep my commandments': A meta-analysis of the effect of religion on crime,'' *J. Res. Crime Delinquency*, vol. 38, no. 1, pp. 3–21, Feb. 2001, doi: 10.1177/0022427801038001001.

[18] Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, ''The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan,'' *Cities*, vol. 79, pp. 45–52, Sep. 2018, doi: 10.1016/j.cities.2018.02.021.

[19] D.-W. Sohn, ''Residential crimes and neighbourhood built environment: Assessing the effectiveness of crime prevention through environmental design (CPTED),'' *Cities*, vol. 52, pp. 86–93, Mar. 2016, doi: 10.1016/j.cities.2015.11.023.

[20] L. G. A. Alves, H. V. Ribeiro, and F. A. Rodrigues, ''Crime prediction through urban metrics and statistical learning,'' *Phys. A, Stat. Mech. Appl.*, vol. 505, pp. 435–443, Sep. 2018, doi: 10.1016/j.physa.2018.03.084.

[21] J. Ma and J. C. P. Cheng, ''Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology,'' *Appl. Energy*, vol. 183, pp. 182–192, Dec. 2016, doi: 10.1016/j.apenergy.2016.08.079.

[22] E. R. Buhi, P. Goodson, and T. B. Neilands, ''Out of sight, not out of mind: Strategies for handling missing data,'' *Amer. J. Healthcare Behav.*, vol. 32, no. 1, pp. 83–92, 2008, doi: info:doi/10.5993/AJHB.32.1.8.

[23] S. Si, H. Zhang, S. S. Keerthi, D. Mahajan, I. S. Dhillon, and C.-J. Hsieh, ''Gradient boosted decision trees for high dimensional sparse output,'' in *Proc. 34th Int. Conf. Mach. Learn. (PMLR)*, Sydney, NSW, Australia, 2017.

[24] Z. Wen, B. He, R. Kotagiri, S. Lu, and J. Shi, "Efficient gradient boosted decision tree training on GPUs," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, May 2018, pp. 234–243, doi: 10.1109/ipdps.2018.00033.
[25] J. Ma and J. C. P. Cheng, "Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining," *J. Cleaner Prod.*, vol. 151, pp. 406–418, May 2017, doi: 10.1016/j.jclepro.2017.03.083.
[26] *Recursive Feature Elimination-Yellowbrick 0.9 Documentation*. Accessed: Nov. 22, 2018. [Online]. Available: http://www.scikit-yb.org/en/latest/api/features/rfecv.html
[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.
[29] R. J. Sampson and J. H. Laub, "Crime in the making: Pathways and turning points through life," *Crime Delinquency*, vol. 39, no. 3, pp. 396–396, Jul. 1993, doi: 10.1177/0011128793039003010.
[30] M. Zoutewelle-Terovan, V. van der Geest, A. Liefbroer, and C. Bijleveld, "Criminality and family formation: Effects of marriage and parenthood on criminal behavior for men and women," *Crime Delinquency*, vol. 60, no. 8, pp. 1209–1234, Dec. 2014, doi: 10.1177/0011128712441745.
[31] Wikipedia. (2018). *Race and crime in the United States*. Accessed: Nov. 26, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Race_and_crime_in_the_United_States&oldid=870529293
[32] *Bureau of Justice Statistics (BJS)-Prison Inmates at Midyear 2009-Statistical Tables*. Accessed: Nov. 26, 2018. [Online]. Available: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=2200
[33] EBSCOhost. *Hidden Intersections: Research on Race, Crime, and Criminal Justice in Canada*. Accessed: Nov. 29, 2018. [Online]. Available: https://web.a.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=00083496&AN=13184710&h=xQpB9sSSpWr4Zk25ucvz5Upigs%2bzitcgRPfYOigjw7YUQpVwWvrAduqNGVPqoOV7wNUl8XsUevrL27FJvBlcHg%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d00083496%26AN%3d13184710
[34] B. R. E. Wright and C. W. Younts, "Reconsidering the relationship between race and crime: Positive and negative predictors of crime among African American youth," *J. Res. Crime Delinquency*, vol. 46, no. 3, pp. 327–352, Aug. 2009, doi: 10.1177/0022427809335170.
[35] K. Drakulich and E. Rodriguez-Whitney, "Intentional inequalities and compounding effects," in *The Handbook of Race, Ethnicity, Crime, and Justice*. Hoboken, NJ, USA: Wiley, 2018, pp. 17–38, doi: 10.1002/9781119113799.ch1.
[36] P. Walberg, M. McKee, V. Shkolnikov, L. Chenet, and D. A. Leon, "Economic change, crime, and mortality crisis in Russia: Regional analysis," *BMJ*, vol. 317, no. 7154, pp. 312–318, Aug. 1998, doi: 10.1136/bmj.317.7154.312.
[37] W. Groot and H. M. van den Brink, "The effects of education on crime," *Appl. Econ.*, vol. 42, no. 3, pp. 279–289, Feb. 2010, doi: 10.1080/00036840701604412.
[38] A. Boessen and J. R. Hipp, "Parks as crime inhibitors or generators: Examining parks and the role of their nearby context," *Social Sci. Res.*, vol. 76, pp. 186–201, Nov. 2018, doi: 10.1016/j.ssresearch.2018.08.008.
[39] E. Groff and E. S. McCord, "The role of neighborhood parks as crime generators," *Secur. J.*, vol. 25, no. 1, pp. 1–24, Feb. 2012, doi: 10.1057/sj.2011.1.
[40] *Close-Ups and the Scale of Ecology: Land Uses and the Geography of Social Context and Crime*. Accessed: Nov. 26, 2018. [Online]. Available: https://www.researchgate.net/publication/279163670_Close-Ups_and_the_Scale_of_Ecology_Land_Uses_and_the_Geography_of_Social_Context_and_Crime
[41] W.-K. Chen, *Linear Networks and Systems*. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

JIANMING ZHOU received the bachelor's degree in mathematics and management from the University of Electronic Science and Technology of China, and the Master of Engineering degree from Tongji University. He is currently with China Unicom Guangzhou Branch. His research interests include big data, cloud computing, the IoT, industrial Internet, 5G, and communication technology.

ZHENG LI received the B.Eng. degree in mechanical engineering from the Huazhong University of Science and Technology, in 2012. He was a Mechanical Engineer with Midea Group, Foshan, China. He is currently an AI Researcher with Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include machine learning, deep learning, and data mining in smart city.

JACK J. MA received the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently the Chief Research Officer of the Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include smart city, urban computing, data mining, and artificial intelligence.

FEIFENG JIANG is currently pursuing the Ph.D. degree with the Department of Architecture and Civil Engineering, City University of Hong Kong. Her main research interests are traffic safety analysis, machine learning, and big data.