doi:10.20944/preprints202002.0108.v1

Article

Machine Learning Algorithms for Visualization and Prediction Modeling of Boston Crime Data

Jiarui Yin[1], Inikuro Afa Michael[2] and Iduabo John Afa[3,*]

[1] Department of Mathematics, Universitat Autonoma de Barcelona 08193, Bellatera, Spain;
[2] Department of Computer Engineering, Taras Shevchenko University of Kyiv 01033, Kyiv, Ukraine;
[2] Department of Physics, Universitat Politecnica de Catalunya, 08222, Terrassa, Spain;
* Correspondence: iduabo.john.afa@alu-etsetb.upc.edu.

Abstract: Machine learning plays a key role in present day crime detection, analysis and prediction. The goal of this work is to propose methods for predicting crimes classified into different categories of severity. We implemented visualization and analysis of crime data statistics in recent years in the city of Boston. We then carried out a comparative study between two supervised learning algorithms, which are decision tree and random forest based on the accuracy and processing time of the models to make predictions using geographical and temporal information provided by splitting the data into training and test sets. The result shows that random forest as expected gives a better result by 1.54% more accuracy in comparison to decision tree, although this comes at a cost of at least 4.37 times the time consumed in processing. The study opens doors to application of similar supervised methods in crime data analytics and other fields of data science.

Keywords: Machine learning; decision tree; random forest; crime data analytics.

---

1. Introduction

In previous years, crime rate in Boston has experienced a significant increase especially in cases of property crimes like burglary, theft and vehicle jacking. Boston is the biggest and most populous city in the commonwealth of Massachusetts comprising of numerous districts (as seen in figure 1) and as a result is currently estimated by the Uniform Crime Reports (UCR) managed by the Federal Bureau Investigation (FBI) to be the leading city in crime compared to its fellow cities. Some of these Boston districts experience more crime than the others. Estimating crime statistics had been a difficult task for law enforcement before UCR was introduced [1, 2]. The UCR program has improved crime data administration, management and statistical analysis in order to control the occurrence of crimes especially the most violent ones. UCR classifies crime into three main parts based on their severity and level of violence.

Law enforcement is looking towards data mining and machine learning to properly analyze crime data and make attempts in predicting possible future incidents based on crime pattern recognition. Machine learning is a branch of artificial intelligence (AI) and data analytics that enables machines perform operations more skillfully through powerful algorithms capable of recognizing patterns and classifying data used in performing designated tasks [3, 4]. There are numerous robust algorithms of machine learning. These different algorithms are either unsupervised (data driven), supervised (task driven) or reinforcement learning. Some of the most commonly implemented ones include (i) Artificial neural networks (ii) Decision tree (iii) Linear regression (iv) Random forests (v) Logistic regression.

Numerous studies have employed these machine learning algorithms for crime data analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

[5, 6]. The main objective of crime data mining is to recognize patterns in criminal behavior to enhance law enforcement prediction of anticipated crime activities in certain areas in order to prevent them in the future. Linear regression has been used in some works [6], although due to drawbacks such as being limited to linear relationships, it only considers the mean of the dependent variable and is sensitive to outliers [7], other methods were implemented to overcome some of these limitations, for example, logistic regression, decision tree, random forest and artificial neural networks.

Antolos *et al* [8, 9] employed logistic regression to analyze burglary by investigating the relationship between specific predicting factors and burglary occurrence probability. The goal of their research was to understand when and where a burglar would choose to strike a particular residence based on previous burglary activities. Other studies have shown crime activities reporting and prediction using similar method [10]. Decision tree and random forest are also popularly used approaches in crime data analytics. Gutierrez and Leroy [11] explored crime reporting prediction using decision trees and crime victimization survey. Bogomolov *et al* [12] trained a variety of classifiers on a training data following a comparison between 5 methods using a 5-fold cross validation strategy, which showed a decision tree classifier based on Breiman's random forest algorithm to give the best performance in comparison to the others. This study was carried out on data from London Boroughs with the aim of predicting crime from demographics and mobile data.

The advantages of considering these supervised learning methods, on one hand is that probability and ranking estimation are slightly more efficient using logistic regression [4]. On the other hand, decision tree is helpful for performing feature selection and variable screening, while random forest considers multiple decision trees and surmises the best possible result. They also have remarkable robustness and quite easier to interpret and explain to a non-statistically inclined expert [3, 4, 13]. In the present study, we consider decision tree and random forest learning algorithms to analyze Boston crime data from recent years. The article is divided into a theoretical background of machine learning and algorithms used in the work. The visualization and analysis of the data set is presented in section 3. In section 4, we discuss the different models and the computational implementation. We talk about the questions the study answers and also the predictions made based on the UCR information given in the data set. A final comparison between the two models is done based on the results obtained.

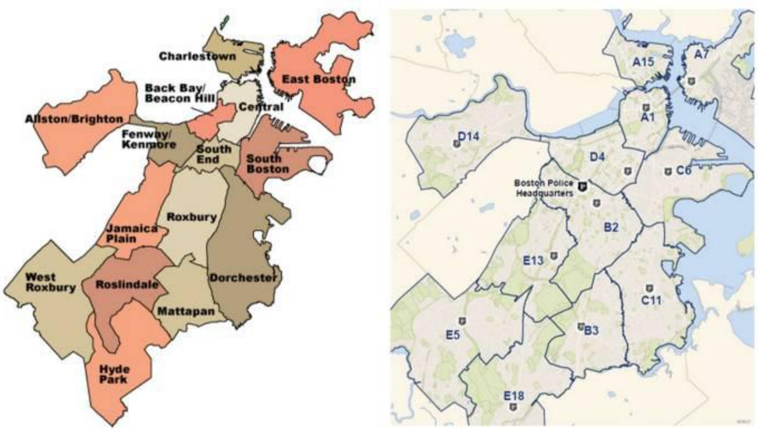


Figure 1. Boston District Map accounting for the main districts in the city [14]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

2. Machine Learning Algorithms

Machine learning as mentioned earlier is a useful tool in analyzing data, performing data extraction and making prediction by implementing efficient algorithms that enables machine to perform their designated tasks cleverly. There are three main classifications of these algorithms (i) Unsupervised learning (ii) Supervised learning (ii) Reinforcement learning algorithms. Supervised learning algorithms are machine learning algorithms that perform tasks based on inferred functions from supervised training data. The objective of these methods is to identify the relationship between input objects (independent variables) and a target attribute (a dependent variable). The relationship is achieved through the algorithm model by predicting the target attribute based on given values of the input objects, this means that in supervised learning the algorithm analyzes the training data to form a basis for accurate description and prediction of an inferred function. The two widely implemented supervised models are classification and regression models.

The learning algorithm is provided with two data sets, the training and the test sets [4, 15]. The task of the algorithm is to establish rules for classifying unlabelled information in the test set by analyzing already labeled information from the training set. The training set comprises of pairs to the *n*th order. Consider a set of measurements of a data point, a_1 and its label, b_1. The training set S_{TRAIN} will be

$$S_{TRAIN} = \langle a_1, b_1 \rangle, \ldots, \langle a_n, b_n \rangle$$

Assume that a_1 is a vector accounting for types of crime classification including severity, area, time of occurrence and other relevant information. The label b_1 could be a classification of the crime with or without shooting. On the other hand, the test sets comprise of unlabelled m measurements as shown below.

$$S_{TEST} = \langle a_{n+1} \rangle, \ldots, \langle a_{n+m} \rangle$$

In our work, we consider two supervised learning algorithms, which are (a) Decision tree and (b) Random Forest. These particular methods were chosen over linear and logistic regression methods due to the type of data set chosen and predictions made.

(I) **Decision Tree:** Decision tree learning is one of the most popular and widely used methods for representing classifiers and inductive inference. It consists of 3 main nodes: root, internal and leaf nodes. Decision tree performs grouping of instances by sorting them from the root to specific leaf nodes. Each leaf node is assigned a class label and has no outgoing branch while in the case of non-leaf nodes, the branches correspond to classifications of instances based on test conditions from posing series of questions about their corresponding characteristics. Some of the advantages that make this method considered as suitable for this data are decision trees can handle heterogeneous data and are easily interpretable [4, 13].

(II) **Random Forest:** Random forest is an ensemble learning method, normally trained with the bagging method that creates a set of decision trees from a random selection of subsets in the training set, which combines the choice from different decision trees to give the test object its final class. This implies that random forest learning combines different learning models to enhance the overall result. The main advantage of random forest over most machine learning algorithm is its applicability in classification and regression problems [16].

3. Data Visualization and Analysis

3.1 Data Source and Description

The dataset selected to carry out this study is a dataset that contains records from recent crime incident report system from the second half of 2015 to the first half of 2018 which classifies the type of incident as well as providing information about the time and geographical location of the incident. The crime data, stored in a csv file, is provided by the Boston department of police and made available on Kaggle Datasets [17]. It contains 17 columns and 328k rows. The format of the data is shown in figure 2 below:

```
> summary(crime)
  INCIDENT_NUMBER    OFFENSE_CODE                       OFFENSE_CODE_GROUP
 I162030584:    13   Min.   : 111   Motor Vehicle Accident Response: 35342
 I152080623:    11   1st Qu.:1001   Larceny                        : 24534
 I172013170:    10   Median :2907   Medical Assistance             : 22351
 I172096394:    10   Mean   :2317   Investigate Person             : 17867
 I162001871:     9   3rd Qu.:3201   Other                          : 17223
 I162071327:     9   Max.   :3831   Drug Violation                 : 15844
 (Other)   :303309                  (Other)                        :170210
                           OFFENSE_DESCRIPTION    DISTRICT      REPORTING_AREA
 INVESTIGATE PERSON                   : 17871   B2     :47770   Min.   :  0.0
 SICK/INJURED/MEDICAL - PERSON        : 17802   C11    :40509   1st Qu.:177.0
 M/V - LEAVING SCENE - PROPERTY DAMAGE: 15556   D4     :39949   Median :343.0
 VANDALISM                            : 14493   A1     :33740   Mean   :383.2
 ASSAULT SIMPLE - BATTERY             : 14051   B3     :33686   3rd Qu.:544.0
 VERBAL DISPUTE                       : 12370   C6     :22133   Max.   :962.0
 (Other)                              :211228   (Other):85584   NA's   :19130
 SHOOTING              OCCURRED_ON_DATE        YEAR           MONTH
  :302402   2017-06-01 00:00:00:    29   Min.   :2015   Min.   : 1.000
 Y:   969   2015-07-01 00:00:00:    27   1st Qu.:2016   1st Qu.: 4.000
            2016-08-01 00:00:00:    27   Median :2016   Median : 7.000
            2015-06-18 05:00:00:    22   Mean   :2016   Mean   : 6.561
            2017-08-01 00:00:00:    22   3rd Qu.:2017   3rd Qu.: 9.000
            2015-12-07 11:38:00:    20   Max.   :2018   Max.   :12.000
            (Other)            :303224

   DAY_OF_WEEK         HOUR              UCR_PART                STREET
 Friday   :46059   Min.   : 0.00               :    90   WASHINGTON ST : 13504
 Monday   :43476   1st Qu.: 9.00   Other       :  1170                 : 10618
 Saturday :42592   Median :14.00   Part One    : 58555   BLUE HILL AVE :  7385
 Sunday   :38262   Mean   :13.12   Part Three:150513   BOYLSTON ST   :  6873
 Thursday :44256   3rd Qu.:18.00   Part Two    : 93043   DORCHESTER AVE:  4907
 Tuesday  :44317   Max.   :23.00                         TREMONT ST    :  4517
 Wednesday:44409                                         (Other)       :255567
      Lat              Long                          Location
 Min.   :-1.00   Min.   :-71.18   (0.00000000, 0.00000000)    : 18839
 1st Qu.:42.30   1st Qu.:-71.10   (42.34862382, -71.08277637):   1183
 Median :42.33   Median :-71.08   (42.36183857, -71.05976489):   1129
 Mean   :42.22   Mean   :-70.92   (42.28482577, -71.09137369):   1072
 3rd Qu.:42.35   3rd Qu.:-71.06   (42.32866284, -71.08563401):    992
 Max.   :42.40   Max.   : -1.00   (42.25621592, -71.12401947):    837
 NA's   :18839   NA's   :18839    (Other)                     :279319
```

Figure 2. Crime Data Summary

Explanation of column names:

INCIDENT_NUMBER: File number registered in the police office

OFFENSE_CODE: Corresponds to a specific kind of crime

OFFENSE_CODE_GROUP: Name of the crime

OFFENSE_DESCRIPTION: More specific name of the crime

DISTRICT: Neighbourhood in Boston

REPORTING_AREA: Place defined by the police

SHOOTING: "Y" stands for cases where shooting occurred

OCCURRED_ON_DATE/YEAR/MONTH/DAY_OF_WEEK/HOUR: Time

UCR_PART: Rate of the crime, part 1 is the highest rank

STREET/LATITUDE/LONGITUDE/LOCATION: Place happened

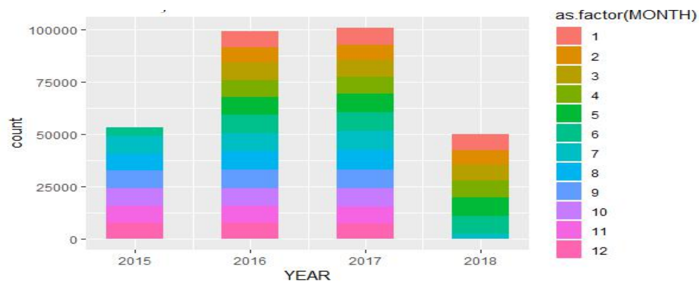
3.2 Visualization method

We used ggplot2 package in R to visualize the data. Different kinds of plot are implemented to analyze the data from different perspectives. In order to accelerate the processing speed we used bigvis package to plot the heat map [18].

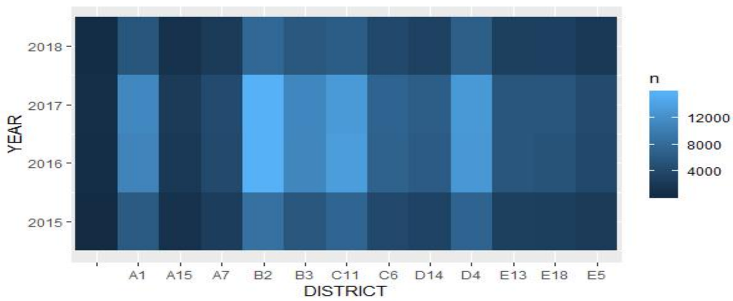
The data can be categorized into 3 subsets. The first is about the place of the crime, for example, street name and coordinates. The second is about the time of the incident, for example, date and hour. The last subset is about the description of the crime, for example, UCR part and offense code group.

3.3 Visualization based on Time

In figure 3a the plot represents the count of crime for each month (in colours) as a function of years. It is important to note that the data of 2015 and 2018 is incomplete. For 2015 the data is from July to December while for 2018 the data is from January to June. Thus in the following visualization if the observations in number of crimes is less in 2015 and 2018, it does not surely imply that the number of crimes decreases these 2 years but the amount of data is less.



(a)



(b)

Figure 3. Number of Crimes by Year as a function of (a) month (in color) and (b) Districts

In figure 3b, the brightness shows the number of crimes in different districts by years. The 2 tiles in the center are brighter than the head and tail due to more data in 2016 and 2017. The first column is the crime cases without district information. If we compare horizontally, the top 3 districts in number of crimes are B-2, D-4 and C-11. The district with the least amount of crimes is A-15.

 doi:10.20944/preprints202002.0108.v1

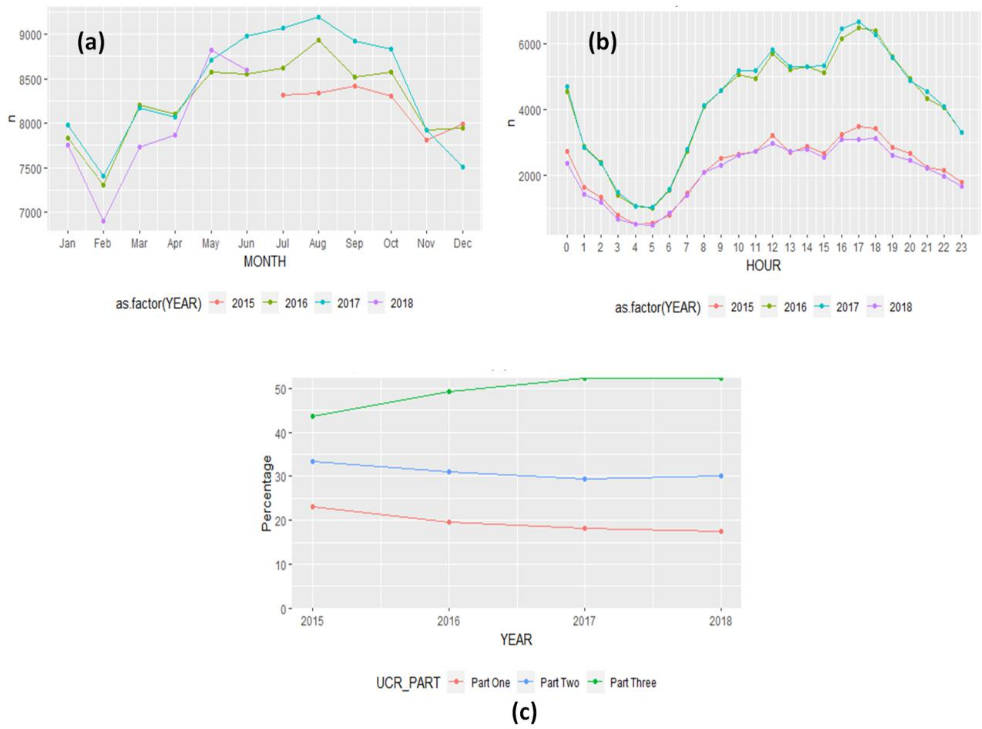


Figure 4. Total number of crimes for each year as a function of (a) year (b) months (c) Percentage of crimes by UCR Parts

In figure 4a, the orange line and purple line are for 2015 and 2018 which do not cover the entire year. The general trend observed is that August and September are the peaks in number of crimes. After these peaks, the number of crime significantly decreases and is at its lowest in February. In figure 4b, which represents crime as a function of the hour, the peak of crime activities shows up in the afternoon at around 17h and it continues decreasing to the lowest point in the early morning around 5h. After 5h, the number of crimes grows and the second peak appears at around 12.

In figure 4c, in order to compensate for the influence of different total numbers of crimes in different years and lack of sufficient data for 2015 and 2018, we convert the crime numbers into percentages by using the UCR parts as categories to see the tendency during the years. In the graph, as the percentages of part one and part two crimes decrease over the years, the percentage of part three crimes increases. Based on the categorization of UCR parts, we can see that part one crimes are the most severe and part three crimes are minor ones. In this case, we can conclude that from 2015 to 2018 the percentage of severe crimes experiences a significant decrease.

3.4 Visualization based on Location

	OFFENSE_CODE_GROUP	n	UCR_PART
	<fct>	<int>	<chr>
1	Larceny	23774	1
2	Larceny From Motor Vehicle	9740	1
3	Aggravated Assault	6904	1
4	Other	15405	2
5	Vandalism	14245	2
6	Simple Assault	14174	2
7	Motor Vehicle Accident Response	29558	3
8	Medical Assistance	21343	3
9	Investigate Person	17172	3

Figure 5. Top 3 Crimes in Each UCR Part

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

Figure 5 shows the top 3 crimes in numbers categorized by UCR parts. Since there are more than 10 kinds of crimes in each UCR part, it is impossible to geographically show all the information in one graph. Hence, only the top 3 crimes are included in the following graphs for each UCR part.

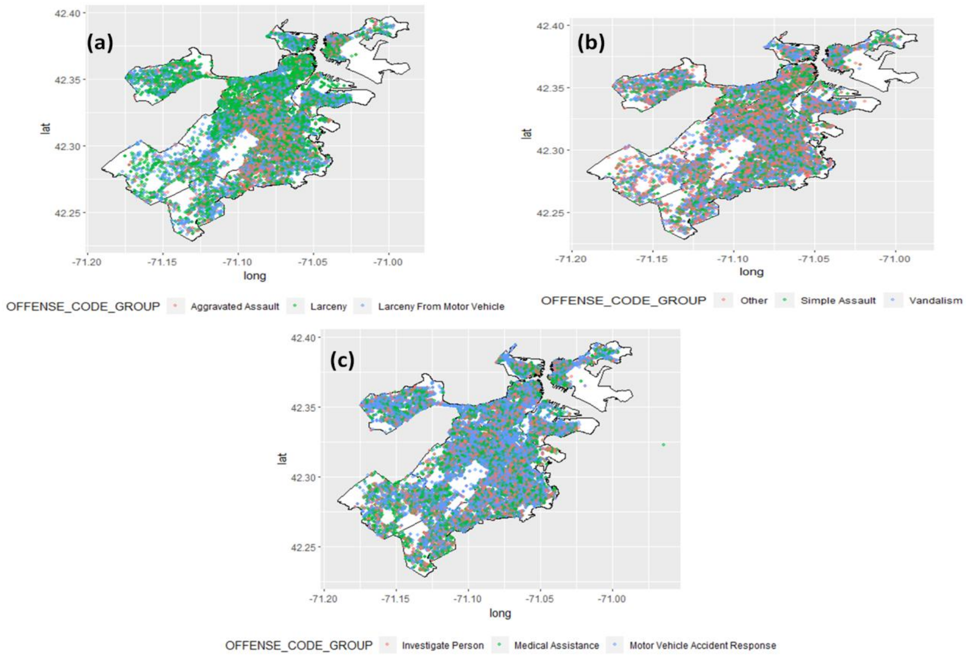


Figure 6. Geographical Distribution of the top 3 crimes in the UCR categories (a) Part 1 (b) Part 2 (c) Part 3

Figure 6 represents the geographical distribution of the crimes in different UCR parts. In figure 6a, the UCR part one crimes, which are the severe crimes, are presented. The color of the dots indicates the type of the crime. From the graph it is clear that larceny is more frequent in Central and Fenway-Kenmore whereas aggravated assault and larceny from motor vehicle are more frequent in Roxbury and South Dorchester. The district names are shown in figure 1. In figure 6b, different colors are more evenly distributed. However, simple assault is more condensed in the center and vandalism is more frequent in the north, for example, Charlestown. In figure 6c, the 3 types of crime are evenly distributed and it is clear that motor vehicle accident response accounts for the biggest percentage on this plot.

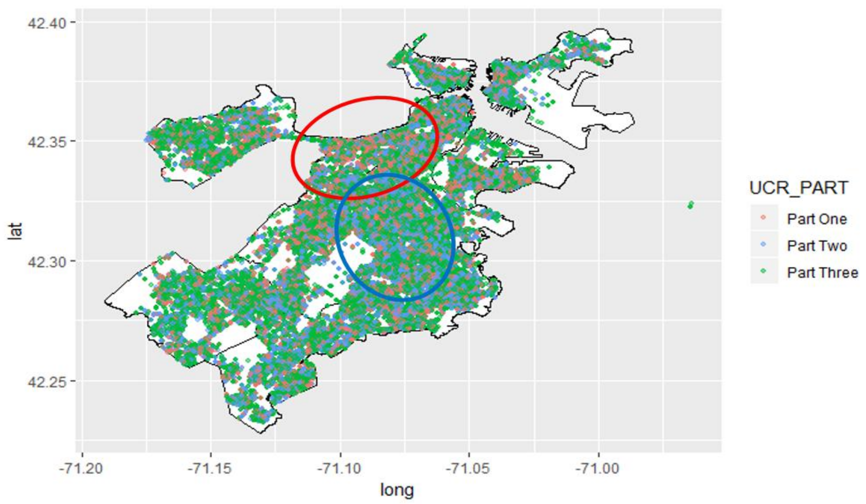


Figure 7. Geographical Distribution of the total number of crimes by all UCR Parts

doi:10.20944/preprints202002.0108.v1

Figure 7 shows a distribution of all crimes categorized in their corresponding UCR part. We can see that part three crimes, which mean minor crimes, are more evenly distributed than the other two. Part one crime is more condensed in Fenway-Kenmore and Back Bay Beacon Hill (the red circle). Meanwhile part two crimes are more aggregated in Roxbury, North Dorchester and South Dorchester.

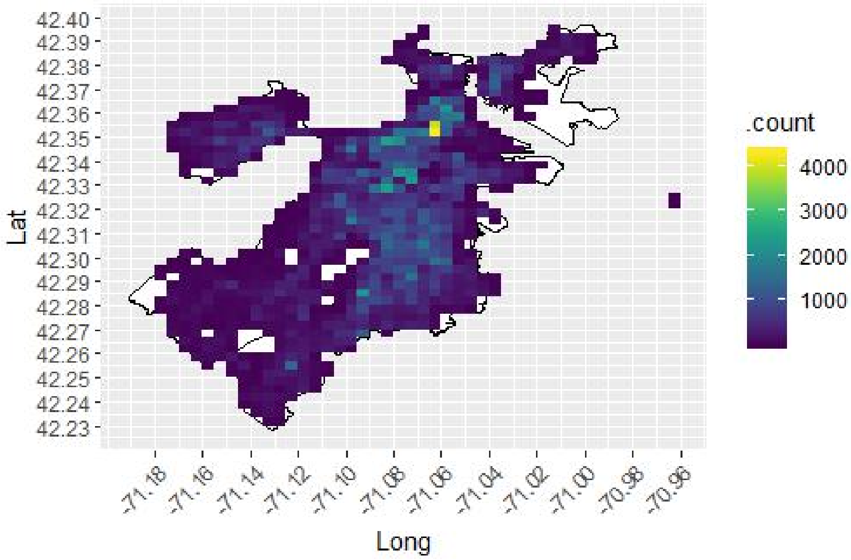


Figure 8. Geographical Crime Heat map in 4 Years with the yellow region being the highest counts of crime incidents

The most frequent place that crimes happened is around (lat: 42.35, lon: -71.06) as seen in figure 8, which is the heat of the city around Park Street Church. Also the surrounding areas of the city have a slightly lower number of crimes because these areas are usually densely populated. As it extends to the fringe, the number of crimes decreases rapidly.

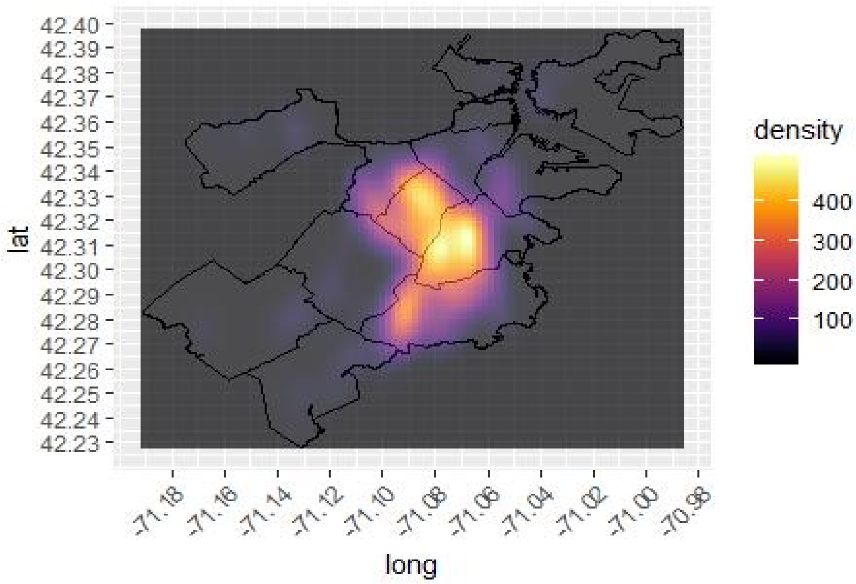


Figure 9. Geographical Shooting Heat map from mid 2015 to mid 2018

The distribution of shooting is very different from the distribution of crimes as a whole as demonstrated in Figure 9. The most frequent districts were shooting occurred are Roxbury and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

South Dorchester.

4. Modeling on Dataset

After trial of different approaches, we decided to use decision tree and random forest methods to model the dataset and predict the outcome. To achieve this using the information provided in the dataset, we proposed a specific question on how the crime type (UCR part) can be predicted from the available data. From the section on visualization we can conclude that crime type is related to location, which is directly linked to the coordinates and also possibly influenced by the time of the day.

4.1 Decision Tree

4.1.1 Tools Used and Model Building Process

To implement decision tree model, we used a package called "rpart" in R. Firstly, we cleaned and prepared the data for modeling. Secondly, we splitted 75% of the data for training the model and the rest 25% for testing the dataset. After that we used rpart() function to implement decision tree mode. After trying different combinations of the independent variables in modeling decision tree, we found the best option, in which case you will get a relatively robust tree that uses longitude, latitude and hour. This is reasonable since from the visualization of the variables it can be concluded that time of the day and the location of the place is highly related with the amount and the type of crime.

In order to generate a robust tree, we set "minsplit" option to 3, based on the dispersive distribution of UCR part one and two data. Minsplit is the minimum number of observations that must exist in a node in order for a split to be attempted. Finally, based on the model built we used data from the test dataset to make predictions and then compare with the real observations in the test dataset. Original data and predictions are visualized in subsequent figures to see the result more clearly.

4.1.2 Results and Analysis

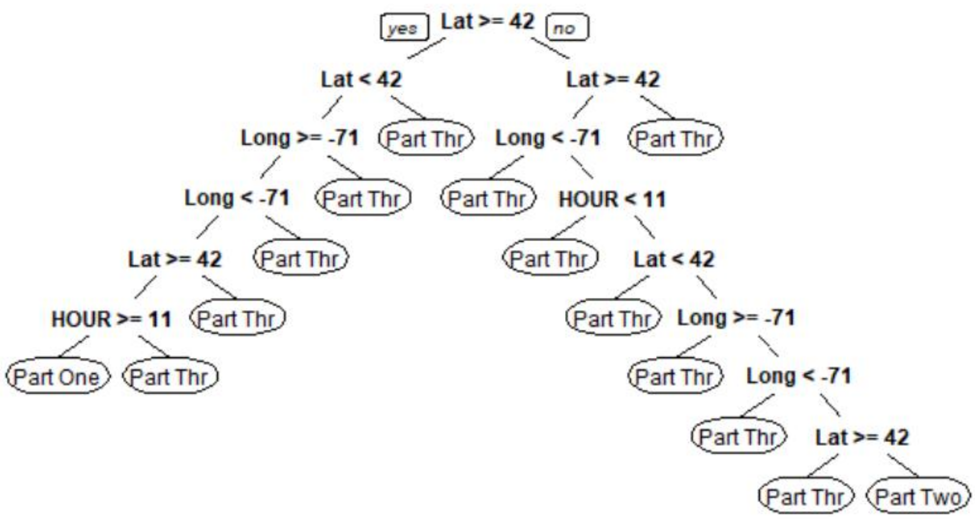


Figure 10. Decision Tree Visualization

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

The majority of the nodes in figure 10 are classifying UCR part three crimes. One possible explanation for this is that the amount of data in UCR part three is more than part one and part two. For training the model, the bigger amount of data can generate more precise prediction.

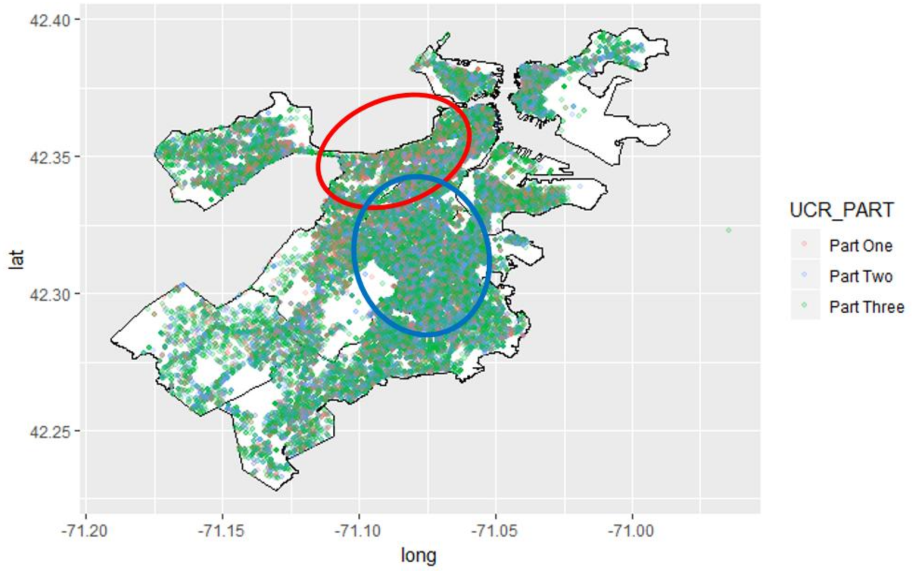


Figure 11. Crime Distribution in Original Test Dataset

Figure 11 is the real observation from the test dataset, which accounts for 25% of the whole data. UCR part three data are more evenly distributed in the whole map whereas we can identify a cluster of part one data in the red circle and a cluster of part two data in the blue circle.

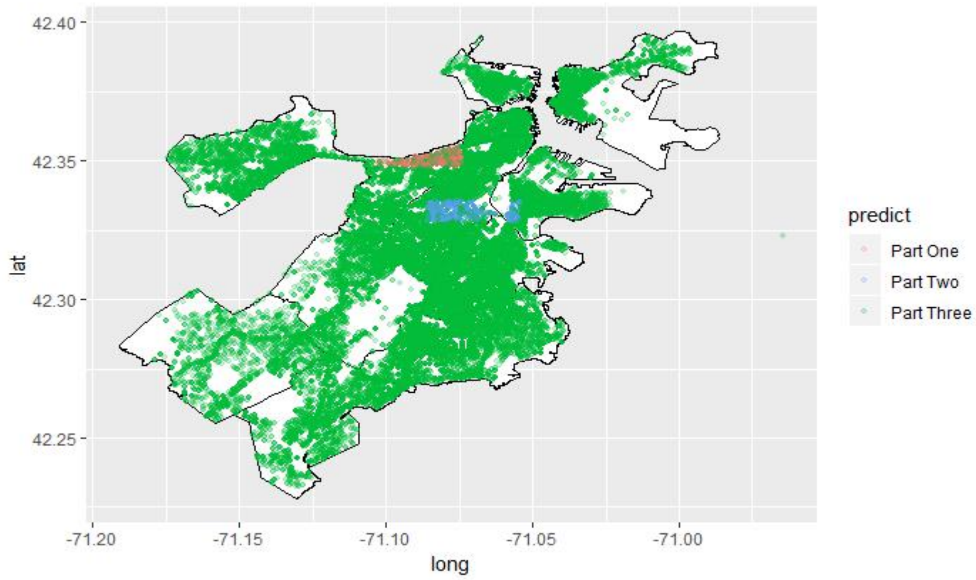


Figure 12. Crime Distribution in Test Dataset Predictions

As seen from the predictions graph in figure 12, the decision tree model approximately identified the most condensed areas for UCR part one and part two. However, the result is not ideal because it greatly simplified the distribution of part one and part two.

4.1.3 Performance of the Model

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

Determination of the performance of the model was judged from 2 perspectives. One is the correctness of the prediction and another is the computational speed of the model. For the first criteria we used confusion matrix. In this case we used errorMatrix() function from the package "rattle". For the second criteria we used system.time() function to calculate the time elapsed during processing.

```
             Predicted
Actual         Part One Part Three Part Two Error
  Part One         1048      14874      524  93.6
  Part Three        692      39400     1144   4.5
  Part Two          444      23680     1338  94.7
```

Figure 13. Confusion Matrix for the Decision Tree Model

The correct percentage for UCR part three is 95.5% and the majority of part one and part two were not accurately predicted as seen from figure 13 with a total accuracy of 50.14%. One possible explanation for the poor result in predicting part one and part two is that the less amount of data in these parts thus the model is not trained well enough in predicting these 2 categories. The time elapsed after testing was 3.17 seconds.

4.2 Random Forest

4.2.1 Tools Used and Model Building Process

To implement decision tree model, we used a package called "randomForest" in R. We employed the same train and test dataset splitted in the previous model. Then we used the function randomForest() to build the model on train dataset. In the first trail we set "ntree" to 100 and "nodesize" to 3. "Ntree" is the number of trees to grow in the random forest model. This should not be set to a very small number, to ensure that every input row gets predicted at least a few times. "Nodesize" is the minimum size of terminal nodes, so setting this number larger generates smaller trees. We used this built model to predict on the test dataset and we visualized the prediction to see the outcome. Afterwards the confusion matrix was generated to check the model performance.

4.2.2 Results and Analysis

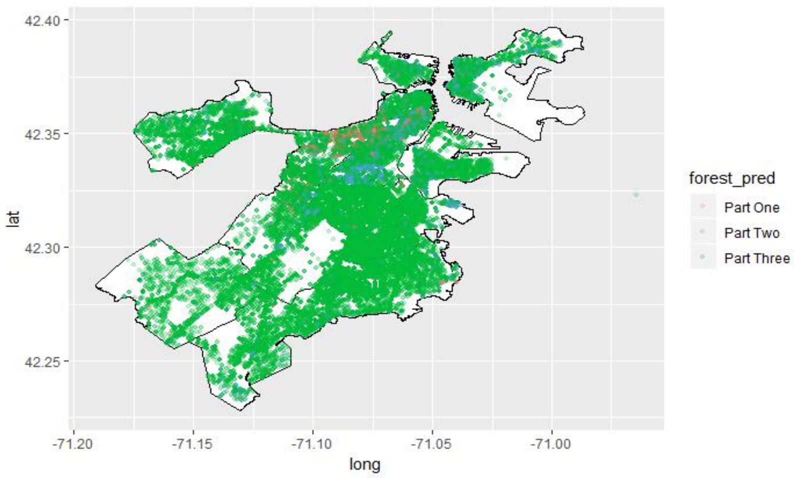


Figure 14. Random Forest Predictions (ntree=100, nodesize=3)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 February 2020 doi:10.20944/preprints202002.0108.v1

In figure 14, it is clear that the condensed areas of part one and part two coincide with the original dataset. However, the number of predictions in part one and part two is less than the real observations in the test dataset. In comparison to the prediction from the decision tree model, the dots for UCR part one and part two are more disperse.

4.2.3 Performance of the Model

```
              Predicted
Actual         Part One Part Three Part Two Error
  Part One        1985       13191     1287  87.9
  Part Three      1109       37362     2590   9.0
  Part Two         842       21235     3541  86.2
```

Figure 15. Confusion Matrix for the Decision Tree Model (ntree=100, nodesize=3)

$$Accuracy = Number\ of\ True\ Predictions/Number\ of\ Total\ Observations$$

(1985+37362+3541)/(1985+13191+1287+1109+37362+2590+842+21235+3541)*100=51.58%

In the case of (ntree=100, nodesize=3), time elapsed is 13.86s. We changed the number of trees generated and the node size of the tree to see how the two factors influence the outcome. We use the same functions to generate the confusion matrix and calculate the time.

Table 1. Summary of the Random Forest Model Performance

Number of Trees	Node Size	Accuracy (%)	Part 1 Error (%)	Part 2 Error (%)	Part 3 Error (%)	Time Elapsed
100	3	51.58	**87.9**	**86.2**	9.0	**13.86s**
200	3	**51.68**	88.3	86.4	8.5	33.95s
100	1	51.59	88.3	86.6	8.6	18.25s
200	1	51.64	88.6	87.3	**7.9**	40.24s

In table 1, the second model is more accurate in predictions whereas the first model consumed less time. Increasing the number of trees can improve the accuracy but the improvement is minor, however, the time consumed increases greatly. Also the smaller node size does not necessarily yield a more accurate result and smaller node size takes more time in processing.

4.3 Models Comparison

4.3.1 Time Consumed and Result Quality

The time consumed in the decision tree model is 3.17 seconds whereas the least time consumed in the four cases of random forest is 13.86 seconds. The accuracy of prediction in the decision tree model is 50.14% and in the cases of random forest, the best result is 51.68% of accuracy. Taking time consumed into account, 970% of increase of time consumed yields a 3.07% improvement in the

 doi:10.20944/preprints202002.0108.v1

accuracy based on the decision tree model.

4.3.2 Discussion

Clearly, results from the both model are not ideal since the accuracy of predictions is around 50%. One possible explanation for this limited performance is the complexity of the crime distribution. This means that only information of location and hour of the day cannot generate an ideal model. Other data information may be needed in order to build a more accurate model, for instance information about victims like gender, age and so on.

Comparing the decision tree model and the random forest model, the difference of accuracy is not significant. Decision tree is one tree whereas random forest is a series of trees generated. Random forest model usually has better outcome compared to decision tree. In our case, it is possible that the decision tree model already took the full potential of the data provided and applying random forest will not improve the outcome significantly since it reaches the bottleneck of the classification competence. Increasing the number of trees in the random forest model can slightly improve the accuracy but this is at the expense of time consumed in processing.

Due to the complexity of the data, for future works one possible approach is to apply more advanced models for example neural networks to generate better models and circumvent the limitations from data information.

5. Conclusions

The present work accounts for four main steps. Firstly, we reviewed theoretical concepts of modeling methods in order to understand the models applied to the dataset. Secondly, we visualized the dataset from different perspectives, this process helped us to better understand the data and generate ideas on how to model the data. Furthermore, based on the information obtained we built the model and used the model to predict on the test dataset.

Finally, we compared the results and came up with some possible explanation for the outcome. In the modeling part, we applied 2 models, one is a decision tree and the other is random forest. Theoretically these models are ideal for classification problems. From the results, we can see that random forest model is slightly better; however, this minor improvement is at the expense of more time consumed. The work presents a preliminary background for future applications in data analysis on crime statistics based on geographical and temporal characteristics and extendable to other data sets in fields of law and order, business and data science.

Supplementary Materials: The dataset, R code implementing the ML models, shapefiles [19] for Boston maps are available online at www.dropbox.com/s/7r05fag4z4vhsh9/Boston_Crime.zip?dl=0.

Author Contributions: J.Y. designed the original codes and contributed in analysing the results and writing the article, I.A.M. wrote the introduction and parts of the result analysis, I.J.A. modified the codes, contributed in writing and reviewing the article.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UCR Crime reporting handbook. Available online (accessed on 25 October 2018). https://ucr.fbi.gov/additional-ucr-publications/ucr_handbook.pdf
2. FBI UCR Crime Statistics – Massachusetts. Available online (accessed on 25 October 2018).

https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/table-6-state-cuts/massachusetts.xls

3. Smola, A. and Vishwanathan, S.V.N. Introduction to Machine learning, 2008, 1st ed., Cambridge University Press, Cambridge: UK, ISBN 0521825830
4. Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997, ISBN: 0070428077
5. Lin, Y-L., Chen, T-Y., Yu, L-C. Using machine learning to assist crime prediction, 6th IEEE-IIAI International Congress on Advanced Applied Informatics, 9-13 July 2017, DOI: 10.1109/IIAI-AAI.2017.46, Hamamatsu, Japan.
6. McClendon, L. and Meghanathan, N (2015), Using machine learning algorithms to analyze crime data, Machine Learning and Applications: An International Journal (MLAIJ), 2(1): 1 – 12. DOI: 10.5121/mlaij.2015.2101.
7. SCIENCING – The disadvantages of linear regression. Available online (accessed on 25 October 2018). https://sciencing.com/disadvantages-linear-regression-8562780.html
8. Antolos, D. (2011), Investigating Factors Associated with Burglary Crime Analysis using Logistic Regression Modeling. Dissertations and Theses. Available online (accessed on 25 October 2018). https://commons.erau.edu/cgi/viewcontent.cgi?article=1014&context=edt
9. Antolos D., Liu D., Ludu A., Vincenzi D. (2013), Burglary Crime Analysis Using Logistic Regression. In: Yamamoto S. (eds) Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business,. HIMI 2013. Lecture Notes in Computer Science, vol 8018. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-39226-9_60
10. Peng, C-Y.J., Lee, K.L., Ingersoll, G.M. (2002) An introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, 96(1): 3 – 14.
11. Gutierrez, J. and Leroy, G. (2007) Predicting crime reporting with decision trees and the national crime victimization survey, Proceedings of the 13th Americas Conference on Information systems (ACIS), 10-12 August, Keystone, Colorado, 2007, pp. 1 – 10.
12. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey — November 12 - 16, 2014, Pages 427 – 434.
13. Rokach, L. and Maimon, O., Data mining with decision trees: Theory and Applications, World Scientific Publishing, NJ, USA (2014), 2nd ed., ISBN 9789814590075, 2014.
14. Map of Boston districts - Where are the safest neighbourhoods in Boston - Jumpshell. Available online (08 January 2020). https://www.jumpshell.com/posts/safest-neighborhoods-in-boston
15. Learned-Miller, E.G. (2014) Introduction to Supervised learning. Available online (accessed on 25 October 2018). https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf
16. Louppe, G. (2015), Understanding Random Forests: From Theory to Practice, PhD dissertation, University of Liege.
17. Crimes in Boston Data set. Available online (accessed on 12 October 2018). https://www.kaggle.com/ankkur13/boston-crime-data/home

 doi:10.20944/preprints202002.0108.v1

18. Wickham, H., Grolemund, G. R for Data Science (Import, Tidy, Transform, Visualize, and Model Data). O'REILLY. 2016
19. Lmullen – GitHub. How I use shapefiles in R with ggplot2 and RGDAL. https://gist.github.com/lmullen/8375785.