Dean&Francis
ISSN 2959-6157

Studying the factors that influence the Incidence of UCR Part One Crimes in Boston

Yanxi He[1, *],

Maomao Xie[2],

Lei Lin[3],

Taoyuan Ding[4]

[1]Alcanta International College GuangZhou, 511400, China, adler3236@163.com
[2]Guangdong Country Garden School, FoShan, 528312, China, maoxie0812@outlook.com
[3]Andrews Osborne Academy, Willoughby, OH, 44094, U.S, lei771117@gmail.com
[4]St. Jude's Academy, Mississauga, ON L5N 2M6, Canada, 2120070910@qq.com

*Corresponding auhor email: adler3236@163.com

Abstract:

Boston, as one of the most populated cities in the United States, while being known for its rich history, humanities and culture, also faces significant challenges with crime. This paper presents research on the critical factors that influence a possibility of a crime in Boston being UCR part 1 crime. If applied to law enforcement agencies, the distribution of police resources in Boston can be more reasonable and enable better urban security. We obtained the dataset from a Kaggle post, which was obtained from the Crime Incident Report of the Boston Police Department. By cleaning the data, converting months to seasons, we made the data more usable. We then split the data in a 70-30% ratio. We ran two models – first for the training session, we used the random forest classifier to identify the most significant factors. We evaluated the model by using metrics such as accuracy, precision, recall, and the ROC-AUC curve. In the testing session, we employed a logistic regression model for cross validation. For both models, the obtained accuracy is over 94%, indicating an extremely high overall performance for both models. Through the data analysis and test in this paper, it can be summarized that both the logistic regression model and random forest classifier can effectively predict and analyze an instance of a UCR part 1 crime in Boston based on the location of the crime, crime code group, season, time of the day, and day of the week. Though the accuracy is significantly high, we should not ignore the fact that we used a slightly older dataset ranging from 2015 to 2018, which, being six years ago, may have intrinsically different crime patterns than those that occur now.

Keywords: Crime, Boston, Cross-validation, Random Forest Classifier, Logistic, Regression

ISSN 2959-6157

1. Introduction

Crime analysis in cities has often been something focused on by researchers and policymakers. The complexity that crime has, being influenced by a mix of factors like social, economic, and environmental ones, requires an approach that both understands and tries to reduce crime. Some older studies usually looked at crime's changes by seasons and its spread over different areas, showing that analyzing things locally is more useful than making general conclusions. For instance, Cohen and Gorr found seasonality in crime varies across neighborhoods, making interventions at a neighborhood level better than broad policies for the whole city. The crime data for Boston, which comes from the Boston Police Department, gives a good foundation for looking into it. This dataset covered more than 320,000 crime events from 2015 until 2018, making the possibility of delving into crime tendencies and aspects. Data cleaning was done after, grouping months into the seasons helped accuracy improvement, and facilitated making better guesses and drawing insights from it.

This research applied two models of machine learning: the random forest and logistic regression. The selection of these models is because of their suitability for handling complex and large data collections. The random forest recognized key elements affecting UCR Part One crimes, whereas logistic regression was used for checking and validating the model, assisting in getting reliable results.

A notable contribution of study happens to be about time and space having effects on crime forecasting. As work from Gorr and others show, crime increases at certain periods during the year[2]. Also, spreading of crimes within city, shaped by urban designs and socio-economic conditions, is important in finding crime places[3].

Boston's crime history, considering big events, has many that affected the city's thinking on law order and also crime controlling. One case in the 1960s, which a man killed 13 women, named Boston Strangler, pointed out the need for better crime investigations[4].

Another example is more recent, in 2013, the bombing that occurred during the Boston Marathon, making people see the importance of emergency responses and having agencies work together[5]. Because of these, Boston has increased its effort toward crime analysis and how to prevent it.

This study found that models like random forest and logistic regression were able to predict UCR Part One crimes at a high rate, more than 94% accuracy was seen. The success shows that these models can be helpful tools for seeing crime data that is complex, also giving useful insights for law enforcement to apply. What was identified as important factors included where crimes happen, the code groups used, seasons, times of day, and what day of the week it is. For instance, districts like B2, C11, and D4 were found to have generally higher crime rates[6].

Crime patterns are influenced by the time in the day. The most crimes, generally, happen in the afternoon time, between 12 PM to 8 PM, during any day of the week, but especially more on Fridays and Sundays. This kind of thing is explained by routine activity theory, which shows that crime happens when the people who commit crimes and targets, with no protection, come together.

Socioeconomic and environmental aspects are connected to how much crime an area has. Places where more crimes happen, like property, violent or public order crimes, usually have lower economic levels, more people living there, and higher numbers of people moving in and out. This idea supports social disorganization theory, saying that weaker communities with less social ties will experience more crime happening.

2. Method

2.1 Data Pre-processing

The dataset we used is from the 2015 to 2018 Boston Police Department's Crime Incident Reports. It includes about 320,000 records of crime incidents, and before using the dataset, we went through several important steps to ensure it was clean and accurate.

The first step was data cleaning. For missing values, we used the median for numerical data and the mode for categorical data to fill in the gaps. Also, duplicate entries were removed, and thus redundancy was prevented. Irrelevant chart variables like administrative codes or internal IDs were discarded from analysis and code.

Data was cleaned and then data transformation was taken up to make dataset in a way that suits analysis better. Time-related variables, for example, date and time of incidents, got converted to categories that are sensible. Months were grouped into seasons (Winter, Spring, Summer, Fall) to take a look at how crime changes with seasons. One-hot encoding was applied to convert categorical variables such as crime type and district to machine learning suitable formats.

Then the next step was feature engineering which means new features creation for models' predictive ability boosting. Some important features have become day of the week and time of day, divided into parts like morning, afternoon, evening, and night parts. We had also looked at the crime code group making it clear what kind of crime it is. Location data was involved like the district place where the crime happened. Such steps are essential for dataset preparation for model training as seen in figures that show

crimes distribution through day parts, district parts, and crime types.

Figure 1

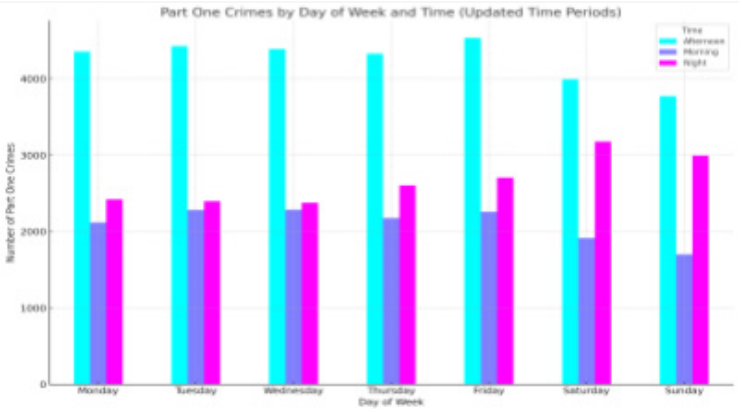
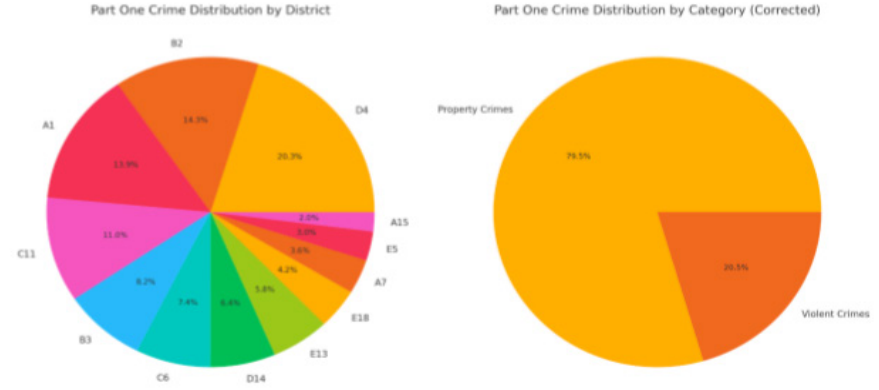


Figure 2



2.2 Model Construction

The random forest classifier was selected for its robustness and ability to handle high-dimensional data. The model construction involved splitting the dataset into training (70%) and testing (30%) sets to evaluate the model's performance. The random forest classifier was trained on the training set, with parameters tuned using grid search and cross-validation to optimize performance.

The random forest classifier was used to identify the most significant features influencing UCR Part One crimes. This model provided insights into which factors were most predictive of crime occurrence, thereby guiding resource allocation for law enforcement.

For the testing phase, we selected logistic regression because it's simple and easy to interpret. We trained this model on the same training set as the random forest. Its performance was then evaluated on the testing set using accuracy, precision, recall, and the F1 score. Logistic regression helped confirm the findings from the random forest, ensuring that the results were consistent across different approaches.

2.3 Model Evaluation

Both models underwent evaluations using some key metrics, no high-tech stuff. The metric like accuracy tells predictions correctly from all. Precision checks how many true positives exist from positive predictions. As for recall, it tells true positive predictions from actual positive cases. F1 score balances out precision with recall to just one value.

True positives, true negatives, false positives, and false negatives placed in a confusion matrix helped in showing model performance. K-fold cross-validation was used for model checking, with 10 as the k value for better judgment. Data split ten times, so training/testing ten times each using different parts as a test set. This is done to reduce overfitting risk and ensure more consistent performance across.

2.4 Implementation Details

Python was what we used to build the models with known data science libraries like Pandas for the data handling, with Scikit-learn we did machine learning, and with Matplotlib we created visualizations. These models were run

ISSN 2959-6157

on a workstation having the capability to manage the big dataset and heavy calculations needed for both training and evaluation.

In closing, the approach taken by this study was thorough which included care in data pre-processing, the strong use of machine learning models, and methods of evaluation were detailed to make sure findings are reliable and accurate. Not only this study gave insights on what factors affect UCR Part One crimes in Boston, but also it showed how advanced analytics can help to improve the strategies in law enforcement and also help to make cities safer.

3. Results and Analysis

3.1 First Trial: Random Forest Classifier

In the initial attempt with the Random Forest Classifier, it was so all features from the dataset were included. Feature importance plot showed "OFFENSE_CODE" and it was almost absolutely correlated to crimes in UCR Part One, being the most dominant feature. This showed the model depended mostly on this one variable, and it was influencing other factors' impact.

In the first trial, classification reports showed very high performance measures, with precision, recall, and F1-score nearing 1.0 for both. Confusion matrix showed hardly any misclassifications, confirming near-perfect model accuracy. This perfect-like performance came mainly because of including the "OFFENSE_CODE" variable.

Despite the model's excellent performance before, the excessive sway by the "OFFENSE_CODE" variable led the model's performance to be skewed. Due to this, for next steps in the analysis, "OFFENSE_CODE" was left out so other variables could contribute more effectively to the model which we aimed for.

Figure 3

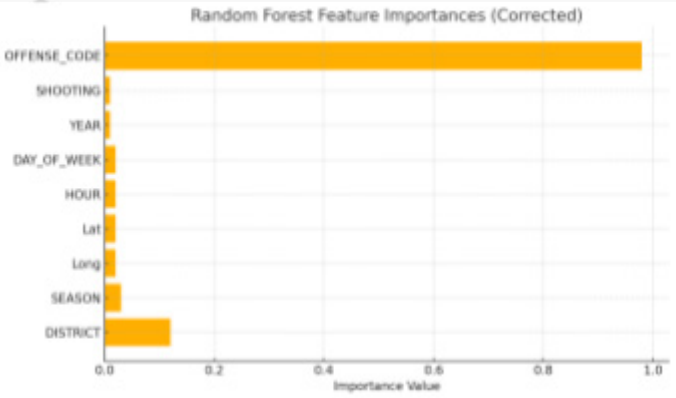


Table 1: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.999973	0.999620	0.999797	73733
1	0.998445	0.999889	0.999166	17980
Accuracy	**0.999673**	**0.999673**	**0.999673**	-
Macro Avg	0.999209	0.999755	0.999481	91713
Weighted Avg	0.999673	0.999673	0.999673	91713

Table 2: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	73705	28
Actual Positive	2	17978

3.2 Second Trial: Random Forest Classifier

During the second go, excluding the "OFFENSE_CODE" variable, the plot showing feature importance now displayed a much more even spread among the remaining variables that were present. "HOUR" of day and "DAY_OF_WEEK" became significant predictors, next to seasonal variables and various districts.

The classification report, which was not as good as the first one but still high in performance metrics. A 95% accuracy was reached, precision and recall balanced both positive and negative classes reasonably well. The confusion matrix showed some misclassification instances, but the model depended more on various features rather than

being too focused on one feature.

With balance improved between features and high metrics noted, it's seen that removing the "OFFENSE_CODE" variable increased the model's capacity in understanding deeper patterns in UCR Part One crimes, leading to results more reliable and easier to interpret.

Figure 4

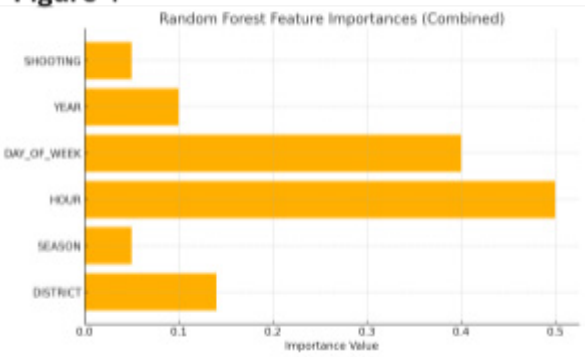


Table 3: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.96	0.98	0.97	73733
1	0.91	0.83	0.87	17980
Accuracy	**0.95**	**0.95**	**0.95**	-
Macro Avg	0.94	0.91	0.92	91713
Weighted Avg	0.95	0.95	0.95	91713

Table 4: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	72058	1675
Actual Positive	3036	14944

3.3 Logistic Regression Model

Following the Random Forest trials, a Logistic Regression model was implemented to validate the findings. The feature importance in logistic regression aligned with those identified in the Random Forest model, with "HOUR," "DAY_OF_WEEK," and specific districts being significant predictors.

The classification report for the Logistic Regression model showed an overall accuracy of 94%. The precision and recall metrics were slightly lower than those of the Random Forest model, with the positive class (UCR Part One crimes) showing a recall of 0.80 and an F1-score of 0.84.

Although the performance of Logistic Regression was lower slightly, yet it still displayed predictive power very strongly. This adds importance to time-related and place-related elements that were focused on in feature analysis.

3.4 Combined Insights

Insights important uncovered by both models, Random Forest also Logistic Regression. Especially hour of the day and week day related to time were main predictors for UCR Part One crimes. Higher frequencies of crimes occurred afternoon and night, with an increase on Fridays and Sundays. This matches with routine activity theory, which tells crime is more likely when public places have offenders and victims meeting.

Important are spatial factors too, with the highest crime rates noticed in B2, C11, and D4 districts. Hinting that parts of Boston become more prone to crime, due to possibly varied socio economic states, density of the population, or how the city is arranged.

To sum up, the insights from Random Forest and Logistic Regression models give a broad view on factors that lead to UCR Part One crimes in Boston. Without the use of the "OFFENSE_CODE" variable, models were balanced, and accuracy increased, stressing on significance of both time and place. These findings present useful pointers for law enforcement to better aim at high-risk zones and times to mitigate crime.

ISSN 2959-6157

4. Conclusion

The study looks in detail at the things affecting UCR Part One crimes all through Boston, and advanced machine learning methods were used which pointed out major insights beneficial for those in law enforcement. Discovering main temporal and spatial patterns, together with socio-economic and environmental aspects, is the study's focus to show the potential for better crime prevention strategies. For a growing Boston, these insights have importance for keeping people safe. Using a mix of advanced analytics and community collaboration and also policy actions, there seems to be a fruitful path to handle urban crime problems in this city.

References

1. Cohen, J., Gorr, W., & Durso, C. (2003). Estimation of crime seasonality: a
cross-sectional extension to time series classical decomposition. *H. John Heinz III Working Paper, (2003-18)*.

2. Towers, S., Chen, S., Malik, A., & Ebert, D. (2018). Factors influencing temporal patterns in crime in a large American city: A predictive analytics perspective. *PLoS one, 13*(10), e0205151.

3. Mapou, A. E., Shendell, D., Ohman-Strickland, P., Madrigano, J., Meng, Q., Whytlaw, J., & Miller, J. (2017). Environmental factors and fluctuations in daily crime rates. *Journal of Environmental Health, 80*(5), 8-22.

4. Boston Police Department. (2018). Crime Incident Reports.

5. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588-608.

6. Hu, J., Hu, X., Lin, Y., Wu, H., & Shen, B. (2024). Exploring the correlation between temperature and crime: A case-crossover study of eight cities in America. *Journal of Safety Science and Resilience*, *5*(1), 13-36.

7. Shaw, C. R., & McKay, H. D. (1969). Juvenile Delinquency and Urban Areas. *University
of Chicago Press*.

8. Delgado, R., & Sánchez-Delgado, H. (2023). The effect of seasonality in predicting the level of crime. A spatial perspective.