Crime Hotspot Analysis

Rashmita Vaggu

December,11 2023

1 Introduction

Predictive Crime Hotspot Analysis in Portland, Oregon

Urban safety and crime prevention are critical concerns in modern cities. In Portland, Oregon, like in many urban areas, understanding and predicting crime patterns plays a crucial role in public safety and resource allocation. This project addresses the specific problem of identifying crime hotspots within Portland. A crime hotspot is defined as a geographical area where crime rates are significantly higher than average. The ability to accurately predict these hotspots can lead to more effective policing strategies, improved resource allocation, and ultimately, enhanced safety for the community.

The approach in this project involves analyzing police call-for-service records to identify patterns that could help predict future crime occurrences. By applying machine learning techniques to this spatial-temporal data, we aim to develop a model that not only highlights current hotspots but also forecasts potential future areas of high crime activity. The outcome of this project holds significant value providing insights to make informed decisions and implement targeted actions to reduce crime rates and improve public safety in Portland.

Motivation of the project

The desire to use data-driven ways to improve urban safety in Portland, Oregon, is the driving force behind this project. Among our objectives are:

- Predicting possible crime hotspots allows law enforcement to transition from reactive to proactive tactics.

- Allocating resources more effectively means putting police and public safety resources where they are most needed.

- Making use of data analytics to give legislators, law enforcement, and urban planners useful information.

- Using well-informed urban safety techniques, Portland neighborhoods' general safety and quality of life are improved.

The goal of this study is to show how machine learning may be an effective tool for predictive policing, with major advantages for community welfare and public safety.

Overview
To locate and forecast Portland, Oregon's crime hotspots, our initiative used a data-driven methodology. To represent the physical dimensions of the city, we organized the police call-for-service records data into a 50x50 grid. Preprocessing the data for consistency, extracting pertinent time-based variables, and using machine learning models for analysis were important stages. The Random Forest Regressor and Gradient Boosting Regressor are two models that were selected because of their ability to handle complex datasets with reliability and efficiency. These models were trained with an emphasis on accuracy and possible hotspot identification to forecast the incidence of crime in each grid cell.

Results
Our models' assessment showed how well they predicted crime rates and pinpointed hotspots. With a large Predictive Efficiency Index (PEI), the Random Forest model showed strong performance and the capacity to precisely identify high-crime locations. In a similar vein, the Gradient Boosting model's R-squared and Mean Absolute Error (MAE) values demonstrated similar accuracy. All things considered, these findings highlight how machine learning may support efforts to promote urban safety by giving law enforcement and municipal planners useful information for allocating resources and making strategic decisions.

2 Related Work

In the last several years, there has been a noticeable increase in interest in the field of machine learning applications in crime investigation and public safety. Several research studies and projects have investigated different facets of hotspot identification and crime prediction, offering insightful information and alternative methods pertinent to our study.

- Spatial-Temporal Crime Prediction: Several studies have focused on the spatial and temporal aspects of crime, similar to our approach. For example, research by Malleson, and Andresen (2015) and Chainey et al. (2008) showed how to detect crime hotspots using spatial data analysis approaches like kernel density estimation. The foundation for incorporating spatial data into crime research was established by these studies.

- Machine Learning in Crime Analysis: Recent research has shown interest in the application of machine learning models, notably those based on decision trees, such as Random Forests and Gradient Boosting. Wang et al. (2012) conducted a noteworthy study that demonstrated the potential of these strategies in predicting crime incidences with a substantial degree of success. Their methodology closely follows ours, demonstrating how well these models handle complicated, non-linear data patterns found in metropolitan crime datasets.

- Predictive Policing: The concept of predictive policing, where data analytics is used to forecast potential criminal activities, has been explored extensively. According to Mohler et al. (2015), a seminal project in this field was the PredPol tool, which used predictive analytics to produce crime estimates. While our study does not employ the same techniques as PredPol (e.g., the self-exciting point process model), the idea of leveraging data-driven predictions to improve policing tactics is shared.

References: Chainey, S., Tompson, L., Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal, 21(1-2), 4-28. DOI: 10.1057/palgrave.sj.8350066

Wang, X., Brown, D. E., Gerber, M. S. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In 2012 IEEE International Conference on Intelligence and Security Informatics (pp. 36-41). IEEE. DOI: 10.1109/ISI.2012.6284084

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., Tita, G. E. (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399-1411.

3 Methods

A large and thorough dataset with a particular focus on Portland, Oregon served as the basis for our crime hotspot study effort. The steps involved in gathering and choosing our data are described below:

Data Source: The Portland Police Bureau provided the police call-for-service records that made up the dataset. This publicly accessible dataset included comprehensive records of all citywide police calls and crime reports.

Time-Period: The data from March through May of 2017 was analyzed. This time frame was chosen to record crime patterns in the spring, a season that is usually linked to variable crime rates because of seasonal variations.

Data Attributes: The crime type, location, and incident date and time were among the dataset's primary characteristics. These characteristics were essential for carrying out geographical and temporal studies.

Data Mining Pipeline The data preprocessing phase was vital to ensure the integrity and usability of the dataset for effective analysis. Here are the steps we took in more detail:

Initial Data Inspection: The dataset was loaded into a Pandas DataFrame and conducted an initial inspection using methods like .info() and .describe() to understand the data's structure, missing values, and basic statistical properties.

This step helped to identify that the ’censustract’ field had missing values, necessitating further scrutiny.

Data Cleaning: The missing values were handled, especially in critical fields. For the ’censustract’, we examined the implications of these missing values and determined the best course of action, which could include imputation or removal, depending on their impact on the analysis. Consistency was ensured in data formats, particularly for categorical data, and checked for any anomalies or irregularities in data entries.

Date and Time Conversion: The ’occdate’ field was converted from string format to a Python datetime format to facilitate temporal analysis. This conversion was crucial for extracting meaningful time-based features and for aggregating data based on periods.

Spatial Mapping to Grid Cells: Spatial grid over Portland was defined with specified dimensions (50x50). Each crime incident, identified by its x and y coordinates, was mapped to a corresponding grid cell. This involved calculating the row and column for each incident based on its coordinates and the grid’s dimensions. This spatial mapping allowed us to transform the data from individual crime records into a format suitable for analyzing spatial patterns.

Feature Engineering - Day of the Week: From the ’occdate’ field, we extracted the day of the week as a new feature, recognizing that crime patterns might vary depending on the day. This feature could provide our models with additional context when making predictions.

Handling Geospatial Data: Using osmnx and geopandas, we visualized the spatial distribution of crimes. This step was not only crucial for initial exploratory analysis but also helped validate the correctness of our spatial mapping.

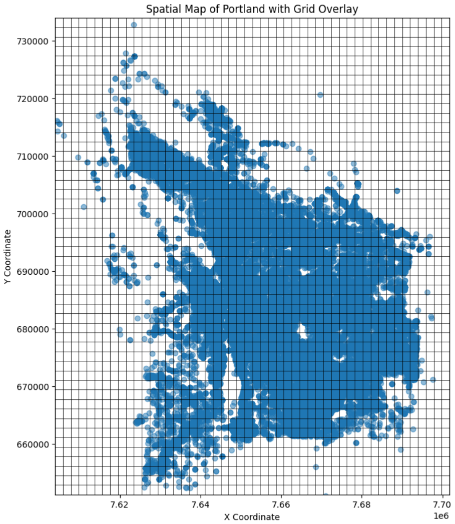


Figure 1: Spatial map of Portland

Aggregating Data for the Space-Time Cube: We aggregated the data by the newly created periods (weekly) and grid cells. This aggregation was performed using a group by operation in Pandas, counting the number of crimes in each grid cell for each period.

The result was a space-time cube, a pivotal structure for our subsequent predictive modeling. By meticulously preprocessing the data, we ensured that our dataset was primed for effective machine learning modeling, providing a reliable foundation for our analyses.

Model Evaluation Analyses

To assess the performance and effectiveness of our predictive models, we conducted comprehensive evaluation analyses. These analyses not only provided insights into the accuracy of the models but also their capability to predict crime hotspots effectively. Here are the detailed steps and methodologies we employed:

1. Evaluation Metrics: Mean Absolute Error (MAE): This metric provided an average of the absolute errors between the predicted and actual crime counts, offering a straightforward measure of prediction accuracy.

R-squared (R^2): We used this metric to assess the proportion of variance in the crime counts that was explained by the models. A higher R^2 indicated better model performance.

Predictive Efficiency Index (PEI): Specifically used for evaluating the model's ability to identify crime hotspots, PEI compared the number of crimes in predicted hotspots with the number in actual hotspots.

2. Model Comparison: We compared the Random Forest and Gradient Boosting models using the above metrics. This comparison allowed us to understand the strengths and weaknesses of each model in the context of crime prediction and hotspot identification.

3. Residual Analysis: By plotting the residuals (the differences between the actual and predicted values), we gained insights into any systematic errors made by the models. This analysis was crucial in identifying any biases or trends that the models might have.

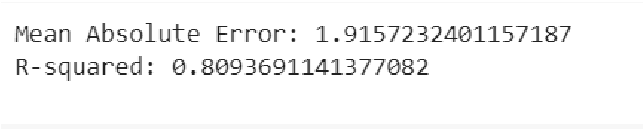
4. Distribution Analysis: We plotted the distribution of actual and predicted crime counts to visually assess how well the models' predictions aligned with the actual data. This step was particularly useful in understanding the models' performance across different ranges of crime counts.

For this project, I utilized a variety of software tools and libraries, predominantly leveraging the capabilities of Google Colab, a cloud-based Python development environment.

4 Results and Discussion

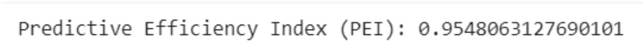
Our project aimed to predict and analyze crime hotspots in Portland using machine learning techniques. The results from the two models we employed, Random Forest and Gradient Boosting, provided insightful findings:

1. Model Performance: The Random Forest Regressor showed a Mean Absolute Error (MAE) of 1.915 and an R-squared value of 0.809, indicating its effectiveness in accurately predicting crime counts. Its Predictive Efficiency Index (PEI) of 0.954 demonstrated a strong capability in identifying crime hotspots.



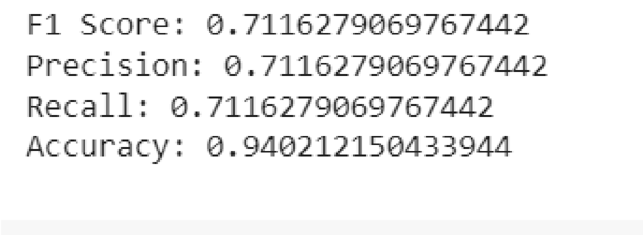
```
Mean Absolute Error: 1.9157232401157187
R-squared: 0.8093691141377082
```

Figure 2: MAE and r2 of Random Forest



```
Predictive Efficiency Index (PEI): 0.9548063127690101
```

Figure 3: PEI of Random Forest



```
F1 Score: 0.7116279069767442
Precision: 0.7116279069767442
Recall: 0.7116279069767442
Accuracy: 0.940212150433944
```

Figure 4: evaluation metrics of Random forest

The Gradient Boosting Regressor displayed a comparable performance with an MAE of 2.822 and an R-squared of 0.659. Its PEI of 0.871 suggested a similarly effective prediction of crime hotspots. These results indicate that both models were effective in capturing the underlying patterns of crime occurrences in Portland, with Gradient Boosting showing a slightly higher sensitivity to variations in the data.

2. Visualizations and Analysis: Distribution Plots: I created plots to visualize the distribution of actual and predicted crime counts. These visualizations were crucial in comparing the models' predictions against the real data, highlighting their accuracy across various crime count ranges.

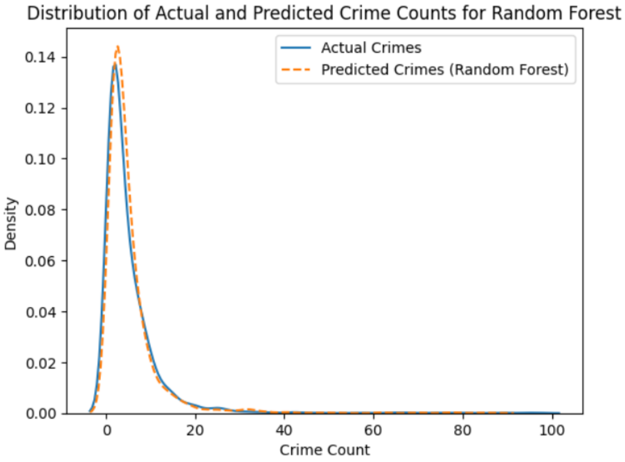


Figure 5: Distribution plot of RF

The distribution plot for the Random Forest model similarly indicates a close alignment between predicted and actual crime counts.

The peaks of the distributions almost coincide, which suggests that the Random Forest model is also effective in predicting the most common crime counts. The slight deviation in the tails of the distribution may indicate a difference in predicting less frequent higher crime counts, but overall, the model appears to perform well.

The distribution plot for the Gradient Boosting model shows that the predicted crimes (dashed line) closely follow the actual crime distribution (solid line). The peaks of both distributions align well, indicating that the Gradient Boosting model has a good predictive performance in terms of matching the overall crime count distribution. The density of the predicted crimes tapers off similarly to the actual crimes, suggesting that the model captures the decrease in frequency of higher crime counts effectively.

Residual Plots: By examining the residuals, I assessed where and how the models might be erring. These plots were instrumental in identifying any patterns or biases in the predictions, such as consistent errors for certain types of crimes or locations.

The residual plots for both the Gradient Boosting and Random Forest models illustrate the errors between the predicted and actual crime counts. For the Random Forest model, the residuals are tightly clustered around the zero line, especially for lower crime counts, suggesting that the model has consistent predictive accuracy for the majority of the data.

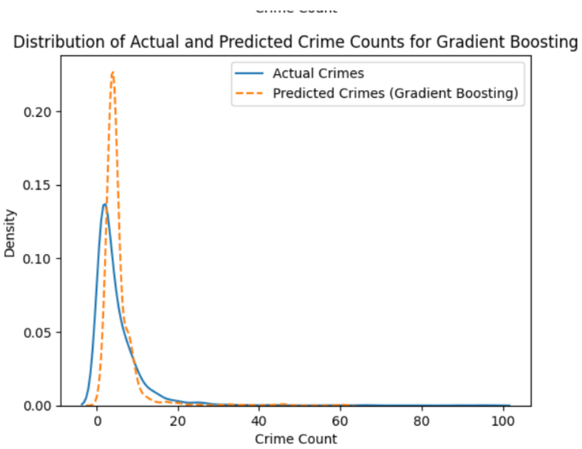


Figure 6: Distribution plot of Gradient Boosting

The Gradient Boosting model also shows a clustering of residuals around the zero line but with a slight tendency for larger errors as the crime count increases, indicated by the spread of residuals. This could suggest that the Gradient Boosting model may not perform as uniformly across the range of crime counts as the Random Forest model does.

3. Discussion of Findings:

- The analysis revealed key hotspots that consistently showed higher crime rates, indicating the models' utility for strategic planning in law enforcement.

- Temporal patterns, such as variations in crime rates on specific days or during certain periods, aligned with my initial hypotheses based on criminological studies.

- These results emphasize the potential of data-driven approaches in enhancing public safety and urban planning efforts.

5 Conclusion

Synopsis of Results:

This project's main goals were to assess the efficacy of two distinct predictive models—the Random Forest Regressor and the Gradient Boosting Regressor—and use machine learning techniques to identify crime hotspots in Portland, Oregon. A tight cluster of residuals in the Random Forest model indicated

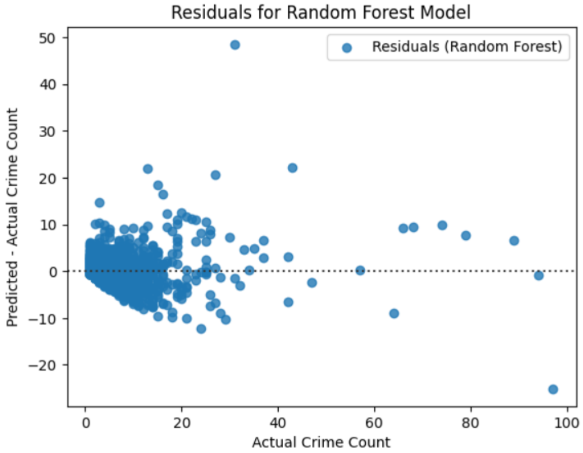


Figure 7: Residual plot for RF

consistent predictive performance over a range of crime counts. Even with its effectiveness, the Gradient Boosting model's forecasts varied a little more when the number of crimes increased. The Predictive Efficiency Index (PEI) for both models was high, indicating that they were good at predicting crime hotspots.

Limitations and Shortcomings:

The project had several issues despite the methods' strengths. Because the dataset was restricted to a three-month window, it might not have adequately captured the yearly and seasonal fluctuations in crime trends. Furthermore, the models did not take socioeconomic, demographic, or other potentially significant characteristics into account; instead, they just used historical crime data. Subjectivity was introduced into the analysis via the binary categorization for model comparison, which used a threshold to define 'high crime' locations and may have oversimplified the complex nature of crime occurrences.

Future Work and Project Extensions:

Building on this work, upcoming studies might incorporate a wider range of data, such as demographic data, socioeconomic indicators, and other environmental elements, to improve the predictive capacity of the models. Predictions could be improved by investigating more complex machine learning methods, such as deep learning models, which can capture intricate nonlinear linkages and interactions. A larger dataset covering more years would also make it possible to look at long-term trends and create models that take seasonal and annual fluctuations in crime patterns into consideration. Developing an interactive web tool that enables users to explore various scenarios and display forecasts could be another way to extend the research and help public safety officials with resource allocation and strategic planning. This tool could serve as a valuable asset for

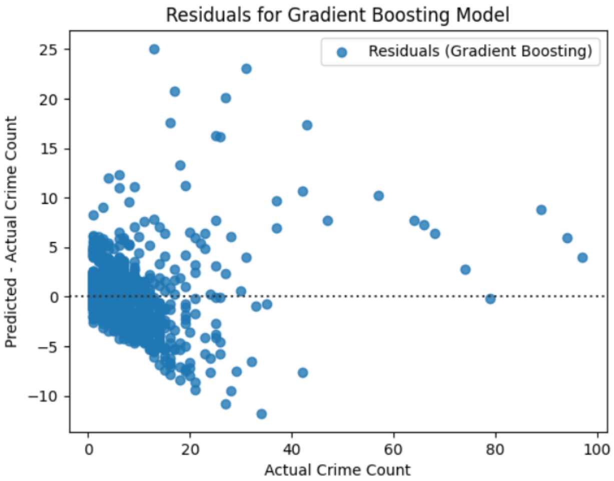


Figure 8: Residual plot for GB

community engagement, allowing residents to understand and contribute to the discussions on urban safety and crime prevention.

6 Data and Software Availability

Code available on:- Github repository - Crime Hotspot Analysis.
Dataset:- Call-for-service data

7 References

Mireia Rosell , Juan Fernández-Recio (2018) Hot-spot analysis for drug discovery targeting protein-protein interactions, Expert Opinion on Drug Discovery, 13:4, 327-338, DOI: 10.1080/17460441.2018.1430763

Jeong, K.-S., Moon, T.-H., , Jeong, J.-H. (2010). Hotspot Analysis of Urban Crime Using Space-Time Scan Statistics. Journal of the Korean Association of Geographic Information Studies, 13(3), 14–28. https://doi.org/10.11108/KAGIS.2010.13.3.014

Wang, D., Ding, W., Lo, H. et al. Crime hotspot mapping using the crime related factors—a spatial data mining approach. Appl Intell 39, 772–781 (2013). https://doi.org/10.1007/s10489-012-0400-x

Ibrahim, N., Wang, S., Zhao, B. (2019). Spatiotemporal Crime Hotspots Analysis and Crime Occurrence Prediction. In: Li, J., Wang, S., Qin, S., Li, X., Wang, S. (eds) Advanced Data Mining and Applications. ADMA 2019. Lecture Notes in Computer Science(), vol 11888. Springer, Cham. https://doi.org/10.1007/978-3-030-35231-842