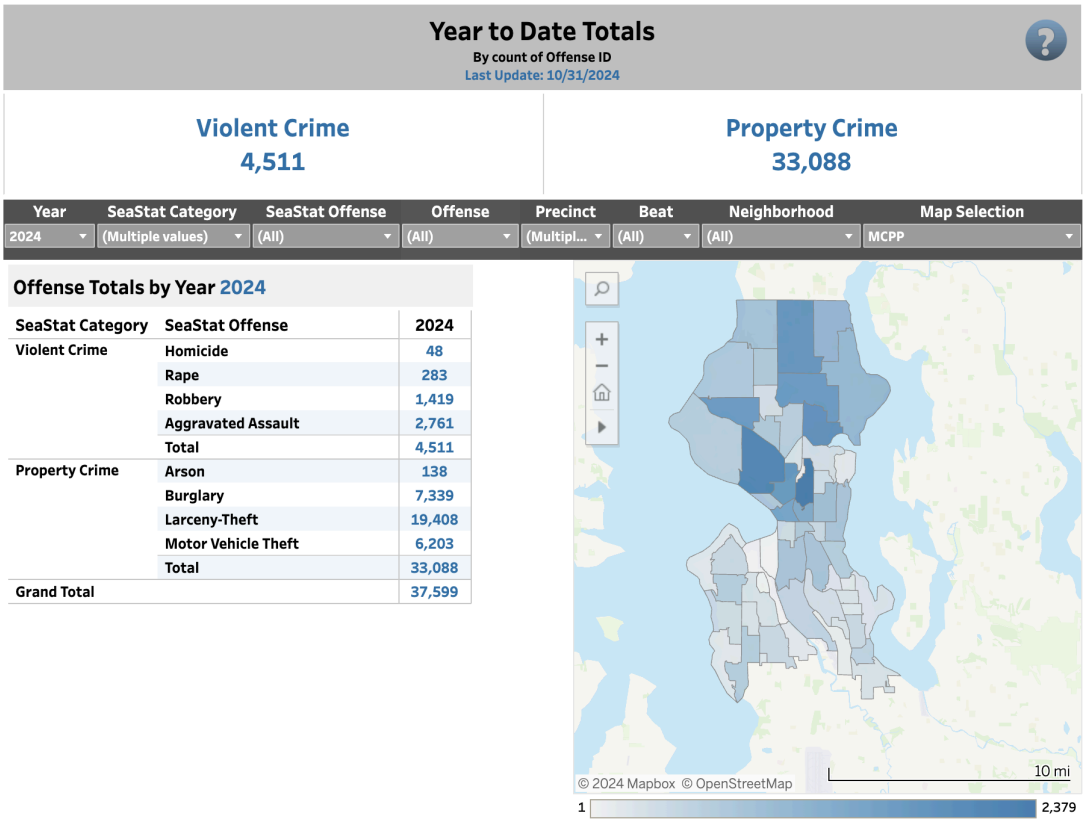
Seattle Crime Risk Prediction
CS7180 Final Report
Xianying Chen, Hongyu Liu
Dec 11, 2024

Introduction

Seattle is a thriving metropolitan center, yet like many urban environments, it faces persistent challenges in public safety. Crime—ranging from property theft and vandalism to violent offenses—creates fear, erodes community trust, and reduces overall quality of life. The city's diverse neighborhoods and dynamic conditions make it difficult to consistently understand, anticipate, and manage these issues.

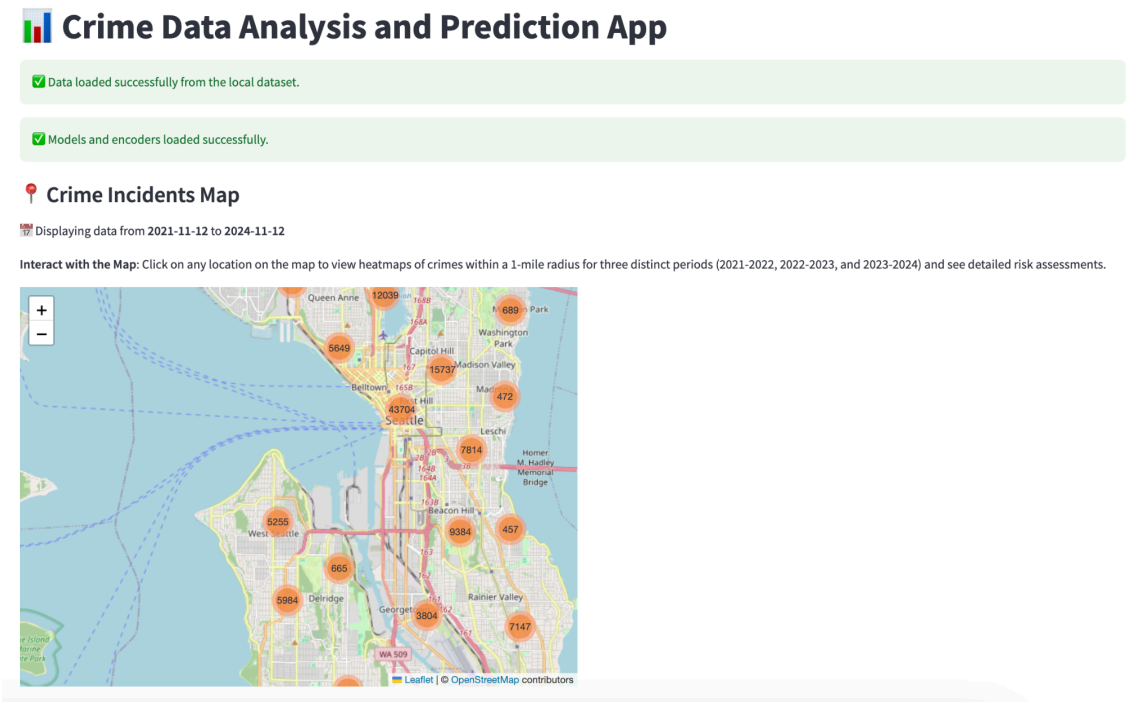


[Figure 1: Year-to-Date Crime Totals]

Although a wealth of crime data is available for Seattle, much of it remains underutilized due to its complexity and the lack of accessible, actionable analytical tools. Existing systems primarily focus on static summaries or retrospective reports, which often fail to capture the dynamic and multifaceted nature of crime. This makes it challenging for decision-makers—such as law enforcement and community leaders—to proactively identify emerging trends, assess risks, and allocate resources effectively.

Our Seattle Crime Risk Prediction System addresses these shortcomings by introducing a novel approach that integrates spatial and temporal data to forecast both crime risk levels and specific crime types. Unlike traditional systems, which typically provide generic crime trends or hotspots, our system predicts the likelihood of specific offenses occurring at particular locations and times. This dual-layer prediction capability enables more targeted interventions and resource allocation.

The predictions are presented through interactive maps and charts, allowing users to visualize high-risk areas, explore historical crime patterns, and derive actionable insights. By leveraging clustering analysis, machine learning model XGBoost, and techniques to address class imbalances, the system offers a robust, dynamic framework that bridges the gap between raw data and informed decision-making. This integrated approach provides a deeper understanding of crime patterns, empowering public safety officials and community stakeholders to implement proactive measures that enhance security across Seattle.



[Figure 2: Screenshot of the Seattle Crime Risk Prediction Application]

Dataset

The primary data source for this study is the Seattle Police Department Crime Dataset, accessible through the City of Seattle's open data portal. Spanning from 2008 to the present, it includes over 1.14 million incident records with 17 attributes, such as offense

start and end dates, location coordinates, and crime classifications. These temporal and spatial details enable nuanced analysis of crime patterns, essential for predicting both risk levels and specific offense types.

A key strength of the dataset is its adherence to National Incident-Based Reporting System (NIBRS) standards, ensuring consistent definitions and reliable comparisons across categories. Daily updates further enhance its utility for near-real-time predictive models, making it highly relevant for capturing current and emerging trends.

The dataset's combination of spatial detail, temporal granularity, and standardized classifications forms a robust foundation for advanced feature engineering and machine learning. By integrating these structured fields, the system moves beyond descriptive statistics to uncover patterns that inform actionable insights, enabling a comprehensive and context-aware crime risk prediction framework.

Column Name	Description	API Field Name	Data Type
Report Number	Primary key/UID for the overall report. One report can contain multiple offenses, as denoted by the Offense ID.	report_number	Text
Offense ID	Distinct identifier to denote when there are multiple offenses associated with a single report.	offense_id	Text
Offense Start DateTime	Start date and time the offense(s) occurred.	offense_start_datetime	Floating Timestamp
Offense End DateTime	End date and time the offense(s) occurred, when applicable.	offense_end_datetime	Floating Timestamp
Report DateTime	Date and time the offense(s) was reported. (Can differ from date of occurrence)	report_datetime	Floating Timestamp
Group A B	Corresponding offense group.	group_a_b	Text
Crime Against Category	Corresponding offense crime against category.	crime_against_category	Text
Offense Parent Group	Offense_Parent_Group	offense_parent_group	Text
Offense	Corresponding offense.	offense	Text
Offense Code	Corresponding offense code.	offense_code	Text
Precinct	Designated police precinct boundary where offense(s) occurred.	precinct	Text
Sector	Designated police sector boundary where offense(s) occurred.	sector	Text
Beat	Designated police sector boundary where offense(s) occurred.	beat	Text
MCPP	Designated Micro-Community Policing Plans (MCPP) boundary where offense(s) occurred.	mcpp	Text
100 Block Address	Offense(s) address location blurred to the one hundred block.	_100_block_address	Text
Longitude	Offense(s) spatial coordinate blurred to the one hundred block.	longitude	Number
Latitude	Offense(s) spatial coordinate blurred to the one hundred block.	latitude	Number

[Figure 3: Dataset Columns and Attributes]

Data Cleaning and Preprocessing

To ensure relevance and model accuracy, we focused on the most recent subset of the dataset. Specifically, we included incidents from the past three years, reflecting current crime patterns and avoiding outdated trends that may no longer be informative. This temporal constraint reduces noise and enhances the generalizability of subsequent predictions.

We also filtered the dataset to emphasize attributes essential for our predictive modeling. Key columns, such as offense_start_datetime, latitude, longitude, precinct, offense_parent_group, and crime_against_category, were retained for their direct influence on spatial-temporal modeling and risk categorization. Less informative or redundant variables were excluded to streamline data processing and reduce model complexity. Similarly, offenses with very few occurrences—outliers that could skew model training—were removed. By refining the dataset to high-impact features and recent records, we established a focused, reliable foundation for predictive analysis.

Handling Missing and Invalid Values

Prior to modeling, we addressed missing or invalid entries within the selected columns. Records lacking essential spatial or temporal information were discarded, as their absence could compromise both the interpretability and accuracy of downstream results. By ensuring that each record included complete coordinates and timestamps, we maintained the integrity of our spatial-temporal inferences.

Feature Engineering

To capture patterns across different scales of time and location, we derived new features from the existing temporal fields. Variables such as month, day_of_week, and hour were extracted from the offense_start_datetime to highlight recurring cycles and seasonality in criminal activity. Additionally, crimes were grouped into “High” or “Low” risk categories based on the offense_parent_group. For example, violent offenses like assault and robbery were designated as high-risk, whereas non-violent or financial crimes were classified as low-risk. These engineered features offer richer input signals that help models discern temporal and categorical patterns.

Categorical Encoding

To ensure model compatibility with categorical data, we employed a combination of one-hot and label encoding. Temporal features such as month, day_of_week, hour were one-hot encoded to preserve their categorical nature without imposing ordinal relationships. In contrast, categories like offense_parent_group and risk_level were label-encoded, converting distinct classes into integer values. This approach reduced

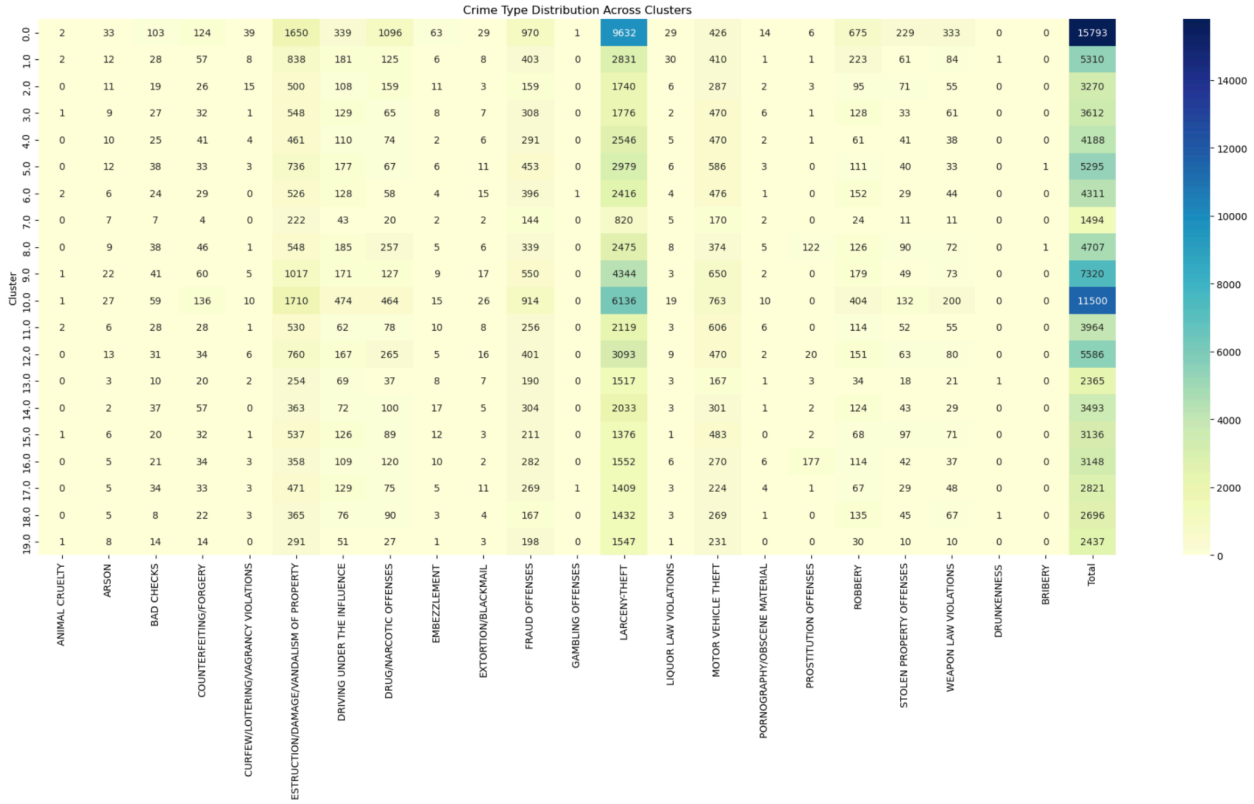
dimensionality and allowed the model to efficiently process essential categorical information.

Imbalance Handling

A prominent challenge was the underrepresentation of high-risk crimes. To mitigate this, we applied the Synthetic Minority Oversampling Technique (SMOTE), generating synthetic examples of minority-class instances. This approach balanced the class distribution, improving the model's ability to identify high-risk events. We acknowledge that synthetic data can introduce biases and may not fully replicate real-world conditions; however, preliminary experimentation suggested that SMOTE improved recall for high-risk predictions, ultimately enhancing the model's operational utility.

Methodology

As part of the exploratory phase, we performed an unsupervised clustering analysis on the dataset to identify natural groupings of crime incidents and uncover hidden patterns. The heatmap visualization shows the distribution of various crime types across these clusters, with color intensity reflecting the frequency of each crime category within each group. This preliminary insight guided subsequent steps in the analysis: it informed feature selection, highlighted dominant crime types associated with particular clusters, and helped us tailor our predictive modeling approach to better capture the nuanced relationships between spatial, temporal, and categorical factors influencing crime in Seattle.



[Figure 4: heatmap visualization]

Model Selection

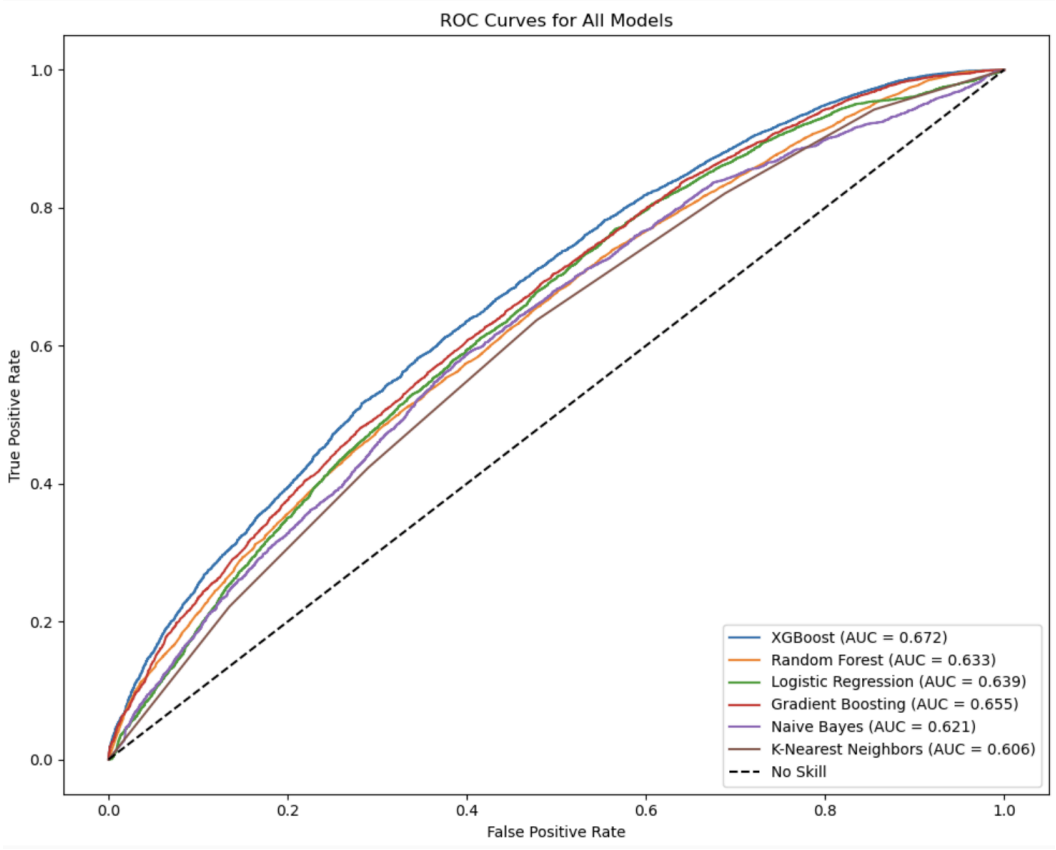
We evaluated six machine learning models to determine their effectiveness in predicting crime risk within urban environments: XGBoost, Gradient Boosting, Random Forest, Logistic Regression, Naive Bayes, and K-Nearest Neighbors (KNN). This selection aimed to capture a broad spectrum of methodological approaches—ranging from linear classifiers to ensemble-based and tree-structured models—thereby accommodating nonlinear feature interactions and varying sensitivities to imbalanced class distributions. By considering diverse algorithms, we aimed to identify a model that most accurately differentiates between “High” and “Low” crime-risk areas.

Training and Validation Procedures

The dataset was partitioned into training and testing subsets. The training set was used to fit each model and tune its hyperparameters, while the testing set was held out to provide an unbiased evaluation of the models' predictive capabilities. We employed cross-validation to mitigate overfitting and ensure that reported performance metrics were representative of generalization to unseen data.

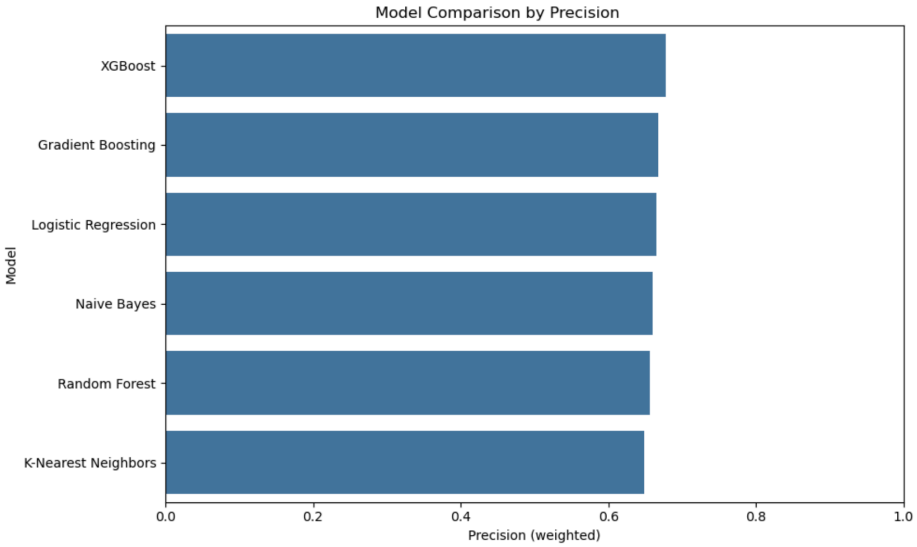
Performance Metrics

Because the dataset contained disproportionately fewer high-risk incidents than low-risk ones, we chose evaluation metrics sensitive to class imbalance. While accuracy provided a baseline measure, it tended to overemphasize the performance of the majority class. To gain a more nuanced understanding, we incorporated recall, precision, and F1-score to assess each model's handling of minority-class predictions. Additionally, we prioritized the AUC-ROC score, which reflects a model's ability to discriminate between classes across varying decision thresholds, making it particularly informative in imbalanced classification tasks. Charts and tables summarizing these metrics allowed for a clear and interpretable comparison.

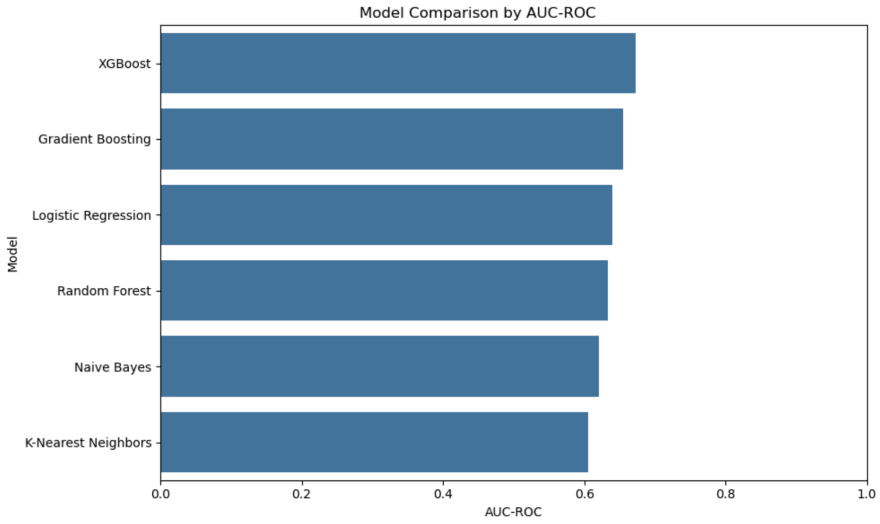


[Figure 5: AUC-ROC Curves for All Models Here]

In Figure 5, we present the AUC-ROC curves for each model, illustrating their relative abilities to distinguish between high and low-risk incidents across various thresholds.



[Figure 6: Precision for All Models Here]



[Figure 7: AUC-ROC Curves for All Models Here]

Figure 6 and Figure 7 visualize the comparative performance of all models based on AUC-ROC and weighted precision, respectively.

XGBoost consistently demonstrated superior performance, achieving an AUC-ROC of 0.762 and a weighted precision of 0.677. Compared to Gradient Boosting and Random Forest, which achieved marginally lower AUC-ROC values, and to Logistic Regression, Naive Bayes, and KNN, which performed less reliably, XGBoost offered a stronger balance between correctly identifying high-risk areas and maintaining accuracy on the majority class. Although it occasionally misclassified some high-risk instances, it surpassed competing models in capturing complex feature interactions and delivering robust discriminatory power.

	Predicted High	Predicted Low
Actual High	3193	2654
Actual Low	4463	9690

Table 1: Confusion Matrix for XGBoost

To better understand the chosen model's strengths and limitations, we examined confusion matrices and classification reports. The confusion matrix for XGBoost (Table 1) reveals important insights into its classification performance. While the model correctly predicted a significant majority of "Low" crime-risk areas, it occasionally misclassified "High" risk areas as "Low." This is a trade-off we made in imbalanced datasets, where the model's emphasis on overall performance can result in reduced sensitivity for minority classes. Despite this, the recall for the "High" class was **0.55**, ensuring that over half of the high-risk areas were accurately identified. Weighted precision and recall values of **0.68** and **0.64**, respectively, underscore the model's reliability in balancing performance across both classes.

```
Evaluating XGBoost...
Accuracy: 0.64415
Precision (weighted): 0.6774300141121701
Classification Report:
              precision    recall  f1-score   support

        High       0.42      0.55      0.47      5847
         Low       0.78      0.68      0.73     14153

    accuracy                           0.64     20000
   macro avg       0.60      0.62      0.60     20000
weighted avg       0.68      0.64      0.66     20000
```

[Figure 8: Classification Report for XGBoost]

The classification report (Figure 8) further highlights these findings, with F1-scores of **0.47** for the "High" class and **0.73** for the "Low" class. These metrics demonstrate XGBoost's effectiveness in identifying "Low" crime-risk areas with high precision and recall, which is critical for resource allocation in crime prevention strategies. The model's overall weighted F1-score of **0.66** indicates robust predictive performance across the dataset.

The evaluation demonstrates that XGBoost's ability to handle imbalanced datasets and capture complex feature interactions makes it the most reliable choice for crime prediction. Its higher AUC-ROC and weighted precision scores indicate strong discriminatory performance, while its recall for high-risk areas ensures that critical zones are effectively identified. However, the occasional misclassification of "High" risk areas underscores the need for future work in addressing class imbalance.

Overall, XGBoost provides a solid foundation for building a scalable and accurate crime-risk prediction system. By prioritizing AUC-ROC and recall metrics, the model aligns well with the operational requirements of public safety applications, ensuring a balance between predictive accuracy and actionable insights.

Results

To assess the practical utility of the crime prediction system, we evaluated the XGBoost model under two spatial frameworks: a precinct-based approach and a 3-mile radius approach. Both were tested using accuracy, precision, recall, F1-score, and AUC-ROC. These metrics offered a multifaceted view of the model's ability to discriminate between "High" and "Low" crime-risk areas, as well as maintain robust performance across class imbalances.

Precinct-Based Performance

The precinct-based model configuration delivered outstanding performance. It achieved a 96.77% accuracy and a near-perfect AUC-ROC of 0.9942, underscoring its exceptional discriminatory power. Crucially, it demonstrated consistent strengths across both classes:

- **High-Risk Areas:** Precision reached 0.91, recall was 0.97, and the F1-score stood at 0.94, indicating that the model accurately identified and flagged high-risk zones with minimal false alarms.
- **Low-Risk Areas:** Precision obtained 0.99, recall was 0.97, and the F1-score was 0.98, confirming the model's reliability in maintaining overall accuracy and minimizing misclassifications for the majority class.

The confusion matrix for this approach showed minimal classification errors, and the balance between high and low-risk class metrics suggests that the precinct-based approach aligns effectively with administrative boundaries and structured data regions. This combination of precision and stability makes the precinct-based model a strong candidate for operational use, guiding resource allocation and proactive interventions.

```
Accuracy: 0.967741935483871
Classification Report:
              precision    recall  f1-score   support

        High       0.91      0.97      0.94        86
         Low       0.99      0.97      0.98       255

    accuracy                           0.97       341
   macro avg       0.95      0.97      0.96       341
weighted avg       0.97      0.97      0.97       341

Confusion Matrix:
 [[ 83   3]
 [  8 247]]
AUC-ROC: 0.9942088463292293
```

[Figure 9: Classification Report for Precinct-Based Performance]

3-Mile Radius Approach

While the 3-mile radius model remained effective, it displayed certain limitations compared to the precinct-based configuration. It achieved 84% accuracy and an AUC-ROC of 0.8721. Although these results are still respectable, the trade-offs become apparent when examining class-specific metrics:

- **High-Risk Areas:** The model demonstrated strong sensitivity (recall = 1.00) but moderate precision (0.62), yielding an F1-score of 0.76. Although it successfully identified all high-risk instances (no false negatives), the reduced precision implies a higher rate of false alarms.
- **Low-Risk Areas:** Precision for low-risk areas was perfect (1.00), but the recall decreased to 0.78, and the F1-score declined to 0.88. The reduced recall indicates that some low-risk areas were misclassified, potentially due to the spatial overlap and less-defined boundaries introduced by the 3-mile radius approach.

The increased complexity in spatial delineation likely contributed to these challenges. The 3-mile radius method imposes arbitrary and overlapping boundaries, which may dilute the distinctiveness of crime patterns and introduce additional noise.

```
Accuracy: 0.84
Classification Report:
              precision    recall  f1-score   support

        High       0.62      1.00      0.76        13
         Low       1.00      0.78      0.88        37

    accuracy                           0.84        50
   macro avg       0.81      0.89      0.82        50
weighted avg       0.90      0.84      0.85        50

Confusion Matrix:
 [[13  0]
 [ 8 29]]
AUC-ROC: 0.8721413721413721
```

[Figure 10: Classification Report for 3 mile-Based Performance]

The precinct-based approach consistently outperformed the 3-mile radius method across all metrics. Its ability to balance precision and recall, combined with its near-perfect AUC-ROC score, makes it the better choice for predicting crime risk. One of the reasons for its success is the alignment of precincts with administrative boundaries, which provides more structured and less noisy data. On the other hand, the 3-mile radius approach introduces spatial overlap, which can dilute the model's precision and lead to more errors.

The classification reports and confusion matrices for both approaches highlight these differences clearly. The precinct-based model's high precision and recall for high-risk areas are particularly important for resource allocation and decision-making in public safety strategies.

Overall, the precinct-based method stands out as the most reliable approach for crime prediction, offering accurate and actionable insights that can be used to enhance public safety efforts. While the 3-mile radius method is still useful, especially for broader analysis, its limitations suggest that the precinct-based model is better suited for practical applications. With real-time data integration and richer contextual features, the system has the potential to become even more robust and effective.

Application

Interactive Features

The Streamlit application provides an intuitive and interactive platform for exploring crime data in Seattle. Users can filter crime incidents by date ranges and specific crime types through an easy-to-use sidebar interface. The app displays an interactive crime map that allows users to click on any location to view detailed crime heatmaps within a 1-mile radius across three time periods (2021-2022, 2022-2023, and 2023-2024). Each heatmap highlights crime density and provides an assessment of risk levels based on historical data. Additional features include charts that compare the top crime types over time and a line chart visualizing monthly trends, giving users a comprehensive view of crime patterns.

Implementation

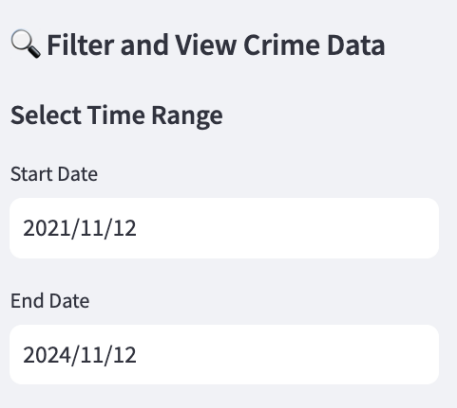
The application is built using a robust tech stack to ensure functionality and scalability:

- Streamlit: For building the web-based user interface and interactive components.
- Folium & Streamlit-Folium: To create interactive maps and heatmaps for crime visualization.
- Scikit-learn: For preprocessing data and encoding categorical variables.
- XGBoost Models: Pre-trained machine learning models are imported to predict crime risks and types, ensuring efficient and accurate predictions without in-app training.
- Pandas & NumPy: For data manipulation, cleaning, and feature engineering.

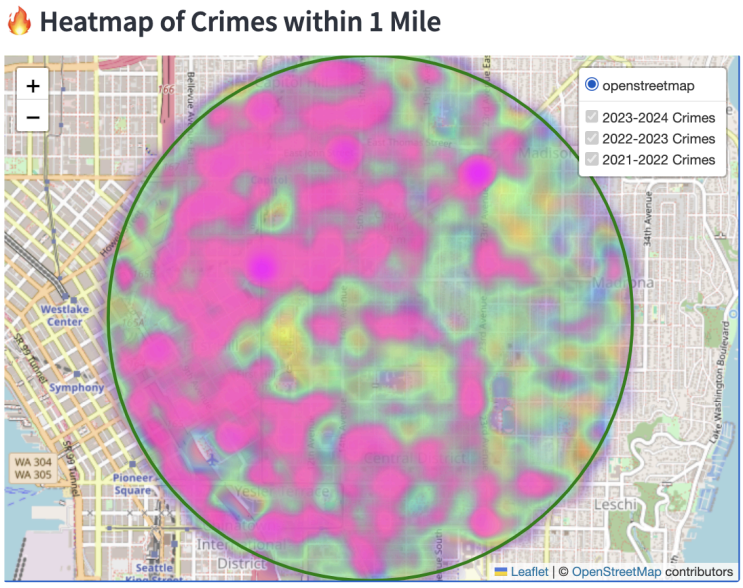
- Plotly Express: To generate dynamic visualizations for crime trends and comparisons.
- Joblib: For loading serialized pre-trained models quickly and efficiently.

User Interaction
The app is designed to be highly customizable, empowering users to explore crime data in ways that suit their needs. Users can:

- Filter data by specific date ranges and crime types via the sidebar.



- Interact with the map to explore high-risk areas and view historical crime density around specific locations.



- Customize their exploration further by focusing on specific types of crimes or areas of interest.
  Each visualization is accompanied by clear and informative explanations to help users interpret the data. For example, the map provides a detailed breakdown of

the risk levels, total incidents, and crime types within a selected radius. Charts and graphs make it easy to spot trends and compare crime patterns over time, enabling users to make informed decisions about safety and resource allocation.

The application bridges the gap between raw crime data and actionable insights, making it a valuable tool for policymakers, law enforcement, and residents alike.

Challenges and Future Work

Challenges

Throughout the development of the Seattle Crime Risk Prediction System, several technical hurdles emerged. One of the most significant challenges was efficiently handling the large dataset, which included over 1.14 million rows. Loading and processing such a large amount of data often led to slower performance and longer computation times. To address this, we implemented Streamlit's caching mechanisms, which significantly improved app responsiveness by reducing redundant computations.

Another challenge was the imbalance in the dataset, with high-risk crimes being underrepresented compared to low-risk crimes. Although we used techniques such as SMOTE to balance the dataset, synthetic data introduced potential biases. Synthetic samples may not fully capture the complexity of real-world data, which could impact the accuracy and generalizability of our predictions. Balancing the need for accurate representation with the risk of overfitting required careful consideration.

Lastly, ensuring model interpretability while maintaining accuracy proved to be a challenge. While XGBoost delivered strong performance, their complexity sometimes made it harder to explain predictions in a way that end-users could easily understand. Simplifying outputs for non-technical users without losing critical insights remains an area for improvement.

Future Enhancements

Looking ahead, there are several opportunities to enhance the system further:

1. Incorporate Additional Data Sources:
   To improve the predictive accuracy and contextual awareness of the system, we plan to integrate supplementary data sources such as social media feeds, real-time weather conditions, and emergency report logs. These data points could provide a richer context for crime predictions, allowing the system to capture external factors that influence criminal activity.
2. Develop Real-Time Prediction Capabilities:
   A major limitation of the current system is the reliance on static historical data.

Implementing real-time data streaming and processing could enable the system to provide live updates on crime trends and risks. This would enhance its utility for immediate decision-making and public safety efforts.

Conclusion

The Seattle Crime Risk Prediction System showcases an important step forward in using data and machine learning to improve public safety in urban areas. By combining spatial and temporal data with XGBoost, the system offers meaningful insights into crime patterns and risk levels across Seattle. With interactive heatmaps and visual charts, it makes complex data easy to understand, empowering city officials, law enforcement, and residents to make smarter decisions. The precinct-based approach proved to be the most effective, achieving high accuracy and precision by leveraging structured administrative boundaries. Although challenges like imbalanced datasets and model complexity posed difficulties, SMOTE helped address these issues, improving the system's ability to identify high-risk areas.

What sets this system apart is its unique approach to predicting both crime risk levels and specific crime types while factoring in spatial and temporal patterns. This dual focus required careful model training and evaluation, with the precinct-based method ultimately delivering the best results. Looking ahead, there are exciting opportunities to enhance the system by integrating real-time data, such as social media and weather updates, for more dynamic and context-aware predictions. Additionally, expanding its accessibility through mobile apps or customizable dashboards could make it even more useful for a wider audience. Overall, the Seattle Crime Risk Prediction System provides a strong foundation for improving public safety and has the potential to become a transformative tool in crime prevention and urban safety.

References

Figure 1: https://seattle.gov/police/information-and-data/data/crime-dashboard
Dataset: Seattle Police Department. (n.d.). *SPD Crime Data: 2008-Present*. Seattle Open Data Portal. Retrieved from https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5/about_data