BIRMINGHAM CITY
University

Predicting Crime Case Resolutions Using Machine Learning

CMP5367 Coursework

Priyanka Khatri

Student Id: 25123827

Bachelors of Computer Science with AI
(Faculty of Computing, Engineering, and the Built Environment)
Birmingham City University (BCU)
Sunway College
Kathmandu, Nepal
priyanka.khatri@mail.bcu.ac.uk

Word Count: 4118

Abstract

Efficient crime cases resolutions are a key indicator of the effectiveness of law enforcement, with ongoing investigations placing sustained pressure on the limited operational resources and, in the long run, may erode public trust in the justice system. Thus,understanding the cases with a longer investigation timeline is important not only to operational planning but also to the support of transparent and accountable practices of policing. This paper will explore the idea of whether machine learning models can be used to forecast the outcomes of crime cases resolution using historical crime data of the Los Angeles Police Department between 2020 and 2024.The problem is stated as a supervised binary classification problem where the data instances are crime reports, and the target variable is binary: resolved or unresolved.

An organized data preparation pipeline was utilized, such as data cleaning, feature engineering, temporal decomposition, and spatial aggregation to handle the dimension of scale, heterogeneity, and imbalance of the real world policing data. Light Gradient Boosting Machine (LightGBM) was used as the major model because it is effective on large-scale tabular data and has a substantial capacity to handle class imbalance. ROC-AUC, precision, recall and analysis of the confusion matrix were used to evaluate model performance.

The proposed model obtained a ROC-AUC of 0.85 which is strong, demonstrating noisy & unbalanced administration crime data. The interpretability analysis of SHAP demonstrates that the crime-context and operational features are the key factors influencing prediction with crime category, the complexity of the modus operandi, the type of premises, the grouping of weapons, and the jurisdiction area have turned out to be the most effective features. Hour, month, and day of the week are examples of temporal variables which have relatively smaller influence in the prediction. These results indicate that decision support systems built on machine learning can help law enforcement agencies to focus investigative effort on the cases that are more likely to go unnoticed, while also recognizing the practical and ethical limitations of predictive modeling in high stakes policing.

Keywords: Area under the curve (AUC), exploratory data analysis(EDA), gradient boosting machine (GBM), Los Angeles Police Department(LAPD), machine learning(ML), receiver operating characteristic (ROC), Shapley additive explanations (SHAP)

Table of Contents

List of Figures

List of Tables

1. Introduction

One of the foundational indicators of the effectiveness of law enforcement agencies is crime resolution efficiency after crimes are reported. Unresolved or protracted investigations may exert a persistent strain on the efforts of the police, make it more challenging to plan operations and influence public confidence in the justice system. Prior studies demonstrate that the results of investigations are influenced by various factors, such as the nature of crime, workload, and the organizational practices in the police departments (Lee, 2020). Further, crime processes tend to be spatially and temporally structured, and not random, expressing trends in the distribution of incidences and resource allocation in the urban landscapes (Mohler et al., 2011).

1.1 Motivation and Research Gap

The majority of the current machine learning studies in crime analytics have been dedicated towards predicting where crimes may occur, risk assessment, or estimating recidivism, and less emphasis was given on what happens once a crime is reported (Berk, 2021). Though administrative crime datasets contain a large amount of information about investigation operational status and timeframes, these data have been rarely utilized in a systematic study of crime outcomes in cases (Burrows and Tarling, 1987). Consequently, little empirical knowledge exists regarding the factors related to long or unsolved cases and how such knowledge can be used to make better cases prioritization and allocate resources.

1.2 Problem Statement

In most of the City police departments, a significant portion of the reported crime cases go unsolved over a long period of time, which indicates a lack of capacity to investigate cases, prioritize, and utilize limited resources (Governing, 2024; Johnson et al., 2023). Though giant police data sets provide a lot of contextual and procedural data, such records are rarely reviewed to uncover the consistent trends associated with the delayed or unsolved case resolutions. This leads to the situation where the allocation of investigative resources are often allocated reactively and are not based on systematic and evidence-based indicators of the risk of resolution.

1.3 Dataset Introduction

The dataset used in this study is taken from Los Angeles Open Data Portal (https://data.lacity.org/), which is officially maintained by the LAPD. The dataset is called Crime Data 2020-Present and comprises crime reports in Los Angeles in the years 2020 to 2025. The dataset has 1,004,991 observations and 28 attributes after the initial data selection and inspection. The data set contains data pertaining to the time of the incidence, the geographical location, the type of crime, the demographics of the victim, and the status of investigation. Since the data is

based on police records of operations, it might have missing values and reporting discrepancies as well as noises as are common in large administrative datasets.

1.4 Machine Learning Type Identification

The issue is presented as a supervised binary classification problem whereby the target variable would be whether a crime investigation remains open (delayed) or has been resolved.

2. Exploratory Data Analysis(EDA)

2.1 Dataset Overview: Grouping Features and Data Integrity

The dataset contains 1,004,991 crime incidents and 28 attributes that are crime reports submitted to the LA Police Department in 2020-2025. To analyze the structure, the variables are categorized into temporal, spatial, crime-context, victim-related, and administrative outcome features, which demonstrates the way the information is reflected in the real police reporting procedures and helps us to receive a complete picture of trends with respect to time, place, types of crimes, and victims demographics.

The distribution of missing values is highly skewed to more peripheral and context-dependent elements (including secondary crime codes, cross-street information and weapon description) where the absence is more of a conditional reporting problem than a random quality issue.

Feature Group	No. of Features	Average Missing(%)	Data Type(s)
Administrative	1	0.00	Integer
Temporal	3	0.00	Date, Integer
Spatial	7	12.09	Integer, Float, Object
Crime_Context	12	36.91	Integer, Float, Object
Victim_Demographics	3	9.59	Integer, Object
Outcome	2	0.00	Object

Table 1: Missingness among Feature Groups and Data Characteristics

2.2 Temporal Characteristics of Crime Reporting

Temporal features are the date of occurrence, date of reporting and time of the day. Analysis of data represents, the crime volume remains fairly constant between 2020 and 2023, then decreases in 2024, and reaches the highest level of sparsity in 2025 representing incompleteness. The 2025 data has less than 100 crime records hence excluded from the further analysis.

The crime intensity within a day varies significantly by hour, with repeated peaks in the middle of the day and at night, but crime volume does not differ across weekdays and years.

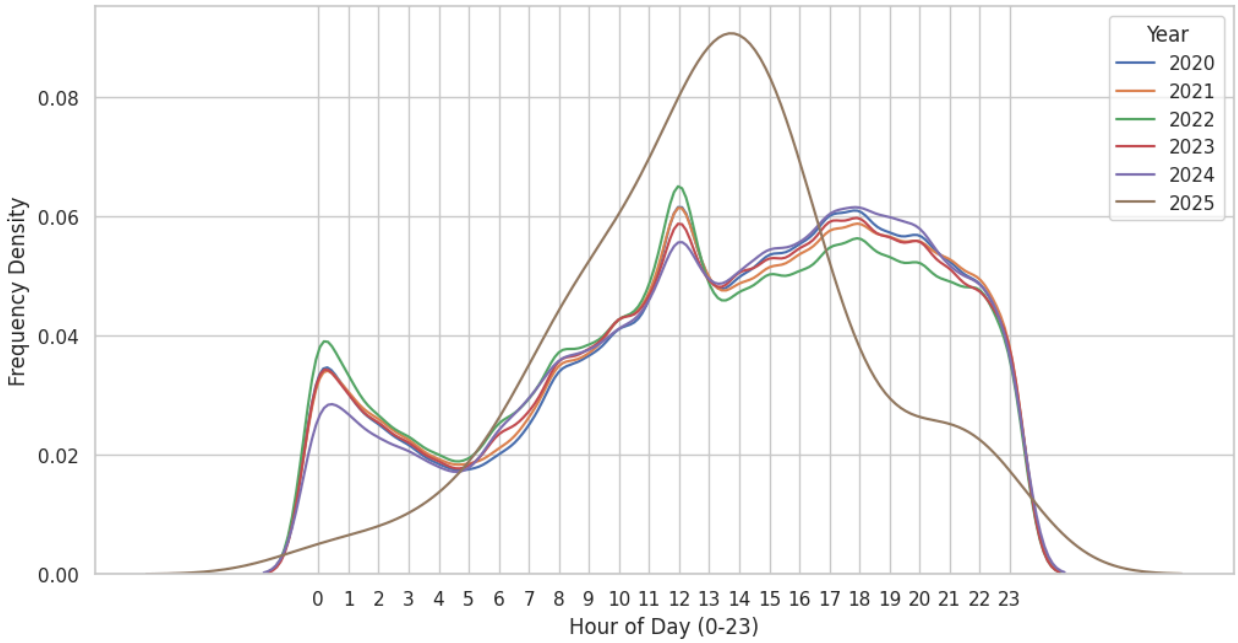

Figure 1:Crime Probability by Hour of Day:Year-over-Year Comparison

2.3 Location Patterns and Spatial Distribution

The spatial features covered are police area, reporting district, latitude and longitude. There is also imbalance in distribution of crime reports in areas, as few districts have shown a high percentage of incidents. This concentration is based on the population density and commercial activity and not reporting bias.

Whereas address level fields such as cross streets have a large number of missing values, which means optional reporting and not location uncertainty, geographic coordinates are mostly complete, allowing for trustworthy spatial analysis.

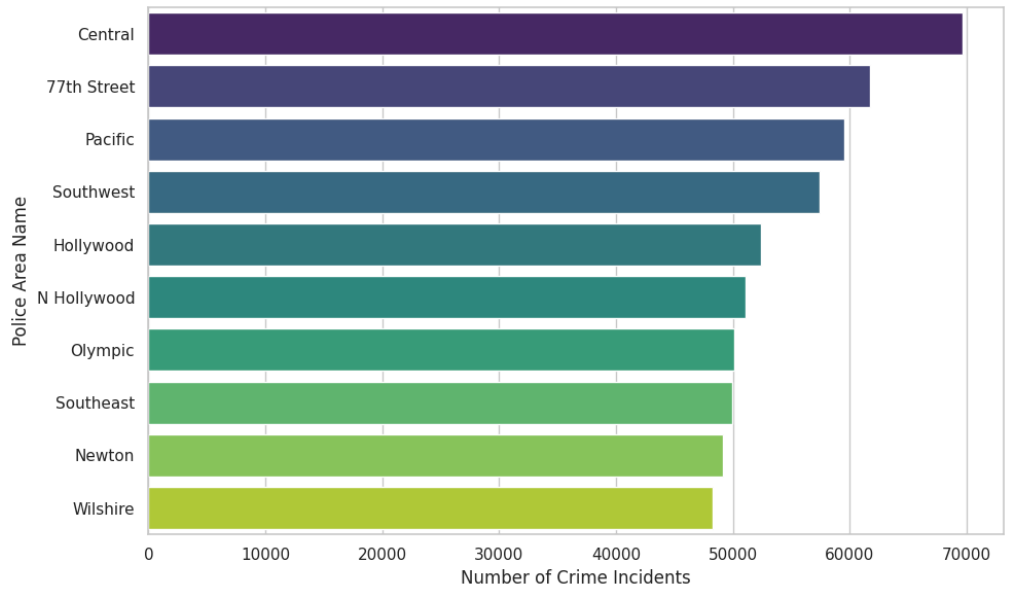

Figure 2: Crime Volume by Area

2.4 Crime Type and Premise Characteristics

The Crime context attributes show strong structural regularities. The dataset is dominated by a few types of offenses, specifically stealing related crimes, assault crimes, breakages, and vehicle related crimes. This level of concentration implies that there is predictability in the patterns of crime and not the wide-ranging patterns across categories.

Likewise, streets, residential houses, and car parks and commercial areas are premises that have been the focal point in most cases, which underscores the significance of public and semi-public places in urban crime processes.

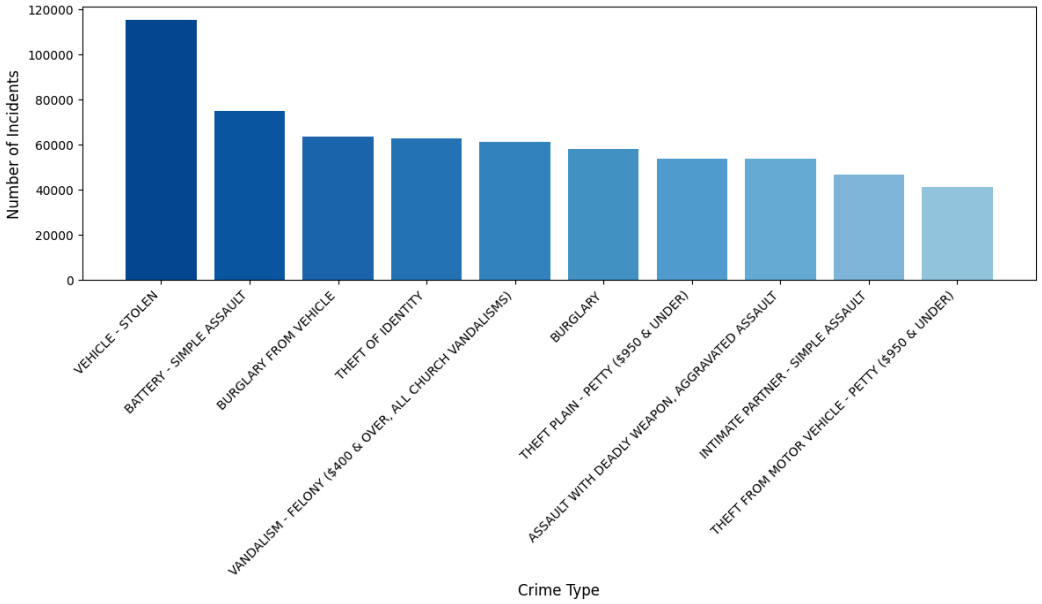

Figure 3: Top 10 most frequent crimes types

2.5 Weapon Usage Patterns

The weapon-related variables have high levels of missingness as it is expected since not all crimes involve a weapon. In the cases reported, bodily force, unknown weapon, and firearms are the most recorded.

The fact that the weapon labels contain the term unknown indicates a lack of information or progress of research, rather than data misinterpretation, which further supports that weapon characteristics should be interpreted carefully.

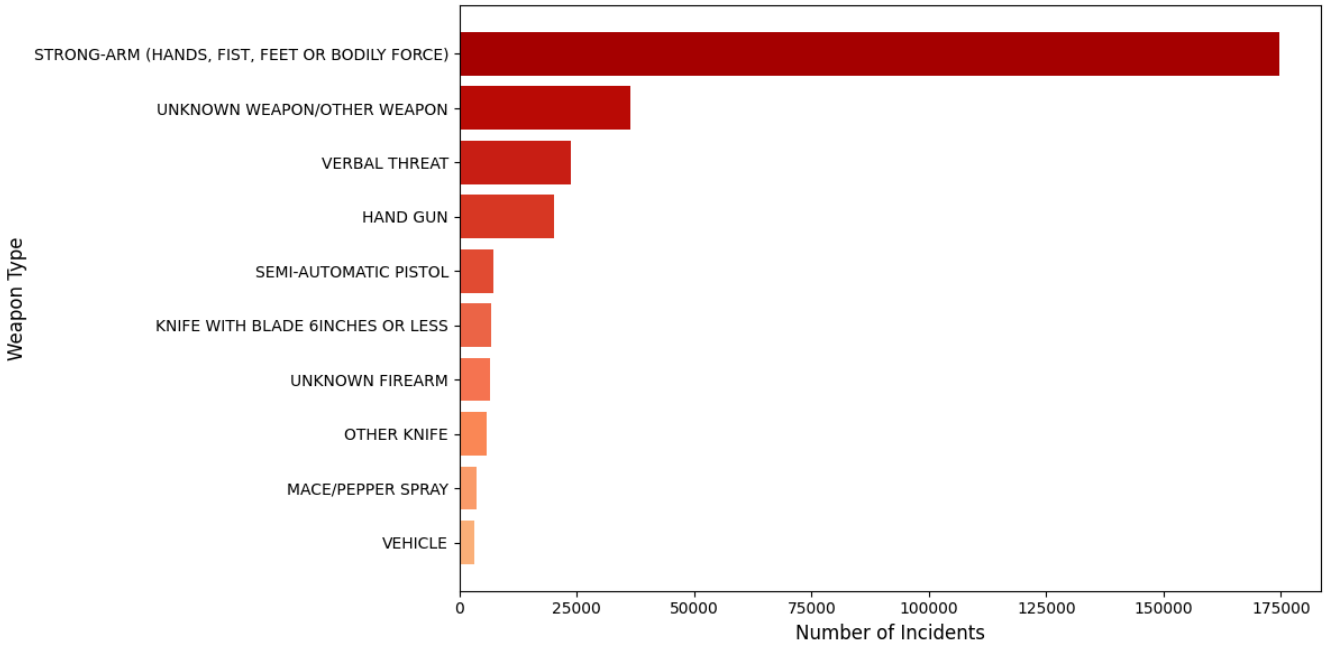

Figure 4: Top 10 Weapons Used

2.6 Victim Demographics and Seriousness of Crimes

There are significant demographic patterns in features related to the victim. Male victim crimes are marginally more than female victim crimes, and a non-trivial percentage of the records have no or unspecified sex information. Victim related missing values are assumed to exist due to non-discloser policies rather than data quality issues. Hispanic, White, and Black categories are dominant on the victim descent distributions, which is generally the population demographics of the reporting population.

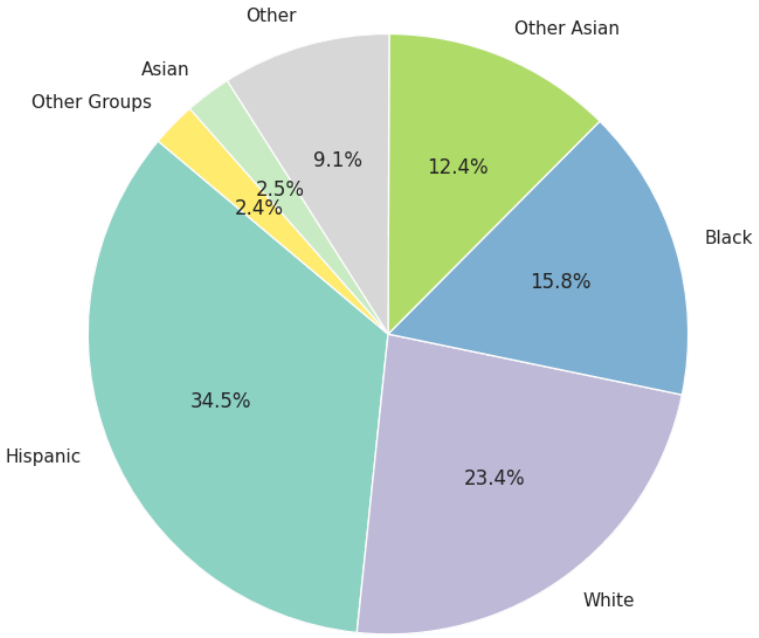

Figure 5: Victim Descent Distributions(Top Groups)

The Part I and Part II types of crime severity depict that serious crimes constitute the majority of those reported, however, non-serious crimes also still constitute a significant proportion of reported crimes. Figure 6 indicates that investigations that have existed over an extended period are more widespread with Part II crimes.

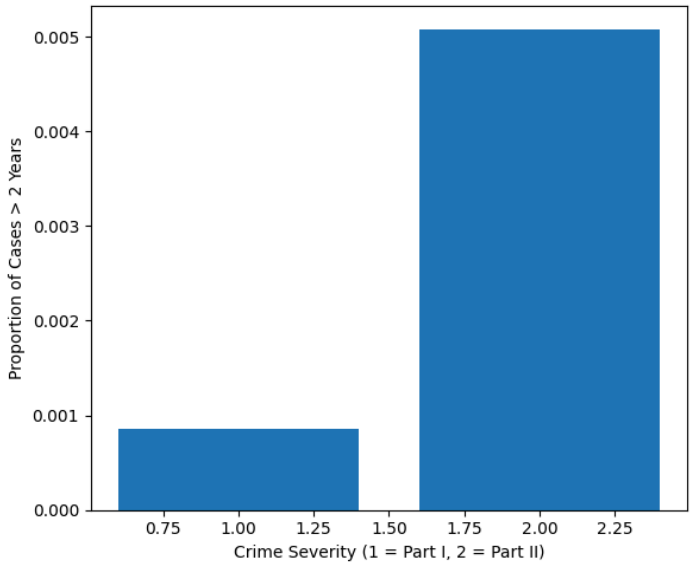

Figure 6: Proportion of Long-Running Investigations by Crime Severity

2.7 Case Resolution Status and Outcome Imbalance

The outcome variable has a high level of class imbalance with most of the cases under investigation continued. This unbalance is a practical representation of the policing procedures in which cases could wait long before being closed due to lack of evidence, reporting, or poor resources.

The outcome labels are hence uncertain in terms of time, which must be considered when using them for predictive modeling.

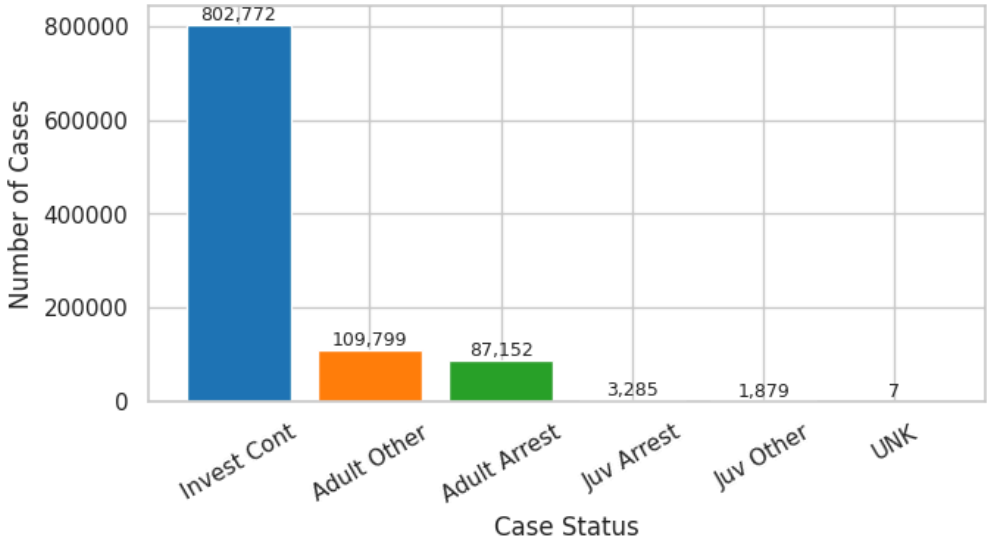

Figure 7: Case Status Distribution

3. Experimental Design

In this section, the modeling approach adopted for the prediction of case outcomes is outlined. It briefly describes the main assumptions on which the experimental arrangement is based and justify the choice of the baseline and advanced machine learning models that will be used in the current study.

3.1 Modeling Assumptions

The problem comes out as a supervised binary classification problem, in which the task is to predict whether a reported crime case remains in an investigation or solved in a small time frame. Only the information available at the time of reporting and very few post incidents information are used in order to avoid leakage in outcomes. Status of investigation registered in the LAPD dataset is considered an operational proxy of investigative delay because administrative and

procedural considerations might affect the timelines of closures. Since the outcome variable is highly imbalanced, the evaluation of the models will focus on discrimination-based measures, including ROC-AUC , Precision, Recall & confusion matrix instead of the overall accuracy, which can be deceptive in disproportionately prepared administrative data (He and Garcia, 2009; Thabtah et al., 2019).

3.2 Baseline Model Justification: Logistic Regression Classifier

The reason why a Logistic Regression classifier is taken as the baseline model is because of its interpretability, stability, and applicability as a source of diagnostic reference when dealing with classification problems that use noisy administrative data (Wang et al., 2013). Being a linear classifier, it furnishes a clear benchmark with regard to evaluating the possibility of explaining the results of the long-term investigation with additive relationships between crime characteristics and the status of cases. The regularization is used to solve problems of correlated predictors and heterogeneous sets of features that are typical of large datasets of police. The aim of the baseline model is not to maximize predictive performance, but to provide an objective lower bound on predictive performance on which the need and usefulness of more expressive models can be judged.

3.3 Advanced Model Justification: LightGBM and XGBoost

Gradient boosting tree ensembles are used to derive non-linear effects and interactions between features that cannot be well modeled using linear models (Ke et al., 2017; Chen and Guestrin, 2016). LightGBM is chosen as the main model because it supports large scale data, is fast on more than one million records, and has the ability to deal with class imbalance because of in-built weighting schemes. It has a histogram-based learner and tree growing via leaf-wise tree to accomplish effective representation of heterogeneous and partially noisy crime data without significant feature-transformation, making it specifically suitable for large administrative data like LAPD crime records (Ke et al., 2017).

XGBoost is considered as a second model in order to make a comparative analysis because it is a well-established gradient boosting framework with high regularization and empirical results on structured data (Chen and Guestrin, 2016). Because of the conceptual overlap between the two methods, XGBoost is utilized more as a benchmark model to determine whether observed performance improvements are consistent among boosting applications. This design choice has the benefits of making conclusions driven by modeling capacity rather than tuning intensity.

4. Data Cleaning and Preprocessing

The administrative crime data sets are gathered to be used as an operational record keeping instead of an analytical modeling and therefore results in heterogeneous fields and reporting artifacts might mislead the learning process when input directly. Based on this, a preprocessing pipeline was implemented with a structured method of preprocessing the data in order to keep the features at the time of reporting, filter out the attributes that cause leakage or noise and convert the raw records to a compact and readable feature space suitable for large scale classification.

4.1 Feature Removal and Variable Transformation

Variables such as administrative identifiers, procedural artifacts, or too high granularity were eliminated such as case numbers, reporting-specific fields, free-text descriptors and raw geographic coordinates. These attributes indicate recording processes instead of substantive crime or investigative characteristics and spurious learning. Direct or indirect variables that are dependent on the results of the cases were also avoided to prevent leakage. Very sparse or overlap representations (like secondary crime codes and excessively fine-grained categorical descriptions) were eliminated to increase the dimensionality and enhance generalization.

Selective transformation of retained variables was done to maintain the relevant information and to increase the analytical clarity. Temporal attributes which were initially saved as timestamps were separated into hour, day of week and month to provide cyclical crime patterns without reference to raw date fields. To reduce noise and enhance robustness, the victim age was grouped into readable age groups. Indicators of presence and complexity were used to summarize the modus operandi information so that the models would not overfit the unusual codes.

The K-Means clustering (k = 15) was used to represent the spatial features based on latitude and longitude. The number of cluster choices strikes a balance between spatial details and model stability. Clustering of raw coordinates was used to eliminate overfitting, and the geo-cluster obtained represented meaningful spatial structure without revealing precise locations

.

```
from sklearn.cluster import KMeans

# Apply KMeans clustering on latitude & longitude
kmeans = KMeans(n_clusters=15, random_state=42)
df['geo_cluster'] = kmeans.fit_predict(df[['latitude', 'longitude']])

# Drop raw coordinates after clustering
df.drop(columns=['latitude', 'longitude'], inplace=True)
```

Figure 8: Code Snippet Formation of (geo_cluster)

4.2 Missing Data and Outlier Analysis

Missing values are primarily indicative of reality reporting rather than problems in data. As an example, the weapon details are absent when there is no weapon. Incomplete values are handled in a manner that does not lose their operational meaning instead of indiscriminate imputation or deleting cases. Explicit *Unknown* value is assigned to categorical variables where they are appropriate. Binary indicators were also used to indicate the presence or absence of key investigative elements.

Outliers were evaluated using boxplots and Interquartile Range (IQR) approach.The distorted values represented actual investigational variability and not recording error. That's why outliers were not eliminated in order to maintain rare though valuable patterns.

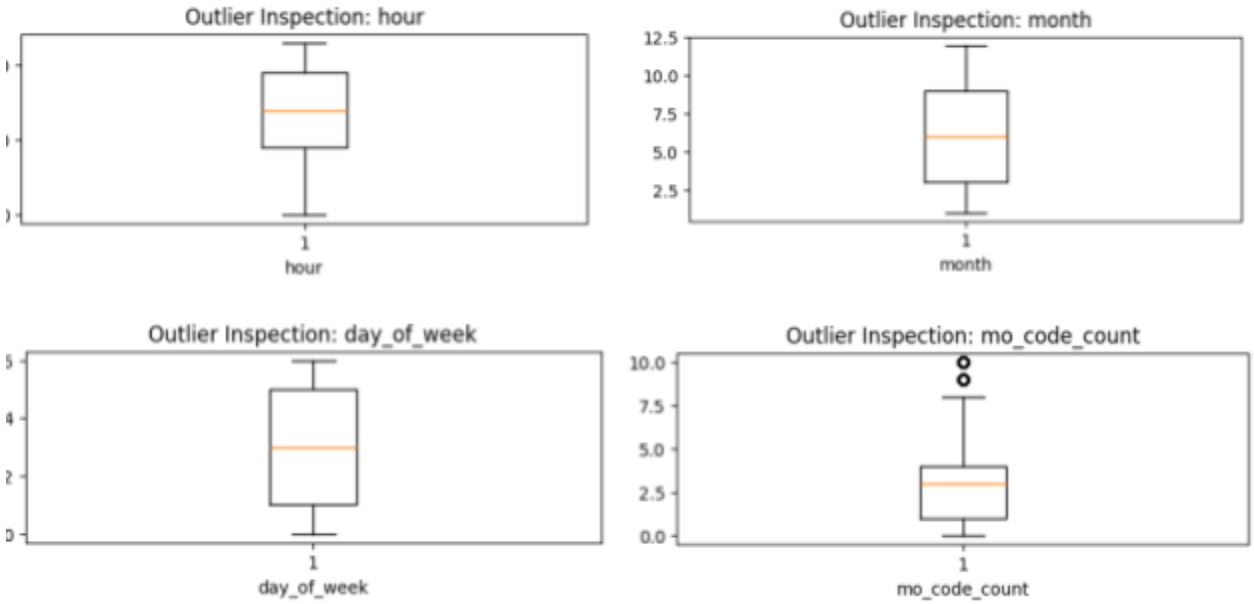

Figure 9: Boxplot-Based Outlier Inspection of Temporal and Behavioral Variables

4.3 Encoding and Feature Scaling

The categorical features were encoded according to cardinality and model requirements. To prevent data leakage, high-cardinality attributes were target encoded whereas low-cardinality features were one-hot encoded explicitly on the training data to prevent leakage. For those models capable of handling categorical data variables natively, original category representations were retained.

```
# Target encode high-cardinality features (train only)
te = TargetEncoder(cols=['crime_category', 'premise_code_grouped'])
X_train[['crime_category', 'premise_code_grouped']] = te.fit_transform(
    X_train[['crime_category', 'premise_code_grouped']], y_train
)
X_test[['crime_category', 'premise_code_grouped']] = te.transform(
    X_test[['crime_category', 'premise_code_grouped']]
)

# One-hot encode low-cardinality features
one_hot_cols = ['victim_sex', 'victim_descent', 'victim_age_group', 'weapon_group']
X_train = pd.get_dummies(X_train, columns=one_hot_cols, drop_first=True)
X_test = pd.get_dummies(X_test, columns=one_hot_cols, drop_first=True)

# Align feature space
X_test = X_test.reindex(columns=X_train.columns, fill_value=0)
```

Figure 10: Categorical Encoding Strategy with Leakage Control

The numerical characteristics were normalized by means of z-score. Scaling parameters were estimated and applied on the training data and used on the test data with methodological rigor and compatible with the linear models, as well as with the consistency of preprocessing of the experiments.

4.4 Data Partitioning and Leakage Control

Stratified random sampling was used as a preservation sampling technique of classes in the dataset to form training and testing subsets. Any transformation of such type that depends on the target (through the use of encoding and scaling) was performed only on the training data subsequently applied to the test data. Such a separation will make the performance estimates indicate true generalization and not inflation due to leakage.

4.5 Final Dataset Characteristics

After preprocessing, the data ended up having 1,003,008 observations and 16 carefully chosen and crafted features as mentioned in the table below.

Attribute	Data Type	Transformation / Construction
area, part	Integer	Retained in original numeric form as structured administrative indicators.
victim_sex, victim_descent	Categorical	Retained as reported; missing values encoded as *unknown* to preserve reporting semantics.
hour , day_of_week, month	Integer	Extracted from reported time fields to capture temporal patterns.
crime_category	Categorical	Aggregated from primary crime codes and descriptions to reduce cardinality and improve generalization
geo_cluster	Integer	Derived via K-Means clustering on latitude and longitude to encode spatial patterns
victim_age_group	Categorical	Constructed by binning victim age into interpretable age ranges
mo_present, mo_code_count	Binary, Integer	Derived from modus operandi codes to capture presence and complexity
premise_code_grouped	Categorical	Grouped from original premise codes and descriptions to address sparsity and high cardinality
weapon_group	Categorical	Aggregated from raw weapon codes into broader weapon classes
weapon_used	Binary	Derived indicator reflecting whether a weapon was involved

Table 2: Model-Ready Features After Cleaning and Feature Engineering

5. Model Development

This section focuses on how models were trained, evaluated, and systematically refined under real-world constraints such as class imbalance and feature redundancy.

5.1 Data Splitting Strategy

The finalized preprocessed dataset was further divided into training and testing groups by 80:20 stratified into two groups to maintain the original distribution of classes. Such imbalance in class distribution required stratification to ensure that the minority class were equally represented in both subsets.

All the preprocessing operations such as categorical encoding, scaling and class-weight configuration were only learned on the training set and later applied on the held-out test set to avoid data leakage. The test set was not used in any of the model selection or hyperparameter tuning steps and was only used in final evaluation.

5.2 Model Training Configuration

5.2.1 Models Implemented in the Study

Three models were trained:

- Logistic Regression served as the baseline classifier and provided cardinality-conscious categorical encoding as well as using class-weight balancing.
- LightGBM is trained using its native categorical feature processing and gradient-boosted decision trees.
- XGBoost, which is a parallel tree based boosting model to be used to compare results.

Logistic Regression is the baseline because it is simplistic, interpretable, and popular in imbalanced binary classification.

5.2.2 Optimization and Loss Functions

Each model optimized the binary log-loss objective. In the case of Logistic Regression, this is the maximum likelihood estimation under a Bernoulli assumption. The log-loss minimization in case of LightGBM and XGBoost is done through gradient-based boosting whereby new trees are trained on the remaining errors left by the previous iterations.

ROC-AUC was used to evaluate model performance mostly because it measures ranking quality independent of classification thresholds and is strong to class imbalance. Precision and recall values that depend on thresholds were evaluated later to estimate trade-offs in operations.

5.2.3 Training Parameters

In the case of Logistic Regression, training was done by convergence-based regularized optimization. In opposed gradient boosting models where training was controlled by:

- Boosting iterations.
- Learning rate
- Leaf limitations and depth of the tree.
- Instance subsampling and feature subsampling.

- Regularization parameters

Since the scale of the data set was large, to achieve stable convergence, fixed boosting rounds were used to train baseline models. Only after the optimization phases the concepts of cross-validation and early stopping were introduced.

5.3 Model Refinement and Experimental Analysis

The refinement of the models was in a gradual and evidence-based approach that was meant to maintain the natural structure of the data in addition to enhancing generalization and interpretability.

5.3.1 Baseline Model Performance (Logistic Regression)

The initial assessment of the baseline Logistic Regression model was made by native class-weight balancing without modifying the original data distribution. In this configuration, the model was able to reach a ROC-AUC of around 0.82, which is a good linear benchmark.

Recall in the minority class was quite high, whereas precision-recall trade-offs indicated the limitations of linear decision boundary to reflect the complicated interaction of features. These findings served as reference points where more expressive models were ranked.

5.3.2 Advanced Model Performance (Tree-Based Boosting)

LightGBM and XGBoost were trained with their native imbalance handling capacity that used their ability to model non-linear interactions and non-uniform feature types.

Both of the models performed better than the baseline logistics in terms of ROC-AUC with a value of 0.85. This advancement points to the fact that higher-order interactions and tree-based partitioning offer a significant predictive value over the assumption of linear modeling.

Since, LightGBM is better in terms of performance, handling large data, training efficiency and native categorical support, it was chosen as the major model to further refine.

5.3.3 Resampling-Based Experiments

The SMOTE-ENC and NearMiss hybrid resampling was tested to address the imbalance in classes to achieve a 60:40 balance(not blind balancing). Although the recall of minority-class improved slightly to about 77, the overall discriminative performance decreased, and ROC-AUC

of light GBM decreased to about 76% compared to about 85% and the same case for Logistic Regression.

The gaps between train-test performance were small and a pointer of no overfitting. The degradation observed is rather due to loss of information caused either by undersampling or geometric distortion due to synthetic samples thus resampling based methods were ruled out in future experiments.

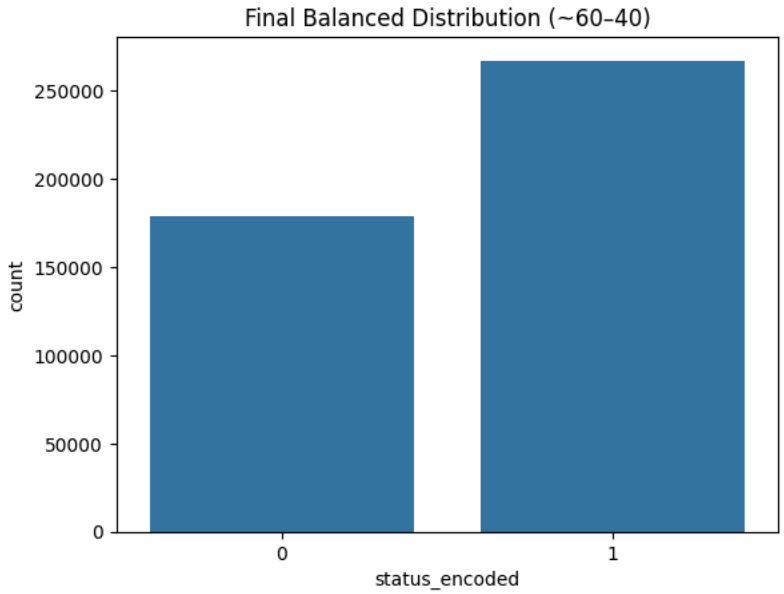

Figure 11: Class distribution before and after SMOTE-ENC and NearMiss

5.3.4 Hyperparameter Optimization

Five-fold stratified cross-validation on Bayesian hyperparameter search was also used to further optimize LightGBM. Optimization was done based on learning rate, depth of the trees, number of leaves, minimum sample on each child and strength of regularization.

```python
def objective(trial):
    return lgb.cv(
        {
            "objective": "binary",
            "metric": "auc",
            "learning_rate": trial.suggest_float("learning_rate", 0.01, 0.1),
            "num_leaves": trial.suggest_int("num_leaves", 20, 150),
            "max_depth": trial.suggest_int("max_depth", 3, 12),
            "min_child_samples": trial.suggest_int("min_child_samples", 20, 200),
            "random_state": 42
        },
        lgb.Dataset(X_train, label=y_train),
        nfold=5
    )["auc-mean"][-1]
```

Figure 12: Bayesian Hyperparameter Tuning for LightGBM

The results of this process were quite steady as they led to the value of ROC-AUC growing by around 0.86 which shows that algorithm-level tuning was more efficient to maximize performance than data-level resampling.

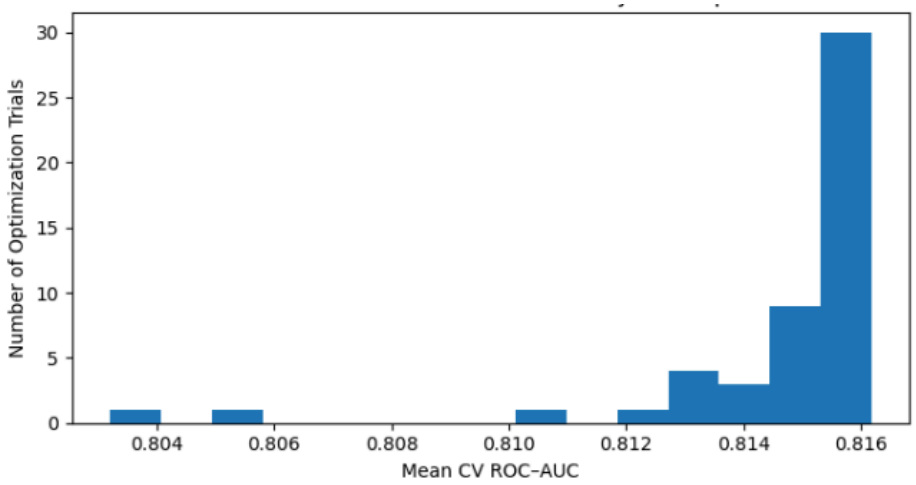

Figure 13: Distribution of Cross-Validated ROC–AUC Scores Across Bayesian Hyperparameter Optimization Trials

5.3.5 Exploratory Experiments: Cost-Sensitive Learning and Feature Redundancy

From the imbalance handling with the purpose of assessing the possible improvements, we trained a cost-sensitive LightGBM model with higher penalties associated with misclassification to the minority-class. Although this somewhat increased minority recall, there was no significant change in model performance as opposed to the baseline.

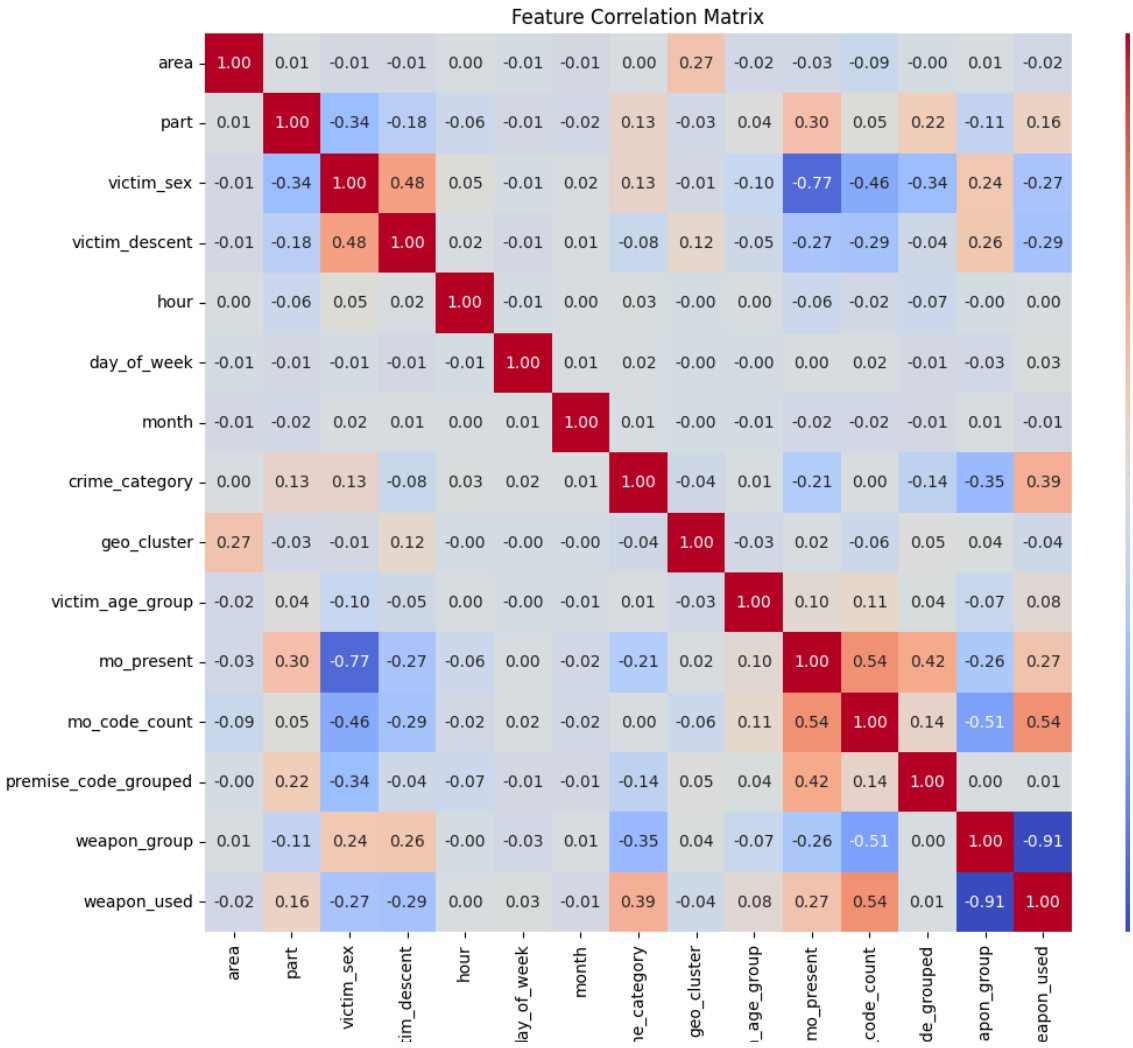

Figure 14: Feature Co-relation Heatmap

Parallel correlation analysis was done to detect highly redundant features in the final dataset. The heat map indicates that attributes such as `weapon_group` and `weapon_used` are negatively related to each other (-0.91) and `mo_present` and `mo_code_count` are strongly correlated (0.54). Upon eliminating the overlap weapon_use column the model had little effect suggesting that predictive power is largely due to larger contextual and temporal variables and not finer-scale redundancies.

In general, the findings of the model development suggest that optimization of algorithms on the original data distribution was always superior to the resampling based methods. After Bayesian tuning, LightGBM was the most successful one based on its balance of performance and robustness and thus it was chosen to be evaluated finally.

6. Evaluation Metrics and Results

This section reports about model performance based on evaluation measures that are suitable with imbalanced binary classification.

6.1 Evaluation Metrics

The ROC-AUC was mainly used to measure model performance since it quantifies the quality of ranking at every decision threshold and is not vulnerable to class imbalance. Since defining the outcomes of a minority is operationally significant, minority-class recall and minority-class precision were studied to reflect false-negative and false-positive trade-offs. Accuracy and F1-score are also reported to be complete but were not in fact used to select the model because they are threshold dependent under imbalance. For the final selected model error patterns are contextualized by use of confusion matrices.

6.2 Model Performance Comparison

The table summarizes test-set performance under native class-weighted training.

Model	ROC-AUC	Minority Recall	Minority Precision	Majority Recall	Majority Precision
Logistic Regression	0.823	0.74	0.45	77	92
LightGBM (Native)	0.856	0.80	0.45	76	94
LightGBM (optimized)	0.860	0.79	0.46	77	94
XGBoost	0.854	0.80	0.45	76	94

Table 3: Model Performance Comparison

Tree-based models were always performing better than the Logistic Regression on ROC-AUC, which means that the tree-based model is a better classifier that relies on non-linear relations among features. Bayesian optimization was found to produce a small yet steady improvement over the native baseline LightGBM with no manipulation of the original data distribution.

6.3 Error Analysis

The analysis of errors of the optimized LightGBM model revealed that the majority of the misclassifications were false negative ones, and most of the cases assigned to the predicted probability were very near the decision threshold (about 0.45-0.55). This trend implies overlapping intrinsic features rather than confident overfitting or misclassification. Comparisons

of performance between baseline, resampled and optimized settings showed consistent train-test behavior, which was explained by the distortion in distributions and not variance inflation.

6.4 Interpretability

SHAP analysis showed the patterns of most frequently occurring features to correctly classified majority class cases indicating that prediction failures are not due to model failure instead due to intrinsic feature overlap. Top contributors to the prediction of the unresolved cases in the Logistic Regression model were `crime_category`, `weapon_use`, `weapon_group_unknown`, and `victim_age_group(0-12)` whereas the effect of operational variables such as area was lower compared to LightGBM.

In the case of the LightGBM, `mo_code_count`, `crimecategory`, and `weapon_group` were very likely to contribute to the probability of an unresolved case whereas some `geo_cluster` values decreased it. Contextual: `area` and `part` moderately affected predictions, which proves that operational and crime-context factors are the largest contributors to the model output.

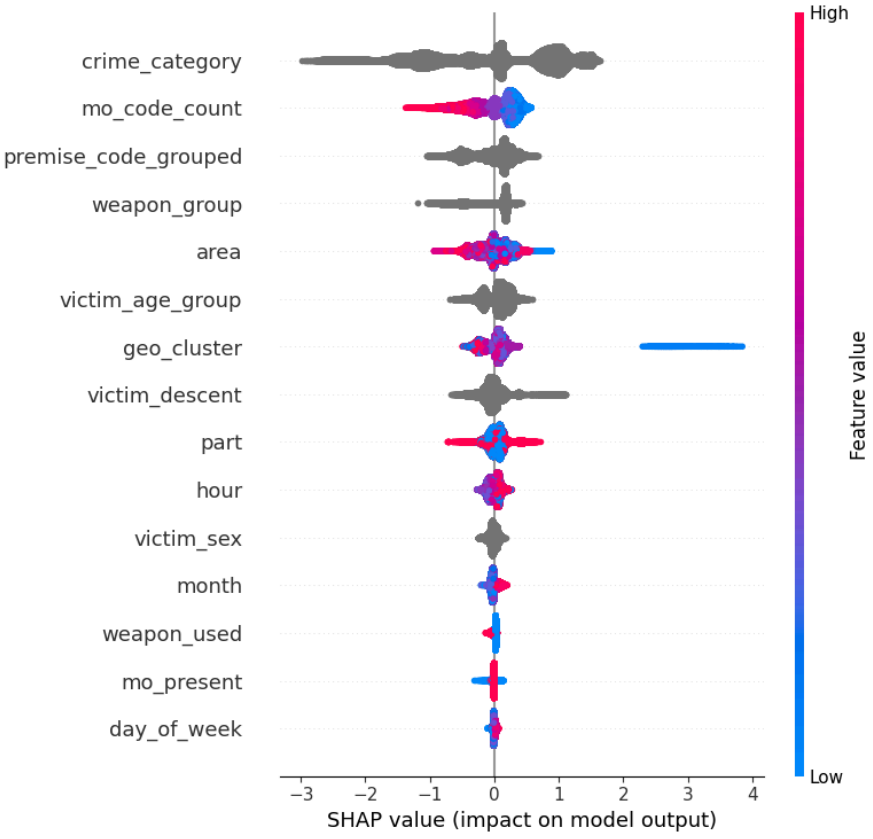

Figure 15: SHAP Summary Plot(lightGBM)

6.5 Model Limitations

The Limitations remain in spite of a good ranking performance. The model is based on the static feature representations, and it cannot represent the dynamics of time and investigation. Additionally, overlapping class characteristics constrain achievable recall, and resampling methods cannot generate truly novel discriminatory structure. Addressing these limitations requires richer contextual and temporal data.

7. Conclusion

7.1 Summary of Findings

This paper shows that machine learning algorithms, especially gradient boosting models such as LightGBM, have the potential to be effective in predicting the probability of crime cases being resolved with the help of administrative data. The use of natural class-weighted learning and preservation of the natural class distribution was consistently better than the resampling strategies, which manipulated the data and impaired the general predictive performance. The interpretability analysis of SHAP showed that the strongest influence was on the operational and crime-context features, including the type of crime, their complexity of the modus operandi, the type of premises, the weapon use, and the area of jurisdiction, and the impact of temporal variables was relatively low. The findings indicate that predictive models may give actionable advice on law enforcement agencies, assist in prioritizing its investigations that might not have been resolved, and assist in resources allocating decisions considering the inherent constraints of overlapping classes features and fixed features representations.

7.2 Future Work and Recommendations.

There are multiple possibilities to develop this work and make it more relevant in its operations. The temporal or sequential modeling methods might be also more appropriate to reflect the changing nature of investigation and the recurring patterns of offenses. Optimization based on cost sensitivity and adaptive decision-levels can also help optimize model outputs by prioritizing policing based on high-severity crime.

The extension of contextual characteristics to incorporate socio-economic, geographic, and policing resource information might be useful to record the latent effects on the case resolution outcomes. Moreover, predictive models, which the police case management systems are equipped with and assisted by interpretable dashboards, may allow making proactive and transparent decisions. Lastly, it is important that future studies be clear on ethical issues and maybe biases in historical data on crime to make sure that predictive instruments are used to facilitate fair and accountable policing procedures.

8. References

Berk, R.A. (2021) 'Artificial intelligence, predictive policing, and risk assessment for law enforcement', *Annual Review of Criminology*, 4(1), pp. 209–237.

Blumstein, A. and Wallman, J. (2006) 'The crime drop and beyond', *Annual Review of Law and Social Science*, 2, pp. 125–146.

Burrows, J. and Tarling, R. (1987) 'The investigation of crime in England and Wales', *The British Journal of Criminology*, 27(3), pp. 229–251.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.

Chen, T. and Guestrin, C. (2016) 'XGBoost: a scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Coupe, T. and Griffiths, M. (1999) 'The influence of police actions on victim satisfaction in burglary investigations', *International Journal of the Sociology of Law*, 27, pp. 413–431.

He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA.

Lee, Y.H. (2020) 'How police policies and practices impact successful crime investigation: factors that enable police departments to "clear" crimes', *Justice System Journal*, 41, pp. 1–25.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA.

Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P. and Tita, G.E. (2011) 'Self-exciting point process modeling of crime', *Journal of the American Statistical Association*, 106(493), pp. 100–108.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) 'CatBoost: unbiased boosting with categorical features', *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada.

Richardson, R., Schultz, J.M. and Crawford, K. (2019) 'Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice', *NYU Law Review Online*, 94, pp. 192–233.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalvesv, A.H. (2019) 'Data imbalance in classification: experimental evaluation', *Information Sciences*, 513, pp. 429–441.

Wang, C., Han, B., Patel, B. and Rudin, C. (2023) 'In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction', *Journal of Quantitative Criminology*, 39, pp. 519–581.

Wang, T., Rudin, C., Wagner, D. and Sevieri, R. (2013) 'Learning to detect patterns of crime', in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML-PKDD 2013, Lecture Notes in Computer Science*, vol. 8190 LNAI, pp. 515–530.

9. Appendix

- Dataset Source: *crime_dataset from 2020 to Present*
- EDA Notebook Source: *crime_case_resolution(LAPD).ipynb*
- Model Training Notebook Source: *modelTrainingNotebook.ipynb*

#	Column	Non-Null Count	Dtype
0	division_number	1004991 non-null	int64
1	date_reported	1004991 non-null	object
2	date_occurred	1004991 non-null	object
3	time_occ	1004991 non-null	int64
4	area	1004991 non-null	int64
5	area_name	1004991 non-null	object
6	reporting_district	1004991 non-null	int64
7	part	1004991 non-null	int64
8	crime_code	1004991 non-null	int64
9	crime_description	1004991 non-null	object
10	modus_operandi	853372 non-null	object
11	victim_age	1004991 non-null	int64
12	victim_sex	860347 non-null	object
13	victim_descent	860335 non-null	object
14	premise_code	1004975 non-null	float64
15	premise_description	1004403 non-null	object
16	weapon_code	327247 non-null	float64
17	weapon_description	327247 non-null	object
18	status	1004990 non-null	object
19	status_description	1004991 non-null	object
20	crime_code_1	1004980 non-null	float64
21	crime_code_2	69160 non-null	float64
22	crime_code_3	2314 non-null	float64
23	crime_code_4	64 non-null	float64
24	location	1004991 non-null	object
25	cross_street	154236 non-null	object
26	latitude	1004991 non-null	float64
27	longitude	1004991 non-null	float64

Table 4: Initial Dataset Inspection [data.info()]

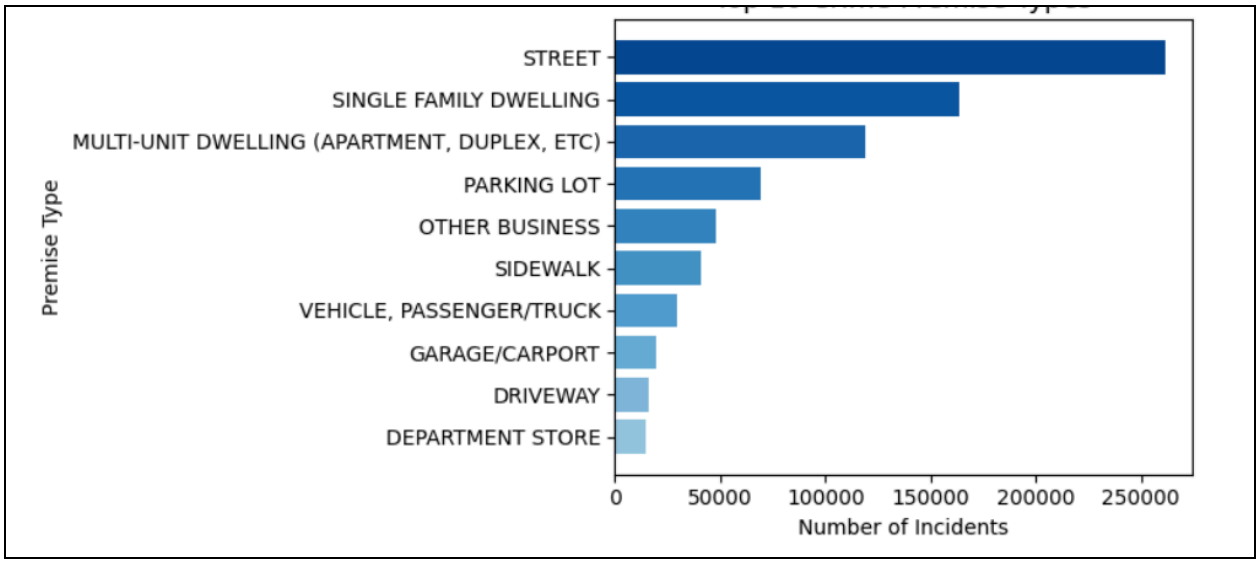

Figure 16: Top 10 Crime Premise Types

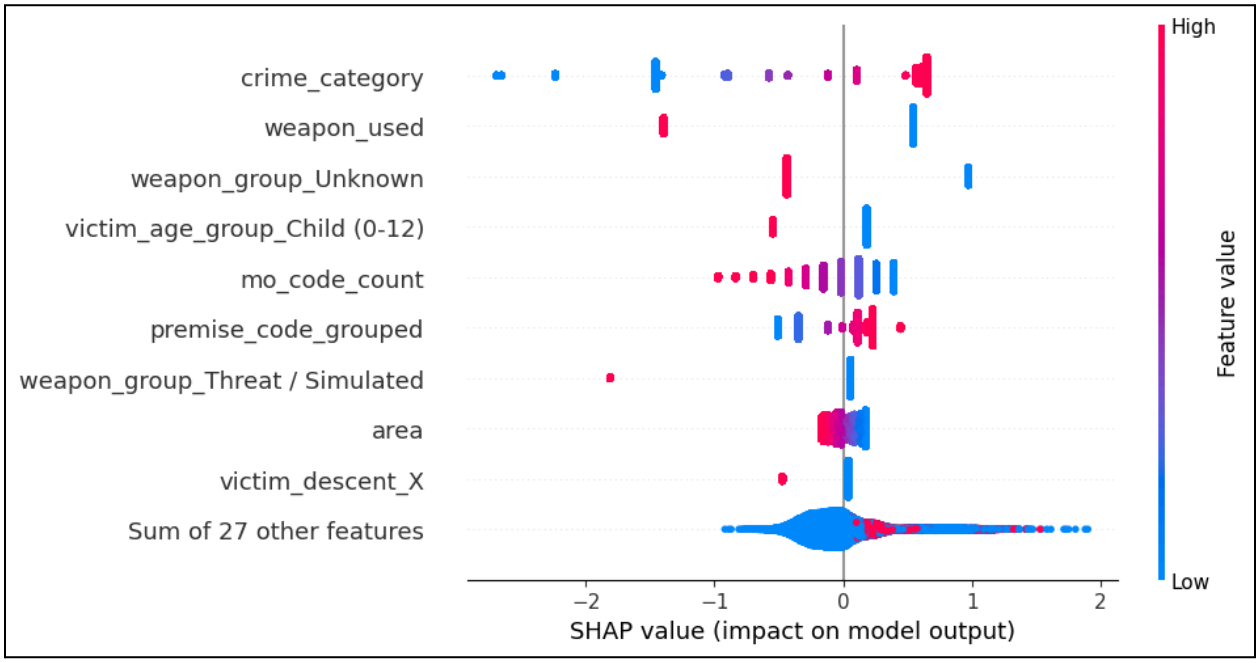

Figure 17: SHAP Summary Plot (logistic Regression)

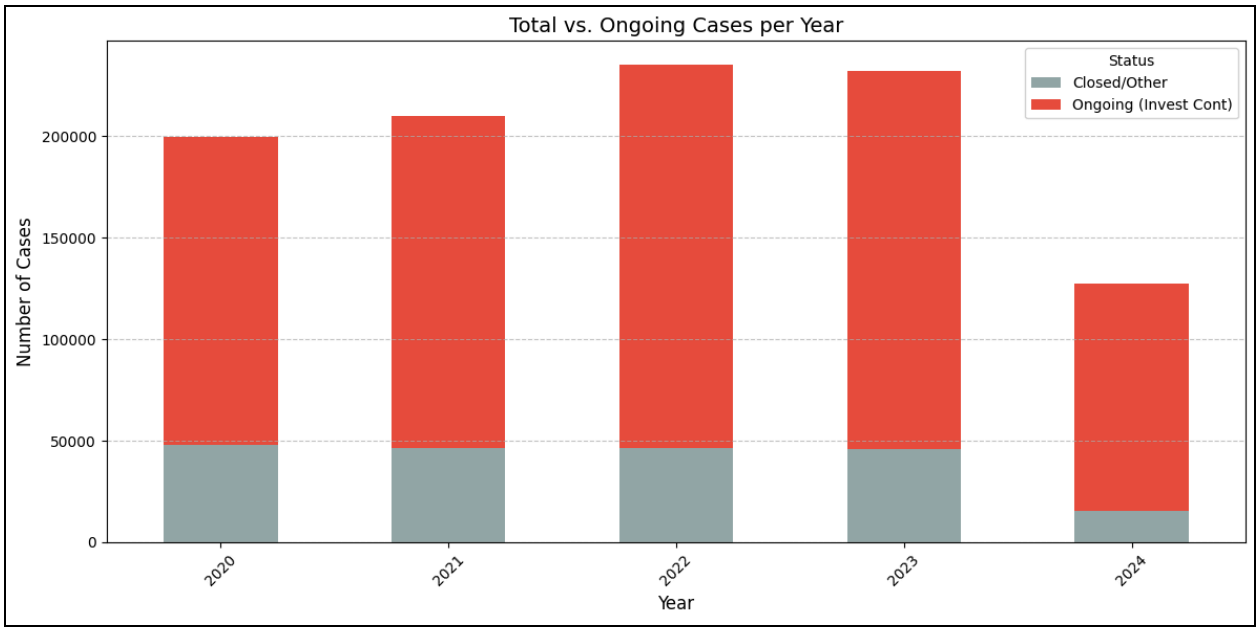

Figure 18: Resolved vs Investigation Ongoing Cases per year

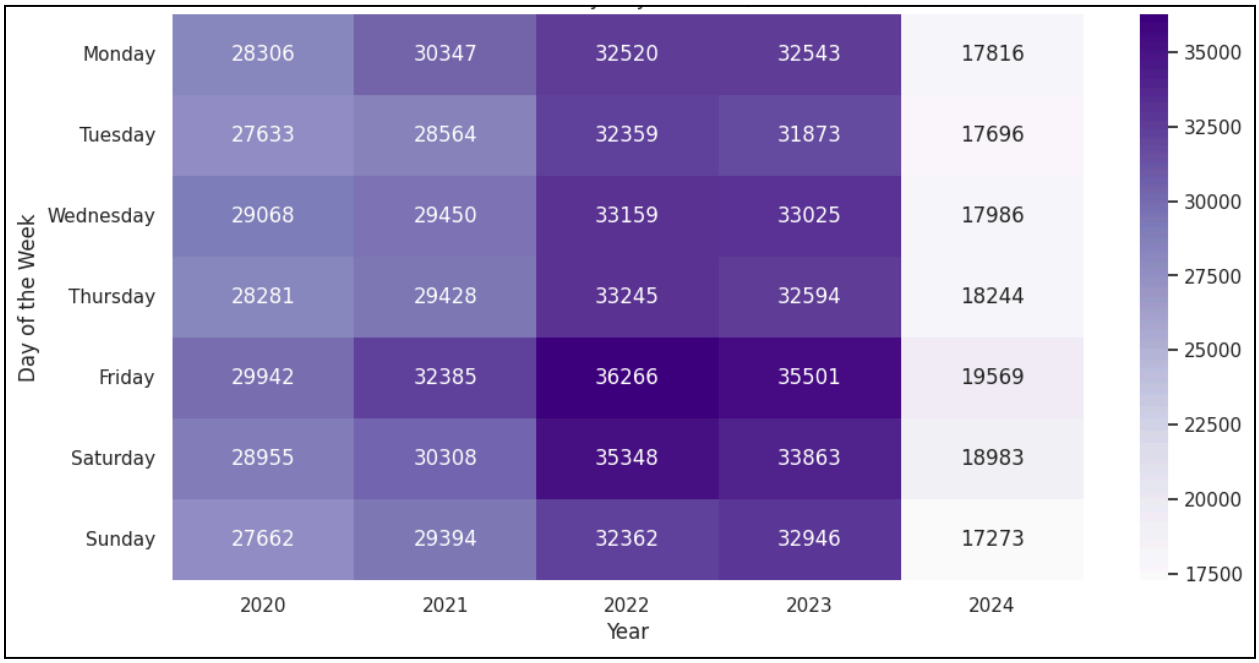

Figure 19: Crime Distribution by Day of Week

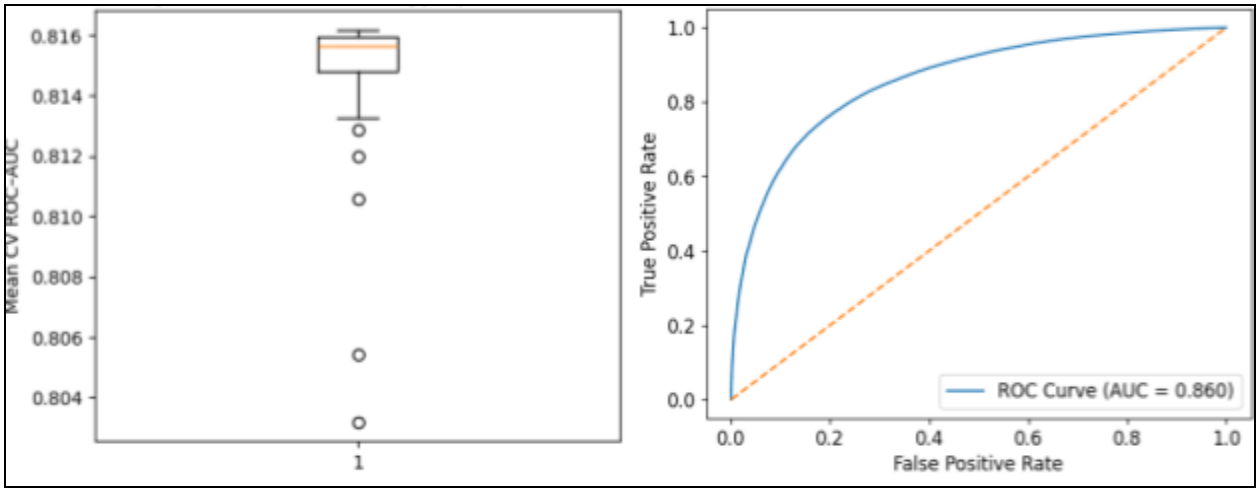

Figure 20 : Stability of ROC-AUC Across Hyperparameter Trials

Figure 21: ROC Curve-Optimized LightGBM

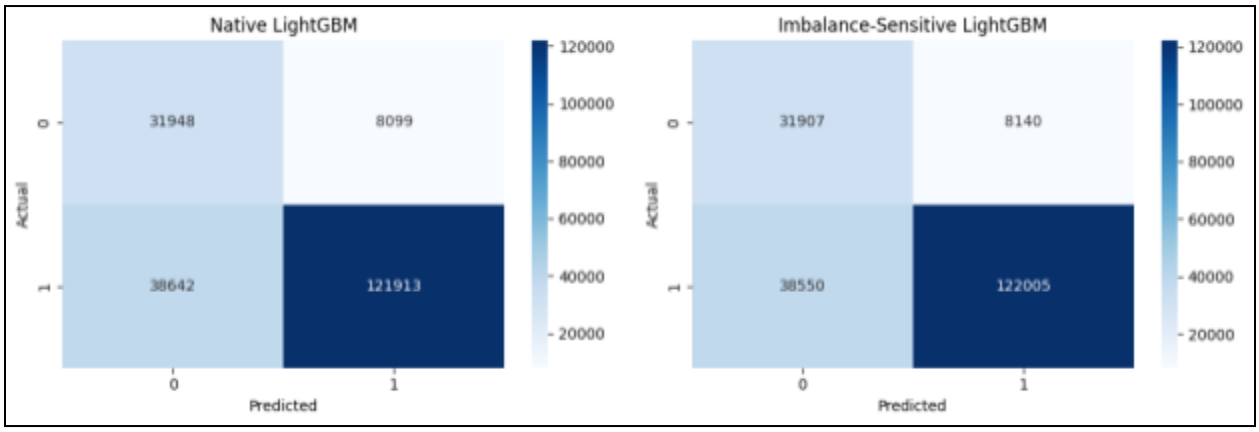

Figure 22: Confusion Matrix Native LightGBM vs Imbalance -Sensitive Performance

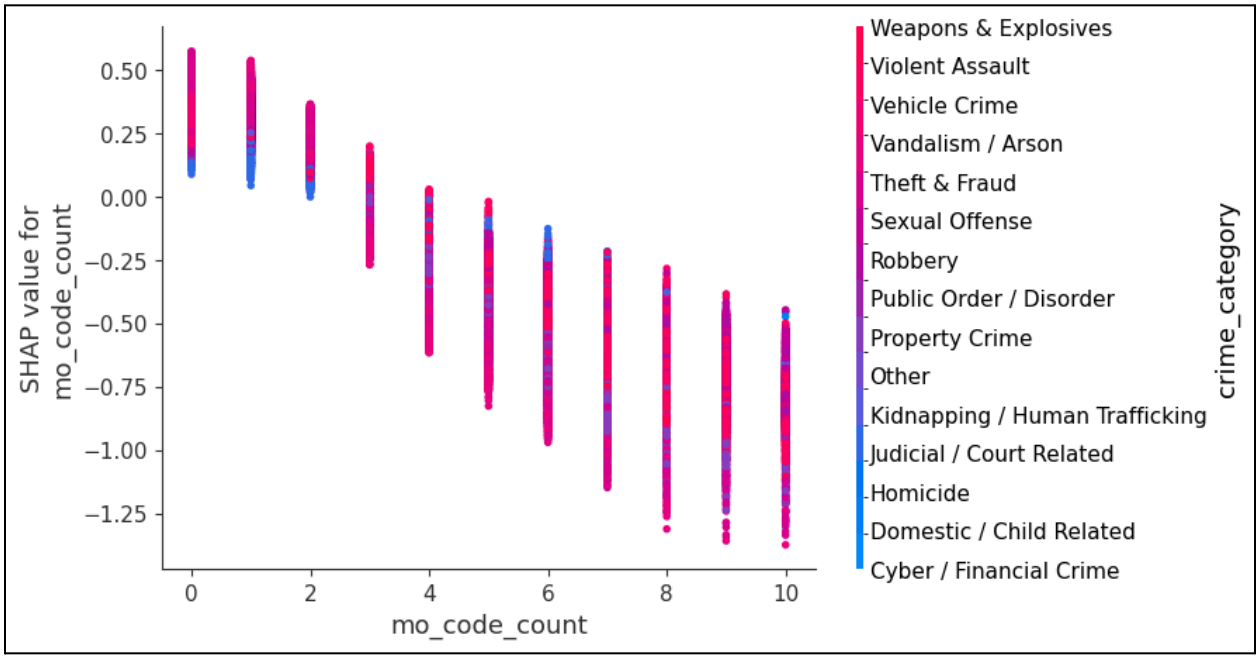

Figure 23: SHAP Dependence Plot (mo_code_count)

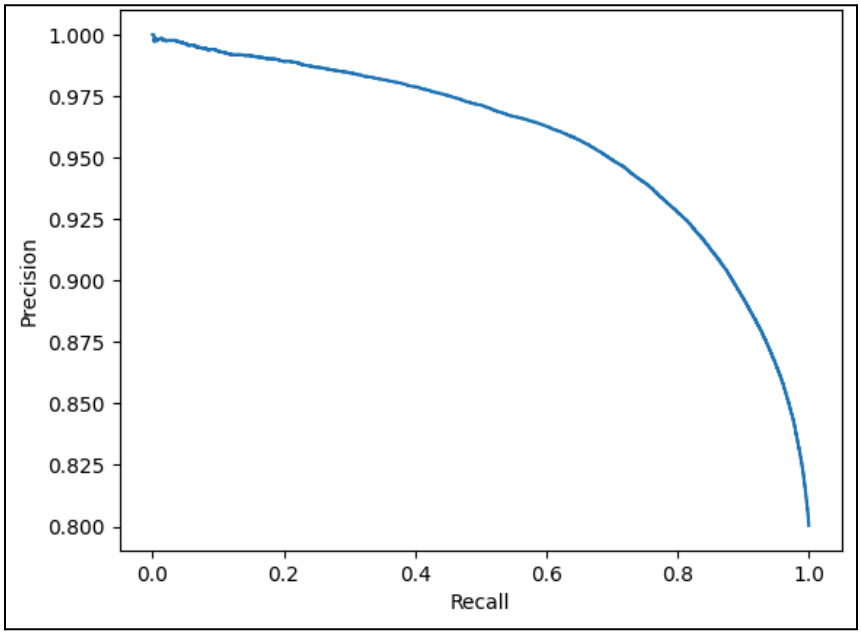

Figure 24: Precision-Recall Curve -LightGBM