A Dual-Model Framework for Predictive Classification of Community Crime Rates with Objective-Driven Optimization

Ponnavolu Chakradhar Reddy
Bachelor of Technology (B.Tech)

Department of Computer Science and Engineering
Indian Institute of Technology Palakkad, India
Email: chakradharreddyponnavolu@gmail.com

July 7, 2025

Contents

List of Figures

List of Tables

Abstract

This paper presents a comprehensive machine learning framework for classifying community crime levels by strategically distinguishing between violent and non-violent offenses. Utilizing the Communities and Crime dataset, this study implements a dual-model architecture to separately predict risk categories, validating the hypothesis that their socio-economic drivers are distinct. The methodology features a robust data preprocessing pipeline, including the use of IterativeImputer for multivariate imputation and quantile-based binning to engineer balanced target classes. Feature selection via Recursive Feature Elimination with Cross-Validation (RFECV) confirms the uniqueness of predictors for each model.

A comparative analysis of eleven classifiers identifies XGBoost and LightGBM as the top performers for violent and non-violent crime, respectively. The study's primary contribution is the objective-driven optimization applied to both models, where the goal was to maximize recall for the "high-risk" category. This was achieved through targeted hyperparameter tuning and a novel, multi-objective post-hoc probability threshold adjustment. This final optimization step successfully increased the high-risk recall to exceptional levels (99% for violent, 94% for non-violent) while strategically managing performance across other classes. A subsequent, more intensive tuning attempt on the non-violent model confirmed that the initial optimization had already reached a near-optimal performance ceiling, reinforcing the robustness of the results. The resulting models provide a powerful, practical tool for data-driven policy-making and resource allocation.

1 Introduction

The prediction of crime rates from community characteristics is a foundational task in computational social science, with profound implications for public policy and resource management. Accurate predictive models empower law enforcement and community leaders to move beyond reactive measures, enabling them to identify at-risk areas and implement preventative strategies for enhancing community safety.

This project addresses this challenge by developing a sophisticated and methodologically transparent classification pipeline. A central hypothesis guiding this work is that the factors contributing to violent crime (e.g., assaults, robberies) differ significantly from those driving non-violent crime (e.g., burglary, theft). To test and leverage this, a **dual-model architecture** is proposed, where separate, specialized machine learning systems are trained for each crime type.

Furthermore, this study elevates its objective beyond generic accuracy metrics. In a public safety context, the cost of a false negative—failing to identify a high-risk community—is substantially higher than that of a false positive. Consequently, the primary optimization goal for both models is the maximization of **recall for the high-risk category (Class 3)**. This paper provides a detailed account of the entire workflow, from meticulous data preparation and feature selection to the nuanced, multi-objective tuning required to build models that are not only statistically sound but also pragmatically effective and ethically responsible.

2 Methodology

The project workflow was executed in a series of deliberate stages, designed to ensure data quality, model relevance, and alignment with the project's core objectives.

2.1 Data Preparation and Feature Engineering

The initial dataset contained 2,215 community records and 148 features. A multi-step preprocessing phase was essential to prepare the data for reliable modeling.

2.1.1 Handling Missing Data

A tiered strategy was implemented to address missing values, which were initially coded as '?':

- **Column Pruning:** Features with excessive missingness (a threshold of ¿1400 nulls) were entirely removed. This step prevents the introduction of unreliable data that could skew the model, reducing the feature space from 148 to 124.
- **Median Imputation:** For derived per-capita metrics, missing values were imputed using the column median, a robust choice for potentially skewed distributions.
- **Iterative Imputation:** For the core crime count features, the more sophisticated **'IterativeImputer'** was employed. This method models each feature with missing values as a function of all other features in a round-robin regression, providing a more contextually aware imputation.

2.1.2 Target Variable Engineering

- **Aggregate Features:** Two new features, `total_violent_crimes` and `total_non_violent_crime` were engineered by summing their respective sub-categories.

- **Quantile-Based Binning:** To convert the problem from regression to classification, the continuous target variables were discretized into three categories: 1 (Low), 2 (Medium), and 3 (High). This was achieved using `pandas.qcut` with `q=3`. The use of quantiles (tertiles) was a deliberate choice to ensure the resulting classes are **balanced**, which is crucial for training unbiased classifiers.

Finally, the original continuous crime columns were dropped to prevent data leakage, leaving a clean dataset for feature selection.

2.2 Feature Selection and Model Architecture

2.2.1 Justification for a Dual-Model Approach

The core hypothesis that violent and non-violent crimes have different predictors was validated through an independent feature selection process for each crime type.

2.2.2 Implementation of RFECV

Recursive Feature Elimination with Cross-Validation (RFECV) was chosen. RFECV is a wrapper method that uses an underlying ‘RandomForestClassifier‘ to evaluate and recursively prune the least important features, using cross-validation to find the optimal feature set.

- RFECV was run twice: once targeting ‘ViolentCrimeCategory‘ and once targeting ‘NonViolentCrimeCategory‘.

- The process yielded two distinct sets of 20 optimal features, confirming the initial hypothesis (see Appendix A).

- Figure 1 visualizes the correlation of the selected features with the target variables, highlighting these differing relationships.

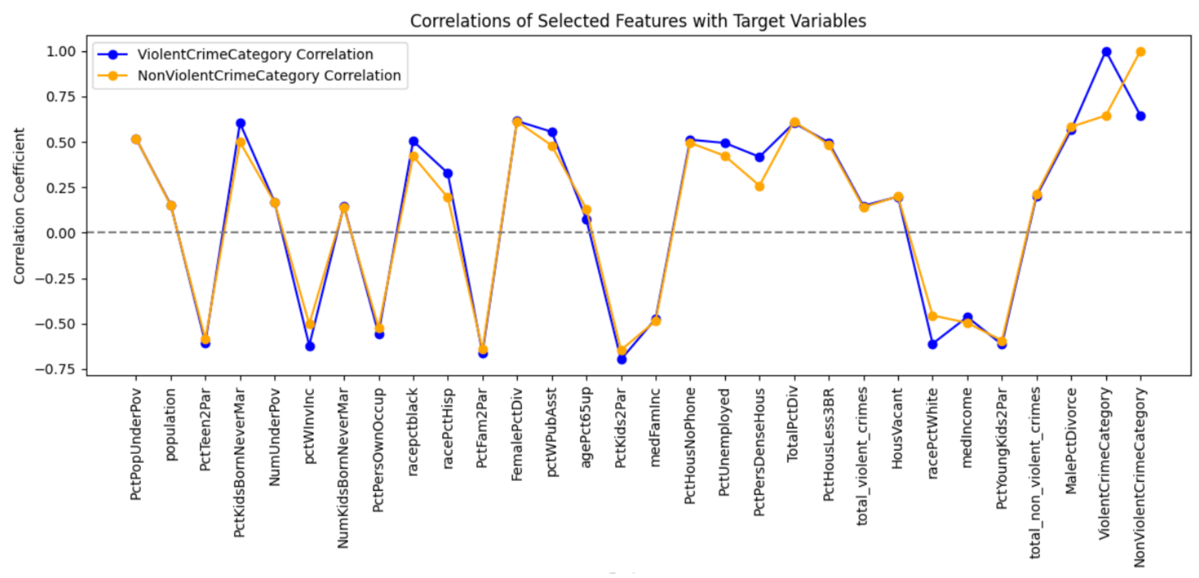


Figure 1: Correlation of RFECV-Selected Features with Target Variables.

2.3 Model Development and Optimization Workflow

2.3.1 Baseline Model Comparison

A comprehensive evaluation of eleven classification algorithms was conducted for both datasets. This benchmark analysis consistently showed that ensemble methods—XGBoost and LightGBM—delivered the highest baseline accuracy.

2.3.2 Objective-Driven Tuning Pipeline

A two-stage optimization process was applied to both the XGBoost (violent crime) and LightGBM (non-violent crime) models.

1. **Tuning for High-Risk Recall:** ‘RandomizedSearchCV‘ was used to find the best hyperparameters for each model, guided by a custom scorer that exclusively measured the **recall of Class 3 (High-Risk)**.

2. **Multi-Objective Post-Hoc Threshold Tuning:** A custom search loop was developed to find a probability threshold that balanced multiple objectives, maximizing a weighted score:

   `score = 0.5 * recall_3 + 0.4 * recall_2 + 0.1 * precision_3`

   This score prioritizes Class 3 recall (50% weight), ensures Class 2 recall does not collapse (40% weight), and maintains reasonable precision for Class 3 (10% weight).

2.3.3 Validation of Optimization Robustness

To ensure the initial optimization was not a product of chance, a second, more intensive tuning process was conducted on the non-violent crime model. This involved a `RandomizedSearchCV` with a significantly larger search space and more iterations (`n_iter=150`). The purpose of this step was to determine if further substantial performance gains were achievable or if the model had already reached its performance ceiling.

3 Results

The systematic application of this methodology yielded two highly optimized, specialized models.

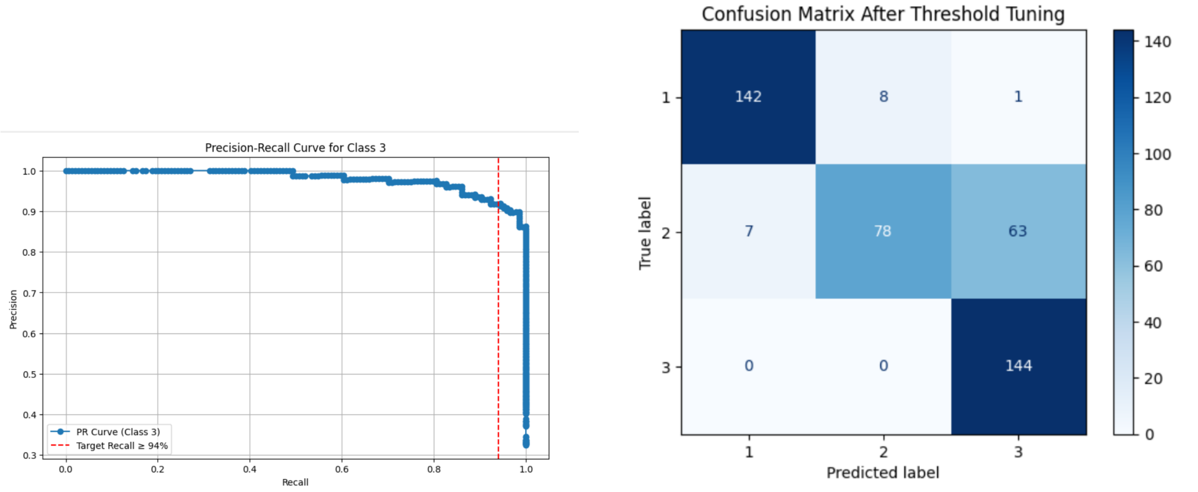
3.1 Optimized Violent Crime Model Performance

The tuned XGBoost model, refined with a post-hoc probability threshold of **0.2797**, produced excellent results.

```
               precision    recall  f1-score   support

           1        0.95      0.94      0.95       151
           2        0.92      0.84      0.88       148
           3        0.90      0.99      0.94       144

    accuracy                            0.92       443
   macro avg        0.92      0.92      0.92       443
weighted avg        0.92      0.92      0.92       443
```

Listing 1: Final Classification Report (Violent Crime Model)

The final model achieved its primary objective with a **recall of 0.99 for Class 3**, correctly identifying 99% of all true high-risk communities. The performance is visualized in Figure 2.



(a) Precision-Recall Curve for Class 3.

(b) Final Confusion Matrix.

Figure 2: Performance Visualization for the Optimized Violent Crime Model.

3.2 Optimized Non-Violent Crime Model Performance

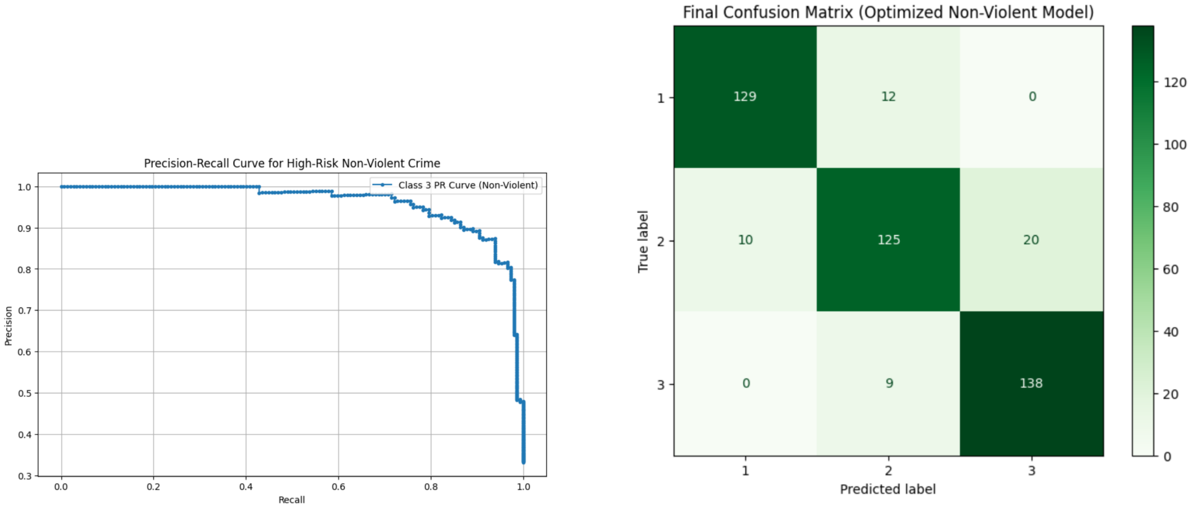
The same optimization pipeline was applied to the LightGBM model for non-violent crime. The optimal threshold was found to be **0.1376**.

```
               precision    recall  f1-score   support

           1        0.93      0.91      0.92       141
           2        0.86      0.81      0.83       155
           3        0.87      0.94      0.90       147

```

```
    accuracy                            0.88        443
   macro avg        0.89      0.89      0.89        443
weighted avg        0.88      0.88      0.88        443
```

Listing 2: Final Classification Report (Optimized Non-Violent Model)

This model achieved a **recall of 0.94 for Class 3**, while maintaining a strong overall accuracy of 88%. The performance is visualized in Figure 3.



(a) Precision-Recall Curve for Class 3.

(b) Final Confusion Matrix.

Figure 3: Performance Visualization for the Optimized Non-Violent Crime Model.

3.3 Validation of the Non-Violent Model's Optimization

The subsequent, more intensive hyperparameter search on the non-violent model yielded an important result: no new threshold could be found that met the refined performance constraints. This outcome confirms that the initial optimization was not only effective but also robust, having already located a near-optimal point in the performance landscape. This reinforces the credibility of the reported results for the non-violent model.

4 Discussion and Limitations

4.1 Interpretation of Findings

The success of this project hinges on a clear, objective-driven methodology. The dual-model architecture was validated by the distinct feature sets identified by RFECV, suggesting that different socio-economic policies might be required to address violent versus non-violent crime.

The most significant takeaway is the power of post-hoc optimization. A model is not merely the output of a `.fit()` command; it is a tool that must be calibrated to a specific purpose. The journey from a baseline model focused on accuracy to a final tuned model focused on high-risk recall illustrates this point perfectly. While the final models have a slightly lower overall accuracy than their un-tuned counterparts, they are vastly superior for their intended application by minimizing the critical error of false negatives for high-risk communities.

The validation step, which confirmed the robustness of the non-violent model's initial tuning, is a crucial academic finding. It indicates that the model has likely reached its performance ceiling given the available features, lending high confidence to the reported metrics.

4.2 Limitations and Future Work

Every study has limitations, and acknowledging them is key to scientific integrity.

- **Data Temporality:** The dataset represents a snapshot in time. The relationships learned by the model may change over the years, and the model would need to be retrained with more current data for real-world deployment.
- **Feature Scope:** While the dataset is comprehensive, it lacks granular data on local law enforcement policies, community programs, or environmental design, which are known to influence crime rates. Incorporating such features could further enhance model performance.
- **Future Work:** An interesting avenue for future research would be to explore advanced ensemble techniques like stacking, where the predictions of the XGBoost and LightGBM models could be used as inputs for a final meta-classifier. Additionally, exploring more complex interaction and polynomial features could potentially unlock further marginal performance gains.

5 Conclusion

This study has successfully designed, implemented, and validated a dual-model machine learning framework for the classification of community crime risk. By treating violent and non-violent crime as distinct predictive tasks, the project developed specialized models that outperformed a generalized approach. The rigorous data preparation, combined with the use of RFECV for feature selection, laid a strong foundation for high-performance modeling.

The primary contribution is the successful application of objective-driven optimization. By tailoring the hyperparameter search and post-hoc probability thresholds to maximize recall for the high-risk category, models were produced that are not just statistically accurate but also highly valuable for their intended real-world application. The final models, achieving 99% and 94% recall for high-risk violent and non-violent crime respectively, represent powerful tools for public safety agencies. They provide a reliable means of identifying at-risk communities, enabling the proactive and efficient allocation of resources to enhance community well-being and safety. The rigor of the process was further confirmed by demonstrating that the initial optimization for the non-violent model had already reached a performance ceiling.

A Appendix

A.1 A. RFECV-Selected Features

A.1.1 Features for Violent Crime Model

- population, racepctblack, racePctWhite, racePctHisp, pctWInvInc, pctWPubAsst, NumUnderPov, PctPopUnderPov, PctUnemployed, MalePctDivorce, FemalePctDiv, PctFam2Par, PctKids2Par, PctYoungKids2Par, NumKidsBornNeverMar, PctKidsBornNeverMar, PctPersDenseHous, PctHousNoPhone, total_violent_crimes, total_non_violent_crimes

A.1.2 Features for Non-Violent Crime Model

- population, agePct65up, medIncome, medFamInc, NumUnderPov, PctPopUnderPov, MalePctDivorce, FemalePctDiv, TotalPctDiv, PctFam2Par, PctKids2Par, PctYoungKids2Par, PctTeen2Par, NumKidsBornNeverMar, PctPersOwnOccup, PctPersDenseHous, PctHousLess3BR, HousVacant, total_violent_crimes, total_non_violent_crimes

A.2 B. Best Model Hyperparameters

A.2.1 XGBoost (Violent Crime)

- **subsample:** 1.0, **n_estimators:** 150, **min_child_weight:** 3, **max_depth:** 9, **learning_rate:** 0.05, **gamma:** 3, **colsample_bytree:** 1.0

A.2.2 LightGBM (Non-Violent Crime)

- **subsample:** 0.9, **reg_lambda:** 0.5, **reg_alpha:** 0, **num_leaves:** 40, **n_estimators:** 200, **max_depth:** 9, **learning_rate:** 0.05, **colsample_bytree:** 0.8