Crime Classification Prediction in San Francisco

Amish Gupta, Lauryn Nakamitsu, Samantha Tang, Alex Wong

IEOR 142 Final Project Fall 2024

1 Introduction

Crime prediction is crucial for enhancing urban safety and public security in San Francisco, a city with over 800,000 residents and nearly 60,000 crimes reported annually. Accurate crime forecasting could significantly improve the effectiveness of law enforcement interventions and ensure community safety. San Francisco faces continual challenges in distributing law enforcement resources effectively, especially given the unpredictable spikes in crime across different districts. With an average of 164 incidents per day, the police force's resources are often stretched thin. This project is driven by the need to leverage data-driven decisions to optimize police deployment and responses, aiming to reduce crime rates and improve the quality of life for all city residents. The research focuses on predicting crime types and patterns in San Francisco by analyzing historical crime data along with sociodemographic and environmental factors. The goal is to develop advanced predictive models that more accurately identify crime categories and adapt to the dynamic urban environment. These models will assist in the strategic positioning of law enforcement, improving resource allocation and the overall efficiency and effectiveness of crime prevention efforts.

1.1 Data Collection

Our dataset came from the "CrimeCast: Forecasting Crime Categories" competition on Kaggle, which includes detailed crime data with incident specifics, victim demographics, and locations.

We cleaned and processed the data to prepare it for analysis, including dummy encoding categorical variables like Victim_Sex and Victim_Descent, while removing less relevant categories such as Victim_Sex_nan. Date fields were converted to date-time objects, enabling the extraction of features like year, month, and day to capture crime trends over time.

Additionally, we removed less predictive columns such as Location, Cross_Street, and Status_Description to focus on more impactful features. These steps improved model accuracy and interpretability by streamlining the dataset and emphasizing key variables.

1.2 Model Selection

Our model selection process was divided into distinct phases to evaluate the impact of various stages in model development while keeping other variables and model changes controlled.

We categorized our models into two groups: baseline and advanced. The baseline models served as a foundation for comparison, enabling us to measure the effectiveness of more sophisticated approaches. For advanced modeling, we employed logistic regression, random forest, gradient boosting, CART, and LDA. The objective was to construct predictive models capable of accurately forecasting the category of crime for each incident. This task involved a multi-class predictive modeling dataset, which introduced additional complexities beyond those encountered in binary classification problems.

1.3 Advanced Modeling

The dataset was split into training and test sets with a 70/30 split to ensure proper eval-

uation. Accuracy was used as the primary performance metric, with the baseline model achieving 57.68% accuracy by always predicting the most frequent class.

For advanced models, categorical variables like "Victim Sex," "Victim Descent," and "Status" were transformed into numerical features through dummy encoding, enabling compatibility with models such as Random Forest, Logistic Regression, and Gradient Boosting. While LDA and CART were initially considered, they struggled to handle the dataset's complexity and delivered lower accuracy compared to the other models.

Random Forest was enhanced using cross-validation, which reduced overfitting and ensured consistent performance across multiple data subsets. We also optimized the model by identifying and removing irrelevant features using its "feature_importance" attribute, retaining only the top 10 most impactful variables. This approach improved both the model's efficiency and predictive accuracy, allowing it to generalize effectively to unseen data, which is critical for crime prediction in a dynamic urban setting like San Francisco.

1.4 Sentiment Analysis

We employed the use of sentiment analysis on the "Weapon_Description" column, which denotes the description of the weapon code in words. We utilized sentiment analysis NLP to predict Crime_Category. In order to do this, we isolated the "Weapon_Description" column as the only dependent variable to separate its impact on our models. In our initial analysis, we can see how words like "threat" can likely predict the Crime_Category due to how it is heavily prevalent in the category "Crimes against Public Order."

In our analysis, the bag of words initially contained a total of 99 terms. We removed words that appeared in fewer than one percent of weapon descriptions in our training dataset (n=0.01×7335=73.35). Additionally, we eliminated "stop words," such as prepositions and articles (e.g., "on" and "the").

This yielded gradually improved accuracy scores in regards to all models on the sentiment columns alone with all advanced models achieving approximately a 70% .

1.5 Performance Metrics

We evaluated each model primarily based on their accuracy score, which measures the proportion of correctly predicted instances out of all predictions made. Accuracy is a straightforward and intuitive metric, making it easy to compare the performance of different models on the same task. Since the goal of this project is to predict crime categories high accuracy score is essential to ensure that the model is reliably identifying the most critical crime hot spots.

1.6 ROC Curves & Model Comparisons

The ROC curve provides a method to evaluate the performance of our logistic regression model to classify crime categories in the data set. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across all thresholds, offering a threshold-independent evaluation. The Area Under the Curve (AUC) summarizes this performance, where higher values indicate better discriminative ability. For our model, the AUC values reveal varying performance across crime categories: "Crimes against Persons" (AUC = 0.7497) and "Crimes against Public Order" (AUC = 0.7180) show strong discrimination, while "Fraud and White-Collar Crimes" (AUC = 0.4330), "Violent Crimes" (AUC = 0.2984), and "Property Crimes" (AUC = 0.2276) indicate poor performance. The ROC curve helps identify these differences, highlighting areas for improvement. By comparing these AUC values to another model's perfor-

mance, we could determine whether an alternative approach might perform better for underperforming categories. This analysis shows the importance of the ROC curve and the AUC in guiding model selection and improving classification accuracy across all classes.

1.7 Results & Discussion

After testing multiple models, we opted for the Gradient Boosting model due to its impressive performance, achieving the highest accuracy of 80.87%. This model outperformed others, including Logistic Regression and Random Forest, in terms of predicting crime categories with the most precision. The strong accuracy of Gradient Boosting highlights its ability to handle complex, non-linear relationships within the data, which is crucial for making accurate predictions in a dynamic environment like crime forecasting. Its technique of sequentially improving predictions by focusing on difficult cases further solidified its effectiveness. Given its superior accuracy, we chose Gradient Boosting as the most reliable model to assist law enforcement in optimizing resource allocation and enhancing crime prevention strategies.

1.8 Potential Benefits

The potential benefits of our machine learning project on crime classification are substantial in enhancing public safety and optimizing law enforcement resources. By accurately predicting crime categories, law enforcement agencies can better allocate personnel and resources, ensuring that certain crimes with similar features are classified as such and are held to the same repercussions through category assignment. Additionally, precise crime classification helps in the legal and judicial processes by providing clearer insights into crime patterns, potentially leading to more informed litigation and policy development aimed at crime reduction.

1.9 Ethical Considerations

Using machine learning for classifying crimes within a dataset, particularly ones from San Francisco, raises significant ethical considerations. Crime data is often biased due to systemic inequalities in law enforcement practices, such as over-policing in marginalized communities or underreporting in wealthier areas. This bias can perpetuate harmful stereotypes or disproportionately target certain groups when incorporated into predictive models. Additionally, if historical data reflects discriminatory practices, the model may reinforce those patterns, leading to unfair treatment or unjust outcomes. Ensuring transparency, addressing bias, and carefully evaluating the societal impacts are critical to mitigating these ethical risks.

1.10 Future Work

This dataset highlights limitations that impact the modeling process. There is limited information about the suspect beyond the "Modus Operandi," while more data is available on the victim. While this reduces demographic bias related to the perpetrator, it shifts the focus to the victim, which may not significantly contribute to categorizing the crime itself. The model also heavily relies on the "Weapon Description" feature through NLP techniques, which introduces a bias toward weapon-related attributes.

Future applications of this work could explore the broader consequences of crime classification in real-world contexts. For instance, assigning a crime to the wrong category might lead to cascading errors or unintended outcomes, especially if misclassifications carry unequal consequences. It would also be essential to evaluate model performance using metrics beyond accuracy, such as precision, recall, or weighted penalties for incorrect classifications, to better understand the impact of false positives and negatives. Additional modeling techniques could include analyzing the like-

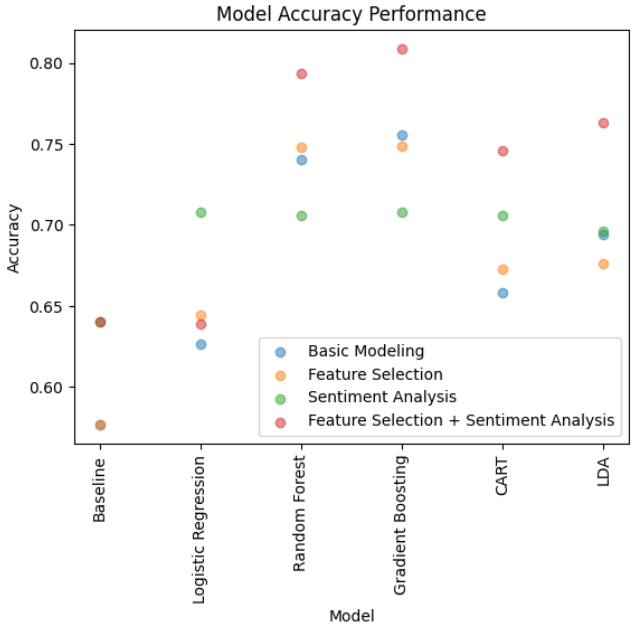

Figure 1: Accuracy Summary Across Models

lihood of each category assignment, predicting outcomes like jail time using linear regression, or incorporating geospatial data to make neighborhood-specific predictions.

1.11 Conclusion

Our multiclass crime category predictor demonstrates potential to enhance precision and efficiency in crime detection strategies. Gradient Boosting proved effective by iteratively improving and addressing misclassifications, helping to capture overlapping characteristics between crime types and significantly boosting accuracy.

Sentiment analysis enriched the model by extracting insights from text, enabling better classification of crimes with similar attributes but differing contexts or severity. Feature selection further improved performance by reducing overfitting, removing redundant features, and streamlining the model to focus on the most relevant variables, resulting in faster training and better generalization.

Overall, the project achieved its goals and laid the groundwork for future enhancements. With more granular data, the model could further support public safety initiatives and enable proactive, data-driven decision-making in crime prediction and prevention.

---

2 Appendix

2.1 Code

Dataset: https://www.kaggle.com/competitions/crime-cast-forecasting-crime-categories/data

Code URL: https://tinyurl.com/indeng142afinalproject

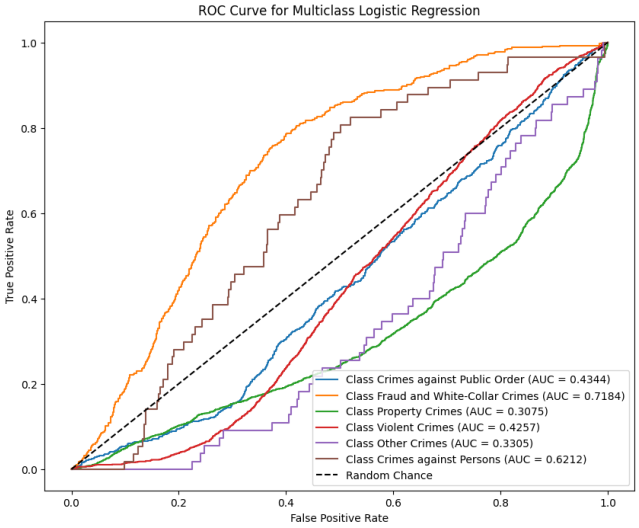

Figure 4: ROC Curve for Multiclass Logistic Regression

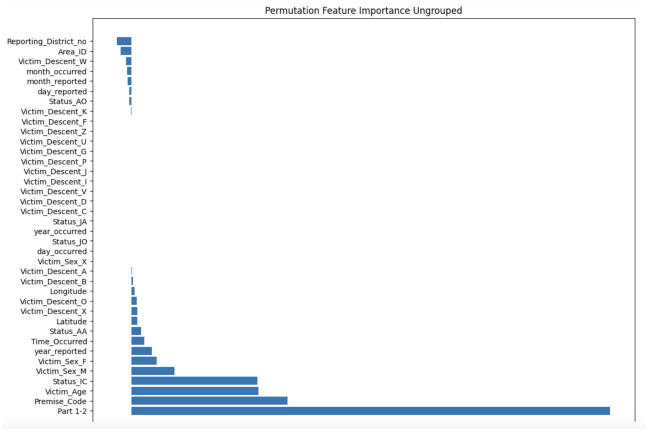

Figure 5: Random Forest Feature Importance

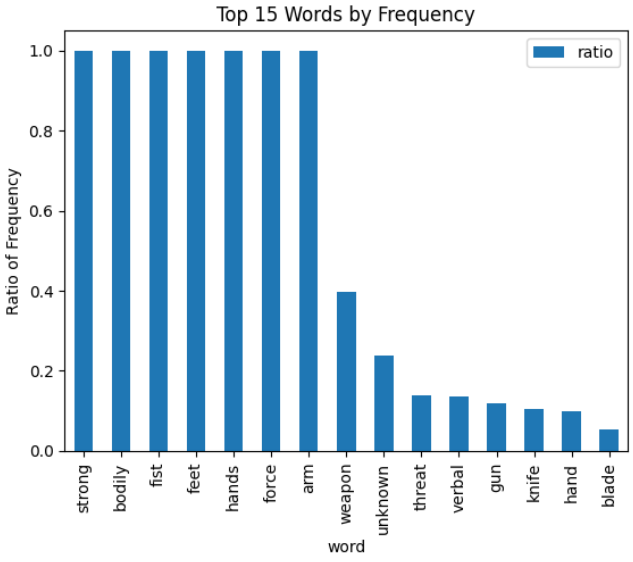

Figure 6: 15 most frequent words to appear across "Weapon_Descriptions"