Crime Data Analysis from 2020 to 2025

Students Kasra Sabertehrani and Arvin Lajani
Teacher Matteo Francia

Introduction

This project analyzes a real-world crime dataset provided by the **Los Angeles Police Department (LAPD)**, containing reported incidents from 2020 onward. Each record represents an individual crime report and captures a wide range of information about the event, including the time and date of occurrence, reporting time, type of crime committed, the location where it occurred, victim characteristics, the status of the investigation, location of the crime, and the weapon used by the criminal.

In the following, you can see the full list of features in this dataset, their data type, and their description.

Column Name	Type	Description
DR_NO	int64	This is the crime unique ID.
Date Rptd	object	The date and time in which the crime was reported.
DATE OCC	object	The date and time in which the crime occurred.
TIME OCC	int64	The time in which crime happened.
AREA	int64	The unique ID of police reporting area.
AREA NAME	object	The area name of the police district.
Rpt Dist No	int64	The sub division ID of the police reporting areas.
Part 1-2	int64	Crimes divided into two parts.
Crm Cd	int64	The unique ID of crime committed.
Crm Cd Desc	object	Description of the crime.
Mocodes	object	Acts or features combined with the criminal.
Vict Age	int64	Victim's age.
Vict Sex	int64	Victim's sex.
Vict Descent	object	Victim's race.
Premis Cd	float64	The unique ID of place where crimes happened.
Premis Desc	object	The whereabouts of the crime.
Weapon Used Cd	float64	The unique ID of weapon used by the criminal.
Weapon Desc	object	The weapon used by the criminal.
Status	object	The case status ID.
Crm Cd 1	float64	Other crimes that the criminal committed.
Crm Cd 2	float64	Other crimes that the criminal committed.
Crm Cd 3	float64	Other crimes that the criminal committed.
Crm Cd 4	float64	Other crimes that the criminal committed.
LOCATION	object	The location of the crime.
Cross Street	object	The cross street nearest to the crime.
LAT	float64	Coordinates of the crime (Latitude).
LON	float64	Coordinates of the crime (Longitude).

Table 1: Dataset column descriptions.

Related Work

(1) In recent years, the application of machine learning techniques to crime data has gained significant attention, particularly for understanding factors associated with arrest outcomes. One notable study by **Dmytro Iakubovsky**, titled *"What crime types are associated with an arrest? Analysis of arrest probabilities for 5 million Los Angeles crimes modeled with Machine Learning,"* analyzes a large dataset of over 5 million recorded crimes in Los Angeles since 2010. Iakubovsky focuses on predicting arrest outcomes by reframing the problem as a binary classification task, where "Adult Arrest" and "Juvenile Arrest" are treated as positive classes and all other statuses as negative.

Due to the highly imbalanced nature of the dataset (approximately 1:150 Arrest to Non-Arrest ratio), the study emphasizes data preprocessing techniques, such as removing duplicates, feature engineering (e.g., extracting year of occurrence, binning victim ages, and encoding categorical variables), and applying a stratified train-test split to maintain class distribution. For model training, a CatBoostClassifier was employed with balanced class weights inversely proportional to class frequencies. The model achieved a weighted ROC AUC score of approximately 0.88 on both the training and testing sets, indicating strong predictive performance without overfitting.

An important contribution of this work lies in its interpretability analysis using SHAP (SHapley Additive exPlanations) values. By interpreting the SHAP values, Iakubovsky identified key features influencing arrest probabilities. Among them, **criminal code description** (e.g., robbery, aggravated assault, shoplifting), **weapon description** (e.g., use of bodily force), and **year of occurrence** (with higher probabilities for older cases such as 2013–2014) were found to be the most influential. Additionally, victim demographics such as **younger age groups** (20 years and below), **Hispanic or Latino descent**, and **non-binary or unspecified genders** were associated with higher arrest probabilities. The type of premise (e.g., group homes, schools) and the geographical area (e.g., Southwest, 77th Street divisions) also emerged as important factors.

This work by Iakubovsky demonstrates the viability of using tree-based ensemble methods like CatBoost, especially when dealing with structured, imbalanced crime datasets. Furthermore, it highlights the importance of feature interpretability in criminal justice applications, ensuring that predictive models remain transparent and actionable. Techniques such as data balancing, feature selection, and model explainability, as presented in this study, are directly applicable to similar predictive modeling tasks in crime analysis, including our current project on predicting arrest probabilities in Los Angeles.

(2) Another relevant study is the project by **Escort Kwon**, titled *"EDA of Crimes Data in LA and Regression Models,"* which conducts an exploratory data analysis (EDA) on Los Angeles crime data and applies regression models for prediction tasks. Kwon systematically explores key variables such as crime types, locations, victim demographics, and temporal patterns to uncover underlying structures in the data. The study employs multiple regression techniques, including Ridge and Lasso regression, focusing particularly on predicting numeric features like the number of crimes or severity scores rather than arrest outcomes. Through careful feature selection and regularization, the models address multicollinearity issues and enhance generalization performance. While Kwon's work focuses more on regression rather than classification, it provides valuable insights into how crime patterns vary across areas, times, and victim characteristics. The feature engineering methods, visualization techniques, and handling of crime categorical variables are particularly useful for informing feature selection and modeling strategies in related crime prediction tasks, such as our current study on arrest probability prediction.

Proposed Method

The goal of this project was to predict the sex of crime victims (Male, Female, or No Victim) using a real-world crime dataset from the Los Angeles Police Department (LAPD) collected between 2020 and the present. To achieve this, we selected **Random Forest Classifier** as our primary modeling approach. Random Forest was chosen because it offers high flexibility, robustness to noisy or messy data, and strong baseline performance, especially in classification tasks involving structured data.

Data Preparation

Before building the model, careful preprocessing of the dataset was necessary to ensure data quality and reliability:

- **Datetime Handling**: The original dataset contained `DATE OCC` and `TIME OCC` columns, representing the date and time of the crime occurrence separately. We combined these into a single `DATETIME_OCC` column, which allowed us to capture temporal patterns more effectively.
- **Feature Selection**: Several columns were dropped, including administrative codes, redundant textual descriptions, and other attributes not relevant to the modeling goal. We focused on features such as crime type, area code, and date/time of occurrence.
- **Handling Missing Data**: For features considered critical (e.g., crime type, weapon description, area), rows with missing values were removed to avoid introducing noise. Features with a high proportion of missing entries were reviewed and filled to keep the main distribution.
- **Encoding**: One of the advantages of Random Forest is that it is largely insensitive to feature scaling and does not require complex encoding schemes for categorical variables. While some light cleaning and label encoding were performed for ease of implementation, extensive one-hot encoding or normalization was avoided. This minimized preprocessing time and maintained model interpretability.

The result was a clean, well-structured dataset containing key predictors that were ready for modeling.

Model Selection: Random Forest Classifier

We opted to use a **Random Forest Classifier** due to several key reasons:

- **Robustness to Outliers**: Tree-based methods like Random Forest are not significantly affected by extreme values in the data, making them suitable for real-world crime datasets where reporting inconsistencies are common.
- **Minimal Data Assumptions**: Unlike linear models, Random Forest does not assume linear relationships between features and outcomes, allowing it to model complex, non-linear interactions naturally.
- **Handling of Categorical Features**: Random Forest can work effectively with categorical data even if it is not heavily encoded, unlike methods like Logistic Regression that require numeric input.
- **Resistance to Overfitting**: By aggregating results from multiple decision trees (bagging), Random Forest reduces the risk of overfitting compared to single decision trees.
- **Scalability and Flexibility**: Random Forest can be easily scaled to larger datasets and, if needed, extended or replaced with more advanced ensemble methods such as XGBoost, depending on future needs.

At this stage, we prioritized building a **strong and stable baseline** model over complex hyperparameter tuning. The model was initialized with default parameters, though important parameters like the number of trees (`n_estimators`) and maximum tree depth (`max_depth`) could be optimized later if necessary.

Training and Evaluation

The dataset was split into a **training set (80%)** and a **testing set (20%)** to evaluate model performance on unseen data. This split helps ensure that the results are a realistic estimate of how the model would perform in real-world scenarios.

The following evaluation metrics were used:

Accuracy: The overall percentage of correct predictions.

Overall, this approach aimed to strike a balance between performance, simplicity, and flexibility. Should future experiments indicate the need for model improvements, Random Forest's structure allows easy hyperparameter tuning, feature engineering, or replacement by a different model without major changes to the preprocessing pipeline.

Results

After training the Random Forest model on the prepared dataset, we evaluated its performance using several metrics including overall accuracy, a confusion matrix, and a detailed classification report.

Model Performance

The model achieved a **high overall accuracy** on the test set, indicating that the Random Forest was able to learn meaningful patterns between the crime features and the victim's sex.

Accuracy: The model achieved an accuracy greater than 68%, confirming that Random Forest is a strong baseline for this task even without hyperparameter tuning.

General Observations

Overall, the Random Forest model provided a strong initial performance without requiring extensive data transformation or engineering. Its robustness to outliers, ability to handle categorical features, and natural feature ranking capabilities made it an ideal choice for this first phase of the project.

Further model improvements could focus on:

- Hyperparameter optimization (e.g., tuning number of trees, maximum depth),
- Balancing the "No Victim" victim class,
- Exploring alternative models like Gradient Boosting for performance comparison.

Conclusions

In this project, we tackled the problem of predicting the sex of crime victims using the Los Angeles Police Department crime dataset spanning from 2020 to the present. The approach involved careful data preparation, feature selection, and model training using a Random Forest Classifier.

Random Forest was specifically chosen for its robustness to noisy real-world data, its ability to handle mixed-type features without heavy preprocessing, and its natural resistance to overfitting through ensemble learning. The model demonstrated strong predictive performance, achieving high overall accuracy and generating meaningful insights into the relationship between crime characteristics and victim demographics.

Analysis of model outputs revealed that certain features, such as crime type, area, and weapon used, played a significant role in the classification task. However, some challenges were observed, notably the lower performance for the minority "No Victim" class, which was expected due to the inherent class imbalance in the dataset.

This work successfully established a strong baseline model that could serve as a foundation for future improvements. Potential next steps include:

- Hyperparameter tuning to further boost performance,
- Applying class balancing techniques to improve minority class prediction,
- Exploring alternative machine learning models such as XGBoost or LightGBM for comparison,
- Conducting deeper feature engineering, particularly on temporal variables and crime descriptions.

Overall, Random Forest proved to be an excellent starting point for the task, offering both strong performance and high flexibility. The methods and results outlined in this project provide a clear pathway for future work aimed at enhancing predictive accuracy and extracting deeper insights into crime victim patterns in Los Angeles.