CS-503: Project Report
Crime Category Prediction

Submitted By:

Sai Sahasra Surkanti

2021MEB1328

Supervised by :

Dr.Sudharshan Iyenger

1. Introduction

This project offers a unique opportunity to delve into a dataset that captures the nuances of crime incidents within a city. Each entry provides critical information about the time, location, victim demographics, and additional attributes that paint a detailed picture of each incident. The primary objective of this project is to use this dataset to build predictive models capable of forecasting the category of crime involved in each case. By leveraging machine learning techniques, I aim to identify patterns and insights within the data that could support law enforcement strategies, optimize resource allocation, and contribute to public safety improvements.

Enhanced Crime Prevention: Identifying high-risk areas and potential crime types can enable proactive strategies to deter criminal activities.

Resource Optimization: Predicting crime categories can guide the allocation of law enforcement resources, ensuring efficient deployment of personnel and equipment.

Improved Public Safety: By understanding the underlying factors contributing to specific crime types, targeted interventions can be implemented to enhance community safety.

Let's dive in and see what stories and patterns this dataset can reveal

2. Dataset Overview

Dataset Source:

State where the data was sourced from Kaggle, providing a comprehensive collection of crime incident reports within the city.

Data Structure:

The dataset contains **20,000** rows and **22** columns. Each row represents a unique crime incident, while the columns encompass various attributes relevant to understanding these incidents. Key features include:

Target Variable:**Data Issues**: Missing Values, Outliers, Imbalance:

2. Dataset and Data Preprocessing

2.1 Importing Libraries

To facilitate data processing and model development, the following libraries were used:

1. **NumPy and Pandas**: For numerical operations and data manipulation.
2. **Matplotlib and Seaborn:** For visual exploration of the dataset.
3. **Scikit-Learn:** Provided tools for data preprocessing, model selection, and evaluation.
4. **XGBoost and LightGBM:** Included as advanced machine learning models for classification.

2.2 Dataset Overview

The dataset provides features that include both categorical and numerical variables, which necessitates preprocessing and encoding. The initial exploration revealed the following key points:

1. **Data Size:** The dataset contains **20,000** rows and **22** columns. Each row represents a unique crime incident, while the columns encompass various attributes relevant to understanding these incidents.
    a. **Features:** Various features covering
    b. **Location:** Street address of the crime incident.
    c. **Cross_Street:** Cross street of the rounded address.
    d. **Latitude:** Latitude coordinates of the crime incident.
    e. **Longitude:** Longitude coordinates of the crime incident.
    f. **Date_Reported:** Date the incident was reported.
    g. **Date_Occurred:** Date the incident occurred.
    h. **Time_Occurred:** Time the incident occurred in 24-hour military time.
    i. **Area_ID:** LAPD's Geographic Area number.
    j. **Area_Name:** Name designation of the LAPD Geographic Area.
    k. **Reporting_District_no:** Reporting district number.
    l. **Part 1-2:** Crime classification.
    m. **Modus_Operandi:** Activities associated with the suspect.
    n. **Victim_Age:** Age of the victim.
    o. **Victim_Sex:** Gender of the victim.
    p. **Victim_Descent:** Descent code of the victim.
    q. **Premise_Code:** Premise code indicating the location of the crime.
    r. **Premise_Description:** Description of the premise code.
    s. **Weapon_Used_Code:** Weapon code indicating the type of weapon used.
    t. **Weapon_Description:** Description of the weapon code.
    u. **Status:** Status of the case.
    v. **Status_Description:** Description of the status code.
    w. **Crime_Category:** The category of the crime (Target Variable).

2. **Target Variable:** The target variable for this analysis is **Crime_Category**, which categorizes each incident into **7** predefined types of crime.

2.3 Data Exploration

1. **Exploratory Data Analysis (EDA)** is explored through descriptive statistics, frequency plots, histograms, count plots, correlation matrix, and box plots. This stage reveals insights into data distribution, feature relationships, and potential outliers.
    a. Checking data shape and information.
    b. Descriptive statistics.
    c. Visualizations like frequency plots, histograms, and count plots to analyze features like Crime Category, Victim Sex, Victim Descent, and Case Status.
    d. Correlation matrix to identify relationships between features.
    e. Box plots to detect potential outliers.
    f. Analysis of unique values in categorical features.

2.4 Data Preprocessing

1. Missing data is handled by filling null values with zeros.
2. Specific data cleaning steps are performed, like replacing zeros in 'Victim_Sex' and 'Victim_Descent' with 'Unknown' to maintain consistency during encoding.
3. Unnecessary categorical features are dropped to simplify the model.
4. Date and time features are engineered:
    a. 'Date_Reported' and 'Date_Occurred' are converted to datetime format.
    b. 'Report_delay' is calculated as the difference between the reported and occurred dates.
    c. Date and time features are split into individual components like day, month, year, hour, and minute.
    d. Outliers in 'Victim_Age' are handled by converting negative values to zero.
    e. The 'Cross_Street' feature is converted to a boolean indicator (1 for presence, 0 for absence).
    f. Numerical features are scaled using MinMaxScaler.
    g. Categorical features are encoded using OneHotEncoder.

2.5 Feature Engineering

1. The 'Modus_Operandi' feature is handled using MultiLabelBinarizer to capture multiple crime activities associated with the suspect.
2. Missing columns are identified and added to both train and test datasets with initial values of 0 to ensure consistency.
3. Column order is standardized between train and test sets.

4. The 'Crime_Category' column is dropped from the test set as it's the target variable.

4. Model Building

1. The train data is split into features (x) and target (y).
2. The data is further split into training and testing sets using train_test_split.
3. Various classification models are trained and evaluated, including:
    a. Ridge Classifier
    b. Logistic Regression
    c. Gradient Boosting Classifier
    d. Random Forest Classifier
    e. Decision Tree Classifier

Model performance is evaluated using accuracy score,precision,Recall,F1 score

5. Model Selection and Improvement

1. Chose the best-performing classification models and their corresponding accuracy scores.
2. Some models are further explored with techniques like **Principal Component Analysis (PCA) and hyperparameter tuning** using RandomizedSearchCV to potentially improve performance.

6. Model Performance

Accuracy

Light GBM	0.95775
XGBoost	0.95675
Bagging	0.95125
Gradient Boosting	0.95275
Extra Trees	0.95275

Precision

	Light GBM	XGBoost	Bagging	Gradient Boosting	Extra Trees
Crimes against Persons	0.63	0.68	0.57	0.62	0.82
Crimes against Public	0.84	0.84	0.8	0.82	0.82

Order					
Fraud and White-Collar Crimes	0.96	0.95	0.94	0.96	0.94
Other Crimes	0.5	0.54	0.53	0.45	0.83
Property Crimes	0.99	0.99	0.99	0.99	0.99
Violent Crimes	0.94	0.93	0.94	0.93	0.9

Recall

	Light GBM	XGBoost	Bagging	Gradient Boosting	Extra Trees
Crimes against Persons	0.66	0.61	0.59	0.59	0.34
Crimes against Public Order	0.89	0.88	0.89	0.85	0.82
Fraud and White-Collar Crimes	0.96	0.96	0.92	0.93	0.94
Other Crimes	0.16	0.18	0.21	0.24	0.83
Property Crimes	0.99	0.99	0.99	0.99	0.99
Violent Crimes	0.95	0.96	0.94	0.96	0.9

F1 score

	Light GBM	XGBoost	Bagging	Gradient Boosting	Extra Trees
Crimes against Persons	0.64	0.64	0.58	0.6	0.48
Crimes against Public Order	0.87	0.86	0.84	0.83	0.82
Fraud and White-Collar Crimes	0.96	0.95	0.93	0.94	0.94
Other Crimes	0.24	0.27	0.3	0.24	0.23
Property Crimes	0.99	0.99	0.99	0.99	0.99
Violent Crimes	0.94	0.94	0.94	0.96	0.94

We got to know that particularly Light GBM for its overall accuracy, and Extra Trees for its ability to identify less frequent crime types, for proactive policing, optimizing resource allocation, and implementing targeted interventions to enhance community safety. By identifying high-risk areas and potential crime types, law enforcement can deploy personnel and resources more effectively, deterring criminal activities and creating safer environments for all.