Machine Learning-Based Classification of Crime in Los Angeles

Aryan Patidar
Data Science Institute, Brown University
October 20, 2023

GitHub Repository: github.com/apatidar/data-1030-project

Introduction

Problem

- Los Angeles, widely considered one of the most dangerous cities in California, experiences a wide range of different crimes
- **Aim**: classification of the type of crime using various determinants of crime
- **Importance**: facilitates a better understanding of crime patterns in the city, informing resource allocation by law enforcement to improve public safety

Source of Dataset

- Kaggle; retrieved from the Open Data repository maintained by the City of Los Angeles
  - Collected from manually-typed crime reports, resulting in widespread missing values
  - Spans crime incidents from January 2010 to April 2023 (2 million data points)

EDA: Evaluation of Target Variable

Target Variable: crime (categorical)

- 144 distinct crime categories
- Most prevalent types occur roughly 200,000 times, which is a major portion of the dataset

```
count                          2827881
unique                             144
top        Battery - Simple Assault
freq                           246420
Name: Crime Code Desc, dtype: object
```

```
Crime Code Desc
Battery - Simple Assault                                  246420
Vehicle - Stolen                                          235962
Burglary From Vehicle                                     206332
Burglary                                                  190916
Theft Plain - Petty ($950 & Under)                        186135
                                                             ...
Till Tap - Attempt                                             4
Firearms Emergency Protective Order (Firearms Epo)             4
Train Wrecking                                                 2
Drunk Roll - Attempt                                           1
Firearms Temporary Restraining Order (Temp Firearms Ro)        1
Name: count, Length: 144, dtype: int64
```

EDA: Evaluation of Features

```
Fraction of Missing Values in Features:
Mocodes              0.115126
Vict Sex             0.102341
Vict Race            0.102359
Premise Code         0.000022
Premise Desc         0.000202
Weapon Used Code     0.661124
Weapon Desc          0.661124
Status               0.000001
Crime Code 1         0.000007
Crime Code 2         0.931666
Crime Code 3         0.998121
Crime Code 4         0.999944
Cross Street         0.833720
dtype: float64
Fraction of Points with Missing Values: 0.9999897449715882
Fraction of Features with Missing Values: 0.4642857142857143
```

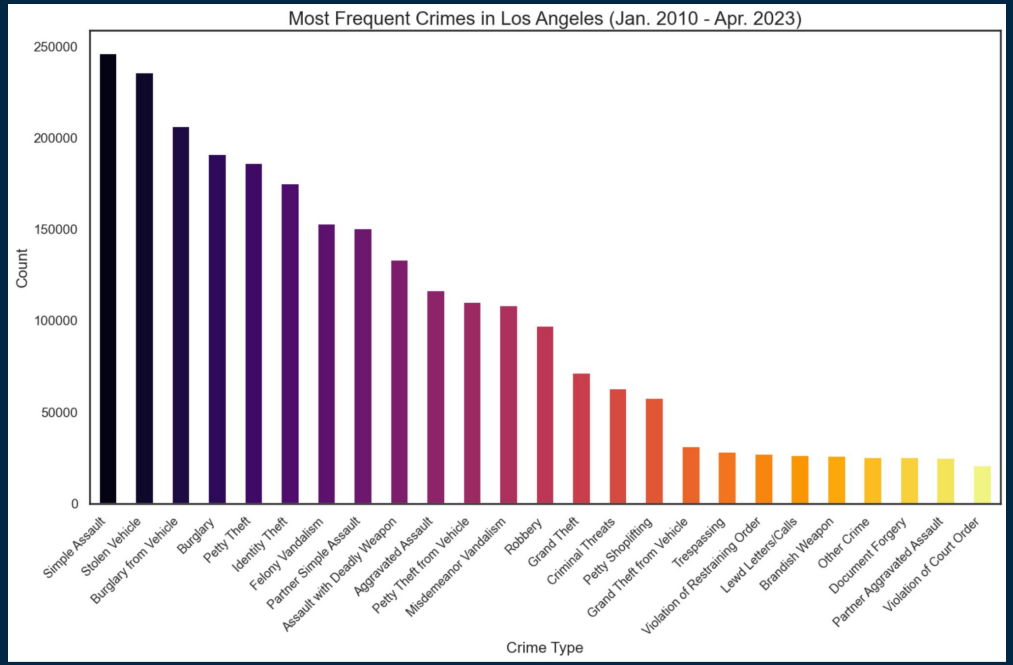
- **Shape**: 2827881 rows, 28 columns
- Many columns are not useful for classification and will likely be omitted in the final analysis
- Useful features to classify crime: 'Date', 'Time', 'Area', 'Victim Age', and 'Victim Sex'
    - Further EDA would help explore these trends

EDA: Inspection of Missing Values

- Several columns are mainly comprised of missing values
- Columns like ‘Crime Code 2’ are applied to provide additional detail and are thus rarely used - easiest to remove these features entirely
- Purpose of other columns, like ‘Part 1-2’, is unclear from the description - they will likely also have to be removed

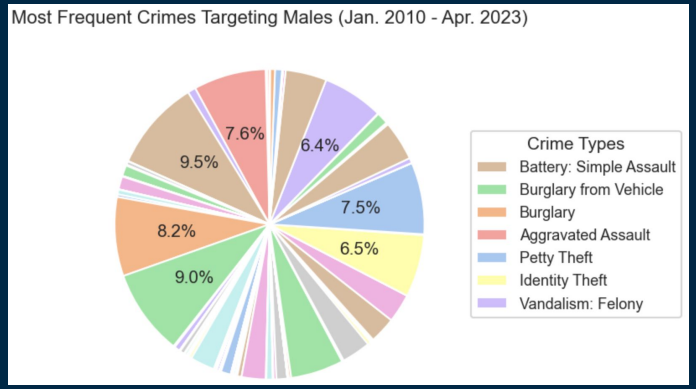
```
Record ID                0
Date Reported            0
Date Occurred            0
Time Occurred            0
Area                     0
Area Name                0
Report Dist No           0
Part 1-2                 0
Crime Code               0
Crime Code Desc          0
Mocodes             325563
Vict Age                 0
Vict Sex            289408
Vict Race           289460
Premise Code            61
Premise Desc           572
Weapon Used Code    1869579
Weapon Desc         1869580
Status                   3
Status Desc              0
Crime Code 1            19
Crime Code 2       2634640
Crime Code 3       2822566
Crime Code 4       2827722
Location                 0
Cross Street       2357662
Latitude                 0
Longitude                0
dtype: int64
```

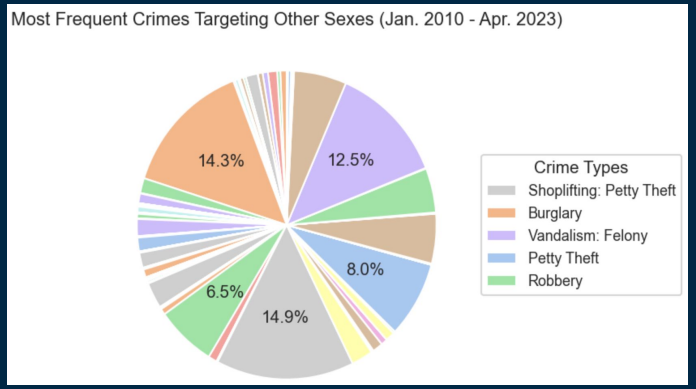
EDA: Visualization of Target Variable

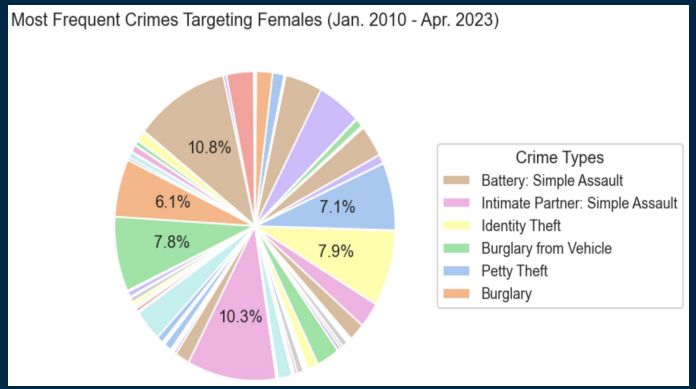


- Reveals the city's most common crimes
- Out of 144 crime types, roughly 20 are responsible for the majority of incidents
- This refines our analysis, making it easier to create a predictive model that can successfully identify common crimes before they occur
- Notably, the most frequent crimes encompass both serious offenses (‘aggravated assault’, ‘stolen vehicles’) and less serious ones (‘petty theft’)

EDA: Exploration of Victim Sex and Crime

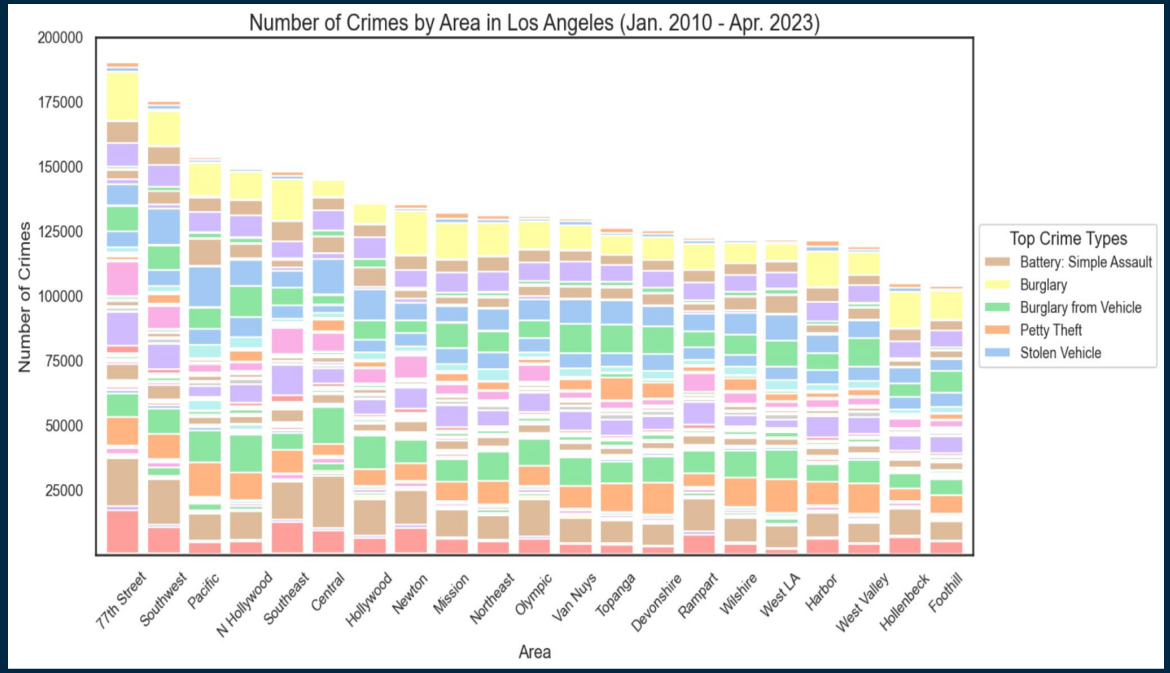






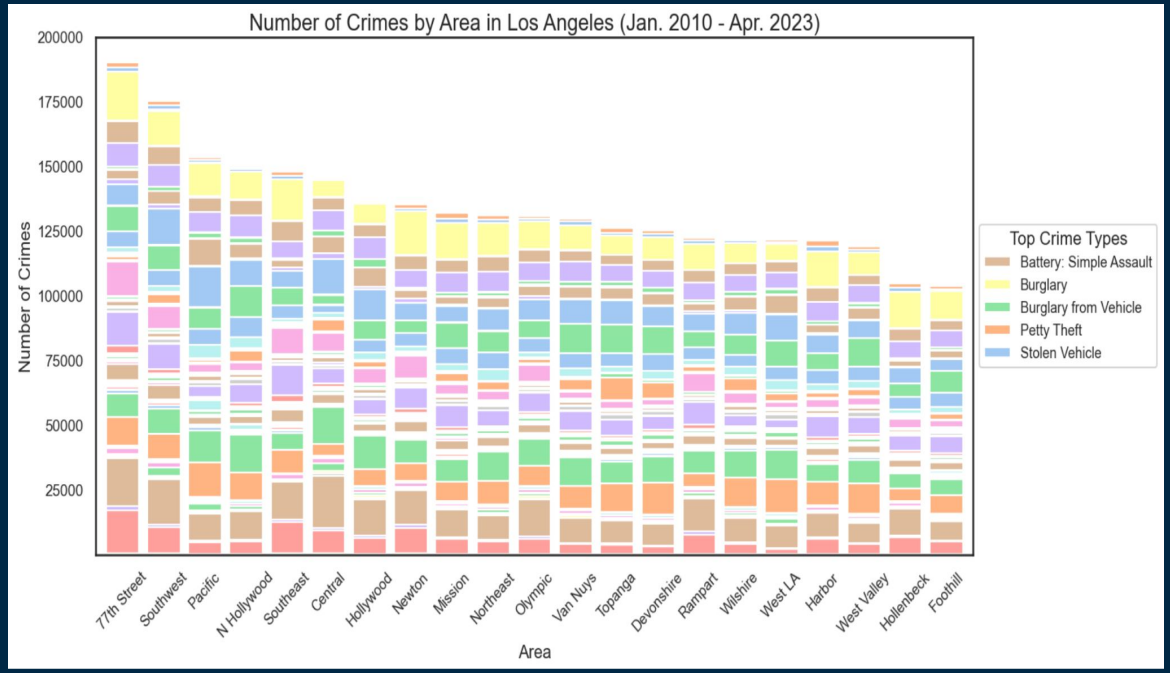
- Visualizes relationship between sex of the victim and type of crime to assess if sex is a strong predictor
- Common crimes like 'simple assault' and 'burglary' are prevalent across different sexes, indicating sex's limited predictive value
- Crimes like 'assault by an intimate partner' disproportionately affect females, warranting a focus by law enforcement

EDA: Analysis of Area and Crime



- Crime is distributed relatively evenly across different areas of Los Angeles
- High crime rates are observed in areas like 77th Street and Southwest, but even relatively safer regions like Hollenbeck and Foothill experience significant crime

EDA: Analysis of Area and Crime (cont.)



- Many crimes from the previous analysis reappear, prompting further investigation into geographical variation of crime
- Notably, 'stolen vehicle' only emerges as a common crime in this analysis, indicating an even distribution in its impact on both sexes

Data Splitting

```
print(X_train.shape)
print(y_train.shape)
print(X_val.shape)
print(X_test.shape)
```

```
(1696728, 27)
(1696728,)
(565576, 27)
(565577, 27)
```

- Split using "train_test_split" for higher efficiency due to the dataset's large size
  - Conserves memory by only training the model once (unlike k-fold cross-validation)
- **Split**: 60% training, 20% validation, and 20% testing
  - Balances the need for sufficient training data with the requirement for statistically significant validation and testing sets
- Since the shape of X_train, X_val, and X_test align with the shape of the original dataframe, the split is accurate
  - Each subset retains the original 27 features, minus the target variable

Data Preprocessing

- **Categorical Features**: 'Weapon Used Code'; applied one-hot encoding to represent the presence of a code as '1' (used a weapon) and no code as '0' (no weapon)
- **Ordinal Features**: 'Victim Sex' and 'Victim Race'; used ordinal encoding
- **Numerical Features**: 'Date Occurred,' 'Time Occurred,' 'Area,' 'Vict Age,' and 'Premise Code'; standardized using standard scaling
  - Date and time data were converted from datetime objects to integers
- After preprocessing, the feature count skyrocketed to 86, necessitating a careful evaluation before beginning the analysis

```
print(X_train_prep.shape)
print(X_val_prep.shape)
print(X_test_prep.shape)
```

```
(1696728, 86)
(565576, 86)
(565577, 86)
```

Appendix: Representation of Top 5 Crimes