LAPD®

LOS ANGELES POLICE DEPT.

TO PROTECT AND TO SERVE

HOW DATA CAN ENHANCE ARREST OUTCOMES

A CLASSIFICATION BASED MODEL

BY OLIVER BOHLER

INTRODUCTION

The Audience

The primary users of this project are the LAPD, the City of Los Angeles, and neighborhood watch groups. Each of these stakeholders has unique challenges and responsibilities, but all can benefit from a system that leverages data to identify trends and predict patterns in crime.

The LAPD, responsible for policing one of the largest jurisdictions in the United States, overseeing over 4 million residents across 469 square miles. Within 21 divisions, the department faces the daunting task of allocating limited resources efficiently. A data-driven solution would help the LAPD prioritize its efforts. Neighborhood watch groups, on the other hand, are smaller-scale stakeholders who can use this data to understand local trends and increase community safety efforts.

Why?

The LAPD is facing a severe shortage of officers, increasing retirements and declining recruitment rates. This resource gap has forced law enforcement to prioritize major crimes, leaving less attention for “smaller” crimes that directly impact residents, including burglary, robbery and assault. This project aims to fill that gap by examining vast amounts of reported crime data, helping to create a predictive model that empowers both law enforcement and community organizations to proactively address these issues.

The Data Source

The dataset used in this project contains every crime reported in Los Angeles between 2010 and 2023 and consists of two csv files. I accessed these files on data.gov. These records were originally handwritten and later digitized into a database, making them prone to human errors such as typos and inconsistencies. Thorough data cleaning and validation are required to ensure accuracy for analysis. Additionally, the LAPD is transitioning to a new system for recording crimes to align with modern standards and improve the efficiency and accuracy of data collection, which will ultimately enhance future crime analysis efforts.

THE DATA

The dataset represents detailed police reports, with over 3 million entries of various crimes reported between 2010 and 2023 across the 17 jurisdiction districts of Los Angeles. Real world data is an exciting field to discover yet it brings some:

Challenges

These records were originally handwritten and later manually entered into a database, leading to numerous missing values, discrepancies, and inconsistencies. Organizing this data into meaningful zones rather than jurisdictions proved to be a challenging yet necessary step to align with the project's goals.

With a dataset of this magnitude, it was critical to gain a clear understanding of each column to determine what could be dropped or combined into new features. Crime data is inherently detailed, yet not every column was relevant for this project.

Furthermore, the diversity of crime types and variability in data formatting across the years created additional hurdles, necessitating significant preprocessing to standardize and clean the entries. Additionally, certain columns, such as victim demographics, contained ambiguous or incomplete values, adding complexity to the cleaning process.

Another challenge involved balancing the class distribution in the data. Since crimes resulting in arrests are often a minority, ensuring a fair representation of all outcomes for analysis was an important consideration.

Opportunities

The dataset includes a vast variety of crime codes, each representing a specific type of offense, which adds to its complexity. While these factors made data cleaning and preprocessing a significant challenge, the dataset’s richness provides an opportunity to extract detailed insights about crimes directly impacting residents, such as burglary, robbery, and assault.

The consistent structure of the dataset, with identical column names and orders, made merging additional datasets straightforward, facilitating a broader analysis.

Beyond the crime codes and police-related factors, the dataset offers an opportunity to analyze victim demographics and explore how crime trends vary across different zones. Understanding these patterns could lead to targeted interventions and strategies tailored to the unique needs of specific communities.

DATA WRANGLING

I started by dropping duplicates, which is always a step I do in the beginning. After renaming and dropping columns the data frame now had a little under 3 million entries and 22 rows.

Organizing the Data frame

Zones: The dataset originally included 17 distinct jurisdiction areas, which needed to be consolidated into a more manageable structure. Using an interactive map, I visualized each jurisdiction and grouped them into four zones based on geographical and socio-economic similarities:

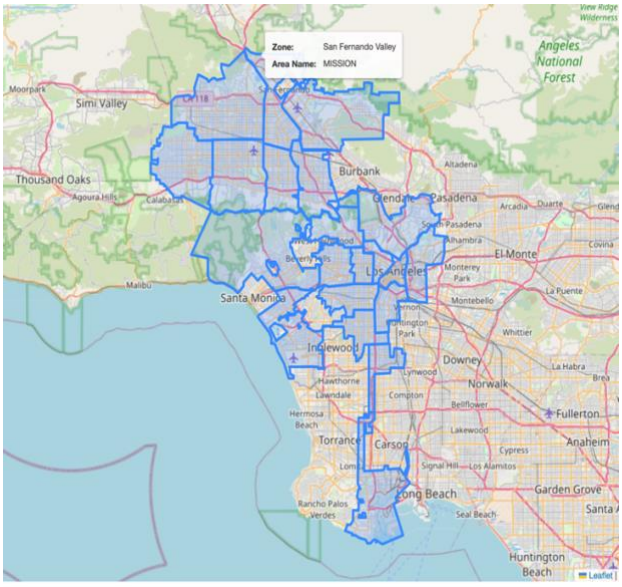
- **Westside & Wilshire**
- **Hollywood/Greater Downtown**
- **San Fernando Valley**
- **South LA**

This zonal organization allowed for more effective aggregation of data and a clearer understanding of crime trends across the city.

As shown in the table on the right, the data includes a wide variety of crime codes, with the top 20 most frequent crimes accounting for a significant portion of all reported incidents over the past 13 years.

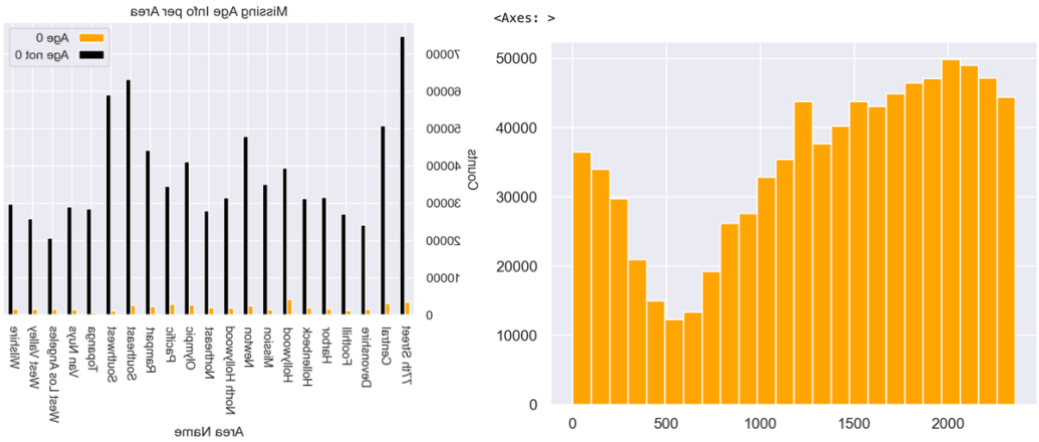
To make the dataset more meaningful for analysis, I grouped crimes into categories based on similar characteristics. These categories include:

- **Sexual Offenses**
- **Assault and Battery**
- **Vandalism and Arson**
- **Robbery and Burglary**
- **Theft (including car theft)**



	Crime Code	Crime Code Description	Counts
63	624	BATTERY - SIMPLE ASSAULT	263432
58	510	VEHICLE - STOLEN	258038
15	330	BURGLARY FROM VEHICLE	219608
13	310	BURGLARY	204164
38	440	THEFT PLAIN - PETTY ($950 & UNDER)	197293
26	354	THEFT OF IDENTITY	187401
81	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	165837
65	626	INTIMATE PARTNER - SIMPLE ASSAULT	160762
6	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	144013
28	420	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	124274
4	210	ROBBERY	115533
82	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	114346
17	341	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LI...	105764
127	930	CRIMINAL THREATS - NO WEAPON DISPLAYED	75298
40	442	SHOPLIFTING - PETTY THEFT ($950 & UNDER)	72770
16	331	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...	64948

I made the data-driven decision to exclude all other crime categories from the dataset, as the remaining categories provided sufficient coverage and relevance for this project's scope. Upon reflection, this step inadvertently introduced a **class imbalance** in the dataset, as these specific crime groups tend to have a higher proportion of **unsolved** cases compared to resolved ones.



The Time Occurred Colum: The recorded times were not in 24-hour military time format, which is a standard for temporal data analysis. To address this, I first created a histogram to visualize the distribution of the recorded times. This step allowed me to gain insights into any existing patterns or irregularities in the data. The conversion resulted in further temporal analysis, and examining trends across different periods of the day.

The Victim Age Colum: As anticipated, the dataset contains missing or erroneous values (e.g., age values recorded as 0) that could negatively impact the analysis by introducing bias. While **5.1%** missing data is a relatively small proportion, simply dropping these entries would result in a loss of valuable information, particularly given the size and diversity of the dataset.

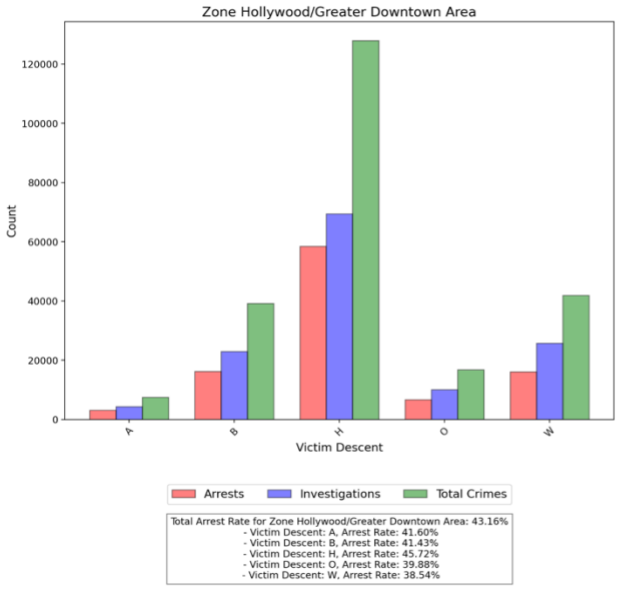
I evaluated two potential imputation strategies:

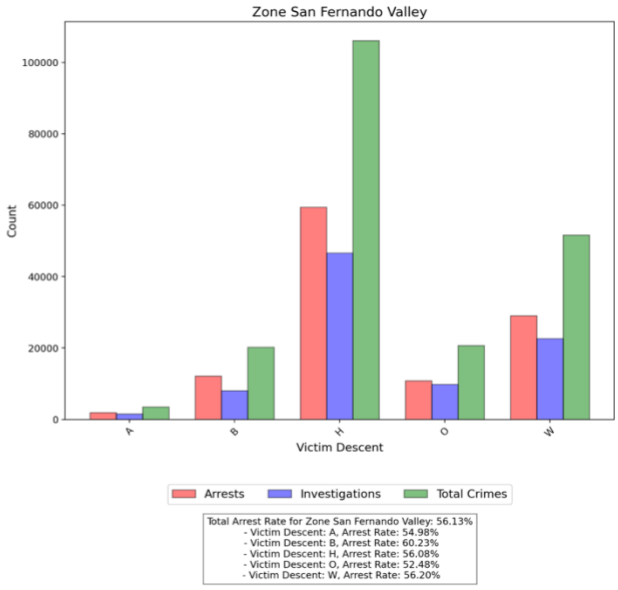
1. **Random Sampling**: Replacing the missing values with randomly generated numbers within a plausible range (18 to 99). This approach introduces variability but risks adding artificial noise to the data.
2. **Mean Imputation**: Calculating and replacing missing values with the mean age for each Area Name. This method preserves the general distribution of the data and accounts for geographic differences, as areas might have distinct demographic patterns.

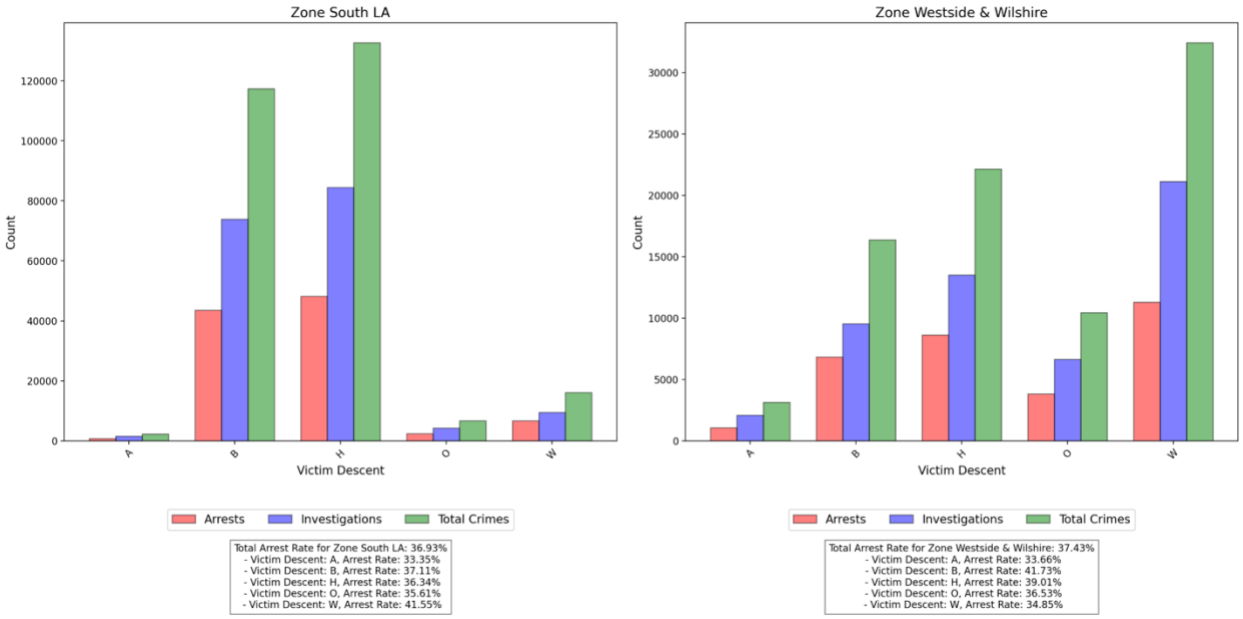
To ensure a more systematic and interpretable approach, I opted for mean imputation, using the Area Name column instead of the Zone column

EDA

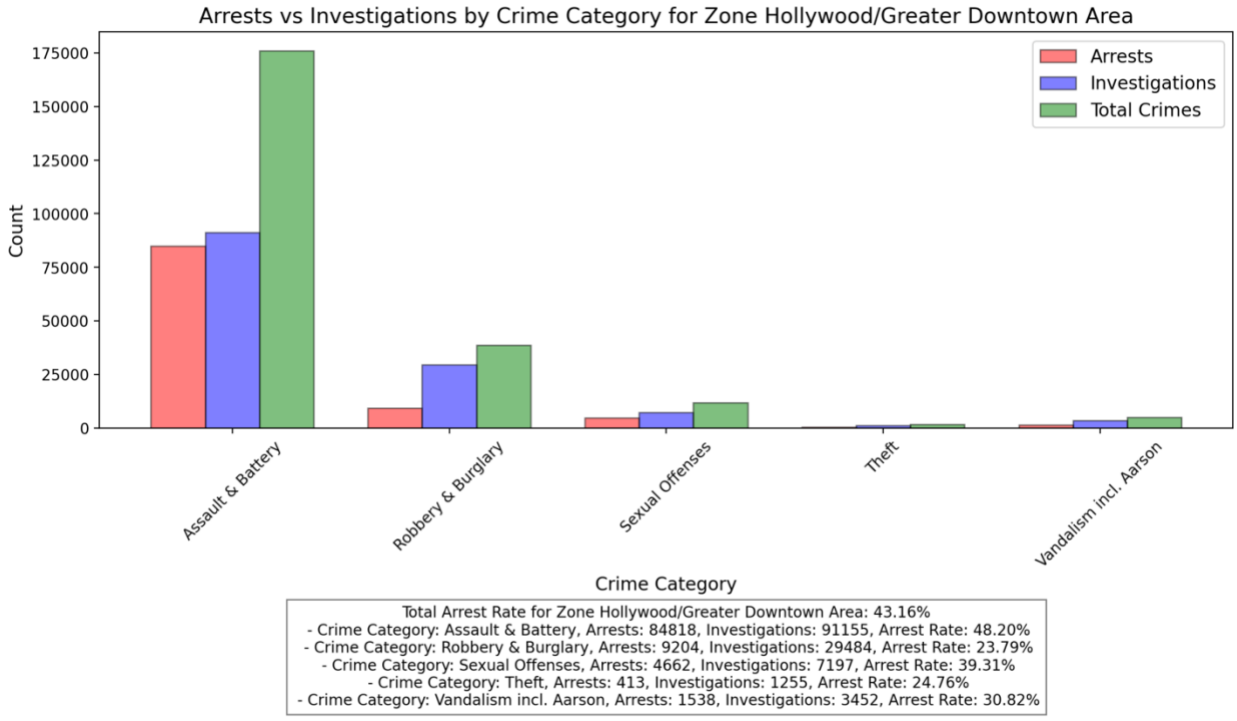
The main goal of the exploratory data analysis was to get better insights into the dataset with a special focus on the different zones and crime categories. While exploring this data I also provide the percentage of arrests based on the victim's race to answer the first Hypothesis: **Does the LAPD arrest rates differ based on the victim's ethnical background? (Please note that the LAPD has been always subject to these accusations by the public)**

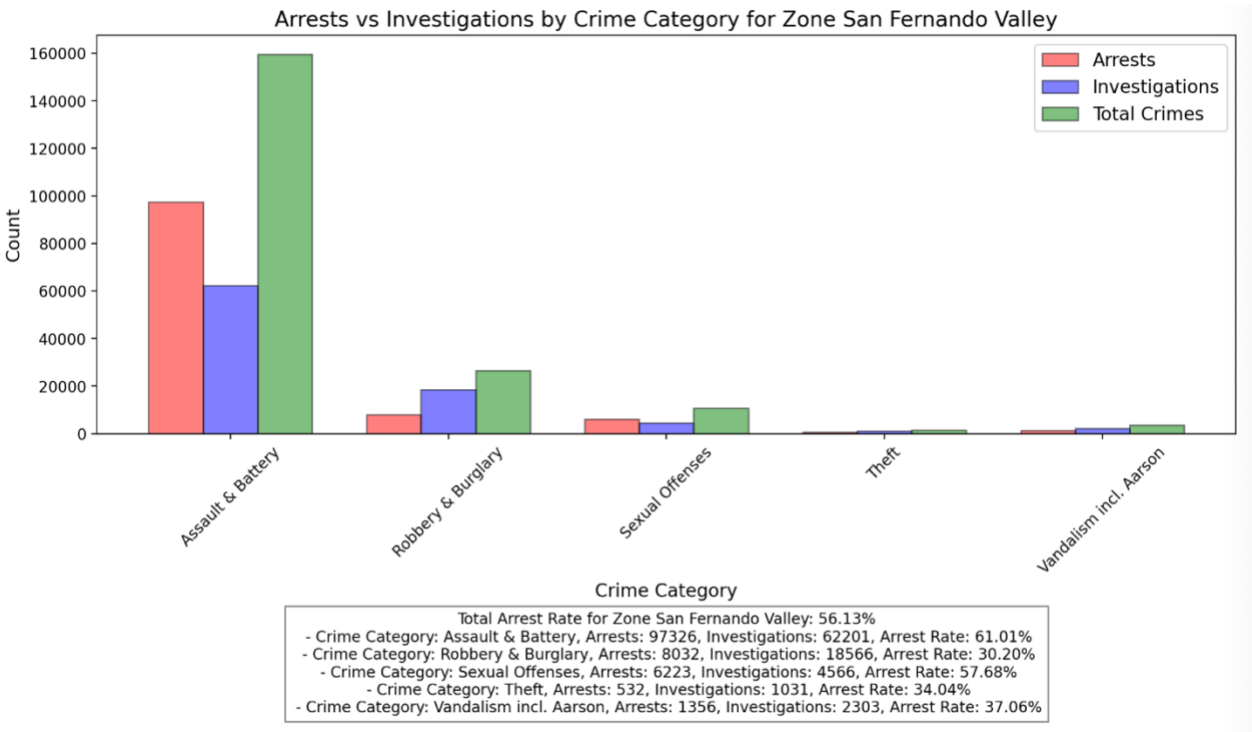




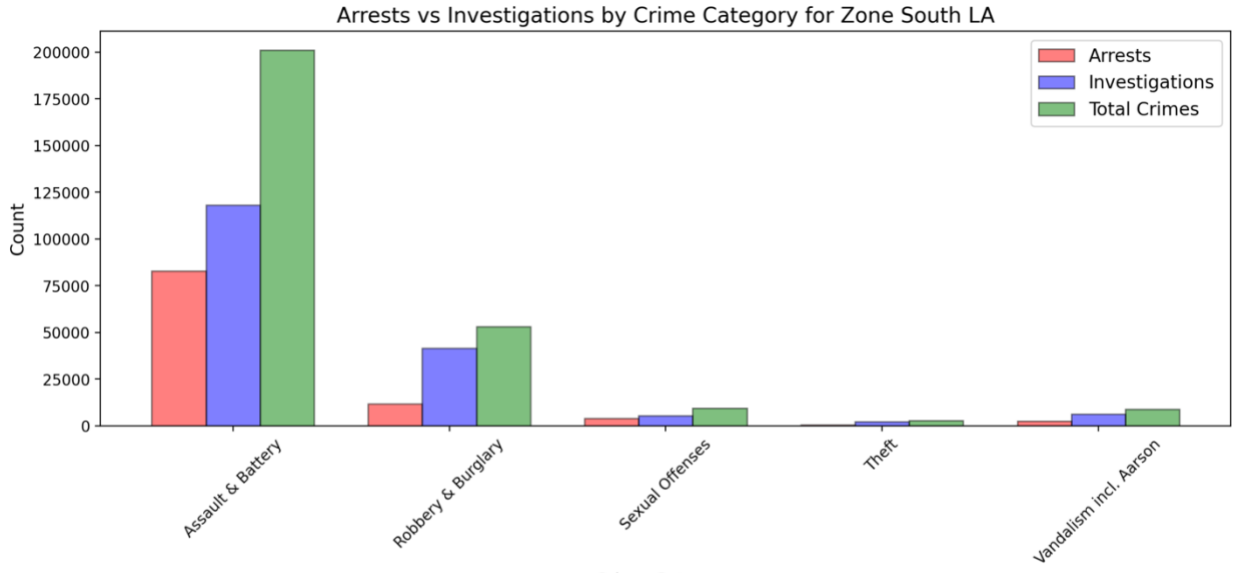


As we can see, there is no bias towards any ethnicity when it comes to getting a crime solved. However, I went on to investigate the total arrest rates based on each crime category as some of the low overall percentage is not a good sign for a balanced dataset.

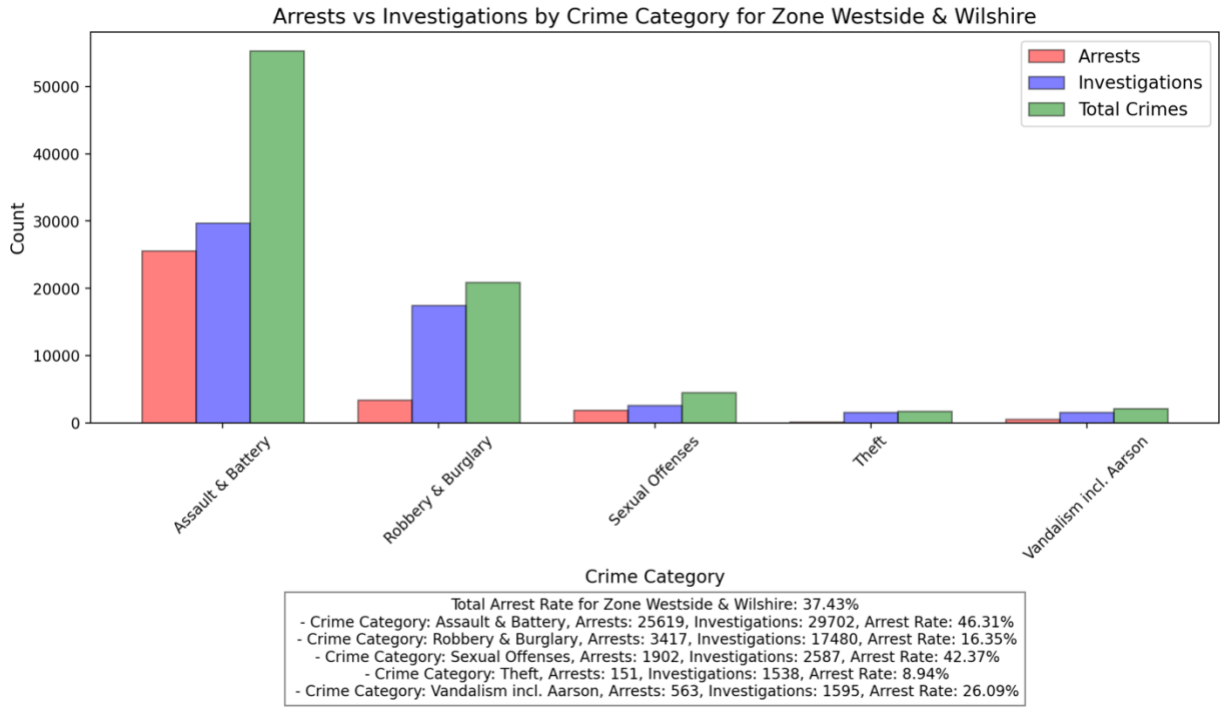




Total Arrest Rate for Zone San Fernando Valley: 56.13%
- Crime Category: Assault & Battery, Arrests: 97326, Investigations: 62201, Arrest Rate: 61.01%
- Crime Category: Robbery & Burglary, Arrests: 8032, Investigations: 18566, Arrest Rate: 30.20%
- Crime Category: Sexual Offenses, Arrests: 6223, Investigations: 4566, Arrest Rate: 57.68%
- Crime Category: Theft, Arrests: 532, Investigations: 1031, Arrest Rate: 34.04%
- Crime Category: Vandalism incl. Aarson, Arrests: 1356, Investigations: 2303, Arrest Rate: 37.06%



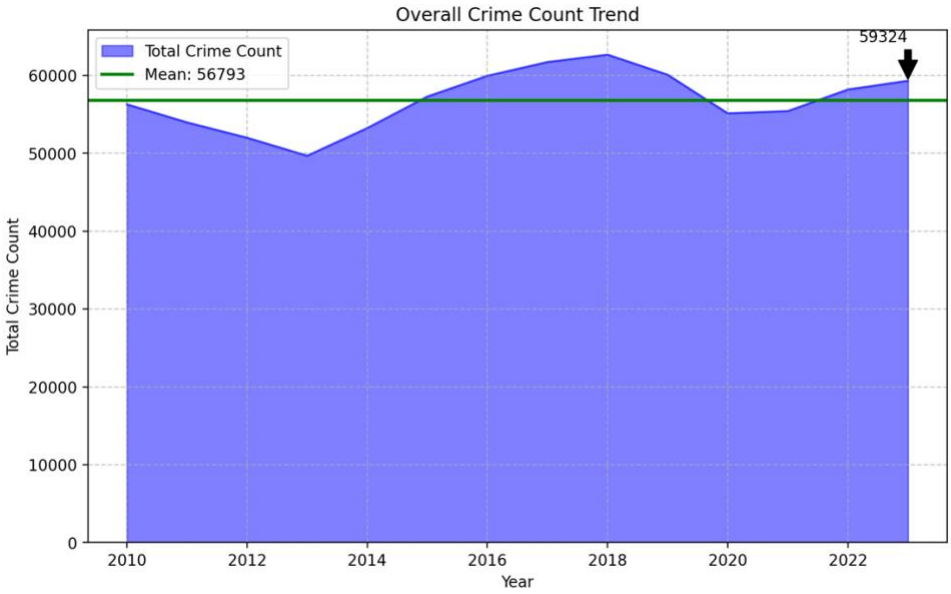
Total Arrest Rate for Zone South LA: 36.93%
- Crime Category: Assault & Battery, Arrests: 82930, Investigations: 118127, Arrest Rate: 41.25%
- Crime Category: Robbery & Burglary, Arrests: 11701, Investigations: 41446, Arrest Rate: 22.02%
- Crime Category: Sexual Offenses, Arrests: 3996, Investigations: 5505, Arrest Rate: 42.06%
- Crime Category: Theft, Arrests: 574, Investigations: 2184, Arrest Rate: 20.81%
- Crime Category: Vandalism incl. Aarson, Arrests: 2434, Investigations: 6324, Arrest Rate: 27.79%



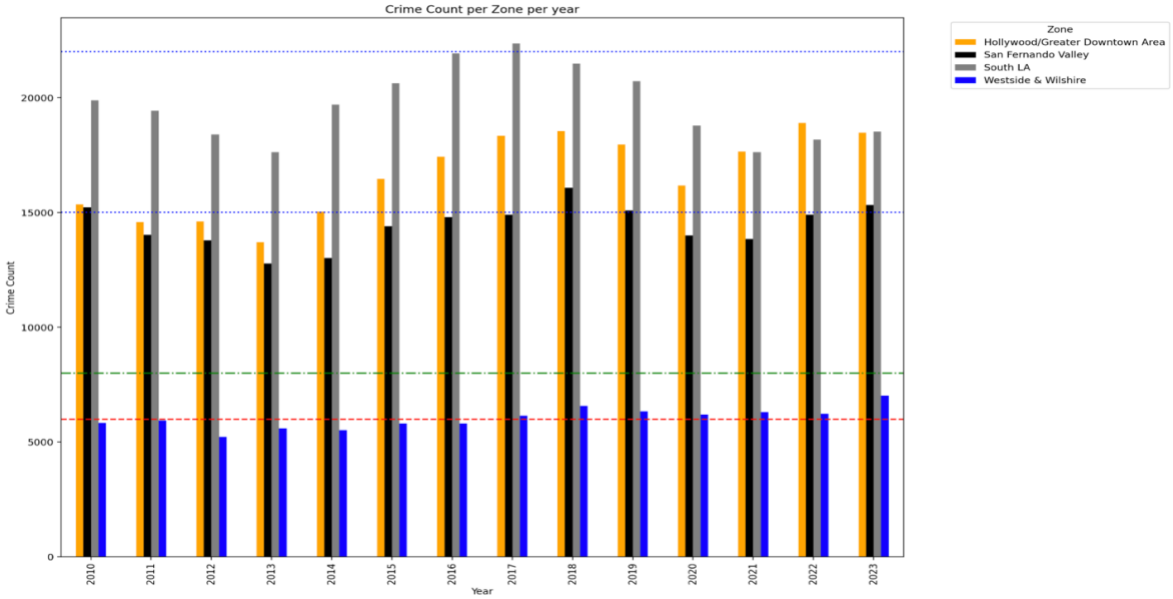
As I dove deeper into the crime categories, we saw arrest rates as low as 9 percent! While this shows the need for machine learning in the fight against crime, we have way more open cases than solved ones. This creates a severe class imbalance, meaning the machine will not have enough arrest cases to learn from.

Hypothesis 2: **Crime is at an all time high**

We all have heard controversial news about arrest rates. Some say crime is up from last year, then it's down 10% from the previous month, just to be at record numbers again in a year . In preparation to this project I did my fare share of research trying to find the answer to that question. This dataset finally gave a more conclusive one.



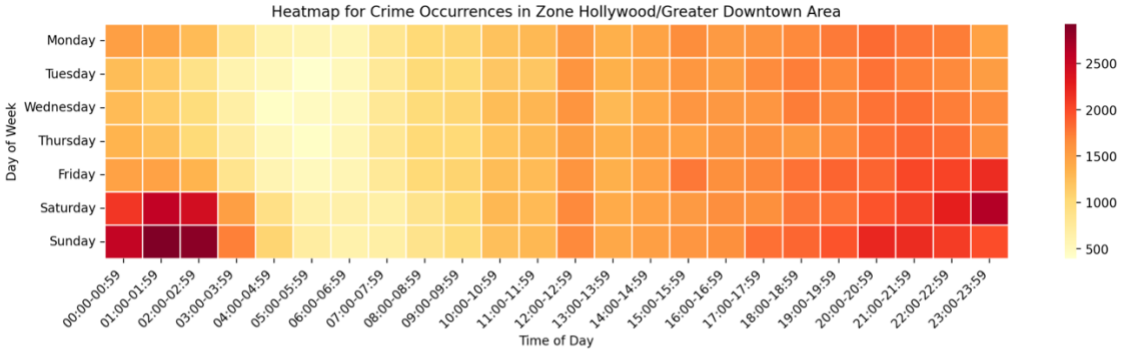
As the graph shows, crime (remember these are only the crimes listed previously) is not at an all time high but, it might get back there as we can see a clear tendency upwards from the years after the pandemic. Another interesting way to look at this is based on crimes per zone per year.

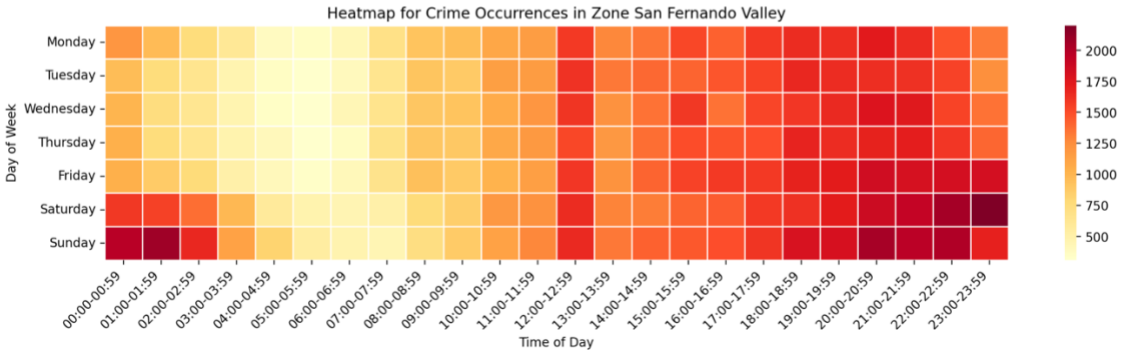


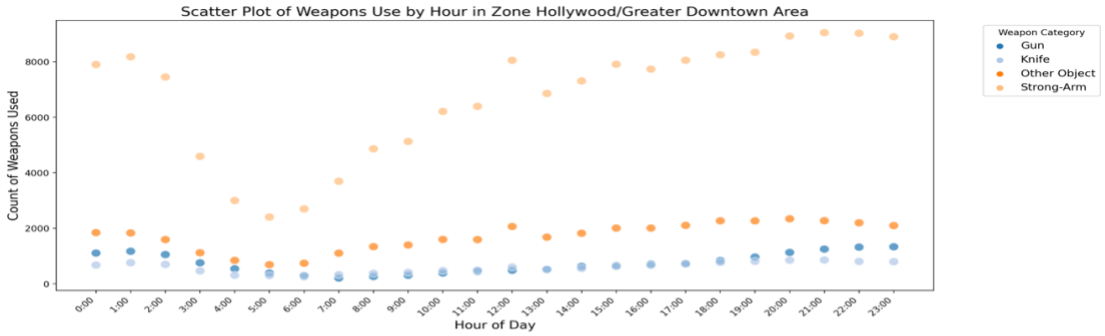
Hypothesis 3: **The use of guns and other weapons increases by night**

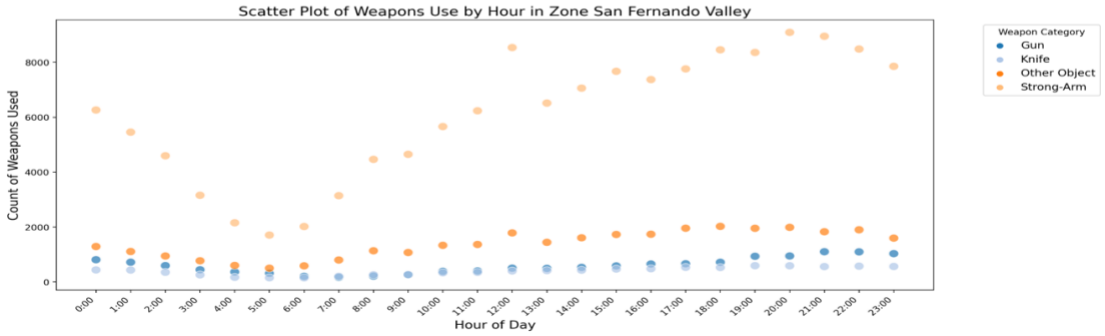
When it comes to officer safety, knowing the likelihood of a gun or other weapon involved can be the difference of live and death, not just for the officers but for potential victims as well. Below I have created heatmaps showing the occurrence of crimes per

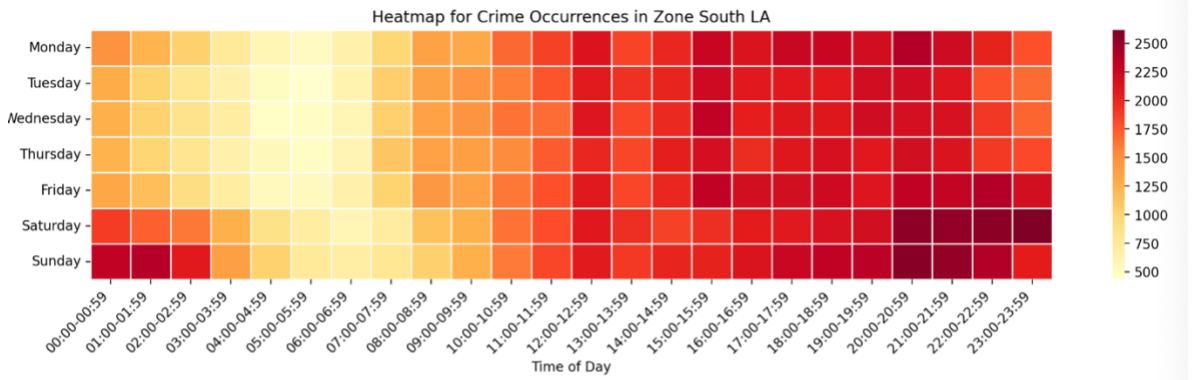
hour of the day. Further a scatterplot of weapon usage per hour was created and compared to the heatmap findings. This is clearly true.

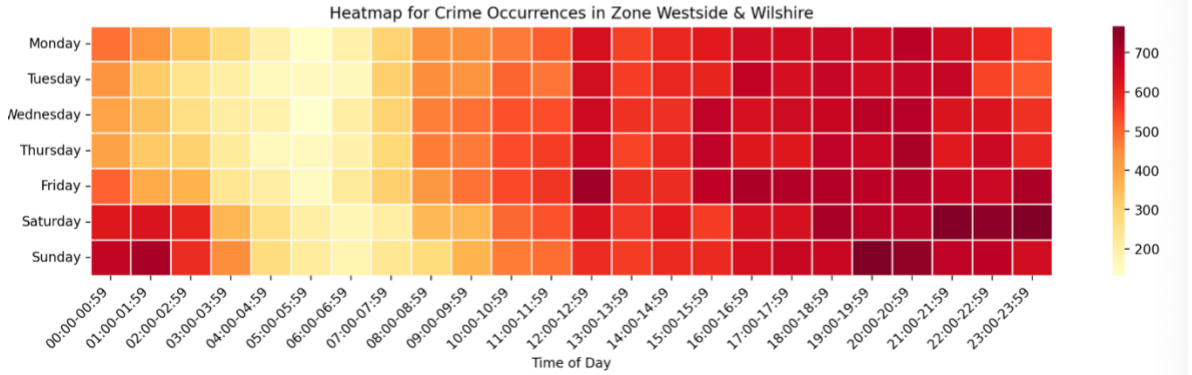


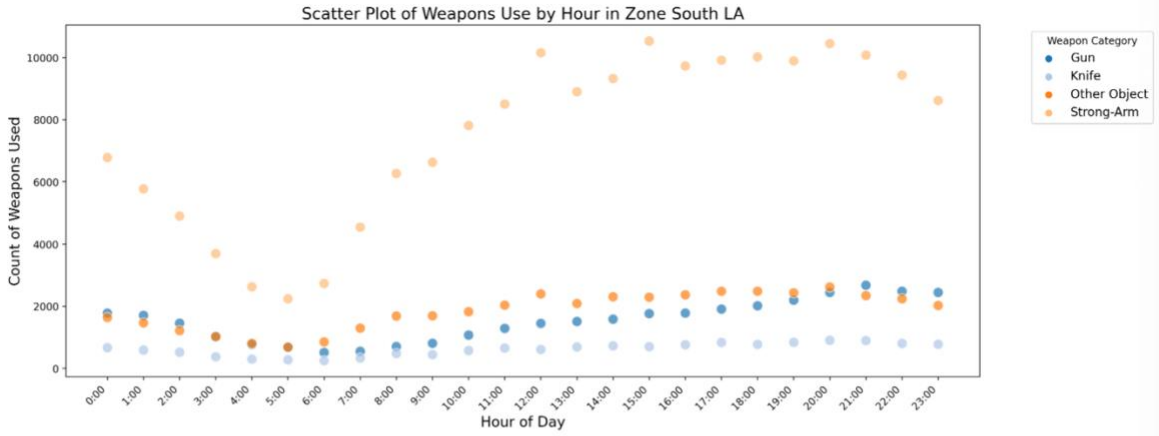


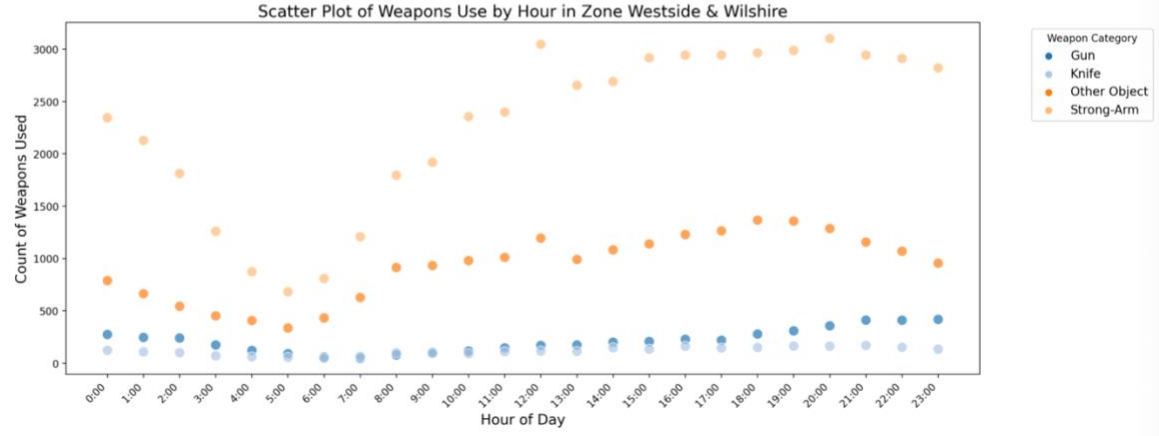





Heatmap for Crime Occurrences in Zone South LA
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
00:00-00:59
01:00-01:59
02:00-02:59
03:00-03:59
04:00-04:59
05:00-05:59
06:00-06:59
07:00-07:59
08:00-08:59
09:00-09:59
10:00-10:59
11:00-11:59
12:00-12:59
13:00-13:59
14:00-14:59
15:00-15:59
16:00-16:59
17:00-17:59
18:00-18:59
19:00-19:59
20:00-20:59
21:00-21:59
22:00-22:59
23:00-23:59
Time of Day
2500
2250
2000
1750
1500
1250
1000
750
500


Heatmap for Crime Occurrences in Zone Westside & Wilshire
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
00:00-00:59
01:00-01:59
02:00-02:59
03:00-03:59
04:00-04:59
05:00-05:59
06:00-06:59
07:00-07:59
08:00-08:59
09:00-09:59
10:00-10:59
11:00-11:59
12:00-12:59
13:00-13:59
14:00-14:59
15:00-15:59
16:00-16:59
17:00-17:59
18:00-18:59
19:00-19:59
20:00-20:59
21:00-21:59
22:00-22:59
23:00-23:59
Time of Day
700
600
500
400
300
200


Scatter Plot of Weapons Use by Hour in Zone South LA
Count of Weapons Used
0
2000
4000
6000
8000
10000
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Hour of Day
Weapon Category
Gun
Knife
Other Object
Strong-Arm


Scatter Plot of Weapons Use by Hour in Zone Westside & Wilshire
Count of Weapons Used
0
500
1000
1500
2000
2500
3000
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Hour of Day
Weapon Category
Gun
Knife
Other Object
Strong-Arm

MODEL SELECTION MACHINE LEARNING

With the initial findings from the EDA, I got a better understanding for crime patterns and arrest rates. But one problem persisted: class imbalance.

My main goal for the feature engineering here was to create a target variable through one hot encoding, using the ‘Status Description’ Column. This enabled me to have a binary outcome for the case ending up in an arrest or investigation continued. Further the data frame was reduced to include only 8 relevant columns. I created the train and test set and added balanced class weights to prevent overfitting.

The first algorithm used was the Random Forest Classifier. The initial findings showed that it was a good choice to add the weights, however, the model struggled when it came to predicting the actual arrests, 60%. No arrest was doing better at roughly 68%.

SMOTE was used to address further class imbalance by generating synthetic samples for the minority class. I decided not to use the method of Bootstrapping as this would have created more samples for both classes, yet I needed to focus more on the minority class.

The results only improved a little bit, but they improved. I used Optuna to implement Bayesian Optimization through a pipeline combined SMOTE and the Random Forest Model and got the most balanced results yet.

```
Test ROC AUC score: 0.7367
Accuracy: 0.68
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.71      0.72     89540
           1       0.62      0.63      0.63     66814

    accuracy                           0.68    156354
   macro avg       0.67      0.67      0.67    156354
weighted avg       0.68      0.68      0.68    156354

Confusion Matrix:
[[63502 26038]
 [24391 42423]]
```

At this point I decided to move on from the Random Forest Classifier and use XGBoost, which is more suited for imbalanced classes feature importance and model efficiency.

```
Accuracy: 0.68
ROC AUC score: 0.7341
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.75      0.73     89540
           1       0.63      0.59      0.61     66814

    accuracy                           0.68    156354
   macro avg       0.67      0.67      0.67    156354
weighted avg       0.68      0.68      0.68    156354

Confusion Matrix:
[[66796 22744]
 [27429 39385]]
```

The first run of this model already had very similar results as the modified Random Forest model. At this point I was ready to see if the Bayesian optimization and Optuna would also

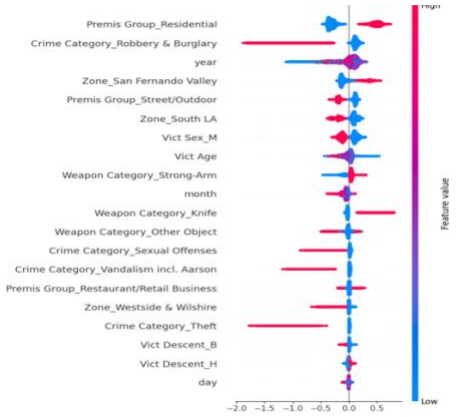
enhance the performance of this model. However, my personal computer and it's RAM capability met their limits and only 25 runs of this optimization were tested. While unfortunate for the further accuracy of this project, I could see slight improvements already after those few runs. Those were slight upticks in the ROC and AUC score as well as in the accuracy of the minority class (arrests) predictions.

Further models that did not exceed previous model classifications were Linear Regression, Gradient Boosting and KNN Nearest Neighbors. The last model I tested was the CatBoost Classifier which had the most balanced precision form all but a slightly lower ROC AUC curve. Therefore, the model XGBoost was chosen for modeling the results and recommendations.

```
Test ROC AUC score: 0.7309
Accuracy: 0.68
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.75      0.73     89540
           1       0.64      0.58      0.60     66814

    accuracy                           0.68    156354
   macro avg       0.67      0.67      0.67    156354
weighted avg       0.68      0.68      0.68    156354

Confusion Matrix:
[[67562 21978]
 [28351 38463]]
```

While the machine learning outcome here is moderate due to a high imbalance of the data set, I am confident that I am on the right track and that data and machine learning can help in the fight against crime and making Los Angeles safer. The perks of XGBoost is that it also creates a histogram of the most important features when it comes to arrests. The results show that the biggest factor for arrests is the location. It seems like crimes that are committed in residential buildings tend to have a higher chance of getting solved.

The plot below highlights the importance and effect of various features in predicting arrests or related outcomes. Premis Group_Residential and Crime Category_Robbery & Burglary are the most impactful features, indicating that crimes in residential areas and these specific categories strongly influence the model's predictions. Temporal features like year and month also contribute significantly. Geographic zones, especially San Fernando Valley and South LA, are important, indicating a strong spatial dimension to crime predictions. Demographic factors like Victim Sex (Male) and Victim Age, as well as weapon categories (e.g., strong-arm and knife), further refine the model's ability to capture crime characteristics.


Premis Group_Residential
Crime Category_Robbery & Burglary
year
Zone_San Fernando Valley
Premis Group_Street/Outdoor
Zone_South LA
Vict Sex_M
Vict Age
Weapon Category_Strong-Arm
month
Weapon Category_Knife
Weapon Category_Other Object
Crime Category_Sexual Offenses
Crime Category_Vandalism incl. Aarson
Premis Group_Restaurant/Retail Business
Zone_Westside & Wilshire
Crime Category_Theft
Vict Descent_B
Vict Descent_H
day
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Feature value
Low

CONCLUSION

The machine learning analysis of the LAPD dataset provided valuable insights into the prediction of arrests and related crime outcomes. Among the models evaluated, XGBoost emerged as the most promising, showcasing its strength in handling imbalanced data and uncovering meaningful patterns from complex feature relationships.

The feature importance analysis revealed that crimes in residential areas, robbery and burglary categories, and certain geographic zones (e.g., San Fernando Valley and South LA) were the strongest predictors of arrest outcomes. Temporal variables, such as year and month, also significantly influenced predictions, along with demographic factors like victim age and sex, and weapon types such as knives and strong-arm tactics. However, the hyperparameter tuning process—limited to only 25 trials—likely hindered the model's performance, leaving room for further optimization.

Despite these limitations, the XGBoost model demonstrated moderate predictive power, with an ROC AUC score reflecting a reasonable ability to differentiate between arrest and no-arrest cases. However, the results indicate that the project's scope could have been refined to achieve better outcomes. Specifically, narrowing the crimes down to violent and non-violent crimes may have reduced complexity and improved model performance. Moreover, the model's challenges with certain categories underscore the critical importance of thoughtful feature selection and precise problem framing in machine learning projects.

Further Improvements

Broader Crime Categorization: A more general approach, such as distinguishing between violent and non-violent crimes, could provide a simpler framework and allow for more targeted analysis. This broader perspective may better capture patterns and reduce noise introduced by granular categories.

Refined Crime Selection: Future iterations of this project could focus on specific crimes with higher arrest rates or consistent reporting standards, such as aggravated assault or theft, to improve model reliability and interpretability.

Incorporate Additional Contextual Features: Incorporating external datasets, such as socio-economic factors, police resource allocation, or historical crime trends, might provide richer contextual information for the model.