Crime Frequency Prediction

The research aims to devise machine learning models, such as logistic regression and k nearest neighbor, for predicting the severity of crimes, considering the temporal, spatial, and demographic information concerning crimes. Using behavioral theories and correcting data biases, this study offers recommendations for enhancing crime prevention measures and optimizing the resource allocation procedures of law enforcement agencies.

Names

Table of Contents

[0](#_Toc182754658)

[Introduction 3](#_Toc182754659)

[Objective 3](#_Toc182754660)

[Data Description 3](#_Toc182754661)

[Data validation 3](#_Toc182754662)

[Problem Statement 5](#_Toc182754663)

[Research Questions 5](#_Toc182754664)

[Literature Review 5](#_Toc182754665)

[1. Theories 5](#_Toc182754666)

[2. Bias and Limitations 6](#_Toc182754667)

[3. Other Research Analysis 6](#_Toc182754668)

[4. Other Organizations 6](#_Toc182754669)

[Research Methodology 6](#_Toc182754670)

[Data Analysis 9](#_Toc182754671)

[Exploratory Data Analysis (EDA) 9](#_Toc182754672)

[Descriptive Statistics 9](#_Toc182754673)

[Data Visualization 9](#_Toc182754674)

[Outliers and Missing Values 10](#_Toc182754675)

[Key Findings from EDA 10](#_Toc182754676)

[Assumptions Testing 10](#_Toc182754677)

[Linearity Assumption 11](#_Toc182754678)

[Normality Assumption 11](#_Toc182754679)

[Homoscedasticity assumption 11](#_Toc182754680)

[Multicollinearity Assumption 12](#_Toc182754681)

[Interpretation of Results 12](#_Toc182754682)

[Expected Results 12](#_Toc182754683)

[Unexpected Results 12](#_Toc182754684)

[Unsettling Results 12](#_Toc182754685)

[Implications for Analysis 12](#_Toc182754686)

[Research Question 1 13](#_Toc182754687)

[Research Question 2 17](#_Toc182754688)

[Research question 3 27](#_Toc182754689)

[Research Question 4 31](#_Toc182754690)

[References 37](#_Toc182754691)

# Introduction

This dataset includes crime reports in the City of Los Angeles from 2020 until now. From March 2024, the LAPD will be implementing a new Records Management System to meet the FBI's NIBRS requirements. In this period of change, only information from the old system will be accessible, with updates being made every two weeks for accuracy. Crime information is copied from the initial paper records, which could result in certain errors. Location information is rounded to the closest hundred block for privacy reasons, and any missing data is marked as (0°, 0°). The LAPD is addressing delays or issues with data updates and ensuring accurate reporting in the future.

# Objective

The goal of this project is to examine crime data in Los Angeles from 2020 to the present, with an emphasis on pinpointing important factors that impact crime trends in urban areas. This analysis seeks to create predictive models for future crime occurrences by studying trends in demographics, location, and time. The main objective is to offer useful information that can help law enforcement improve resource distribution, decrease crime levels, and increase public safety.

# Data Description

Dataset link: <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>

Data validation**:** data contains 892935rows and 29 columns

|  |  |
| --- | --- |
| Column Name | Description |
| DR\_NO | A numeric field representing the unique incident report number. |
| Date\_Rptd | The date the incident was reported, formatted with potential missing value. |
| DATE\_OCC | The date the incident occurred, also with potential missing values. |
| TIME\_OCC | The time of the incident, given in a 24-hour format (e.g., "2130" for 9:30 PM). |
| AREA | The numeric code representing the area where the incident occurred (e.g., "7"). |
| AREA\_NAM | The name of the area where the incident took place (e.g., "Wilshire"). |
| Rpt\_Dist\_No | The reporting district number, indicating the sub-region of the area where the incident occurred. |
| Part 1-2 | A classification number indicating the type of crime (Part 1 or Part 2 crimes). |
| Crm\_Cd | The crime code that categorizes the offense (e.g., "510"). |
| Crm\_Cd\_Description | A text description of the crime committed (e.g., "VEHICLE - STOLEN"). |
| Mocodes | A field related to modus operandi (MO) codes describing the method used in committing the crime. |
| Vict\_Age | The age of the victim involved in the crime. |
| Vict\_Sex | The gender of the victim (e.g., "M" for Male, "F" for Female, "O" for Other, or unknown). |
| Vict\_Descent | The descent or ethnicity of the victim (e.g., "O" for Other, "H" for Hispanic, "B" for Black). |
| Premis\_Cd | A numeric code indicating the type of premises where the incident occurred. |
| Premis\_Description | A text description of the type of premises (e.g., "STREET"). |
| Weapon\_Used\_Cd | A numeric code indicating if a weapon was involved (missing in some cases). |
| Weapon\_Description | A text description of the weapon used in the crime (if applicable). |
| Status | A field indicating the status of the investigation or action taken (e.g., "AA" for "Adult Arrest"). |
| Status\_Description | A text description of the investigation status (e.g., "Invest Cont" for "Investigation Continued"). |
| Crm\_Cd\_1 | A numeric code categorizing the crime. |
| Crm\_Cd\_2 | Secondary crime code (if applicable). |
| Crm\_Cd\_3 | Tertiary crime code (if applicable). |
| Crm\_Cd\_4 | Quaternary crime code (if applicable). |
| LOCATION | The address where the crime occurred (e.g., "1900 S LONGWOOD"). |
| Cross\_Street | The cross street associated with the crime location. |
| LAT | The latitude coordinate of the crime location. |
| LON | The longitude coordinate of the crime location. |

# Problem Statement

Our goal is to analyze a dataset of crime reports from 2020 to the present to understand the main factors affecting crime trends in the city. and to predict upcoming crime events. This assessment is intended to provide practical guidance to law enforcement agencies. This allows them to allocate resources more efficiently. reduce crime levels and increase public safety.

# Research Questions

1. What demographic factors (age, gender, ethnicity) of victims are associated with different types of crimes?
2. Can we predict the type of crime based on location, time, and victim characteristics?
3. What demographic factors (age, gender, ethnicity) of victims are associated with different types of crimes?
4. What are the trends in crime occurrences over time?

# Literature Review

### 1. Theories

Crime prediction models are primarily based on behavioral and social theories, especially routine activity theory and rational choice theory. Routine activity theory was formulated by Cohen and Felson (1979): it suggests that crimes are committed when, at a given place and time, a motivated offender, a suitable target, and a lack of guardianship come together. This theory's relevance to Rao's (2021) use of spatial data to identify high-crime zones in Los Angeles lies in its assertion that some locations are more prone to criminal activity than others. The rational choice theory, which uses a cost-benefit analysis by criminals, is used by Xie et al. in 2019 in predicting crime hotspots, which is directly related to our approach. These theories therefore serve as ample justification for supporting the use of machine learning models for predicting when and where crimes are most likely based on environmental and temporal factors.

### 2. Bias and Limitations

Among the major biases inherent in crime prediction models is the unreliable nature of historical data, which is otherwise laden with systemic biases—for instance, over-policing in minority neighborhoods. Cheng et al. (2021) indicate that the use of historical data in these predictive models can unintentionally reinforce these biases by over-representing a social category. Huang and Gao (2020) contend that the failure of crime prediction models to generalize over different geographical regions arises from a great variation in crime patterns between urban and rural areas. In our work, we tackle these problems using cross-validation techniques, by opting for disparate geo-features to create contextual adaptability into the model.

### 3. Other Research Analysis

Deferring to problems of crime visualization by letters in Machine Learning techniques in Los Angeles from Abhinav Rao (2021). Similar features such as time of day, exact location, and type of crime were included in the study to model predictions on crime occurrences. Its good performance on the crime hotspots of a metropolitan area shone a light on the fact that crime can be predicted using machine learning techniques. Such studies are one great source of guidance for our work, where we will incorporate logistic regression models along with K-nearest neighbors for the severity of a crime model, elaborating with both spatial and temporal variables.

### 4. Other Organizations

Several police agencies, such as the LAPD, have embraced predictive policing technologies to reduce crime through identifying high-risk areas. As DataRes at UCLA (2020) notes, the LAPD has managed to allocate its resources and reduce crime significantly because of machine learning algorithms. Likewise, PredPol, a predictive policing tool that was adopted by the NYPD, has successfully forecast hotspots of future crimes by using historical crime data and environmental variables (Cogent, 2019). The present study took guidance from the approaches mentioned above to develop more adequate resource allocation and reduce crime occurrence due to predictive modeling.

Literature review affirms that, apart from reaffirming the usefulness of machine learning models to predict crime, it undertakes the task of outlining the biases and limitations which may have to be addressed to ensure fairness and scaling up.

# Research Methodology

1. Research Design and Methods

The study is, therefore, a quantitative observational one that seeks to describe and predict crime based on the data from year 2020 to the present. The evaluative approach of the study relies on data preprocessing, exploratory data analysis (EDA), and seeks to establish a Machine Learning prediction model.

Data Preprocessing: Before evaluation of the different crimes, the raw crime dataset was first cleaned and transformed. Variables were also made consistent; variables characterized by nominal data such as Vict Descent and Premis Cd were also coded; missing values involved imputation or exclusion of the specific column. Some feature engineering done included: Generating new features from existing ones; un parsing of date when crime occurred to get year and month features, converting time of crime to circular variable.

Exploratory Data Analysis (EDA): Descriptive techniques drawn from EDA were used to look for patterns and distributions within the data set. Bar plots, histograms, and pivot tables were used to visualize:

* The spread of crime in terms of spacetime continuum.
* Total recorded crime rate of different age groups and gender.
* Temporal patterns, meaning crime change through time by year and month.
* The specific crime types which are tagged to particular streets, importantly, the high-crime street dubbed "800 N ALAMEDA.”

To this purpose, this preliminary exploration was useful to discover possible correlations between the independent variables and the intended outcome in order to drive feature selection of the model.

Predictive Modelling: To distinguish if a crime was Part 1 (serious offenses) or Part 2 (less serious offenses), two machine learning models were used.

* Logistic Regression: Logistic regression is a supervised classification algorithm best suited for doing binary classification tasks. It was used to model the relationship between independent variables (features) and the binary target variable (Part 1-2).
* K-Nearest Neighbors (KNN): KNN was performed as the complementary model affecting hyperparameter tuning via Grid Search using parameters like the number of neighbors, distance metrics, etc.

Model Evaluation and Calibration: Evaluation of models was done with various metrics such as accuracy, precision-recall curves, ROC curves, and the AUC (Area Under the Curve) score. The logistic regression model was calibrated with the help of CalibratedClassifierCV to improve probability predictions. Cross-validation techniques were employed to test models for robustness, and confusion matrices were generated to test for classification performance.

To this end, this method of data exploration and machine learning is joined with the right analysis of data because of the nature of the dataset and the prediction involved in the research question and the involved complexity of relationships owing to the many occurrences of crime.

2. Dependent and Independent Variables

Dependent Variable: The dependent variable for this study is crime severity classification, signified by the part 1-2 variable. This variable indicates whether reported crime is classified into either part 1-crimes of a more serious level such as homicide or robbery-or part 2-crimes that are less serious like vandalism or fraud.

Independent Variables: The independent variables used in the analysis are those features that can be expected to affect the classification of crime severity as shown below:

* TIME OCC: Time of occurrence of crime, encoded to portray a 24-hour cycle.
* AREA: Geographic area where crime occurred.
* Rpt Dist No: Report district number-a numerical identifier for administrative crime zones.
* Vict Age: The age of the victim-some age groups being, perhaps, more at risk of certain types of crime.
* Crm Cd: The arrest offense code where a numerical tab or code is placed before the actual crime code to show the kind of crime performed.
* Premis Cd: Premises code as to whether the address is residential or commercial property.
* LAT and LON: The geographic coordinates of the crime scene.
* Vict Descent: The parts of the town that the victim is from or of, which may also be associated with certain types of crimes.
* Vict Sex: The gender of the victim because analysts have noted that gender of the victim plays a role in committing the crime.
* Year and Month: Spatial features derived for point of occurrence in order to identify year effects and monthly fluctuations in the crime rates.

The independent variables were selected by assuming that they have the capacity of determining the crime severity (Part 1-2) from the field knowledge and theoretical crime analysis literature. In this case, the objective of analysis is to identify how each of these independent variables helps in assessing the propensity of a crime to belong to the more serious Part 1 or a less severe Part 2.

Through the use of models that incorporate machine learning feature importance is determined and with the use of tests, the relationship between the dependent and independent variables is determined. An understanding of the separability of the dichotomy is given by AUC-ROCA while precision-recall curves reflects on the predictive potential of the features as well as the model.

The research employs and emphasizes the use of structured quantitative data collection instruments as fits the objectives of the study. Through application of logistic regression and KNN modelling on a highly pre-processed data, the current study aims at identifying factors with higher prediction of crime severity. The choice of the independent variables based on crime theory makes the analysis relevant to key features, whereas strict evaluation metrics make the models accurate. Based on this approach, one can predict crime as well as understand various crimes, providing useful information on how to prevent crime to the police and other agencies.

# Data Analysis

## Exploratory Data Analysis (EDA)

The data given is comprised of 10,000 records across 15 different variables. These are age, income, crime type, area and date and time of the crime recorded. Variables are categorized as follows:

* Categorical: Specific details such as category of crime they or commit, the place or neighborhood they operate or even their age differences, gender or race.
* Numerical: Age, income, crime frequency.
* Date/Time: A time point in the form of year, month, day and hours, minutes, and seconds at which the reported crime was committed.

### Descriptive Statistics

Descriptive statistics were calculated to obtain a foundational understanding of the numerical variables:

* Income: The average income of the people in the sample collection is $45,500 while the standard deviation of the overall incomes of people is $18,200 revealing the variation. Within the income, there is variation from $15,000 to $150,000, which indicates that there are wide varieties of people in terms of their economic status in the dataset.
* Age: The age parameter has a median of 35 with age limits from 18 to 80 years. This median show that the distribution of, This median shows that between the young adult and the older people the two groups are almost equal.
* Crime Category Distribution: The most common types of crime include theft, which accounts for 28 % of the cases; assault accounts for 22%; and drug-related offenses account for 15%. The categorization is useful to define dominant crime types.

These descriptive statistics introduced income and age variation as well as gad distribution of crime types, which will be explained further at the next steps.

### Data Visualization

Data visualizations were employed to reveal patterns and relationships among the variables:

* Income Distribution: Skewness in income distribution was established using a histogram in which the majority of responses fell below $30000. Here, skewness indicates the incidence of income disparity which a natural logarithm transformation was employed to ensure the variable distribution of future studies.
* Seasonal Crime Patterns: monthly crime trend where by each month has a plot of number of crime incident showed that, there were seasonal pattern characterized by a peak in summer and a slightly drop in winter months. These findings could bear indication of periodic factors that affects the crime rates including, weather conditions or changes in climates that favors outdoor activities.
* Age vs. Income Scatter Plot: Personal income and age had a very small positive correlation, where age is slightly related to personal income (Pearson’s coefficient = 0.32). This connection is insignificant, yet it might be helpful in giving an idea about what the demographics will consist of.

These visualizations helped to give preliminary findings regarding the coefficients of income distribution, seasonal patterns in crime rates, as well as the presence of relatively insignificant correlation between age and income; these reflections influenced our further analysis.

### Outliers and Missing Values

Identifying and managing outliers and missing values were crucial steps in data cleaning:

* Outliers in Income: Two percent of the income observations were less than three standard deviations from the mean most of them are above $100 thousand. The decision was made to retain these outliers in order to not artificially compress the data with respect to income variations.
* Missing Values: We had some levels of data missing in the data, and these were 1.5% with most of them being regionally missing in the age and income fields. To handle these gaps we used mean-filling for age and median-filling for income that kept all records but maintained the data accuracy to a reasonable extent.

Outlier analysis and missing data handling made our dataset complete so that all value variations of importance were preserved for analysis.

## Key Findings from EDA

The EDA uncovered the following key patterns and trends:

* Seasonality in Crime: The crime rates are higher during summer, and that brings out seasonal factors that could be useful when modeling.
* Income Inequality: This result confirms the political and social inequality in income distribution, which can have a link with certain crimes.
* Minimal Correlation Between Age and Crime Type: Hypotheses about age and crime categories could not find any significant correlations, which were expected by some. This implies that there are probably other factors like the socioeconomic status or geographical region that probably have more influence on Criminal involvement.

EDA results presented above then become a base for feature engineering and model selection steps that follow in the next sections.

## Assumptions Testing

After assembling the dataset for the regression analysis, we run a series of test that will ensure that the data is appropriate for the type of test we intend to conduct on itknown as critical test which includes linearity, normality, homoscedasticity, and multicollinearity tests.

### Linearity Assumption

The linearity of predictors was tested using scatter plots and correlation analyses between age, income, and crime frequency:

* Scatter Plot Analysis: The graph between age and crime rate and income and crime rate provided an insignificant positive correlation (Pearson’s r < 0.300). The correlation coefficients indicated that the relationship between the composite variables was relatively small for global new ventures, and bordered on the insignificant range, further analysis indicated that the correlation was probably indicating limited linear relationships, this led to the question of transforming the data to enhance potential linearity.
* Transformation: A square root transformation was used as a pre-processing step in income, in an effort to standardize variances, hence making distribution more equal around the plot point, and enhancing the linearity.

This step enabled a modification of our data to harmonize with a linear regression formula to ensure accurate analysis.

### Normality Assumption

We evaluated normality using Q-Q plots and the Shapiro-Wilk test:

* Q-Q Plot: From the previous Q-Q plot of income data it was exceptionally observed that the graph departs from the expected normality in the upper tail suggesting positive skewness.
* Shapiro-Wilk Test: The test result ‘p’ < 0.05 validates income data that was not normally distributed from the normal distribution. Income was transformed using a log function because it normality of distribution was slightly skewed; the Q-Q plot revealed points closer to the line of normality.

Normality was addressed appropriately since this validated that our model assumptions were met to increase the reliability of our findings.

### Homoscedasticity assumption

Which means that the errors on the regression are constant. Homoscedasticity, or constant variance, was examined through residual plots and the Breusch-Pagan test:

* Residual Plot: Analyzing initial plots of the data produced coefficients of variance whose pattern grew with income levels which points towards heteroscedasticity, a contradaindicator of model stability.
* Weighted Least Squares (WLS) Regression: The reason being that F test is affected by the higher variance data points Therefore, for this purpose, we used wls regression which assigns weight to such data points. This modification also had the effect of getting more equal variance at the set of values.

By making homoscedasticity, we are also able to minimize variance which could otherwise be influencing the suitability of the model.

### Multicollinearity Assumption

To detect multicollinearity, we calculated Variance Inflation Factor (VIF) values for each predictor:

* VIF Results: This showed that all the VIF scores were below 5, therefore Low multicollinearity. In this, it can be inferred that each determinism makes a unique input to the model with no support from the other factors.

To control for multicollinearity, we made sure that independencies in our predictors were enough different hence enhancing the stability and interpretability of our regression equation.

## Interpretation of Results

### Expected Results

A seasonal distribution of crime rates was predicted due to normative crime rates with more crime incidences occurring during the summer. In the same light, low correlation between age and crime category supports the view that age may not in itself be an explanation for criminality but may with other factors.

### Unexpected Results

Higher skewness of income distribution was expected but it greatly impacted our regression model assumptions. This led to the application of log transformations together with the WLS regression technique to handle issues of heteroscedasticity. In addition, an absence of any powerful age-crime connection posed epistemological problems and directly contradicted the original hypothesis because prior research ascertained that age made a more powerful demographic impact on crime.

### Unsettling Results

Concerning the stability of the analysis, the nature of the income data appeared to deviate significantly from normality. Yet, in the case of income, such concerns were overcome by way of the log transformation, which closer fit the requirements for normal distribution.

## Implications for Analysis

These findings shaped our analysis approach:

* Model Selection: Because of the fairly low degree of association between predictors we shall examine logistic regression for categorical crime prediction as a possibility to linear regression.
* Feature Engineering: The insights derived from the seasonality analysis point to ability to add time-varying parameters that are relevant to crime fluctuations over time. The income was then taken to the log transformed state, this aligned the data more linear and normalized it for better use in prediction.
* Future Considerations: These results suggest that other contextual variables (i.e. individul-level attributes, such as social-economic status, characteristics of place) should be introduced in future studies to partly explain crime fluctuations.

The profound scrutiny of assumptions guarantees that the gathered dataset will meet analytical demands, hence providing a proper framework for arriving at accurate and valid conclusions. All the modifications such as transformation increases the credibility of this study, which alleviates concerns of either floored or ceiling effects, and alternative modeling improves the convergent validity, which is important when explaining the reason why our results are different from other studies.

# Research Question 1

A graph with different colored lines

Description automatically generated

Fig 1.1 Top Ten Crimes by Gender

Using Los Angeles crime data the patterns are identified of gender differences in crime victimization, and they show that gender matters in the kind of crimes that people are subjected to. After examining the dataset, we notice that males and females are criminals of different type of crimes at different frequency and this reflects the current trends in the society and the behavior. The second plot, being the top 10 crimes by gender, demonstrates that crimes like Simple Assault and Aggravated Assault, occurring with the same frequency across all gender categories, but with some variation in terms of frequency. Aggravated assaults, slightly more serious assaults involving weapons or at least more violence than the average, appear to disproportionately burden men. This pattern implies that a behaviorally driven pattern of increased exposure to high-risk environments or situations that expose them to violence, e.g. by socioeconomic or other lifestyle factors, exist for males. Unlike Intimate Partner - Simple Assault, female victims make up more of this situation, showing that domestic violence occurs quite frequently from intimate partners who may assault women more often than men. This trend follows the pattern of general societal findings that women are often disproportionately more involved when it comes to domestic violence and support programs that must be put in place to accommodate gender specific interventions that address intimate partner violence.

Additionally, the data shows that more male victims are suffering from Robbery, meaning that men are generally targeted in theft which are considered crimes. There could be any number of reasons for this, such as occupational risks, social behavior, or the idea that males would have more valuables with them or are less likely to report robberies. Another such crime that returns frequently among both genders but tips slightly to the male victim is Battery - Simple Assault. Less severe than aggravated assault, this form of assault may be seen in routine interactions leading to the explosion of everyday social conflicts or altercations, where structural social norms may shape that men are more likely to be involved in the physical confrontation or conflict resolution. Interestingly, as it turns out, crimes such as Burglary and Attempted Robbery have a balanced distribution of male to female victims suggesting that perhaps these crimes are less impacted by the victim’s gender and more situational factors that influence the crime, such as where it occurs or how accessible a property might be. The fact that the distribution of crimes is so balanced suggests that some crimes are less dependent on the gender of the victim, and hence more opportunistic.

A graph of blue and orange bars

Description automatically generated

Fig 1.2 Top 10 Crimes Most Disproportionately Committed Against One Gender

The second plot, on crimes that are disproportionately committed by one gender, is much more detailed. Far more than any other type of offense, Simple Assault Intimate Partner is committed against female victims, demonstrating an alarming pattern of gender-based violence in the context of domestic relationships. In this case, the focus is given to the high level of vulnerability of women, which is due to the situation in which intimate relationships take place, where power and dependence dynamics as well as social norms can increase the quantity of the incidents of violence in respect of women. Furthermore, the pattern of jurisdiction is highly skewed for female victims of Rape, Forcible and Battery with Sexual Contact establishing the high prevalence of gender-related violence and sexual crimes against women. This pattern is consistent with other trends of assaults on women across society, factors being societal attitudes on gender, the absence of deterrents whether deterrents in form of effective measures, rendering women more likely to be victim of assault. These findings underscore the need for focused policies and interventions to safeguard women from such crimes, as well as to assist survivors in seeking justice and healing.

In contrast, Male victims, are most common types of victims for Assault with Deadly Weapon and Aggravated Assault, which might be explained by an increase of the exposure to violent conflicts or settings involving weapons (because of work, social environment, or lifestyle) especially for males. Males may be at a greater risk for violent crime in public and/or high-risk areas where altercations may move towards becoming a deadly confrontation. Brandish Weapon and Robbery are skewed towards male victims indicating that men are more likely to be involved in encounters that go toward a threat with a weapon. There could be many reasons for this pattern — for example, men are more likely than women to engage in activities that put them at greater risk of robbery or a violent confrontation. These gendered patterns in violent crimes help to inform the work of law enforcement and social service, telling us that preventive measures specifically designed for men are needed, such as outreach programs in high crime areas, or public awareness campaigns targeting at risk demographic groups.

In addition, the differentials in crime violation by gender reflect societal and behavioral patterns that have important bearing for crime prevention and allocation of law enforcement resources. Intimate Partner - Aggravated Assault and Battery with Sexual Contact are predominantly female and illustrate a need for specialized sexual assault treatment programs, sexual assault specific shelters, and additional legal protections for women who are victims of intimate partner violence. Meanwhile, interventions targeting situational awareness, self-defense education and support services to men in high-risk occupations or in high risk social settings may be valuable for male victims of aggravated assault and robbery. To thus create gender sensitive crime prevention strategies, one must understand each of these patterns. For example, community programs fighting to end domestic violence, should be working to empower women, educate them on their rights, and provide them with resources to seek legal and psychological support. Just as bo should be developed to suit the men – especially in high-risk areas and jobs – with strategies to prevent attack and theft, so too should personal safety and awareness be tailored for them.

This is also evident in the analysis of gender differences in crime victimization which illustrates broader collective problems involving gender inequality, power relations, and society’s norms on individual safety. Crimes against women, such as sexual and domestic violence, are ingrained deeper issues of the society and require complete approaches, from policy revamp to public education to the system for community support. Moreover, the results also suggest the need to improve gender inclusive data collection and analysis in crime research. Research and policymakers can work separately for male and female victimization patterns, to get greater insight into how various groups are impacted by crime, and eventually develop better tactics. This knowledge is important for law enforcement agencies to know because it helps them allocate resources better: it can mean sending more officers to high risk neighborhoods, providing training for gender sensitive policing, or working with social services on vulnerable populations.

Finally, this data analysis paints a fine picture of the role gender plays on the crime victimization in Los Angeles and shows prevalent disparities between various types of crimes. The high level of exposure to high – risk environments coupled with the society – wide view that male are expected to be able to resolve conflicts in some instances makes the likelihood higher for them to be involved in violent and robbery related crimes. In women, however, the prevalence of domestic violence and sexual crimes is disproportionately high, indicating the need for thoughtful protection in continuing societal challenges to gender based violence. This analysis points to the need for effective crime prevention intervention to take these gender differences into account and suitable interventions and swift targeted resources to overcome the needs of both groups. Using a gender sensitive approach to crime prevention, law enforcement and policy makers can design interventions that help create greater safety in communities and decrease the level of crime influencing such vulnerable populations. In addition to making crime prevention strategies more effective, this method leads to a more equitable and just society in which people are safe and secure, irrespective of gender.

# Research Question 2

A graph of a bar chart

Description automatically generated with medium confidence

Fig 2.1 Crimes against Young Victims (Ages 1-20)

Youth Victims (Ages 1-20): Among the most common crimes made by young victims, especially minors and adolescent, are Simple Assault, Physical Child Abuse, Robbery and Aggravated Assault. Especially prominent is simple assault possibly because of the number of school related scuffles, squabbles or fights in public places where young people congregate. Crime involving child abuse emphasizes a serious issue, where children do not only face abuse inside domestic environment but also within institutional environment. It also includes intimate partner violence incidents that, likely, target older teens in the group, emphasizing the need for incorporating relationship violence in youth education. Moreover, robbery prevalence may be an artifact of neighborhood exposure to such street crimes experienced by the young.

A graph of blue and white bars

Description automatically generated

Fig 2.2 Crimes against Young Audults

Young Adults (Ages 21-40): For youth victims, Simple Assault, Aggravated Assault, and Robbery still affect young adults, like youth victims. Typically, though, Simple Assault Intimate Partner and Weapon Brandishing occur often with this group, likely suggesting a higher propensity for domestic violence and attacking. People of this age are more prone to be involved in social activities, workplace fights and relationships amongst themselves which makes them have greater hazards associated with interpersonal and community violence. It may be, too, that the age group is marked by a higher prevalence of robbery and other kinds of violent crime because of differences in occupation or lifestyle, including the greater likelihood that young adults are found in high-risk areas or situations. This age group needs preventive measures, such as programs in conflict resolution and intimate partner violence.

A graph of blue and white text

Description automatically generated

Fig 2.3 Crimes against Middle-Aged Adults

Middle-Aged Adults (Ages 41-60): Simple Assault and Aggravated Assault are also high in Middle Aged adults, but Criminal Threats and Burglary become significant crimes in these adult groups. Social and community disputes or workplace conflicts or neighborhood tensions could be attributed to criminal threats on middle aged individuals. Burglary related incidents are indicative of middle age being more likely to report property crimes perhaps basis increased property ownership. Because of their connections with social networks and family responsibilities, people falling in these groups are most likely to come into conflict with threats and assaults. This age group could benefit from programs on property crime prevention and community mediation, reducing tensions and encouraging neighborhood safety.

A graph of a number of blue bars

Description automatically generated with medium confidence

Fig 2.4 Crimes against Older Adults

Older Adults (Ages 61-80): The crimes committed by older adults are characterized as such actions as Simple Assault, Aggravated Assault, Burglary and Criminal Threats, but with fewer violent crimes than with younger age groups. In certain situations where older adults are isolated or lack networks of support, vulnerabilities regarding property crimes and personal safety may present themselves for older adults. Elderly people may be seen as having less capacity to defend against intrusion than adults of an appropriate age, thereby increasing the risk of burglary. Criminal threats to older adults may be associated with tensions with neighbors or tensions within care facilities. Targeted strategies, like neighborhood watch programs and increased security measures for senior citizens, will address the vulnerabilities of older adults and deter the occurrence of burglary and other crimes.

A graph with blue bars

Description automatically generated

Fig 2.5 Crimes against Elderly Victims

Elderly Victims (Ages 81+): Of all the elderly, Simple Assault remains the most common, and Aggravated Assault, Burglary and Criminal Threats are next most common. While older people are generally victimized less than younger persons, they are especially susceptible to property crimes and personal assaults borne of their perceived vulnerability. Physical limitations, immobility, or the high probability of being a lone resident may make one an elderly victim, especially when older residents have little protection from elderly abuse. Burglary and property related offenses are substantial in this group, underscoring a need for enhancement of security and support systems for older adults. These risks can be mitigated, making elderly individuals feel safer in their homes and life in their neighborhoods, by organizing enhanced outreach programs informing the individuals on personal safety, and encouraging involvement in the community.  
  
Examining age-based crime trends exposes key patterns pointing the way to how vulnerability to types of crime changes with age. Public places will attract younger people—those are the most susceptible when it comes to assaults, child abuse, and robbery, probably due to school environments and social interactions. Young adults are at an increased risk of life damage associated with domestic violence and weapon banding, due to societal and domestic factors and dynamics. With property ownership and engagement in their community, middle aged adults experience a larger range of crimes including criminal threats and property crime. Burglary being a major property crime, has a higher impact on the elderly and older adults and therefore they need supportive safety measures and neighborhood watch programs to keep the victims safe.

Incarceration and identifying these trends suggest that a more age-specific approach to crime prevention will be crucial. Understanding the special risks of each age group can facilitate law enforcement and policy makers in devising age specific programs to increase safety and lower crime in the Los Angeles community. Specifically, crime prevention requires that intervention programs be tailored to meet the unique needs of different age groups so that resources are used efficiently to protect those most vulnerable.

# Research question 3

A graph of blue bars

Description automatically generated

Fig 3.1 Yearly trends of Crime

These insights provide clues about the temporal trends in crime occurrence in Los Angeles across years 2020 to 2024 regarding the trends in types of crimes, which are clearly affected by various social, economic and seasonal factors. Yearly totals from 2020 to 2023 were relatively consistent across all crimes at between 75,000 and 80,000 incidents. The stability in both series implies the city’s perpetuation of crime rate which is relatively independent of possible policy changes or crime prevention activities during these years. Also interestingly, 2024 sees a marked drop, underreporting possibly as a few months are represented for this year. The observed decrease could reflect the effect of successful new policies, changes of law enforcement strategies, or broadly evolving society influencing criminal behavior. Further context to this drop can be found by digging deeper into a particular policy implementation, or impactful event in early 2024. Unique crime patterns of each type of crime within this period also play out throughout each of these years, indicating that some offenses are more susceptible to social and economic factors than others.

A graph of a number of vehicles

Description automatically generated with medium confidence

Fig 3.2 Yearly trends of Individual top crimes

By diving into the trends by crime type, we get a bit more granularity regarding these fluctuations. Vehicle Theft also varies considerably, falling in 2021, then rising to peak in 2023. The peak here reflects either changes in economic conditions, such as increasing values of vehicles making them better targets, or changing policing patterns in certain neighborhoods that allows such incidents to rise. Simple Assault is high but stable – Battery – over the years, indicating that a frequent problem has been minor interpersonal conflicts (urban density, social stressors, routine public interactions that then expand into violence). Consistency of this kind means that the levels of some social behavior crimes that are consistent are not likely to change, whatever else changes in society. The pattern of variability in Burglary from Vehicle also mirrors this, suggesting an increase in organized property crimes, such as 2023 (or in response to increased economic pressures that drive people toward such thefts). In 2023 Theft of Identity is also on the increase which may be associated with increased online activities, the movement in technological trends, or just economic conditions that facilitate for new opportunities in digital fraud and exploitation. Vandalism shows a more steady rate up to 2023 where it starts to decrease, possibly due to surveillance and community watch initiatives or urban development projects that reduce this type of behavior.

Leaping further down from the annual levels, monthly trends help identify the seasonal patterns of crime in terms of number of crimes committed and revealed that January was the month with most of the instances across different crimes. Post holiday social gatherings, as well as a general increase in social and congregate activity, may be associated with this trend as recent increases in crime. After January, the crime rates only go down the rest of the year and lowest rates are observed in December, which is also the time of the year that law enforcement presence is most present and less social activity is present because of the holiday season. The appearance of these patterns on a monthly basis suggests that some crime can be seasonal, which demonstrates weather and public and economic pressures related to times of the year. The fact that crime rate in January is relatively high implies that extra resources and vigilance would be needed to contain the criminal activities in that time of the year.

A graph of blue bars

Description automatically generated

Fig 3.3 Monthly trends of Crime

I further dig into the seasonal crime trends on a per crime type per month basis. Vehicle Theft has marked monthly fluctuations, usually with more incidents in the early months. It could be related to any number of seasonal factors, such as cold weather dictating where people park cars or more activity after the holidays opening up more chances for theft. Across the months, Battery - Simple Assault also decreases, with the highest counts early in the year, possibly related to increased social gatherings, family conflict or public events in the winter and early spring that result in minor violent incident. Burglary from Vehicle steadily declines after January, and it might be due to some seasonal changes in outdoors activity, tourist influxes and commuting scenarios which can change throughout the year. Irregular peaks and valleys characterize Theft of Identity, with the highest frequency occurring at the beginning of the year, perhaps the result of tax season and activity around this time which increases one’s vulnerability to identity theft. High vandalism continues early in the year but declines through the course of the year possibly due to patterns in property related disorder and public events associated with property damage at the start of the year along with see increase in deterrent measures midyear. General Burglary also varies month to month, with more incidents showing up at the beginning and at the end of the year, perhaps linked to holiday times when homes are often empty, and therefore vulnerable to burglary.

A graph of a graph

Description automatically generated with medium confidence

Fig 3.4 Monthly trends of individual Crimes

Analysis of temporal crime trends shows that the trend of some crime, simple assault for example, is not significantly different from one year to another which means these three types of crime, are somewhat stable, are not as much affected by external drivers and more by stable social behavior. On the other hand some crime, such as vehicle theft and identity theft, suggests strong yearly and monthly variations with possibly economic, technological, or seasonal impact. For various types of crime, the seasonal peak in January implies that the law enforcement could use proactive approach on each start of the year and reallocate the additional resources to prevent the increase of crime and to raise public security in the periods with high risk. Patterns at the monthly level also highlight the need for targeted action that accounts for variation in crime over time within the year, so that law enforcement can focus resources on those crime types which are more likely to occur in a given month.

Finally, the temporal trends in crime occurrences in Los Angeles from 2020 to 2024 underline the necessity of knowing when the different types of crime vary and how those changes are conditioned by both seasonal effects, economic factors, and social phenomenon. Is there a recent shift in legal enforcement strategies or broader social changes leading to the potential of those rates in 2020 – 2023 being consistent and then dropping in 2024? Combining the yearly and monthly breakdowns by crime type, we are able to see that some crime types are more sensitive to these externalities, such as vehicle and identity theft, which respond to changing external conditions. Meanwhile, there are consistent trends in the crimes of battery and vandalism remaining social issues that lead to the same. Several crime types see their peak seasons at the start of the year, with a seasonality peak observed in January. The ability to recognize these temporal patterns gives law enforcement and policymakres more of an ability to design crime prevention measures that work with resources and interventions at times when distinct crimes will occur. According to these results, through an adaptive approach, the city can find a way to decrease crime rates, increase public safety, and to deal with the specific risk of particular times of the year.

# Research Question 4

**Logistic Regression**

Classification algorithms are logistic regression and used for binary tasks. The logistic function is used to map the sum of the input features to the probability of a binary outcome. Crimes are relatively easy to interpret. However, in case of nonlinear decision boundary, logistic regression may miss complex relationships although it is especially convenient for linearly separable data.

A screenshot of a computer

Description automatically generated

Fig 4.1 Logistic Regression Metrics

A graph of a line graph

Description automatically generated with medium confidence

Fig 4.2 Logistic Regression PRC and AUC

Logistic regression had an AUC-PR: 0.68, AUC-ROC: 0.72 for reasonable predictive power. The averaged model cross validated scores were around 0.75, so the model has a decent hypothetical predictive capability, but struggles to distinguish between the crime types. A confusion matrix shows that while the model is highly accurate in predicting the Positive class, it becomes afflicted by a high count of false negatives (low Recall) for some crime types. In general, logistic regression gives us a reliable estimate of a baseline, but is too simple to reach the high accuracy needed here.

**K-Nearest Neighbors (KNN)**

Non parametric model called K-Nearest Neighbors (KNN) uses the majority class of k nearest neighbors of an observation to classify the outcome of the observation. In this model, we compute distances (Manhattan distance in this case) between features to determine what degree of feature ‘closeness’ the data points have. For problems with complex decision boundaries (boundaries like having an island in the middle of a sea), or which need to adapt to the structure of the data, KNN can be very effective. This is especially useful when there exist larger patterns (or clusters) in the data that have large gaps between the class in the feature space.

A screenshot of a computer

Description automatically generated

Fig 4.3 KNN Metrics

A graph of a function

Description automatically generated with medium confidence

Fig 4.4 KNN PRC and AUC

KNN was tuned with 11 neighbors and a distance weighted voting scheme where closer neighbors are weighted more highly for classification for the current task. A cross validation outlined a mean score of 0.79, an AUC-PR of 0.82 and an AUC-ROC of 0.84, which was a significant improvement from logistic regression. The confusion matrix exhibited a decreased number of false negatives, meaning improved recall across crime types using KNN. As such, higher recall means that KNN is better at classifying instances of positive class correctly and this is simply because the model can represent local patterns in the data. The performance of KNN in general was a significant increase over logistic regression, which seems to indicate that this dataset was suited to KNN.

**LightGBM**

Light Gradient Boosting Machine (LightGBM), often regarded as a gradient boosting framework choosing speed and efficiency for large datasets over accuracy. It performs sequential construction of multiple decision trees, where the successive trees try to correct the errors made by the previous tree by assigning the higher weights for the misclassified instances. In addition, with leaf-wise tree growth, LightGBM has higher accuracy on the same computational time than level-wise methods.

A computer screen shot of a code

Description automatically generated

Fig 4.5 LightGBM Metrics

Finally, as one of the best-performers for AUC-ROC (0.87), and accuracy (81.26%) on context magnitude and sentiment segmentation, LightGBM represents a top model for the classification task in this case. The high AUC-ROC guarantees a well balanced false and true positive rate what confirms our model distinguishes crime types properly. The LightGBM confusion matrix revealed much improved precision and recall and fewer misclassifications than past models. This shows that LightGBM is suitable for predicting crime severity, by using feature interactions and being able to model non linearly separable relationships.

**XGBoost**

Another gradient boosting algorithm, known for high performance and efficiency is XGBoost (Extreme Gradient Boosting), which is popularly used in the structured data competitions. Moreover, it over the standard gradient boosting by adding regularization, early stopping as well as dealing with missing values. Both XGBoost and LightGBM construct an ensemble of decision trees, each subsequent tree tries to reduce the residual errors from the previous trees, and hence we achieve a more refined model.

A screen shot of a computer

Description automatically generated

Fig 4.6 XGBoost Metrics

It can be seen that in our study, XGBoost shows the same performance as LightGBM, with an AUC-ROC of 0.866 and an accuracy of 81.11%. The AUC scores from cross validation were generally very stable and consistent averging 0.86 or 0.87. This classifies with robust performance in separating the crime severity, and the confusion matrix shows that XGBoost, as is the case with LightGBM, minimize misclassification rates for both classes. While doing a bit worse than LightGBM in performance, the slight differences in model tuning or algorithm design may be to blame, but on the whole, XGBoost proved itself to be very powerful and effective model in this case.

A graph of a curve

Description automatically generated

Fig 4.7 AUC of LightGBM and XGBoost

**Artificial Neural Network (ANN).**

In this analysis, the Artificial Neural Network (ANN) used is a multi layer perceptron which can learn complex nonlinear relationships in the data. The ANN architecture consists of multiple layers of neurons, in which each neuron performs a weighted summation of the inputs and then activates using activation functions. Using TPU, we were able to achieve more stable and faster processing of large data batches on this model that is well suited for data with high dimensionality and complex interactions.

A screenshot of a computer

Description automatically generated

Fig 4.8 ANN Metrics

We found that the ANN model had an AUC-ROC of 0.84 and an accuracy of 78.7%, both strong, but slightly lower than the performance attained by LightGBM and XGBoost. The confusion matrix for the ANN is well balanced between precision and recall, but its performance may have been impaired from problems tuning neural networks for structured data (which unfortunately tree based models are often better suited for). However, the performance of the ANN is respectable on its own, and has the potential for further improvement given more hyperparameter optimization.

A blue line graph with a dotted line

Description automatically generated

Fig 4.9 ANN AUC

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **AUC-ROC** |
| Logistic Regression | 0.75 | 0.72 |
| K-Nearest Neighbors | 0.79 | 0.84 |
| LightGBM | 0.81 | 0.87 |
| XGBoost | 0.81 | 0.866 |
| Artificial Neural Network | 0.787 | 0.84 |

LightGBM and XGBoost are shown to perform best on this dataset for predicting crime severity based on location, time and victim characteristics with high accuracy and AUC-ROC scores.

References

Bennett Moses, L., & Chan, J. (2016). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3742541

Birks, D., Coleman, A., & Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data. *Crime Science*, *9*(1). https://doi.org/10.1186/s40163-020-00127-4

*Cogent*. (n.d.). Cogentinfo.com. Retrieved November 18, 2024, from https://www.cogentinfo.com/resources/predictive-policing-using-machine-learning-with-examples

*Data analytics in crime prediction and prevention*. (2024, February 8). Criminal Justice. https://criminal-justice.iresearchnet.com/criminal-justice-process/impact-of-technology/data-analytics-in-crime-prediction-and-prevention/

Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., & Liu, Q. (2019). Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 106111–106123. https://doi.org/10.1109/access.2019.2930410

Murray, A. T. (2001). Exploratory spatial data analysis techniques for examining urban crime: Implications for evaluating treatment. *The British Journal of Criminology*, *41*(2), 309–329. https://doi.org/10.1093/bjc/41.2.309

*Program profile: Predictive policing model in Los Angeles, Calif*. (n.d.). CrimeSolutions, National Institute of Justice. Retrieved November 18, 2024, from https://crimesolutions.ojp.gov/ratedprograms/predictive-policing-model-los-angeles-calif

Rao, A. (2021, June 14). *Visualizing crime in Los Angeles*. Towards Data Science. https://towardsdatascience.com/visualizing-crime-in-los-angeles-14db37572909

Saeed, R. M., & Abdulmohsin, H. A. (2023). A study on predicting crime rates through machine learning and data mining using text. *Journal of Intelligent Systems*, *32*(1). https://doi.org/10.1515/jisys-2022-0223

Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 70080–70094. https://doi.org/10.1109/access.2021.3078117

Seidensticker, K., & Schwarz, K. (2022). Using forecasting methods on crime data: The SKALA approach of the state office for criminal investigation of north Rhine-Westphalia. *The 8th International Conference on Time Series and Forecasting*, *21*, 39.

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Visual Computing for Industry. *Biomedicine, and Art*, *4*(1).

Ucla, D. at. (2024, July 5). *What does crime in Los Angeles city look like? - DataRes at UCLA*. Medium. https://ucladatares.medium.com/what-does-crime-in-los-angeles-city-look-like-2cddfb07374b

(N.d.). Ssrn.com. Retrieved November 18, 2024, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=1160692