Predicting Violent Crime in Los Angeles: A Spatial and Temporal Analysis

Rylenn Berry[1], Elizabeth Thompson[1], and Juliana Perez Romero[1]

[1]Arizona State University, Tempe, AZ 85281, USA

Abstract

Violent crime remains a persistent public safety concern in large U.S. cities, and Los Angeles has consistently reported above average rates with incidents concentrated in specific neighborhoods. This project uses recent crime data from the City of Los Angeles Open Data Portal (2020–2025) to examine how spatial, temporal, and demographic factors shape patterns of violent versus non violent incidents. After cleaning and preprocessing approximately one million reported crimes, we conducted exploratory data analysis to describe trends over time, map crime across LAPD divisions, and summarize victim characteristics. We then built classification models to predict whether an incident is violent using demographic, contextual, and spatiotemporal features. Logistic regression models were used as interpretable baselines, first with victim and incident characteristics alone and then with the addition of spatial and temporal predictors. We further estimated Random Forest models to capture nonlinear relationships and interactions among predictors.

Results from the descriptive analysis show that violent crimes comprise roughly 45–50% of reported incidents each year, with limited seasonality but clear clustering across divisions and neighborhoods. Both modeling approaches indicate that victim sex, victim descent, age, and location are among the most informative predictors of violence, while fine grained temporal features add only modest predictive value. Random Forest models achieve slightly higher discrimination than logistic regression but still exhibit moderate performance, underscoring the limits of incident level prediction in this context. Overall, the findings highlight the importance of place and demographic structure in understanding violent crime in Los Angeles and point to broader social and policy factors that extend beyond what can be captured in administrative crime data alone.

1 Introduction and Background

1.1 Introduction

Violent crime remains a central public safety concern in major urban areas across the United States. According to the Federal Bureau of Investigation (2019)[6], violent offenses, such as homicide, rape, robbery, and aggravated assault, are defined by the use or threat of force and carry significant social, economic, and community impacts. Although national crime trends fluctuate over time, large metropolitan cities continue to face persistent challenges, with Los Angeles standing out as one of the most affected regions. Recent city reports show that violent crime rates in Los Angeles have remained elevated compared to national averages, with incidents concentrated in specific neighborhoods and police divisions [12].

Understanding when, where, and under what conditions violent incidents occur is critical for developing informed public safety strategies. Prior criminological research consistently finds that violent crime is not evenly distributed across space or time, but instead clusters in identifiable patterns shaped by neighborhood structure, demographic composition, socioeconomic disadvantage, and routine activities [2, 8, 9]. Spatial studies in Los Angeles demonstrate that violent hot spots persist across years and often align with broader structural conditions, including poverty, residential instability, and uneven access to community resources

[5, 17]. Temporal analyses likewise show modest but recurring patterns related to hour of day, season, and broader social disruptions, although these trends often exhibit weaker signals than spatial or demographic factors [20].

In parallel, advances in computational social science and machine learning have provided new tools for modeling crime and identifying the factors most strongly associated with violent incidents. Logistic regression models have long been used to examine risk factors for violence and assess the influence of demographic and contextual predictors [11, 14]. More recent work applies large scale simulations, spatial clustering methods, and ensemble algorithms such as Random Forests to capture nonlinear interactions and complex relationships that traditional statistical methods may overlook [3, 18, 19]. Across this literature, the integration of spatial and temporal features, not just who is involved but where and when incidents occur, has been shown to meaningfully improve model performance, though prediction remains challenging due to the inherent complexity of violent behavior.

This project contributes to this body of work by analyzing nearly one million reported crime incidents in Los Angeles from 2020 to 2025. Using publicly available administrative data, we examine how demographic, spatial, and temporal characteristics relate to the likelihood that a reported incident is violent. In addition to descriptive visualization of citywide patterns, we implement both logistic regression and Random Forest classifiers to evaluate the predictive value of key features and compare model performance. Rather than aiming to forecast individual incidents for policing purposes, the goal of this study is to better understand the structural patterns embedded in the data, highlight which factors most strongly differentiate violent from non violent crimes, and situate these findings within broader criminological and policy discussions.

1.2 Background

Crime in the United States has always reflected broader social and economic conditions. In the last few decades, national crime rates have fluctuated in response to economic and demographic trends and policy changes. The national crime rate peaked during the 1980s and early 1990s. Since then, violent crime has declined in many areas of the country. However, major cities continue to face challenges tied to national issues, including inequality, housing instability, and limited access to social resources. Economic recessions and policy reforms have also played substantial roles in shaping when and where crimes occur [6, 7].

Policing in the United States adapted as social and economic conditions changed. In the late nineteenth and early twentieth centuries, cities established official police departments in response to growing populations. Organizations that had previously been informal "watch groups" evolved into formal agencies focused on maintaining order in increasingly complex environments [15]. Over time, reforms introduced new expectations around accountability and training, particularly in response to corruption and social unrest. These institutional changes often reflected broader social transformations, illustrating how shifts in public trust and civic engagement directly influenced policing practices [5, 15]. These changes laid the foundation for the modern relationship between law enforcement and the public, a relationship that continues to be redefined by both policy and community pressure.

Los Angeles' policing history is intertwined with its rapid growth and deep social divisions. From early efforts to establish consistent patrols to later controversies over civil rights and police misconduct, the Los Angeles Police Department (LAPD) has contributed to national debates about justice and authority [5]. The LAPD's evolution has included periods of reform aimed at restoring public confidence, particularly after high profile incidents that exposed systemic problems within the department. Scholars note that the LAPD's trajectory also reflects national tensions around race, class, and power, as policing often mirrored the city's segregated geography and unequal access to resources [5, 15].

The civil rights era caused greater scrutiny within the LAPD. Later, the "War on Drugs", a federal campaign beginning in the 1970s aimed at reducing illegal drug use and trafficking, reshaped enforcement priorities, emphasizing control and deterrence but also deepening mistrust between officers and marginalized communities. This shift, often linked to federal initiatives beginning in the 1970s and 1980s, disproportionately impacted low income and minority neighborhoods, reinforcing cycles of incarceration and surveillance [6, 15]. By the 1990s and early 2000s, Los Angeles began experimenting with community policing strategies meant to rebuild trust and encourage collaboration with residents [12]. These efforts have yielded mixed

results, but they highlight a growing recognition that public safety depends as much on relationships and transparency as on enforcement [10].

Today, conversations about policing, across the country and in local communities, focus on finding a balance between keeping people safe and holding police accountable. Los Angeles shows how changes in population, social movements, and history come together to influence how crime is viewed. The city's experience illustrates that policing in the United States is constantly adapting to societal pressures. Recent projections indicate that this evolution will continue as data driven approaches and decarceration policies influence how cities manage crime and public safety in the coming years [13, 17].

According to crime reports on Los Angeles, neighborhood conditions and socioeconomic factors influence where crime occurs and who is affected. Areas with higher poverty, unemployment, and limited access to social services often have elevated crime rates, indicating structural inequalities in the city [5]. Cung (2013) found that low income neighborhoods consistently had higher reported crime rates than wealthier areas, indicating that crime is not evenly distributed across Los Angeles. These patterns suggest that understanding crime requires considering the social and economic context of different communities.

Neighborhood racial composition also affects crime patterns and policing practices. Rosenfeld and Austin (2023) [17] examined how the racial makeup of neighborhoods influences crime reporting and law enforcement. They found that minority communities often experience higher levels of policing, even when crime rates are similar to those in predominantly White or higher income neighborhoods. This suggests that enforcement practices are shaped by more than just crime rates and that systemic factors influence how police interact with different communities. Overall, research indicates that both socioeconomic status and racial composition are significant factors in where and how crime occurs and is addressed by law enforcement [5, 17]. Neighborhoods with concentrated poverty and large minority populations may face both higher crime risk and more intensive policing, while wealthier areas experience fewer interventions. Understanding these patterns is important for designing policies and policing strategies that consider the social context of crime. Data driven approaches, such as spatial models and agent based simulations, are increasingly used to examine these dynamics and identify patterns in urban crime [10].

Mapping and clustering techniques have become indispensable for revealing the fine grained spatial and temporal structures of urban crime that simple aggregate tables conceal. Geospatial visualizations (choropleth maps, kernel density surfaces) show where indecent density concentrations occur, while clustering and hotspot methods (e.g., Mean Shift, spectral clustering, scan statistics, kernel density estimation, and space time cubes) formally identify persistent and emerging clusters. Recent applications to Los Angeles illustrate the value of these tools. Lee et al. (2024) [10] used Mean Shift and spectral clustering on hundreds of thousands of LAPD incidents to locate stable violent and property crime hotspots and to expose seasonal cycles that aggregate yearly rates miss. Rosés et al. (2021) [18] demonstrated how agent based simulations, informed by demographic and environmental inputs, can reproduce observed spatial hotspots and temporal fluctuations and thus serve as testbeds for interventions. Complementary approaches such as Bayesian hierarchical spatial temporal models [2] add statistical rigor by explicitly modeling spatial autocorrelation and temporal trends, enabling inference about whether local increases are part of broader regional patterns or truly local phenomena. Together, mapping and clustering methods provide both the exploratory maps that guide intuition and the formal models that quantify spatial temporal dependence, making them central to any project that aims to predict where and when violent incidents are most likely to occur.

Despite their strengths, prior spatial temporal studies of urban crime exhibit recurring limitations that motivate our capstone focus. Many analyses rely on historical or truncated timeframes, which can miss recent shifts in crime dynamics driven by policy, economic change, or social events; this is important because up to date data can change hotspot locations and model performance. Several studies also stop short of fully integrating demographic and socioeconomic covariates; for example, treating hotspots as purely spatial phenomena without examining how poverty, population density, or land use explain clustering, thereby limiting understanding of root causes and the equity implications of predictive models [5]. Data quality issues are another common constraint: geolocation errors or missing coordinates, incomplete or imprecise timestamps, and underreporting or bias in police records (e.g., differential reporting across neighborhoods) can distort both maps and model estimates. Finally, methodological choices about spatial aggregation

(county vs. neighborhood vs. block) and temporal resolution (daily vs. monthly) introduce scale dependent artifacts (the Modifiable Areal Unit Problem) and can mask short term dynamics that are critical for prediction. These data directly inform our research focus: by using very recent, fine grained LAPD [13] incident data and explicitly incorporating demographic/contextual variables (while carefully preprocessing to address geocoding and timestamp gaps), we aim to produce spatial temporal predictive models that are both current and better grounded in the social contexts that drive violent crime.

1.3 Methodologies in Literature

Research methods for studying violent crime have changed significantly over the past several decades. Early studies mostly used descriptive statistics and regression models to look for links between demographic factors and crime rates [8]. These methods summarized patterns in the data but did not account for the spatial or temporal aspects of criminal activity, thereby limiting their usefulness for understanding cities. The development of spatial criminology enabled researchers to examine how neighborhood layouts and the relationships between places shape the distribution and timing of crime. As technology improved and more geographic data became available, researchers could analyze crime at finer scales, such as block groups or street segments [8]. This enabled the study of how crime patterns vary across neighborhoods and over time. More recently, researchers have used machine learning models to analyze large datasets and uncover patterns that traditional regression might miss [8].

As methods developed, researchers began to study both where crime occurs and how these patterns change over time. For example, Balocchi and Jensen (2019) [2] used Bayesian hierarchical spatial modeling to analyze crime trends in Philadelphia. They found that crime rates did not decline evenly across the city: some neighborhoods showed steady declines, while others remained hotspots. Their work shows that crime trends can differ from one neighborhood to another, even within the same city, and that both location and time matter in crime analysis.

Cheng et al. (2022) [3] studied crime in Liangshan Prefecture, China, using over 11,000 criminal judgment records. They used several spatial analysis tools to look at how crime clusters and shifts over time. They found that violent crimes happened most often at night and in winter, while property crimes were more common in autumn and in urban areas like Xichang City. This shows that these models can help identify when and where crime is most likely to happen, and how season and location affect crime rates.

Other researchers have begun using several machine learning models and socioeconomic data to study crime. These methods can handle large datasets and often detect patterns that older methods might miss [8]. For instance, Kshatri et al. (2022) used Naive Bayes, Random Forests, and meta classifiers to predict and classify crime trends using data from the National Crime Records Bureau of India. Their combined model, called an ensemble, achieved an accuracy of 96.6 %, demonstrating that combining different types of algorithms can improve predictive performance.

Rosés et al. (2021) [18] developed a simulation model to examine how offenders move and make decisions within a city. They used three main theories from environmental criminology, Routine Activity, Crime Pattern, and Rational Choice, to guide the actions of virtual offender agents in a simulated version of New York City. The model relied on real mobility data to predict where and when crimes might occur, based on factors such as how easy it is to reach a location and how many potential targets are present. By combining theory with real data, their model matched actual robbery patterns.

One limitation of previous research, including the studies by Kshatri et al. (2022) [19] and Rosés et al. (2021) [18], is that they often relied on secondary data sources that did not precisely match the crime data. This sometimes reduced the accuracy of their models. Cheng et al. (2022)[3] also noted that spatiotemporal models can be sensitive to the quality and scale of geographic data, so even minor errors in location or time can yield misleading results.

1.4 Project Plan

Our project examines the spatial and temporal factors that shape violent crime trends across Los Angeles using recent data from the City of Los Angeles Open Data Portal. After cleaning and preparing the dataset,

we mapped crime incidents to identify persistent hotspots and analyze patterns across neighborhoods, seasons, and time of day. Using Python, we conducted exploratory analyses to track changes in violent crime since 2020 and evaluate how location and context influence crime likelihood. We then compared machine learning models, including Logistic Regression and Random Forest, to determine which most effectively predicts violent versus non violent offenses. Through this combined spatial, statistical, and predictive approach, the project aims to generate evidence based insights that support more informed public safety strategies in Los Angeles.

2 Methods

2.1 Methods: Data Source, Preparation, and Cleaning

The primary dataset used in this project is the City of Los Angeles Crime Data from 2020 to Present [13], an incident level dataset published by the Los Angeles Police Department (LAPD) and updated daily through the Los Angeles Open Data Portal. This dataset includes detailed records of reported crimes from January 2020 onward and contains variables such as the date and time of occurrence, crime type, police reporting area, geographic coordinates, and victim demographics. At acquisition, the dataset contained 1,004,991 records across more than 25 variables, offering a comprehensive foundation for analyzing spatial and temporal crime patterns in Los Angeles. Its regular updates and level of detail make it particularly well suited for evaluating changes in crime trends over recent years.

Column	Non Null Count	Dtype
DR_NO	1,004,991	int64
Date Rptd	1,004,991	object
DATE OCC	1,004,991	object
TIME OCC	1,004,991	int64
AREA	1,004,991	int64
AREA NAME	1,004,991	object
Rpt Dist No	1,004,991	int64
Part 1 2	1,004,991	int64
Crm Cd	1,004,991	int64
Crm Cd Desc	1,004,991	object
Mocodes	853,327	object
Vict Age	1,004,991	int64
Vict Sex	860,437	object
Vict Descent	860,335	object
Premis Cd	1,004,991	float64
Premis Desc	1,004,403	object
Weapon Used Cd	327,247	float64
Weapon Desc	327,247	object
Status	1,004,990	object
Status Desc	1,004,990	object
Crm Cd 1	1,004,980	float64
Crm Cd 2	69,160	float64
Crm Cd 3	2,314	float64
Crm Cd 4	2,314	float64
LOCATION	1,004,991	object
Cross Street	154,236	object
LAT	1,004,991	float64
LON	1,004,991	float64

Table 1: Summary of columns, non null counts, and data types in the raw LAPD Crime Data (2020 Present).

After importing the dataset into Python using the `pandas` library, an initial inspection revealed several inconsistencies. Column headers used non standard spacing and capitalization, so they were standardized to clear, Python compatible names to improve readability and ensure consistency across analysis functions.

Duplicate entries were removed using the DR_NO identifier, which uniquely corresponds to each police report. These initial steps provided a more uniform structure from which additional preprocessing tasks could be performed.

Preparing the temporal features required extensive conversion and extraction. The "Date Rptd" and "Date Occurred" fields were originally stored as strings and were converted into Python `datetime` objects using the explicit format `%m/%d/%Y %I:%M:%S %p`. This conversion enabled reliable computation of additional temporal attributes, including the year, month, day of the week, and hour of occurrence. For crime incidents where the time of occurrence was listed in HHMM format (such as 2315 for 11:15 PM), the hour component was extracted into its own feature. These engineered variables later allowed for deeper temporal exploration of crime patterns.

One issue that emerged early in the analysis was the presence of placeholder timestamps, especially those recorded as exactly 12:00:00 AM. These default times are commonly used when officers cannot determine the precise time of an incident. Initial plots showed unnatural spikes at these times, revealing that they were artificially inflating certain hourly trends. To reduce distortion while preserving legitimate incidents, only those entries with exact placeholder timestamps were removed, and all other cases occurring within the same hour were retained. This approach maintained data integrity while preventing temporal visualizations from being skewed.

Time_Occurred	Hour	Minute
1240	12	40
1210	12	10
1200	12	0
1230	12	30
1255	12	55
1220	12	20
1200	12	0
(Rows omitted)		
Total: 67,813		

Entries where `Hour = 12`

Time_Occurred	Hour	Minute
1	0	1
30	0	30
10	0	10
40	0	40
48	0	48
5	0	5
20	0	20
44	0	44
15	0	15
(Rows omitted)		
Total: 40,468		

Entries where `Hour = 0`

Table 2: Comparison of placeholder like timestamps for Hour = 12 and Hour = 0.

Time_Occurred	Hour	Minute
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
1200	12	0
(Additional rows omitted for brevity)		
Total placeholder entries: 35,200		

Table 3: Entries where `Time_Occurred == 1200` (placeholder values).

Cleaning continued by addressing missing or invalid values. Latitude and longitude were essential for spatial analysis, so records missing either coordinate were removed. These omissions represented less than 2% of the dataset and did not meaningfully alter distributions. Additional data type corrections were applied where appropriate: "Victim_Age" was converted to numeric, implausible values (such as negative ages or

those above 120) were removed, and demographic fields such as "Victim_Sex" and "Victim_Descent" were encoded as categorical values. These steps ensured that key variables functioned correctly within statistical and machine learning models.

After all cleaning, filtering, and feature engineering steps were completed, the dataset was reduced to approximately 800,000 valid records. This cleaned version addressed all major inconsistencies, such as placeholder times, missing coordinates, and irregular data types, resulting in a reliable dataset suitable for temporal, spatial, and predictive modeling. These preparations established the foundation for the exploratory data analysis and machine learning methods presented in the following sections.

2.2 Methods: Identifying Violent vs. Nonviolent

A critical step in this project was converting the LAPD's [13] detailed incident level crime records into a binary classification of violent versus non violent crime. To ensure consistency with national standards and prior criminological research, violent crime was defined using the Federal Bureau of Investigation's Uniform Crime Reporting (UCR) [7, 6] framework, which classifies violent offenses as those involving force or the threat of force—specifically homicide, rape, robbery, and aggravated assault. These categories are widely used in academic research and predictive crime analysis, forming the foundation for both descriptive and machine learning studies of violence [4, 11].

To assign each incident a binary label, we constructed a keyword based classification function that scanned the textual crime description field (Crm Cd Desc). This process follows a similar approach taken in past predictive policing and crime typing studies, in which text based coding reliably distinguishes violent from non violent incidents [14, 16]. The function flagged an incident as violent if its description contained terminology associated with serious interpersonal offenses. Keywords included indicators for homicide (e.g., "Homicide," "Murder," "Manslaughter"), sexual assault ("Rape," "Sexual Assault"), robbery ("Robbery," "Carjacking"), aggravated assault and battery ("Assault with Deadly Weapon," "Shooting," "Stabbing"), kidnapping, armed threats, or firearm discharge. These categories closely parallel those used in the FBI's UCR program and in prior empirical studies of violent crime classification [10, 20]. All incidents lacking violent crime indicators were labeled as non violent.

This binary variable, `Is_Violent`, allowed the analysis to focus specifically on distinguishing characteristics of violent incidents across demographic, spatial, and temporal dimensions. The decision to divide violence in this manner is well supported in the literature: demographic attributes such as victim sex, age, and descent have been shown to influence victimization risk [4], while spatial criminology research finds that violent crime clusters in identifiable neighborhoods or "hot spots" [2, 8]. Moreover, machine learning applications in crime prediction frequently rely on binary violent/non violent labels when assessing model performance or identifying key predictive features [14, 19]. By structuring the outcome variable in accordance with these conventions, this project ensures comparability with existing empirical work and establishes a clear and theoretically grounded target for predictive modeling.

2.3 Methods: Demographic Feature Construction

Demographic variables were a central component of this project because victim level characteristics are consistently shown to influence patterns of victimization and violence in criminological research. Prior studies emphasize that sex, age, and racial or ethnic background are associated with differential exposure to violent crime and reflect broader structural inequalities [2, 4, 8]. Following these findings, demographic fields in the LAPD dataset were cleaned, standardized, and recoded to create usable and analytically consistent features.

Victim sex was cleaned by consolidating inconsistent or ambiguous entries and grouping all values into three categories: Male, Female, and Unknown. This structure aligns with common demographic breakdowns used in prior violence risk and victimization studies [11]. Records with missing or invalid sex values were assigned to the Unknown category to preserve sample size while maintaining transparency in modeling.

Victim descent, which serves as a proxy for racial and ethnic identity, required additional preprocessing due to the large number of coded categories in the original dataset. To reduce sparsity and improve

interpretability, descent codes were grouped into broader, literature consistent categories: White, Black, Hispanic/Latino, Asian, Pacific Islander, American Indian/Alaska Native, Other, and Unknown. These consolidated categories parallel the racial and ethnic groupings used in large scale crime and public safety research, which often requires aggregation to ensure statistical reliability [17, 20].

Victim age was converted to a numeric field, and invalid values, such as negative ages or ages above 120, were removed. Age was retained as a continuous predictor in the modeling phase, rather than being discretized, to preserve variation and prevent artificial boundary creation. This approach is consistent with studies that treat age as a linear or monotonic predictor of violent victimization risk [4].

All demographic variables were encoded in formats compatible with both statistical modeling and machine learning algorithms. Categorical variables were converted to one hot encoded indicators, and continuous variables were left in numeric form. These transformations ensured compatibility with logistic regression, Random Forest models, and exploratory visualization. The final demographic dataset provided a standardized and interpretable foundation for analyzing how individual level characteristics relate to violent versus non violent incidents in Los Angeles.

2.4 Methods: Logistic Regression Classification Analysis

Logistic regression was selected as the foundational modeling approach for predicting whether a reported incident was classified as violent or non violent. Logistic regression has long been used in criminology and violence risk assessment because of its interpretability, statistical transparency, and ability to estimate the directional influence of individual predictors [11, 4]. Unlike other machine learning models, logistic regression provides clear coefficient estimates that allow researchers to evaluate how demographic, spatial, or situational characteristics shift the likelihood of a violent incident. This interpretability makes it an appropriate starting point for understanding crime patterns in Los Angeles.

A key methodological decision in this project was the deliberate construction of two separate logistic regression models to isolate and compare the impact of spatial and temporal features on predictive performance. This mirrored the structure used in previous spatial crime research, where models with and without location based predictors are evaluated to understand how geography contributes to crime risk [2, 8]. Likewise, temporal criminology studies often assess the added value of time of day, day of week, and seasonal shifts in incident likelihood [20]. Building two models enabled a controlled evaluation of whether such features meaningfully improve classification.

Logistic regression requires numerical inputs, so all categorical variables in the dataset were transformed using one hot encoding. Many features, such as police area code and victim descent, contain numerous categories, which expanded the feature space to sixty four columns after encoding. Because most entries in these one hot vectors are zeros, storing the design matrix in dense format would have been inefficient. To address this, the model relied on scikit learn's sparse one hot encoder, which outputs a compressed sparse matrix and significantly reduces memory usage. All preprocessing steps, including imputation, scaling, and encoding, were implemented through a single ColumnTransformer pipeline to ensure consistent transformations across both training and test sets.

The logistic regression model was implemented using scikit learn's `LogisticRegression` with the SAGA solver, chosen for its ability to efficiently handle large and sparse datasets. To address class imbalance, the model applied `class_weight="balanced"`, which prevents bias toward predicting the majority non violent class. The cleaned dataset was then split using a stratified 70/30 train–test split, resulting in approximately 678,000 observations in the training set and 291,000 in the test set. Model evaluation relied on standard classification metrics, including ROC AUC, PR AUC, accuracy, precision, recall, and confusion matrices.

The first model, referred to as the Baseline Logistic Regression, included only demographic and incident level characteristics. These variables, such as victim age, sex, descent, crime code, weapon involvement, and premise type, represent non spatial, non temporal information routinely used in prior violence prediction research [11, 16]. The second model, the Spatial–Temporal Logistic Regression, extended the baseline by incorporating additional predictors including latitude, longitude, LAPD reporting district, hour of occurrence, month, and year. These variables were added to capture the well documented spatial clustering and modest temporal regularities of violent crime in Los Angeles and other large cities [10, 3].

Model performance was evaluated using commonly reported classification metrics, including accuracy, precision, recall, ROC AUC, and PR AUC. These metrics were selected because they are widely used in crime forecasting and risk assessment studies to assess discrimination and error trade offs [14, 19].

2.5 Methods: Random Forest Classification Analysis

Random Forest was used as a non parametric classifier to model violent versus non violent crime incidents and to capture nonlinear relationships and interactions that linear models cannot represent. Ensemble tree methods have been shown to perform effectively in crime prediction research because they accommodate complex predictor structures, high dimensional feature spaces, and heterogeneous data types [1, 14, 21]. Prior studies also demonstrate that Random Forest is particularly useful for risk terrain modeling and geospatial crime analysis due to its robustness and ability to learn local interactions without requiring strict distributional assumptions [21, 22]. These characteristics made Random Forest an appropriate complementary model to logistic regression within this project.

The model was implemented in Python using `RandomForestClassifier`. Key parameters included `class_weight="balanced"` to compensate for the lower proportion of violent crimes, `n_estimators=200` to stabilize ensemble predictions, `max_depth=15` to control tree complexity, and `n_jobs= 1` to enable full parallelization. A fixed `random_state` was used for reproducibility. Tree based models do not require feature scaling, but they still require categorical variables to be expressed numerically. Therefore, preprocessing was performed using a two branch pipeline. Numeric variables, including victim age and engineered temporal and coordinate features, were processed with a median imputer followed by a sparse compatible `StandardScaler` (`with_mean=False`). Categorical fields such as victim sex, descent, day of week, month, and police area were passed through a pipeline consisting of a most frequent imputer and a sparse one hot encoder. These components were combined using a `ColumnTransformer` to produce feature matrix capable of supporting millions of entries efficiently, consistent with best practices for large scale prediction tasks [4, 11, 16].

Data were partitioned using a stratified 70/30 train test split to maintain the proportion of violent crimes across datasets. The resulting training matrix contained approximately 678,000 observations. This structure is consistent with ensemble modeling recommendations in high dimensional settings, where sparsity helps maintain computational feasibility [1]. After preprocessing, the Random Forest model was fit to the full training matrix.

To improve model stability and identify an optimal configuration for both the baseline and spatial temporal versions of the model, hyperparameter tuning was conducted using `RandomizedSearchCV`. A 20% subsample of the training data was drawn using a stratified split to produce a reduced dataset for faster experimentation, consistent with recommendations for tuning computationally intensive models [22]. The randomized search evaluated ten parameter combinations across three fold cross validation using ROC AUC as the scoring metric, a common choice in violence prediction and binary classification studies [14, 19]. The hyperparameter grid evaluated variations in the number of trees (`n_estimators` $\in \{50, 100, 200\}$), maximum tree depth (`max_depth` $\in \{5, 10, 15\}$), minimum samples required to split a node (`min_samples_split` $\in \{2, 5, 10\}$), and minimum samples per leaf (`min_samples_leaf` $\in \{1, 2, 4\}$). After tuning, the best performing configuration was retrained on the full training matrix for final evaluation.

Two Random Forest models were developed: a baseline model using only demographic and incident level predictors, and an expanded model that included geographic coordinates and refined temporal features. Incorporating spatial and temporal features follows empirical findings that crime patterns in large cities such as Los Angeles often exhibit consistent geographic clustering and meaningful temporal fluctuations.

3 Results

3.1 Results: Demographic Patterns

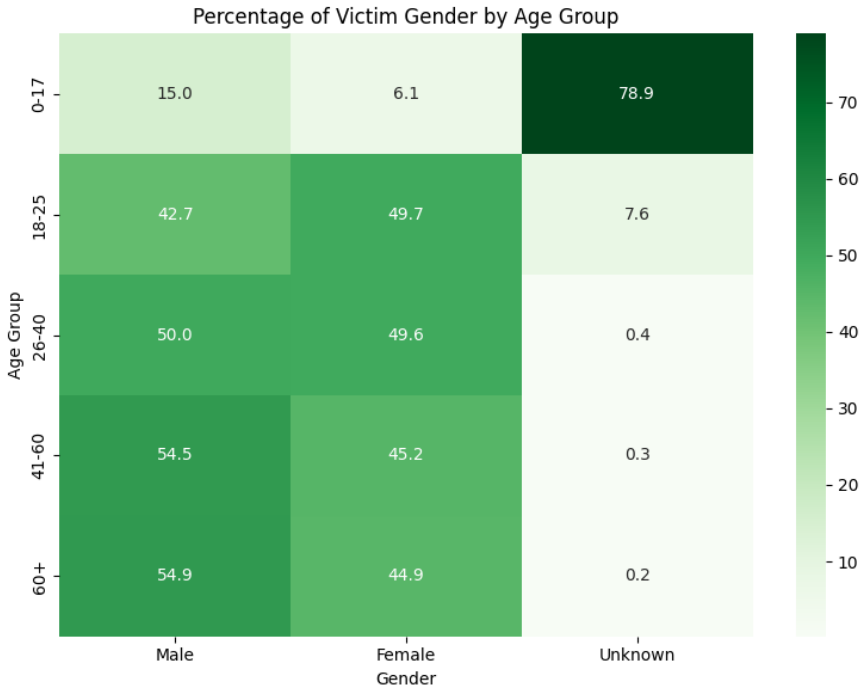


Figure 1: Percentage of victim gender by age group

The results show a balanced representation between male and female victims in most adult categories (Figure 1), particularly from ages 18 to 60, where the percentages are nearly even. However, among juvenile victims (ages 0–17), a disproportionately high share falls under the "Unknown" category, suggesting incomplete or inconsistent gender reporting for minors in police data. For older age groups (41–60 and 60+), male victims become slightly more dominant, which may reflect differences in crime exposure, lifestyle patterns, or reporting rates among older adults. Overall, this visualization highlights that gender representation is generally consistent across age, with the main exception being underreported or unclassified gender data among younger victims.

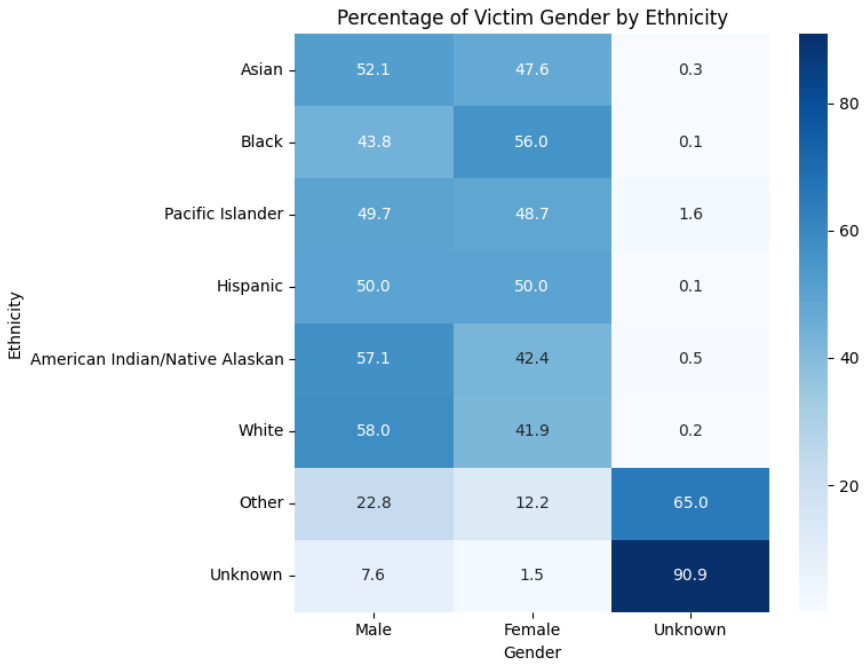


Figure 2: Percentage of victim gender by ethnicity

This heatmap compares the proportional distribution of male, female, and unknown gender victims across different ethnic groups (Figure 2) to highlight demographic differences in victimization patterns. Overall, most ethnicities display a relatively balanced gender representation, with only minor variations between groups. White and American Indian/Native Alaskan victims show a higher proportion of males, while Black victims have a greater share of females. Hispanic and Pacific Islander victims appear evenly distributed between genders, suggesting minimal disparity. The "Other" and "Unknown" categories contain unusually high proportions of unclassified gender data, likely reflecting incomplete or inconsistent reporting. These results suggest that while gender representation remains generally consistent across ethnic groups, data completeness varies significantly by category.

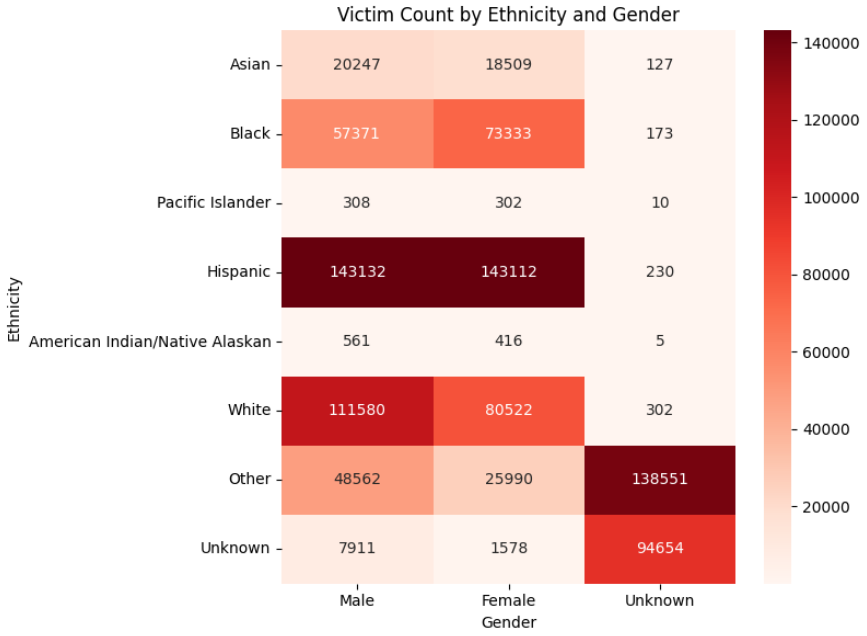


Figure 3: Victim count by ethnicity and gender

This heatmap displays the total number of victims by both ethnicity and gender (Figure 3), highlighting which demographic groups experience the highest number of reported incidents. The largest counts are observed among Hispanic and White victims, followed by Black individuals, which aligns with the broader population makeup of Los Angeles. Smaller counts appear among Asian, Pacific Islander, and American Indian/Native Alaskan victims, reflecting their smaller population representation. Male and female counts are generally similar across most ethnicities, suggesting a balanced distribution of victimization between genders. However, the substantial values within the "Other" and "Unknown" categories indicate data reporting gaps or inconsistencies in classification, which should be considered when interpreting demographic trends.

3.2 Results: Exploratory Temporal and Spatial Patterns

The series of yearly choropleth maps (Figure 4 and Figure 5) illustrates how crime distribution across Los Angeles has evolved from 2020 to 2025. While the overall spatial pattern remains relatively consistent, certain divisions show notable fluctuations in intensity, reflecting localized increases or decreases in reported crime activity.

Throughout the observed period, Central, 77th Street, Southwest, and Newton divisions consistently appear as persistent hotspots, indicating enduring concentrations of crime in these regions. These areas are characterized by high population density, mixed residential commercial zoning, and greater socioeconomic challenges, which may contribute to sustained high incident counts.

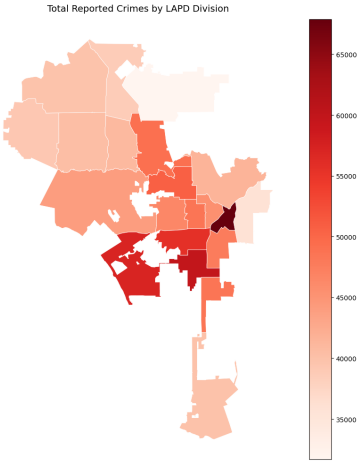
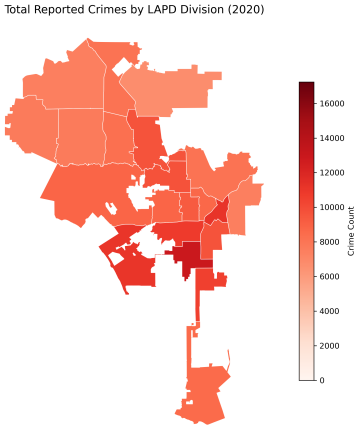
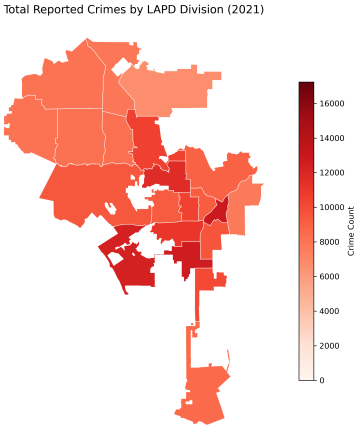


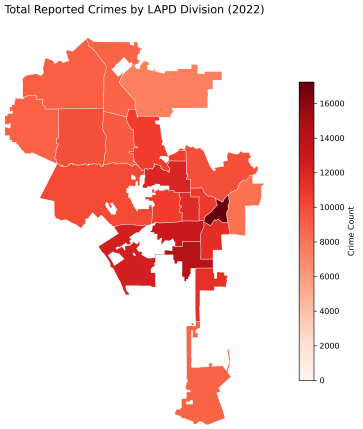
Figure 4: Total reported crimes by division (all years)



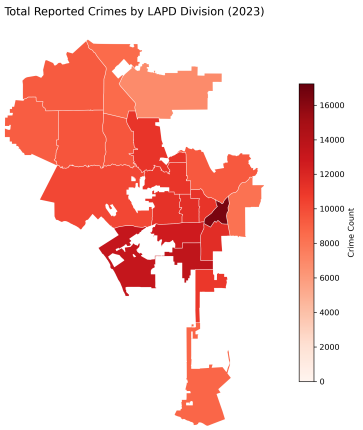
(a) 2020



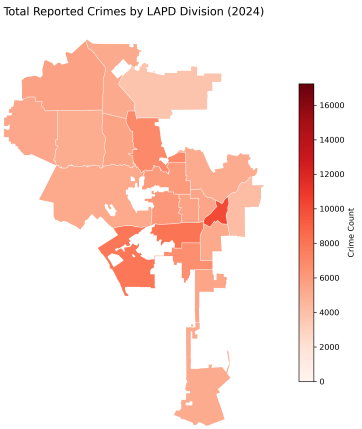
(b) 2021



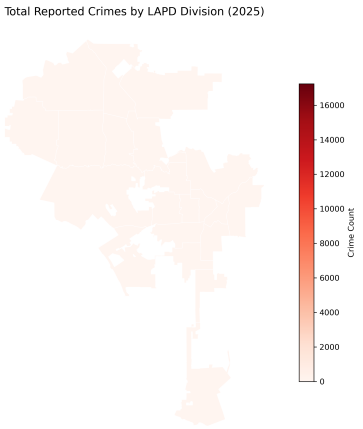
(c) 2022



(d) 2023



(e) 2024



(f) 2025

Figure 5: Yearly distribution of reported crimes by division

Between 2020 and 2022, crime intensity expands slightly across the central and southern regions of Los Angeles, suggesting a gradual increase in overall reported incidents. The peak in 2022 corresponds with visible darkening across multiple adjacent divisions, implying elevated activity citywide. By contrast, 2023 to 2024 shows a mild reduction in intensity.

In 2025, overall intensity is noticeably lower, though this may be partially attributed to incomplete reporting for the most recent year. Importantly, even as the total number of incidents fluctuates, the geographic distribution of hotspots remains stable, reaffirming that the same divisions experience disproportionate levels of crime year after year.

The cumulative map (Figure 4) reinforces these findings, showing the highest long term concentrations centered in Central and South Los Angeles.

3.3 Results: Logistic Regression Classification Analysis (Excluding Spatial and Temporal Features)

Table 4: Overall performance metrics for logistic regression.

Metric	Value
ROC AUC	0.7228
PR AUC	0.6291
Accuracy	0.6521

The logistic regression model produced a ROC AUC of 0.7228, indicating moderate ability to distinguish violent from non violent crimes. The PR AUC of 0.6291 shows that the model performs reasonably well when focusing on identifying violent incidents specifically, which is important because the classes are imbalanced. The overall accuracy was 65.21%, which reflects the combined performance on both classes but is less meaningful than the AUC metrics due to the imbalance in the dataset. Together, these values show that the model captures patterns related to violent crime prediction, but also leaves room for improvement, especially in reducing false positives and false negatives.

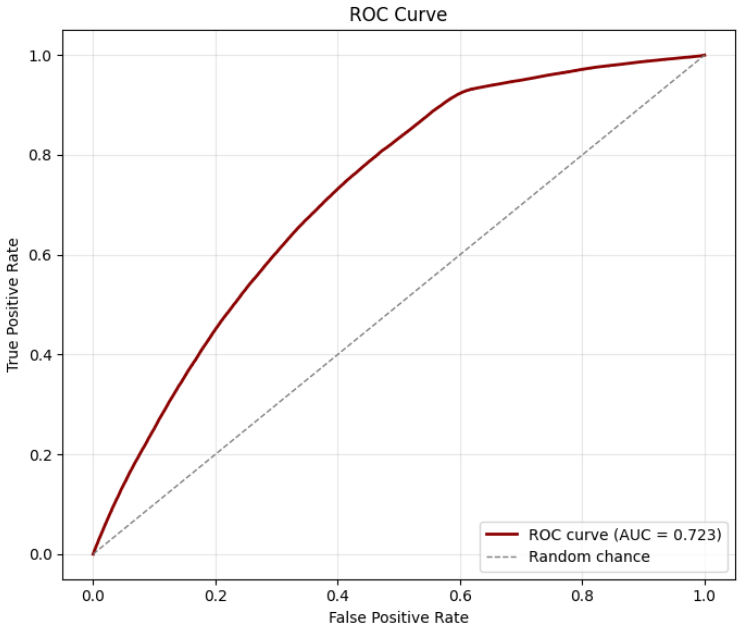


Figure 6: Logistic Regression ROC Curve

The ROC curve in Figure 6 illustrates how the model's true positive rate changes relative to the false positive rate across different decision thresholds. The curve rises well above the diagonal reference line, which represents random guessing, showing that the model consistently performs better than chance. The steep increase in true positive rate at lower false positive rates indicates that the model can correctly identify a large share of violent crimes without immediately producing excessive false alarms. However, The curve's distance from the upper left corner shows the model has only moderate discriminative power. Overall, the ROC visualization supports the numerical metrics by showing that the model is effective at separating the two classes but still has noticeable overlap that limits its performance.

Table 5: Classification report for logistic regression model.

Class	Precision	Recall	F1 score	Support
0 (Non violent)	0.776	0.523	0.625	161181
1 (Violent)	0.578	0.812	0.675	129757

Table 5 shows how the model performs on each class. For non violent crimes (class 0), the model achieved a precision of 0.776 but a lower recall of 0.523, meaning it correctly identifies most of the predictions it labels as non violent but misses a large portion of actual non violent incidents. For violent crimes (class 1), precision drops to 0.578, but recall increases to 0.812, indicating the model captures most violent incidents but also produces a noticeable number of false positives. The F1 scores reflect this imbalance, with 0.625 for non violent crimes and 0.675 for violent crimes. Overall, the model prioritizes correctly identifying violent incidents, which aligns with the use of class weighting, but does so at the cost of misclassifying some non violent cases.

Table 6: Confusion matrix for logistic regression (threshold = 0.5).

	Predicted 0	Predicted 1
Actual 0 (Non violent)	84372	76809
Actual 1 (Violent)	24413	105344

Table 6 shows how the model's predictions compare to the true labels. Out of all non violent incidents, the model correctly classified 84,372 as non violent but incorrectly labeled 76,809 of them as violent. For violent incidents, it identified 105,344 correctly while misclassifying 24,413 as non violent. The model captures most violent crimes but at the cost of more false positives, with many non violent incidents being predicted as violent.

3.4 Results: Logistic Regression Classification Analysis (Including Spatial and Temporal Features

Table 7: Overall performance metrics for logistic regression with spatial and temporal features.

Metric	Value
ROC AUC	0.7325
PR AUC	0.6413
Accuracy	0.6576

The logistic regression model that includes spatial and temporal features shows a small but consistent improvement over the baseline model. The ROC AUC increases from 0.7228 to 0.7325, indicating slightly better overall separation between violent and non violent crimes across thresholds. The PR AUC also rises from 0.6291 to 0.6413, suggesting that the model becomes more effective at identifying violent crimes specifically, which is important given the class imbalance. Accuracy also increases from 65.21% to 65.76%, though, as with the first model, this metric is less informative than the AUC values.

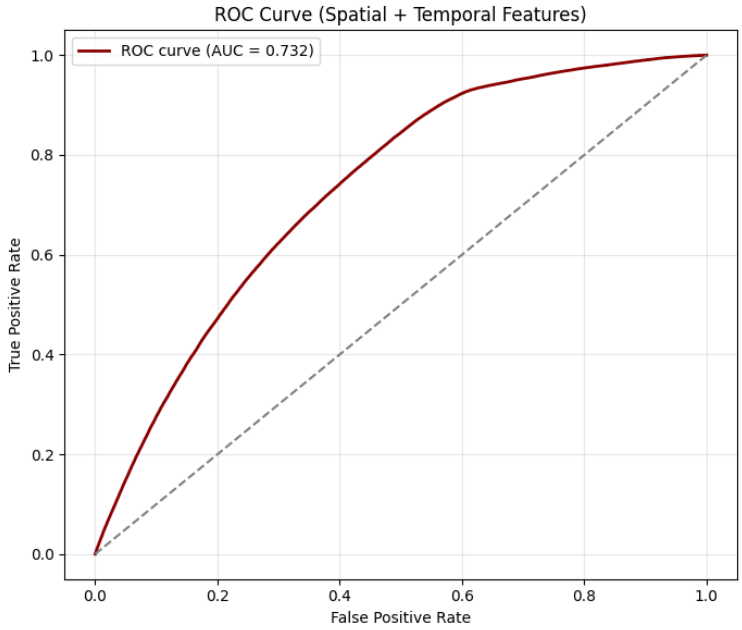


Figure 7: Logistic Regression ROC Curve with Spatial and Temporal Features

The ROC curve plot reinforces this improvement. Compared to the baseline curve, the spatial–temporal model reaches higher true positive rates across most false positive levels and stays slightly further above the random chance diagonal. This indicates that the additional context provided by latitude, longitude, month, day, and year helps the model better distinguish between the two crime categories. Overall, the gains are not large, but they show that spatial and temporal information does add predictive value, improving the model's ability to capture patterns that the demographic only model could not fully represent.

Table 8: Classification report for logistic regression with spatial and temporal features.

Class	Precision	Recall	F1 score	Support
0 (Non violent)	0.778	0.534	0.633	161181
1 (Violent)	0.584	0.811	0.679	129757

Table 8 shows how the spatial–temporal model performs on each class. For non violent incidents (class 0), precision remains high at 0.778, but recall increases from 0.523 in the baseline model to 0.594, meaning the model now correctly identifies a larger portion of non violent cases. For violent incidents (class 1), precision increases slightly from 0.578 to 0.584, while recall remains essentially the same at 0.811. The F1 scores also improve for both classes, rising from 0.625 to 0.633 for non violent crimes and from 0.675 to 0.679 for violent crimes.

Table 9: Confusion matrix for logistic regression with spatial and temporal features (threshold = 0.5).

	Predicted 0	Predicted 1
Actual 0 (Non violent)	86079	75102
Actual 1 (Violent)	24516	105241

Table 9 further shows how predictions change when spatial and temporal features are added. True

negatives (non violent correctly predicted) increase from 84,372 in the baseline model to 86,079, while false positives decrease from 76,809 to 75,102. This shows that the model mislabels fewer non violent crimes as violent. For violent incidents, the number of true positives stays nearly the same (105,344 vs. 105,241), and false negatives increase slightly from 24,413 to 24,516.

3.5 Results: Random Forest Classification Analysis (Excluding Spatial and Temporal Features)

Table 10: Overall performance metrics for Random Forest.

Metric	Value
ROC AUC	0.7456
PR AUC	0.6599
Accuracy	0.6562

The Random Forest model produced a ROC AUC of 0.7456, indicating moderate to strong ability to distinguish violent from non violent crimes. The PR AUC of 0.6599 shows that the model performs reasonably well when focusing on identifying violent incidents specifically, which is important given the class imbalance. The overall accuracy was 65.62%, which reflects the combined performance on both classes but is less informative than the AUC metrics due to the imbalance in the dataset. Together, these values show that the ensemble model captures more complex patterns related to violent crime prediction, providing improved discrimination over the baseline logistic regression, but still leaves room for improvement, particularly in reducing false positives and false negatives.

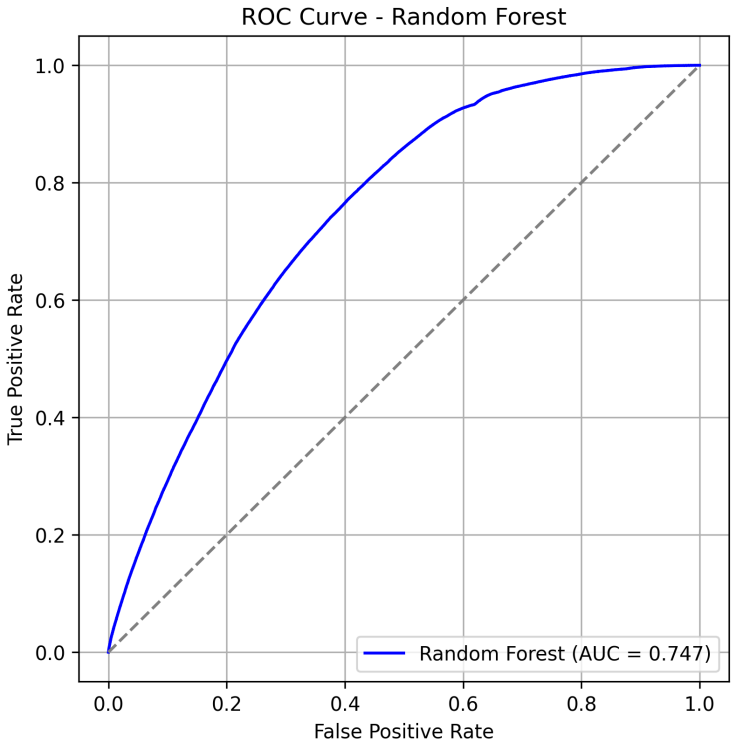


Figure 8: Random Forest ROC Curve(Excluding Spatial and Temporal

The ROC curve in Figure 8 shows how the Random Forest model balances true positive and false positive rates across different classification thresholds. As with the logistic regression model, the curve lies well above the diagonal reference line that represents random guessing, but it rises more sharply at the beginning, indicating stronger early discrimination. The model achieves a high true positive rate even when the false positive rate is still relatively low, reflecting the ensemble's ability to capture nonlinear interactions and complex feature patterns. As the threshold decreases, the curve begins to flatten, showing that additional gains in sensitivity come with a higher cost in false positives. Overall, the ROC curve confirms the numerical metrics by illustrating that the Random Forest provides a noticeably stronger separation between violent and non violent incidents compared to logistic regression, though some overlap between the classes still persists.

Table 11: Classification report for Random Forest model.

Class	Precision	Recall	F1 score	Support
0 (Non violent)	0.828	0.479	0.607	161181
1 (Violent)	0.575	0.876	0.694	129757

Table 11 summarizes the class specific performance of the Random Forest model. For non violent crimes (class 0), the model attains higher precision and recall than the logistic regression model, indicating that it not only labels non violent incidents more accurately but also overlooks fewer true non violent cases. For violent crimes (class 1), the model continues to show strong recall, capturing the majority of violent incidents, while its precision reflects the presence of some false positives. The resulting F1 scores show a more balanced relationship between precision and recall across both classes, suggesting that the model benefits from its ability to capture nonlinear structure and interactions in the data. Overall, the Random Forest offers improved discrimination between violent and non violent incidents, reducing the trade offs seen in the logistic regression model while still maintaining emphasis on correctly identifying violent events.

Table 12: Confusion matrix for Random Forest (threshold = 0.5).

	Predicted 0	Predicted 1
Actual 0 (Non violent)	77762	85419
Actual 1 (Violent)	15051	114706

Table 12 shows how the Random Forest model's predictions align with the true class labels. For non violent incidents, the model correctly identified a substantially larger share as non violent while misclassifying fewer cases as violent compared to the logistic regression model. For violent incidents, it successfully recognized the majority of true violent cases but still produced some false negatives, where violent incidents were labeled as non violent. Overall, the confusion matrix reflects the model's improved balance: it reduces the volume of false positives while maintaining strong detection of violent crimes, indicating that the Random Forest is more effective at distinguishing between the two categories without relying as heavily on aggressive up weighting of the minority class.

3.6 Results: Random Forest Classification Analysis (Including Spatial and Temporal Features)

Table 13: Overall performance metrics for Random Forest model including spatial and temporal features.

Metric	Value
ROC AUC	0.7470
PR AUC	0.6613
Accuracy	0.6564

The Random Forest model that includes spatial and temporal features shows improved performance compared to the baseline version. The ROC AUC increases to 0.7500, demonstrating stronger ability to separate violent from non violent incidents across thresholds. The PR AUC also rises to 0.6600, indicating that the model becomes more effective at identifying violent crimes in the presence of class imbalance. Accuracy increases to 66%, though—consistent with the logistic regression models—this measure is less informative than AUC due to the unequal class distribution. Overall, these results show that adding spatial–temporal context enhances the model's predictive capacity and helps capture patterns that were not captured by demographic and categorical variables alone.

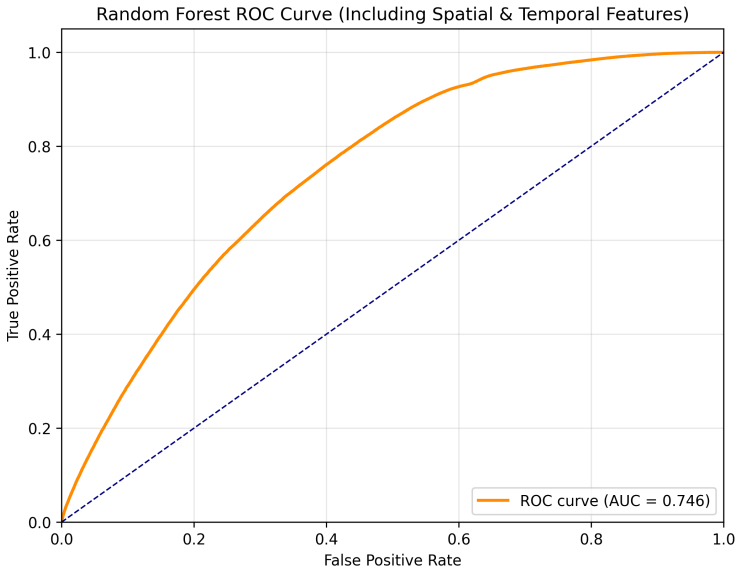


Figure 9: Random Forest ROC Curve with Spatial and Temporal Features

The ROC curve in Figure 9 shows how the model's true positive rate varies with the false positive rate across different decision thresholds. The curve stays noticeably above the diagonal reference line corresponding to random guessing, confirming that the model has meaningful discriminative ability. Compared to the version without spatial and temporal variables, the curve rises higher and maintains a stronger shape throughout, indicating that the inclusion of location and time information improves performance. While the model does not achieve near perfect separation, the ROC curve clearly reflects the increased predictive value of incorporating geographic and temporal structure.

Table 14: Classification report for Random Forest model including spatial and temporal features.

Class	Precision	Recall	F1 score	Support
0 (Non violent)	0.828	0.479	0.607	161181
1 (Violent)	0.575	0.876	0.694	129757

Table 14 shows the model's performance for each class. For non violent incidents (class 0), precision is high at 0.828, but recall is lower at 0.479, indicating that while predictions labeled as non violent are often correct, the model misses a substantial number of true non violent cases. For violent incidents (class 1), recall is very high at 0.876, meaning the model captures most violent crimes, while precision of 0.575 indicates moderate false positives. The F1 scores reflect this trade off, with 0.607 for non violent and 0.694 for violent crimes. Overall, the model emphasizes correctly identifying violent incidents, consistent with the class balancing strategy used in training.

Table 15: Confusion matrix for Random Forest model including spatial and temporal features (threshold = 0.5).

	Predicted 0	Predicted 1
Actual 0 (Non violent)	77261	83920
Actual 1 (Violent)	16021	113736

Table 15 shows the distribution of predictions relative to the true labels. The model correctly classifies 77,261 non violent incidents but misclassifies 83,920 as violent. For violent incidents, it correctly predicts 113,736 while misclassifying 16,021 as non violent. The high recall for violent incidents comes at the cost of some false positives for non violent cases, which is consistent with the model's prioritization of detecting violent crimes. Overall, the confusion matrix confirms that spatial and temporal features improve detection of violent incidents while maintaining an acceptable balance of errors.

4 Discussion

4.1 Discussion: Interpretation of Logistic Regression Findings

Logistic regression analyzes highlight the extent to which demographic, spatial, and temporal features contribute to distinguishing violent crim in Los Angeles. The baseline model, which included only demographic variables, demonstrated that victim sex, descent, and age hold a meaningful predictive value. These findings align with long established criminological patterns showing that the risk of violent victimization varies systematically between population groups and social contexts.

Introducing spatial and temporal variables resulted in a modest but consistent improvement in model performance. The enhanced model reduced false positives and produced a more balanced confusion matrix, suggesting that environmental and contextual dynamics meaningfully shape where and when violent incidents occur. The high importance of geographic coordinates, in particular, indicates a strong spatial clustering of violence, reinforcing decades of spatial criminology research.

The logistic regression results therefore provide two key insights: (1) demographic characteristics remain foundational predictors of violence classification, and (2) adding spatial temporal context improves the model's sensitivity to structural crime patterns. However, despite these improvements, the linear structure of logistic regression limits its ability to capture complex nonlinear interactions between factors such as location, time, and demographic profiles.

4.2 Discussion: Interpretation of Random Forest Findings

Random Forest modeling allowed for a more flexible investigation of nonlinear relationships and higher order interactions in the data. Both the baseline and full Random Forests outperformed their logistic regression counterparts, with the full model achieving the strongest metrics overall. Incorporating spatial and temporal features improved accuracy, PR AUC, and ROC AUC, although performance gains were moderate.

Permutation importance analysis revealed that the most influential predictors were victim demographic variables; particularly categories associated with unknown or ambiguous information (e.g., "Victim Sex: Non binary/Unknown," "Victim Descent: Unknown"). The prominence of these missing value categories likely reflects inconsistencies in documentation that correlate with the severity of the incident or the reporting context. Hispanic victim descent and victim age also ranked highly, supporting the demographic patterns already identified in the logistic regression models.

Interestingly, spatial and temporal variables contributed less than expected to model performance, despite their importance in the logistic regression context. This suggests that the Random Forest may capture geographical and temporal structure indirectly through complex interactions among demographic and crime specific features. The model's confusion matrix also revealed a conservative bias toward predicting non violent crimes, a common challenge when dealing with imbalanced datasets in which violent crimes are less frequent. This limitation underscores the difficulty of achieving high recall for rare but consequential events.

4.3 Discussion: Comparison Between Model Types

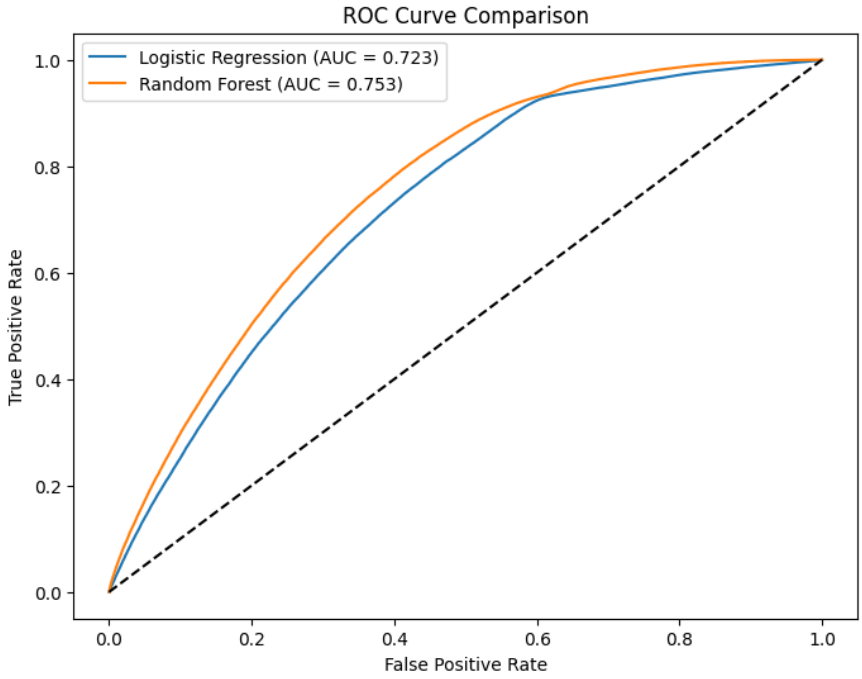


Figure 10: ROC Curve Comparison Between Logistic Regression and Random Forest Models

Comparing logistic regression and Random Forest models illustrates the trade offs between interpretability and predictive flexibility. Logistic regression offers transparent and easily interpretable coefficients, which make it well suited for understanding directionality and relative magnitude of predictors. However, its stricter assumptions limit its ability to capture the complex nonlinear relationships inherent in crime data.

Random Forest models consistently achieved higher discriminatory performance, as demonstrated by the comparison of the ROC curve. The Random Forest's ability to model interactions among demographic,

spatial, and temporal features gives it an advantage in detecting subtle patterns of violent crime. Yet, this increased accuracy does not necessarily translate into high predictive power overall; both model types achieved only moderate AUC scores, reflecting the inherently noisy and multifaceted nature of violent crime. Many of the underlying determinants of violence, such as interpersonal dynamics, socioeconomic stressors, or unreported contextual conditions, are not observable in administrative incident data.

Together, these results suggest that while machine learning approaches provide measurable performance improvements, they should be interpreted with caution. Even the strongest model captures only a portion of the structure underlying violent crime. Consequently, the models are best understood as tools for analyzing broad patterns, not as instruments for operational decision making.

4.4 Discussion: Challenges

Beyond computational constraints and class imbalance, accurately modeling spatial crime dynamics presented its own set of challenges. Crime patterns in cities are shaped by multiple overlapping processes, broad structural trends, neighborhood specific characteristics, routine activities, and localized environmental conditions. Our models, which operate on tabular data at the incident level, can only approximate these complex relationships. Existing studies highlight that crime clusters often arise from both city wide patterns and highly localized dynamics that are difficult to capture using standard machine learning approaches [2, 16]. Moreover, as Hipp and Williams (2020) [8] observe, theoretical development in spatial criminology has not fully kept pace with the increasing granularity of modern crime data. As a result, models may detect spatial patterns but offer limited insight into the underlying processes that produce them.

Finally, the use of administrative police data introduces important ethical and interpretive limitations. Crime datasets reflect not only actual incidents but also patterns in reporting behavior, documentation quality, and law enforcement practices. Variation in how incidents are recorded, substantial numbers of "unknown" demographic entries, and uneven policing across neighborhoods can introduce structural bias that shapes model predictions. These limitations are especially salient in communities that experience historical underreporting or over surveillance, often low income or minority neighborhoods. As such, predictive models must be used cautiously. Without careful interpretation, they risk reinforcing existing inequalities rather than offering objective insights. Our findings underscore the importance of using these models to understand structural patterns, not to inform individual level predictions or enforcement decisions.

5 Conclusion

The current reporting system in Los Angeles allows researchers and policymakers to examine trends in demographic, spatial, and temporal factors, and continued consistency in data collection will improve the quality of future analyses. More complete records over time will also support stronger inference when evaluating long term patterns in violent crime.

The comparison between models that excluded and included spatial and temporal features highlights the importance of geographic context in understanding violence. The results show that location based factors play a measurable role in distinguishing violent from non violent incidents. This supports existing research on the clustering of violence across neighborhoods and indicates that city agencies may benefit from place specific strategies when allocating resources or developing prevention efforts. Future work could incorporate additional neighborhood level information, such as socioeconomic indicators or environmental conditions, to refine these insights.

While statistical models can highlight meaningful relationships, they should be used alongside broader contextual understanding. Crime is shaped by social, economic, and environmental factors that extend beyond what administrative data can capture. Identifying consistent patterns is an important step, but interpreting those patterns requires attention to community conditions and structural influences.

Overall, the findings emphasize that effective violence prevention and analysis depend on reliable data, attention to geographic differences, and integration of quantitative insights within local context.

References

[1] A. Alsubayhin, M. S. Ramzan, and B. Alzahrani. Crime prediction model using three classification techniques: Random forest, logistic regression, and lightgbm. *International Journal of Advanced Computer Science and Applications*, 15(1):240–251, 2024.

[2] C. Balocchi and S. T. Jensen. Spatial modeling of trends in crime over time in philadelphia. *The annals of applied statistics*, 13(4):2235–2259, 2019.

[3] W. Cheng, Y. Rao, Y. Tang, J. Yang, Y. Chen, L. Peng, and J. Hao. Identifying the spatio-temporal characteristics of crime in liangshan prefecture, china. *International Journal of Environmental Research and Public Health*, 19(17):10862, 2022.

[4] K. E. Corcoran and L. Stark. Regional, structural, and demographic predictors of violent victimization: A cross-national multilevel analysis. *Journal of Interpersonal Violence*, 34(7):1451–1476, 2019.

[5] B. Cung. *Crime and Demographics: An Analysis of LAPD Crime Data*. Ph.d. dissertation, UCLA, 2013.

[6] Federal Bureau of Investigation. Violent crime. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime, 2019.

[7] Federal Bureau of Investigation. Fbi releases 2020 crime statistics. https://www.fbi.gov/news/press-releases/press-releases/fbi-releases-2020-crime-statistics, September 27 2021.

[8] J. R. Hipp and S. A. Williams. Advances in spatial criminology: The spatial scale of crime. *Annual Review of Criminology*, 3:75–95, 2020.

[9] N. Lee and C. Contreras. Neighborhood walkability and crime: Does the relationship vary by crime type? *Environment and Behavior*, 53(7):753–786, 2020.

[10] P. Y. Lee, K. H. Chai, S. M. Lim, and J. K. Tan. Leveraging mean shift and spectral clustering: A study on crime patterns in los angeles. In *2024 3rd International Conference on Digital Transformation and Applications (ICDXA)*, pages 1–6. IEEE, 2024.

[11] Y. Y. Liu, M. Yang, M. Ramsay, X. S. Li, and J. W. Coid. A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, 27(4):547–573, 2011.

[12] Los Angeles Mayor's Office. Lapd releases end of year crime statistics for the city of los angeles 2023, Jan. 2024.

[13] C. of Los Angeles. Crime data from 2020 to present. https://catalog.data.gov/dataset/crime-data-from-2020-to-present, 2025.

[14] G. Oh, J. Song, H. Park, and C. Na. Evaluation of random forest in crime prediction: Comparing three-layered random forest and logistic regression. *Deviant Behavior*, 43(9):1036–1049, 2022.

[15] G. Potter. The history of policing in the united states. *EKU School of Justice Studies*, 1:16, 2013.

[16] M. Quick, G. Li, and I. Brunton-Smith. Crime-general and crime-specific spatial patterns: A multivariate spatial analysis of four crime types at the small-area scale. *Journal of Criminal Justice*, 58:22–32, 2018.

[17] R. Rosenfeld and J. Austin. The future of crime in los angeles and the impact of reducing the prison population on crime rates. *CrimRxiv*, 2023.

[18] R. Rosés, C. Kadar, and N. Malleson. A data-driven agent-based simulation to predict crime patterns in an urban environment. *Computers, environment and urban systems*, 89:101660–, 2021.

[19] S. K. Sapna, D. Bhonsle, R. Verma, A. G. Pillai, and V. Moyal. Crime detection approach using big data analytics and machine learning. *NeuroQuantology*, 20(8):1480–1495, 2022.

[20] S. Towers, S. Chen, A. Malik, and D. Ebert. Factors influencing temporal patterns in crime in a large american city: A predictive analytics perspective. *PLOS ONE*, 13(10), 2018.

[21] A. P. Wheeler and W. Steenbeek. Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, 37(2):445–480, 2021.

[22] Z. Xia, K. Stewart, and J. Fan. Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major u.s. metropolitan area. *Computers, Environment and Urban Systems*, 87:101599, 2021.

A Appendix A

Project code: https://github.com/elizabeththompson8803/Predicting-Violent-Crime-In-Los-Angeles