Predictive Risk Modeling for Safety Interventions in Transportation Networks Using Spatial Crime History

Nazmus Sakib Pallab
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
npallab2@illinois.edu

Jiarui Yu
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
jiaruiy9@illinois.edu

Favour Jack
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
frjack2@illinois.edu

Muhammad Fahad Ali
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
mali19@illinois.edu

***Abstract*—Integrating personal safety into transportation and pedestrian planning requires systematic use of crime data. Information on crime location, time, and type can be analyzed to identify unsafe streets, intersections, and transit hubs, uncovering vulnerable areas in the urban network. Such insights enable engineers to propose design interventions such as reducing dead-end streets, improving pedestrian connectivity, and strategically relocating public transit drop-off points to enhance safety.**

In this study, raw crime record data will be transformed into actionable hotspot maps and predictive risk models to optimize the allocation of traffic police and patrol routes, ensuring coverage in the areas of highest need. Using advanced machine learning techniques, the study predicts crime types based on factors such as location, time of day, victim profile, and premises description. These results provide Civil Engineers and Urban Planners with evidence-based tools to prioritize infrastructure improvements and safety investments, while also identifying specific locations likely to evolve into future hotspots for proactive deployment of patrols, surveillance, and safety infrastructure.

Together, these predictive and spatial approaches are expected to enhance response efficiency and guide long-term city planning initiatives—from upgrading street lighting and redesigning public spaces to improving transit accessibility and targeting community resources—thereby strengthening the overall resilience and safety of urban infrastructure.

***Index Terms*—Transportation safety, Crime data analysis, Predictive risk modeling, Hotspot mapping, Machine learning, Urban infrastructure planning, Pedestrian safety**

I. Description of Dataset

The dataset used for this project is the ***"Crime Data from 2020 to Present"*** dataset for the City of Los Angeles, which is publicly available on DATA.GOV [1]. It is maintained and released by the Los Angeles Police Department (LAPD) as part of the city's open-data initiative, based on official crime reports filed by law enforcement officers.

The dataset is provided in CSV format and contains over 1 million rows of crime incidents. The full dataset consists of 28 columns, while our project will focus on the following 12 key attributes:

- DR_NO (Text): Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.
- Date Rptd (Floating and Timestamp): Date the incident was reported.
- DATE OCC (Floating and Timestamp): Date the incident occurred.
- TIME OCC (Text): Time the incident occurred.
- AREA (Text): Area where the incident occurred.
- AREA NAME (Text): ID of the area where the incident occurred.
- Rpt Dist No (Text): A four-digit code that represents a sub-area within a Geographic Area.
- Vict Age (Text): Age of the victim.
- Vict Sex (Text): Sex of the victim.
- Vict Descent (Text): Descent of the victim.
    - LAT (Number): Latitude coordinate of the incident.
    - LON (Number): Longitude coordinate of the incident.

This dataset provides both spatial (latitude/longitude, area, district) and socio-demographic (victim age, sex, descent) attributes, along with temporal information (date and time of crime occurrence), enabling spatial, temporal, and predictive risk modeling for transportation safety interventions.

II. Exploratory Data Analysis (EDA)

A. *Crime Type Distribution*

The Fig. 1 ranks the most common crimes reported across Los Angeles between 2020 and the present.

Vehicle-related crimes (particularly Vehicle Stolen, Burglary from Vehicle, and Theft from Motor Vehicle) dominate the dataset, together accounting for nearly half of all recorded incidents. These are followed by Battery – Simple Assault and Identity Theft, highlighting both property security and personal safety as key urban vulnerabilities.

To identify which areas experience the highest concentration of specific crimes, we generated a heatmap showing the top ten most frequent crime types across all police districts. Each cell represents the number of incidents for a given crime type within an area.

The heatmap reveals that crime intensity is not evenly distributed across the city. Property-related crimes such as Theft, Burglary from Vehicle, and Motor Vehicle Theft dominate the overall dataset but are heavily concentrated in a few areas—particularly Central, 77th Street, and Newton divisions. Conversely, violent offenses like Assault with a Deadly Weapon and Robbery cluster around Southeast and Southwest areas, indicating localized vulnerability.

The color gradients in the heatmap highlight that these high-frequency zones consistently report a wider mix of offenses than low-crime areas, suggesting persistent multi-type crime exposure. This uneven distribution underscores the importance of strategic resource allocation, as new or expanded police stations in these hotspots could improve response times and deter repeated offenses. These findings, combined with temporal and demographic analyses in later sections, provide quantitative evidence for prioritizing areas that face both high crime density and crime diversity, strengthening the case for targeted infrastructure and patrol expansion.

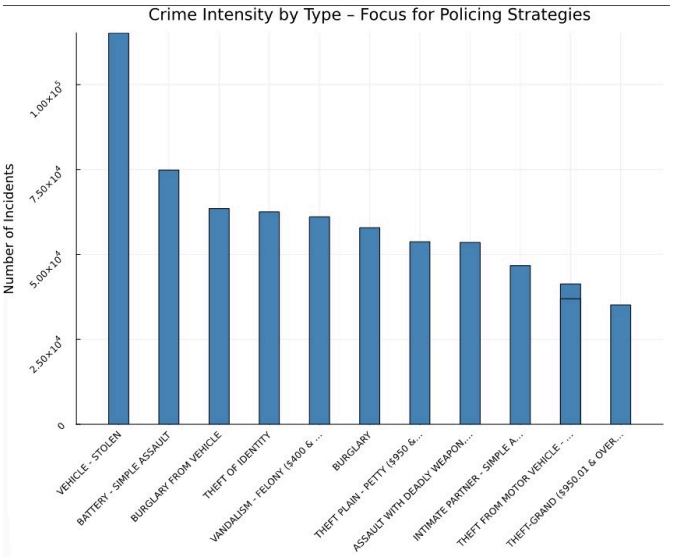

Fig. 1. Top 10 Crime Types in Los Angeles (2020–2024).

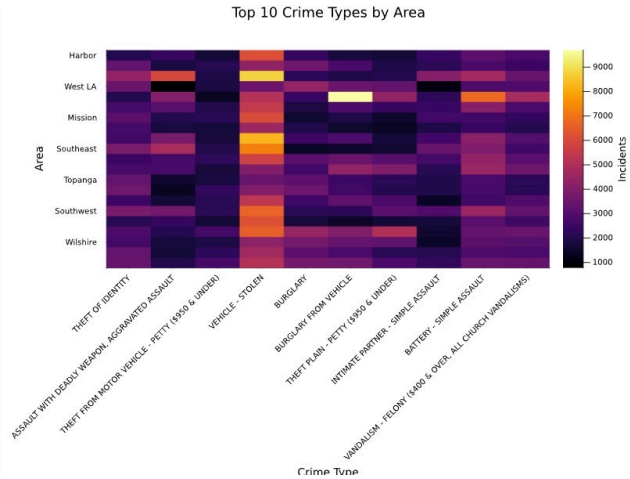

Fig. 2. Top 10 Crime Types in Los Angeles vs. Areas

B. *Temporal Analysis*

(i) Temporal Patterns of Different Crime Categories:

The temporal profile of generalized crime categories reveals that motor vehicle and bicycle theft consistently rank as the most frequent crime types, averaging 30,000–32,000 cases per year. These are followed by simple assault, personal or retail theft, and vandalism, which collectively account for a substantial share of the total crime volume. Aggravated assault and theft from vehicle form the next major tier, reflecting a stable yet diversified pattern of property and personal crimes. While total incident counts remained relatively stable from 2020–2023, a modest uptick was observed in 2022–2023, coinciding with post-pandemic normalization of urban activity. A slight decline in 2024 may partially reflect data latency or reporting lag rather than an actual reduction in crime rates.

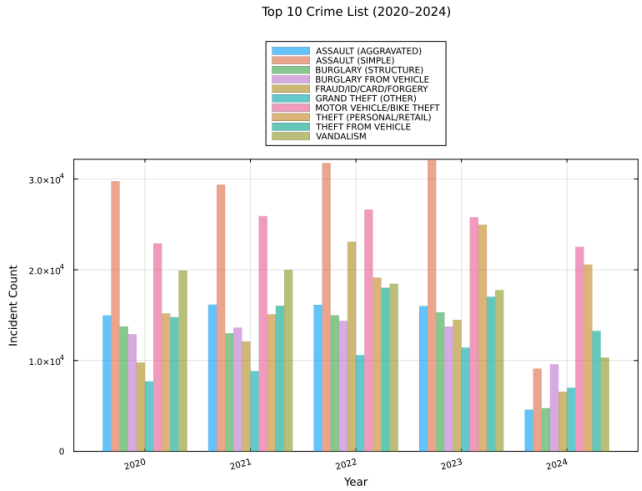

Fig. 3. Top ten generalized crime categories by year (2020–2024).

(ii) Incident Frequency Heatmap:

Incident frequency exhibits a strong diurnal and weekly rhythm, consistent with human activity cycles in an urban environment. As shown in the hourly heatmap, the lowest activity occurs during the early morning hours (03:00–

06:00), followed by a sharp increase after 08:00 that persists throughout the day. Evening hours remain active, peaking around typical commuting and social periods, and the highest overall frequencies are observed on Fridays and weekends. This pattern aligns with nightlife, leisure, and mobility trends, highlighting the influence of temporal human behavior on incident dynamics.

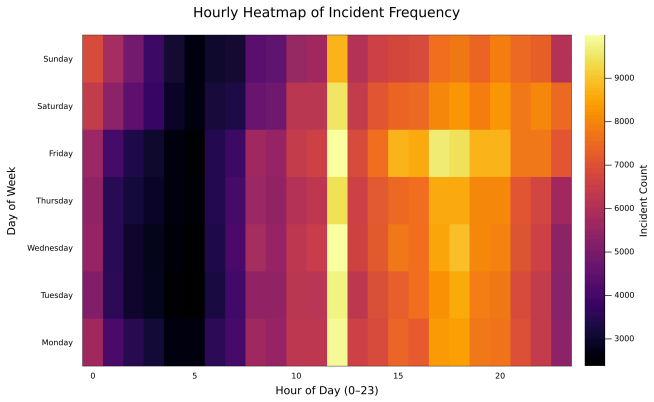

Fig. 4. Hourly heatmap showing diurnal and weekly variations in incident frequency.

(iii) Reporting Delay Distribution:

The analysis of reporting delay is defined as the difference between the date reported and the date of occurrence (Date Rptd – DATE OCC) — reveals a pronounced right-skewed distribution. Most incidents are reported either on the same day or within a few days of occurrence, with a mean delay of 2.96 days and a median delay of 1 day. Approximately 90% of all incidents are reported within five days, indicating generally prompt reporting behavior across most crime categories. The long upper tail in the delay distribution likely reflects crimes with delayed discovery or complex administrative workflows, such as fraud, forgery, or identity-theft–related offenses. This temporal asymmetry underscores the importance of considering both immediate and delayed reporting in operational planning and predictive modeling.

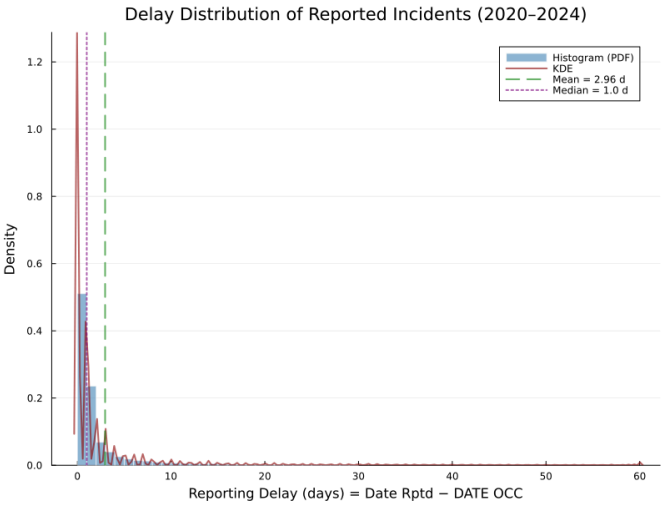

Fig. 5. Distribution of reporting delays (All recorded incidents, 2020–2024).

(iv) Spatial Distribution and Temporal Stability:

Spatially, incident clusters remain highly consistent across the five-year period, concentrating in Downtown, Hollywood, Westlake, and South Los Angeles. These areas exhibit persistent activity regardless of month or year, suggesting enduring socioeconomic and infrastructural factors driving higher incident density. The six representative panels display the months of peak activity for each year between 2020 and 2024. The scatter of points lies almost entirely within the official Los Angeles city boundary, confirming the spatial integrity and proper geocoding of the dataset. Such spatial persistence provides a reliable foundation for hotspot-based predictive modeling and targeted resource deployment.

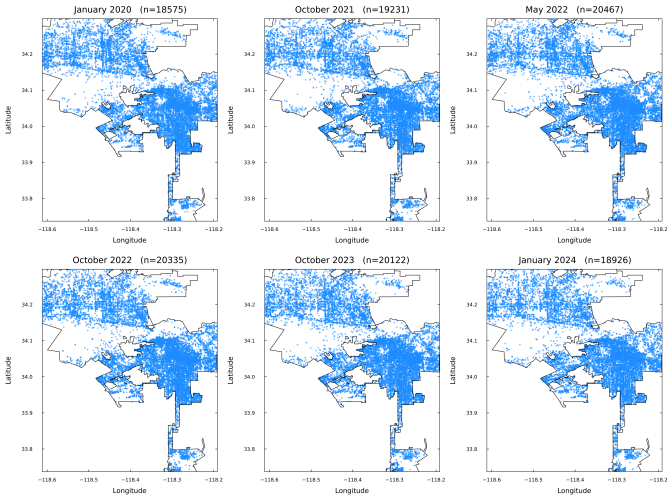

Fig. 6. Annual Peak Months of Incident Distributions (2020–2024)

C. *Spatial Analysis*

Another important aspect of our exploratory data analysis is the spatial distribution of crimes across Los Angeles. By mapping the latitude–longitude of each incident, we visualize hotspots and identify areas with high crime density.

We work with three spatial datasets: the set of crime points $C = \{C_i\}_{\{i=1\}}^{\{N_c\}}$, the set of street-lamp points $L = \{L_l\}_{\{l=1\}}^{\{N_l\}}$, and the city-boundary polygon B. We ensure longitudes and latitudes are numeric, finite, and within plausible ranges so that subsequent geometry remains meaningful. The street-light dataset is sourced from the City of Los Angeles GeoHub [2].

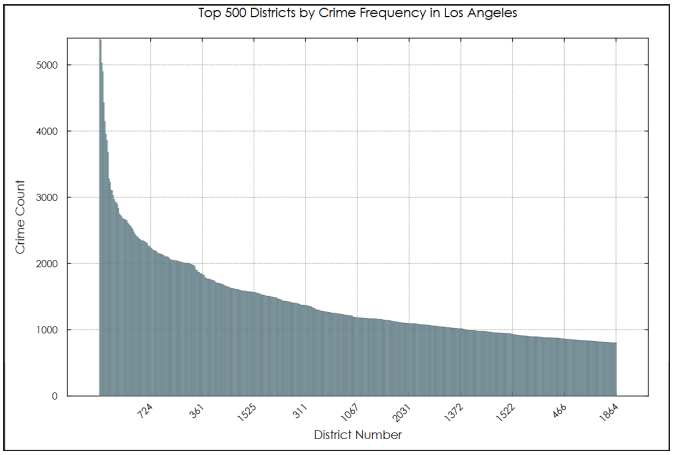

Fig. 7. Top 500 crime count in each designated LAPD district.

Figure Fig. 7 ranks Los Angeles (city) LAPD reporting districts by total reported incidents (2020–present). The x-axis lists districts in descending order (leftmost = highest), and the y-axis shows incident counts. The distribution is heavy-tailed: a few districts concentrate many incidents, followed by a long tail of moderate activity.

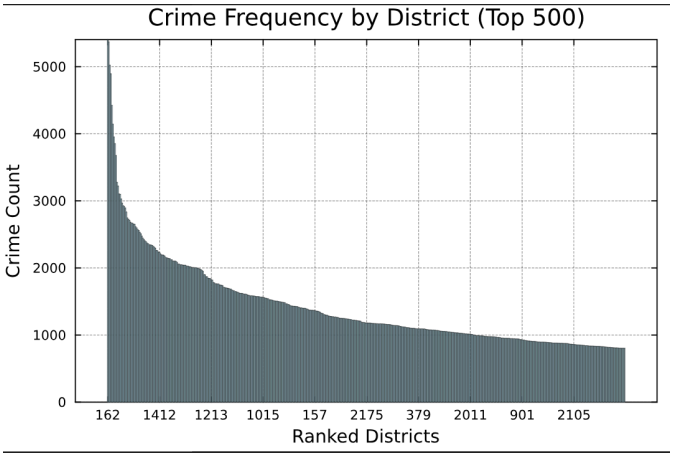

Fig. 8. LA top-500 crime counts (ranked).

Figure Fig. 8 shows the top 500 LAPD reporting districts sorted by total incidents, with the x-axis as rank (left = highest). The y-axis is the incident count. The curve is heavy-tailed: a few districts account for many incidents, followed by a long taper of moderate counts across the remaining ranked districts.

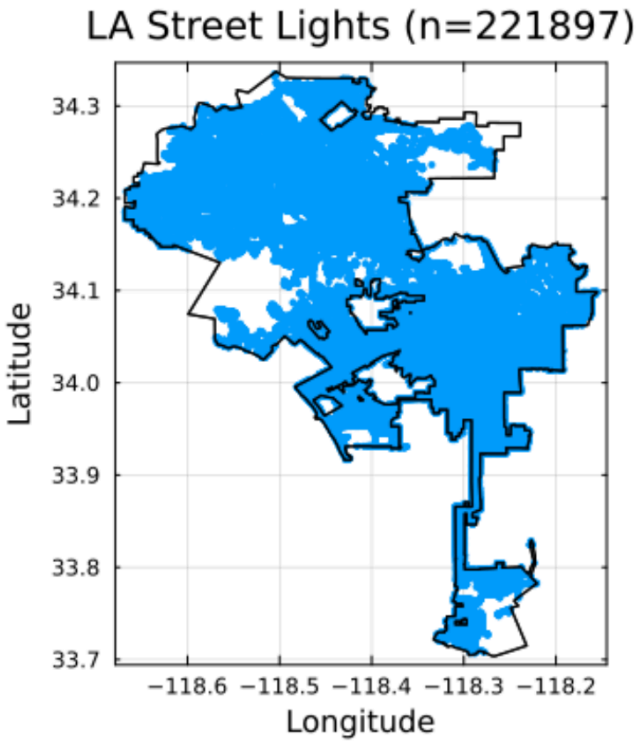

Fig. 9. LA street lights within the city boundary (n=221,897).

This map Fig. 9 plots the reported street-light point locations for Los Angeles in geographic coordinates (longitude/latitude). The high point density makes many neighborhoods appear as solid filled regions, revealing broad coverage across the city and sparser coverage near edges and open spaces. Axes are in degrees to match the source data.

a) *Finding relationship between crime locations and street lights:*

We project all coordinates into a single metric CRS so distances are measured in meters:

$((x_i, y_i) = \Phi(\lambda_i, \varphi_i),\ (u_l, v_l) = \Phi(\lambda'_l, \varphi'_l))$

where:

- (λ_i, φ_i) and (λ'_l, φ'_l) are geographic (lon/lat, degrees) for crime C_i and lamp L_l,
- Φ denotes the projection to a local metric CRS,
- (x_i, y_i) and (u_l, v_l) are projected (planar) coordinates in meters.

A local projected CRS is used because Euclidean lengths in degrees are not physically meaningful, and a single metric CRS avoids unit mismatches.

We restrict the analysis to the jurisdiction by clipping points to the city polygon B, keeping only crimes and lamps whose projected coordinates fall inside B. This removes out-of-area points and reduces boundary artifacts that would otherwise inflate nearest-distance values.

We define planar Euclidean distance between a crime and a lamp:

$$d(C_i, L_l) = \sqrt{(x_i - u_l)^2 + (y_i - v_l)^2} \qquad (1)$$

For each crime, we keep its nearest-lamp distance: $d_i = m \underset{\{1lelleN_l\}}{\in} d(C_i, L_l), q \quad i = 1, ..., n$, where n is the number

of crimes inside B. This yields a single, interpretable proximity value per incident. A spatial index keeps these queries fast.

From $\{d_i\}_{\{i=1\}}^{n}$ we build the empirical cumulative distribution function (ECDF): $F(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{d_i \le r\}, \quad r \ge 0,$ so the coverage at an operational radius R is simply: $\text{Coverage}(R) = F(R)$.

We report robust summaries of proximity (median) and tail behavior (p90, p99) along with the maximum distance. Quantiles are preferred over the mean because the distribution is typically right-skewed and outlier-sensitive.

To guard against faulty records, we flag implausible distances above a conservative cap and, if needed, compute robust z-scores using the median and MAD to identify unusual d_i values.

Finally, to see how coverage accumulates with distance, we partition r into analyst-chosen bands (e.g., 0–50–100–250 m, ...) and tabulate the share of crimes whose nearest-lamp distance falls into each band. This "coverage by band" view highlights where most gains occur (very small radii) and where diminishing returns set in as the radius grows.

Table II summarizes the distance from each reported crime in Los Angeles to its nearest street light. It shows total valid records, distribution percentiles (p25–p99), the maximum, and how many / what share fall within the working radius **R = 100 m**. Most crimes are close to a light, with a long right tail driven by a small set of far-out points. More such comparison will be carried out based upon availability of the data.

D. *Demographic Anlysis*

Besides analyzing crime types and its patterns over time and space, exmining the demographic characteristics of crime victims might provide some useful sights. We analyzed age, sex, and descent compositions of victims. Overall, the victim population is 40.19% male, 35.68% female, and 24.13% unknown or missing. The age distribution of victims is shown in Fig. 10. It shows that the age group of 30-34 has the highest number of victims, followed by the age group of 25-29. The descent distribution of victims is shown in Fig. 11. It shows that the major victim descent groups are Hispanic/ Latin/Mexican, White, and Black, with the percentages of 34.45%, 23.41%, and 15.79%, respectively.

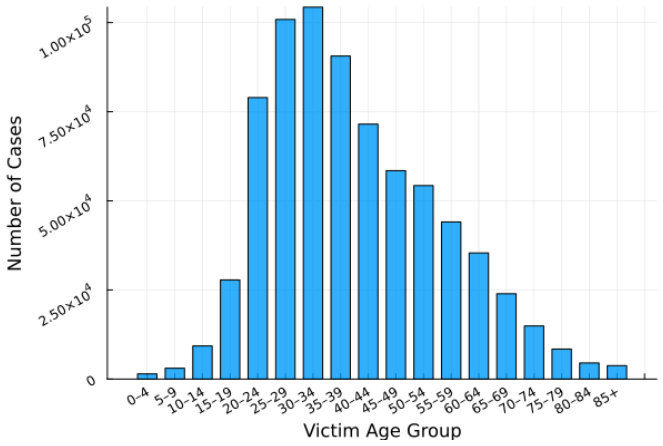

Fig. 10. Age Distribution of Crime Victims

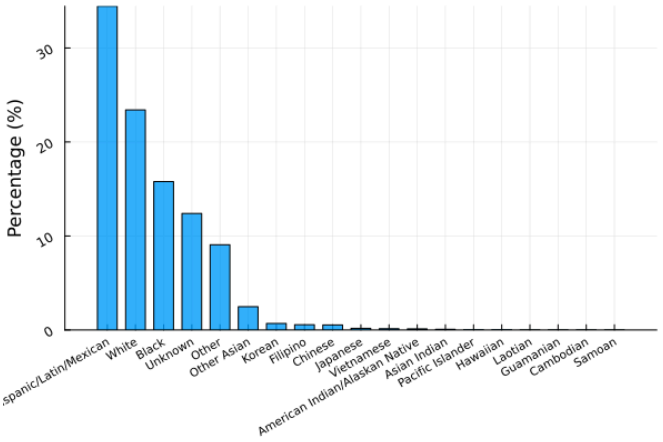

Fig. 11. Descent Distribution of Crime Victims

To see the correlation between age and descent, we created a heatmap shown in Fig. 12. The pattern of age among all descent groups is similar, with the age group of 25-34 having the highest number of victims across all descent groups.

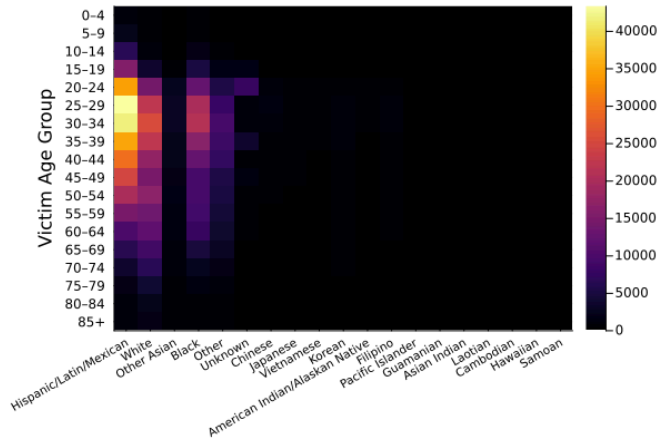

Fig. 12. Heatmap of Victim Age vs. Descent

E. *Summary Statistics of Dataset*

TABLE I
OVERVIEW OF KEY STATISTICS OF CRIME DATA FROM 2020 TO 2024

Metric	Value
Total records (2020–2024)	≈ 1,004,991 incidents
Average monthly incidents	≈ 20,000–23,000
Distinct crime types	≈ 140
Generalized crime types	37 aggregated categories
Earliest date of Dataset	2020-01-01
Latest date of Dataset	2024-03-31
Mean Hotspot for Crime (lat / lon)	34.05 / −118.32 (≈ Downtown LA)
Mean reporting delay after incident	2.96 days (mean), 1 day (median)
Victim Sex Composition	40.19% male, 35.68% female
Largest Victim Age Group	30–34 years
Top Victim Descent	Hispanic/Latin/Mexican

TABLE II
LA DISTANCE-TO-LIGHT SUMMARY.

Metric	Value
N (valid)	1,004,996
min (m)	0.0852
mean (m)	25,702.8
std (m)	5.43e3
p25 (m)	10.8821
median (m)	14.2868
p75 (m)	20.8421
p90 (m)	81.4454
p95 (m)	152.942
p99 (m)	342.602
max (m)	1.15167e7
≤R (count)	922,403
≤R (%)	91.7822

III. PREDICTIVE MODELING

A. *Problem Statement and Modeling Strategy*

The Los Angeles Police Department (LAPD) "Crime Data from 2020 to Present" dataset provides rich temporal, spatial, demographic, and contextual information about crime incidents. Our overarching goal in this deliverable is to use machine learning to understand and predict crime patterns that are relevant for urban safety, resource allocation, and transportation systems.

Our main predictive research question is:

Can we use structured temporal, spatial, demographic, and accessibility features to reliably predict (i) what kind of crime occurs, (ii) where it is most likely to occur, and (iii) whether it is vehicle-related in Los Angeles?

To answer this overarching question, we organize our predictive modeling into four related sub-questions:

1) Crime Type Prediction with Demographic and Temporal–Spatial Features – Can we predict the detailed crime type from victim demographics or basic temporal–spatial information?
2) Temporal–Spatial Crime Hotspot Prediction with Softmax Neural Network – Can we predict which spatial hotspot a crime will occur in, given when it happens and broad categorical context?
3) Spatial Crime-Risk Modeling with Random Forests and Accessibility Features – Can we predict the probability of crime at arbitrary locations using their accessibility to facilities such as mental health centers, food assistance, parks, libraries, and police stations?
4) Vehicle Crime Prediction Model – Given the temporal, spatial, and demographic context of a reported incident, can we predict whether the crime is vehicle-related?

Together, these models move from fine-grained crime-type prediction (which proves difficult) toward region-level hotspot and risk prediction (which is more successful), and finally to category-specific prediction for vehicle-related crime. Each modeling effort informs our understanding of what is and is not predictable from the available features.

B. *Crime Type Prediction with Decision Tree Classifiers*

*i. **Prediction from Demographic Features**:*

We first examined the relationship between demographic features of victims and the types of crimes that occur. The characteristics of victims include age, gender, and descent. After extracting necessary data from the original dataset, we set:

- Features: victim age, victim gender, victim descent
- Label: crime type (top 20 most frequent crime types)

Data preprocessing and encoding included:

- Gender: encoded as male and female
- Age: grouped into 9 age categories
- Descent: original categorical codes retained

- Crime type: restricted to the top 20 categories by frequency

We used an 80/20 train–test split. A standard decision tree classifier from DecisionTree.jl was trained on the training portion and evaluated on the held-out test set.

- Full feature set (age, gender, descent): test accuracy ≈ 16.14%
- Pairwise feature subsets:
  - Age + gender ⟶ 12.48%
  - Age + descent ⟶ 15.06%
  - Gender + descent ⟶ 13.45%

Given that random guessing among 20 classes corresponds to a baseline accuracy of 5%, the model performs better than random but is still far from practically useful. These results suggest that:

- Either decision trees are not well-suited to predicting detailed crime types from these features alone, or
- Victim demographics are only weakly related to the specific crime type experienced, at least in this dataset and encoding.

This negative result helps refine our focus: predicting detailed crime type from demographics alone is challenging and calls for richer context.

ii. ***Prediction from Temporal–Spatial Features****:*

Next, we tested whether crime types are more strongly related to temporal–spatial characteristics. In the dataset:

- The city is divided into 21 LAPD reporting areas, and
- The exact hour of the crime is recorded.

We used:

- Features: area index, time of occurrence rounded to the nearest hour (0–23)
- Label: top 20 crime types (same set as above)

Again, 80% of the data was used for training and 20% for testing. For the decision tree classifier, we used:

- Maximum depth = 10
- Minimum samples per leaf = 50
- Minimum samples per split = 100
- Minimum purity increase = 0.001

Under these settings, the test accuracy was:

- Test accuracy: ≈ 16.75%

We further tuned the tree hyperparameters (depth, minimum leaf size, etc.), but the accuracy remained below 17%. This indicates that:

- Either decision trees are not ideal for predicting detailed crime type from such coarse temporal–spatial features, or
- Crime types in this dataset are not strongly determined by just hour-of-day and LAPD area, at least at the level of the 20-category label space.

From a modeling perspective, these first experiments show that predicting fine-grained crime type is hard, even when combining demographics and basic temporal–spatial features. This motivated a shift toward predicting coarser spatial outcomes (hotspot regions) rather than exact crime categories.

C. ***Temporal–Spatial Crime Hotspot Prediction Model***

Motivated by urban safety and resource allocation in Los Angeles, we then focused on a different predictive task: we aim to predict which spatial crime hotspot an incident will occur in using only the time of occurrence (hour of day, day of week, and month) and basic categorical context.

This task is practically relevant because police, EMS, and traffic management agencies often plan at the level of hotspot regions, not exact crime categories, and their deployment strategies are strongly shaped by time-of-day patterns.

i. ***Spatial Clustering of Crime Locations (Unsupervised)****:*

To create spatial hotspots, we applied k-means clustering to three years of historical crime data (2020–2022), using only geographic coordinates (latitude and longitude). After experimentation, we selected:

- Number of clusters: K = 8

This choice produced meaningful and reasonably well-distributed hotspot regions across Los Angeles.

For crimes occurring in the test period (2023–2024), each incident was assigned to the nearest cluster centroid using Euclidean distance. Because the clusters were learned only from 2020–2022 data, the test incidents did not influence the cluster definitions, preventing information leakage. This step is fully unsupervised, as hotspot labels emerge solely from spatial density patterns.

ii. ***Predicting Hotspot Membership (Supervised Softmax Classifier)****:*

After generating cluster IDs from k-means, we treated the cluster index as the label for a supervised classifier. The following features were used:

- Hour of day
- Day of week
- Month
- Crime type (broad category)
- LAPD reporting area

Numeric features were normalized; categorical features were encoded consistently using training-set levels.

We then trained a softmax neural network classifier with:

- Cross-entropy loss
- Gradient descent optimization
- Accuracy and loss tracked across iterations

The dataset split was chronological:

- Training: 2020–2022
- Testing: 2023–2024

iii. ***Evaluation and Performance****:*

Model performance was:

- Training accuracy: 87%
- Test accuracy: 88%
- Baseline accuracy: 1/8 = 12.5% (random guessing among 8 clusters)

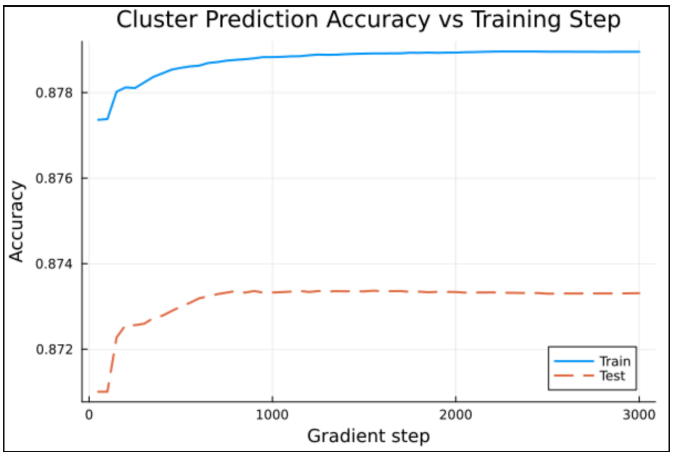

Fig. 13. Training and testing accuracy over gradient steps.

The learning curves show stable convergence with no evidence of overfitting: training and testing accuracies are closely matched across iterations.

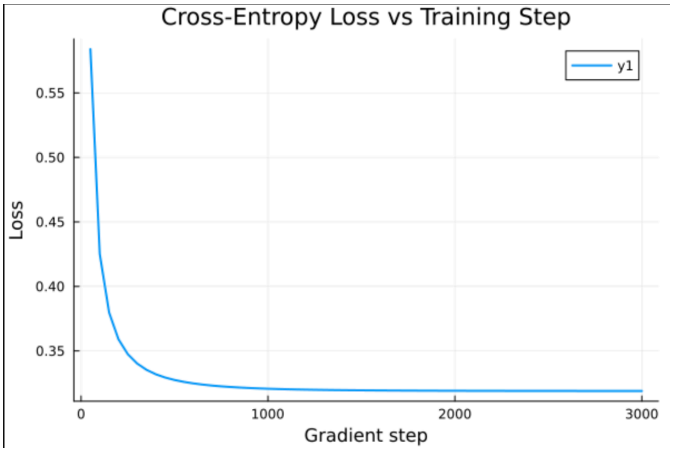

Fig. 14. Cross-entropy loss during training.

The cross-entropy loss decreases sharply during the first 300 gradient updates, indicating rapid learning of the primary temporal–spatial decision boundaries. After that, the loss continues to decline more gradually and stabilizes around ≈ 0.32, suggesting smooth convergence and a well-chosen learning rate. The absence of large spikes in the loss indicates that the model is not overfitting to rare or noisy samples.

Overall, this hotspot prediction model shows that temporal patterns (hour, day, month) combined with broad crime context and LAPD areas are highly predictive of which spatial hotspot is activated, in contrast to the much weaker predictability of detailed crime type.

iv. ***Hotspot Visualization and Interpretation****:*

To improve interpretability, we generated geographic visualizations that overlay predicted hotspot centroids on the Los Angeles city boundary shapefile for different temporal queries.

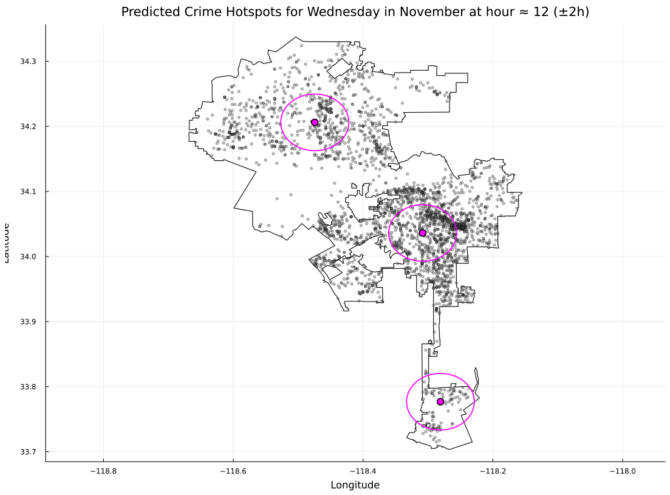

Fig. 15. Spatial distribution of predicted crime hotspots for a sample temporal query (Friday, July, 20:00 ±1 h).

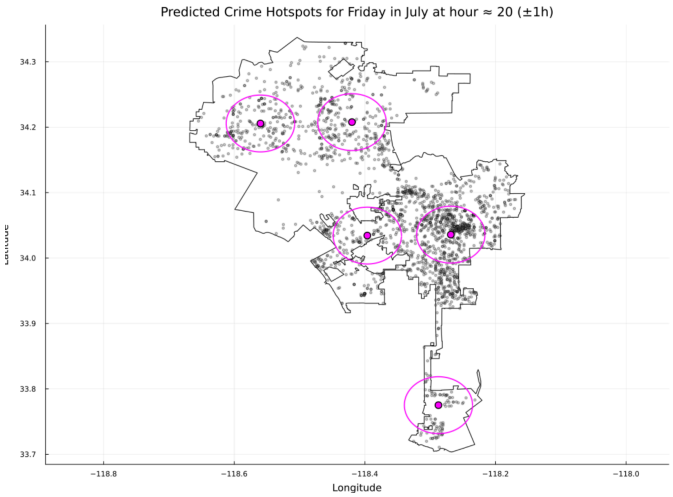

Fig. 16. Spatial distribution of predicted crime hotspots for a sample temporal query (Wednesday, November, 12:00 ±2 h).

For each user-defined query (e.g., “Fridays in July at 20:00 ± 1 hour” or “Wednesdays in November at 12:00 ± 2 hours”), the visualizations display:

- Historical crime points within the specified time window (gray)
- The LA city boundary outline
- The model’s top predicted hotspot centroids (magenta points)
- Approximately 1-mile radius highlight zones around each centroid (magenta circles)

These visualizations demonstrate how predicted hotspot regions shift with time-of-day and season, providing interpretable and actionable spatial insights for planning patrols and other safety resources.

v. ***Future Extensions for Hotspot Modeling****:*

Potential improvements include:

- Incorporating exogenous data such as weather, special events, and traffic volume

- Trying alternative clustering methods (e.g., DBSCAN, Gaussian Mixture Models) for more flexible hotspot shapes
- Exploring deeper neural networks or ensemble methods for the supervised stage
- Deploying the model as an interactive tool to support real-time decision-making

D. *Spatial Crime-Risk Modeling with Random Forests and Spatial Accessibility*

The hotspot model predicts membership in a discrete set of regions. As a complementary approach, we next asked:

Can we predict the probability of crime at arbitrary points in space using spatial accessibility to public facilities (e.g., mental health centers, food assistance providers, parks, libraries, and police stations)?

Here, the goal is to move from cluster-level predictions to a continuous crime-risk surface over Los Angeles.

1. ***Data, Notation, and Study Region:***

1.1 Crime and Non-Crime Points:

Let each location be represented by geographic coordinates:

$$p_i = (\text{lat}_i, \text{lon}_i) \in \mathbb{R}^2, i = 1, ..., N. \tag{2}$$

Each point has a binary crime label $y_i \in \{0, 1\}$, where:

- $y_i = 1$ if a crime is observed at p_i,
- $y_i = 0$ if the point is treated as a non-crime location.

The final analysis dataset contains:

- $N = 2,009,982$ total records,
- 1,004,991 crime locations ($y_i = 1$),
- 1,004,991 non-crime locations ($y_i = 0$),

so that the crime proportion is:

$$\frac{1}{N} * \sum_{i=1}^{N} y_i = 0.50 \quad (50\% \text{ crimes, } 50\% \text{ non-crimes}). \tag{3}$$

1.2 Generating Non-Crime Locations:

The original crime dataset provides only locations where crime occurred. To train a binary classifier, we need contrasting "non-crime" points. The notebook generates these using a mixture of two strategies:

1) **Nearby perturbations** around crime points (simulating "similar" areas where crime did not happen).
2) **Random points** uniformly spread across the city's bounding box.

Let N_c be the number of original crime locations. We create:

- $N_{\text{near}} = \lfloor 0.7 N_c \rfloor$ nearby points, and
- $N_{\text{rand}} = N_c - N_{\text{near}}$ random points.

For a subset of crime locations p_i, nearby non-crime points are generated as:

$$p_i^{(\text{near})} = p_i + \Delta p_i, \tag{4}$$

where Δp_i is a random perturbation with typical magnitude

$$\|\Delta p_i\| \approx 0.01°, \tag{5}$$

corresponding to roughly 1.1 km in latitude.

Random points are sampled uniformly inside the geographic bounds:

$$\text{lat}_{\min} \le \text{lat} \le \text{lat}_{\max}, \quad \text{lon}_{\min} \le \text{lon} \le \text{lon}_{\max}. \tag{6}$$

1.3 Geographic Coverage:

The combined dataset spans:

- Latitude range: 0.0° to 34.3343°,
- Longitude range: −118.6676° to 0.0°.

Using the standard "111 km per degree" approximation:

- North–south size:

$$L_{\text{NS}} \approx (\text{lat}_{\max} - \text{lat}_{\min}) * 111 \ \text{ km} \approx 3811.1 \ \text{ km} \tag{7}$$

- East–west size:

$$L_{\text{EW}} \approx (\text{lon}_{\max} - \text{lon}_{\min}) * 111 \cos(|(\text{lat})) \ \text{ km} \approx 12585.3 \text{ km}, \tag{8}$$

where $|(\text{lat})$ is the mean latitude. This is a conceptual bounding box; the actual points are concentrated near Los Angeles.

2. ***External Spatial Datasets and KD-Trees:***

The notebook loads several facility layers, each represented as a set of points:

- Mental health centers,
- Food assistance providers,
- Public libraries,
- County parks,
- City parks,
- Metro lines / stations (where available),
- Police stations.

For each facility type k, we denote its locations as:

$$F^k = \left\{ f_j^k \in \mathbb{R}^2 : j = 1, ..., M_k \right\}. \tag{9}$$

To efficiently query distances and neighbors, each F^k is indexed with a KD-tree using NearestNeighbors.jl. This enables:

- **Nearest neighbour**:

$$f_{\text{nn}}^k(p) = \text{argmin}_{f \in F^k} \| \, p - f \, \|_2, \tag{10}$$

- **Radius search**:

$$\left\{ f \in F^k : \| \, p - f \, \|_2 \le r \right\}. \tag{11}$$

3. ***Spatial Accessibility Features:***

For each point p_i and each facility type k, two feature families are computed:

1) **Nearest distance**:

$$d_i^k = \min_{1 \le j \le M_k} \| \, p_i - f_j^k \, \|_2. \tag{12}$$

2) **Count within radius r**:

$$c_i^k(r) = \sum_{j=1}^{M_k} \mathbf{1}\Big(\| \, p_i - f_j^k \, \|_2 \le r \Big), \tag{13}$$

where $\mathbf{1}(\cdot)$ is an indicator function.

The implementation uses:

- knn(tree, query_coords, 1, true) for nearest distances,
- inrange(tree, query_point, radius, false) to count facilities within the radius.

We choose:

$$r = 0.01^\circ \approx 1.1 \ \text{ km}, \tag{14}$$

interpreted as a local neighborhood scale.

For seven facility types, we obtain:

- 7 nearest-distance features,
- 7 radius-count features,

for a total of 14 spatial features:

$$p = 14. \tag{15}$$

The feature vector for point p_i is:

$$x_i = \left(d_i^{\text{mh}}, c_i^{\text{mh}}, d_i^{\text{food}}, c_i^{\text{food}}, ..., d_i^{\text{police}}, c_i^{\text{police}}\right) \in \mathbb{R}^{14}. \tag{16}$$

4. ***Correlation Analysis and Descriptive Statistics****:*

Before modeling, we compute Pearson correlations between each spatial feature and the crime label.

Given feature vector $X = (X_1, ..., X_N)$ and labels $Y = (Y_1, ..., Y_N)$ with $Y_i \in \{0, 1\}$, the Pearson correlation is:

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{17}$$

Features with higher absolute correlation are inspected more closely. Additionally, we summarize feature means for crime vs non-crime points:

- Mean among crime points:

$$\mu_{\text{crime}}(X) = \frac{1}{N_1} \sum_{i:y_i=1} X_i, \tag{18}$$

- Mean among non-crime points:

$$\mu_{\text{no-crime}}(X) = \frac{1}{N_0} \sum_{i:y_i=0} X_i. \tag{19}$$

This helps interpret whether crime locations tend to be closer or farther from certain facilities than non-crime points.

5. ***Machine-Learning Dataset Construction****:*

We construct:

- $X \in \mathbb{R}^{N \times p}$ as the design matrix,
- $y \in \{0, 1\}^N$ as the crime indicator.

The prepare_ml_data function:

1) Selects all spatial accessibility features and any lighting features (columns containing "light").
2) Replaces Inf values with a large constant (e.g., 10^{10}).
3) Imputes NaNs with feature medians.
4) Encodes the target as "0" or "1" for DecisionTree.jl.
5) Creates a random train–test split with test ratio $\alpha = 0.2$.

With $N = 2,009,982$ and $\alpha = 0.2$:

- Training size:

$$N_{\text{train}} = 0.8N = 1,607,986 \tag{20}$$

- Test size:

$$N_{\text{test}} = 0.2N = 401,996 \tag{21}$$

Class counts:

- Train: 803,826 crimes, 804,160 non-crimes,
- Test: 201,165 crimes, 200,831 non-crimes.

Both splits are almost perfectly balanced.

6. ***Random Forest Classifier****:*

6.1 Single Decision Tree (Conceptual):

A single decision tree partitions $\mathbb{R}^p$ using axis-aligned splits of the form $x_j \leq \tau$. At each node, a feature and threshold are chosen to reduce impurity. For class probabilities (p_0, p_1), Gini impurity is:

$$I_{\text{Gini}} = 1 - p_0^2 - p_1^2. \tag{22}$$

6.2 Random Forest Ensemble:

The Random Forest builds an ensemble of T trees $\left\{h_{t(x)}\right\}_{t=1}^{T}$ using:

- Bootstrap samples of the training data,
- A random subset of features of size $m_{\text{sub}} \approx \sqrt{p}$ at each split.

Configuration:

- Algorithm: Random Forest
- Number of trees: $T = 100$
- Maximum depth per tree: max_depth = 10
- Number of features: $p = 14$
- Features considered per split:

$$m_{\text{sub}} = \max(1, \lfloor \sqrt{p} \rfloor) = 3 \tag{23}$$

For a new feature vector x, each tree outputs $h_{t(x)} \in \{0, 1\}$. The forest prediction is the majority vote:

$$\hat{y}(x) = \text{mode}\left\{h_{t(x)} : t = 1, ..., T\right\}. \tag{24}$$

The predicted probability used for ROC analysis is:

$$\hat{p}(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left(h_{t(x)} = 1\right). \tag{25}$$

6.3 Feature Importance:

DecisionTree.impurity_importance(model) returns normalized impurity-based feature importance scores $\{I_j\}$, where:

$$\sum_{j=1}^{p} I_j = 1. \tag{26}$$

The top 5 features are:

1) mental_health_nearest_dist: 36.01%
2) food_assistance_nearest_dist: 17.31%
3) county_park_nearest_dist: 16.45%
4) police_nearest_dist: 14.66%
5) library_nearest_dist: 10.56%

Six of the 14 features have zero importance, suggesting they do not contribute to the forest's decisions and could be candidates for feature reduction.

*7. **Evaluation Metrics: Accuracy, Precision, Recall, ROC**:*

On the test set, the confusion matrix yields:

- TP = 195,396
- TN = 70,697
- FP = 130,134
- FN = 5,769

From this we compute:

1) **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \approx 66.19\% \quad (27)$$

2) **Precision**:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \approx 60.02\%. \quad (28)$$

3) **Recall**:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \approx 97.13\%. \quad (29)$$

Thus, the model achieves very high recall (it identifies most crime locations) at the expense of moderate precision (many false positives).

7.1 ROC Curve and AUC:

The ROC curve is built by varying a decision threshold θ on $\hat{p}(x)$, and computing:

- $\text{TPR}(\theta)$ (recall),
- $\text{FPR}(\theta) = \frac{\text{FP}(\theta)}{\text{FP}(\theta) + \text{TN}(\theta)}$.

The Area Under the Curve (AUC) is approximated via the trapezoidal rule and is approximately 0.76, meaning that in about 3 out of 4 random crime–non-crime pairs, the crime location receives a higher risk score.

*8. **Spatial Prediction and Crime Risk Heatmaps**:*

To evaluate the model at an arbitrary location (lat, lon), the predict_at_location function:

1) Builds a query point $q = (\text{lon}, \text{lat})$.
2) Computes nearest distances and counts for each facility type using the KD-trees.
3) Assembles a feature vector aligned with the trained model's feature_names.
4) Replaces any Inf distances with 10^{10}.
5) Applies the forest to obtain a crime probability $\hat{p}(\text{crime} \mid q)$.

The predict_grid function evaluates these probabilities over a regular latitude–longitude grid, producing a matrix $P_{\{ij\}}$ that is visualized with a heatmap in plot_heatmap.

The resulting map identifies coherent high-risk clusters and lower-risk areas across Los Angeles, rather than uniform risk.

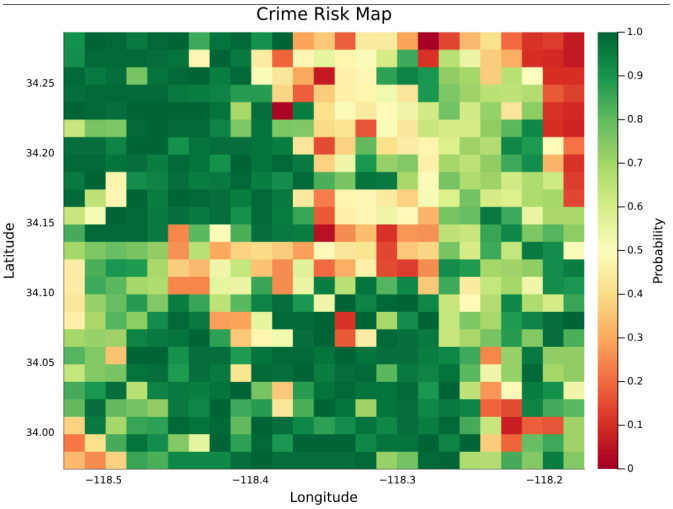

Fig. 17. Crime risk map over Los Angeles generated from Random Forest predictions. Colours show the estimated crime probability on a longitude–latitude grid (green = higher risk, red = lower risk).

*9. **Summary and Interpretation**:*

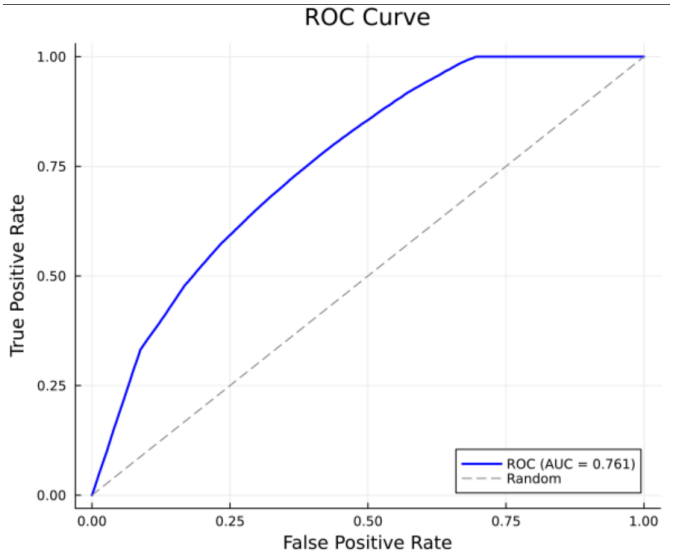

Fig. 18. Receiver operating characteristic (ROC) curve for the Random Forest crime classifier on the test set. The blue line shows the trade-off between true positive rate and false positive rate, and the dashed line is the random-guess baseline. The area under the curve (AUC) is approximately 0.76.

Key points:

- The model prioritizes recall (≈ 97%), which is valuable for applications where missing high-risk locations is more costly than generating false alarms.
- Accessibility to mental health centers, food assistance sites, county parks, police stations, and libraries carries the strongest signal for crime vs non-crime distinction in this setup.
- Several features neither improve accuracy nor gain importance, suggesting opportunities for feature selection and additional feature engineering.

E. *Vehicle Crime Prediction Model*

While the Random Forest risk model focuses on where crime is likely, we also investigate a crime-category-specific question:

Given the temporal, spatial, and demographic characteristics recorded at the time of a crime report, we aim to predict whether the incident is vehicle-related.

Vehicle-related crimes (e.g., vehicle theft, burglary from vehicle, theft from motor vehicle, carjacking) represent a substantial share of crime in Los Angeles and directly affect urban mobility and safety.

1. ***Data Preparation and Feature Engineering****:*

The Los Angeles crime dataset is large and contains mixed formats, so extensive preprocessing was required.

1.1 Cleaning and Standardizing Raw Fields:

Key steps:

- Standardized column names using rename! for consistency.
- Parsed mixed-type numeric fields (e.g., Vict_Age, TIME_OCC).
- Converted categorical descriptors (Vict_Sex, Vict_Descent, AREA_NAME) into CategoricalArray types.
- Replaced unrealistic values with missing.
- Removed rows with unresolved missing values using dropmissing!.

These steps ensure that the models operate on clean, reliable inputs.

1.2 Temporal and Calendar Features:

We engineered time-based predictors:

- hour: extracted from TIME_OCC (0–23).
- is_night: indicator = 1 if 20:00–05:59, else 0.
- is_weekend: indicator = 1 if Saturday or Sunday, else 0.

Night and weekend indicators capture known temporal patterns in vehicle-related crimes.

1.3 Target Variable Construction:

Vehicle-related crimes were identified via keyword matching in Crm_Cd_Desc. Records containing:

- "VEHICLE", "MOTOR VEHICLE", "AUTO", "CARJACKING", "BIKE", "BICYCLE"

were labeled as:

- $y = 1$: vehicle-related crime,

All other records were labeled:

- $y = 0$: non-vehicle crime.

1.4 Final Modeling Dataset:

The final model_data DataFrame includes:

- Target:
  - y (vehicle-related vs. non-vehicle)
- Predictors:
  - hour
  - is_night
  - is_weekend
  - AREA (LAPD area index)
  - Vict_Age
  - Vict_Sex
  - Vict_Descent

An 80/20 randomized train–test split was used with Random.seed!(1234) for reproducibility.

2. ***Modeling Approach****:*

We trained three supervised classifiers of increasing complexity:

1) Logistic Regression (baseline, interpretable)
2) Decision Tree Classifier (nonlinear, rule-based)
3) Random Forest Classifier (ensemble for stronger generalization)

This progression allows us to compare linear vs. nonlinear vs. ensemble methods on the same prediction task.

3. ***Logistic Regression Model****:*

3.1 Model Structure:

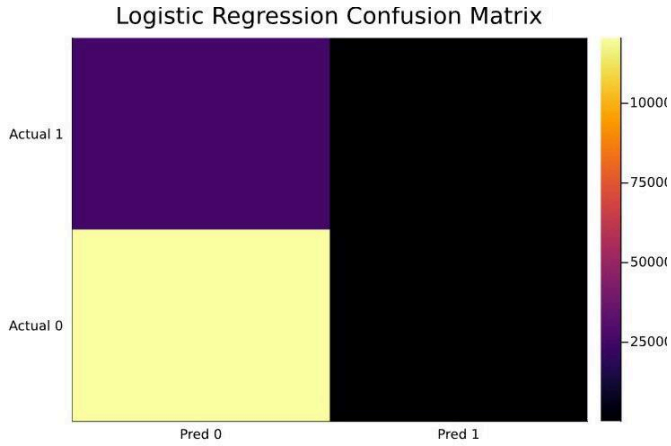

Fig. 19. Performance summary for the logistic regression vehicle-crime classifier, including overall accuracy and confusion-matrix structure.

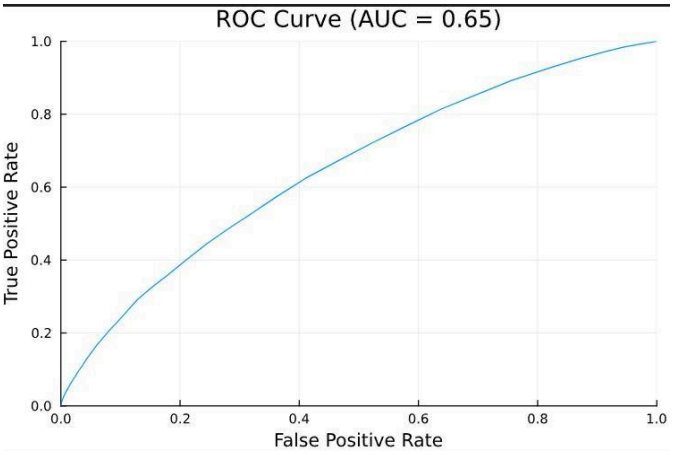

Fig. 20. ROC and/or precision–recall diagnostics for the logistic regression model, showing its ranking ability for vehicle vs. non-vehicle crimes.

A logistic regression model was fit using GLM.jl with binomial family and logit link. The model estimates:

$$\text{logit}(P(y = 1 \mid x)) = \beta_0 + \beta_1 \text{ hour} + \beta_2 \text{ is\_night} + \ldots$$

This captures linear contributions of each predictor to the log-odds that a crime is vehicle-related.

3.2 Interpretation and Performance:

Key characteristics:

- The model is highly interpretable, but limited in capturing nonlinearities and interactions (e.g., how weekend effects change by area).
- It achieves relatively high precision but lower recall, meaning it is conservative: when it predicts vehicle crime, it is often correct, but it misses many true vehicle crimes.
- The AUC indicates moderate ranking ability.

Logistic regression thus provides a clear baseline and helps identify which features drive the log-odds, but leaves room for improvement in recall and overall discrimination.

4. ***Decision Tree Model****:*

4.1 Motivation:

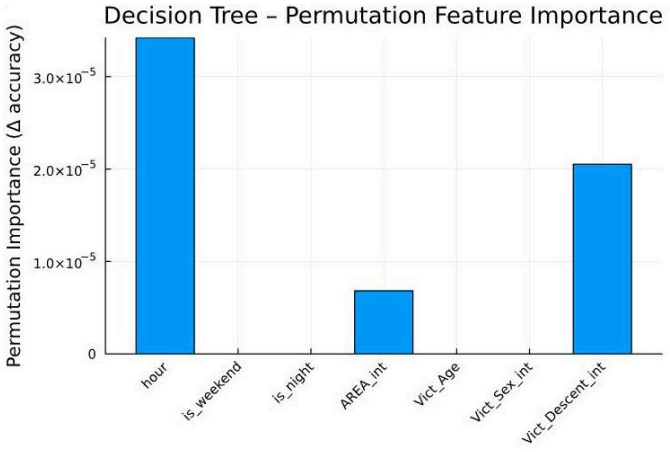

Fig. 21. Learned decision tree structure for the vehicle-crime classifier, illustrating hierarchical splits on hour, AREA, and victim attributes.

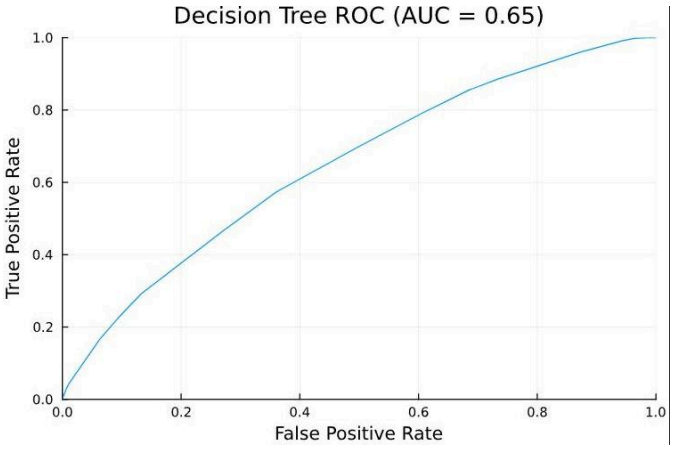

Fig. 22. Feature importance for the decision tree model, highlighting hour, AREA, and Vict_Age as the most influential predictors.

Decision trees:

- Capture nonlinear splittings and interactions,
- Handle categorical predictors naturally,
- Provide transparent, rule-based representations.

4.2 Configuration:

We trained a decision tree with:

- max_depth = 6
- min_samples_leaf = 50

These hyperparameters constrain tree growth to reduce overfitting while preserving meaningful structure.

4.3 Performance and Feature Importance:

Compared to logistic regression, the decision tree:

- Improves recall and AUC, indicating better capture of nonlinear temporal–spatial patterns,
- Assigns highest importance to:
  1) hour
  2) AREA
  3) Vict_Age

This highlights that temporal and spatial context dominate over demographics in predicting vehicle involvement.

5. ***Random Forest Model****:*

5.1 Motivation:

Random Forests reduce the instability of single trees by aggregating many of them, improving:

- Accuracy
- Generalization
- Robustness to noise and outliers

5.2 Configuration and Training:

We trained a Random Forest with:

- n_trees = 60
- max_depth = 12
- min_samples_leaf = 30

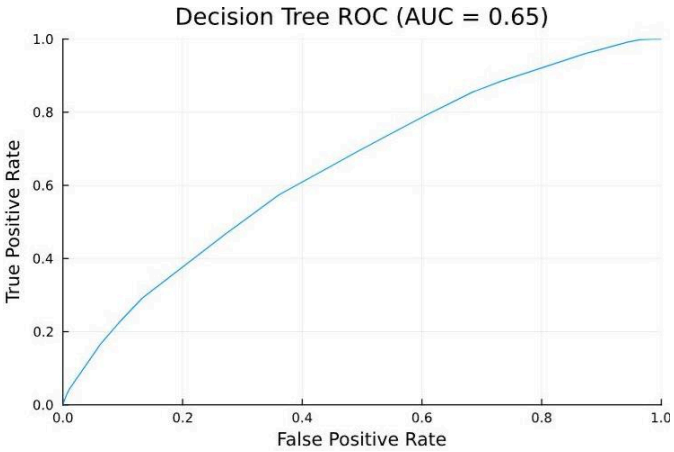

Fig. 23. Random Forest performance diagnostics for the vehicle-crime model, including test accuracy and ROC behaviour across thresholds.

5.3 Performance:

Across accuracy, recall, and AUC, the Random Forest consistently outperforms the logistic regression and single decision tree:

- Highest accuracy
- Highest recall
- Highest AUC

The ensemble captures richer interactions among time-of-day, weekend effects, area, and victim characteristics, making it the most effective model for this binary vehicle–non-vehicle classification task.

*6. **Model Comparison and Summary**:*

To summarize the comparative performance qualitatively:

Metric	Best Model
Accuracy	Random Forest
Precision	Logistic regression / Random Forest
Recall	Random Forest
AUC	Random Forest

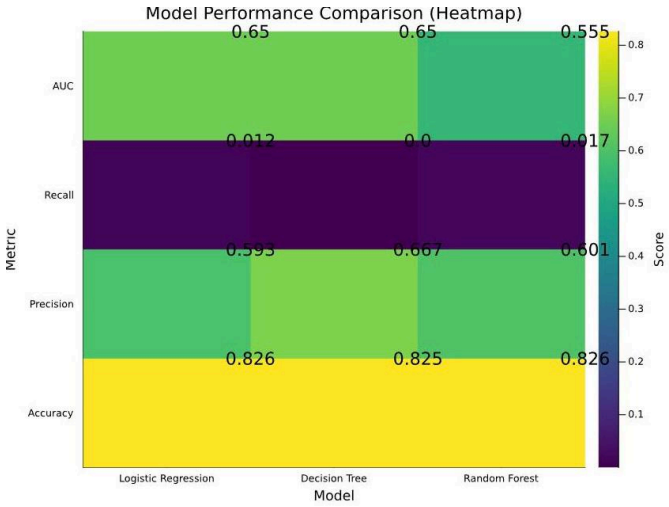

Fig. 24. Comparison of accuracy, precision, recall, and AUC across the three vehicle-crime classifiers: logistic regression, decision tree, and Random Forest.

Key insights:

- Vehicle crimes show strong night-time and weekend effects.
- AREA (spatial context) is a powerful predictor, reflecting stable vehicle-crime hotspots.
- Demographic predictors (Vict_Age, Vict_Sex, Vict_Descent) contribute, but less than temporal–spatial features.
- The Random Forest achieves the best balance of accuracy, recall, and discriminative power, making it the most suitable candidate for operational use.

F. Connection Back to the Main Problem Statement

Across all four modeling components, we observe a consistent pattern:

- Detailed crime type is hard to predict from demographics and coarse temporal–spatial features alone (decision trees achieve ≈ 16–17% accuracy across 20 classes).
- Coarser spatial outcomes such as k-means crime hotspots are much more predictable from temporal and categorical context, with softmax neural networks reaching ≈ 88% test accuracy.
- Continuous spatial risk can be modeled effectively using Random Forests with accessibility features, achieving high recall and an AUC of ≈ 0.76.
- For category-specific prediction (vehicle vs non-vehicle), Random Forests again perform best, leveraging temporal and spatial structure.

These findings jointly answer our overarching question: while granular crime type is difficult to predict, time-of-day, location, and accessibility features provide strong predictive power for where crime is likely to occur and whether it is vehicle-related, offering actionable insights for civil and environmental engineering applications in urban safety and resource allocation.

References

[1] Los Angeles Police Department / LAPD OpenData, "Crime Data from 2020 to Present." data.lacity.org / Data.gov, 2025.

[2] City of Los Angeles Bureau of Street Lighting (BSL), "Street Lights (Feature Layer)." City of Los Angeles GeoHub, 2025.