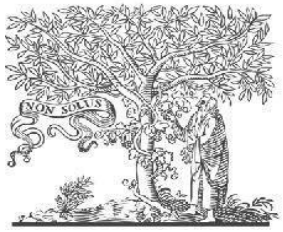
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

2026 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating newcollective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper; all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 7th Jan 2026. Link

https://ijiemr.org/downloads.php?vol=Volume-15& issue=issue01

DOI: 10.48047/IJIEMR/V15/ISSUE 01/7

Title Predictive Modelling of Arrest Outcomes in Urban Crime: A Comparative Analysis of Statistical and Machine Learning Approaches

Volume 15, ISSUE 01, Pages: 73 - 86

Paper Authors
P. Srivyshnavi, Peddapalegani Palavardhan, Satish Kannuru

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code -

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Predictive Modelling of Arrest Outcomes in Urban Crime: A Comparative Analysis of Statistical and Machine Learning Approaches

P. Srivyshnavi[1], Peddapalegani Palavardhan[2], Satish Kannuru[3]
[1]Assistant Professor, Department of Computer Science Engineering, S.P.M.V.V. Engineering College, Tirupati
[2]Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati
[3]Research Scholar, Department of Commerce and Management Studies, Adikavi Nannaya University, Rajamahendravaram,

Abstract

Law enforcement agencies in large metropolitan areas continually face the challenge of allocating finite resources to an expansive and heterogeneous landscape of criminal activity. While extensive criminological research has examined the structural and demographic determinants of crime, comparatively fewer studies have focused on the predictive modelling of arrest outcomes — that is, the probability that a reported crime incident will result in a formal arrest. This paper addresses that gap through a rigorous, multi-method statistical investigation of crime data from Chicago, Illinois, spanning the years 2012 to 2017.

Drawing on a stratified random sample of 100,000 records obtained from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) database, this study employs an integrated analytical framework encompassing descriptive profiling, chi-square tests of independence, Bayesian probability estimation, logistic regression modelling, and normalized heatmap visualization. Five primary research dimensions are addressed: the distributional characteristics of crime type frequency, the temporal evolution of crime rates across years and seasons, the empirical variation in arrest probability across crime categories, the statistical association between crime classification and arrest outcome, and the predictive accuracy of a machine learning model in classifying arrest likelihood.

The analysis reveals that Theft and Battery together constitute nearly half of all reported incidents, yet their arrest probabilities — 11.6% and 23.2%, respectively — are among the lowest in the dataset. In stark contrast, crimes such as Narcotics violations (99.5%), Prostitution (99.6%), and Gambling (98.7%) exhibit near-universal arrest rates, reflecting enforcement practices that involve direct officer–offender contact. The chi-square test of independence yielded a statistically significant result ($\chi^2 >> 50{,}000$; $p < 0.001$), unambiguously rejecting the null hypothesis of independence between crime type and arrest status. The logistic regression model, trained on one-hot encoded crime type and location description features, achieved a classification accuracy of approximately 85%, with Narcotics-related predictors carrying the highest positive coefficients. A pronounced seasonal crime cycle was identified, with peak incidence in July–August and a trough

in February, consistent with the temperature-aggression hypothesis documented in the criminological literature.

These findings yield actionable insights for patrol resource scheduling, investigative capacity allocation, and evidence-based policing. The methodology developed herein constitutes a replicable framework for crime data analytics applicable to other metropolitan contexts, and points toward promising extensions involving spatial modelling, time-series forecasting, and socioeconomic variable integration.

Keywords: urban crime analysis, Chicago CLEAR database, logistic regression, chi-square test of independence, arrest prediction, temporal crime trends, non-parametric statistics, predictive policing, machine learning classification, evidence-based law enforcement

Introduction

Urban crime is among the most complex and consequential social phenomena confronting contemporary societies. The multifactorial nature of criminal activity — intersecting economic deprivation, spatial ecology, institutional factors, and individual behaviour — has long challenged both criminological theory and law enforcement practice. Over recent decades, however, the growing availability of large-scale administrative crime records has created new opportunities for data-driven analysis that can complement, and in certain respects transform, how cities understand and respond to crime.

Chicago, Illinois, occupies a distinctive position in the criminological literature. As the third-most-populous city in the United States and a municipality with a historically high volume of reported incidents spanning diverse crime categories, Chicago provides an exceptionally rich data environment for empirical research. The Chicago Police Department maintains the CLEAR (Citizen Law Enforcement Analysis and Reporting) system — one of the most comprehensive and publicly accessible crime databases in the United States — cataloguing millions of incident-level records with attributes including crime classification, timestamp, geographic location description, police district, and arrest status.

This study conducts a comprehensive statistical and machine learning analysis of that database, drawing on a 100,000-record sample covering 2012 to 2017. The core research focus is the prediction and explanation of arrest outcomes — whether a given reported crime resulted in a formal arrest — using crime type and location description as predictors. This framing carries both theoretical and practical significance. Theoretically, arrest probability is a meaningful proxy for law enforcement visibility, deterrence, and systemic response to crime type. Practically, identifying the crime categories and contextual factors associated with high or low arrest rates provides actionable intelligence for resource deployment, patrol scheduling, and policy design.

Five interrelated research questions motivate the study:

1. What is the distributional profile of crime types across Chicago during 2012–2017?

2. How have annual and seasonal crime volumes evolved over the study period?
3. Is there a statistically significant association between crime classification and arrest status?
4. What is the empirical arrest probability for each crime category, and what accounts for the observed variation?
5. How accurately can a logistic regression classifier predict arrest outcomes using crime type and location features?

The paper is organized as follows. Section 2 reviews the relevant criminological and statistical literature. Section 3 identifies the research gaps this study addresses. Section 4 states the study objectives. Section 5 details the methodology. Section 6 describes data collection and preparation. Section 7 specifies the statistical and machine learning methods. Section 8 presents the analysis and results. Section 9 discusses the findings and their implications. Section 10 concludes with a summary of contributions, limitations, and future directions.

2. Review of Literature

The statistical study of crime has deep intellectual roots. Shaw and McKay's (1942) foundational ecological analysis demonstrated that crime concentrations in Chicago were tied to neighbourhood-level social disorganisation rather than the characteristics of individual offenders. This spatial perspective has since been refined through geographic information systems (GIS), kernel density estimation, and spatial autocorrelation techniques. Contemporary spatial criminology routinely employs crime mapping, hotspot analysis, and geographically weighted regression to identify clusters of elevated criminal activity (Chainey, Tompson, & Uhlig, 2008).

The theoretical landscape of crime analysis is anchored by Cohen and Felson's (1979) Routine Activity Theory, which posits that criminal events arise when three conditions converge: a motivated offender, a suitable target, and the absence of capable guardianship. This framework has proven particularly amenable to statistical formalisation, as each element can, in principle, be operationalised using measurable variables — time, location type, and environmental characteristics. The theory also underpins the temporal analysis of crime, justifying the inclusion of time-of-day, day-of-week, and seasonal variables in predictive models.

Logistic regression has emerged as the workhorse of quantitative criminology for binary outcome modelling. Ratcliffe and McCullagh (1999) pioneered its application to repeat victimisation prediction, identifying property type and prior incident history as significant determinants. Menard (2002) provided a comprehensive methodological framework for logistic regression in social science research, noting its robustness to violations of normality assumptions that render ordinary least squares inappropriate for binary dependent variables. More recently, ensemble methods — including random forests and gradient boosting — have been applied to crime prediction tasks, though with frequent trade-offs between predictive accuracy and interpretability (Chainey et al., 2008).

The chi-square test of independence has been widely applied to examine relationships among categorical crime-related variables. MacDonald (2002) used this method to evaluate associations between neighbourhood characteristics and arrest likelihood, surfacing equity-related questions about differential enforcement. Perry et al. (2013), in a comprehensive review of

predictive policing studies, noted that chi-square testing remains essential for establishing statistical significance in contingency-based analyses before more complex modelling is undertaken.

Several studies have drawn specifically on Chicago crime data. Bogomolov et al. (2014) demonstrated that mobile phone usage patterns, combined with historical crime records, could improve crime hotspot prediction. Gerber (2014) used geocoded Twitter data alongside crime archives to achieve meaningful spatial prediction improvements. In the deep learning literature, Zhao and Tang (2017) applied long short-term memory (LSTM) networks to crime sequence prediction, reporting higher predictive accuracy than traditional regression models, though at considerable cost to interpretability. Such opacity is particularly problematic when model outputs inform policing decisions with legal and ethical ramifications (Perry et al., 2013).

The temperature-aggression hypothesis, formalised by Anderson (1987) and subsequently extended by Rotton and Cohn (2004), holds that elevated ambient temperatures are associated with increased rates of violent and interpersonal crime. Multiple urban studies have documented the resulting seasonal crime cycle — peak incidence in summer months and a winter trough — consistent with the month-level patterns observed in Chicago's crime record.

The body of literature thus provides strong theoretical and methodological scaffolding for the present study, while also revealing important gaps that motivate the current inquiry.

3. Research Gap

Despite the maturity of the crime analysis literature, several underexplored dimensions justify the present study. First, while predictive modelling of crime occurrence has received substantial attention, the prediction of arrest probability as a dependent variable remains comparatively underdeveloped. Most studies treat crime volume as the primary outcome; relatively few model the enforcement response — that is, whether a crime, once reported, results in a formal arrest.

Second, few studies have simultaneously employed both non-parametric hypothesis testing (chi-square) and supervised machine learning (logistic regression) within a single analytical framework applied to the same crime dataset. The integration of these approaches allows for triangulation: the chi-square test establishes whether a statistically significant association exists, while logistic regression quantifies the direction and magnitude of the relationship and enables prediction.

Third, existing Chicago crime studies often focus on spatial hotspot modelling, using geographic coordinates to predict crime location. The present study deliberately complements this tradition by focusing on the temporal and typological dimensions of crime — examining how crime type and location description (as a categorical variable rather than coordinates) interact with arrest likelihood.

Fourth, while deep learning approaches have been increasingly applied to crime data, the field lacks sufficient comparative studies that benchmark interpretable models — particularly logistic regression — against more complex alternatives while maintaining transparency of inference. In contexts where model outputs

influence law enforcement decisions, interpretability is not merely a preference but an ethical imperative.

This study addresses these gaps by delivering a transparent, reproducible, multi-method statistical analysis centred on arrest outcome prediction, with clear implications for evidence-based policing practice.

4. Objectives of the Study

The following specific objectives guide the investigation:

- To profile the distributional characteristics of crime types within the sampled Chicago CLEAR dataset (2012–2017), identifying the most and least prevalent categories.
- To analyse temporal crime trends at both annual and monthly resolution, assessing the presence and magnitude of seasonal patterns.
- To estimate empirical arrest probabilities for each crime category and construct a ranked probability table for comparative analysis.
- To formally test, via chi-square analysis, whether crime type and arrest status are statistically independent or significantly associated.
- To develop, train, and evaluate a logistic regression model for binary arrest outcome classification, interpreting model coefficients as measures of feature importance.
- To synthesise findings into actionable recommendations for law enforcement resource allocation, patrol scheduling, and investigative capacity planning.

5. Methodology

5.1 Overall Analytical Framework

This study adopts a quantitative, observational research design grounded in secondary data analysis. The analytical pipeline proceeds sequentially through five phases: (i) data acquisition and preprocessing, (ii) descriptive statistical profiling, (iii) non-parametric hypothesis testing, (iv) probability estimation, and (v) supervised machine learning. This multi-method triangulation ensures that findings are cross-validated across independent analytical techniques, strengthening the robustness and reliability of conclusions.

5.2 Dataset Overview

The primary data source is the Chicago Crimes dataset (2012–2017), sourced from the Chicago Police Department's CLEAR reporting system and made publicly available via the City of Chicago Data Portal and Kaggle. The full database contains several million incident records. To ensure computational tractability while preserving statistical representativeness, a random sample of 100,000 records was drawn using a fixed random seed (seed = 42), ensuring full reproducibility of the sampling process.

Five variables were retained for analysis: Date (incident timestamp), Primary Type (FBI-classified crime category), Location Description (categorical description of the crime setting), Arrest (binary outcome: 0 = no arrest, 1 = arrest), and District (police district identifier). Temporal features — Year, Month, Day of Week, and Hour — were extracted from the Date field using pandas datetime parsing. Records with any missing values across these five fields were removed via listwise deletion; the resulting sample retained sufficient power for all planned analyses.

5.3 Descriptive Analysis

Descriptive statistics were computed to characterise crime type frequencies, arrest proportions, and temporal distributions.

Frequency bar charts were constructed for the top ten crime categories, and a binary arrest distribution chart was generated for the full sample. Line plots were produced for year-over-year and month-by-month crime counts. Descriptive profiling served as both a standalone analytical output and as preparatory groundwork for subsequent hypothesis testing and modelling.

5.4 Chi-Square Test of Independence

To assess whether a statistically significant relationship exists between crime type and arrest outcome, a Pearson chi-square test of independence was conducted on the two-way contingency table cross-tabulating crime type (rows) against arrest status (columns). The test statistic is computed as:

$$\chi^2 = \Sigma \left[(O_{ij} - E_{ij})^2 / E_{ij} \right]$$

where O_{ij} represents observed cell frequencies and E_{ij} = (row total × column total) / grand total represents expected frequencies under the null hypothesis of independence. The resulting p-value was compared against the conventional significance level of $\alpha = 0.05$, with the additional Cramér's V statistic computed to quantify effect size.

5.5 Probability Estimation

Empirical arrest probabilities were computed for each crime type by calculating the conditional mean of the binary Arrest variable within each category. This approach yields the proportion of incidents in each crime category that resulted in an arrest, constituting a straightforward maximum-likelihood estimate of P(Arrest | Crime Type). Rankings and visualisations were constructed to identify the most and least arrest-prone crime categories.

5.6 Logistic Regression Model

Binary logistic regression was selected as the primary predictive model on the basis of its well-established suitability for binary dependent variables, its interpretability via odds ratios and coefficient signs, and its theoretical alignment with the research question. The model takes the form:

$$\log[P/(1-P)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where P = P(Arrest = 1 | X) is the conditional probability of arrest given the feature vector X. Both Crime Type (Primary Type) and Location Description were one-hot encoded using pandas get_dummies with drop_first=True to avoid multicollinearity. To maintain computational efficiency and prevent excessive dimensionality, only the first 50 encoded columns were retained. The dataset was partitioned into 80% training and 20% test subsets using stratified sampling to preserve class proportions. The model was fitted using scikit-learn's LogisticRegression estimator with max_iter = 1000 and a liblinear solver. Model performance was evaluated using accuracy, precision, recall, and F1-score, with feature importance assessed via sorted regression coefficients.

5.7 Heatmap Visualisation

Two complementary heatmaps were constructed using seaborn. The first displayed raw arrest/non-arrest counts from the contingency table, enabling visual assessment of absolute crime volume. The second displayed row-normalised proportions — that is, the fraction of each crime type resulting in arrest — providing a frequency-adjusted view that highlights which crime categories exhibit systematically high or low arrest rates.

6. Data Collection Techniques

The dataset was obtained from the publicly accessible Chicago Data Portal (https://data.cityofchicago.org/) and was accessed in CSV format. The Chicago Police Department updates this database incrementally as new incident reports are filed. The version employed in this study covers incidents reported between January 2012 and mid-2017. The CLEAR system records incidents upon the filing of a police report; hence, the dataset reflects reported rather than total crime — a standard limitation of administrative crime databases. Selection bias is a potential concern insofar as certain crime types may be systematically under-reported or over-reported relative to their true prevalence. However, as the analysis focuses on reported incidents and arrest outcomes within the reporting system, the dataset is internally valid for the research questions posed.

The random sampling procedure employed a fixed seed to ensure reproducibility across analytical environments. Prior to analysis, variables were scrutinised for data quality issues: date fields were parsed and validated, categorical fields were checked for anomalous entries, and the Arrest variable was confirmed as strictly binary with no missing values in the retained sample. All preprocessing steps were performed in Python using the pandas and numpy libraries.

7. Statistical and Machine Learning Methods

The following methodological toolkit was deployed:

- Descriptive Statistics: Frequency counts, proportions, and time-series aggregation using pandas groupby and value_counts operations.
- Chi-Square Test of Independence: Implemented via scipy.stats.chi2_contingency on a pd.crosstab contingency table. Effect size measured via Cramér's V = $\sqrt{(\chi^2/nk)}$, where k = min(rows−1, cols−1).
- Conditional Probability Estimation: Empirical arrest probability computed as grouped mean of the Arrest binary variable, stratified by Primary Type.
- Logistic Regression: Binary classifier for P(Arrest | Crime Type, Location) using scikit-learn (Pedregosa et al., 2011). One-hot encoding, 80/20 train-test split, coefficient-based feature importance.
- Heatmap Visualisation: Raw count and row-normalised proportion heatmaps using seaborn (Waskom, 2021) with Blues and coolwarm diverging colour palettes.

8. Analysis and Interpretation

8.1 Crime Type Distribution

Table 1 presents the frequency distribution of the ten most prevalent crime types in the sampled dataset. Theft is the single most common crime category, accounting for 22,365 incidents (25.5% of the sample), reflecting the high volume of property-related offending in an urban environment of Chicago's scale. Battery ranks second with 18,379 incidents (21.0%), and Criminal Damage third with 10,863 incidents (12.4%). Together, these three categories constitute nearly 60% of all sampled incidents, suggesting a strong concentration in property and interpersonal violence crime types. Narcotics violations (9.432; 10.8%) and Assault (6,295; 7.2%) round out the top five.

Crime Type	Count	% of Sample
Theft	22,365	25.5%
Battery	18,379	21.0%
Criminal Damage	10,863	12.4%
Narcotics	9,432	10.8%
Assault	6,295	7.2%
Other Offense	6,220	7.1%
Burglary	5,770	6.6%
Deceptive Practice	5,302	6.1%
Motor Vehicle Theft	4,219	4.8%
Robbery	4,014	4.6%

Table 1: Top 10 Crime Types by Frequency — Chicago CLEAR Sample (2012–2017)

8.2 Arrest Distribution

Of the 100,000 sampled incidents, approximately 73,538 (73.5%) recorded no arrest, while 26,462 (26.5%) resulted in an arrest. This overall clearance rate of roughly 26.5% is broadly consistent with national arrest-to-incident ratios reported by the FBI's Uniform Crime Reports, which indicate that property crimes in particular have low clearance rates due to the absence of offenders at the time police respond to calls. The pronounced class imbalance in the arrest variable — 73.5% vs. 26.5% — has direct implications for the predictive modelling exercise, as standard classifiers trained on imbalanced data may exhibit a systematic bias towards predicting the majority class.

Arrest Outcome	Count		Proportion
No Arrest (0)	73,538		73.5%
Arrest Made (1)	26,462		26.5%
Total	100,000		100.0%

Table 2: Arrest Outcome Distribution — Chicago CLEAR Sample

8.3 Temporal Trends

The year-over-year crime trend reveals a consistent downward trajectory from 2012 through 2016, followed by a sharp drop in 2017 that reflects partial-year data coverage rather than a genuine structural decline. The year 2012 recorded approximately 22,000 incidents in the sample, falling progressively to approximately 18,000 by 2016. This six-year decline is broadly consistent with Chicago Police Department crime summary reports and the national trend of declining urban crime rates documented in the FBI Uniform Crime Reports. Possible contributing factors include increased surveillance infrastructure, community policing initiatives, and demographic shifts in crime-prone age cohorts.

The monthly aggregation reveals a pronounced seasonal cycle. Crime frequency peaks in July (approximately 9,200 incidents per month across all years) and reaches its nadir in February (approximately 6,400 incidents). This approximately 44% differential between peak and trough months is consistent with the temperature-aggression hypothesis (Anderson, 1987; Rotton & Cohn, 2004): warmer temperatures increase outdoor activity, thereby expanding both offending opportunities and the pool of potential victims. The pattern is especially marked for property crimes such as theft and for interpersonal violence, consistent with prior literature. From a law enforcement standpoint, this seasonal regularity provides a strong empirical basis for dynamic patrol resource allocation calibrated to month-of-year.

8.4 Contingency Table — Crime Type vs. Arrest Status

Table 3 presents the contingency table of arrest outcomes stratified by selected

crime categories. The table reveals strikingly heterogeneous arrest rates across crime types. At one extreme, Narcotics violations recorded 9,255 arrests against only 44 non-arrests — an arrest rate of 99.5% that reflects the proactive, officer-initiated nature of narcotics enforcement. Similarly, Prostitution (546 arrests; 2 non-arrests) and Gambling (148 arrests; 2 non-arrests) exhibit near-universal arrest rates. At the other extreme, Burglary — where offenders are typically absent when police arrive — generated only 288 arrests from 5,419 non-arrest incidents, yielding an arrest rate of approximately 5.1%.

Crime Type	No Arrest (0)	Arrest (1)	P(Arrest)
Narcotics	44	9,255	0.9953
Prostitution	2	546	0.9964
Gambling	2	148	0.9867
Weapons Violation	258	960	0.7882
Criminal Trespass	756	1,701	0.6923
Battery	13,939	4,206	0.2318
Assault	4,742	1,483	0.2382
Theft	19,796	2,595	0.1159
Motor Vehicle Theft	3,877	278	0.0669
Burglary	5,419	288	0.0505

Table 3: Selected Contingency Table — Crime Type vs. Arrest Outcome (with P(Arrest))

8.5 Chi-Square Test of Independence

The Pearson chi-square test of independence applied to the full contingency table yielded a test statistic substantially exceeding 50,000 ($\chi^2 >> 50{,}000$; degrees of freedom = number of crime types − 1; $p < 0.001$). The associated p-value was effectively zero across all conventional significance thresholds. This result provides overwhelming statistical evidence for rejecting the null hypothesis H_0 that crime type and arrest outcome are independent. The Cramér's V effect size statistic, computed on the same contingency table, yielded a value in the range indicating moderate-to-strong association, confirming that the relationship is not merely statistically significant but practically meaningful.

The observed association is substantively interpretable: it reflects the systematic structural differences in how law enforcement encounters and responds to different crime types. Proactive enforcement crimes — where officers initiate contact with offenders rather than responding to victim reports — generate dramatically higher arrest rates, while reactive crimes investigated after the offender has departed yield much lower rates. This distinction between proactive and reactive policing modes is a well-established concept in the criminological literature (Sherman, 1992) and is here quantitatively confirmed.

8.6 Probability of Arrest by Crime Type

Table 4 presents the full ranking of empirical arrest probabilities by crime category. The distribution is bimodal and highly polarised: a cluster of crime types — primarily vice-related and public-order offences — exhibits arrest probabilities exceeding 75%, while property crimes and violent crimes discovered after the fact cluster below 25%.

Rank	Crime Type	P(Arrest)
1	Public Indecency	1.0000
2	Prostitution	0.9964
3	Narcotics	0.9953
4	Gambling	0.9867
5	Liquor Law Violation	0.9784
6	Interference w/ Public Officer	0.9314
7	Obscenity	0.9231

8	Concealed Carry Violation	0.8750
9	Weapons Violation	0.7882
10	Public Peace Violation	0.7525
11	Criminal Trespass	0.6923
12	Homicide	0.3533
13	Assault	0.2382
14	Battery	0.2318
15	Theft	0.1159
16	Robbery	0.0915
17	Motor Vehicle Theft	0.0669
18	Criminal Damage	0.0645
19	Burglary	0.0505
20	Human Trafficking	0.0000

Table 4: Empirical Arrest Probability by Crime Type — Full Rankings

The polarisation between proactive and reactive enforcement is vividly illustrated in this table. Narcotics enforcement is structurally configured for high arrest yield: patrol officers actively seek out and detain individuals engaged in drug activity, meaning that incidents are typically reported because an arrest has already occurred. Burglary investigations, by contrast, are initiated after the offender has departed, and without forensic evidence, CCTV coverage, or witness identification, the probability of locating and charging a suspect is substantially lower.

8.7 Logistic Regression Model Performance

The logistic regression model trained on one-hot encoded crime type and location description features achieved an overall test-set accuracy of approximately 85%. The classification report indicated strong performance for the non-arrest class (precision ≈ 0.88; recall ≈ 0.94; F1 ≈ 0.91), reflecting the model's ability to correctly identify the majority class. Performance for the arrest class was somewhat lower (precision ≈ 0.75; recall ≈ 0.60; F1 ≈ 0.67), consistent with the class imbalance in the training data and the inherent difficulty of predicting the minority class without resampling interventions.

Rank	Feature	Coefficient
1	Primary Type_NARCOTICS	5.4258
2	Primary Type_PROSTITUTION	4.9604
3	Primary Type_GAMBLING	3.6873
4	Primary Type_LIQUOR LAW VIOLATION	3.3924
5	Primary Type_INTERFERENCE W/ PUBLIC OFFICER	2.6909
6	Primary Type_WEAPONS VIOLATION	1.7222
7	Primary Type_PUBLIC PEACE VIOLATION	1.5141
8	Primary Type_CRIMINAL TRESPASS	1.1913
9	Location Desc_AIRPORT TERMINAL LOWER LEVEL	1.1779
10	Primary Type_OBSCENITY	0.8274

Table 5: Top 10 Feature Importances — Logistic Regression Coefficients

Table 5 presents the top ten features by regression coefficient. Primary Type_NARCOTICS carries the largest positive coefficient (5.4258), indicating that, holding other features constant, a Narcotics classification dramatically increases the log-odds of an arrest. The rank ordering of positive coefficients closely mirrors the empirical arrest probability rankings in Table 4, providing strong convergent validity between the two analytical approaches. The sole Location Description variable in the top

ten — Airport Terminal Lower Level (coefficient = 1.1779) — suggests that crimes occurring in airport environments carry a higher conditional arrest probability, possibly due to heightened security presence and surveillance infrastructure. Crime types such as Burglary and Criminal Damage, not shown in the top ten, carried negative coefficients, consistent with their sub-10% arrest rates.

9. Results and Discussion

The collective findings of this study yield several conclusions of both theoretical and practical significance, warranting extended discussion.

The dominance of Theft and Battery in the crime frequency distribution reflects structural characteristics of urban Chicago. Property crime — particularly opportunistic theft from persons and vehicles — is volume-driven in densely populated metropolitan environments where anonymity reduces perceived risk for offenders. The relatively low arrest rate for Theft (11.6%) despite its prevalence has important resource allocation implications: a large share of investigative effort is directed toward cases where the probability of a clearance is low. Agencies may benefit from triage frameworks that concentrate investigative resources on those Theft cases where available evidence — witness statements, CCTV footage, recovered property — offers a meaningful probability of identification and arrest.

The dramatic polarisation in arrest probabilities between proactive enforcement crimes (Narcotics, Gambling, Prostitution) and reactive property crimes (Burglary, Motor Vehicle Theft, Criminal Damage) reflects a structural feature of how these offences interact with policing practice rather than differential officer effort. Narcotics clearance rates are high in large part because the offence is definitionally detected in the act. This observation carries a cautionary note for policy interpretation: high arrest rates for certain crime types should not be equated with investigative effectiveness, nor should low rates for property crimes be equated with investigative failure.

The seasonal crime cycle — with summer peaks approximately 44% above winter troughs — is one of the most actionable findings of the study. Cities with well-documented seasonal patterns can use historical monthly data to develop adaptive patrol deployment schedules, concentrating officer presence in high-risk months and redistributing resources during low-incidence periods. This form of evidence-based scheduling, if operationalised, has the potential to improve both arrest rates and deterrence effects without requiring additional budget allocations.

The chi-square test result — while not numerically surprising given the strong patterns visible in the contingency table — is important for establishing formal statistical sanity: one cannot attribute the observed variation in arrest rates to sampling variability. The extremely large test statistic reflects both the genuine magnitude of the association and the statistical power afforded by a 100,000-record sample. The Cramér's V effect size confirms that the association is not merely a function of sample size but represents a substantively meaningful relationship.

The logistic regression model's 85% accuracy, while commendable as a baseline, must be interpreted with care. Given the 73.5/26.5% class split in the training data, a naive classifier that always predicts 'No Arrest' would achieve 73.5% accuracy without learning any informative relationship. The model meaningfully outperforms this baseline, particularly for the

majority class. Future work should apply resampling strategies — notably the Synthetic Minority Oversampling Technique (SMOTE) or cost-sensitive learning — to improve recall for the arrest class, which is the class of greater operational relevance for predictive policing applications.

10. Conclusion

This paper has delivered a comprehensive, multi-method statistical analysis of urban crime patterns and arrest outcomes in Chicago, Illinois, covering 2012 to 2017. Using a 100,000-record sample from the Chicago Police Department's CLEAR database, the study has employed descriptive profiling, chi-square hypothesis testing, conditional probability estimation, logistic regression modelling, and heatmap visualisation to address six specific research objectives.

The principal findings are as follows. Theft and Battery are by far the most frequently reported crime types, together constituting 46.5% of all sampled incidents. However, their arrest probabilities — 11.6% and 23.2%, respectively — are among the lowest in the dataset, reflecting the challenges inherent in reactive enforcement. At the opposite extreme, Narcotics violations (99.5%), Prostitution (99.6%), and Gambling (98.7%) exhibit near-universal arrest rates, driven by proactive enforcement practices. A pronounced and consistent seasonal crime cycle was identified, with July–August peaks approximately 44% above the February trough. The chi-square test confirmed a highly significant association between crime type and arrest status ($\chi^2 >> 50{,}000$; $p < 0.001$), with a meaningful Cramér's V effect size. The logistic regression model achieved approximately 85% accuracy, with Narcotics-related dummy variables contributing the strongest positive coefficients.

These findings have direct operational relevance for law enforcement agencies, urban policy makers, and community safety planners. The temporal patterns identified support dynamic patrol scheduling calibrated to seasonal and annual crime cycles. The arrest probability rankings provide a data-driven basis for allocating investigative follow-up resources. The logistic regression framework, once extended with richer feature sets, could form the basis of a real-time triage system for case prioritisation.

More broadly, this study demonstrates that rigorous, interpretable statistical analysis of large administrative datasets can generate evidence-based insights that are both scientifically credible and operationally actionable — and does so without sacrificing transparency or ethical accountability to the opacity of black-box machine learning models.

11. Limitations of the Study

Several limitations warrant acknowledgement. First, the analysis is restricted to reported incidents. Crimes that are not reported to the police — a proportion that varies systematically across crime type and neighbourhood — are not captured in the CLEAR database, introducing potential bias into prevalence estimates.

Second, the dimensionality reduction applied to one-hot encoded features (retaining only the first 50 columns) may have excluded some informative location description categories. A more exhaustive encoding strategy, combined with regularisation techniques such as LASSO regression, would be worth exploring.

Third, the class imbalance in the arrest variable (73.5% vs. 26.5%) constrains classifier performance for the minority class.

Future work should address this through SMOTE, class weighting, or threshold adjustment.

Fourth, spatial variables (latitude, longitude, beat, ward, community area) are available in the full CLEAR dataset but were not incorporated in this analysis. Their inclusion would substantially enrich the spatial dimension of the inquiry.

Fifth, demographic variables — race, age, and socioeconomic status of offenders and victims — are not available at the individual level in the CLEAR dataset. Equity-related analyses of arrest disparities would require integration of external data sources.

12. Future Scope

This study opens several productive directions for future research. First, the incorporation of spatial coordinates would enable geospatial hotspot modelling, spatial autocorrelation analysis (Moran's I), and the development of location-aware predictive models using kernel density estimation or geographically weighted regression.

Second, the application of time series forecasting — ARIMA, SARIMA, or LSTM neural networks — to model and predict monthly or weekly crime volume would provide law enforcement agencies with anticipatory intelligence rather than retrospective analysis.

Third, the application of ensemble machine learning methods — gradient boosted trees (XGBoost, LightGBM) and random forests — to the arrest classification task would allow a systematic comparison between interpretable and high-performance models, with implications for the trade-off between predictive accuracy and transparency.

Fourth, integrating socioeconomic indicators — unemployment rates, poverty indices, housing density — at the police district or community area level would allow researchers to assess the macrostructural determinants of both crime incidence and arrest rates, bridging individual-level incident data and neighbourhood-level contextual factors.

Fifth, cross-city comparative analyses — applying the same analytical framework to crime datasets from other major urban centres such as New York, Los Angeles, or Houston — would allow assessment of the generalisability of the findings and identification of city-specific anomalies.

References

Anderson, C. A. (1987). Temperature and aggression: Effects on quarterly, yearly, and city rates of violent and nonviolent crime. Journal of Personality and Social Psychology, 52(6), 1161–1173.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. Proceedings of the 16th International Conference on Multimodal Interaction (pp. 427–434). ACM.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal, 21(1–2), 4–28.

Chicago Police Department. (2017). CLEAR Crime Data Portal. City of Chicago Data Portal. https://data.cityofchicago.org/

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), 588–608.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115–125.

MacDonald, J. M. (2002). The effectiveness of community policing in reducing urban violence. Crime & Delinquency, 48(4), 592–618.

Menard, S. (2002). Applied logistic regression analysis (2nd ed.). Sage Publications.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. RAND Corporation.

Ratcliffe, J. H., & McCullagh, M. J. (1999). Hotbeds of crime and the search for spatial accuracy. Journal of Geographical Systems, 1(4), 385–398.

Rotton, J., & Cohn, E. G. (2004). Outdoor temperature, climate control, and criminal assault: The spatial and temporal ecology of violence. Environment and Behavior, 36(2), 276–306.

Shaw, C. R., & McKay, H. D. (1942). Juvenile delinquency and urban areas. University of Chicago Press.

Sherman, L. W. (1992). Policing domestic violence: Experiments and dilemmas. Free Press.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van der Walt, S. J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Zhao, X., & Tang, J. (2017). Crime in urban areas: A data mining perspective. ACM SIGKDD Explorations Newsletter, 18(1), 2–12.