Enterprise Cloud Computing and Big Data (BUDT 737)

Predictive Crime Analytics for Optimizing Business Security Investments in Los Angeles

Abdoulay Lashley, Dante Caraballo, Goutham Korada, Leo Petrucci

ORIGINAL WORK STATEMENT

"We, the undersigned, certify that we composed this proposal and that it is original work."

Team Member	**Signed Signature**
Abdoulay Lashley	*Abdoulay Lashley*
Dante Caraballo	*Dante Caraballo*
Goutham Korada	*Goutham Korada*
Leo Petrucci	*Leo Petrucci*

Case Study: Sentinel Retail – Using Predictive Crime Analytics to Safeguard Urban Operations

Introduction

Alexis Reed is the Chief Operating Officer of Sentinel Retail, a fast-growing chain of convenience stores operating across Los Angeles County. With over 120 locations, Sentinel serves a high-volume urban customer base, often in neighborhoods with varied crime profiles. As the company scales, Alexis faces increasing concerns over the safety of frontline employees and customers, particularly after a series of late-night thefts and vandalism incidents in 2023.

Although Alexis has previously relied on historical police reports and anecdotal store manager feedback to guide security spending, these methods lack precision. After attending a University of Maryland workshop on predictive modeling in public safety, she realized the potential of machine learning to transform how Sentinel anticipates and mitigates crime risk. She engaged a data science consulting team to analyze roughly three years of LAPD crime data and design a risk-informed strategy to optimize Sentinel's safety investments.

Company Background: Sentinel Retail

Founded in 2015, Sentinel Retail focuses on community-based convenience, operating stores in dense urban centers often underserved by larger retail chains. With a lean operating model and a local hiring preference, Sentinel prides itself on accessibility and affordability.

As part of its mission, Sentinel aims to balance profitability with social responsibility, including ensuring the physical safety of its staff and customers. Historically, stores have been equipped with standard alarm systems and security cameras. However, incidents of burglary, armed robbery, and physical assault have challenged the adequacy of this one-size-fits-all model.

Alexis's executive team is now exploring a data-driven framework that can:

- Identify which stores face elevated risks.
- Predict which crime types are most probable by location and time.
- Suggest proactive investments in security personnel or physical barriers like plexiglass installations.

Strategic Initiative: Predictive Crime Modeling

With access to three years of LAPD crime data, Sentinel partnered with the analytics firm to conduct the following:

1. Historical Risk Assessment: Mapped all Sentinel locations against a geospatial database of reported crimes.
2. Model Development: Used machine learning algorithms (e.g., Random Forest and a Multiclass Perceptron Classifier) to model the most likely crime per location.
3. Suggest Security New Recommendations

Challenges and Opportunities

Sentinel's current approach leads to uneven protection: Some high-risk stores lack proper safeguards, while low-risk locations overspend on unnecessary measures. Security costs are rising, but incidents continue, highlighting the need for more intelligent allocation.

Questions

- What kinds of techniques can be used to predict which locations are likely to experience specific types of crime?
- What kind of workshop did Alexis Reed attend? What was the main skill taught there?
- What kind of data was used to generate the insights, and where was it from?
- What is an example of a kind of security upgrade that the model would recommend?

Business Challenge and Question

In an era characterized by data-driven strategic decision-making, businesses are confronted with the growing challenge of managing operational risks associated with criminal activities. Criminal activity in urban regions, such as Los Angeles, influences not only financial performance but also the safety of employees, the experience of customers, and the long-term sustainability of the organization. While numerous firms depend on subjective perceptions or

anecdotal evidence to evaluate their exposure to crime, a substantial opportunity exists to utilize public data and predictive modeling for the formulation of objective, localized crime risk assessments.

This case study explores a comprehensive approach to using machine learning and crime data to help businesses answer a critical question: *"Should we invest in additional safety infrastructure, such as hiring security guards or installing plexiglass barriers, based on the crime risk at our commercial site?"* By tapping into 368,505 crime reports spanning 2023 to the present, our goal was to build a predictive system that predicts crime probability, identifies high-risk areas, and advises business owners on our critical question.

I. Business Context

Depending on their location, industry, and hours of operation, businesses in Los Angeles experience different exposure to theft, assault, vandalism, and other criminal activity. Despite this, many companies follow standardized security protocols that may not align with the risk of operating a business, even in certain areas that are more prone to crime.

In reviewing LAPD crime data from **2023 to present**, we observed:

- 368,505 recorded incidents across 28 different categories
- A disproportionately high rate of thefts and assaults near retail shops, convenience stores, and fast-food establishments
- Temporal patterns showing increased criminal activity in the late afternoon and evening hours, particularly on weekends

With this robust dataset, businesses can shift from reactive to proactive, anticipating criminal activity and preparing accordingly.

II. Business Problem

Urban businesses must make complex, cost-effective decisions about security investments. A retail chain may ask:

- Is spending $2500-$4000 per month on a security guard for this location?
- Should we invest $5000 in plexiglass barriers for staff safety?
- Are there specific hours we should reduce operations to decrease risk?

These decisions, when uniformed, either result in over-investment leading to wasted capital or under-investment resulting in loss and liability. The goal of this project is to offer informed ways to:

- Quantify crime risk at the location level
- Determine the most probable type and timing of crime
- Provide tailored safety recommendations based on predicted risk.

Model Solution and Implementation

I. Approach and Methodology

Our methodology leverages distributed data processing using PySpark to handle the large-scale crime dataset efficiently. We adopted a robust and scalable pipeline that supports multiclass classification and interpretable insights.

II. Data Acquisition and Preparation

- **Data Source:** Obtained LAPD crime dataset covering 2023-2025 from the LA City Crime Data Portal. Filtered from 2023 – present.
- **Platform and Tools:** Apache Spark (PySpark), ensuring efficient handling of large datasets.
- **Initial Data Ingestion:** Read CSV data via Spark's CSV reader with automatic schema inference.

III. Data Cleaning and Preprocessing

- Dropped rows with NAs in critical columns: `Crm Cd Desc`, `Premis Cd`, `DATE OCC`, `"TIME OCC"`, `"LAT"`, `"LON"`, `"Vict Age"`, `"Vict Sex"`, `"Vict Descent"`, `"Cross Street"`
- Imputed remaining missing values for non-critical fields in `Weapon Desc`, `Premis Desc`, and `Mocodes` with `"Unknown"`
- Resulted in approximately 38,566 usable records.

IV. Feature Engineering

- **Temporal Features**
    - Parsed timestamp fields (`DATE OCC`, `TIME OCC`) into structured datetime formats
        - Extracted hour, month, and day_of_week from timestamp fields (`DATE OCC`, `TIME OCC`)

    - Night Indicator: A binary `is_night` variable was created (1 for crimes between 8 PM and 6 AM; 0 otherwise).
- **Spatial Features**
    - Converted `LAT` and `LON` into numerical features for spatial modeling
    - Applied K-Means clustering (`k=10` clusters) to identify geographic "crime hotspots".
- **Demographic Features**
    - Created Bins of `Victim Age`
        - `"Under 18"`, `"18-35"`, `"36-65"`, and `"65+"` using conditional logic.
        - This created a new categorical variable, `vict_age_group`, improving model interpretability.
- **Categorical Encoding**
    - Encoded categorical fields such as `Crm Cd Desc` using StringIndexer
    - Employed StringIndexer and OneHotEncoder on categorical variables such as:
        - `Vict Sex`, `Vict Descent`, `Premis Desc`, `Weapon Desc`, `vict_age_group`, and `AREA NAME`.

Primary Model: Random Forest Multiclass Classification

Objective: Framed a multiclass classification problem to predict Crm Cd Desc (a specific crime type).

Used **Spark ML pipeline** with the following steps:

Categorical Encoding: Indexing and One-hot Encoding.

- `StringIndexer`
- `OneHotEncoder`

Label Encoding: Crime descriptions indexed for multiclass classification.

Vector Assembly (`VectorAssembler`): Combined numerical, categorical, and spatial features into a single feature vector.

Random Forest Classifier (`RandomForestClassifier`):

I. Initial Model Training

- Configured with 100 trees.
- Trained using an 80/20 train-test split.
- Initial evaluation indicated moderate predictive performance:
    - **F1 Score:** 0.2318
    - **Accuracy:** 36.97%
    - **Precision:** 28.24%
    - **Recall:** 36.97%

II. Top-10 Crime Type Selection

Identified the top 10 most frequent crimes for improved predictive accuracy by the count():

1. |ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
    a. |5772 |
2. |BATTERY - SIMPLE ASSAULT
    a. |4653 |
3. |BURGLARY FROM VEHICLE
    a. |3828 |
4. |ROBBERY
    a. |3555 |
5. |VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
    a. |2907 |
6. |THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)
    a. |1781 |
7. |INTIMATE PARTNER - SIMPLE ASSAULT
    a. |1697 |
8. |THEFT PLAIN - PETTY ($950 & UNDER)
    a. |1558 |
9. |BRANDISH WEAPON
    a. |1072 |
10. |THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
    a. |1068 |

Filtered the dataset to include only these crimes (**focused dataset**).

III. **Enhanced Model Implementation (Random Forest)**

- Re-trained the Random Forest classifier on the focused dataset.
- `Pipeline`: Indexing → Encoding → Label Indexing → Feature Assembling → Random Forest.
- Configured with **100 trees**.

IV. **Cross-Validation and Hyperparameter Tuning**

- Performed 5-fold Cross-Validation.
- The hyperparameters were tuned: `maxDepth (5, 8, 10)`, `maxBins (35)`, and `impurity ("gini", "entropy")`.
- Selected the best hyperparameters based on CV results.

Secondary Model: Multiclass Perceptron Classifier

- `Pipeline`: Indexing → Encoding → Label Indexing → Feature Assembling → **Multiclass Perceptron Classifier**.
- Takes Predicted Labels
  - Recommends security measures based on labels
  - Structure:
    - Input Layer: Labels
    - Single Hidden Layer with 64 Perceptrons
    - Output Layer: Predicted Labels

Model Evaluation and Results: Random Forest Multiclass Classification

Evaluated using multiclass metrics: `F1`, `Accuracy`, `weightedPrecision`, `weightedRecall`. Feature importance extracted from Random Forest for interpretability.

I. **Evaluation Metrics (Focused Dataset - Initial Re-Evaluation)**

- **F1 Score:** 0.5704
- **Accuracy:** 63.71%
- **Precision:** 53.97%
- **Recall:** 63.71%

II. **Post Cross-Validation Metrics (Optimized Model)**

- **F1 Score:** 0.5793
- **Accuracy:** 64.88%
- **Precision:** 54.13%
- **Recall:** 64.88%
- **Top-5 Accuracy:** 88.79%

III. Key Insights from Evaluation

- Substantial improvement from initial modeling to the optimized model (accuracy improvement from ~37% to ~65%).
- Strong Top-5 accuracy indicates reliability for practical recommendations.

IV. Geospatial and Crime Matching Application

- Integrated crime prediction with geographic clusters to identify the most common crime types in each area.
- Enabled personalized safety recommendations based on dominant crimes in each business area.
- Used static mapping of crime frequency within the area to evaluate high-risk hotspots.
- Determine the location of crimes committed and consider actions for nearby stores, using the mapping tool as a complementary method to the multiclass perceptron classifier.
- Used a LA location GeoJson to map data on boundaries within LA for cluster data mapping.

V. Recommended Security Measures

- **High Assault Risk Areas:** Employ security personnel (monthly cost: ~$3,000–$4,000).
- **High Vehicle-related Crime Areas:** Install advanced lighting and surveillance (initial cost: ~$5,000–$7,500).
- **Retail Theft-prone Areas:** Plexiglass barriers or visible anti-theft installations (cost: ~$5,000).

Model Evaluation and Results: Multiclass Perceptron Classifier

- **F1 Score:** 0.5759
- **Accuracy:** 60.86%
- **Precision:** 55.40%
- **Recall:** 60.86%

- Recommended Security Measures (Same as Random Forest)

Streaming Implementation

We wanted to implement a live recommendation system via streaming. We do not have a speech-to-text model implemented right now, but in the future, this would aid our producer side of the code in analyzing police scanner data.

Consumer

- **Takes the crimes from the broker as input**
  - **Outputs Recommended Security Actions**

Producer

- **Takes crimes from data, but in future implementations will take from a police scanner**
  - **Stores crimes in the broker**

Annexures - Data Overview

I. Data Source

The dataset used in this study was obtained from the City of Los Angeles Open Data Portal, specifically from the Los Angeles Police Department's crime dataset:

Crime Data from 2020 to Present – LAPD

This dataset is maintained by the Los Angeles Police Department (LAPD) and reflects incidents of crime reported across all 21 LAPD community police divisions. It is updated bi-monthly and openly licensed under Creative Commons.

II. Description of the Data

The dataset contains structured crime report data transcribed from official LAPD records. Each row corresponds to a reported crime incident, with variables capturing temporal, spatial, demographic, and contextual attributes. Below is a detailed breakdown:

Variable	Description	Data Type	Treatment in the Model

DATE OCC	Date the crime occurred (MM/DD/YYYY)	Timestamp	Parsed to extract month and day of week
TIME OCC	Time of crime in 24-hour format	Integer	Transformed into padded hour + binary night flag
AREA	LAPD division number (1–21)	Categorical	Mapped to AREA NAME; not used directly
AREA NAME	Division name (e.g., "Hollywood", "Northeast")	Categorical	One-hot encoded
Crm Cd Desc	Primary crime description (target label)	Target	Indexed via StringIndexer
Vict Age	Victim’s age	Numerical	Binned into age groups (Under 18, 18–35, 36–65, 65+)
Vict Sex	Victim gender (M, F, X)	Categorical	One-hot encoded
Vict Descent	Victim’s ethnicity/descent code	Categorical	One-hot encoded
Premis Desc	Description of the place where the crime occurred	Categorical	One-hot encoded
Weapon Desc	Description of weapon used, if any	Categorical	One-hot encoded

LAT / LON	Latitude and longitude of the crime	Numerical	Used in geo-clustering (KMeans, geo_cluster)
geo_cluster (derived)	Cluster ID indicating geographic grouping of incidents	Categorical	Added as a numerical feature from KMeans
Rpt Dist No	LAPD reporting district code	Categorical	Dropped (high cardinality with limited added value)
Mocodes	Modus Operandi codes describing methods used by suspect(s)	Categorical	Dropped for initial model due to multi-label complexity
Crm Cd 1–4	Crime codes (1 = most serious, 2–4 = associated crimes)	Categorical	Used only Crm Cd 1 through Crm Cd Desc
Part 1-2	FBI crime classification (1 = more serious)	Categorical	Not included in model directly
Premis Cd	Code for Premis Desc	Categorical	Dropped in favor of description
Weapon Used Cd	Code for Weapon Desc	Categorical	Dropped in favor of description
Status / Status Desc	Case disposition (e.g., open, closed, cleared)	Categorical	Not used; model focuses on incident prediction

LOCATION	Nearest 100-block street location	Categorical	Dropped (privacy-masked; overlaps with geo coords)
Cross Street	Intersecting road at crime location	Categorical	Dropped (many nulls and irregular formats)

III. Sample Size and Variables

- **Filtered Dataset Range**: January 1, 2023 – Present
- **Sample Size (n)**: ~38,566 rows after cleaning and filtering for valid geographic and temporal data
- **Number of Variables (k)**: 15 key variables used in modeling, including both original and engineered features (e.g., hour, is_night, geo_cluster)

IV. Sample Observations

DATE OCC	TIME OCC	AREA NAME	Crm Cd Desc	Vict Age	Vict Sex	Weapon Desc	Premis Desc	LAT	LON
01/05/2023	2130	Southeast	BATTERY - SIMPLE ASSAULT	27	M	STRONG-ARM	SIDEWALK	33.95	-118.25
01/10/2023	1330	Northeast	BURGLARY FROM VEHICLE	35	F	UNKNOWN	STREET	34.11	-118.28
02/02/2023	0230	Hollywood	VANDALISM - FELONY ($400 & OVER…)	18	M	UNKNOWN	FREEWAY	34.09	-118.33
03/15/2023	1800	Olympic	THEFT PLAIN - PETTY ($950 & UNDER)	22	F	UNKNOWN	PARKING LOT	34.05	-118.29

04/22/2023	0200	Foothill	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	30	M	KNIFE	STREET	34.25	-118.45

V. Why This Data Is of Interest

Crime is a pressing concern for urban business owners, affecting operations, safety, and profitability. This dataset offers:

- **High Granularity**: Daily crime incidents are timestamped and geolocated.
- **Broad Scope**: Covers multiple crime types and community areas.
- **Real-World Relevance**: Enables predictive insights tied to specific locations and demographics.
- **Open Licensing**: Freely accessible and compliant with ethical data usage standards.

Note: This makes it ideal for training machine learning models that assist in real-time decision-making about ***investments in safety infrastructure****, tailored to* ***location-specific crime risks****.*