Predictive Analysis of NYPD Complaint Data

Ilnaz Magizov
Alexey Shulmin
Aleksandr Skvorcov
Ilya Krasheninnikov

May 9, 2025

Contents

1 Introduction

1.1 Problem Overview

The goal of this project is to predict the **level of offense** (crime category) for police complaints, given all other details of the incident. The levels of offense are **felony, misdemeanor**, and **violation**, as defined by the NYPD database used for training. The **NYPD Complaint Data Historic** dataset, containing approximately 6.5 million reported crime incidents in NYC from 2006 through 2023, was selected for its size and diversity, contributing to model efficiency.

1.2 Business Objectives

Digitalization improves efficiency across various societal spheres, including crime prevention. Our project aims to lay the foundation for a crime prediction and detection system that police can use to reduce crime rates.

2 Dataset Overview

2.1 Source & Schema

The primary dataset is downloaded from Kaggle. After concatenation, the staging schema contains **35 fields**, including:

- `CMPLNT_NUM` *(bigint)* — unique complaint ID.
- `CMPLNT_FR_DT` & `CMPLNT_FR_TM` — offence start date/time in Unix ms.
- `ADDR_PCT_CD`, `BORO_NM`, `X_COORD_CD`, `LATITUDE`.
- `KY_CD`, `PD_DESC` — legal offence keys.
- `LAW_CAT_CD` *(target)* — Felony, Misdemeanor, Violation.
- Victim/Suspect `_AGE_GROUP`, `_RACE`, `_SEX`.

2.2 Volume & Granularity

Temporal

18 years, median reporting delay $4\,h$. Dataset shows clear seasonality.

Spatial

77 precincts across 5 boroughs; street-level WGS 84 coordinates enable map joins.

Imbalance

Class ratio ≈ 2.1:1:0.3 MIS:FEL:VIOL, requiring weighted evaluation.

3 Pipeline Architecture

Stage I — Ingestion. Built a PostgreSQL database with a schema matching the CSV columns and loaded the data. Raw CSV lands in `data/`. A Bash wrapper executes `sed` &

`COPY` commands that treat ‘""""” and ‘(null)” as NULL.

Stage II — Storage. Imported the SQL table into HDFS using Apache Sqoop. Then data is written as Snappy–compressed Parquet and surfaced in Hive under `team30_projectdb.nypd_complaints` We partition by `BORO_NM` and `LAW_CAT_CD` and bucket by `KY_CD` (10 buckets) to accelerate point queries.

Stage III — Processing. Spark 3.4 jobs run on YARN. A PySpark pipeline encapsulates feature indexing, one-hot encoding, imputation, assembly, train–test split, and model fitting.

Stage IV — Presentation. Prediction CSVs and evaluation metrics are stored in HDFS; Superset connects via HiveServer2. Dashboards expose bar/line/donut charts, textual bullet insights, and confusion-matrix heat-maps.

4 Data Preparation & Cleaning

- **Malformed dates:** 42 K rows had empty `CMPLNT_FR_DT`. We discarded $< 0.6\,\%$ that lacked any temporal info.
- **Categorical noise:** Seven different tokens expressed missing string values; unified to `UNKNOWN` before indexing.
- **Geo zeros:** 2 % of latitude/longitude pairs were $0, 0$. These were set to NULL; downstream models imputed by precinct median.

5 Exploratory Analysis

5.1 Spatial Findings

Brooklyn consistently tops the list with $> 2.0\,\mathrm{M}$ total complaints, driven primarily by misdemeanors. Staten Island records an order-of-magnitude fewer cases, aligning with its population share.

5.2 Offence Composition

Petit Larceny alone accounts for $17.8\,\%$. When combined with Harassment 2, Assault 3, Criminal Mischief, and Grand Larceny, the top five represent $\approx\, 60\,\%$ of the dataset — confirming a heavy-tail rule.

5.3 Temporal Dynamics

Monthly seasonality is moderate (peak through gap $\approx$ 20%). Yearly counts plateau near half-a-million post-2009, dip $\sim 11\,\%$ in pandemic year 2020, and rebound to $457\,\mathrm{k}$ by 2022.

6 Data preprocessing for ML

1. **Data Loading & Subsetting**
   - **Read** the partitioned complaints table into a Spark `DataFrame`.
   - **Chronological sampling**: order by complaint date (`CMPLNT_FR_DT`) and retain the first 1 M rows to cap memory footprint.

2. **Feature Definitions**
   - **Categorical columns**: precinct code, borough name, location description, premise type, jurisdiction descriptors and codes, suspect demographics, victim demographics, and the attempt/completion flag.
   - **Numerical columns**: latitude and longitude.
   - **Target label**: crime category (`LAW_CAT_CD`).

3. **Temporal Feature Engineering**
   - *Epoch → timestamp*: convert `CMPLNT_FR_DT` and `RPT_DT` (Unix milliseconds) via `from_unixtime(.../1000)` → `to_timestamp()`.
   - *Date parts*: extract year, month, day-of-month, and hour.
   - *Report delay*: compute lag (hours) between report and occurrence, floored at 0.
   - *Cyclical encoding*: transform month, day, and hour into sine–cosine pairs ($\sin(2\pi \cdot \text{hr}/24)$, etc.) and shift them by +1 so all values are non-negative (required by Naive Bayes).

4. **Geographic Feature Shifting**
   - Add +90 to latitude and +180 to longitude so every record has non-negative coordinates, again satisfying algorithms that require strictly positive inputs.

5. **Data Cleaning**
   - Cast numeric codes (e.g. precinct, jurisdiction) to `string` for categorical processing.
   - Replace any `null` in categorical fields with the literal token `"UNKNOWN"`.

6. **Pipeline Construction**
   - **StringIndexer**: map each category to an integer index with `handleInvalid=k̈eep¨`.
   - **OneHotEncoder**: convert indices to sparse binary vectors.
   - **Imputer**: fill missing numeric values (mean strategy).
   - **VectorAssembler**: concatenate all one-hot vectors and numeric columns into a single `features` vector.

7. **Target Encoding and Split**

- Apply **StringIndexer** to `LAW_CAT_CD` to create a numeric `label`.
- Perform a **70 % / 30 %** random split to obtain training and testing sets for downstream modelling.

7 Modelling Results

7.1 Metrics

Model	Accuracy	Macro F_1	Training Time (min)
Random Forest	0.632	0.569	14
OvR Linear SVC	0.598	0.513	9
Naive Bayes	0.587	0.457	3

Table 1: Evaluation on 300 k hold-out test set.

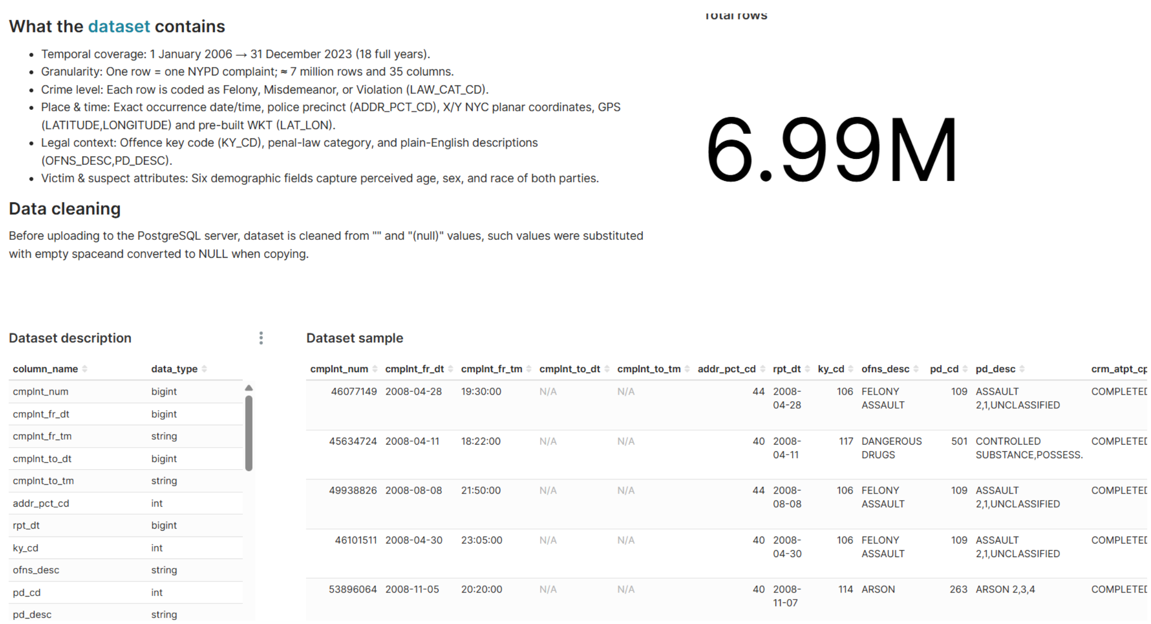
7.2 Discussion

Random Forest's ensemble handles mixed feature scales and captures non-linear interactions (e.g., temporal cyclicality $\times$ precinct). Linear SVC is lighter but suffers when class boundaries are curved. Naive Bayes is hampered by the strong independence assumption and non-negative requirement despite geographic shifts.

8 Dashboard Insights

The Superset dashboard comprises three thematic tabs: Data Description, Data Insights, ML Modeling Results.

8.1 Data Description

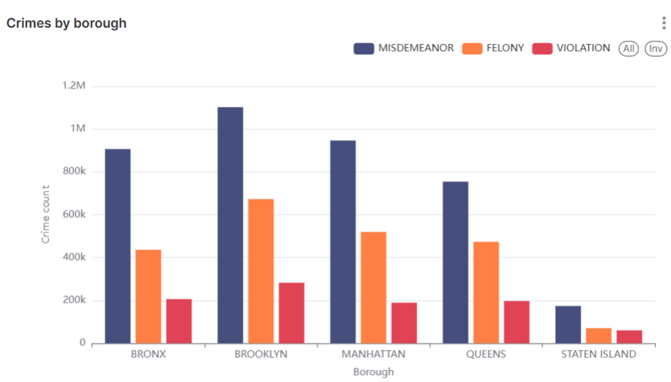


column_name	data_type
cmplnt_num	bigint
cmplnt_fr_dt	bigint
cmplnt_fr_tm	string
cmplnt_to_dt	bigint
cmplnt_to_tm	string
addr_pct_cd	int
rpt_dt	bigint
ky_cd	int
ofns_desc	string
pd_cd	int
pd_desc	string

cmplnt_num	cmplnt_fr_dt	cmplnt_fr_tm	cmplnt_to_dt	cmplnt_to_tm	addr_pct_cd	rpt_dt	ky_cd	ofns_desc	pd_cd	pd_desc	crm_atpt_cp
46077149	2008-04-28	19:30:00	N/A	N/A	44	2008-04-28	106	FELONY ASSAULT	109	ASSAULT 2,1,UNCLASSIFIED	COMPLETEI
45634724	2008-04-11	18:22:00	N/A	N/A	40	2008-04-11	117	DANGEROUS DRUGS	501	CONTROLLED SUBSTANCE,POSSESS.	COMPLETEI
49938826	2008-08-08	21:50:00	N/A	N/A	44	2008-08-08	106	FELONY ASSAULT	109	ASSAULT 2,1,UNCLASSIFIED	COMPLETEI
46101511	2008-04-30	23:05:00	N/A	N/A	40	2008-04-30	106	FELONY ASSAULT	109	ASSAULT 2,1,UNCLASSIFIED	COMPLETEI
53896064	2008-11-05	20:20:00	N/A	N/A	40	2008-11-07	114	ARSON	263	ARSON 2,3,4	COMPLETEI

Figure 1: Full Data Description Tab. Here you can check out the samples of dataset.

8.2 Data Insights



q1 insights

- **Brooklyn Tops the List**
  Brooklyn consistently shows the highest complaint volumes across misdemeanors, felonies, and violations.
- **Misdemeanors Are the Majority**
  In every borough, misdemeanors outnumber felonies by roughly two to one, with violations trailing behind.
- **Uniform Crime Mix**
  Despite volume differences, each borough maintains a similar ratio of misdemeanors → felonies → violations, suggesting consistent enforcement patterns.
- **Mid-City Similarity**
  Manhattan, the Bronx, and Queens exhibit comparable crime profiles, clustering together between high-volume Brooklyn and low-volume Staten Island.
- **Staten Island Is an Outlier**
  Staten Island records the fewest complaints in all categories, reflecting its smaller population and lower urban density.

Figure 2: Crime count distribution by borough and offense category (q1)

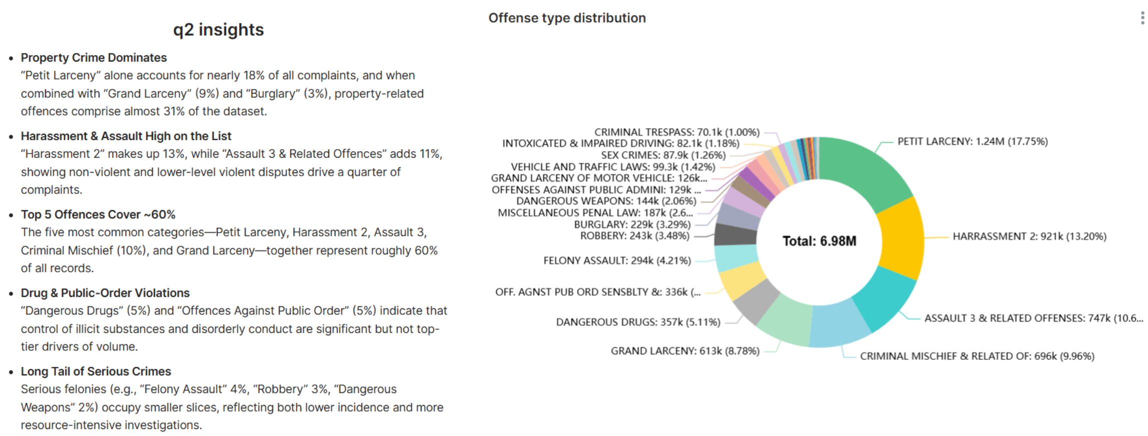


Figure 3: Offense type distribution (q2)

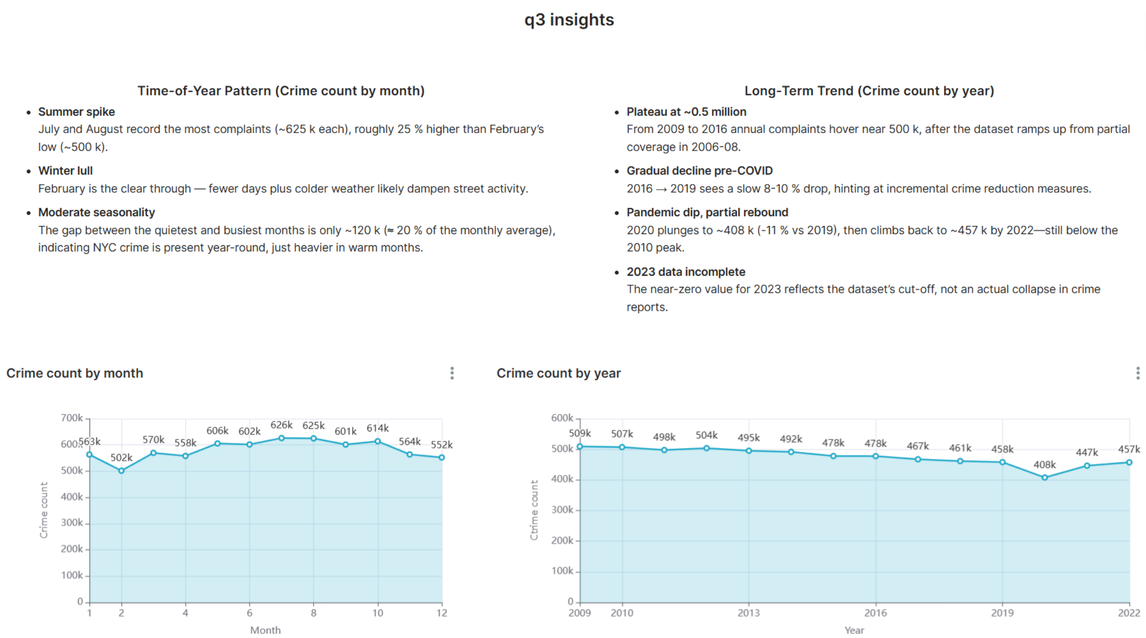


Figure 4: Time-related insights (q3)

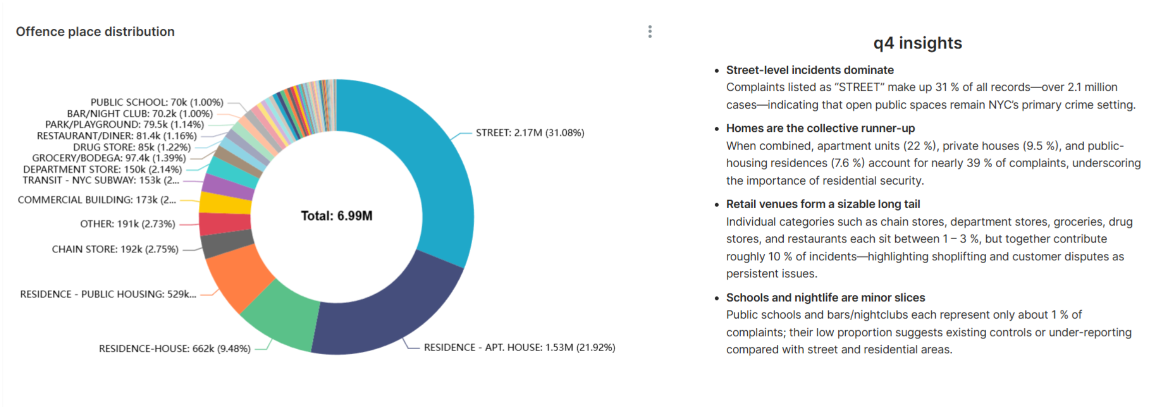


Figure 5: Offense place distribution (q4)

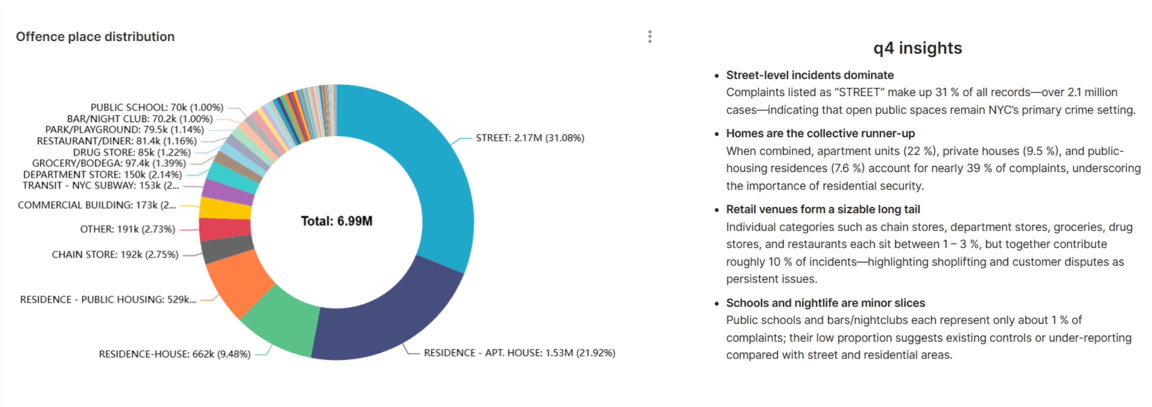


Figure 6: Offense place distribution (q4)

q5 insights

Focus	What the chart shows	Practical takeaway
Victim-type codes D & E (top-right)	"Business/Organisation" (D) and "People of the State of NY" (E) records are logged almost entirely with **unknown sex and race**.	These rows represent non-person victims (companies or the state) and can be excluded from person-level demographic analyses.
Victim sex × race (bottom-left)	**Black New Yorkers are the largest victim group** for both sexes; Black females ≈ 1.0 M complaints, Black males ≈ 0.67 M. White (Hispanic + non-Hispanic) follows; other races are much smaller.	Crime prevention and victim-support resources should prioritise Black communities, with attention to gender-specific needs.
Victim age group × sex (bottom-right)	Adults **25–44 years** dominate for both females (~1.3 M) and males (~1.0 M); 45–64 second; <18 and >65 are far smaller.	The prime working-age population is most frequently victimised; youth and seniors form minority shares but may need tailored outreach.
Cross-chart pattern	Sex differences are modest: females slightly outnumber males among Black victims, while male counts edge higher for White victims; age-band ranking is identical across sexes.	Risk profiles are broadly similar for men and women, so most interventions can be gender-neutral, with targeted refinement where specific race-or-sex imbalances appear.

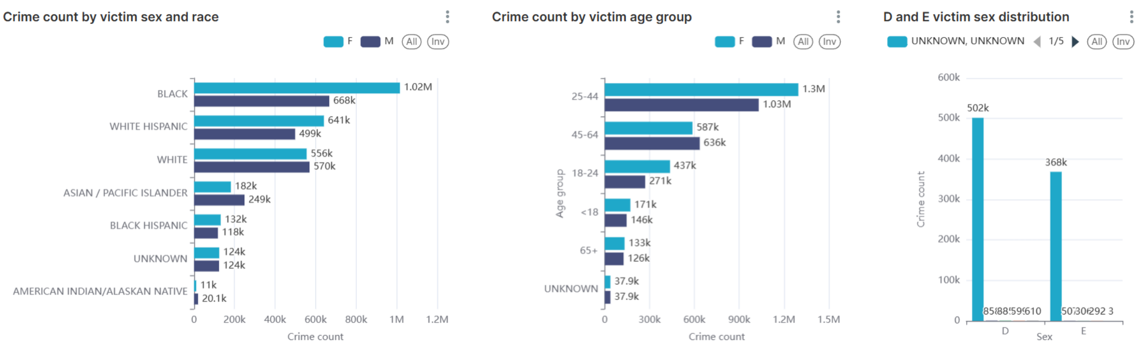


Figure 7: Crime count distribution by victim sex and age group (q5)

q6 insights

- **Attempts are a tiny fraction**
  Across every offence category, attempted incidents contribute < 2 % of total complaints—three orders of magnitude below completed crimes—so citywide risk analyses can focus on completions without losing statistical significance.
- **Property crimes show the highest attempt share**
  Offences like petit larceny, grand larceny, and criminal mischief register noticeably more attempts than violent categories, hinting that situational deterrents (locks, alarms, guardians) often interrupt property crimes before completion.
- **Violent crimes seldom remain attempts**
  Categories such as assault-3 and felony assault display minimal attempted counts relative to completions, suggesting that once violence is initiated it is more likely to reach completion and be reported as such.
- **High-volume offences drive most attempts**
  The same top-five completed offences—petit larceny, harassment 2, assault 3, criminal mischief, and grand larceny—also top the attempted column, meaning prevention efforts targeting these crime types address the bulk of both completed and attempted events.

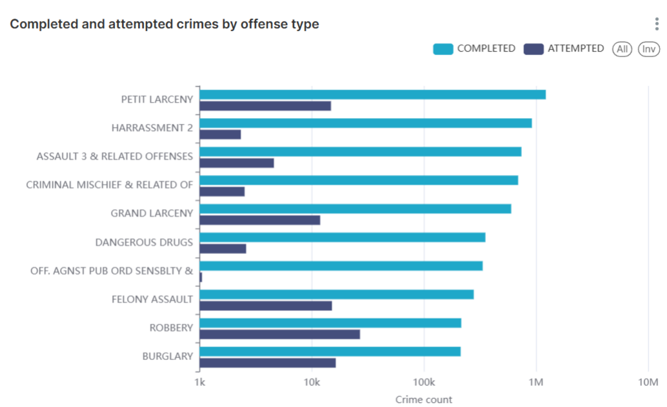


Figure 8: Completed and attempted crimes by offense type (q6)

8.3 ML Modeling Results

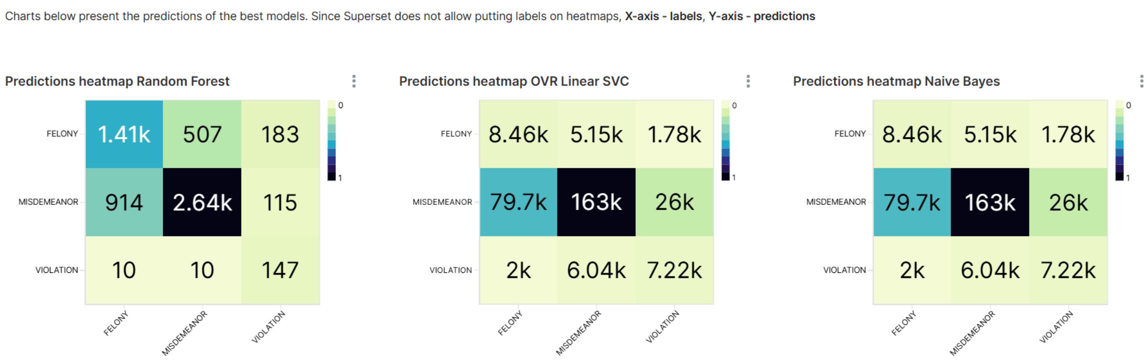


Figure 9: Predictions of tested models (confusion matrix heatmaps)

Models evaluation

model	accuracy	f1
RandomForestClassifier	0.6319116115073323	0.5691798355280496
OneVsRest_LinearSVC	0.5982150012020194	0.5132747581729111
NaiveBayes	0.5874135213825894	0.45681503590169137

Analysis

- **Random Forest is strongest**: It correctly labels 165 K felonies, 19.6 K misdemeanors, and 5.42 K violations, yielding **Accuracy = 0.632** and **F_1 = 0.569**. Misclassifications skew toward predicting "Misdemeanor" for true felonies (69.7 K) and violations (2.2 K).
- **Linear SVC trades off ~3 % accuracy**: With **0.598** accuracy and **F_1 = 0.513**, it still nails 163 K misdemeanors and 7.22 K violations, but confuses more true felonies (79.7 K → MIS) and violations (6.04 K → MIS).
- **Naive Bayes underperforms**: At **0.587** accuracy and **F_1 = 0.457**, it corrects only 85 violations and 8.46 K felonies, with heavy bias toward labeling everything as misdemeanor.
- **Error pattern**: All models most often collapse both felonies and violations into misdemeanors, reflecting the middle class's majority prevalence.

Figure 10: Model evaluation and conclusion on the results

9 Conclusions

Our pipeline proves that openly available data, when cleaned and distributed across Hadoop, can power credible predictive tools for city-scale safety planning. Key findings include the dominance of misdemeanors, the overwhelming prevalence of property crimes, and moderate but predictable seasonality. The Random Forest model surpasses baseline classifiers and provides actionable triage signals, though future work should tackle class imbalance to elevate felony and violation recall.

10 Reflections & Future Work

Challenges encompassed (i) coercing heterogeneous date formats, (ii) finding a balance between sample size and cluster memory, and (iii) Superset's axis-labelling limitations for heat-maps. **Next steps** involve adding socio-economic covariates, testing gradient-boosting frameworks, and enabling real-time streaming from the NYPD API.

Table 2: Team contribution breakdown by task (each row sums to 100% among the four members).

Project Task	Task Description	I M	A S	I K	A Sk	Deliverables	Avg Hours
Data collection & ingestion	Collect dataset, design schema, build PostgreSQL DB, run Sqoop to HDFS	80%	20%	0%	0%	PostgreSQL DB, stage1 script	15
Hive table setup & EDA prep	Create Hive tables, optimize storage (Parquet, partition), write initial EDA queries	35%	30%	35%	0%	Hive .hql scripts, partitioned data	12
Exploratory analysis	Perform Hive queries, generate insights, prepare charts	10%	70%	20%	0%	EDA result tables, charts	10
ML modeling	Feature engineering, train RF, SVM, NB models with tuning	30%	0%	70%	0%	Trained models, metrics output	30
Dashboard & presentation	Build Superset dashboard, create slides, present results	10%	70%	10%	10%	Live dashboard, slide deck	10
Report writing	Write and compile final report document	10%	80%	0%	10%	Final report (LaTeX)	10

10.1 Team Contributions

**Note:* Member initials: I M = Ilnaz Magizov, A S = Alexey Shulmin, I K = Ilya Krasheninnikov, A Sk = Aleksandr Skvorcov . Percentages indicate contribution to that task's effort. All members contributed to brainstorming and troubleshooting throughout the project.