Chicago Crime Prediction

Melissa Tobias, Natalie Pegues, Malaika Galvan, Daynise Escudero, and Sirjana Yadav

CRIMINAL JUSTICE

Background information of Chicago

Baseline models set a basic level of performance that more complex models should aim to exceed.

Final conclusion and summary of the project

Crucial steps in preparing raw data for analysis and modeling

The process of creating a blueprint or representation of how data is organized, structured, and related

Background

Chicago has long faced challenges with public safety, particularly in specific neighborhoods. Their violent crime rate is around 540 per 100,000 people and is one of the top 20 cities in the US. with the highest crime rates. Understanding patterns in criminal activity like what types of crime occur most frequently and where they are concentrated is essential for law enforcement and community leaders.

The dataset includes over five years of reported crime data from 2012 to 2017 in the city of Chicago. Providing detailed information such as the type of crime, location, time, arrest status, and more. By analyzing this data, we are aiming to identify high risk areas and trends, offering insights that could inform and create future crime prevention strategies.

Objective & research questions

Our project focuses on using classification and other techniques to answer the following questions:

Can we predict whether an arrest will occur for a reported crime in Chicago?

- How frequently do arrest occur for reported crimes?

Can historical data help inform police decision making?

- How can we use past patterns to support smarter deployment and improve justice outcomes?

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	...	24.0
1	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	...	20.0
2	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False	...	37.0
3	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	0460	BATTERY	SIMPLE	SIDEWALK	False	...	28.0
4	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	0820	THEFT	$500 AND UNDER	RESIDENCE	False	...	28.0

	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
	29.0	08B	1154907.0	1893681.0	2016	05/10/2016 03:56:50 PM	41.864073	-87.706819	(41.864073157, -87.706818608)
	42.0	08B	1183066.0	1864330.0	2016	05/10/2016 03:56:50 PM	41.782922	-87.604363	(41.782921527, -87.60436317)
	25.0	24	1140789.0	1904819.0	2016	05/10/2016 03:56:50 PM	41.894908	-87.758372	(41.894908283, -87.758371958)
	25.0	08B	1143223.0	1901475.0	2016	05/10/2016 03:56:50 PM	41.885687	-87.749516	(41.885686845, -87.749515983)
	25.0	06	1139890.0	1901675.0	2016	05/10/2016 03:56:50 PM	41.886297	-87.761751	(41.886297242, -87.761750709)

Unnamed: 0	int64
ID	int64
Case Number	object
Date	object
Block	object
IUCR	object
Primary Type	object
Description	object
Location Description	object
Arrest	bool
Domestic	bool
Beat	int64
District	float64
Ward	float64
Community Area	float64
FBI Code	object
X Coordinate	float64
Y Coordinate	float64
Year	int64
Updated On	object
Latitude	float64
Longitude	float64
Location	object

dtype: object

```
(1456714, 23)
```

```
Arrest
False    1079242
True      377472
Name: count, dtype: int64
```

Primary Type	
THEFT	329460
BATTERY	263700
CRIMINAL DAMAGE	155455
NARCOTICS	135240
ASSAULT	91289
OTHER OFFENSE	87874
BURGLARY	83397
DECEPTIVE PRACTICE	75495
MOTOR VEHICLE THEFT	61138
ROBBERY	57313
CRIMINAL TRESPASS	36912
WEAPONS VIOLATION	17233
PUBLIC PEACE VIOLATION	13122
OFFENSE INVOLVING CHILDREN	11398
PROSTITUTION	7633
CRIM SEXUAL ASSAULT	6823
INTERFERENCE WITH PUBLIC OFFICER	6195
SEX OFFENSE	4885
HOMICIDE	2649
ARSON	2217
GAMBLING	2212
LIQUOR LAW VIOLATION	1953
KIDNAPPING	1099
STALKING	828
INTIMIDATION	662
OBSCENITY	187
NON-CRIMINAL	93
CONCEALED CARRY LICENSE VIOLATION	90
PUBLIC INDECENCY	62
NON - CRIMINAL	38
OTHER NARCOTIC VIOLATION	30
HUMAN TRAFFICKING	28
NON-CRIMINAL (SUBJECT SPECIFIED)	4

Name: count, dtype: int64

Methodology

Problem Type

- Binary classification task: Predict whether a crime will result in an arrest

Tools & Libraries

- Pandas, NumPy – Data manipulation and cleaning
- Scikit-learn – Modeling, evaluation, and preprocessing
- Matplotlib, Seaborn – Visualizations
- XGBoost – Advanced tree-based classification

Models Used

- Logistic Regression - Interpretable baseline model, and good for linearly separable classes
- Decision Tree - Simple non-linear classifier and helps visualize decision paths
- XGBoost Classifier - Handles class imbalance, overfitting and is a strong performance on structured/tabular data
- Random forest - To handle large data and class imbalance better

Exploratory Data Analysis

Feature Categorization, grouped them into:

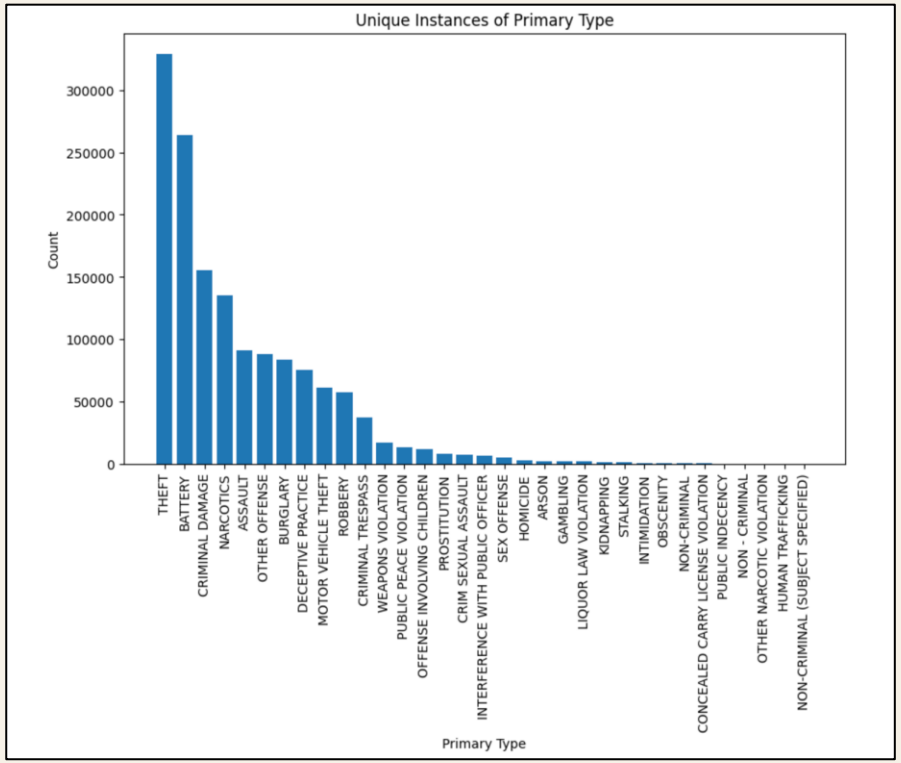
- Categorical: Crime type, arrest status, location description
- Numerical: District, Beat, Ward, Community Area
- Date/Time: Reported date, last update
- Location: Latitude, Longitude

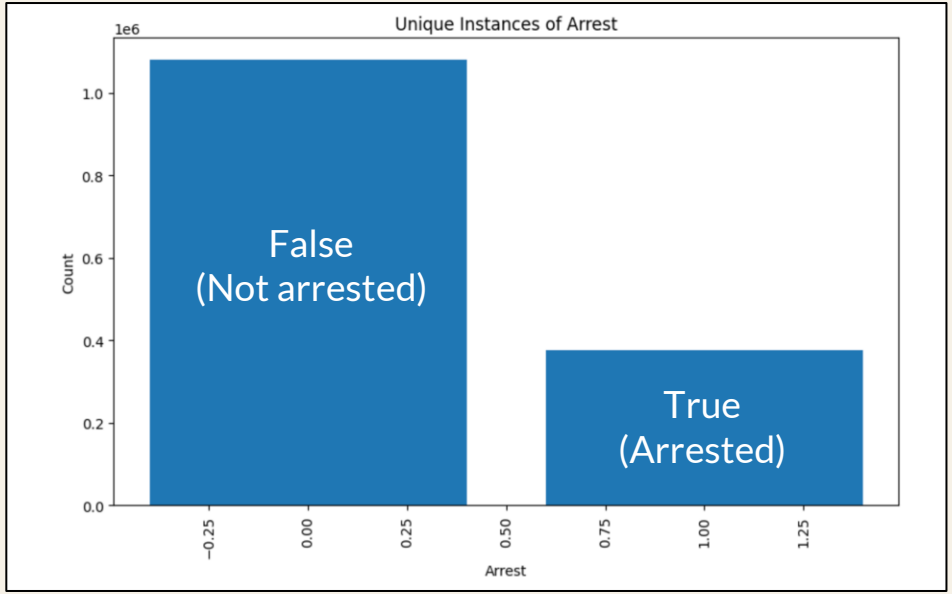
Initial Insights

- Theft and battery are the most common crimes.
- Most crimes are not domestic and don’t result in an arrest.
- Certain community areas consistently show higher crime counts.

Visual Exploration

- Used bar plots to visualize categorical feature frequencies.
- Noted clutter in visuals due to long category names, future visual improvements planned.


Unique Instances of Primary Type
Count
300000
250000
200000
150000
100000
50000
0
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
ASSAULT
OTHER OFFENSE
BURGLARY
DECEPTIVE PRACTICE
MOTOR VEHICLE THEFT
ROBBERY
CRIMINAL TRESPASS
WEAPONS VIOLATION
PUBLIC PEACE VIOLATION
OFFENSE INVOLVING CHILDREN
PROSTITUTION
CRIM SEXUAL ASSAULT
INTERFERENCE WITH PUBLIC OFFICER
SEX OFFENSE
HOMICIDE
ARSON
GAMBLING
LIQUOR LAW VIOLATION
KIDNAPPING
STALKING
INTIMIDATION
OBSCENITY
NON-CRIMINAL
CONCEALED CARRY LICENSE VIOLATION
PUBLIC INDECENCY
NON - CRIMINAL
OTHER NARCOTIC VIOLATION
HUMAN TRAFFICKING
NON-CRIMINAL (SUBJECT SPECIFIED)
Primary Type


Unique Instances of Arrest
1e6
Count
1.0
0.8
0.6
0.4
0.2
0.0
False
(Not arrested)
True
(Arrested)
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Arrest

Data Preprocessing - Cleaning & Selection

Loaded raw dataset (2012-2017 Chicago crime data)

Checked for missing values

- Dropped all rows with any missing values
- Still retained 1M+ records

Removed unnecessary features:

- Dropped: Unnamed: 0, Case Number, Date, Description, Updated On, Location, Latitude, Longitude
- Reason: Redundant identifiers or overly specific info that's not critical for prediction

Target variable: Arrest (binary classification: True/False)

- Predicting whether an arrest was make for a reported crime

Unnamed: 0	0
ID	0
Case Number	1
Date	0
Block	0
IUCR	0
Primary Type	0
Description	0
Location Description	1658
Arrest	0
Domestic	0
Beat	0
District	1
Ward	14
Community Area	40
FBI Code	0
X Coordinate	37083
Y Coordinate	37083
Year	0
Updated On	0
Latitude	37083
Longitude	37083
Location	37083

dtype: int64

(1456714, 23)

-38349

Unnamed: 0	0
ID	0
Case Number	0
Date	0
Block	0
IUCR	0
Primary Type	0
Description	0
Location Description	0
Arrest	0
Domestic	0
Beat	0
District	0
Ward	0
Community Area	0
FBI Code	0
X Coordinate	0
Y Coordinate	0
Year	0
Updated On	0
Latitude	0
Longitude	0
Location	0

dtype: int64

(1418365, 23)

Data Preprocessing - Feature Engineering & Encoding

Categorized columns into:

- Categorical: Primary Type, Location Description, Domestic
- Numerical: IUCR, Ward, District, Beat, Community Area, Year
- Dropped unnecessary date/time and location feature for this model

Feature encoding:

- Applied one-hot encoding to categorical features
- Converted boolean Domestic and Arrest values for modeling
- Normalized numerical data using StandardScaler for consistency

Result:

- Clean and well structured dataset ready for modeling and regression/classification tasks

	IUCR	Primary Type	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	Year
0	0486	BATTERY	APARTMENT	1	1	1022	10.0	24.0	29.0	2016
1	0486	BATTERY	RESIDENCE	0	1	313	3.0	20.0	42.0	2016
2	0470	PUBLIC PEACE VIOLATION	STREET	0	0	1524	15.0	37.0	25.0	2016
3	0460	BATTERY	SIDEWALK	0	0	1532	15.0	28.0	25.0	2016
4	0820	THEFT	RESIDENCE	0	1	1523	15.0	28.0	25.0	2016

	Arrest	Domestic	Beat	District	Ward	Community Area	Year	Primary Type_ARSON	Primary Type_ASSAULT	Primary Type_BATTERY
0	1	1	-0.185476	-0.181792	0.083027	-0.395454	1.510226	0	0	1
1	0	1	-1.210577	-1.195622	-0.206959	0.211174	1.510226	0	0	1
2	0	0	0.540336	0.542373	1.025482	-0.582108	1.510226	0	0	0
3	0	0	0.551903	0.542373	0.373013	-0.582108	1.510226	0	0	1
4	0	1	0.538890	0.542373	0.373013	-0.582108	1.510226	0	0	0

...	IUCR_5094	IUCR_5110	IUCR_5111	IUCR_5112	IUCR_5113	IUCR_5114	IUCR_5121	IUCR_5130	IUCR_5131	IUCR_5132
...	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0

Baseline model

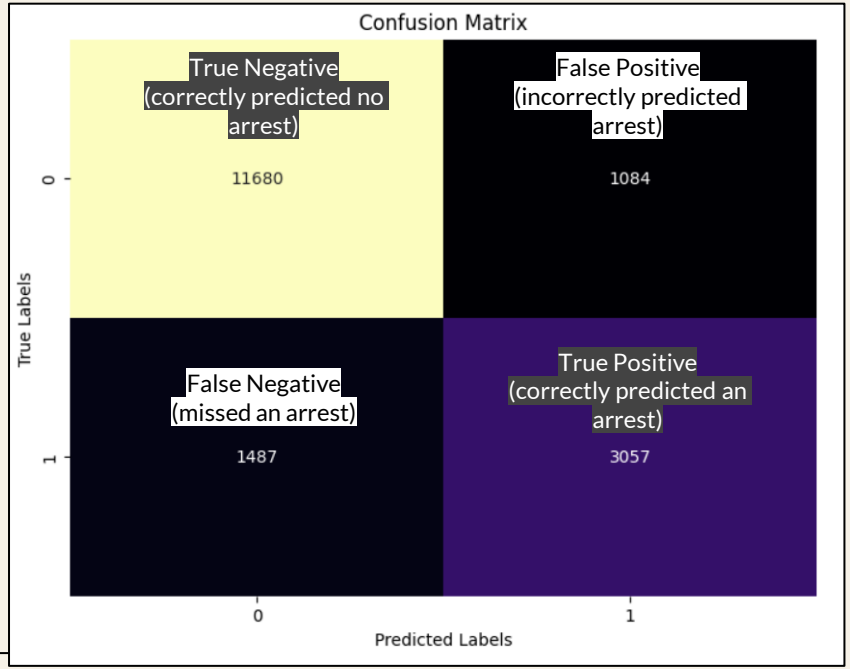
Data Preparation

- Using the cleaned
- Split into features (X) and target (y = Arrest)
- Applied an 80/20 train-test split

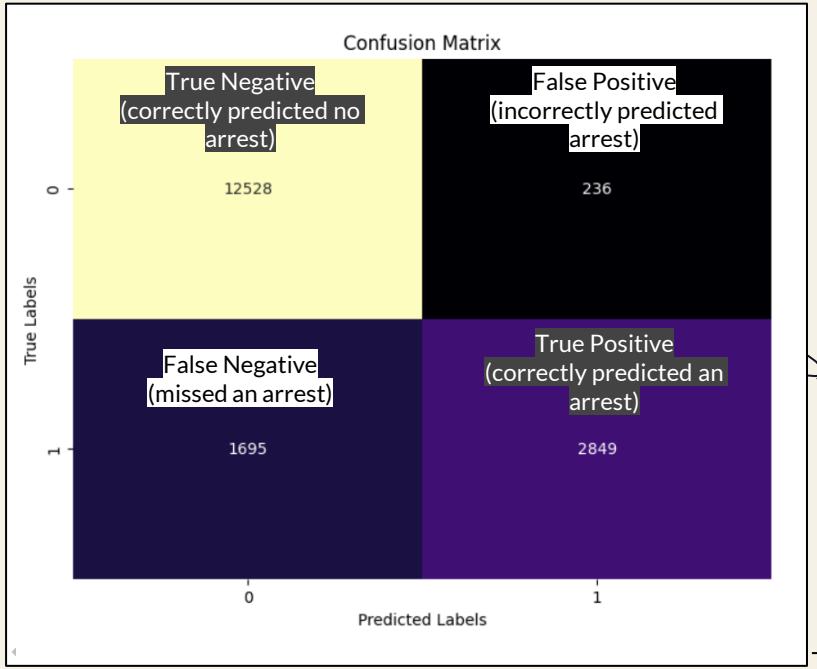
Models trained

- Decision Tree Classifier
  - Trained on full feature set
  - Evaluated using accuracy
- Confusion Matrix
  - True Positives: Correctly predicted arrests
  - False Positives: Predicted arrest when none happened
- Logistic Regression
  - Baseline linear classifier for binary target
  - Set max_iter=1000 to ensure convergence
  - Evaluated using accuracy, classification report, and confusion matrix

Decision Tree Classifier:



Logistic Regression:



Decision Tree Classifier:

```
F1: 0.703972366148532
```

Logistic Regression:

```
F1: 0.7468868790142876
              precision    recall  f1-score   support

           0       0.88      0.98      0.93     12764
           1       0.92      0.63      0.75      4544

    accuracy                           0.89     17308
   macro avg       0.90      0.80      0.84     17308
weighted avg       0.89      0.89      0.88     17308
```

Resized data for 2015

	Arrest	Domestic	Beat	District	Ward	Community Area	Year	Primary Type_ARSON	Primary Type_ASSAULT
0	0	0	0.550457	0.542373	1.025482	-0.582108	0.807598	0	0
1	0	1	1.260365	1.266538	1.822943	1.844403	0.807598	0	0
2	0	0	1.131685	1.121705	1.532957	-1.468719	0.807598	0	0
3	0	0	-0.027879	-0.036959	0.083027	-0.535445	0.807598	0	0
4	1	0	-0.314156	-0.326625	-1.439399	1.097784	0.807598	0	0

Primary Type_BATTERY	...	IUCR_5094	IUCR_5110	IUCR_5111	IUCR_5112	IUCR_5113	IUCR_5114	IUCR_5121	IUCR_5130	IUCR_5131	IUCR_5132
0	...	0	0	0	0	0	0	0	0	0	0
1	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0

XGBoost Model Summary

Why XGBoost?

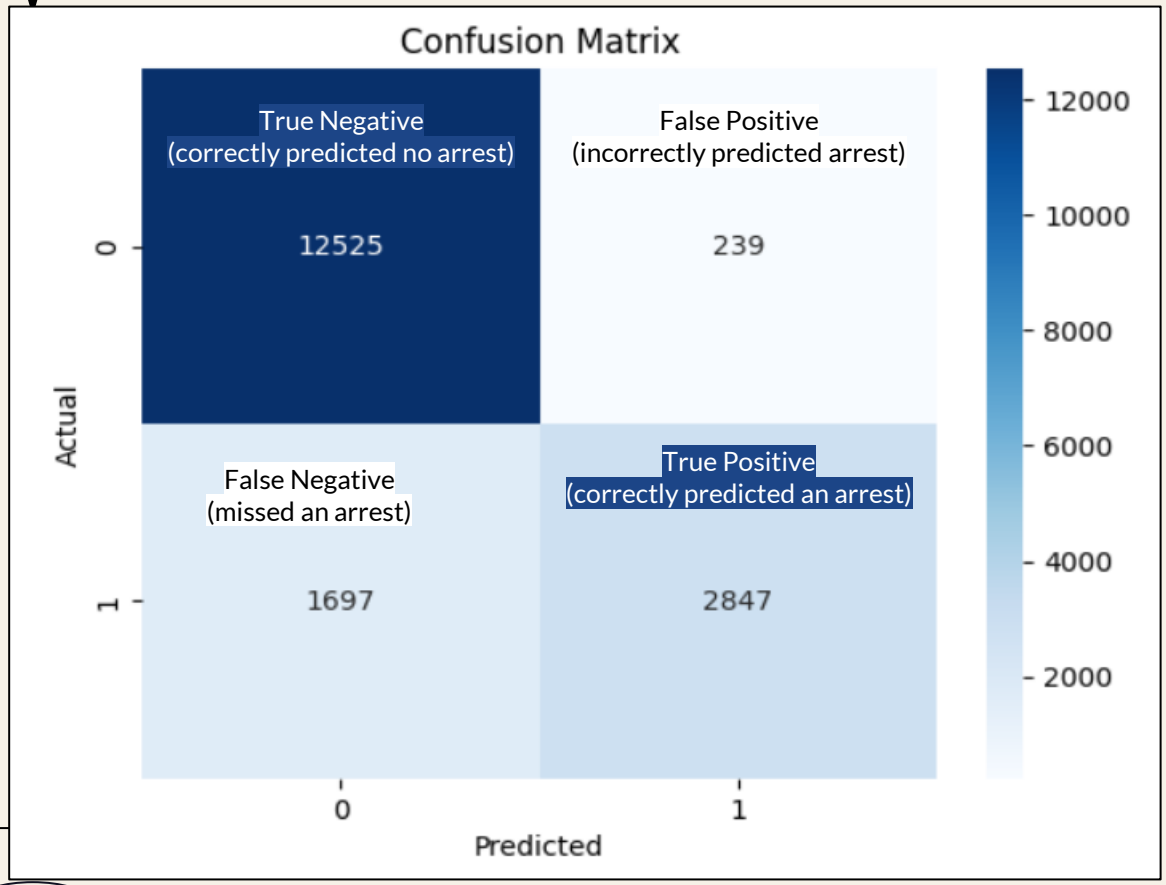
- Gradient boosting algorithm known for high performance and speed, great for large, imbalanced datasets.

Training Strategy:

- Tested on 5,000 and 10,000 randomly sampled rows
- Used XGBClassifier with binary logistic objective to predict Arrest

Model Evaluation:

- Confusion matrix used to visualize classification results
- Improved over baseline and performed competitively with Random Forest

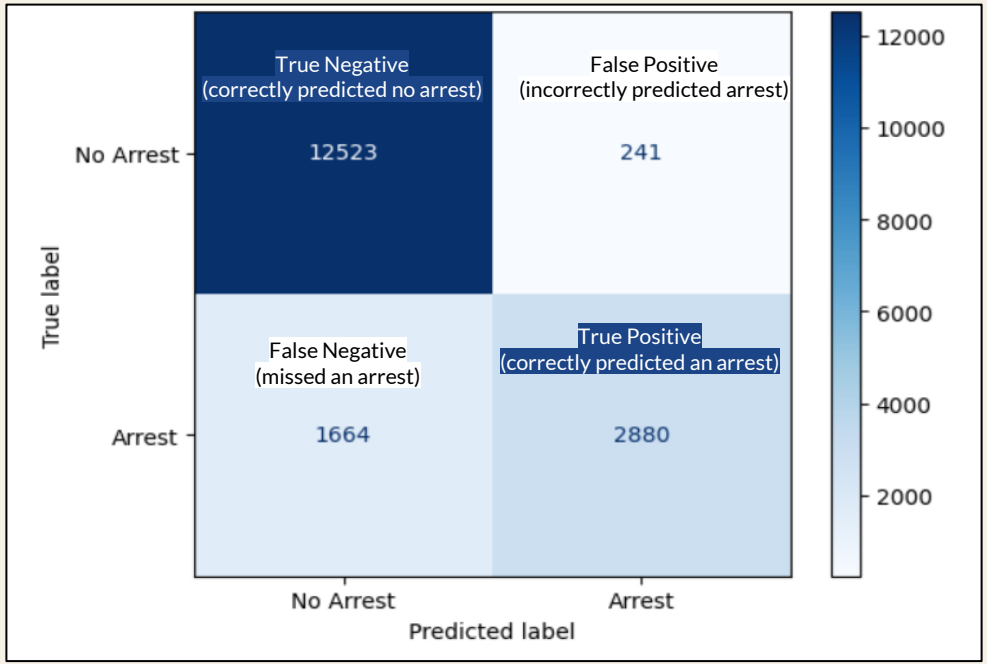


Scores (Whole dataset) rows:
Accuracy: 0.8881
F1 Score: 0.8373
Precision: 0.9016
recall: 0.8039

Scores with 5000 rows:
Accuracy: 0.8870
F1 Score: 0.8398
Precision: 0.8904
recall: 0.8109

Scores with 10,000 rows:
Accuracy: 0.8830
F1 Score: 0.8341
Precision: 0.8746
recall: 0.8090

GridSearch XG boosting - after tuning




```
Best parameters found:  {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300}
Best F1 score found:  0.7468436391118257
```

```
Best cross-validated F1 score: 0.7468436391118257
```

Random Forest Model Summary

Why Random Forest?

- A more advanced ensemble model compared to logistic regression, handles large datasets, and class imbalance better.

Training Strategy:

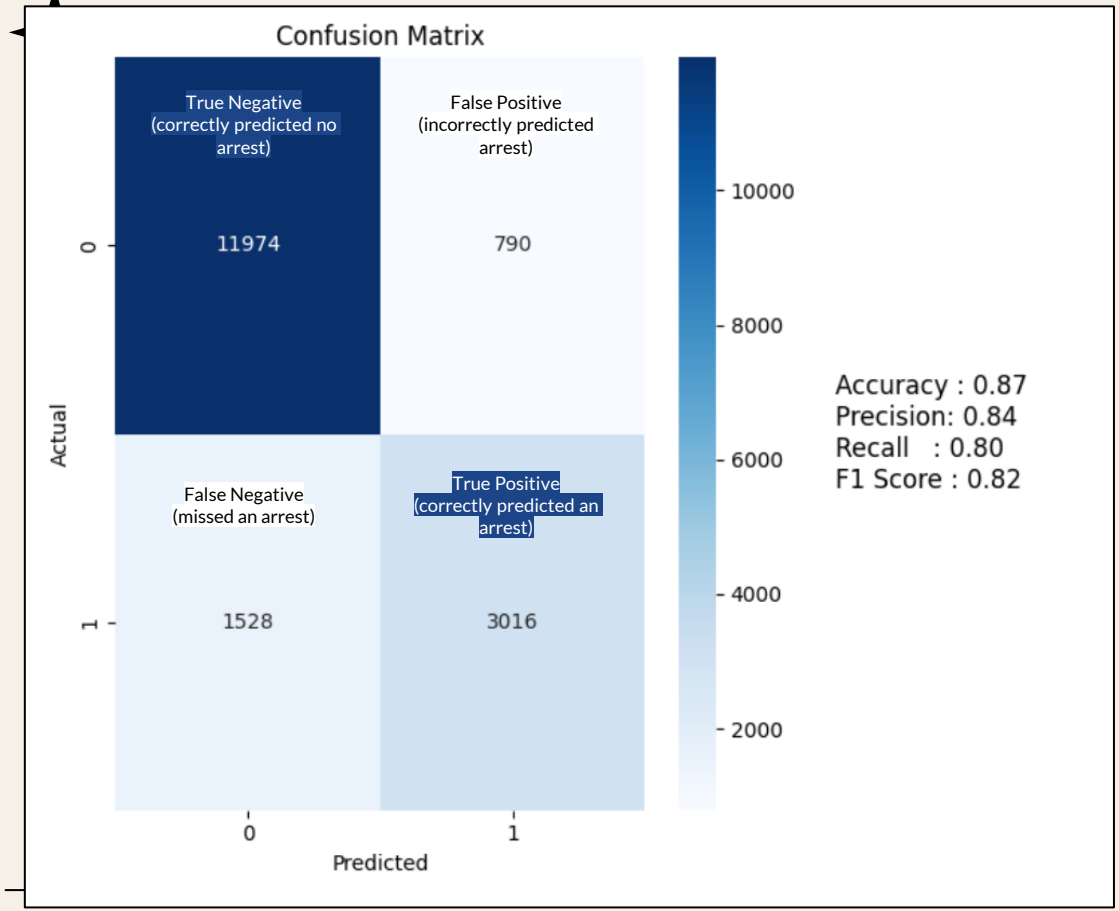
- Tested on two sample sizes: 5,000 and 10,000 records for quicker evaluation
- Final model trained on reduced dataset (around 80,000 rows)

Confusion Matrix:

- Visualized predictions vs. actual arrest outcomes
- Helps identify false arrests and missed arrests in prediction

Hyperparameter Grid Search:

- Searches over a grid of RF hyperparameters using 3-fold cross-validation to find the combination that gives the best F1 score on the training data set.



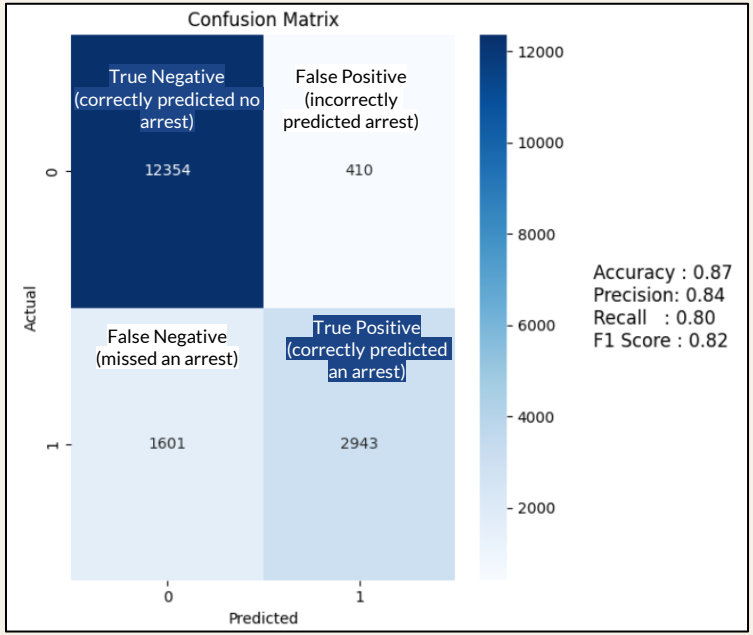
(86537, 543)

Original Metrics:
Accuracy : 0.8661
Precision: 0.8396
Recall : 0.8009
F1 Score : 0.8171

Sample 1 Metrics:
Accuracy : 0.8640
Precision: 0.8399
Recall : 0.7965
F1 Score : 0.8141

Sample 2 Metrics:
Accuracy : 0.8700
Precision: 0.8418
Recall : 0.8077
F1 Score : 0.8223

Grid Search Random Forest - after tuning




```
Fitting 3 folds for each of 12 candidates, totalling 36 fits
Best Hyperparameters: {'max_depth': None, 'min_samples_split': 10, 'n_estimators': 300}
Best CV F1:0.7455

Test F1 : 0.7453463340509054
Confusion Matrix:
 [[12354   410]
 [ 1601  2943]]
```

Logistic Regression Model Summary

Why Logistic Regression?

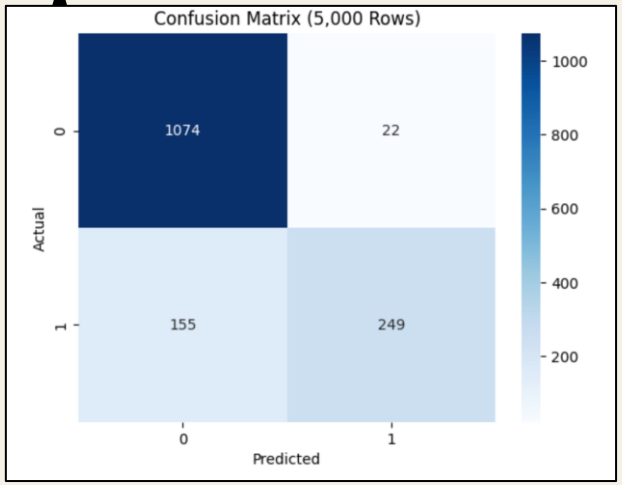
- Simple, interpretable for binary classification. Useful for understanding feature impact.

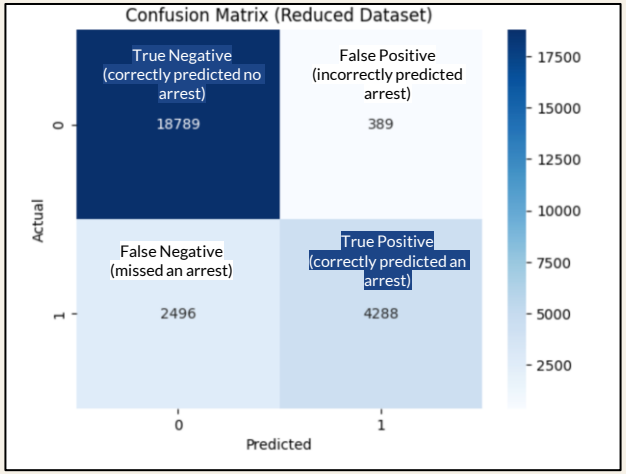
Training Strategy:

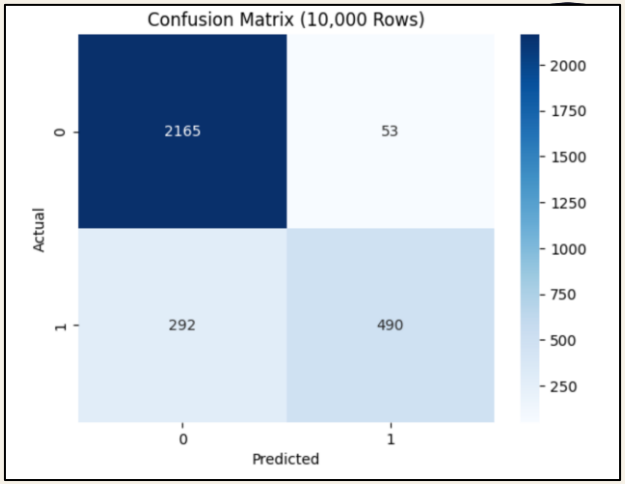
- Tested on the whole reduced dataset, 5,000 and 10,000 randomly sampled rows for efficiency.
- Used LogisticRegression from scikit-learn to predict likelihood of arrest.

Model Evaluation:

- Confusion matrix, F1 score, and classification report used to assess performance.
- Showed reasonable F1 score, but limited in handling complex patterns and imbalance.



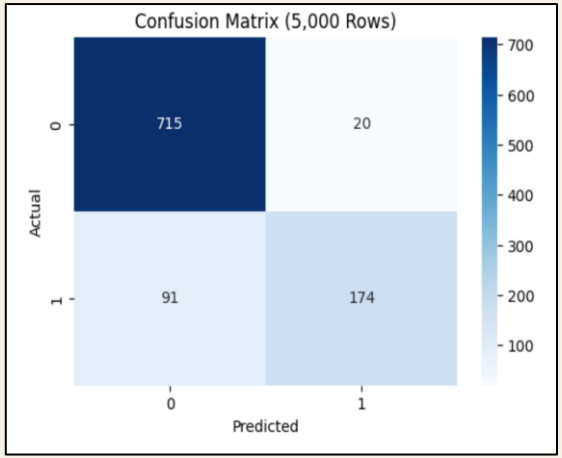


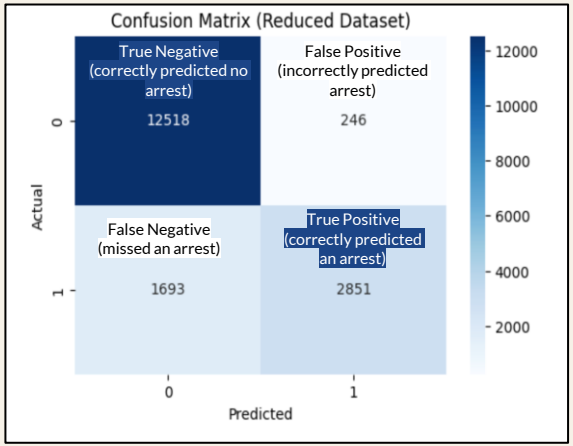


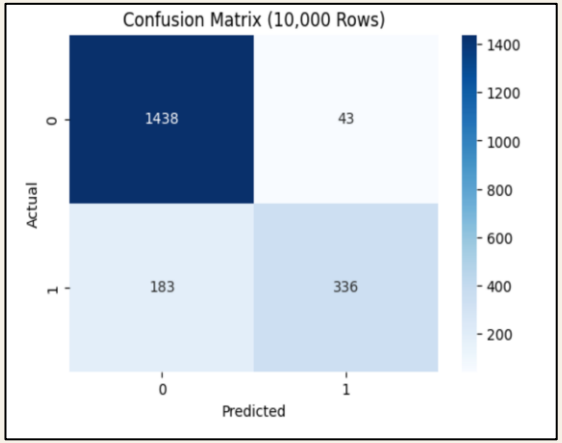
Model Performance Summary:

Dataset	Accuracy	Precision	Recall	F1 Score
Reduced Dataset	0.888876	0.916827	0.632075	0.748277
10,000 Rows	0.885000	0.902394	0.626598	0.739623
5,000 Rows	0.882000	0.918819	0.616337	0.737778

Grid Search Logistic Regression - after tuning



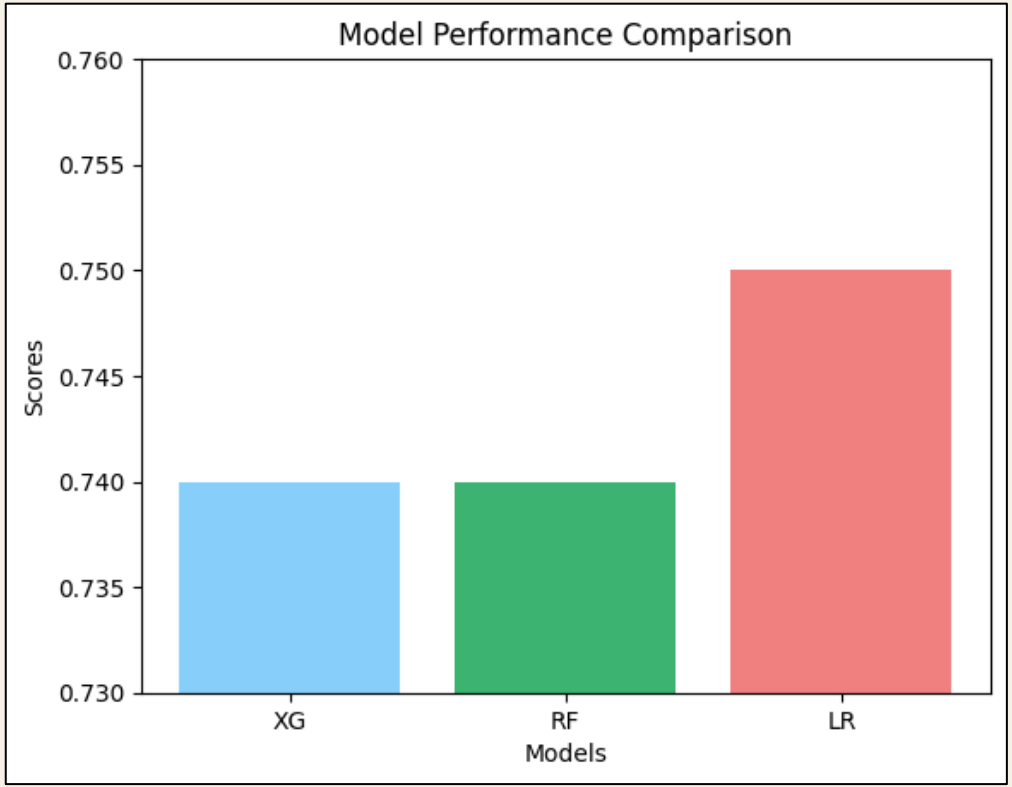




Tuned Model Performance Summary:

Dataset	Accuracy	Precision	Recall	F1 Score
Reduced Dataset	0.887971	0.920568	0.627421	0.746237
10,000 Rows	0.887000	0.886544	0.647399	0.748330
5,000 Rows	0.889000	0.896907	0.656604	0.758170

Final Model Selection



Best model: Logistic Regression

Based on: F1 score

Why?

- Metric that combines recall & precision

Challenges

Data Quality

- Missing values required row dropping
- Some features (like location, descriptions) too noisy or inconsistent

Forced to Reduce

- Due to its size, we needed to reduce the dataset which affected our overall results
- The loss of the data limited how much we were able to thoroughly conclude

Imbalanced Classes

- Most crimes did not lead to arrest so we had a skewed target distribution

Limitations & Future Work

What Could Be Improved

- Include more recent data (post-2017) for current trends
- Add external features: demographic info, police staffing, surveillance
- Use feature selection to reduce dimensionality and multicollinearity

Future Modeling Ideas

- Try ensemble models, neural networks, or cost sensitive learning
- Apply clustering to identify hidden crime/arrest patterns
- Explore spatial/temporal modeling for dynamic predictions

Real-World Impact

- Help police departments allocate resources more efficiently
- Improve transparency on what factors lead to arrests
- Inform policy changes to reduce disparities in law enforcement

Questions? Comments?

Thank you!