Predicting Arrests in Chicago Crime Data

In Chicago, predicting arrests based on historical crime data is crucial for improving public safety and policing strategies. Leveraging advanced statistical and machine learning methods can uncover patterns and key factors influencing arrest likelihood in reported crime incidents.

The team of researchers includes
Ankita Tripathy
Fabrizio Petrozzi
Muhammad Hammaz,
Priyanka Jammu
Sai Praneetha Sigharam
Subham Mohanty

Importance of Public Safety

Predicting Arrest Likelihood

Using crime data to identify key factors that increase the chances of an arrest, such as crime type, location, and time.

Smarter Policing Strategies

Helps allocate resources effectively, focusing on high-risk locations and times for proactive intervention to reduce crime.

Enhanced Community Safety

Data-driven insights enable faster responses, improving arrest rates and ultimately leading to a safer community.

Key Questions Driving Our Analysis

1 **What factors influence the likelihood of an arrest?**

Analyzing crime characteristics and contextual variables to identify the key predictors of arrests.

2 **How can we create accurate arrest prediction models?**

Evaluating different modeling approaches to find the most effective technique for forecasting arrests.

3 **What is the business value of accurate arrest predictions?**

Examining how this analysis can support public safety initiatives and resource allocation decisions.

Data Characteristics

Data Sources

The dataset is sourced from **the Chicago Data Portal**, a comprehensive repository of public data related to the city of Chicago.

Data Types

The dataset contains a mix of data types, including boolean, float, integer, and object.

Data Size

The Chicago Crime dataset contains **221,635 rows and 24 columns,** providing a substantial amount of information for analysis.

1

2

3

4

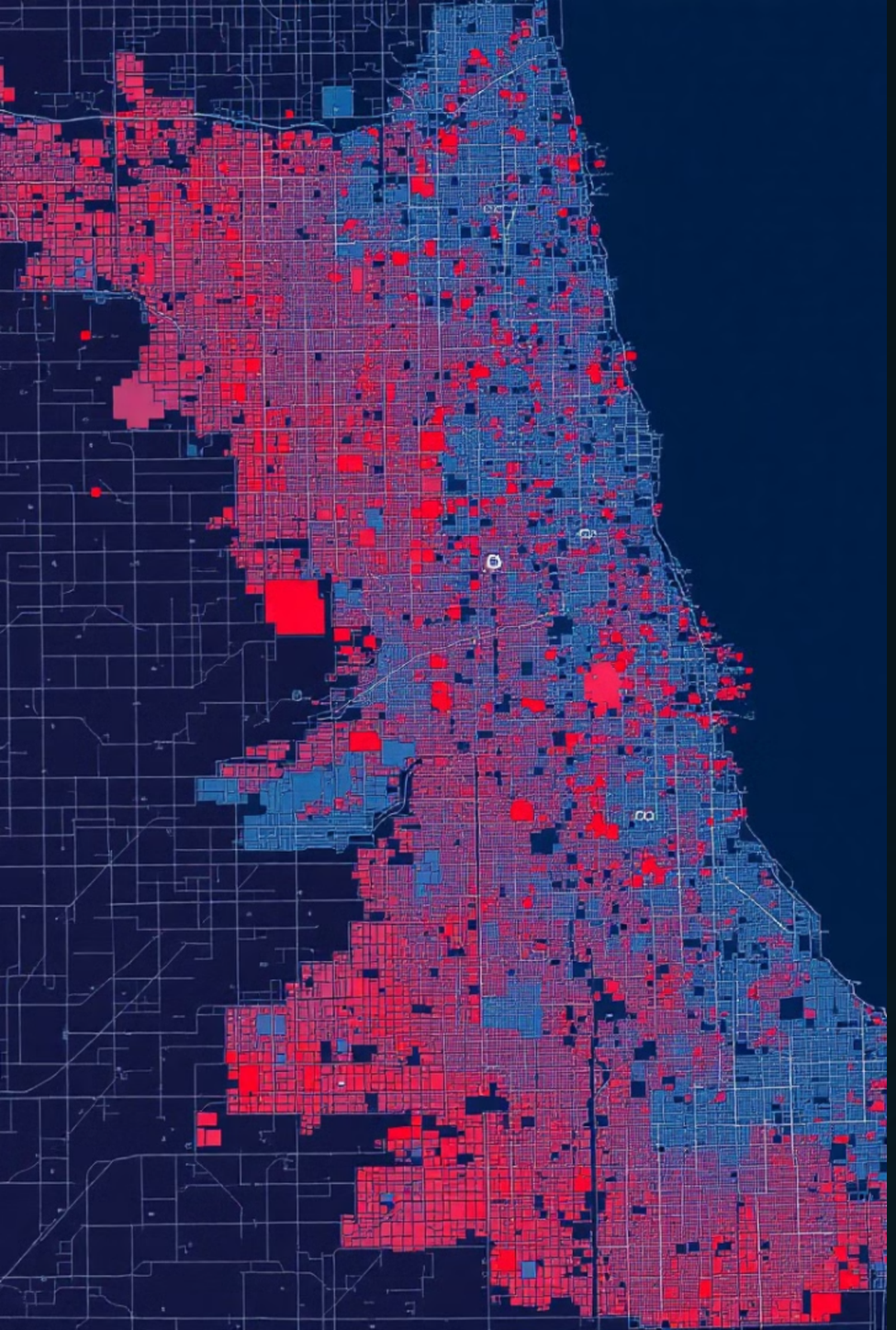
5

Dependent Variables

The main dependent variable in the dataset is Arrest, which indicates whether an arrest was made for a given incident.

Independent Variables

The key independent variables (DV's) include Primary Type, Domestic, Community Area, Year, and Location Description.



Data Preparation Process

1

Exploratory Data Analysis

Analyze the dataset to understand its characteristics, identify patterns, and uncover potential issues.

2

Class Imbalance Handling

Use oversampling techniques to balance the classes and improve the model's ability to predict the minority class outcomes..

3

One-Hot Encoding

Convert categorical features into a numerical format using one-hot encoding to prepare the data for modeling.

4

Standardization

Standardize the features to ensure they are on a similar scale, which can improve the model's performance.

5

Train-Test Split

Divide the dataset into training and testing sets to evaluate the model's performance on unseen data.

Modeling Approaches

Logistic Regression

Logistic Regression is a simple yet effective classification algorithm that is easy to interpret. It helps in understanding the significance of different predictors, such as "Primary Type" and "Domestic," and their impact on the likelihood of an arrest.

Random Forest

Random Forest is a robust ensemble model that effectively manages both numerical and categorical features. It is resistant to noise and overfitting due to its aggregation of multiple decision trees.

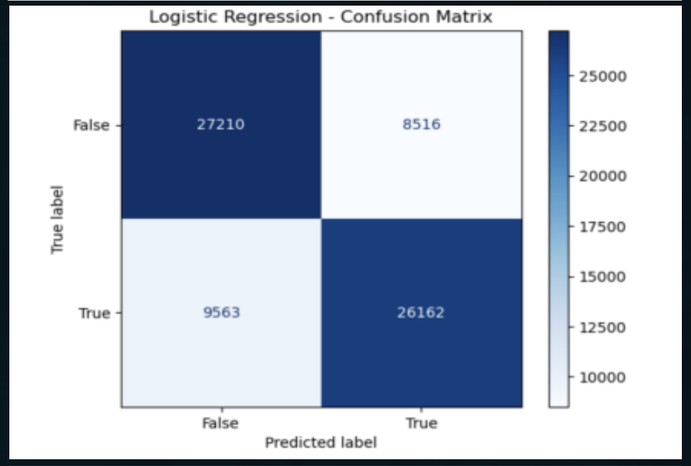
Neural Networks

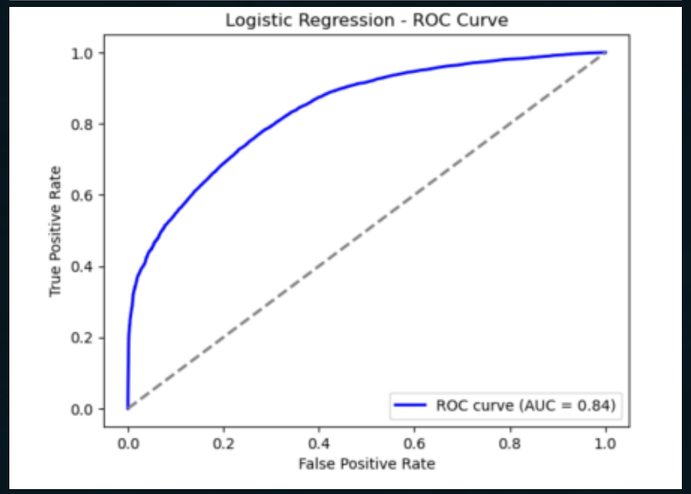
Neural Networks are adept at modelling complex, non-linear relationships between features and the target variable, making them suitable for high-dimensional datasets.

Evaluation of Modeling Approaches

Logistic Regression

	precision	recall	f1-score	support
False	0.74	0.76	0.75	35726
True	0.75	0.73	0.74	35725
accuracy			0.75	71451
macro avg	0.75	0.75	0.75	71451
weighted avg	0.75	0.75	0.75	71451

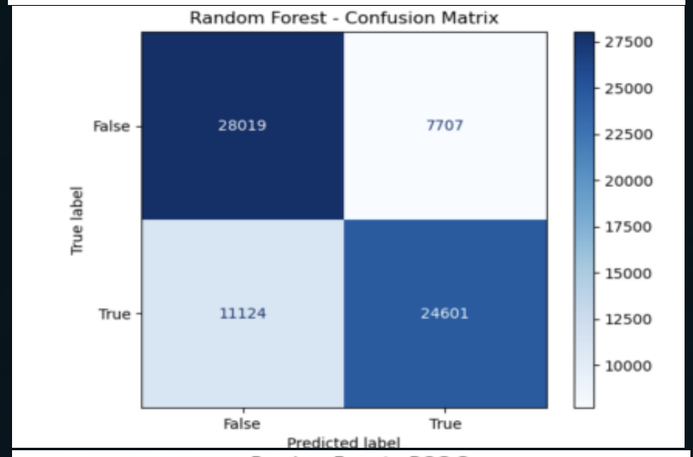


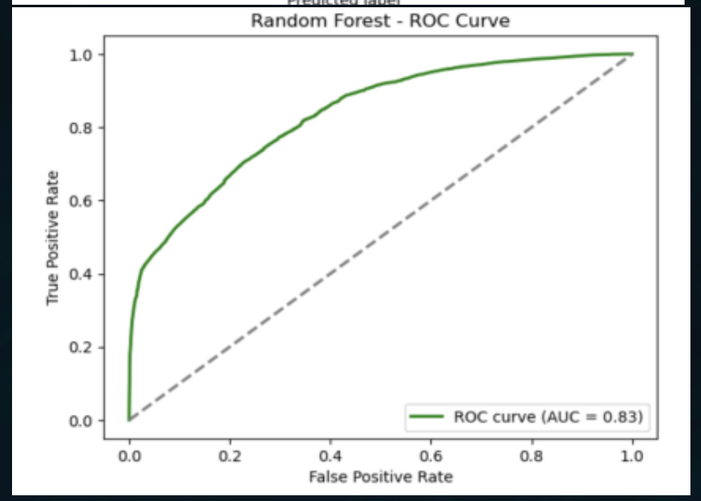


Random Forest

Random Forest Classification Report:

	precision	recall	f1-score	support
False	0.72	0.78	0.75	35726
True	0.76	0.69	0.72	35725
accuracy			0.74	71451
macro avg	0.74	0.74	0.74	71451
weighted avg	0.74	0.74	0.74	71451

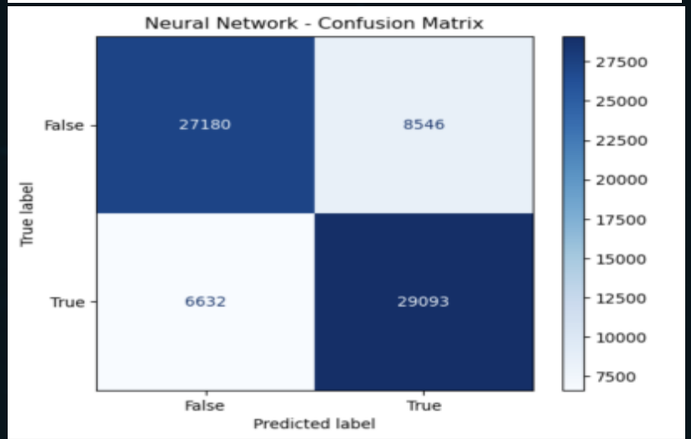


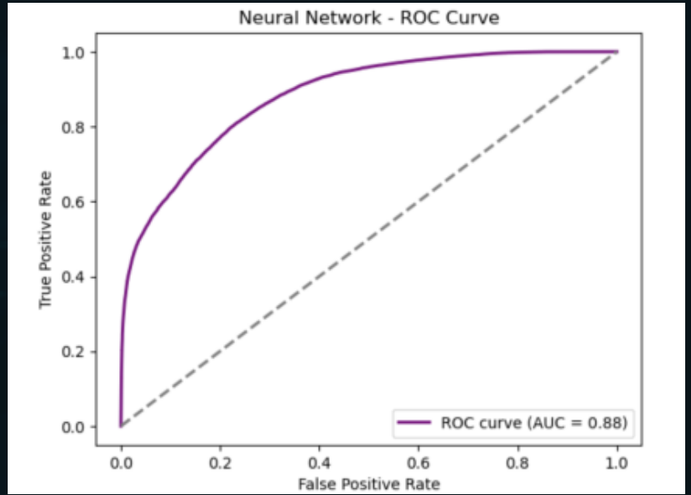


Neural Network

Neural Network Classification Report:

	precision	recall	f1-score	support
False	0.80	0.76	0.78	35726
True	0.77	0.81	0.79	35725
accuracy			0.79	71451
macro avg	0.79	0.79	0.79	71451
weighted avg	0.79	0.79	0.79	71451





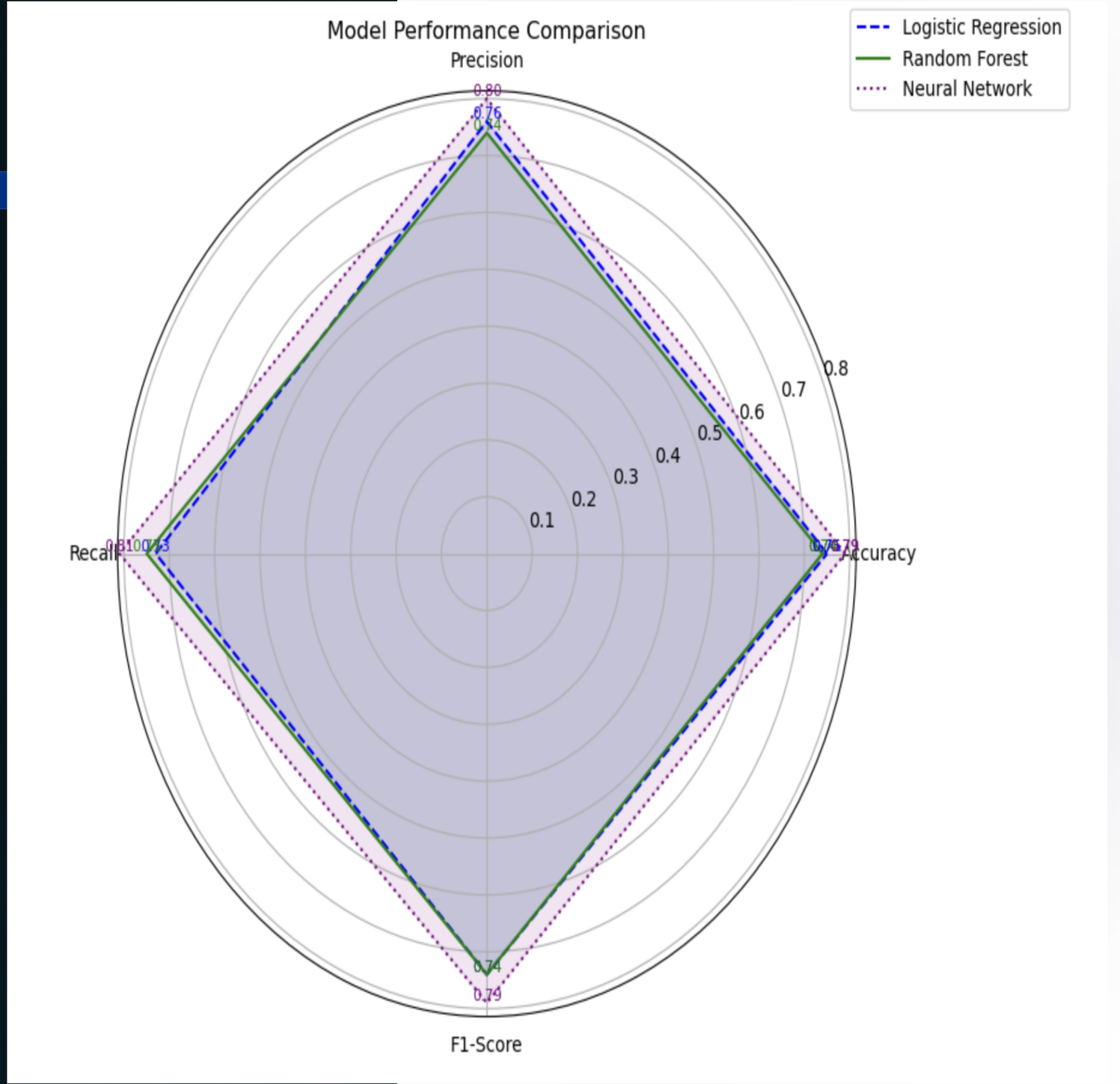
Best Performing Model

Neural Network Model

The Neural Network model emerges as the best-performing option, with the highest values across all the key metrics. Its balanced performance and ability to generalize well make it the optimal choice for this analysis.

Reasons for Selection

- Superior Metrics: Highest Accuracy, Precision, Recall, and F1-Score
- Balanced Performance: Minimizes false positives while capturing true positives
- Generalized Performance: Demonstrates strong ability to generalize across the dataset



Business Value of Accurate Arrest Prediction

1

Resource Allocation

Targeted deployment of police officers and other public safety resources to high-risk areas.

2

Proactive Interventions

Identifying potential crime hotspots and implementing preventive measures.

3

Improved Outcomes

Reducing crime rates, increasing arrest rates, and enhancing overall public safety.

4

Operational Efficiency

Automate analysis to save time and costs

5

Policy and Planning Support

Inform long-term decisions on resource distribution

Conclusion and Next Steps

1 Continuous Model Refinement

Incorporating new data and feedback to iteratively improve the arrest prediction models.

2 Expanded Analysis

Exploring the drivers of different crime types and developing specialized models for each.

3 Collaborative Partnerships

Engaging with law enforcement agencies to integrate the models into their decision-making processes.

References

- https://medium.com/analytics-vidhya/predicting-arrests-looking-into-chicagos-crime-through-machine-learning-78697cc930b9
- https://www.kaggle.com/datasets/chicago/chicago-crime
- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data
- https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html
- https://github.com/Mayank-004/Boston-Crime
- https://github.com/rahulbordoloi/Predict-Crime-Rate-in-Chicago/blob/master/Predict_Crime_Rate_in_Chicago.
- https://www.kaggle.com/code/datajack1234/predicting-crime-rate-in-chicago-using-prophet
- https://www.kaggle.com/code/threadid/chicago-crimes-regression-neural-network

TASK DISTRIBUTION	CONTRIBUTED BY
DATASET, CODE AND MODEL SELECTION	PRIYANKA & HAMMAZ
DOCUMENTATION	ANKITA & SUBHAM
PRESENTATION	SUBHAM & FABRIZIO
SCRIPT FOR THE PRESENTATION	SAI PRANEETHA & FABRIZIO