Final Group Project
Predicting Arrests in Chicago Crime Data

Presented by
Ankita Tripathy
Fabrizio Petrozzi
Muhammad Hammaz
Priyanka Jammu
Sai Praneetha Sigharam
Subham Mohanty

MS in Artificial Intelligence (AI) and Business Analytics,

University of South Florida

ISM6251.004F24.96281 Machine Learning

Under the Guidance of Reza Ebrahimi

November 20th, 2024

INTRODUCTION

Crime prediction and prevention are critical for ensuring public safety and effective police administration. Chicago, one of the largest cities in the United States, these challenges are particularly complex due to the city's diverse socio-economic landscape and varying crime rates across neighbourhood. By leveraging historical crime data, it becomes possible to predict certain outcomes, such as the likelihood of arrests in reported crime incidents. This capability is invaluable for enhancing policing strategies and shaping policies. Analysing and predicting arrests based on historical crime records from Chicago data that includes details such as crime type, location,date, time, and arrest status can be effectively approached using advanced statistical and machine learning methods. These techniques help uncover patterns and identify significant factors influencing arrests.

The target variable in this project utilizes the "Arrest" field to identify whether an arrest could be made for a given crime using the method of supervised learning. This objective is achieved by the use of models such as Logistic Regression, Random Forest, and Neural Networks, whose performances are evaluated on metrics such as accuracy, precision, recall, and F1-score.

OBJECTIVE

This paper aims to utilize machine learning models to predict arrests based on Chicago's crime data. Specifically, it seeks to achieve the following objectives:

- Identifying patterns related to arrests within historical crime records.
- Developing predictive models to classify incidents as resulting in an arrest or not.
- Assessing the performance of these models to identify the most effective approach for the task.

By accomplishing these goals, law enforcement agencies can enhance their ability to prioritize cases and allocate resources more efficiently.

INDUSTRY TARGETED: PUBLIC SAFETY & LAW ENFORCEMENT-WHY IT MATTERS

The public safety and law enforcement industry is critical for ensuring community well-being, reducing crime rates, and maintaining trust between citizens and authorities. By leveraging predictive analytics, this project addresses some of the most pressing challenges faced by law enforcement agencies:

Predicting Arrest Probabilities- Analysing crime data—such as type, location, date and time of occurrence—helps identify situations where arrests are more likely. This empowers law enforcement to make informed decisions, enabling better planning and timely interventions.

Optimized Policing Strategies-Data-driven insights support the efficient allocation of limited resources, allowing authorities to focus on high-risk areas and times. This improves both the effectiveness and impact of policing efforts.

Strengthened Community Safety-Leveraging predictive analytics enhances law enforcement's ability to respond quickly and effectively, increasing arrest rates and reducing overall crime. These advancements foster public trust and contribute to safer communities.

ATTRIBUTES AND DATA OVERVIEW

The data is taken from **Chicago Data Portal**

Main Crime Data Source - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present-Map/ahwe-kpsy

The dataset comprises 220408 rows and 22 features that capture details about crimes and their outcomes.

Description of features:

- ID: Unique identifier for the crime record.
- Case Number: Unique Chicago Police Department record ID.
- Date: Timestamp indicating when the crime occurred.
- Block: Partially redacted address of the crime location.
- IUCR & Primary Type: Illinois Uniform Crime Reporting code and general classification of the crime (e.g., THEFT, BATTERY).
- Description: Detailed subcategory of the crime.
- Location Description: Type of location where the crime occurred
- Arrest & Domestic: Flags indicating if an arrest was made or if the incident was domestic-related.
- Beat & District: Police geographic areas where the crime occurred.
- Ward & Community Area: City council district and community area of the crime.
- FBI Code: FBI classification of the crime type.
- Coordinates: X, Y, latitude, and longitude of the crime location
- Year & Updated On: Year of the crime and the last update to the record.
- Location: Combined latitude and longitude for mapping.

To obtain community names, we are loading the community data and mapping it to the crime dataset based on the community area number.
Community Data Source- https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
Description of features:

- AREA_NUMBE: Represents the unique identifier for each community area.
- COMMUNITY: Name of the corresponding community area.

For this analysis, the most relevant attributes include:
Independent Variables (IV):

- Primary Type: General classification of crime (e.g., theft, assault).
- Location Description: Type of location where the crime occurred.
- Community Area: Geographic location identifier.
- Year: Year of the incident.
- Domestic: Whether the crime involved domestic violence.

Dependent Variable (DV):

Arrest: Indicates whether an arrest was made (True/False).

KEY QUESTIONS DRIVING OUR ANALYSIS

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

This question focuses on identifying the key predictors of arrest outcomes. By analysing crime characteristics (e.g., crime type, location, time) and contextual variables, we aim to uncover patterns that influence whether an arrest is made. Understanding these factors is critical for refining predictive models and improving decision-making.

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

The objective here is to evaluate different machine learning approaches to develop effective and reliable predictive models. This involves testing algorithms such as Logistic Regression,

Random Forest, and Neural Networks to identify the most accurate and robust model for forecasting arrests.

3. **What is the business value of accurate arrest predictions?**
Accurate predictions have practical implications for public safety and resource optimization. They can guide smarter allocation of law enforcement resources, enable proactive interventions in high-risk areas, and ultimately support policies that enhance community safety and operational efficiency.

DATA PREPROCESSING AND ANALYSIS

1. Exploratory Data Analysis (EDA): Theft and battery were the most common types of crimes. Domestic-related crimes showed a higher likelihood of resulting in arrests. Certain community areas consistently exhibited higher crime and arrest rates, marking them as potential hotspots for interventions. Arrest rates varied across years, possibly reflecting changes in policing strategies or crime reporting practices.

2.Handling Missing Data: A small percentage (<1%) of values were missing in features like Location Description and Coordinates. Due to the minimal proportion of missing data, we opted to remove the missing values, resulting in an insignificant loss of data.

3.Handling Class Imbalance (Oversampling Method): The target variable, Arrest, was highly imbalanced, with 80% non-arrest cases and 20% arrest cases. Used resampling techniques to oversample the minority class (Arrest = True). Balanced the dataset to achieve equal representation of both classes. Balancing significantly enhanced the models' ability to predict the minority class, improving overall performance.

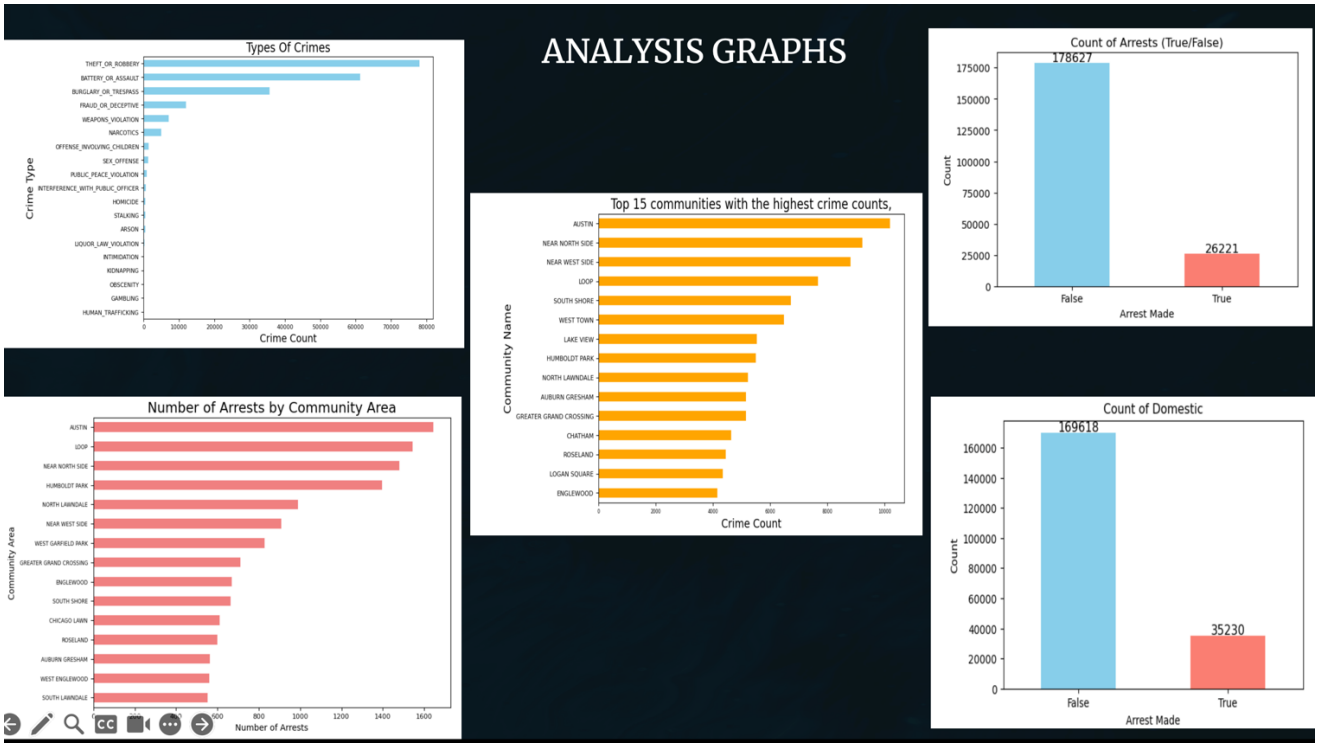
4.Encoding Categorical Variables: Machine learning algorithms cannot directly process categorical variables like Primary Type and Location Description. Applied one-hot encoding to transform categorical variables into binary features. Created separate binary columns for each category to enable efficient model processing. New binary columns were added, representing categorical variables numerically for effective analysis.

5.**Feature Scaling:** Numerical features like Community Area and Year have varying scales, which could bias distance-based models such as Neural Networks. Used Min-Max Scaling to normalize numerical features within a 0–1 range. Ensured all features contributed equally to the models. Feature scaling improved model performance and accelerated convergence during training.

6. Train and Test Split: Evaluate model performance on unseen data. Divided the dataset into training (80%) and testing (20%) sets. Ensured stratified splits to maintain class distribution in both sets. Created robust datasets for training and evaluating models effectively.

Below are the multiple graphs, each analysing different aspects of crime data which helps in understanding the data better:

1. Types of Crimes(Top Left)
2. Top 15 Communities with Highest Crime Counts(Top Centre)
3. Count of Arrests - True/False (Top Right)
4. Number of Arrests by Community Area(Bottom Left)
5. Count of Domestic Crimes - True/False(Bottom Right)



MODELS UTILIZED

Logistic Regression

- **Reason for Selection:** 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.
- **Optimal Use Case:** This model performs well when the relationship between predictors and the target variable is predominantly linear.
- **Primary Advantage:** Its interpretability makes it valuable for stakeholders seeking insights into the most influential factors.

Random Forest

- **Reason for Selection:** 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.
- **Optimal Use Case:** It excels in datasets with complex feature interactions and non-linear relationships.
- **Primary Advantage:** Random Forest provides rankings of feature importance, aiding in identifying the key factors affecting arrest probability.

Neural Networks (MLP Classifier)

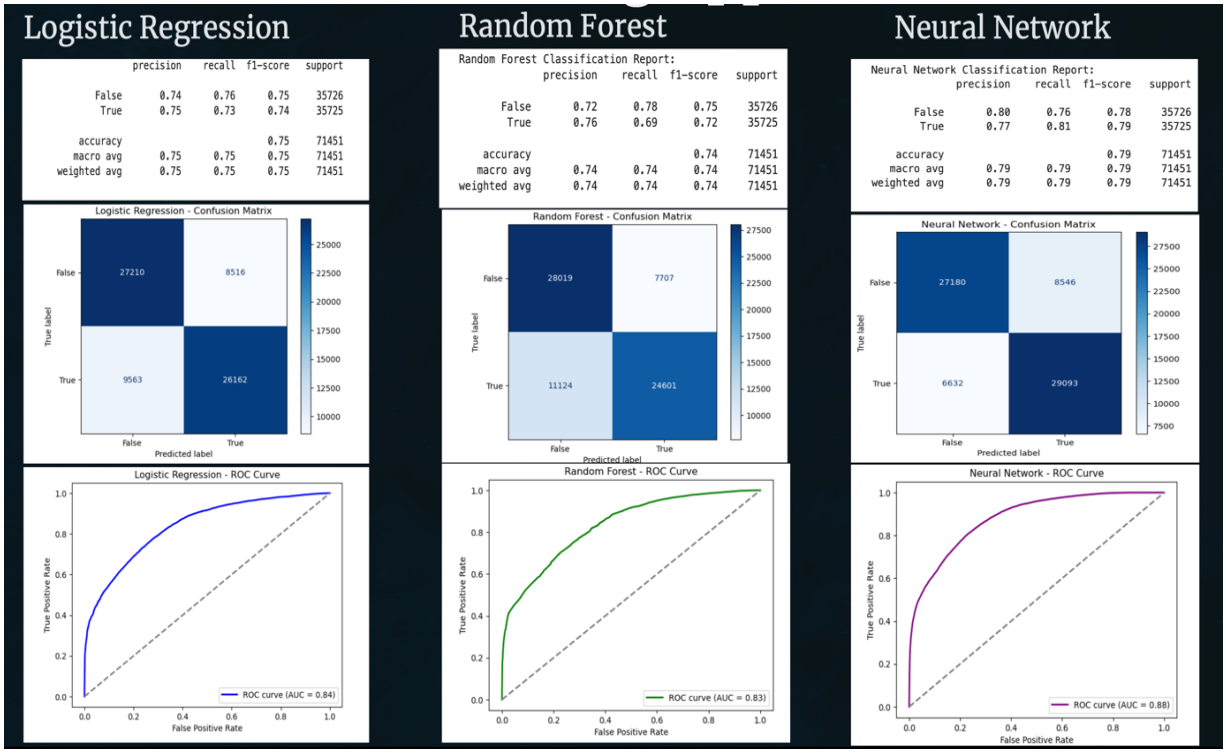
- **Reason for Selection:** Neural Networks are adept at modelling complex, non-linear relationships between features and the target variable, making them suitable for high-dimensional datasets.
- **Optimal Use Case:** These models perform best with large, well-preprocessed datasets that address issues such as scaling and class imbalance.
- **Primary Advantage:** Neural Networks are highly powerful compared to simpler models but have limited interpretability, which can be a drawback.

EVALUATION METRICS

The following metrics were used to assess the performance of the models:

- **Accuracy:** Represents the proportion of correctly classified outcomes, including both arrests and non-arrests.
- **Precision:** Measures the proportion of true positives (correctly predicted arrests) out of all predicted positives. High precision ensures a lower rate of false positives.
- **Recall (Sensitivity):**Indicates the proportion of actual arrests correctly identified by the model. A high recall implies fewer false negatives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.
- **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):** Evaluates model performance across various classification thresholds. A higher AUC signifies better discrimination between the positive (arrest) and negative (non-arrest) classes.

MODEL PERFORMANCE



Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	75%	0.74	0.76	0.75	0.84
Random Forest	74%	0.72	0.78	0.75	0.83
Neural Network	79%	0.80	0.76	0.78	0.88

The performance of the models is summarized as follows, providing insights into their effectiveness based on various evaluation metrics:

Logistic Regression

- **Accuracy (75%)**: Indicates that 75% of all cases (arrests and non-arrests) were correctly classified.

- **Precision (0.74)**: About 74% of cases predicted as arrests were true arrests, meaning a moderate rate of false positives.
- **Recall (0.76)**: The model correctly identified 76% of actual arrests, but some arrests were missed (false negatives).
- **F1-Score (0.75)**: Balances precision and recall, showing that the model performs well overall but could improve in handling false positives or false negatives.
- **ROC-AUC (0.84)**: The model has decent discrimination ability between arrest and non-arrest cases but lags behind more advanced models.

Random Forest

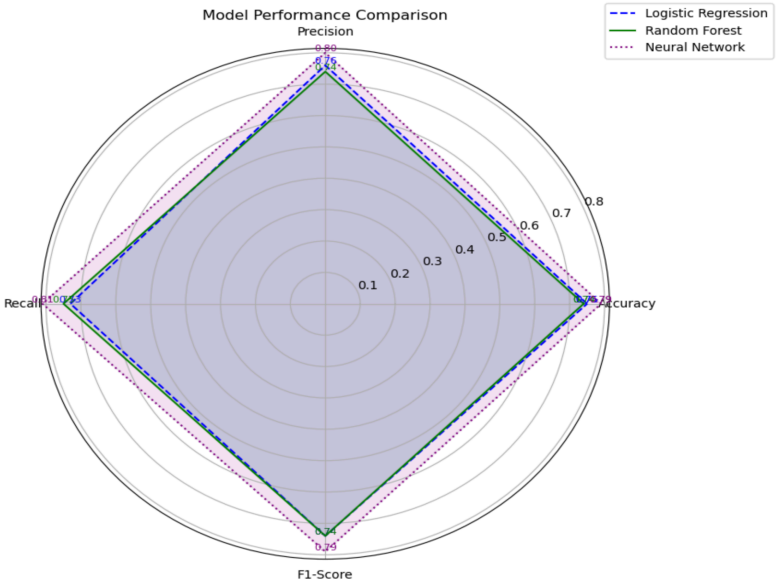
- **Accuracy (74%)**: Demonstrates higher accuracy, with 79% of cases correctly classified.
- **Precision (0.72)**: With 72% precision, it has fewer false positives compared to Logistic Regression.
- **Recall (0.78)**: The model successfully identifies 78% of actual arrests, improving recall compared to Logistic Regression.
- **F1-Score (0.75)**: Shows a good balance between precision and recall, making it a strong performer.
- **ROC-AUC (0.83)**: Indicates better separation of arrest and non-arrest cases.

Neural Network

- **Accuracy (79%)**: Matches the accuracy of Random Forest, classifying 79% of cases correctly.
- **Precision (0.80)**: The highest precision among the three models, meaning it predicts arrests more reliably with fewer false positives.
- **Recall (0.76)**: Performs on par with Random Forest in identifying actual arrests.
- **F1-Score (0.78)**: Maintains a balanced performance similar to Random Forest, excelling in both precision and recall.
- **ROC-AUC (0.88)**: The highest AUC score, showing the best ability to distinguish between arrest and non-arrest cases across different thresholds.

BEST PERFORMING MODEL- RADAR CHART

The radar chart compares the performance of three machine learning models (Logistic Regression, Random Forest, and Neural Network) across key metrics: Precision, Recall, F1-Score, and Accuracy.



- Logistic Regression (Blue - Dashed Line): A reliable baseline model, but lacks the capacity to capture complex patterns in the data.
- Random Forest (Green - Solid Line): A solid improvement over Logistic Regression, especially in precision and overall accuracy.
- Neural Network (Purple - Dotted Line): Despite slightly lower precision, the Neural Network demonstrates the best overall performance due to its superior recall and ability to generalize complex relationships in the data.

Reasons for Neural Network's Selection:

- Superior Recall: Excels in identifying true arrests, which is critical for crime prediction applications.
- Balanced Metrics: While precision is slightly lower, the overall F1-score and recall compensate by reducing false negatives.
- Generalization Capability: The Neural Network handles complex, non-linear patterns effectively, making it the most robust option for diverse scenarios in Chicago crime data.

HOW THE MODEL ANSWERS THE PROPOSED QUESTIONS

1.What factors influence the likelihood of an arrest?

The models (Logistic Regression, Random Forest, and Neural Network) analyse historical crime data, identifying key variables that impact arrest likelihood, such as: Primary Type of crime (e.g., theft, assault),Domestic Violence involvement, Location Description (e.g., street, residence),Time and Year of the crime.

2.How can we create accurate arrest prediction models?

By training and testing three models, Logistic Regression serves as a baseline to provide simple, interpretable results. Random Forest introduces robust ensemble learning for handling non-linear relationships. Neural Networks excel at modelling complex, high-dimensional interactions in the data.

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

The chosen Neural Network model provides actionable insights:
Resource Allocation: High-risk times and locations identified by the model guide targeted deployment of law enforcement.
Proactive Intervention: Predictive insights help prevent crimes by focusing on areas with a higher likelihood of arrests.
Improved Public Safety: Enhanced arrest predictions can reduce crime rates and improve community trust.
Operational Efficiency: Automates data analysis, saving time and enabling data-driven decision-making for law enforcement and policymakers

BUSINESS VALUES

- Accurate arrest predictions provide significant value by helping law enforcement focus resources on high-probability cases, improving efficiency and response times.
- By utilizing insights from key factors such as crime type, location, and time, police departments can develop proactive strategies that enhance arrest rates and public safety.
- These predictions enable data-driven decisions that create safer communities while optimizing operational effectiveness.

CONCLUSION

In conclusion, our analysis highlighted the importance of predictive analytics in improving arrest likelihood predictions, guided by factors such as crime type, location, and time, which were critical in enhancing model performance. Crime type provided insight into the nature and severity of incidents, location helped identify high-risk areas where law enforcement resources could be better concentrated, and time revealed patterns that allowed for optimized scheduling of patrols. Advanced models like Neural Networks effectively captured these relationships, offering insights for smarter resource allocation and proactive policing strategies. These insights not only improve arrest predictions but also provide a practical way for law enforcement to make data-driven decisions that enhance public safety. Moving forward, refining the model with additional data, focusing on specific crime categories, and collaborating with police departments will help integrate these tools into real-world applications, creating safer and more secure communities.

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