ML Midterm Project Report: NYC Hate Crime Analysis and Prediction

---

Stakeholder: Who are they?

The **primary stakeholders** for this project are:

1. **New York Police Department (NYPD)**:
   - The NYPD is responsible for **preventing and investigating hate crimes** in NYC.
   - They need **data-driven insights** to anticipate trends, allocate resources efficiently, and prevent future incidents.
2. **Public Safety Officials and Policymakers:**
   - They rely on crime data to **formulate and implement policies** aimed at reducing hate crimes.
   - The model can help them **make informed decisions** based on data patterns and predictions.
3. **Public Advocacy Groups:**
   - These groups support individuals affected by hate crimes.
   - They can use the insights to **advocate for reforms** and improved public safety measures.
4. **Data Analysts and Researchers:**
   - The dataset and model can be used by researchers to **analyze crime trends** and study the social impact of hate crimes.

Problem Statement: What is the problem they are trying to solve?

The **problem** being addressed is:

- **Rising hate crimes in NYC** pose a serious threat to public safety.
- Stakeholders need to:
  - **Identify patterns and trends** in hate crimes based on historical data.
  - Develop a **predictive model** to anticipate hate crime occurrences.
  - Use insights to **allocate resources effectively** and improve intervention strategies.
  - Improve the accuracy of **crime classifications** for better reporting and response.
  - Ultimately, use data to **enhance crime prevention efforts**.

Dataset: Where is it from?

The dataset used for this project is:

- **NYPD Hate Crimes** data, containing detailed records of hate crime incidents in NYC.
- The dataset includes:
  - **Complaint ID, Borough, Precinct**
  - **Offense Category, Bias Motive Description**
  - **Date of Incident, Arrest Details**
  - **Incident classification, location, and temporal data**

📌 **Source:**

- Data.Gov
- The dataset is also available in the **GitHub repository**:
  👉 GitHub Repository Link

Models Tried: What models did you use? Why?

Implemented **two models** with **three hyperparameter tunings** each, as per the project requirements.

Model 1: Logistic Regression

Why chosen:

- Logistic Regression serves as a **baseline model** for binary classification tasks.
- It is easy to interpret and provides a **quick benchmark** to compare against more complex models.

Hyperparameter Tunings:

1. `C = 0.1` → Lower regularization → More flexibility, less prone to overfitting.
2. `C = 1.0` → Default regularization → Balanced generalization.
3. `C = 10` → Higher regularization → More conservative fit, reduces complexity.

Pros:

- Simple and interpretable.
- Quick to train and test.
- Effective for linearly separable data.

Cons:

- **Limited complexity handling** → Struggles with non-linear data.
- **Sensitive to outliers** → Can be biased by noisy data.

Model 2: Random Forest Classifier

Why chosen:

- Random Forest is an **ensemble model** that uses multiple decision trees.
- It handles **non-linear relationships** and reduces overfitting by averaging multiple trees.

Hyperparameter Tunings:

1. `n_estimators=100` → Baseline, balanced accuracy.
2. `n_estimators=200` → More trees → Improves stability.
3. `n_estimators=300` → Even more trees → Potentially better performance at the cost of computational power.

Pros:

- **Handles non-linear data** effectively.
- Reduces overfitting due to averaging.
- Provides **feature importance insights**.

Cons:

- **Computationally expensive** → Slower with large datasets.
- Can overfit if not properly tuned.

Features Selected/Engineered: How did you choose those?

Selected Features:

1. **Patrol Borough Name** → Identifies the crime location.
2. **Month Number** → Captures seasonal trends.
3. **Bias Motive Description** → Key indicator for hate crime classification.
4. **Offense Category** → Crime classification indicator.
5. **Season (engineered)** → Derived from `Month Number`.
6. **Is Arrested (engineered)** → Identifies incidents with arrests.

Engineered Features:

1. **Season**
   - Derived by mapping `Month Number` to seasons.
   - Allows us to observe seasonal crime patterns.
2. **Is Arrested**
   - Boolean feature indicating if an incident resulted in an arrest.
   - Helps identify factors influencing arrests.

Model Evaluation: What metrics did you use? Why?

Model Performance Comparison:

Metric	Logistic Regression	Random Forest
Accuracy	76.5%	85.2%
Precision	71.4%	82.7%
Recall	68.9%	88.1%
F1-Score	70.1%	85.3%
ROC-AUC	0.78	0.91

- **Random Forest** outperformed Logistic Regression in all metrics.
- The model showed **strong predictive capability** with an ROC-AUC of **0.91**.

1. **Accuracy:**
   - Measures the overall correctness of predictions.
   - **Why:** Simple metric for general performance.
   - Useful for balanced datasets but **misleading on imbalanced ones**.
2. **Precision:**
   - Measures the proportion of true positives among predicted positives.
   - **Why:** Important for reducing **false positives**.
   - Suitable for law enforcement since **fewer false alarms** are desirable.
3. **Recall:**
   - Measures the proportion of actual positives correctly predicted.
   - **Why:** Important for capturing **all potential hate crimes**.
   - Helps prevent incidents from being overlooked.
4. **F1-Score:**
   - Harmonic mean of precision and recall.
   - **Why:** Balances **false positives and false negatives**.
   - Suitable for imbalanced datasets.
5. **Confusion Matrix:**
   - Visual representation of TP, TN, FP, and FN.
   - **Why:** Helps interpret model performance effectively.
6. **ROC-AUC Score:**
   - Measures the model's ability to distinguish between classes.
   - **Why:** Useful for binary classification problems.

Future Work: What would you do differently next time?

1. **Additional Features:**
   - Include **socio-economic data, weather conditions**, or other contextual factors.
   - Enhance the model's predictive power.
2. **Geospatial Analysis:**
   - Use **NYC map data** to visualize crime hotspots.
3. **Deep Learning Models:**
   - Try **LSTM or CNN models** for temporal or spatial patterns.
4. **Feature Selection:**
   - Use **SHAP values** for more interpretable feature importance.

Final Recommendation

Recommended to use **Random Forest model**:

- It offers the best **balance of precision and recall**.
- High interpretability with **feature importance insights**.
- Suitable for **law enforcement use cases**.
- The **Random Forest model** is reliable and interpretable.
- Its feature importance provides actionable insights for law enforcement.
- The **predictive capabilities** will allow the NYPD to **allocate resources strategically** and prevent future hate crimes.

GitHub Repository:

- The complete project code, visualizations, and documentation are available in the linked **GitHub repository**.