Machine Learning Midterm Project Report: NYPD Hate Crimes Analysis

---

1. Introduction

Stakeholder and Problem Statement:

Our stakeholder for this project is the **NYPD Hate Crimes Task Force**, which aims to identify and address patterns of hate crimes in New York City. The task force needs a machine learning model to:

- **Predict hate crime occurrences** based on historical data.
- Identify patterns and features that contribute to these crimes.
- Improve resource allocation and crime prevention strategies.

Dataset Source and Description:

- The dataset used: **NYPD_Hate_Crimes.csv**
- Source: NYPD Open Data Portal
- The dataset contains **3,255 records** with **14 columns**, including:
    - **Crime details**: Complaint ID, year, month, precinct, and borough.
    - **Incident descriptions**: Law code, offense type, bias motive.
    - **Arrest information**: Arrest date and ID.

---

2. Data Preparation

Data Cleaning:

1. **Date formatting:**
    - Converted `Record Create Date` and `Arrest Date` to `datetime` format.
2. **Missing values:**
    - Handled missing values by imputing or removing where appropriate.
3. **Feature consistency:**
    - Standardized column names for easier referencing.

Feature Engineering:

1. **Seasons:**
    - Created a `Season` column from the `Month Number`.
    - Mapped months to corresponding seasons (Winter, Spring, Summer, Fall).
2. **Crime-to-arrest lag:**

    - Engineered a `Crime-to-Arrest Lag` column, measuring the time difference between the incident and the arrest date.
3. **Encoding:**
    - Applied one-hot encoding for categorical features.
4. **Scaling:**
    - Scaled numerical features using `StandardScaler`.

---

3. Exploratory Data Analysis (EDA)

Key Visualizations and Insights:

- **Seasonal Trends:**
    - Most hate crimes occurred in **Winter** and **Summer**.
- **Crime Distribution by Borough:**
    - Manhattan and Brooklyn had the highest reported hate crimes.
- **Offense and Bias Motive Analysis:**
    - Most common bias: **Anti-Jewish** and **Anti-Black**.
- **Arrests vs. Non-Arrests:**
    - A significant number of hate crime cases did not lead to arrests.

---

4. Modeling Phase

Model Selection and Reasoning:

We used the following models:

1. **Logistic Regression:**
    - Chosen for its simplicity and interpretability.
    - Good baseline for classification tasks.
2. **Random Forest Classifier:**
    - Chosen for its ability to handle complex, non-linear data.
    - Robustness and feature importance capabilities.

Hyperparameter Tuning:

We performed a `GridSearchCV` with **5-fold cross-validation** for both models.

- **Logistic Regression:**
    - `C`: [0.1, 1.0, 10.0]
    - `Penalty`: L1 and L2 regularization.
- **Random Forest:**
    - `n_estimators`: [100, 200, 300]
    - `max_depth`: [5, 10, 15]

Model Evaluation Metrics:

1. **Accuracy:** Measures overall correctness.
2. **Precision:** Percentage of correct positive predictions.
3. **Recall:** Ability to identify all positive cases.
4. **F1-Score:** Harmonic mean of precision and recall.
5. **ROC-AUC:** Measures model discrimination.

---

5. Results and Insights

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**.

Feature Importance:

- The most influential features were:
  - **Precinct Code**: Highly correlated with crime occurrences.
  - **Season (Winter & Summer)**: Seasonal trends significantly impacted crime occurrences.
  - **Crime-to-Arrest Lag**: Lag influenced the likelihood of arrest.

---

6. Recommendations and Future Work

Key Recommendations:

1. **Model Selection:**
   - The **Random Forest model** is recommended for deployment due to its superior performance and accuracy.
   - It offers higher recall, making it better for identifying hate crime patterns.
2. **Feature Insights:**
   - **Seasonal trends** indicate that law enforcement should allocate more resources during **Winter and Summer**.
   - **Precinct-level analysis** can help prioritize high-crime regions.
3. **Production Deployment:**
   - The model can be integrated into **real-time crime monitoring systems**.

Future Improvements:

1. **More Advanced Models:**
   - Try **XGBoost or LightGBM** for improved performance.
2. **Time Series Analysis:**
   - Model temporal patterns for better predictions.
3. **Geospatial Analysis:**
   - Use NYC geolocation data for deeper spatial insights.
4. **Data Augmentation:**
   - Incorporate external socio-economic factors (population, income) to enhance predictions.

Conclusion:

This project demonstrated the end-to-end machine learning workflow, including:

- **Data cleaning, feature engineering, and EDA**.
- **Model building, hyperparameter tuning, and evaluation**.
- **Insights and stakeholder recommendations**.

The **Random Forest model** emerged as the superior performer, making it the recommended solution for the stakeholder's use case.

GitHub Repository:

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