**Chicago Crime Rate Analysis – Machine Learning Evaluation Report**

**1. Project Objective**

The main goal of this project is to identify the factors influencing crime rates in Chicago and evaluate their impact. By analyzing crime data alongside weather conditions, population density, location, and temporal features, the project aims to determine under which circumstances crime increases, which types of crime are more sensitive to these factors, and whether future crime rates can be predicted accurately.

**2. Machine Learning Methodology**

The model used for prediction was the GradientBoostingRegressor from scikit-learn. Input features included:

* **Meteorological variables**: temperature, humidity, precipitation, windspeed, visibility, UV index
* **Temporal variables**: month, day of the week, holiday status, daylight hours
* **Lag features**: crime rate 1 and 7 days prior
* **Location**: Community Area ID

Data was standardized using StandardScaler, and categorical variables (e.g., day name) were encoded using LabelEncoder. The dataset was split into **80% training** and **20% testing** sets.

**3. Model Performance**

The model’s effectiveness was evaluated with the following metrics:

* **R² Score:** **0.30**  
  This score indicates that the model explains around 30% of the variance in daily crime rates. While this is a meaningful starting point, the relatively low R² suggests that many influential factors are either missing from the dataset or highly complex.
* **MAE (Mean Absolute Error):** **~1.55 crimes per 100,000 population**  
  On average, the model's predictions deviate from the true value by approximately 1.55 crimes per 100,000 people. This level of accuracy may be acceptable for broader trends but is limited for precise, localized forecasting.

**4. Prediction Example and Observations**

An illustrative forecast was made for **May 12, 2025**. For **Community Area 18**, the results were:

* **Actual Crime Count**: *X*
* **Predicted Crime Count**: *Y*
* **Absolute Error**: *Z*

While some community areas showed close alignment between actual and predicted values, others demonstrated significant deviations. These gaps may be due to factors not captured in the dataset, such as local events, police activity, or socioeconomic disruptions.

**5. Key Limitations**

* **Lag Features**: While useful, lag features can propagate errors and rely on temporal continuity, which may not always hold.
* **Missing Variables**: Real-time social factors (e.g., public events, economic stress) are not included.
* **Data Granularity**: Weather data was aggregated daily, which may obscure sub-daily crime trends.
* **Imputation**: Missing values were handled using default methods (fillna), potentially introducing bias.

**6. Hypothesis Evaluation**

* **H1 (Weather Impact):** Supported – Higher temperatures are associated with increased crime; rain slightly reduces crime but shifts its type.
* **H2 (Demographics & Location):** Partially supported – Longitude correlates with crime patterns, but variables like income and race show weak predictive power.
* **H3 (Crime Type Sensitivity):** Supported statistically, but not reflected in the ML model (which only predicts total crime rate, not by type).

**7. Recommendations**

* Incorporate hourly weather and crime data for finer granularity.
* Build classification models for different crime types.
* Use interpretable ML techniques (e.g., SHAP) to identify feature importance.
* Apply cross-validation and hyperparameter tuning.
* Test time-series models like SARIMA or Facebook Prophet.

**8. Conclusion**

The model provides a solid baseline for understanding how crime rates are influenced by environmental and temporal factors. However, its limited predictive power underscores the complexity of crime as a phenomenon. Future work should focus on incorporating richer data, improving interpretability, and modeling specific crime categories for more actionable insights.