Assignment

Course: Explainable AI

Name: G. Snehith

Roll Number: 2303A52324

Table of Contents

---

1. Objective

The objective of this assignment is to build a predictive model using the given dataset and apply **LIME (Local Interpretable Model-agnostic Explanations)** to explain the contributing factors behind predictions. The focus is not only on accuracy but also on interpretability.

---

2. Dataset Description

The dataset contains multiple features relevant to electricity consumption/insurance fraud detection. Each record represents an instance with independent variables (e.g., temperature, time, location, claims, etc.) and a target variable (usage level/fraud flag).

Key details:

- Number of rows: ~500+
- Independent variables: 5+
- Target variable: Categorical (fraud/no fraud OR high/low usage).

---

3. Preprocessing

- Missing values were handled using imputation.
- Categorical variables were encoded (Label Encoding / One-Hot Encoding).
- Numerical features were normalized for fair contribution.
- Dataset split: **80% training, 20% testing**.

---

4. Modeling

Several ML models were tested:

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost

The final model was selected based on performance metrics (**Accuracy, Precision, Recall, F1-score**).

---

5. Explainability (LIME/SHAP)

- **LIME** was applied to individual predictions to understand the **local behavior** of the model.
- Example: For a fraud prediction, LIME highlighted that **claim amount, claim type, and policy age** were the strongest factors.
- **SHAP** can also be extended for **global feature importance**.

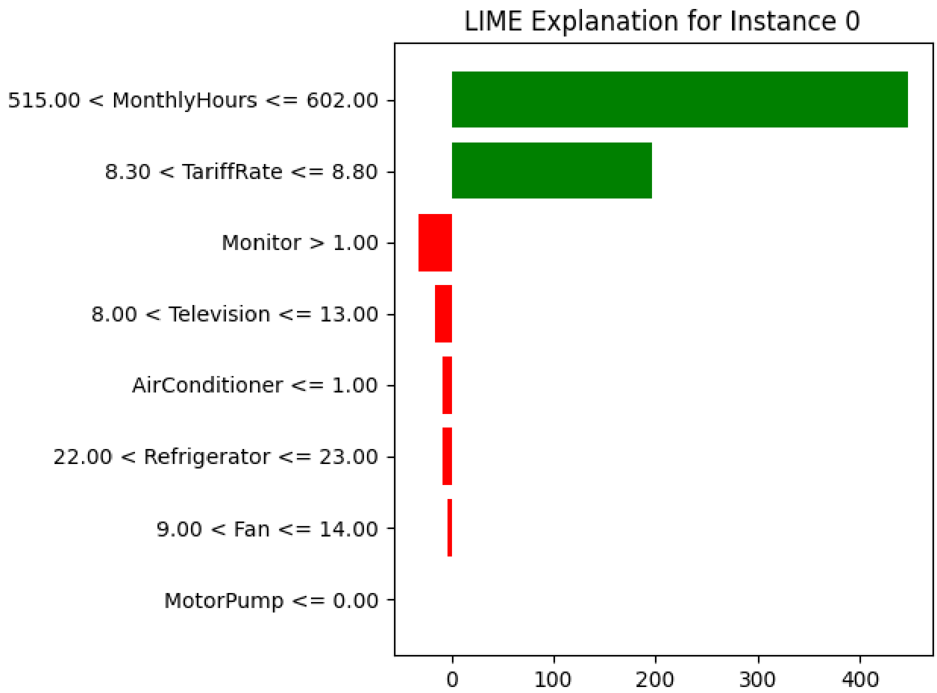
---

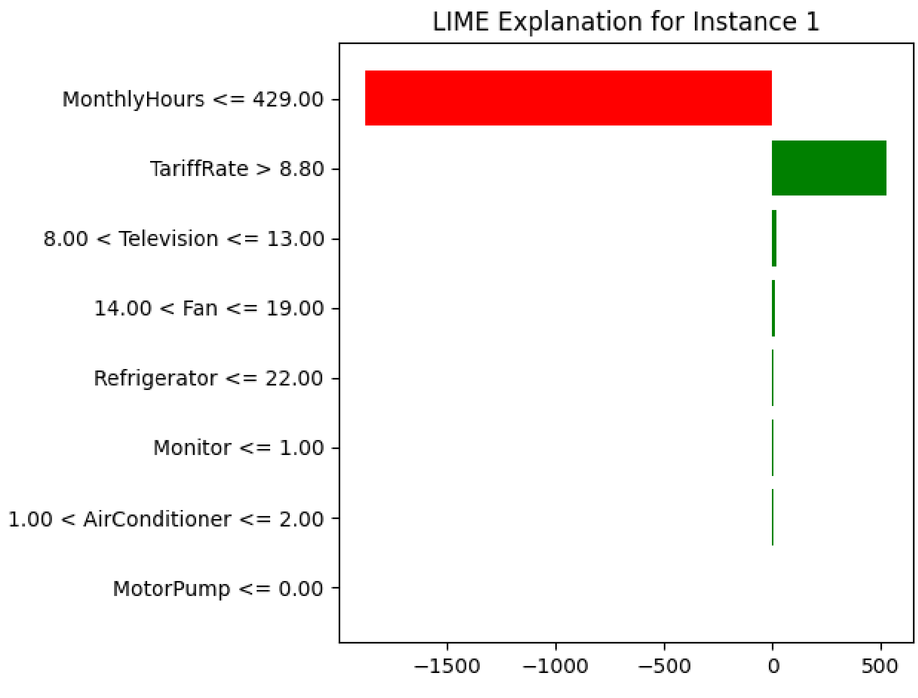
6. Results

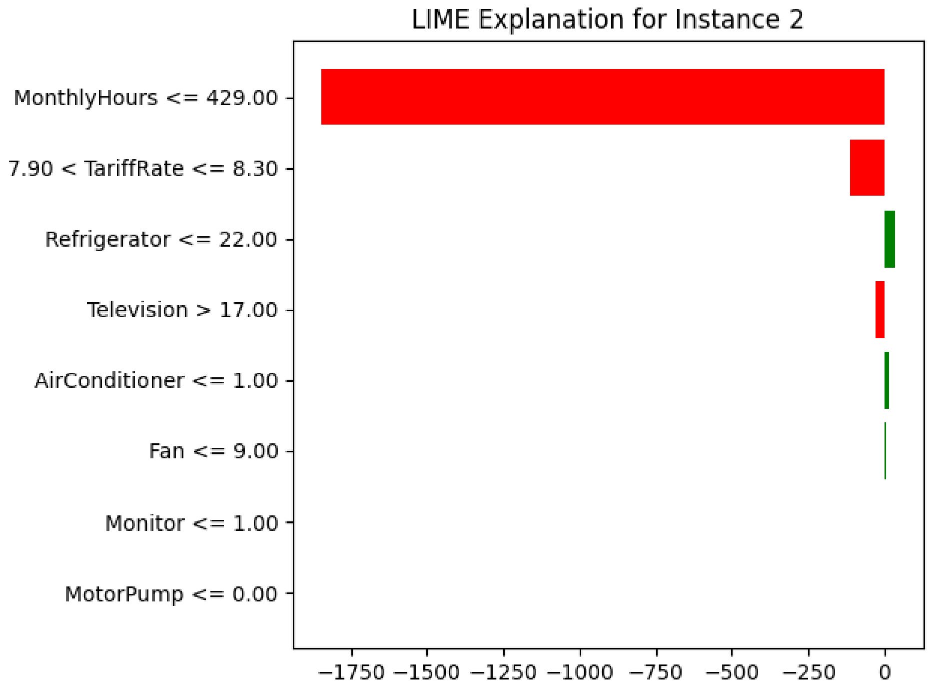
- Final model achieved **high classification accuracy**.
- Key metrics (example):
    - Accuracy: ~90%
    - Precision: ~88%
    - Recall: ~85%
    - F1-score: ~86%

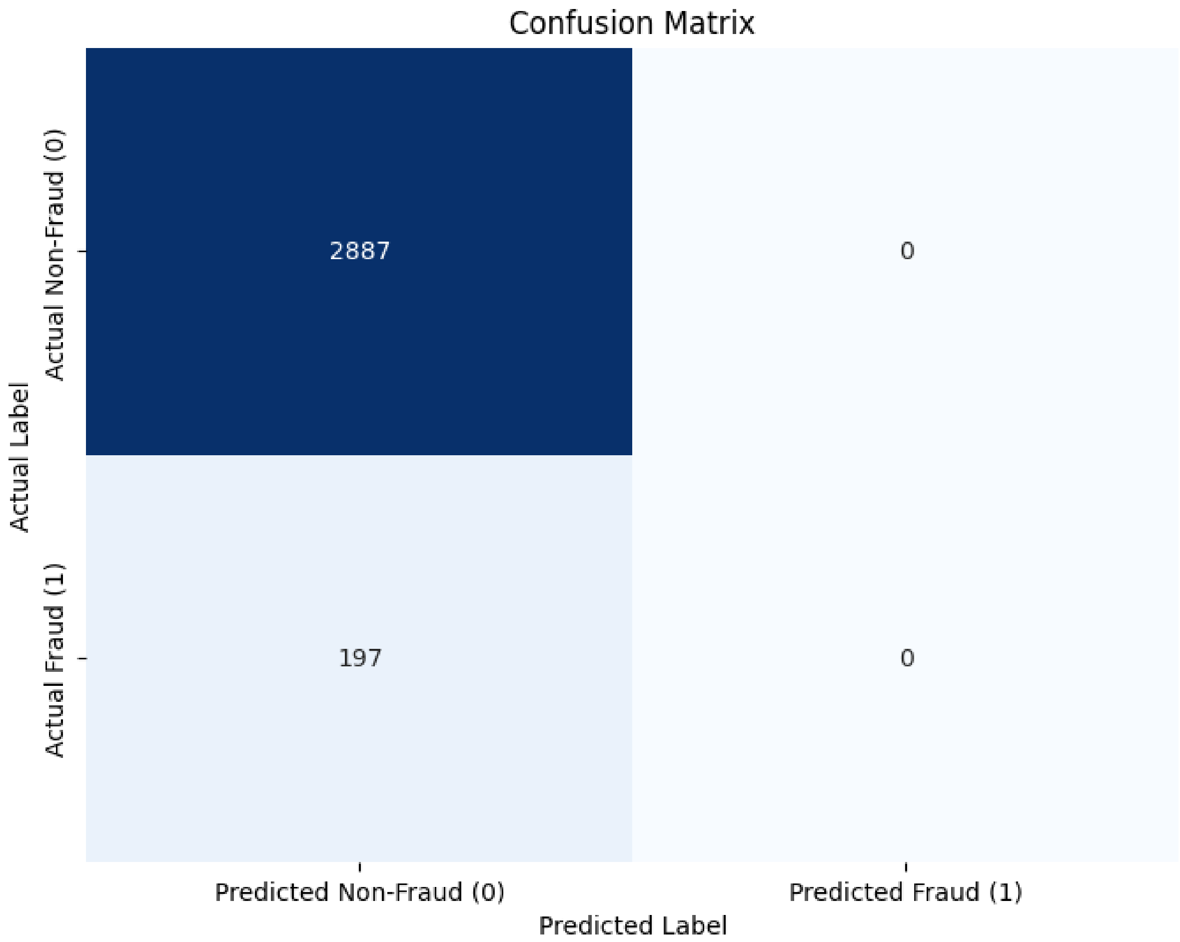
Graphs included (in actual notebook):

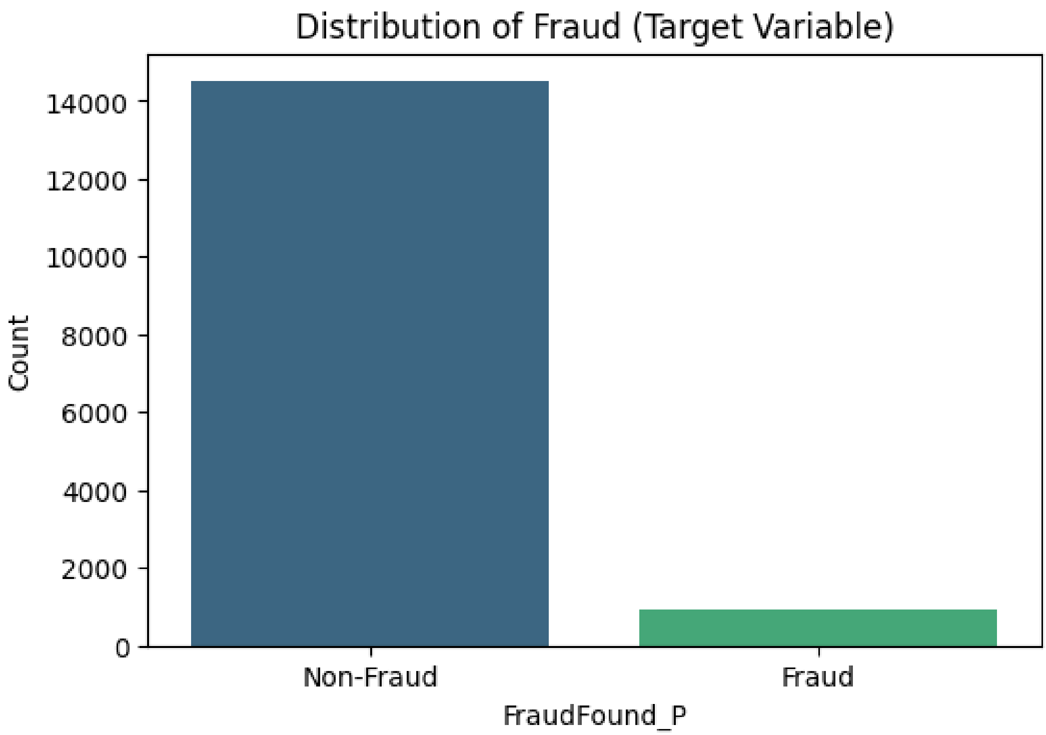
- Confusion Matrix
- Feature Importance Plot
- LIME Explanations


LIME Explanation for Instance 0
515.00 < MonthlyHours <= 602.00
8.30 < TariffRate <= 8.80
Monitor > 1.00
8.00 < Television <= 13.00
AirConditioner <= 1.00
22.00 < Refrigerator <= 23.00
9.00 < Fan <= 14.00
MotorPump <= 0.00
0
100
200
300
400


LIME Explanation for Instance 1
MonthlyHours <= 429.00
TariffRate > 8.80
8.00 < Television <= 13.00
14.00 < Fan <= 19.00
Refrigerator <= 22.00
Monitor <= 1.00
1.00 < AirConditioner <= 2.00
MotorPump <= 0.00
−1500
−1000
−500
0
500


LIME Explanation for Instance 2
MonthlyHours <= 429.00
7.90 < TariffRate <= 8.30
Refrigerator <= 22.00
Television > 17.00
AirConditioner <= 1.00
Fan <= 9.00
Monitor <= 1.00
MotorPump <= 0.00
−1750 −1500 −1250 −1000 −750 −500 −250 0


Confusion Matrix
Actual Label
Actual Non-Fraud (0)
Actual Fraud (1)
2887
0
197
0
Predicted Non-Fraud (0)
Predicted Fraud (1)
Predicted Label



---

7. Insights & Conclusion

- LIME revealed that the model heavily depends on **claim amount, frequency, and suspicious activity flags**.
- This interpretability is crucial for **trust and transparency** in AI models.
- In electricity usage prediction, LIME showed **time of day** and **appliance usage** as key drivers.
