ABSTRACT

In modern digital banking, the detection of fraudulent transactions is vital to maintaining user trust and financial security. However, conventional machine learning systems often lack transparency, making it difficult for end-users to understand why certain transactions are flagged. This project presents an **Explainable Fraud Detection System** that not only identifies anomalous transactions using the **Isolation Forest algorithm**, but also enhances interpretability through a **rule-based trust scoring mechanism** and **natural language explanations**.

The system analyzes key transaction features—such as amount, time, and location—to model user behavior and detect anomalies in real-time. A custom trust scoring function evaluates the contextual risk of each transaction based on known heuristics (e.g., unusual hours, untrusted locations, large amounts). For transactions identified as suspicious, the system leverages **Hugging Face's distilgpt2 language model** to generate concise, human-readable explanations that clarify the reasons behind the alert.

By integrating anomaly detection with explainable AI techniques, this system bridges the gap between technical fraud detection methods and user comprehension. It offers a transparent and user-friendly approach to fraud prevention, promoting both security and trust in AI-driven financial services.

ROGHAN KUMAR R

Department of Computer Science and Engineering

RANJITH SRI KUMAR MR

Department of Computer Science and Engineering

RAMESH R

Department of Computer Science and Engineering

RAMANA GIRI M

Department of Computer Science and Engineering