Project Abstract

Explainable Fraud Alert System

In the modern financial ecosystem, fraud detection is essential to ensure the security and integrity of banking operations. Traditional AI systems used by banks to flag suspicious transactions often suffer from a lack of transparency, leaving customers confused and frustrated about why certain actions are marked as fraudulent. This project addresses this issue by developing an **Explainable Fraud Alert System** that not only detects anomalies but also justifies its decisions using natural language explanations powered by **Hugging Face Transformers**.

The system incorporates four key components: behavior-based transaction modeling, anomaly detection, a trust-based scoring mechanism, and a natural language explanation interface. It begins by analyzing transaction features such as amount, time, and location to understand typical user behavior. Using the **Isolation Forest algorithm**, it identifies outliers in real-time and flags them as potentially fraudulent.

To enhance interpretability, the system employs a rule-based trust scoring mechanism that evaluates each transaction against contextual factors—such as whether it occurred at an unusual time or in an uncommon location. Transactions with lower trust scores are passed into a **language generation model**, which explains the suspicious activity in plain English. The model, based on `distilgpt2` from Hugging Face, crafts brief, understandable explanations that help users grasp the reasons behind alerts.

Ultimately, the system bridges the gap between advanced fraud detection techniques and the human need for transparency. By incorporating natural language processing into the fraud alert workflow, it empowers users with insights into their transaction history while promoting trust in AI-driven banking systems. This approach is not only user-friendly but also aligns with emerging regulatory demands for explainable and accountable AI.

Yogeshwaran M

Department of Computer Science and Engineering

Yogendra R

Department of Computer Science and Engineering

Abhishekh V R

Department of Computer Science and Engineering