# Explainable ML-Based False Positive Transaction Filtering System

Fraudulent transactions can occur not only through physical theft of bank account details but also via digital means such as phishing links, fake login portals, OTP manipulation, and exploitation of vulnerable banking servers during transactions. To mitigate financial loss and maintain customer trust, fintech companies employ rule-based and automated fraud detection systems to identify and block suspicious transactions.

However, fraud detection systems operate under the transaction confusion matrix comprising True Positives, False Positives, True Negatives, and False Negatives. Among these, False Positives, where legitimate transactions are incorrectly blocked, pose a significant challenge as they disrupt genuine users and negatively impact customer experience. Strict rule-based verification mechanisms often result in both sender and receiver accounts being blocked, leading to transaction complexity and dissatisfaction.

Traditional approaches adopted by financial institutions to reduce false positives rely on manually engineered rules and transaction parameters, achieving a reduction rate of approximately 50–60%. With the advancement of Artificial Intelligence, Machine Learning–based detection models such as Logistic Regression, Decision Trees, Random Forests, and Isolation Forests have been introduced, significantly improving fraud detection accuracy and reducing false positives to nearly 90%.

Despite these improvements, a critical limitation persists in the form of lack of explainability. Most ML-based systems function as black-box models, producing risk scores without consistently explaining the underlying reasons for flagging transactions. This lack of transparency, combined with model-specific behavior and inconsistent interpretation across systems, limits effective false-positive resolution.

To address this gap, this project proposes an ML-Based False Positive Filtering System that acts as an explainable filtering layer over existing fraud detection models. The system analyzes transaction behavior, historical patterns, and risk indicators to distinguish legitimate transactions from

Done by,
KAVIN VENTHAN J
23TD0038
Y3/S6/B