Adversarial Robustness and Anomaly Detection for Financial ML Models

Introduction

In financial systems, defending against adversarial inputs and detecting out-of-distribution (OOD) anomalies is crucial. Adversarial attacks can craft subtle perturbations to evade detection, while OOD anomalies represent rare, unexpected market behaviors (e.g., fraud, flash crashes).

Algorithm Overview

Mahalanobis Distance: Measures distance to the in-distribution mean under estimated covariance, highlighting deviations.

$$D_M(x) = \sqrt{(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

Autoencoder Reconstruction Error: Trains to minimize self-reconstruction loss; large error indicates anomaly.

$$E(x) = \|x - \hat{x}\|^2$$

Ensemble Score: Combines Mahalanobis and autoencoder signals to improve detection across linear and non-linear structures.

Thresholding

Uses Extreme Value Theory (EVT) to model score tails and determine principled thresholds rather than arbitrary percentiles, providing statistical rigor.

Robustness Tests

Includes bootstrap confidence intervals on AUC metrics and adversarial perturbation tests (FGSM) to evaluate model resilience.

Conclusion

This module provides a comprehensive and theoretically sound framework to secure financial ML models against unexpected and adversarially crafted anomalies.