CyberAntifraud Evaluation Protocol

Temporal validation and cost sensitive evaluation for acquiring fraud detection

Purpose

Define a time aware and leakage resistant evaluation protocol for acquiring fraud detection models, with explicit emphasis on false positives and false negatives.

This document is designed to be reviewer friendly. It specifies split logic, windowing strategy, metrics, and release gates.

Data and splits

Use blocked time series validation. No random shuffles. Training uses past windows only. Validation and test windows are strictly future relative to training.

Primary split: rolling window backtests. Secondary split: out of time holdout that simulates a go live moment.

Rolling windows

Define a sequence of evaluation windows W1...Wk. For each window, train on a fixed or expanding history and test on the next period. Track variance across windows as a stability signal.

Report both mean and dispersion across windows, not only a single point estimate.

Leakage control

Feature computation must be time consistent. Any aggregation must be computed only from information available up to the transaction timestamp.

Prevent label leakage via post event fields. Enforce field catalog with allowed and prohibited variables.

Metrics

Primary: AUPRC, recall at fixed precision, precision at fixed recall, and cost based expected net benefit under operational constraints.

Secondary: calibration drift, alert volume, analyst load proxies, and stability metrics across windows.

Release gates

A model cannot be promoted if it improves AUPRC but increases false positive volume beyond capacity or degrades stability under drift.

Each promotion requires model card completion, audit log traceability, and a documented threshold policy.