Explainable Machine Learning for Forensic Accounting Analytics: Fraud Detection across Narrative Disclosures and Numeric Audit Indicators

Dhananjay Sudhakarrao Naphade
Cognizant Technologies Inc USA, Austin, Texas

Abstract:

Financial statement fraud is rare, costly, and buried in evidence of two very different kinds: the free-text narrative of regulatory filings and the compact numeric indicators produced by audits. We present a unified, leakage-aware, and explainable machine-learning study that measures how much fraud signal can be extracted from each evidence type and translates model outputs into auditor-readable red flags. Across three public datasets—US SEC 10-K filings, Indian firm audit indicators, and European card transactions—we evaluate sparse lexical, dense finance-transformer, and forensic-linguistic feature sets, three class-imbalance strategies, and both supervised and label-free models, under a capacity-constrained protocol that scores precision and recall inside a fixed audit budget rather than reporting accuracy. On filing text, lexical features carry strong signal (ROC-AUC 0.954) and, contrary to a common assumption, outperform dense FinBERT embeddings (0.856; DeLong p=0.026), while purely stylistic features alone reach 0.889. Under 0.17% prevalence, majority-preserving strategies keep ranking quality high (PR-AUC up to 0.806, top-1% recall 0.852) whereas aggressive undersampling collapses precision (PR-AUC 0.226)—a failure invisible to ROC-AUC. On numeric audit evidence we expose and remove a label-leakage trap, then use TreeSHAP to produce one-page red-flag cards that attribute each flagged firm to the SAS 99/ISA 240 fraud-triangle categories; the explanations pass permutation and deletion sanity checks (rank correlation 0.976). The red-flag card, not the raw score, is the artifact an audit team consumes, and is the study's practical contribution.

Keywords:

Forensic accounting, Financial statement Fraud, explainable AI, SHAP, Fraud triangle, Class imbalance, FinBERT, Audit analytics.