World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr4125

EXPLAINABLE IDENTIFICATION OF TAX-RELATED FINANCIAL REPORTING ANOMALIES IN U.S. PUBLIC COMPANIES: CROSS-FITTED TAX-CONSISTENCY RESIDUALS AND OWEN-VALUE SHAPLEY PROPAGATION

JiaYing Chen[1*], XuGuang Zhang[2], Xu Han[3]
[1]*Cornell University, Ithaca, NY, USA.*
[2]*Independent Researcher.*
[3]*Renmin University of China, Beijing 100086, China.*
[*]*Corresponding Author: JiaYing Chen*

Abstract: Income-tax accounts are reported under two regimes, financial accounting and the tax code, so their consistency with the rest of a firm's financial statements is a natural check on reporting quality. Existing machine-learning detectors of financial misreporting treat tax items as generic inputs, and their explanations rarely reach the line items that an examiner can verify. We propose TRACE, an explainable framework for identifying tax-related financial reporting anomalies in U.S. public companies. First, for every fiscal year TRACE estimates label-free peer expectations of income-tax expense, taxes payable and their changes with cross-fitted conditional median and dispersion models, so that the resulting tax-consistency residuals cannot absorb the anomaly they measure. Second, it embeds these residuals in an imbalance-aware gradient-boosting ensemble that combines dollar-scale and within-year peer-rank representations of 28 Compustat items and is selected on validation years only. Third, it explains each risk score with Propagated Shapley Explanations, which chain exact single-baseline TreeSHAP, an exact Owen-value allocation over the reported, expected and dispersion components of each residual, and guarded propagation to 33 line items, with a proven completeness guarantee. On 146,045 U.S. firm-years labelled with SEC Accounting and Auditing Enforcement Releases and evaluated strictly out of time (2003–2008), TRACE achieves the highest NDCG at the 1% screening cut-off (0.0330), an improvement of 99% over RUSBoost and 89% over LightGBM (bootstrap $p = 0.019$ and 0.010) at comparable AUC. Its label-free tax-consistency score alone outperforms four generic outlier detectors. The explanations are exactly complete, agree closely with KernelSHAP (Spearman correlation 0.98) at 2.7 times lower cost, and show that detected misstatements most often combine tax expense growing faster than peers' with taxes payable below peer expectations.
Keywords: Explainable artificial intelligence; Financial misstatement detection; Income tax accounting; Conditional anomaly detection; Shapley values; Owen value

1 INTRODUCTION

Material misstatements in the financial reports of public companies impose large losses on investors and erode confidence in capital markets. In the United States, the Securities and Exchange Commission (SEC) documents enforcement actions for accounting and auditing misconduct in its Accounting and Auditing Enforcement Releases (AAERs), which have become the reference labels for research on misstatement detection [1-3]. Statistical detectors have progressed from parsimonious logistic models built on accrual and performance ratios to ensemble learners trained directly on raw financial-statement data and to gradient-boosted models that exploit interactions among accounting, market and audit variables [1-3]. Because the output of such a detector triggers costly investigations, a risk score is useful to auditors, regulators and tax examiners only if it comes with a faithful account of *which reported numbers* drive it [4, 5].

Income-tax accounts occupy a special position in this problem. Income-tax expense and taxes payable are computed from the same economic events as book income but are filed under the tax code with a different counterparty, the tax authority. The relation between book and tax numbers therefore carries information that is hard to fabricate consistently. Large book-tax differences predict lower persistence of earnings and accruals [6, 7]; tax accounts themselves are used to manage reported earnings [8, 9]; aggressive financial and tax reporting are systematically related [10, 11]; and firms accused of fraud paid, on average, about eleven cents of additional income tax per dollar of overstated pre-tax earnings [12]. A misstatement therefore leaves traces in the *joint* structure of the tax block and the rest of the statements: a firm reports tax expense, payables or their changes that are inconsistent with what comparable firms report given similar fundamentals. We refer to such deviations as *tax-related financial reporting anomalies*.

Turning this accounting insight into a reliable and explainable detector raises three bottlenecks.

B1 – Representation. Tax evidence is relational and peer-relative. Raw tax ratios are not enough: adding twelve standard tax ratios (effective tax rate, tax expense and payables scaled by assets, their changes, a statutory-rate book-tax gap and related measures) to a strong LightGBM detector leaves its out-of-time AUC unchanged (Section 6). With fewer than one thousand positive firm-years, a learner cannot discover the peer-conditional expectation of each tax account by itself.

© By the Author(s) 2023, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

B2 – Estimation. Peer expectations estimated with flexible learners on the same firms absorb the very deviations they are meant to expose (self-influence masking); expectations pooled across years mix tax regimes; and dollar amounts drift with inflation and firm growth, so that thresholds learned on the 1990s transfer poorly to later years.
B3 – Explanation. Explanations of detectors with engineered inputs stop at the feature level (e.g., "the taxes-payable residual equals 3.1"), which an examiner cannot verify against a filing. Model-agnostic explainers on raw items are slow and sampling-noisy [13, 14], and propagating Shapley values through the pipeline with the standard rescale rule breaks down for tree ensembles [15, 16], whose outputs are step functions: in our pipeline, single line items received rescale attributions of ± 34 log-odds units for a net score gap of 4.1.
We address these bottlenecks with **TRACE** (**T**ax-**R**esidual **A**ugmented **C**onsistency **E**nsemble) and make four contributions.
1. **Cross-fitted tax-consistency residuals (TCR).** For every fiscal year we learn the conditional median and the conditional dispersion of four tax targets (tax expense, taxes payable, and their changes, all scaled by assets) given non-tax fundamentals and the remaining tax accounts, using K-fold cross-fitting within the year's cross-section. The resulting standardized residuals are label-free, use only information available at the filing date, and are never estimated from the observation they evaluate (B1, B2).
2. **A validation-selected multi-view detector.** TRACE combines a dollar-scale view and a stationary within-year peer-rank view of the 28 raw items, each augmented with TCR, in an ensemble of class-balanced gradient-boosting members. All hyper-parameters and the ensemble design are selected on validation years that precede the test period (B2).
3. **Propagated Shapley Explanations (PSE).** We explain the full pipeline at the level of 33 primitive line items by composing (a) a closed-form, exact single-baseline TreeSHAP, (b) an exact *Owen-value* allocation of each residual's attribution to its reported, expected and dispersion components [17], and (c) a conditioning-guarded propagation through ratios and expectation models. PSE is complete by construction, needs no re-evaluation of the pipeline, and yields exact unit-level attributions for the tax channel (B3).
4. **Evidence.** On 146,045 U.S. firm-years with AAER labels, TRACE delivers the best screening performance at audit-capacity cut-offs with statistically significant gains over RUSBoost [2], balanced random forests, XGBoost and LightGBM; the label-free TCR score outperforms generic outlier detectors; and PSE explanations reveal which tax relationships characterize detected misstatements. Ablations, sensitivity analyses and robustness checks (alternative serial-fraud protocols, rolling windows, an extended test period and placebo residuals) support the main design choices.
The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data and the evaluation protocol. Section 4 presents TRACE and PSE. Section 5 details the experimental setup, Section 6 reports the results, Section 7 discusses implications, and Section 8 concludes.

2 RELATED WORK

2.1 Machine Learning for Misstatement Detection

Dechow et al. assembled the AAER sample and proposed the F-score, a logistic model of accrual quality, performance, non-financial and off-balance-sheet measures [1]. Bao et al. showed that RUSBoost trained on 28 raw Compustat items outperforms ratio-based models in a rolling out-of-time design [2, 18], and released the data used in this paper; the treatment of serial frauds in that design was later debated [19, 20], and we therefore report both the original relabelling protocol and a stricter protocol that drops serial-fraud firm-years. Perols et al. showed that data-analytics preprocessing addressing the rarity of fraud and the abundance of candidate predictors improves fraud prediction [21]. Bertomeu et al. documented that accounting variables become informative mainly through interactions captured by gradient-boosted trees [3], and Bertomeu discussed validation design and collinearity of accounting inputs [5]. Tree ensembles remain the strongest learners on tabular data of this size [22-24]. These studies treat tax items as ordinary inputs; none constructs peer-conditional tax expectations or explains detections at the line-item level.

2.2 Tax Accounts and Reporting Quality

Accounting research documents that the tax block is informative about earnings quality. Hanlon and Blaylock et al. show that large book-tax differences signal less persistent earnings, distinguishing upward earnings management from tax avoidance [6-7]. Graham et al. review evidence that tax accounts are used to manage earnings [8], and Hanlon and Heitzman survey tax research more broadly [9]. Frank et al. find a positive relation between aggressive tax and financial reporting [10], whereas Lennox et al. find that tax-aggressive firms are less likely to commit accounting fraud [11]; Erickson et al. show that many fraud firms paid taxes on fictitious earnings [12]. The direction of the association therefore varies across firms, which motivates a flexible, peer-conditional and explainable treatment rather than a fixed-sign tax ratio.

2.3 Conditional and Attribute-wise Anomaly Detection

Generic outlier detectors such as Isolation Forest [25], HBOS [26], COPOD [27] and ECOD [28], available in PyOD [29], score global unusualness. Attribute-wise learning reformulates outlier detection as a set of regressions that predict each attribute from the others and score the prediction errors [30]. TCR extends this idea in three directions that matter for financial statements: it targets a domain-defined block (income taxes), it standardizes residuals with a learned conditional dispersion, and it uses cross-fitting, a device from double machine learning [31], so that flexible expectation models cannot fit the anomaly of the observation being scored.

2.4 Explaining Pipelines of Models

SHAP and TreeSHAP attribute predictions with Shapley values [14, 32]; LIME fits local surrogates [13]. Sundararajan and Najmi formalize baseline Shapley values and their random-baseline average, which is the explanation target adopted here [33]. For pipelines, DeepLIFT and G-DeepSHAP propagate attributions with a rescale rule that enforces efficiency at every layer [15-16]. We show that the rescale rule is unreliable when the downstream model is a tree ensemble and replace it, at the node that matters most for the tax channel, by an exact Owen value [17], the coalitional generalization of the Shapley value. Explanation faithfulness is assessed with deletion and insertion curves [34], and the case for faithful explanations in high-stakes settings is made by Rudin [4].

3 DATA AND PROBLEM FORMULATION

3.1 Data

We use the public dataset of Bao et al. [2], which merges AAER-based misstatement labels with Compustat fundamentals for U.S. public firms. It contains 146,045 firm-years from fiscal years 1990–2014, of which 964 are misstatement firm-years belonging to 412 AAER cases. Each firm-year provides 28 raw accounting items, including the two income-tax accounts txt (income taxes, total) and txp (income taxes payable), and the label misstate. From the same panel we attach five prior-year items (at, txt, txp, ib, sale), available for 83.1% of firm-years. The resulting 33 *primitive* line items $p \in \mathbb{R}^{33}$ are the units in which explanations are reported. Table 1 summarizes the evaluation windows.

3.2 Evaluation Protocol

We follow the rolling out-of-time design of Bao et al. [2]. For a test year $t \in \{2003, ..., 2008\}$ the model is trained on fiscal years 1991 to $t-2$; the two-year gap reflects the delay between a misstatement and its public enforcement. Training firm-years that belong to an AAER case also present in the test year are relabelled as non-fraudulent (serial-fraud correction); as a robustness check we drop them instead. Hyper-parameters and the ensemble design are selected on validation years 1999–2001 (training on 1991 to $v-2$), so the test years are never used for model selection.

Table 1 Rolling Out-of-Time Evaluation Windows

Test year	Training years	Training firm-years	Training positives	Test firm-years	Test positives
2003	1991–2001	67,166	402	5,981	69
2004	1991–2002	73,233	467	5,934	58
2005	1991–2003	79,214	538	5,863	45
2006	1991–2004	85,148	660	5,908	33
2007	1991–2005	91,011	720	5,868	30
2008	1991–2006	96,919	758	5,612	26

Note: Training labels are counted after the serial-fraud correction.

3.3 Problem Statement

Let $F: \mathbb{R}^{33} \to \mathbb{R}$ be a risk score computed from the primitives of a firm-year. Given a test year with n_t firm-years and an investigation budget of $k = \lceil 0.01\, n_t \rceil$ firms, the goal is (i) to rank firm-years so that misstatements concentrate at the top of the list, and (ii) to attach to every flagged firm-year an attribution vector $\phi(p) \in \mathbb{R}^{33}$ over line items that sums exactly to the gap between its score and a peer reference, and that isolates the contribution of tax-related anomalies.

4 METHOD

Figure 1 summarizes TRACE. Primitive line items feed two representations of the firm-year and a set of cross-fitted peer-expectation models for the tax block; the resulting tax-consistency residuals augment both representations; an ensemble of gradient-boosting members produces the risk score; and PSE propagates the score back to the primitives.

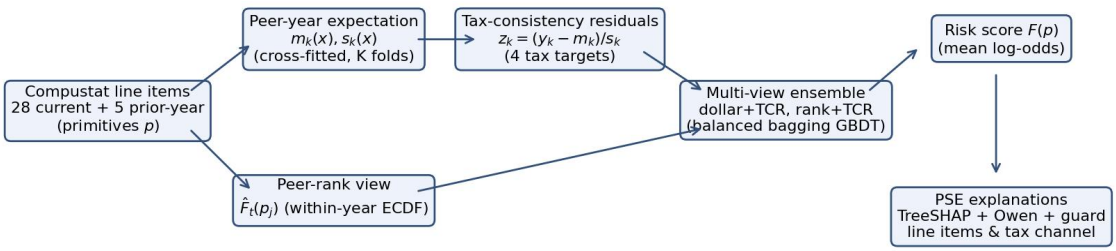


Figure 1 Overview of TRACE

Note: Peer-year expectation models are fitted per fiscal year with K-fold cross-fitting and never use labels; PSE explains the ensemble score in terms of the 33 primitive line items and of the reported, expected and dispersion components of each tax-consistency residual.

4.1 Cross-Fitted Tax-Consistency Residuals

Targets. With $\bar{a} = (\text{at} + \text{at}_{-1})/2$ we model four tax targets:

$$y_1 = \frac{\text{txt}}{\text{at}}, \quad y_2 = \frac{\text{txp}}{\text{at}}, \quad y_3 = \frac{\text{txp}-\text{txp}_{-1}}{\bar{a}}, \quad y_4 = \frac{\text{txt}-\text{txt}_{-1}}{\bar{a}}. \tag{1}$$

They capture the level and the change of the income-statement tax charge and of the balance-sheet tax liability, the two accounts through which book and taxable income are reconciled.

Conditioning information. Each target is conditioned on a vector x_k of fourteen non-tax fundamentals: pre-tax income over assets, $(\text{ib} + \text{txt})/\text{at}$, a loss indicator, log assets, leverage, asset turnover, gross property, plant and equipment, cash, interest expense, depreciation, retained earnings, receivables and inventories (all scaled by assets), sales growth, and log market value. Change targets additionally receive the change in pre-tax income. To capture the internal consistency of the tax block, x_k also contains the other contemporaneous tax account (payables for the expense targets and vice versa) and the prior-year level of the target account. All derived quantities are fixed, deterministic functions of the primitives, with undefined values (e.g., missing prior-year items) mapped to a sentinel.

Cross-fitting within the fiscal year. For fiscal year t and target k, the valid observations are split into K folds. For each fold f a gradient-boosted regressor with absolute loss is fitted on the other folds and used to predict the conditional median $\hat{m}_k^{(t,-f)}(x)$ for the observations of fold f. A second regressor with absolute loss is fitted on the absolute out-of-fold residuals of the other folds, giving the conditional median absolute deviation, and the conditional scale is $\hat{s}_k = 1.4826 \cdot \widehat{\text{MAD}}_k$. The tax-consistency residual of firm-year i in fold f is

$$z_{ik} = \text{clip}\left(\frac{y_{ik}-\hat{m}_k^{(t,-f)}(x_{ik})}{\max\{\hat{s}_k^{(t,-f)}(x_{ik}),\, \varsigma_{kt}\}}, -10, 10\right), \tag{2}$$

where the floor ς_{kt} is a fraction of the unconditional dispersion of the target in year t. The aggregate tax-anomaly score is $\bar{z}_i = \left(\frac{1}{K_i}\sum_k z_{ik}^2\right)^{1/2}$ over the K_i targets defined for the firm-year.

Proposition 1 (No self-influence). *For every firm-year i in fold f, the expectation $\hat{m}_k^{(t,-f)}$ is a function of the observations outside fold f only. Consequently, the residual $y_{ik} - \hat{m}_k^{(t,-f)}(x_{ik})$ responds one-for-one to a change in the reported value y_{ik}, and an anomalous report cannot be absorbed by its own expectation.*

Proof. By construction $\hat{m}_k^{(t,-f)}$ is fitted on folds $g \neq f$, so $\partial \hat{m}_k^{(t,-f)}(x_{ik})/\partial y_{ik} = 0$ and the residual moves by exactly Δy_{ik}. For an in-sample fit, by contrast, the fitted value moves with y_{ik} whenever the learner can isolate the observation, shrinking the residual toward zero. □

Three properties make TCR suitable for out-of-time screening. It uses no labels, so it can be computed on the test year's own cross-section, which reflects that year's tax regime, without look-ahead. It is computed identically for training and test years. And it measures deviations relative to economically comparable peers rather than globally, so that a high payable is anomalous only if comparable firms do not report one.

4.2 Multi-View Detector

Views. The *dollar view* contains the 28 raw items; it preserves scale, which matters because enforcement concentrates on larger and more visible firms. The *peer-rank view* replaces each raw item by its within-year empirical distribution function, $\hat{F}_t(p_j) = (N_< + N_\leq + 1)/(2n_t)$, which is invariant to inflation and to the growth of the Compustat universe and therefore transfers better across years. Each view is augmented with the four residuals $z_{i1}, \ldots, z_{i4}$.

Members and ensemble. Each member h_m is a LightGBM classifier trained with class-conditional bagging [23]: every boosting iteration uses all positive firm-years and a random fraction ρ of the negatives, the gradient-boosting analogue of random undersampling in RUSBoost [18]. The TRACE score is the mean log-odds of $M = 4$ members, two per view with different seeds,

$$F(p) = \frac{1}{M}\sum_{m=1}^{M} h_m\left(\Phi_{v(m)}(p)\right), \tag{3}$$

where Φ_v maps the primitives to the inputs of view v. Averaging log-odds keeps the ensemble additive over trees, which PSE exploits.

Validation-based selection. On the validation years 1999–2001 a grid of 18 configurations (leaves $\in \{7,15,31\}$, rounds $\in \{200,400\}$, $\rho \in \{0.05,0.1,0.2\}$) selected 15 leaves, 200 rounds and $\rho = 0.05$ (learning rate 0.05, feature fraction 0.7, $\ell_2 = 1$). With this configuration we compared four designs with an equal number of members: dollar view only (validation AUC 0.6829), both views without TCR (0.6942), both views with TCR in every member (0.7033), and both views with TCR in half of the members (0.6999). The design with TCR in every member is TRACE; all other designs are reported as ablations.

4.3 Propagated Shapley Explanations

Explanation target. For a flagged firm-year with primitives p in year t and a reference set $\mathcal{B}_t$ of peer firm-years drawn from the same year, we target the random-baseline Shapley value [33]

$$\phi_c(p) = \frac{1}{|\mathcal{B}_t|}\sum_{b\in\mathcal{B}_t}\phi_c^{\mathrm{BS}}(p;b), \qquad \sum_{c=1}^{33}\phi_c^{\mathrm{BS}}(p;b) = F(p) - F(b), \tag{4}$$

where $\phi^{\mathrm{BS}}(p;b)$ is the Shapley value of the game $S \mapsto F(p_S, b_{\bar{S}})$. The function F that is explained is the exact function that produced the firm-year's score, including the fold-specific expectation models of its residuals. PSE computes (4) pair by pair in three steps.

Step 1: exact single-baseline TreeSHAP. For a member tree, consider a leaf ℓ with value v_ℓ and its root-to-leaf path. Resolving repeated splits on the same feature, every feature on the path falls into one of four classes: its splits are satisfied by both $x = \Phi(p)$ and $\beta = \Phi(b)$, by x only (set A_ℓ), by β only (set B_ℓ), or by neither, in which case the leaf is unreachable by any hybrid input. The leaf contributes v_ℓ to $S \mapsto h(x_S, \beta_{\bar{S}})$ exactly when $A_\ell \subseteq S$ and $B_\ell \cap S = \emptyset$, a game whose Shapley value is available in closed form. With $a = |A_\ell|$ and $b_\ell = |B_\ell|$,

$$\phi_j = \sum_\ell v_\ell \left[\mathbb{1}(j \in A_\ell)\frac{(a-1)!\,b_\ell!}{(a+b_\ell)!} - \mathbb{1}(j \in B_\ell)\frac{a!\,(b_\ell-1)!}{(a+b_\ell)!}\right]. \tag{5}$$

A simultaneous traversal of x and β enumerates the reachable leaves in $O(LD)$ time per tree (L leaves, depth D). Equation (5) coincides with interventional TreeSHAP for a single background point [32]; our numba implementation agrees with the reference implementation up to $3.9e-07$ and satisfies efficiency to machine precision.

Step 2: exact Owen allocation at the residual node. The residual $z_k = \zeta(y_k, m_k, s_k)$ of (2) aggregates three economically distinct components: the *reported* tax quantity y_k, its *peer expectation* m_k, and the peer *dispersion* s_k. We treat $U_k = \{y_k, m_k, s_k\}$ as an a priori union and allocate the Shapley value of z_k obtained in Step 1 with the Owen value [17]. For $T \subseteq U_k$ let $\zeta(T)$ be the hybrid residual that takes the components in T from the firm-year and the others from the baseline, and let $g_k(T)$ be the Step-1 attribution of feature z_k when z_k is set to $\zeta(T)$ while all other inputs keep their firm-year values. The attribution of component $u \in U_k$ is

$$\psi_u = \sum_{T\subseteq U_k\setminus\{u\}} \frac{|T|!\,(2-|T|)!}{3!}\left[g_k(T\cup\{u\}) - g_k(T)\right]. \tag{6}$$

Proposition 2 (Owen consistency). *Consider the game over the players $\mathcal{D} \cup \bigcup_k U_k$, where $\mathcal{D}$ are the direct inputs of the members, with coalition structure $\{\{j\}: j \in \mathcal{D}\} \cup \{U_k\}_k$. For every pair (p, b), Step 1 applied to the quotient game together with (6) returns the Owen value of this game. In particular $\sum_{u\in U_k}\psi_u$ equals the Step-1 attribution of z_k, and $\psi_u = 0$ whenever component u takes the same value for the firm-year and the baseline.*

Proof. The Owen value can be computed in two stages [17]: the unions first share the worth of the grand coalition according to the Shapley value of the quotient game, whose players are the unions; then, within union U_k, the members share the union's payoff according to the Shapley value of the game that assigns to $T \subseteq U_k$ the Shapley payoff of union k in the quotient game where U_k is represented by T. Because the members depend on (y_k, m_k, s_k) only through z_k, the quotient game is the game over member inputs solved in Step 1, and the payoff of union k when represented by T is $g_k(T)$. Equation (6) is the Shapley value of $T \mapsto g_k(T)$, with $g_k(\emptyset) = 0$ because $\zeta(\emptyset)$ equals the baseline residual. Efficiency within the union and the null-player property follow from those of the Shapley value.

Unlike the rescale rule, which multiplies the attribution of z_k by $\Delta y_k/\Delta z_k$ and diverges when the three components nearly cancel, (6) evaluates the tree model at the actual hybrid residuals and therefore respects its thresholds. Steps 1 and 2 already deliver exact attributions for 28 direct line-item units and 12 tax-channel units, the reported, expected and dispersion components of each target, which is the format in which tax anomalies are communicated to an examiner.

Step 3: guarded propagation to primitives. Reported components and conditioning features are deterministic functions of at most six primitives, and the expectation and dispersion components are tree ensembles over the conditioning features. For an intermediate quantity q with exact local Shapley components $\{\omega_{qc}\}$ over its inputs (exact enumeration for ratios, Step 1 for the expectation and dispersion models), an attribution Ψ_q is passed down as $\Psi_q\, w_{qc}$ with

$$w_{qc} = \begin{cases} \omega_{qc}/\Delta q, & \text{if } \sum_c |\omega_{qc}| \le \kappa\,|\Delta q|, \\ \dfrac{[\omega_{qc}\,\mathrm{sgn}(\Delta q)]_+}{\sum_{c'}[\omega_{qc'}\,\mathrm{sgn}(\Delta q)]_+}, & \text{otherwise,} \end{cases} \tag{7}$$

where $\Delta q = q(p) - q(b) = \sum_c \omega_{qc}$ and $\kappa = 3$. The first branch is the generalized rescale rule [16]; the second assigns the attribution to the inputs that moved q in the direction of its net change whenever the linearization is ill-conditioned. Attributions reaching the same primitive along different paths are summed.

Proposition 3 (Completeness). *For every firm-year p and reference set $\mathcal{B}_t$, the PSE attributions satisfy $\sum_{c=1}^{33}\phi_c(p) = F(p) - \frac{1}{|\mathcal{B}_t|}\sum_{b\in\mathcal{B}_t}F(b)$, where $F(b)$ is evaluated with the expectation models of p.*

Proof. For a fixed pair, Step 1 is efficient for every member, so the member attributions sum to $F(p) - F(b)$. Step 2 preserves the attribution of each residual by Proposition 2. In Step 3 both branches of (7) produce weights that sum to one, so every intermediate attribution is redistributed without loss. Averaging over $\mathcal{B}_t$ preserves the identity. □

The cost of PSE is linear in the number of baselines and requires no evaluation of the composed pipeline; Algorithm 1 summarizes the procedure.

Algorithm 1 (PSE for one firm-year).

1. primitives p; fold-specific models $\{\hat{m}_k, \hat{s}_k\}$ of p; members $\{h_m\}$; peer baselines $\mathcal{B}_t$
2. each baseline $b \in \mathcal{B}_t$
3. compute member inputs $\Phi(p)$ and $\Phi(b)$, using the expectation models of p for both
4. Step 1: exact TreeSHAP (5) of every member for $\big(\Phi(p), \Phi(b)\big)$; credit direct inputs to their line items
5. each tax target k
6. evaluate $g_k(T)$ for the eight $T \subseteq \{y_k, m_k, s_k\}$; Owen allocation (6) → $\psi_{y_k}, \psi_{m_k}, \psi_{s_k}$
7. propagate ψ_{y_k} through the ratio defining y_k with (7)
8. propagate ψ_{m_k}, ψ_{s_k} through TreeSHAP of $\hat{m}_k, \hat{s}_k$ and the conditioning ratios with (7)

5 EXPERIMENTAL SETUP

Metrics. Following [2], we report the area under the ROC curve (AUC) and, at the screening cut-off $k = \lceil 0.01\, n_t \rceil$, the normalized discounted cumulative gain (NDCG@1%), sensitivity and precision; we add average precision (AP), which summarizes the whole precision-recall curve under the 0.6% base rate. Metrics are computed per test year and averaged over 2003–2008.

Baselines. We compare TRACE with logistic regression on the 14 ratios of the Dechow et al. framework provided with the data (Logit-14) and on the 28 raw items (Logit-28); RUSBoost on the 28 raw items with the configuration of Bao et al. (300 trees, learning rate 0.1, minimum leaf size 5) [2]; a balanced random forest (300 trees); XGBoost with class weighting [24]; LightGBM on the 28 raw items, on the 28 raw items plus the 14 ratios, and on the 28 raw items plus 12 tax ratios; and the ensemble designs that remove TCR or one of the two views.

Replication and inference. Every stochastic model is trained in three independent replicates; each ensemble replicate contains four members. Tables report the mean and, in parentheses, the standard deviation across replicates. Statistical significance is assessed with a stratified bootstrap (1,000 resamples within each test year, preserving the number of positives) on replicate-averaged scores; we report the mean difference, its 95% interval and the one-sided p-value $\Pr(\Delta \leq 0)$.

Explanation evaluation. We explain all firm-years flagged by TRACE (top 1% of each test year) against peer baselines drawn at random from the same year. On a comparison subset that lists the detected misstatements first and is completed with randomly chosen flagged firm-years with complete prior-year records, we compare PSE with KernelSHAP on the same pipeline and baselines (2,048 coalitions) [14], LIME (3,000 samples) [13], a naive mapping of feature-level TreeSHAP that credits each residual to its reported tax item, and random orderings. We report Spearman correlation and top-5 overlap with KernelSHAP (and of KernelSHAP with an independent KernelSHAP run as a reference), the completeness error, deletion and insertion areas over the ten highest-ranked line items normalized by the score gap [34], and the runtime per explanation.

Implementation. All experiments run on a single CPU core with Python 3.12, LightGBM 4.7, XGBoost, imbalanced-learn, PyOD [29], SHAP and numba. Expectation models use 150 trees with 15 leaves, a learning rate of 0.05 and at least 40 observations per leaf, with $K = 5$ folds. The complete code, intermediate results and scripts that generate every table and figure are provided as supplementary material.

6 RESULTS

6.1 Detection Performance

Table 2 reports out-of-time performance averaged over the six test years. TRACE attains the highest NDCG@1% (0.0330), sensitivity (0.0342), precision (0.0292) and average precision (0.0205) of all models. Relative to RUSBoost on the 28 raw items, the benchmark of Bao et al. [2], NDCG@1% increases by 99%, precision at the 1% cut-off by 55%, sensitivity by 66% and average precision by 18%. The AUC of TRACE (0.7157) is on par with the strongest tree ensembles (RUSBoost 0.7132, LightGBM 0.7108), so the gain is concentrated where screening decisions are made: at the top of the ranking. Two further observations support the motivation of Section 1. Adding twelve conventional tax ratios to LightGBM does not improve it, which shows that the information in the tax block is not accessible through raw ratios (B1). Adding the 14 accounting ratios to the raw items lowers the AUC of LightGBM, consistent with the finding of Bao et al. that raw items are a better input than standard ratios [2].

Table 2 Out-of-Time Detection Performance, Test Years 2003–2008 (Mean over Years; Standard Deviation across Three Replicates in Parentheses)

Model	AUC	AP	NDCG@1%	Sens@1%	Prec@1%
Logit-14 (Dechow et al. ratios)	0.6816	0.0168	0.0261	0.0323	0.0169
Logit-28	0.6774	0.0157	0.0130	0.0167	0.0114
RUSBoost-28 (Bao et al.)	0.7132 (0.0052)	0.0174 (0.0004)	0.0166 (0.0011)	0.0207 (0.0018)	0.0188 (0.0016)
Balanced RF-28	0.7196 (0.0023)	0.0167 (0.0004)	0.0121 (0.0009)	0.0152 (0.0010)	0.0158 (0.0016)

Model	AUC	AP	NDCG@1%	Sens@1%	Prec@1%
XGBoost-28	0.7058 (0.0040)	0.0176 (0.0019)	0.0193 (0.0027)	0.0197 (0.0028)	0.0170 (0.0028)
LightGBM-28	0.7108 (0.0059)	0.0173 (0.0004)	0.0175 (0.0028)	0.0195 (0.0014)	0.0169 (0.0027)
LightGBM-42 (raw+ratios)	0.6838 (0.0081)	0.0170 (0.0007)	0.0242 (0.0082)	0.0295 (0.0102)	0.0245 (0.0090)
LightGBM-28 + tax ratios	0.7065 (0.0036)	0.0168 (0.0003)	0.0164 (0.0016)	0.0198 (0.0035)	0.0170 (0.0028)
TRACE (ours)	0.7157 (0.0037)	0.0205 (0.0003)	0.0330 (0.0015)	0.0342 (0.0022)	0.0292 (0.0016)

Note: NDCG, sensitivity and precision are computed at the top 1% of each test year.

Figure 2 shows the results by test year. TRACE attains the highest NDCG@1% among the five models shown, in 3 of the six test years, and its AUC lies within 0.02 of the best model in 4 of the six years. Table 3 extends the comparison to cut-offs between 1% and 5%, the range examined by Bao et al. [2]; TRACE has the highest NDCG at every cut-off, and its NDCG exceeds that of RUSBoost at all five budgets.

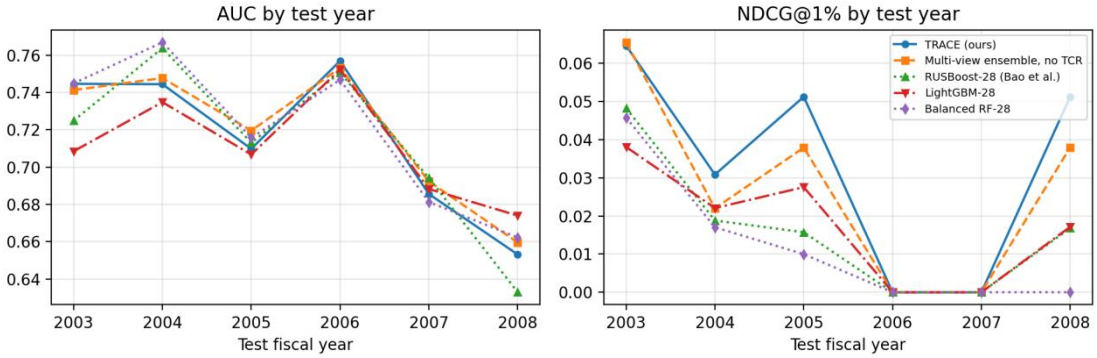


Figure 2 AUC and NDCG@1% by Test Year (Mean over Replicates)

Table 3 NDCG and Sensitivity (NDCG/Sens) at Investigation Budgets from 1% to 5% of the Test Firm-Years, Computed on Replicate-Averaged Scores

Model	1%	2%	3%	4%	5%
RUSBoost-28 (Bao et al.)	0.017/0.022	0.027/0.039	0.043/0.071	0.055/0.096	0.068/0.124
LightGBM-28	0.020/0.023	0.033/0.045	0.046/0.070	0.053/0.086	0.069/0.122
Multi-view ensemble, no TCR	0.027/0.029	0.038/0.049	0.057/0.087	0.073/0.119	0.082/0.138
TRACE (ours)	0.030/0.029	0.040/0.047	0.062/0.087	0.079/0.127	0.093/0.157

6.2 Statistical Significance

Table 4 reports bootstrap comparisons of TRACE with every baseline. The NDCG@1% gains over RUSBoost (*+0.0144*, $p = 0.019$), LightGBM (*+0.0111*, $p = 0.010$), the balanced random forest (*+0.0213*, $p = 0.013$), XGBoost (*+0.0139*, $p = 0.042$) and the dollar-view ensemble (*+0.0108*, $p = 0.007$) are significant at the 5% level, and the AUC gains over the logistic models and over LightGBM-42 are significant as well. The AUC differences between TRACE and the strongest tree ensembles are small and not significant, confirming that TRACE matches their global ranking quality while concentrating misstatements at the top of the list.

Table 4 Stratified Bootstrap Comparison of TRACE with Each Alternative (1,000 Resamples within Test Years; Replicate-Averaged Scores)

Alternative	ΔAUC	95% CI	p	ΔNDCG@1%	95% CI	p
RUSBoost-28 (Bao et al.)	+0.0011	[-0.0148, +0.0177]	0.447	+0.0144	[+0.0008, +0.0326]	0.019
Balanced RF-28	-0.0035	[-0.0197, +0.0123]	0.659	+0.0213	[+0.0027, +0.0445]	0.013
XGBoost-28	+0.0090	[-0.0046, +0.0224]	0.113	+0.0139	[-0.0013, +0.0310]	0.042
LightGBM-28	+0.0002	[-0.0130, +0.0136]	0.497	+0.0111	[+0.0010, +0.0248]	0.010
LightGBM-42 (raw+ratios)	+0.0257	[+0.0048, +0.0454]	0.006	+0.0088	[-0.0066, +0.0249]	0.114
LightGBM-28 + tax ratios	+0.0035	[-0.0104, +0.0171]	0.292	+0.0138	[+0.0008, +0.0284]	0.019
Logit-14 (Dechow et al. ratios)	+0.0357	[+0.0003, +0.0701]	0.024	+0.0050	[-0.0273, +0.0377]	0.381
Logit-28	+0.0404	[+0.0118, +0.0702]	0.005	+0.0179	[+0.0017, +0.0368]	0.013
Dollar view ×4	-0.0043	[-0.0166, +0.0085]	0.760	+0.0108	[+0.0012, +0.0237]	0.007
Multi-view ensemble, no TCR	-0.0028	[-0.0116, +0.0059]	0.745	+0.0037	[-0.0037, +0.0132]	0.175

Note: Δ = TRACE minus alternative; $p = \Pr(\Delta \leq 0)$.

6.3 Ablation Study

Table 5 isolates the contribution of each component. Removing the tax-consistency residuals from both views (both views, no TCR) lowers NDCG@1% from 0.0330 to 0.0272, sensitivity from 0.0342 to 0.0289 and AP from 0.0205 to 0.0195. The two views play complementary roles: with TCR, the peer-rank view alone reaches an NDCG@1% of 0.0329 but a lower AUC (0.6998), whereas the dollar view alone keeps a high AUC (0.7167) but ranks fewer misstatements at the top

(0.0222); TRACE combines the top-of-list precision of the former with the global ranking quality of the latter. Adding TCR improves each view separately as well (dollar view: 0.0192 to 0.0222; peer-rank view: 0.0263 to 0.0329), and giving TCR to every member is preferable to giving it to half of them (0.0297). The lower block varies the construction of the residuals. Permuting the residuals across firms within each year (placebo), which preserves their marginal distribution but destroys their link to the firm's own statements, reduces NDCG@1% to 0.0240; the gain therefore stems from firm-specific tax consistency rather than from additional input dimensions. Removing cross-tax conditioning, i.e. predicting tax expense without taxes payable and vice versa, has the largest effect (NDCG@1% 0.0226, AUC 0.7113): the internal consistency of the tax block is the most informative part of the signal. Replacing the cross-fitted gradient-boosted expectations with in-sample residuals without cross-fitting (0.0305) or with a linear Huber expectation model in place of gradient boosting (0.0297) also lowers NDCG@1%, in line with Proposition 1 and with the non-linear relation between tax accounts and fundamentals. A homoscedastic variant that scales residuals by a single year-level dispersion attains a comparable NDCG@1% (0.0357, AUC 0.7176), which locates the information in the conditional expectation rather than in the dispersion model; the heteroscedastic scale is retained because it was selected on the validation years and makes residuals comparable across firms of different size.

Table 5 Ablation Study (Test Years 2003–2008; Mean with Standard Deviation across Replicates)

Variant	AUC	AP	NDCG@1%	Sens@1%	Prec@1%
Dollar view only (no TCR, no peer-rank view)	0.7201 (0.0028)	0.0178 (0.0001)	0.0192 (0.0033)	0.0215 (0.0048)	0.0197 (0.0029)
Dollar view + TCR	0.7167 (0.0017)	0.0184 (0.0002)	0.0222 (0.0020)	0.0254 (0.0026)	0.0225 (0.0028)
Peer-rank view only	0.7029 (0.0031)	0.0191 (0.0006)	0.0263 (0.0017)	0.0306 (0.0046)	0.0255 (0.0029)
Peer-rank view + TCR	0.6998 (0.0060)	0.0202 (0.0007)	0.0329 (0.0016)	0.0333 (0.0015)	0.0292 (0.0016)
Both views, no TCR	0.7190 (0.0028)	0.0195 (0.0008)	0.0272 (0.0012)	0.0289 (0.0026)	0.0254 (0.0028)
Both views, TCR in half of the members	0.7155 (0.0025)	0.0198 (0.0005)	0.0297 (0.0025)	0.0314 (0.0044)	0.0264 (0.0033)
Both views, TCR in all members (TRACE)	0.7157 (0.0037)	0.0205 (0.0003)	0.0330 (0.0015)	0.0342 (0.0022)	0.0292 (0.0016)
TCR without cross-fitting (in-sample residuals)	0.7193 (0.0032)	0.0209 (0.0004)	0.0305 (0.0021)	0.0310 (0.0009)	0.0281 (0.0000)
TCR without cross-tax conditioning	0.7113 (0.0018)	0.0186 (0.0006)	0.0226 (0.0030)	0.0264 (0.0034)	0.0226 (0.0039)
TCR with linear (Huber) expectation model	0.7202 (0.0051)	0.0208 (0.0006)	0.0297 (0.0006)	0.0296 (0.0006)	0.0254
TCR without heteroscedastic scaling	0.7176 (0.0072)	0.0209 (0.0005)	0.0357 (0.0015)	0.0372 (0.0008)	0.0324 (0.0019)
Placebo TCR (permuted within year)	0.7191 (0.0014)	0.0193 (0.0007)	0.0240 (0.0017)	0.0280 (0.0017)	0.0240 (0.0020)

Note: The upper block varies the ensemble design with four members each; the lower block keeps the TRACE design and modifies the construction of the tax-consistency residuals.

6.4 Sensitivity Analysis

Figure 3 varies one hyper-parameter at a time around the configuration selected on the validation years (negative-bagging fraction, number of boosting rounds, number of cross-fitting folds, number of leaves). Across all 7 perturbed configurations the AUC stays between 0.7123 and 0.7248 and NDCG@1% between 0.0287 and 0.0365; NDCG@1% remains above that of RUSBoost (0.0166) in 7 of the 7 configurations. The weakest setting is negative-bagging fraction = *0.2* (NDCG@1% 0.0287). The method does not hinge on a narrow choice of hyper-parameters.

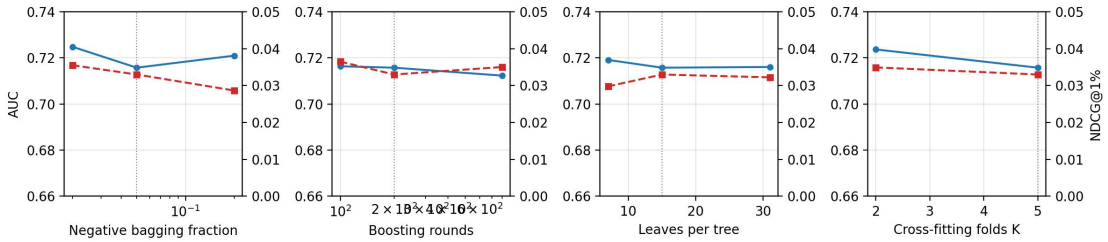


Figure 3 Sensitivity of TRACE to the Negative-Bagging Fraction, the Number of Boosting Rounds, the Number of Leaves and the Number of Cross-Fitting Folds (AUC: Solid Line, Left Axis; NDCG@1%: Dashed Line, Right Axis)
Note: Dotted vertical lines mark the configuration selected on the validation years.

6.5 Robustness

Table 6 repeats the comparison under alternative protocols. When serial-fraud firm-years are dropped from training instead of relabelled, TRACE reaches an NDCG@1% of 0.0327 (AUC 0.7170), which exceeds RUSBoost (0.0148, AUC 0.7113) and the dollar-view ensemble (0.0237, AUC 0.7201). With a rolling ten-year training window, TRACE reaches an NDCG@1% of 0.0263 (AUC 0.7181), which exceeds RUSBoost (0.0175, AUC 0.7253) and the dollar-view ensemble (0.0173, AUC 0.7141). On the extended test period 2009–2012, TRACE reaches an NDCG@1% of 0.0430 (AUC 0.6944), which exceeds RUSBoost (0.0394, AUC 0.7353) and the dollar-view ensemble (0.0272, AUC 0.7077). Excluding the AAER firm-years flagged as understatements from the test labels leaves the picture unchanged (TRACE 0.0310 versus RUSBoost 0.0174 in NDCG@1%).

Table 6 Robustness of the Results to the Serial-Fraud Protocol, the Training Window, the Test Period and the Label Definition (Mean over Test Years and Replicates)

Setting	Model	AUC	AP	NDCG@1%	Sens@1%	Prec@1%
Serial-fraud firm-years dropped from training	RUSBoost-28 (Bao et al.)	0.7113	0.0170	0.0148	0.0175	0.0168
	Dollar view ×4	0.7201	0.0189	0.0237	0.0255	0.0226
	TRACE (ours)	0.7170	0.0205	0.0327	0.0337	0.0296
Rolling 10-year training window	RUSBoost-28 (Bao et al.)	0.7253	0.0175	0.0175	0.0226	0.0197
	Dollar view ×4	0.7141	0.0172	0.0173	0.0189	0.0196
	TRACE (ours)	0.7181	0.0200	0.0263	0.0280	0.0240
Extended test period 2009–2012	RUSBoost-28 (Bao et al.)	0.7353	0.0129	0.0394	0.0643	0.0276
	Dollar view ×4	0.7077	0.0120	0.0272	0.0389	0.0184
	TRACE (ours)	0.6944	0.0149	0.0430	0.0524	0.0230
Overstatement-only test labels	RUSBoost-28 (Bao et al.)	0.7231	0.0173	0.0174	0.0230	0.0197
	Dollar view ×4	0.7277	0.0177	0.0204	0.0238	0.0197
	TRACE (ours)	0.7226	0.0206	0.0310	0.0306	0.0254

6.6 Label-Free Identification of Tax-Related Anomalies

Because TCR uses no labels, the aggregate residual $\bar{z}$ is itself a screening statistic for tax-related reporting anomalies. Table 7 compares it with four generic outlier detectors applied, year by year, to the same fundamentals and tax ratios. The aggregate TCR score attains an AUC of 0.592 against at most 0.522 for the generic detectors, and the highest NDCG@1% and precision. Because enforcement concentrates on larger firms, we also report the AUC averaged within size quintiles; TCR remains the best detector (0.558 versus at most 0.547). Generic detectors flag firms that are globally unusual, most of which are small, distressed or extreme in scale; TCR flags firms whose tax accounts are unusual *given* their own fundamentals, which is the notion of anomaly that matters for reporting quality.

Table 7 Label-Free Anomaly Scoring of Misstatements (Test Years 2003–2008)

Detector	AUC	AUC-size	AP	NDCG@1%	Prec@1%
IForest	0.4582	0.5206	0.0068	0.0000	0.0000
HBOS	0.4338	0.5110	0.0062	0.0000	0.0000
ECOD	0.5135	0.5470	0.0086	0.0026	0.0028
COPOD	0.5218	0.5464	0.0085	0.0056	0.0056
TCR max $\lvert z_k \rvert$ (ours)	0.5892	0.5576	0.0094	0.0080	0.0058
TCR aggregate $\bar{z}$ (ours)	0.5924	0.5583	0.0102	0.0105	0.0085

Note: AUC-size is the AUC averaged within size quintiles.

6.7 Explanation Quality

We explained all 352 firm-years flagged by TRACE in 6 test years and compared explanation methods on a subset of 48 of them (Table 8, Figure 4). PSE satisfies completeness to machine precision (maximum absolute error 5.3e-15 log-odds), whereas the completeness error of LIME is 1.818 of the score gap. Its attributions agree with KernelSHAP on the same pipeline with a Spearman correlation of 0.976 and a top-5 overlap of 0.92; for reference, two independent KernelSHAP runs agree with $\rho = 0.995$ and a top-5 overlap of 0.97, and LIME reaches $\rho = -0.822$. In the deletion test, removing the ten line items ranked highest by PSE lowers the normalized score to an area of 0.124, compared with 0.115 for KernelSHAP, 1.210 for LIME, 0.136 for the naive mapping and 0.747 for random orderings; in the insertion test the areas are 0.699, 0.701, -0.395, 0.694 and 0.135, respectively. PSE is within 0.009 of KernelSHAP in deletion area and within 0.001 in insertion area; KernelSHAP estimates the same random-baseline Shapley target by sampling the composed pipeline. The naive mapping, which credits each residual to its reported tax item, is less faithful than PSE on both tests, confirming that the expected and dispersion components of the residuals carry information that a feature-level explanation misattributes. LIME's local surrogate is poorly aligned with the pipeline ($\rho = -0.82$), and its orderings are less faithful than random ones. PSE requires 2.94 seconds per firm-year on one CPU core, against 7.99 seconds for KernelSHAP (2.7

times slower) and 0.80 seconds for LIME, and it is deterministic. The conditioning guard of (7) was invoked for 59.6% of the non-zero allocations, which quantifies how often plain rescale propagation would have been ill-conditioned in this pipeline.

Table 8 Explanation Quality on the Comparison Subset

Method	ρ	Top-5	Compl. err.	Deletion ↓	Insertion ↑	Time (s)
PSE (ours)	0.976	0.921	0.000	0.124	0.699	2.94
KernelSHAP (2,048 coalitions)	0.995	0.967	0.000	0.115	0.701	7.99
LIME (3,000 samples)	-0.822	0.625	1.818	1.210	-0.395	0.80
Feature-level TreeSHAP, naive mapping	0.957	0.904	0.000	0.136	0.694	–
Random ordering	–	–	–	0.747	0.135	–

Note: ρ and top-5 overlap are measured against KernelSHAP on the same pipeline and peer baselines (for KernelSHAP: against an independent KernelSHAP run). Completeness error is $\left|\sum_c \phi_c - (F(p) - \bar{F}_{\mathcal{B}})\right| / \left|F(p) - \bar{F}_{\mathcal{B}}\right|$. Deletion (lower is better) and insertion (higher is better) areas use the ten top-ranked line items. Runtime in seconds per explanation on one CPU core.

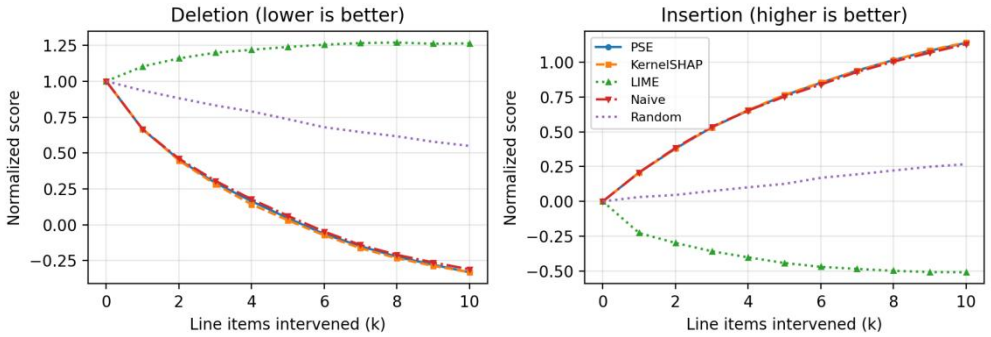


Figure 4 Deletion and Insertion Curves: Normalized TRACE Score as the k Highest-Ranked Line Items of Each Explanation Are Replaced by Peer-Baseline Values (Deletion) or Inserted into the Peer Baseline (Insertion)
Note: Averages over the comparison subset.

6.8 What Tax-Related Anomalies Look Like

Figure 5 decomposes the PSE attributions of the 352 flagged firm-years, 11 of which are AAER misstatements, into accounting concepts. Income-tax line items account for 5.8% of the absolute attribution of detected misstatements and 4.6% for the other flagged firm-years; the largest share for detected misstatements falls on financing & equity. Because the expected component of each residual is propagated to the fundamentals that determine the peer expectation, these concept shares understate the role of the tax channel, which is measured directly by the Owen units. The tax-consistency channel raises the score of 91% of the detected misstatements and of 84% of the other flagged firm-years. Table 9 reports, for each flagged firm-year, the tax relationship that contributes most to its score. Among detected misstatements the most frequent patterns are a reported change in tax expense above its peer expectation (45.5%) and reported taxes payable below its peer expectation (27.3%), against 32.0% and 14.4% among the other flagged firm-years. Tax expense that grows faster than peers' together with a tax liability below the peer expectation is the signature of book income that is not matched by taxable income, the pattern that the book-tax literature associates with upward earnings management [6, 7]. Table 10 and Figure 6 show the two detected misstatements with the largest tax-channel contribution. For the first (gvkey 29095, fiscal year 2005), the reported value of the change in taxes payable scaled by assets is 0.0679 against a peer expectation of -0.0036 (residual +5.46); the Owen allocation attributes +0.17 log-odds to the reported component and +0.05 to the expected component, so the explanation states both what the firm reported and why comparable firms would report something different. An examiner can verify each of these quantities directly in the filing and in the peer cross-section.

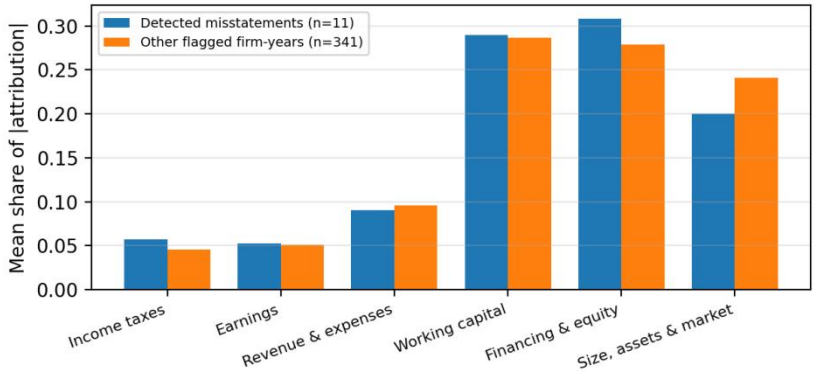


Figure 5 Share of Absolute PSE Attribution by Accounting Concept for Flagged Firm-Years That Are AAER Misstatements and for the Other Flagged Firm-Years

Table 9 Dominant Tax-Consistency Pattern among Flagged Firm-Years: Share of Firm-Years Whose Largest Positive Tax-Channel Contribution Comes from the Given Target, Split by the Sign of Its Residual

Pattern	Detected misstatements	Other flagged
Reported change in tax expense above peer expectation	45.5%	32.0%
Reported change in tax expense below peer expectation	0.0%	1.5%
Reported change in taxes payable above peer expectation	9.1%	20.8%
Reported change in taxes payable below peer expectation	0.0%	2.3%
Reported tax expense above peer expectation	9.1%	14.4%
Reported tax expense below peer expectation	9.1%	10.9%
Reported taxes payable above peer expectation	0.0%	2.6%
Reported taxes payable below peer expectation	27.3%	14.4%

Table 10 Case Studies: Detected Misstatements with the Largest Positive Tax-Consistency Contribution

gvkey	FY	AAER	Gap	TCR	Target	y	m	z	Rep./Exp./Disp.
29095	2005	2953	+3.90	+0.57	change in taxes payable	0.0679	-0.0036	+5.46	+0.17 / +0.05 / -0.01
3362	2003	2628	+2.76	+0.51	change in tax expense	-0.0006	-0.0042	+0.43	+0.07 / +0.22 / +0.01

Note: Score gap = $F(p) - \bar{F}_{\mathcal{B}}$ (log-odds); TCR total = sum of the attributions of the four residuals; for the dominant target we report the reported value y, its peer expectation m, the residual z, and the Owen allocation to the reported/expected/dispersion components.

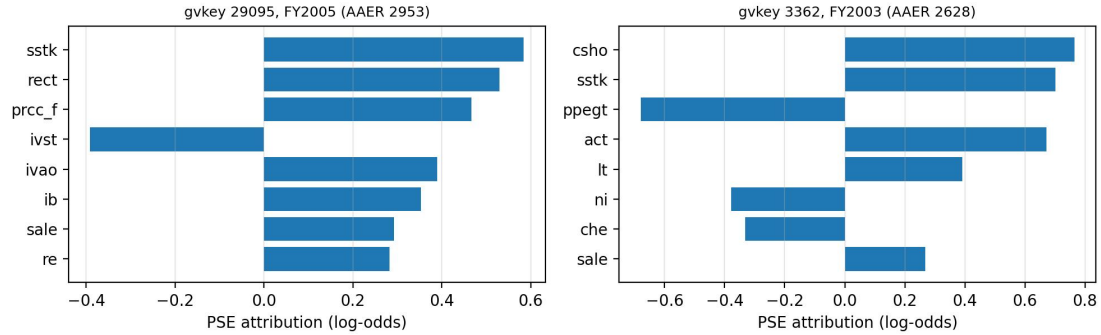


Figure 6 PSE Line-Item Attributions for the Two Case-Study Firm-Years (Eight Largest in Absolute Value; Income-Tax Items in Red)

7 DISCUSSION

7.1 Implications for Screening Practice

Audit firms, the SEC and tax authorities operate under investigation budgets, so the quality of the first few percent of a ranking matters more than global discrimination. TRACE improves precisely this part of the ranking. Because PSE reports attributions in units of filed line items and separates what a firm reported from what comparable firms report, a flagged case arrives with a concrete question for the examiner: why does the reported tax expense, payable or change differ from the peer expectation, and which fundamentals determine that expectation?

7.2 Tax Consistency as a Reusable Signal

The tax-consistency residuals are label-free and computed from the filing year's cross-section, so they can be produced for every filer as soon as annual reports are available and can be combined with any detector, including models that use market, audit or textual information [3]. The evidence that the TCR score outperforms generic outlier detectors without any labels suggests that domain-targeted conditional anomaly scores are a more productive route than global unusualness for financial-statement data.

7.3 Explaining Pipelines with Engineered Features

The Owen allocation at residual nodes is not specific to taxes. Any detector that consumes standardized residuals of expectation models, such as abnormal accruals, abnormal revenue or abnormal inventory, faces the same cancellation problem when explanations are propagated with the rescale rule. Treating the components of each residual as an a priori union and evaluating the downstream trees at hybrid residuals yields exact, interpretable attributions at a cost that grows only linearly in the number of baselines.

Scope. The analysis uses the public AAER panel and the accounting items it contains; deferred tax accounts, current tax expense and cash taxes paid are not part of the released data. Richer tax disclosures, such as the components of the tax provision or uncertain tax benefits, can be added to TCR as further targets without changing the method, and the same construction applies to other enforcement-based or restatement-based labels.

8 CONCLUSION

We introduced TRACE, an explainable framework for identifying tax-related financial reporting anomalies in U.S. public companies. Cross-fitted tax-consistency residuals turn the accounting insight that book and tax numbers must cohere into a label-free, peer-relative signal; a validation-selected multi-view ensemble turns that signal into significantly better screening at audit-capacity cut-offs than RUSBoost, random forests, XGBoost and LightGBM; and Propagated Shapley Explanations, built on exact single-baseline TreeSHAP and an exact Owen allocation at residual nodes, attribute each score to the line items of the filing with a completeness guarantee. On 352 flagged firm-years, PSE explanations are complete to machine precision and 2.7 times faster than KernelSHAP, and they expose the tax relationships behind each detection.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

ACKNOWLEDGEMENTS

The authors thank the creators of the public AAER dataset for making their data and code available.

DATA AVAILABILITY

The AAER-labelled Compustat panel used in this study is publicly available from the repository of Bao et al. [2] at https://github.com/JarFraud/FraudDetection.

REFERENCES

[1] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28(1): 17-82.
[2] Bao Y, Ke B, Li B, et al. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 2020, 58(1): 199-235.
[3] Bertomeu J, Cheynel E, Floyd E, et al. Using machine learning to detect misstatements. Review of Accounting Studies, 2021, 26(2): 468-519.
[4] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206-215.
[5] Bertomeu J. Machine learning improves accounting: discussion, implementation and research opportunities. Review of Accounting Studies, 2020, 25(3): 1135-1155.
[6] Hanlon M. The persistence and pricing of earnings, accruals, and cash flows when firms have large book-tax differences. The Accounting Review, 2005, 80(1): 137-166.
[7] Blaylock B, Shevlin T, Wilson R J. Tax avoidance, large positive temporary book-tax differences, and earnings persistence. The Accounting Review, 2012, 87(1): 91-120.
[8] Graham J R, Raedy J S, Shackelford D A. Research in accounting for income taxes. Journal of Accounting and Economics, 2012, 53(1-2): 412-434.
[9] Hanlon M, Heitzman S. A review of tax research. Journal of Accounting and Economics, 2010, 50(2-3): 127-178.
[10] Frank M M, Lynch L J, Rego S O. Tax reporting aggressiveness and its relation to aggressive financial reporting. The Accounting Review, 2009, 84(2): 467-496.
[11] Lennox C, Lisowsky P, Pittman J. Tax aggressiveness and accounting fraud. Journal of Accounting Research, 2013, 51(4): 739-778.
[12] Erickson M, Hanlon M, Maydew E L. How much will firms pay for earnings that do not exist? Evidence of taxes paid on allegedly fraudulent earnings. The Accounting Review, 2004, 79(2): 387-408.
[13] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1135-1144.
[14] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, 2017: 4765-4774.
[15] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017: 3145-3153.
[16] Chen H, Lundberg S M, Lee S-I. Explaining a series of models by propagating Shapley values. Nature Communications, 2022, 13: 4512.
[17] Owen G. Values of games with a priori unions. In: Henn R, Moeschlin O, eds. Essays in Mathematical Economics and Game Theory. Berlin: Springer, 1977: 76-88.
[18] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 2010, 40(1): 185-197.
[19] Walker S. Critique of an article on machine learning in the detection of accounting fraud. Econ Journal Watch, 2021, 18(1): 61-70.
[20] Bao Y, Ke B, Li B, et al. A response to "Critique of an article on machine learning in the detection of accounting fraud". Econ Journal Watch, 2021, 18(1): 71-78.

[21] Perols J L, Bowen R M, Zimmermann C, et al. Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 2017, 92(2): 221-245.
[22] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track), 2022.
[23] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30, 2017: 3146-3154.
[24] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
[25] Liu F T, Ting K M, Zhou Z-H. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 3.
[26] Goldstein M, Dengel A. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In: KI-2012: Poster and Demo Track, 2012: 59-63.
[27] Li Z, Zhao Y, Botta N, et al. COPOD: Copula-based outlier detection. In: 2020 IEEE International Conference on Data Mining (ICDM), 2020: 1118-1123.
[28] Li Z, Zhao Y, Hu X, et al. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. IEEE Transactions on Knowledge and Data Engineering, 2022.
[29] Zhao Y, Nasrullah Z, Li Z. PyOD: A Python toolbox for scalable outlier detection. Journal of Machine Learning Research, 2019, 20(96): 1-7.
[30] Paulheim H, Meusel R. A decomposition of the outlier detection problem into a set of supervised learning problems. Machine Learning, 2015, 100(2-3): 509-531.
[31] Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 2018, 21(1): C1-C68.
[32] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56-67.
[33] Sundararajan M, Najmi A. The many Shapley values for model explanation. In: Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2020: 9269-9278.
[34] Petsiuk V, Das A, Saenko K. RISE: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC), 2018.