Harvard Data Science Review • Issue 8.2, Spring 2026

Explainability in the Wild, or Wild Explanations? Evidence From Predicting Tax Evasion

Isabel O. Gallegos[1] Benjamin B. Seiler[1,2] Jacob Goldin[3] Carlos Guestrin[1] Daniel E. Ho[1]

[1]Stanford University, Stanford, California, United States of America,
[2]Williams College, Williamstown, Massachusetts, United States of America,
[3]University of Chicago, Chicago, United States of America

The MIT Press

Published on: Feb 18, 2026

DOI: https://doi.org/10.1162/99608f92.c0e75932

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

ABSTRACT

As artificial intelligence (AI) algorithms become more prevalent in high-stakes risk assessment, policymakers have increasingly relied on *explainability* tools for interpretability. Despite growing mandates that AI-based decisions include explanations, there remains little empirical evidence demonstrating the effectiveness of these techniques in real-world applications. This gap often stems from the absence of a clear ground truth for evaluating explanations. In this work, we present an empirical evaluation of explanation techniques in collaboration with the United States Internal Revenue Service (IRS). Using real, line-by-line IRS audits from randomly selected taxpayers, we decompose one component of aggregate tax underreporting into its constituent line item misreporting and apply explainability techniques to recover these risks; the aggregate risk is a function of the constituent risk. Our study makes three contributions. First, we empirically evaluate how well *local explanation models* recover true constituent risks. Second, we compare local explanation models to estimating constituent risks directly. Finally, we situate these findings in a practical setting where explanations are critical not only for transparency but also as guidance for the users of the model's predictions. Our analysis reveals that the quality of local explanations is tied to the quality of the underlying model. Yet even with a theoretically perfect underlying model, local explanations still fail to accurately capture the true risk. While directly estimating constituent risks may yield more accurate results, simplistic rule-based heuristics often overlook the complexity of risk. These findings highlight the need for thoughtful application of explanation techniques in high-risk domains, where errors can have significant consequences.

Keywords: artificial intelligence, explainability, risk assessment

Media Summary

As artificial intelligence (AI) algorithms play a growing role in high-stakes decision-making, policymakers increasingly require that AI decisions be explainable to ensure transparency and interpretability in automated decisions. Explanations can provide information about how an AI model arrives at each individual prediction. However, there is little evidence showing whether these explanations work as intended.

In collaboration with the United States Internal Revenue Service (IRS), we examine AI explanation techniques in a real-world tax audit setting. Using IRS audit data, we empirically assess how accurately model explanations capture the constituent risks of tax underreporting that comprise an individual's overall risk. We find that these explanation methods often fail to accurately capture the true risk, even when the AI model itself is highly accurate. As alternatives, we also explore estimating constituent risks directly instead of relying on explanations. We find that, while some direct methods can provide better accuracy, overly simplistic approaches may ignore important complexities in tax underreporting.

These findings raise important questions about how AI explanations should be used in high-stakes settings like tax audits. If explanations are inaccurate, they can mislead decision-makers, with significant consequences. As decision-making processes increasingly rely on AI, this study highlights the need for careful consideration of how AI models and their explanations are designed and applied.

---

1. Introduction

Artificial intelligence (AI) algorithms have been increasingly integrated into a vast array of decision-making processes with material consequences on people's lives and their outcomes. As the stakes of these decisions have risen, so too have the calls for transparency in these automated systems. Models designed or selected to be interpretable from the start can, of course, offer transparency, but with the thrust of recent AI advancements toward uninterpretable deep neural networks and other complex models, policymakers have sought tools to better understand the underlying mechanisms driving these decisions. At the forefront of these tools are *explanation models*, supported by an extensive suite of different methods that seek to provide interpretable insights into the predictive processes of uninterpretable, black-box models. There are two primary classes of explanation models: (i) *global* explanations, which provide an understanding of what the model has learned and how it makes predictions in general, and (ii) *local* explanations, which seek to explain the model's predictions for individual instances.

The importance of explanations is apparent through regulations that make them not only desirable but also legally required. This is not a new development; in the United States, for instance, laws like the Fair Credit Reporting Act (1970) and the Equal Credit Opportunity Act (1974) have long mandated that creditors provide an explanation when requested if they deny a person a loan. What is new, however, is the extension of such principles to the paradigm of automated and often opaque decision-making processes, with recent AI regulations defining new requirements for explainability (Nannini et al., 2023; Panigutti et al., 2023). For instance, Article 86 of the European Union's Artificial Intelligence Act (Regulation 2024/1689), one of the primary AI regulatory frameworks to date, defines a right to explanation, where any person subject to a decision made by a high-risk AI system is entitled to "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken". Even in settings where explanations are not legally required, a lack of explainability can hinder the ability of downstream users to utilize an algorithm's predictions. This issue is demonstrated in the empirical application we discuss below.

Despite the influx of explainability-based policy proposals and regulations, there exists little empirical evidence of the efficacy of such techniques to provide the desired and required interpretability in real-world settings. Though prior work has provided important insights into tensions that can arise from explainability techniques, existing analyses have tended to be theoretical or abstract critiques that highlight cases where explainability techniques may fail (e.g. Aivodji et al., 2019; Balagopalan et al., 2022; Dai et al., 2022; Rudin,

2019; Shamsabadi et al., 2022; Slack et al., 2020). What remains missing, however, is an understanding of these tensions in the wild, a challenge made difficult by the absence of ground truth for explanations in many settings (Gilpin et al., 2018; Leavitt & Morcos, 2020; Nauta et al., 2023; Yang et al., 2019), with prior work often relying on synthetic data sets to derive some notion of ground truth explanations (e.g., Agarwal et al., 2022; Arras et al., 2022; Guidotti, 2021; Liu et al., 2021; Miró-Nicolau et al., 2024), if not lacking ground truth or systematic quantitative assessment altogether.

In this article, we provide an empirical evaluation of local explanation models in a high-stakes setting for risk assessment, with a well-defined notion of ground truth for local explanations. In this setting, overall risk can be decomposed into a set of constituent elements, which are used to inform decision-making; a local explanation model should recover each of the constituent components. We focus on the setting of tax audit selection by the United States Internal Revenue Service (IRS), the agency responsible for collecting the revenue to fund the federal government. Here, an individual taxpayer's total risk is defined by the difference, or *misreporting*, between taxes owed and taxes paid for their entire tax return. The amount of taxes owed is calculated based on the net total of individual *line items*, such as total wages, total itemized deductions, and net profits from self-employment; for example, gross sales, supplies expenses, and utilities expenses are all line items that constitute the aggregate profit. The misreporting for each line item is similarly defined by the difference between reported and true line item values. In this unique setting, the sum of the line item misreporting amounts equals the total misreporting amount, making the overall risk a composite of the individual line item risks. Typically, auditors examine only a small number of line items due to auditor specialization, audit complexity, or cost constraints; thus, the availability of accurate constituent risks can also guide auditors using the model's predictions about total risk. Leveraging a stratified random sample of real, line-by-line IRS audits from randomly selected taxpayers, this context provides a unique opportunity to explore local explanation models with several key advantages: (i) a concrete definition of ground truth provided by the constituent risk misreporting amounts, (ii) a practical setting where explanations are critical not only for transparency but also as actionable outputs guiding auditors on which constituent risks to prioritize, and (iii) a data set representative of the filing population that avoids the selective labels problems (Lakkaraju et al., 2017).

With this context in mind, our work makes three primary contributions. First, *we empirically answer the question, to what extent do machine learning models with local explanation models recover the true constituent risks?* The first research question is a matter of *fidelity*, or how accurately local explanations match the true underlying decision process. The importance of fidelity is two-fold given access to the true constituent risk values: (i) explanations can have high fidelity to a model but low accuracy with respect to the ground truth if the model fails to predict risk well; and (ii) explanations can have low fidelity to both the model and to the ground truth if the explanations fail to represent the model well, thereby introducing erroneous interpretations of the model's decision-making process. Using a machine learning model to predict individuals' total tax misreporting, we apply local explanation models to identify each constituent misreporting amount and assess

the quality of these explanations by comparing them to the actual (ground truth) constituent misreporting amount imposed upon audit. Our findings reveal that in the IRS tax audit setting, these local explanations often fall short of accurately representing individuals' true constituent risk.

Second, *we compare local explanation models to alternative approaches that estimate the constituent risks directly.* Rudin (2019) has advocated for inherently interpretable models instead of relying on inadequate explanations for black-box models in high-stakes decision-making. In this vein, and in light of the shortcomings of local explanations we observe, the second research question explores the possibility of estimating constituent risks directly, rather than deriving these values from explanations of a complex system. We consider two alternative paradigms: (i) modeling each line item misreporting amount with a separate machine learning model, and (ii) estimating each line item misreporting amount with a simple and interpretable heuristic. Our findings demonstrate that directly estimating each line item misreporting amount separately with a machine learning model yields results comparable to or marginally more faithful to the true risk than those derived from local explanation models. However, simple rule-based heuristics may fail to capture the complexity of individual risk, with the coarsest heuristics performing worst.

Finally, *we demonstrate these findings on real audit data collected by the IRS*, an important case study that reflects a real-world setting where machine learning and AI techniques may be used for audit selection (Black et al., 2022; Elzayn et al., 2024; Henderson et al., 2023; Internal Revenue Service, 2023a) and a setting where explainability-focused regulatory action also applies.[1] This setting is unique because the true values that local explanation models should capture are precisely known in our random audit data, and the explanations that describe high-risk taxpayers are directly actionable, a critical necessity given the nearly $700B annual tax gap, or difference between taxes owed and taxes paid, in the United States (Internal Revenue Service, 2023c). Our work provides an important perspective on the limitations of local explainability techniques and the trade-offs with alternative modeling and prediction approaches at a critical moment when the regulatory space is ripe with growing concerns about the transparency and trustworthiness of AI models for high-risk decisions. Our findings highlight the need for careful consideration about when and how to apply local explanation models in high-risk settings where the consequences of modeling and explanation errors are significant.

2. Explainable AI

Explanation methods in machine learning offer an understanding of the decision-making process of complex or uninterpretable models. This broad set of techniques has garnered significant attention from policymakers and practitioners as they strive to make high-stakes, automated decisions more trustworthy and interpretable. In this work, we focus primarily on local explanation models that can provide *additive feature attributions*. Consider a black-box prediction model f that maps from an input $\mathbf{x}$ to a prediction $f(\mathbf{x})$. These techniques aim to learn an explanation model g such that, given some query point $\mathbf{x}^*$, $g(\mathbf{x}^*) \approx f(\mathbf{x}^*)$ at and in the local neighborhood of $\mathbf{x}^*$. This objective is often slightly modified to employ simplified, more interpretable inputs

$\mathbf{x}'$ (e.g., a binarized version of $\mathbf{x}$), such that $g(\mathbf{x}') \approx f(\mathbf{x})$ at and near $\mathbf{x}'$. Given these notions, additive feature attribution methods learn an explanation model that is a linear function of the simplified features

$$g(\mathbf{x}') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' \qquad (2.1)$$

where $\mathbf{x}' \in \{0,1\}^M$ is the simplified input, M is the number of input features, and $\phi_i \in \mathbb{R}$. An *attribution* of the prediction at input $\mathbf{x}$ is the vector $[\phi_1, \cdots, \phi_M] \in \mathbb{R}^M$, where ϕ_i is the contribution of x_i to the prediction $f(\mathbf{x})$.

We consider three additive feature attribution explanation methods: SHAP (SHapley Additive exPlanation) (Lundberg & Lee, 2017), Cohort Shapley (Mase et al., 2019), and LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016). We include SHAP and LIME because they are the most commonly used off-the-shelf explanation methods, and we include Cohort Shapley for its applicability to black-box models. SHAP values are adapted from Shapley values in cooperative game theory (Shapley, 1953). Intuitively, SHAP values compute the marginal contribution of a feature, based on all feature subsets, comparing model predictions with and without each feature present. Lundberg and Lee (2017) show that the only explanation model g that satisfies properties of *local accuracy*, *missingness*, and *consistency* uses Shapley regression values to determination feature attributions

$$\phi_i(f, \mathbf{x}) = \sum_{\mathbf{z}' \subseteq \mathbf{x}'} \frac{|\mathbf{z}'|!(M - |\mathbf{z}'| - 1)!}{M!} [f_{\mathbf{x}}(\mathbf{z}') - f_{\mathbf{x}}(\mathbf{z}' \setminus i)] \qquad (2.2)$$

where $\mathbf{z}' \in \{0,1\}^M$, $|\mathbf{z}'|$ is the number of nonzero entries in $\mathbf{z}'$, and $\mathbf{z}' \subseteq \mathbf{x}'$ represents all $\mathbf{z}'$ vectors where the nonzero entries are a subset of the nonzero entries in $\mathbf{x}'$. Local accuracy means that the explanation model matches the original model for the simplified input $\mathbf{x}'$, such that $f(\mathbf{x}) = g(\mathbf{x}')$. Missingness means that features with $x_i' = 0$ have no attribution. Consistency means that, if the contribution of a feature to a prediction is higher in a different model, then its corresponding SHAP value will be higher as well.

Perturbations to an individual's feature vectors in SHAP can create hybrid data points that are physically or logically impossible. Instead, using only observed values, Cohort Shapley forms *similarity cohorts* from subjects similar to the target for each feature, with 2^d sets of features on which subjects can be similar to the target for d features, where

$$z_{tj}\left(x_{ij}\right) = \begin{cases} 1, & |x_{ij} - x_{tj}| \leq \delta_t j \\ 0, & \text{else} \end{cases} \qquad (2.3)$$

represents the similarity between subjects i and t for feature j. For a set of features u on which the cohort members are similar and similarity cohort

$$C_{t,u} = \{i | z_{tj}\left(x_{ij}\right) = 1, \forall j \in u\}, \tag{2.4}$$

Cohort Shapley computes the average output of each cohort as

$$\bar{y}_{t,u} = \frac{1}{|C_{t,u}|} \sum_{i \in C_{t,u}} y_i. \tag{2.5}$$

Then, Cohort Shapley applies SHAP to the cohort averages:

$$\phi_j^{CS} = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \left(\bar{y}_{t,u+j} - \bar{y}_{t,u}\right) \tag{2.6}$$

This variant of SHAP is particularly useful for black-box models, as it only requires a single prediction for each subject. In this work, we use the Tree SHAP algorithm, which efficiently computes exact SHAP values for tree-based models (Lundberg et al., 2020). The background data set is the training data used to build the trees within the model. Tree SHAP leverages the number of training examples that traverse each leaf node to represent the background distribution, eliminating the need for explicit sampling. We do not apply any approximation settings; all SHAP values are computed exactly.

Finally, LIME derives explanations by learning an interpretable model, such as a linear model or decision tree, locally around a prediction. To do so, the algorithm (1) generates a set of samples by perturbing the input data around the query point $\mathbf{x}^*$, (2) predicts labels for each sample with the original model f, and (3) learns a simple, interpretable model on the weighted samples and their predicted labels. The weighting of each sample $\mathbf{z}$ when training the local surrogate model is governed by a proximity measure $\pi_{\mathbf{x}}$, defined by a kernel that assigns higher weights to points closer to $\mathbf{x}^*$, and the kernel width σ that defines the size of the local region. In this work, we use the default exponential kernel

$$\pi_{\mathbf{x}}(\mathbf{z}) = \exp\left(\frac{-D(\mathbf{x}, \mathbf{z})^2}{\sigma^2}\right) \tag{2.7}$$

where D is Euclidean distance and $\sigma = 0.75\sqrt{m}$, where m is the number of features in the data set. For each query point, the neighborhood contains 5,000 sample points. We deliberately choose not to standardize features to preserve the interpretability and additive nature of the explanations in dollar amounts. Standardization would obscure the economically meaningful interpretation of attributions. This design choice prioritizes substantive interpretability over purely statistical considerations. Additionally, maintaining features in their original dollar scale allows us to directly evaluate explanation accuracy using metrics like root mean squared error (RMSE), which require dollar-denominated predictions to assess revenue recovery and audit selection performance.

3. Issue Selection

Table 1. Summary of key notation.

Notation	Definition
$i \in 1, \cdots, N$	Index of taxpayers
$\mathbf{x}_i$	Features for taxpayer i
$j \in 1, \cdots, J$	Index of line items; $j = \Sigma$ is the net combination of $j = 1, \cdots, J$
$\ell_i^{(j)}$	True value of line item j
$\tilde{\ell}_i^{(j)}$	Reported value of line item j
$c_i^{(j)}$	Cost to audit line item j
$\delta_i^{(j)}$	Misreporting amount for line item j
$\mathbf{z}_i$	Vector of misreporting amounts for all line items $j = 1, \cdots, J$
τ	Threshold for minimum misreporting amount
$\pi_i^{(j)}$	Probability whether $\delta_i^{(j)}$ exceeds τ, given X_i
a_i	Indicator whether taxpayer i is audited
$l_i^{(j)}$	Indicator whether line item j is audited
k	Budget for number of line items to examine
K	Budget for percentage of population to audit

We now formalize the problem of issue selection for income tax audits, with notation summarized in Table 1. Issue selection is a key component of the process through which the IRS allocates enforcement resources. For many of its audit selection work streams, the IRS employs algorithmic models to predict tax underreporting, the focus of our work here, along with simple rules to flag potentially noncompliant returns.

Consider a set of taxpayers indexed by $i \in 1, \cdots, N$, each with features $\mathbf{x}_i$. We focus on a subset of issues where multiple line items contribute to a top-line total, and only this total matters for tax liability. Examples include the components of self-employment income or itemized deductions. Stated formally, for every line item indexed by $j \in 1, \cdots, J$, with $j = \Sigma$ representing the net aggregation (e.g., net profit or loss) of the constituent $1, \cdots, J$-th items (e.g., itemized profit and expenses), a taxpayer reports a value $\tilde{\ell}_i^{(j)}$, which may be different than the true value of that line item $\ell_i^{(j)}$:

$$\tilde{\ell}_i^{(\Sigma)} = \sum_{j=1}^{J} \tilde{\ell}_i^{(j)} \qquad (3.1)$$

$$\ell_i^{(\Sigma)} = \sum_{j=1}^{J} \ell_i^{(j)} \qquad (3.2)$$

This formulation fits naturally with our notion of local explanations, where a predictor of $\ell_i^{(\Sigma)}$ can be explained by additive feature attributions that corresponded to $\ell_i^{(j)}$ for $j = 1, \cdots, J$.

The set of reported values $\{\tilde{\ell}_i^{(\Sigma)}, \tilde{\ell}_i^{(1)} \cdots, \tilde{\ell}_i^{(J)}\}$ comprises a subset of features in $\mathbf{x}_i$. Every line item is associated with some auditing cost $c_i^{(j)}$, which we assume here to be constant. Let $\delta_i^{(j)} = d^{(j)} \cdot (\ell_i^{(j)} - \tilde{\ell}_i^{(j)})$ be the misreporting amount, where $d^{(j)}$ describes the direction of misreporting, with a value of +1 for a line item corresponding to a source of income or profit and -1 for a line item corresponding to an expense or deduction. The indicator $\mathbf{1}[\delta_i^{(j)} > \tau]$ denotes whether the misreporting amount exceeds some threshold τ, with $\pi_i^{(j)} = \Pr[\delta_i^{(j)} \geq \tau | \mathbf{x}_i]$. Positive values of $\delta_i^{(j)}$ generate revenue if discovered during an audit, which we refer to as *positive adjustments*. To denote predicted values, we use $\hat{\cdot}$ such as $\hat{\delta}_i^{(j)}$ for the predicted misreporting amount for line item j.

Issue selection is concerned with selecting a set of line items to audit. Let a_i be an indicator for whether taxpayer i is audited, and let $l_i^{(j)}$ be an indicator for whether taxpayer i's j-th line item is audited. Depending on the audit selection process, the determination of a_i may or may not be independent of $l_i^{(j)}$. If a_i is independent of $l_i^{(j)}$, then the issue selection process applies to the set individuals $\{i : a_i = 1\}$ who have already been selected by some other mechanism. On the other hand, a_i may instead be informed by the issue selection process, where

$$a^* = \arg\max_a \sum_{i=1}^{n} \sum_{j=1}^{J} a_i \cdot l_i^{(j)} \cdot \delta_i^{(j)}. \qquad (3.3)$$

Here, we assume the latter, namely, that return selection a_i is determined by taxpayers' k line items with the greatest misreporting.

Audits are done under constrained resources, operationalized by limiting both the total audit volume, as well as narrowing the line items on an individual's return that an auditor can examine. We use k to represent a fixed number of line items on a return that can be audited, and K to represent a fixed percentage of the population that can be audited. Under this fixed budget, various objectives may align with different auditing priorities, such as maximizing the amount of detected underreporting, maximizing return-on-investment, minimizing the proportion of returns audited without a true positive adjustment, called the *no-change rate*, or maximizing deterrence. In any of these cases, we seek to find some optimal allocation of l and a to select which line items of which returns to audit.

4. Data

Through collaboration with the IRS, our data set consists of a stratified random sample of taxpayer returns collected from the IRS's National Research Program (NRP) audits. In contrast to standard operational audits, which typically consider only a small number of issues where noncompliance is suspected, NRP audits provide a representative sample that aims to determine the net tax misreporting on the overall return based on a nearly line-by-line audit approach that generates labeled adjustment amounts for most lines on the return (not just the selective subset for operational audits). We treat these NRP-determined line-level adjustments as ground truth, abstracting from interexaminer variation and potential gaps between true misreporting and what auditors uncover.[2] Therefore, NRP data provides us with both the set of reported values $\{\tilde{\ell}_i^{(\Sigma)}, \tilde{\ell}_i^{(1)} \cdots, \tilde{\ell}_i^{(J)}\}$, as well as the complete set of true values $\{\ell_i^{(\Sigma)}, \ell_i^{(1)}, \cdots, \ell_i^{(J)}\}$.

Our data set contains n = 15,655 NRP audits of the Schedule C (Form 1040), Profit or Loss from Business (Sole Proprietorship) from the tax period from April 2012 to December 2015. For each audited return, we focus on $J = 34$ line item expenses, in addition to the aggregate misreporting at index $j = \Sigma$. Each line item expense is fungible, such that only $\delta_i^{(\Sigma)}$ matters for tax liability. We note that IRS guidelines do not necessarily require that NRP audits of the Schedule C investigate every line item. While Schedule C income items are generally classified for mandatory examination, expense items are examined under the "large, unusual or questionable (LUQ) criteria," which takes into account the character and the comparative and absolute size of the line item (Internal Revenue Service, 2023b). Despite this potentially incomplete coverage of line items, NRP audits are still more comprehensive than operational audits, and we emphasize two points: first, there is likely meaningful overlap with the types of issues operational audits would also target; and second, empirical evidence of the high correction rates among line items with positive reported values indicates that NRP auditors do in fact examine a substantial number of nonzero expense items in practice, even if not strictly required by guidelines. All analyses account for NRP sampling weights.

The distribution of misreporting detected through audit is positively skewed, with a median and modal misreporting amount among all line items of $0 and a mean misreporting amount around $200. The aggregate

misreporting is higher, with a median misreporting amount of approximately $2,500 and a mean near $6,500. Though these represent typical values, some misreporting amounts can reach into the millions of dollars. Given this distribution, noncompliance on individual line items is a relatively rare occurrence, posing well-known modeling challenges associated with imbalanced or long-tail learning scenarios (Feng et al., 2023; King & Zeng, 2001). For instance, as shown by the red bars in Figure 1, 28 of the 34 line items have positive adjustments in fewer than 10% of taxpayers, with an average misreporting rate of 6.1% across all line items. In contrast, aggregate misreporting exhibits considerably less sparsity compared to individual line items, with a misreporting rate of 63.6%. However, in either scenario, both line item misreporting amounts and aggregate misreporting amounts exhibit a long right tail, where a few high-value outlier misreporting amounts can significantly impact the fit of predictive models. Despite their rarity, this high misreporting is crucial, as it generates substantial revenue and is therefore important for the model to accurately capture. We observe that these challenges can be partially mitigated by only considering the set of line items for which a taxpayer has reported a positive value $\tilde{\ell}_i^{(j)} > 0$. The grey bars in Figure 1 shows the proportion of positive adjustment values conditional on $\tilde{\ell}_i^{(j)} > 0$ for $j = \Sigma, 1, \cdots, J$. On average across line items, 40.8% of taxpayers with a positive reported value also have a positive adjustment value for that line item.

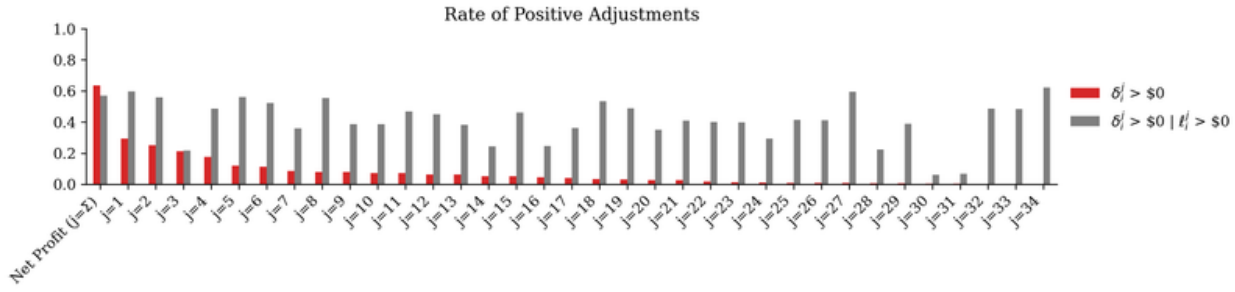


Figure 1. Positive adjustment rates for the aggregate misreporting $j = \Sigma$ and each line item $j = 1, \cdots, 34$. The red bars illustrate that positive adjustments for each line item are uncommon events, with an average misreporting rate of only 6.1%. In contrast, misreporting of the aggregate amount is more common, with a rate of 63.6%. When conditioning on the presence of a positive reported value ($\tilde{\ell}_i^{(j)} > 0$), shown in grey, the rate of positive adjustments increases significantly, averaging 40.8% across the 34 line items. This highlights that while individual line item misreporting is rare, its frequency is higher when only considering taxpayers who report nonzero values.

5. Models

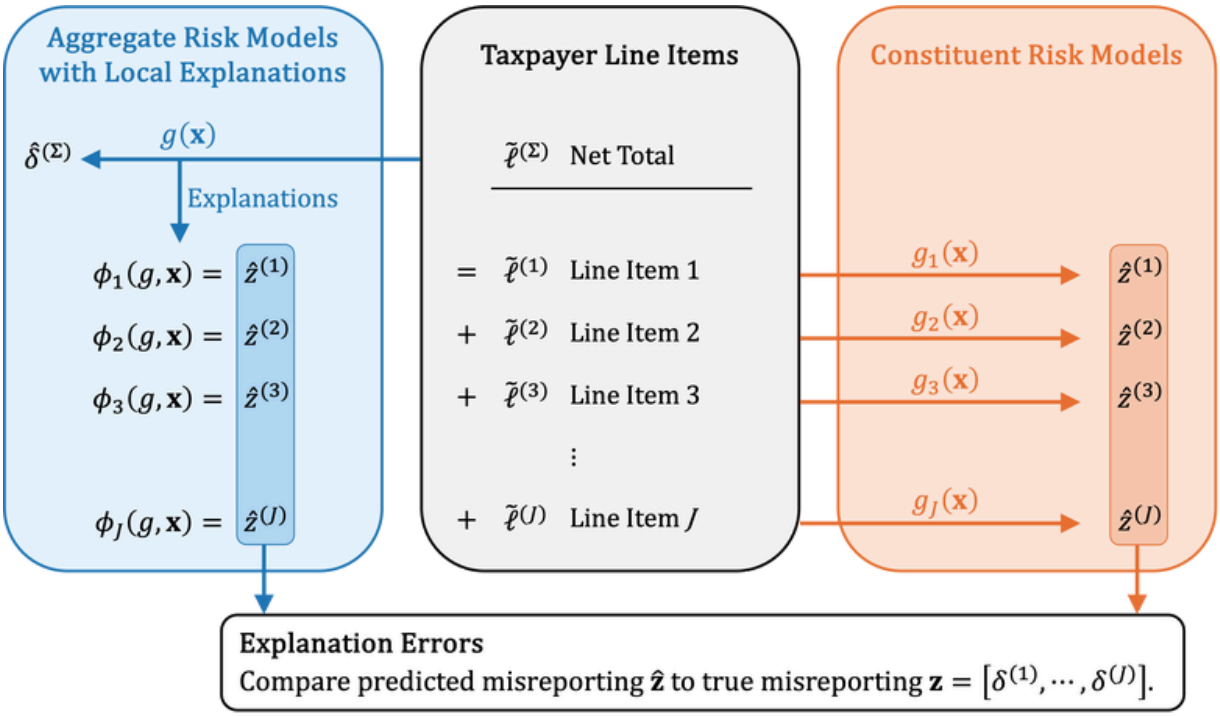


Figure 2. Aggregate and constituent risk modeling approaches. While the aggregate technique first estimates the total misreporting amount $\hat{\delta}_i^{(\Sigma)}$ and using local explanations to derive the constituent misreporting $\hat{\mathbf{z}}$, the constituent technique models each constituent amount directly.

We consider several different techniques to estimate a taxpayer's constituent risk values, with each approach representing a different trade-off with respect to modeling or explanation decisions. At a high level, we consider two different approaches: *aggregate* and *constituent*. We use the term aggregate to describe any top-down technique that first estimates the total misreporting amount $\hat{\delta}_i^{(\Sigma)}$ with an underlying aggregate risk model, and from this derives estimates for the constituent misreporting amounts $\hat{\delta}_i^{(j)}, j = 1, \cdots, J$ with a local explanation model. Constituent risk approaches, on the other hand, describe bottom-up techniques that estimate the constituent misreporting directly, without considering $\hat{\delta}_i^{(\Sigma)}$. We summarize the aggregate and constituent approaches in Figure 2.

While it might seem intuitive to directly model the constituent risks in every case given that these are the primary quantities of interest, there are several reasons to consider an aggregate approach. First, training several constituent risk models requires ground truth misreporting data for every constituent item, which is often not available, as it can be resource-intensive to collect systematically. In Fair Credit Reporting, for instance, human assessments do provide a decomposition of factors that explain a credit denial decision.[3] Our access to the constituent data is precisely what makes our setting unique for evaluating explainability (Gilpin et al., 2018; Leavitt & Morcos, 2020; Nauta et al., 2023; Yang et al., 2019). Second, individual line item

misreporting is much rarer than aggregate misreporting, leading to significant imbalanced and long-tail modeling challenges. These challenges are not unique to our tax audit setting; they can also arise in other contexts where only the overall risk is known, but not the constituent risks, or where individual risk factors are rare on their own. Third, even when constituent item information is available in other settings, it is often only available for a more limited sample or can be of lower reliability than the overall assessment (e.g., Braithwaite & Braithwaite, 1995; Cockburn et al., 2003; Feinstein, 1989; Office of the Inspector General, U.S. Nuclear Regulatory Commission, 1995; U.S. Government Accountability Office, 2005).

In either approach, the objective is to learn some mapping $f : \mathbb{R}^d \to \mathbb{R}^J$ that maps from a d-dimensional feature vector $\mathbf{x}_i$ to a J-dimensional output vector $\hat{\mathbf{z}}_i$, where each element represents some measure of the risk of noncompliance for line item j. The goal is to recover the ground truth misreporting vector

$$\mathbf{z}_i = \left[\delta_i^{(1)}, \cdots, \delta_i^{(J)}\right] \qquad (5.1)$$

where

$$\delta_i^{(\Sigma)} = \sum_{j=1}^{J} z_i^{(j)}. \qquad (5.2)$$

The k elements of $\hat{\mathbf{z}}_i$ with the largest predicted values are selected for audit, with return selection based on the total misreporting among these k line items; this prioritizes positively valued, revenue-increasing predictions. We denote the indices of this set of k line items with $\texttt{top}_k(\cdot)$. For each approach, we can also define an *oracle* model, which assumes that the total misreporting is known exactly (*i.e.*, $\hat{\delta}_i^{(\Sigma)} = \delta_i^{(\Sigma)}$) in the aggregate risk setting, or that the constituent misreporting amounts are known exactly (*i.e.*, $\hat{\mathbf{z}}_i = \mathbf{z}_i$) in the constituent risk setting.

For all approaches, we perform five-fold cross-validation, splitting the data into five folds and holding each fold out as a separate test set. For machine learning models, we perform a hyperparameter grid search on the remaining four folds with an inner five-fold cross-validation, and train a model with the optimal hyperparameters on the four folds. With this cross-validation procedure, each of the five models may have different hyperparameters and thus be fundamentally different models, which can help to understand the stability of model fit and downstream impacts on the evaluation metrics for the data set. For further details of the training procedures, see Appendix A.

5.1. Aggregate Risk Models With Local Explanations

We begin by training underlying aggregate machine learning models and applying local explanations to answer the question, *to what extent do machine learning models with local explanation models recover the true constituent risks?* To generate f in this setting, we first train an underlying model $g : \mathbb{R}^d \to \mathbb{R}$ to predict

some measure of the risk of aggregate noncompliance $\hat{\delta}_i^{(\Sigma)}$. For instance, in the regression setting, we predict $\hat{\delta}_i^{(\Sigma)}$ with an XGBoost regressor, a gradient-boosted decision tree algorithm. To focus on revenue-generating misreporting, we define the response for each line item to be $y^{(j)} = \max\{0, \delta^{(\Sigma)}\}$. The regression models are trained with mean squared error loss. To derive line item attributions, we use SHAP, Cohort Shapley, or LIME to generate the vector

$$\hat{\mathbf{z}}_i = [\phi_1(g, \mathbf{x}_i), \cdots, \phi_J(g, \mathbf{x}_i)]\,. \qquad (5.3)$$

Given the imbalanced, long-tailed nature of the data, it is reasonable to expect the underlying machine learning model to fit the data poorly. Therefore, it is possible that the explanation vector $\hat{\mathbf{z}}_i$ may not closely match the true misreporting $\mathbf{z}_i$ due to the underlying modeling limitations. To overcome this challenge and understand the *best possible performance of local explanation models*, we construct the aggregate oracle model. The oracle model assumes $g_{\text{oracle}}(\mathbf{x}_i) = \delta_i^{(\Sigma)}$ to map exactly to the total misreporting amount. While SHAP requires full access to the model and LIME requires a predictor to generate labels for the local neighborhood, Cohort Shapley can apply to black-box models such as the oracle. It does so by using only a single prediction per subject while leveraging features to group similar subjects into cohorts. In the case of the aggregate oracle, we assume perfect recovery of the underlying aggregate misreporting. Since the oracle itself lacks features, the cohorts are formed based on observed taxpayer characteristics, not model-internal representations.

5.2. Constituent Risk Models

Because the constituent risk values are the primary interest, an alternative to local explanation models is estimating the line item misreporting directly. Constituent risk models can answer the question, *how do local explanation models compare to estimating the constituent risks directly?* To generate f for the the constituent risk models, we train J models $g_j : \mathbb{R}^d \to \mathbb{R}$, where the final output is given by

$$\hat{\mathbf{z}}_i = [g_1(\mathbf{x}_i), \cdots, g_J(\mathbf{x}_i)]\,. \qquad (5.4)$$

In the regression setting, we predict $\hat{\delta}_i^{(j)}$ with an XGBoost regressor for $j = 1, \cdots, J$. We can again use an oracle model to represent the *best possible performance of constituent risk methods*. The constituent risk oracle model assumes $g_{j_{\text{oracle}}}(\mathbf{x}_i) = \delta_i^{(j)}$, which recovers $\hat{\mathbf{z}}_i = \mathbf{z}_i$ across all line items. This oracle model provides the optimal allocation of audit resources.

5.3. Heuristic Models

Rather than relying on complex prediction models, we can use simple, interpretable decision rules to guide decision-making. This approach is consistent with recommendations, such as those by Rudin (2019), which advocate for the use of inherently understandable techniques in decision-making processes. Rules-based approaches follow the older traditions of techniques like business rules or business logic, which formalize

decision logic in a structured and easily understandable manner (Morgan, 2002; von Halle, 2001), and logical AI, which uses formal logic to represent knowledge and infer decisions in a way that is transparent and explainable (Domingos et al., 2016; McCarthy, 1959).

We explore two types of heuristics: *global* and *local*. Global heuristics apply the same rule to all taxpayers, resulting in a uniform set of selected issues. Here, we select the k line items with the highest rate of misreporting (defined by the proportion of taxpayers with a positive adjustment for a given line item) among all taxpayers. In contrast, a local heuristic applies a rule that is consistent across taxpayers but may lead to different issues being selected for each individual taxpayer. For each taxpayer, we select the k line items with the greatest reported value $\tilde{\ell}_i^{(j)}$.

6. Evaluation

When evaluating the different modeling approaches, there are two primary questions we would like to answer: (i) *how closely each constituent risk estimation* $\hat{\mathbf{z}}_i$ *matches to the ground truth* $\mathbf{z}_i$ (Section 6.1); and (ii) *how useful each* $\hat{\mathbf{z}}_i$ *is operationally for issue selection, as a determinant of return selection* (Section 6.2). We describe each metric here.

6.1. Validity of Constituent Risk Estimations

The first question we are interested in is, *to what extent do our estimations of constituent risk recover the true values?* In machine learning, the *fidelity* of explanations refers to how accurately local explanations match an underlying decision process (Balagopalan et al., 2022). Because ground truth explanations are often lacking, fidelity measures the prediction error of the explanations, comparing the aggregate prediction of an explanation model Φ to that of the underlying machine learning model g, for some metric L:

$$L\left(g\left(\mathbf{x}_i\right), \Phi\left(\mathbf{x}_i\right)\right).$$

We introduce an analogous concept called *explanation error* to describe the fidelity of the values predicted by the explanation model $\hat{\mathbf{z}}_i$ to the *true* underlying explanations $\mathbf{z}_i$. Distinct from prediction error, explanation error provides an important conceptual contribution that emphasizes the need to benchmark explanations against a concrete notion of ground truth. For our measure of explanation error, we use RMSE

$$\sqrt{\frac{\sum_{i=1}^{N}\left(z_i^{(j)} - \hat{z}_i^{(j)}\right)^2}{N}} \qquad (6.1)$$

to compare the estimated risk $\hat{z}_i^{(j)}$ to the true risk $z_i^{(j)}$ for a given line item j.

6.2. Operational Utility for Issue Selection

While RMSE provides an understanding of the accuracy of the constituent risk estimations, we are also interested in *how useful the constituent risk estimations are in practice for auditors*, where they must identify a subset of issues to examine. The operational utility considers the selection of line items to audit under an auditing budget K to simulate the process of issue selection in practice, for some fixed k that determines how many line items an auditor can examine. Recovery@k measures the revenue recovered (before deducting auditing costs) and is defined as follows. Given

$$\hat{l}_i^{(j)} = \begin{cases} 1, & j \in \texttt{top}_k(\hat{\mathbf{z}}_i) \\ 0, & \text{else} \end{cases} \tag{6.2}$$

then we can select taxpayers to audit with

$$\hat{a} = \arg\max_a \sum_{i=1}^{N}\sum_{j=1}^{J} a_i \cdot \hat{l}_i^{(j)} \cdot \hat{z}_i^{(j)} \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^{N} \hat{a}_i < K. \tag{6.3}$$

Then, we can generate recovery@k as the actual revenue recovered from $\hat{a}$ and $\hat{l}$ as

$$\text{recovery@}k = \sum_{i=1}^{N}\sum_{j=1}^{J} \hat{a}_i \cdot \hat{l}_i^{(j)} \cdot \delta_i^{(j)}. \tag{6.4}$$

While we focus on recovery@k for its operational utility, we include two additional metrics, the no-change rate@k and r-precision@k, in Appendix B. The no-change rate@k describes the fraction of audits with misreporting less than a de minimus amount $\tau = \$100$. R-precision@$k$ measures how well the model identifies the true high-risk line items among the top-ranked predictions, which functions as a precision-recall-type metric tailored to our ranking task.

7. Results

In this section, we address two primary questions: (i) *how well does each constituent risk estimation approach recover the true risk* (Section 7.1), and (ii) *how useful are each of these estimations operationally for issue selection under a fixed budget* (Section 7.2)? We compare the aggregate risk models with local explanations to the constituent risk models and simple heuristics. We show that local explanation models fail to recover the true misreporting amounts accurately, whereas the constituent models can marginally better estimate true misreporting to provide better guidance for issue selection within budget constraints. Here, we focus on five regression approaches—(i) the aggregate regression model explained by SHAP, (ii) the aggregate regression model explained by Cohort Shapley, (iii) the aggregate oracle model explained by Cohort Shapley, and (iv) the constituent regression models. The observed trends and implications are consistent across our other modeling

approaches. The results for LIME and the additional approaches can be found in Appendix C. We also include results for the additional metrics, no-change rate@k and r-precision@k, in Appendix C.

7.1. Constituent Risk Estimation

We begin by assessing the explanation errors of the risk estimations, comparing the predicted line item misreporting amounts $\hat{\mathbf{z}}_i$ to the true values $\mathbf{z}_i$.

Explanation errors are high across all models, compared to the ground truth constituent misreporting. The left panel of Figure 3 compares the RMSE across the five regression approaches across all line items, as well as for the five expense line items with the highest misreporting rates. Across the five approaches, the average RMSE is $3,118 (SE = $166), driven by incorrect predictions for taxpayers with the highest misreporting (see Appendix C for details). This is an expected challenge in modeling long-tailed data. Yet even the Cohort Shapley explanations with the perfect underlying oracle model have a comparable RMSE: on average across folds, Cohort Shapley with the oracle regression model has a RMSE of $3,211 (SE = $434), compared to $2,996 (SE = $353) for the same explanation approach with the aggregate regression model.

Local explanation models tend to have higher explanation errors than the constituent models. On the right side of Figure 3, we directly compare the aggregate approaches with local explanations to the constituent risk models. We show the percent change in RMSE when switching from the constituent risk model to local explanation models, with comparisons made for both the average across all line items and the five line items with the highest misreporting rates. Explanation-based approaches have on average 6.6% higher RMSE than the constituent risk model.

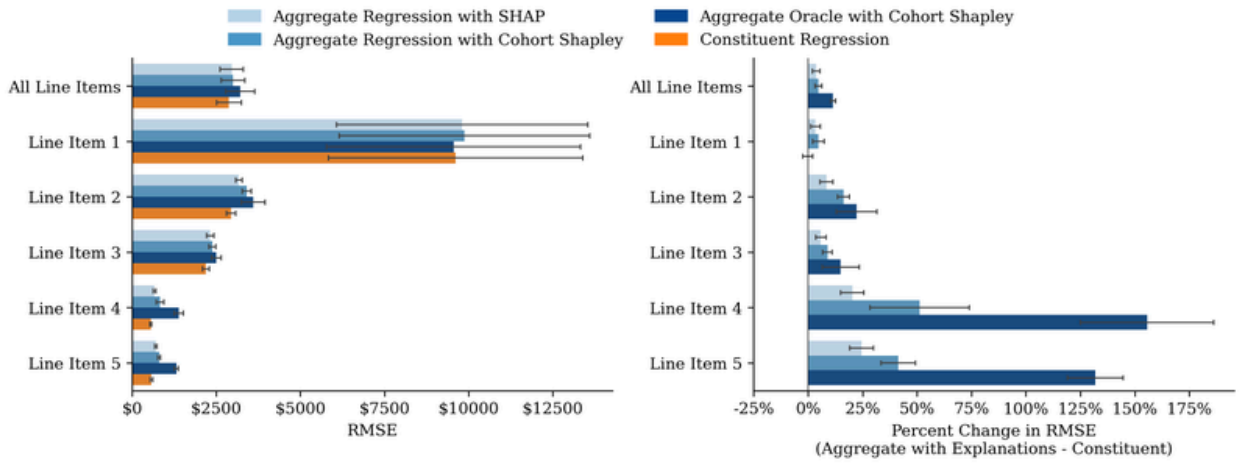


Figure 3. Comparison of predicted versus true line item misreporting for the five regression approaches. The left side shows the mean root mean squared error (RMSE) and standard error across five data folds, both overall and for the five expense line items with the highest misreporting rates. RMSE is relatively high across all models, driven by explanation errors in the upper quantiles. The right side compares RMSE percent change between the constituent regression and local explanation models. Explanation-based approaches have, on average, 6.6% higher RMSE than the constituent risk model. SHAP = SHapley Additive explanation.

7.2. Operational Utility

While RMSE provides insights into the overall performance of each approach, it does not answer the question of how useful the risk estimations are operationally, in a real use case where auditors use these estimations to judiciously select only the highest risk taxpayers and line items to investigate. To understand the operational utility, we turn next to recovery@k, with an auditing budget of $K \leq 3.0\%$ and a line item budget of $k = 5$ for each return.

Constituent regression models tend to perform better than local explanations models. Figure 4 shows the cumulative revenue as the audit budget increases from K = 0.0% to 3.0%, with comparisons between the local explanations models and the constituent risk model. The leftmost plot shows the cumulative revenue for the oracle model and the five regression approaches; the right three figures show the mean difference and standard error across the five folds between the constituent regression models and each of the explanation-based approaches. While the constituent regression models do not accumulate as much revenue as the constituent oracle, the cumulative difference shows that the constituent regression models consistently generate higher revenue than any of the local explanations from an actual underlying aggregate model.

Explanations improve when the underlying model improves. When deriving local explanations from the underlying oracle aggregate misreporting amounts, rather than an actual underlying aggregate prediction model, Cohort Shapley explanations outperform any other explanation technique for all values of K. This

illustrates that local explanations are only as good as the underlying model: when we move to the perfect oracle model, the cumulative revenue recovered increases for the same local explanation technique.

Pure heuristics tend to be inferior to modeling approaches. Table 2 shows the revenue recovered when exactly 3.0% of taxpayers are audited. The local heuristic, personalized to each taxpayer, achieves revenue recovery similar to the local explanation models. In contrast, the global heuristics apply the same rule to all taxpayers, selecting the same issues across all individuals. The global heuristic performs worse than all local explanation-based approaches. This indicates that overly simplistic rules can fail to capture the complexity of individual risk, underscoring the limitations of broad, generalized heuristics.

Table 2. Recovery@k for the oracle, constituent regression, aggregate regression, and heuristic risk estimation models, at $k = 5$ and $K = 3.0\%$. Values show mean and standard error across the five folds. SHAP = SHapley Additive explanation; CS = Cohort Shapley.

Model	Recovery@5
Constituent Oracle	$17,841,294 ($770,405)
Constituent Regression	$10,517,700 ($1,246,249)
Aggregate Regression with SHAP	$8,975,128 ($973,356)
Aggregate Regression with CS	$8,882,328 ($877,816)
Aggregate Oracle with CS	**$10,772,925 ($737,071)**
Global Heuristic	$5,246,358 ($518,340)
Local Heuristic	$9,137,485 ($955,828)

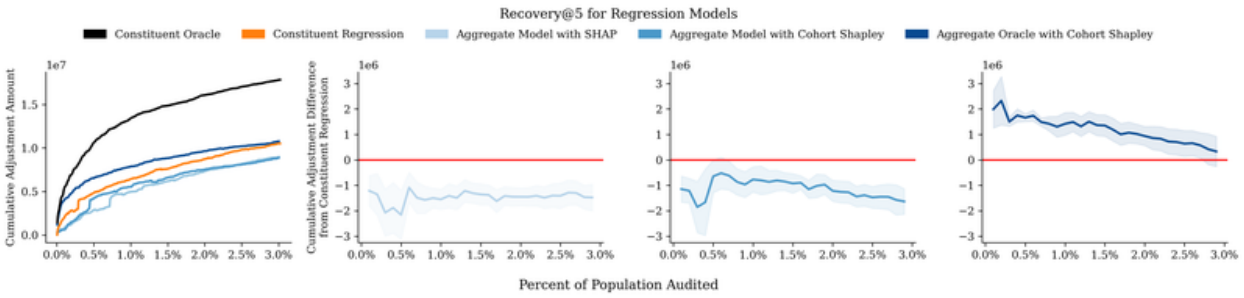


Figure 4. Recovery@k as the audit budget increases from 0.0% to 3.0%, with comparison to the constituent regression model. The left figure shows cumulative revenue for the oracle, constituent regression, local explanation, and heuristic risk estimation models, while the right figures show mean differences and standard errors between the constituent and local explanation models. SHAP = SHapley Additive explanation.

8. Discussion

In collaboration with the IRS, with access to real audit data, we have provided an empirical evaluation of the most commonly used local explanation models in a high-stakes setting for risk assessment. In this unique setting, we know precisely the true values that explanations should capture, providing a concrete definition of ground truth often missing in prior assessments of explainability tools. Our findings suggest that local explanation models depend on the underlying model quality: if the model struggles to capture rare or long-tail events, the explanations are likely to fail as well. While explanations generated by an oracle model align more closely with the true risk compared to those from actual models, the improvements, surprisingly, remain marginal even with the perfect oracle model. This highlights a critical limitation: explanations may fail to bridge the gap between a model's predictions and their interpretability, even when the model itself performs exceptionally well. This shows that the explanation methods themselves have fundamental limitations in our setting, independent of the quality of the underlying predictive model. Moreover, we highlight a subtlety in the dynamics between models and their explanations: significantly better models do not inherently lead to significantly better explanations.

An alternative evaluation strategy is to directly compare aggregate model predictions with the sum of constituent model predictions. We do not pursue this direction, however, as it sidesteps the central question of whether explanations can guide operational decisions. In practice, explanations must drive both return selection and issue selection. In this operationally relevant setting, where we select the top-k line items based on these explanations and then use the sum of these k line items to rank returns for audit selection, the quality of explanations directly affects which returns are selected, making accurate decomposition essential.

Given that local explanation models fail to recover the true risk well, we have also compared explanation techniques to alternative risk models: constituent risk models that predict each line item's risk separately, and heuristics crafted from simple decision rules. We show that directly estimating each line item misreporting

amount individually with separate machine learning models produces results comparable to or marginally better than those obtained through explanations. Local heuristics may also be comparable to explanations, even outperforming the aggregate risk model with local explanations. There are several factors at play. First, some constituent line item models perform quite well. However, other line item models, particularly for those with the least frequent misreporting, can be quite noisy. Second, and relatedly, for some line items, the constituent models may be so sparse that they essentially replicate the heuristic behavior (selecting the highest reported values), providing little additional information beyond the simple rule. The local heuristic's strong performance suggests that in some domains, especially those with sparse, imbalanced data like ours, simple rules based on reported values may capture much of the signal that complex models learn, at least for the purpose of ranking high-risk items. However, coarser global heuristics are the least faithful to the true risk. Thus, the design choice of an alternative, more interpretable approach matters; in particular, some rules-based approaches may be too simple to capture the nuanced complexity of individual risk. This illustrates the real-world trade-offs of using inherently interpretable models rather than relying on local explanation models to make a complex, black-box AI algorithm understandable.

We note several limitations and areas for future work. First, we consider only a limited set of modeling and explanation techniques. Though SHAP and LIME are among the most commonly used methods, the range of local explainability tools is much broader. In particular, prior work has provided empirical evidence that SHAP and LIME have less stability under class imbalance (Chen et al., 2024); thus, long-tailed distributions, characterized by infrequent events, may amplify this issue. We further deliberately did not standardize features to preserve the original dollar scale, but this may contribute to LIME's weaker performance, particularly given the presence of dollar-valued features spanning different scales. Future work should explore alternative explanation methods and their properties under class imbalance and long-tail distributions. Second, in this work, we consider each line item risk independently. However, misreporting may be correlated and may often arise together in clusters. Neither the aggregate nor constituent risk models account for these relationships directly. Such an approach overlooks important dependencies between misreporting that could influence both the predictions and their explanations. Capturing these correlations could provide a more holistic understanding of risk and lead to improved model performance and interpretability. Future research should investigate methods that explicitly model these relationships, such as multitask or sequential approaches, to better account for the interconnected nature of risk factors. Finally, while our findings offer valuable insights into an example real-world risk assessment setting, the results may not generalize well to other data sets that do not contain similar imbalanced, long-tailed, and sparse properties as ours.

Despite these limitations, this work offers an important perspective on the limitations of explainability techniques and the trade-offs with alternative modeling approaches. We provide these insights at a time when regulatory concerns about the transparency and trustworthiness of AI models in high-risk decision-making are

intensifying. Our findings emphasize the importance of carefully evaluating the efficacy of explanation methods in high-stakes environments, where the impact of modeling and explanation errors can be substantial.

Disclosure Statement

The views expressed in this article are those of the authors and do not necessarily reflect the views of the Internal Revenue Service. Any taxpayer data used in this research was kept in a secured IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed. I.O.G. is supported by the Fannie & John Hertz Foundation and Stanford Knight-Hennessy Scholars graduate fellowship.

References

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a transparent evaluation of post hoc model explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 15784–15799). Curran Associates.

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. *Proceedings of Machine Learning Research*, *97*, 161–170.

Arras, L., Osman, A., & Samek, W. (2022). CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, *81*, 14–40. https://doi.org/10.1016/j.inffus.2021.11.008

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1194–1206). ACM. https://doi.org/10.1145/3531146.3533179

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1479–1503). ACM. https://doi.org/10.1145/3531146.3533204

Braithwaite, J., & Braithwaite, V. (1995). The politics of legalism: Rules versus standards in nursing-home regulation. *Social & Legal Studies*, *4*(3), 307–341. https://doi.org/10.1177/096466399500400301

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Cockburn, I., Kortum, S., & Stern, S. (2003). Are all patent examiners equal? Examiners, patent characteristics, and litigation outcomes. In W. M. Cohen & S. A. Merrill (Eds.), *Patents in the knowledge-based economy* (pp. 19–53). The National Academies Press.

Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H., & Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 203–214). ACM. https://doi.org/10.1145/3514094.3534159

Domingos, P., Lowd, D., Kok, S., Nath, A., Poon, H., Richardson, M., & Singla, P. (2016). Unifying logical and statistical AI. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science* (pp. 1–11). ACM. https://doi.org/10.1145/2933575.2935321

Elzayn, H., Smith, E., Hertz, T., Guage, C., Ramesh, A., Fisher, R., Ho, D. E., & Goldin, J. (2024). Measuring and mitigating racial disparities in tax audits. *Quarterly Journal of Economics*, *140*(1), 113–163. https://doi.org/10.1093/qje/qjae027

Equal Credit Opportunity Act, Pub. L. No. 94-239, 90 Stat. 251 (1974). https://www.govinfo.gov/content/pkg/USCODE- 2011-title15/pdf/USCODE-2011-title15-chap41-subchapIV.pdf

Fair Credit Reporting Act , Pub. L. No. 91-508, 84 Stat. 1114 (1970). https://www.govinfo.gov/content/pkg/USCODE-2011-title15/pdf/USCODE-2011-title15-chap41-subchapIII.pdf

Feinstein, J. S. (1989). The safety regulation of US nuclear power plants: Violations, inspections, and abnormal occurrences. *Journal of Political Economy*, *97*(1), 115–154. https://doi.org/10.1086/261595

Feng, C., Li, L., & Xu, C. (2023). Advancements in predicting and modeling rare event outcomes for enhanced decision-making. *BMC Medical Research Methodology*, *23*(1), Article 243. https://doi.org/10.1186/s12874-023-02060-x

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In F. Bonchi, F. Provost, T. Eliassi-Rad, W. Wang, C. Cattito, & R. Ghani (Eds.), *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (pp. 80–89). IEEE. https://doi.org/10.1109/DSAA.2018.00018

Guidotti, R. (2021). Evaluating local explanation methods on ground truth. *Artificial Intelligence*, *291*, Article 103428. https://doi.org/10.1016/j.artint.2020.103428

Henderson, P., Chugg, B., Anderson, B., Altenburger, K., Turk, A., Guyton, J., Goldin, J., & Ho, D. E. (2023). Integrating reward maximization and population estimation: Sequential decision-making for Internal Revenue Service audit selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(4), 5087–5095. https://doi.org/10.1609/aaai.v37i4.25637

Holtzblatt, J., & Engler, A. (2022, June 22). *Machine learning and tax enforcement*. Tax Policy Center, Urban Institute & Brookings Institution. https://taxpolicycenter.org/sites/default/files/publication/163980/machine_learning_and_tax_enforcement.pdf

Internal Revenue Service. (2023a). *Inflation Reduction Act strategic operating plan, FY 2023-2031*. https://www.irs.gov/pub/irs-pdf/p3744.pdf

Internal Revenue Service. (2023b). *Internal Revenue manuals: Classification of NRP returns*. https://www.irs.gov/irm/part4/irm_04-022-003

Internal Revenue Service. (2023c). *IRS Newsroom: The Tax Gap*. https://www.irs.gov/statistics/irs-the-tax-gap

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 275–284). ACM. https://doi.org/10.1145/3097983.3098066

Leavitt, M. L., & Morcos, A. (2020). *Towards falsifiable interpretability research*. ArXiv. https://doi.org/10.48550/arXiv.2010.12016

Liu, Y., Khandagale, S., White, C., & Neiswanger, W. (2021). *Synthetic benchmarks for scientific research in explainable machine learning*. ArXiv. https://doi.org/10.48550/arXiv.2106.12543

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Curran Associates.

Mase, M., Owen, A. B., & Seiler, B. B. (2019). *Explaining black box decisions by Shapley cohort refinement*. ArXiv. https://doi.org/10.48550/arXiv.1911.00467

McCarthy, J. (1959). Programs with common sense. *Proceedings of the Symposium on the Mechanization of Thought Processes* (Vol. 10, pp. 75–92). National Physics Laboratory.

Miró-Nicolau, M., Jaume-i-Capó, A., & Moyà-Alcover, G. (2024). Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence*, *335*, Article 104179. https://doi.org/10.1016/j.artint.2024.104179

Morgan, T. (2002). *Business rules and information systems: Aligning IT with business goals*. Addison-Wesley Professional.

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1198–1212). ACM. https://doi.org/10.1145/3593013.3594074

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), Article 295. https://doi.org/10.1145/3583558

Office of the Inspector General, U.S. Nuclear Regulatory Commission. (1995). *Factors contributing to inconsistency in the operating reactor inspection program* (No. OIG-95A-04).

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1139–1150). ACM. https://doi.org/10.1145/3593013.3594069

Regulation 2024/1689. *Laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* European Parliament, Council of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Shamsabadi, A. S., Yaghini, M., Dullerud, N., Wyllie, S., Aıvodji, U., Alaagib, A., Gambs, S., & Papernot, N. (2022). Washing the unwashable: On the (im)possibility of fairwashing detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 14170–14182). Curran Associates.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton University Press.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). ACM. https://doi.org/10.1145/3375627.3375830

U.S. Government Accountability Office. (2005). *Nursing homes: Despite increased oversight, challenges remain in ensuring high-quality care and resident safety* (No. GAO-06-117). https://www.gao.gov/assets/gao-06-117.pdf

von Halle, B. (2001). *Business rules applied: Building better systems using the business rules approach*. Wiley Publishing.

Yang, F., Du, M., & Hu, X. (2019). *Evaluating explanation without ground truth in interpretable machine learning*. ArXiv. https://doi.org/10.48550/arXiv.1907.06831

Appendices

Appendix A. Training Procedure

As described in Section 4, the misreporting amounts (*i.e.*, our response variable) contain several high outliers that we want our models to capture well from a revenue-maximizing perspective, as these outliers correspond to high-valued positive adjustments. These high-valued misreporting values are fairly sparse, however, and thus any split of the data may contain different distributions in the upper quantiles; moreover, under our mean squared error training objective (which is more sensitive to outliers than mean absolute error or other alternatives), the outliers may heavily influence model fit.

We want our evaluation procedure to reflect any instability in the models that may arise due to the presence of these outliers, and thus we perform a nested five-fold cross-validation to train five distinct models, each with

validation sets that we can evaluate individually or compare to derive uncertainties in our evaluation metrics. To do so, we first randomly split the data into five groups. We hold out each of the five groups as a test set and take the remaining four groups as a training set. Then, among the four training groups, we perform five-fold cross-validation to select the hyperparameters with the best mean score on the validation folds. For the XGBoost models, we perform a grid search over:

- Learning rate: [0.001, 0.005, 0.01, 0.05]
- Number of estimators: [1,000, 1,500, 2,000]
- Maximum tree depth: [3, 5, 7]
- Subsample ratio of training instances: [0.5, 0.75, 1]
- Subsample ratio of columns: [0.75, 1]
- Gamma: [0, 1, 5, 10]

Appendix B. Additional Metrics

In addition to recovery@k, we introduce two additional metrics to measure operational utility: no-change rate@k and r-precision@k. Like recovery@k, these metrics consider the selection of line items to audit under an overall auditing budget K and some fixed k that determines how many line items an auditor can examine.

B.1. No-Change Rate@k

It is undesirable for both auditors and taxpayers to conduct an audit that results in no meaningful adjustment, as each audit can cost significant resources, time, and effort. The no-change rate describes the fraction of audits with misreporting less than τ, given by

$$\frac{\sum_i a_i \cdot \mathbf{1}\left[\sum_{j=1}^{J} \hat{l}_i^{(j)} \cdot \delta_i^{(j)} \leq \tau\right]}{\sum_i a_i}. \qquad \text{(B.1)}$$

We select a de minimis amount of $\tau = \$100$.

B.2. R-precision@k

We aim to maximize precision to avoid auditing issues that do not have a positive adjustment. Conversely, for taxpayers with very little misreporting, we aim to maximize recall to capture as many true misreported line items as possible. R-precision@k, adapted from information retrieval for ranked retrieval result evaluation, compares the number of selected line items in the top $\min\{k, r_i\}$ to the number of true positive adjustments in the top $\min\{k, r_i\}$. Each return has some value $r_i = |\{j : \delta_i^{(j)} > 0\}|$ that defines the number of line items with a true positive adjustment.

$$\text{r-precision@}k = \frac{\texttt{top}_{\min\{k,r_i\}}(\mathbf{z}_i) \cap \texttt{top}_{\min\{k,r_i\}}(\hat{\mathbf{z}}_i)}{\min\{k, r_i\}} \tag{B.2}$$

$$= \begin{cases} \text{recall@r} & r_i < k \\ \text{precision@k} & r_i > k \end{cases} \tag{B.3}$$

where

$$\text{recall@}k = \frac{\texttt{top}_k(\mathbf{z}_i) \cap \texttt{top}_k(\hat{\mathbf{z}}_i)}{r_i} \tag{B.4}$$

$$\text{precision@}k = \frac{\texttt{top}_k(\mathbf{z}_i) \cap \texttt{top}_k(\hat{\mathbf{z}}_i)}{k} \tag{B.5}$$

Appendix C. Extended Results

In these extended results, we consider the following machine learning approaches. The aggregate risk models with local explanations use the following for the underlying model to predict the aggregate misreporting at $j = \Sigma$, while the constituent risk approach trains separate models for $j = 1, \cdots, J$.

1. **Regression model.** We predict $\hat{\delta}_i^{(j)}$ with an XGBoost regressor. We define the response for each line item to be $y^{(j)} = \max\{0, \delta^{(j)}\}$. The regression models are trained with mean squared error loss.
2. **Classification model.** We predict $\hat{\pi}_i^{(j)}$ with an XGBoost classifier. We consider a de minimis amount of $\tau = \$100$, with $y^{(j)} = \mathbf{1}[\delta_i^{(j)} > \$100]$ as the response label with cross-entropy loss.
3. **Positive reported value conditional model.** Given that nearly all misreporting originates from a positive reported value (see Section 4), we train models on only the subset of training instances for which $\tilde{\ell}_i^{(j)} > 0$, and perform inference only on taxpayers with $\tilde{\ell}_i^{(j)} > 0$.

We look first more closely at constituent risk estimation. Explanation errors are driven by poor predictions for taxpayers with the highest misreporting. Figure C.1 shows the root mean squared error (RMSE) binned by the true misreporting amount for the aggregate regression model explained by SHAP (SHapley Additive exPlanation).

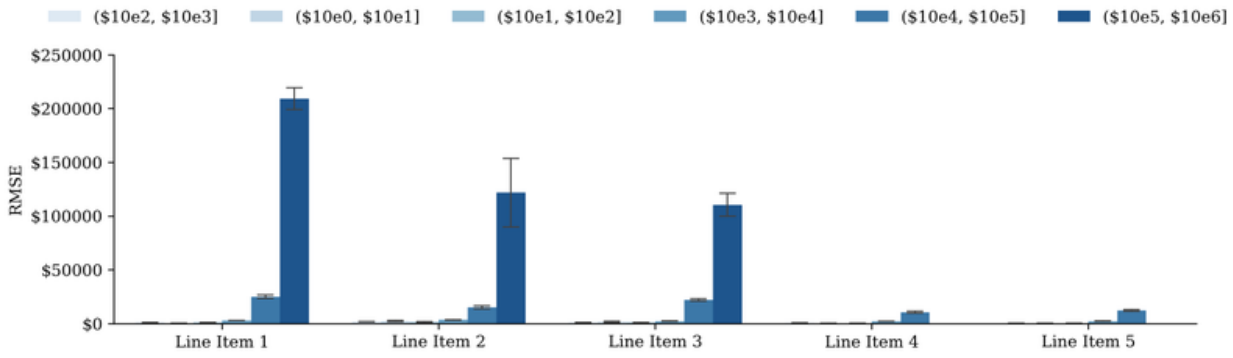


Figure C.1. Root mean squared error (RMSE) binned by misreporting amount for the aggregate regression model with SHAP (SHapley Additive exPlanation), for the five expense line items with the highest misreporting rates. This illustrates that RMSE is driven by poor risk estimation for taxpayers with the largest misreporting amounts.

Next, we turn to operational utility. We summarize performance in Table C.1. We show recovery@k in Figure C.2 and no-change rate@k in Figure C.3. First, as discussed in Section 7, we observe that constituent risk modeling tends to recover more revenue than local explanation models, while simple heuristics tend to perform worse. This trend is reflected in both the regression models, as well as the positive reported value conditional models. The classification models present a notable exception, which suggests that the predicted class probabilities provide poor signal for ranking high-risk issues and taxpayers.

The superior risk recovery of constituent models compared to local explanation models is reflected in no-change rate@k and r-precision@k. The no-change rate illustrates another priority besides revenue generation, where auditors aim to avoid auditing a tax return that generates minimal or no revenue; the optimal no-change rate is 0. The constituent models tend to have a comparable or marginally lower no-change rate than the local explanation models. Similar to recovery@k, the quality of the local explanation models improves when moving from an actual underlying aggregate prediction model to the oracle aggregate model. Similar also to recovery@k, the local and global heuristics tend to have comparable or worse no-change rates compared to the local explanation models.

Turning next to r-precision@k, we see that the constituent oracle model, which selects all issues perfectly, obtains a perfect score at 1.000, as expected. The local heuristic, where line items are selected for audit in decreasing order of reported value, tends to have the highest r-precision@5 of 0.751 (SE = 0.004), which can be interpreted to mean that approximately 75% of the selected line items align with the top five issues with the greatest positive adjustments. The constituent models achieve a similar r-precision@5 of 0.702 (SE = 0.002) in the regression setting and 0.752 (SE = 0.002) in the classification setting. The SHAP explanation models follow in performance, but with a drop to 0.602 (SE = 0.011) with the regression models; we observe similar results with the classification and positive reported value conditional models. Thus, by this metric, the constituent risk model and local heuristic tend to obtain better performance than the explanation-based

approaches. Note that, while r-precision@k provides a summary of issue selection over the entire taxpayer population, it does not take the auditing budget K into account.

Table C.1. R-precision@k, trajectory@k, and no-change rate@k for the oracle, constituent regression, aggregate regression, and heuristic risk estimation models, at $k = 5$ and $K = 3.0\%$. Values show mean and standard error across the five folds. SHAP = SHapley Additive explanation; CS = Cohort Shapley; LIME = Local Interpretable Model-Agnostic Explanations.

Model	R-Precision@5	Trajectory@5	No-Change Rate@5
Constituent Oracle	1.000 (0.000)	$17,841,294 ($770,405)	0.000 (0.000)
Constituent Regression	0.702 (0.002)	$10,517,700 ($1,246,249)	0.225 (0.019)
Aggregate Regression with SHAP	0.602 (0.011)	$8,975,128 ($973,356)	0.225 (0.018)
Aggregate Regression with CS	0.510 (0.014)	$8,882,328 ($877,816)	0.268 (0.014)
Aggregate Oracle with CS	0.284 (0.006)	**$10,772,925 ($737,071)**	**0.118 (0.025)**
Aggregate Regression with LIME	0.229 (0.007)	$537,583 ($143,284)	0.682 (0.028)
Global Heuristic	0.451 (0.006)	$5,246,358 ($518,340)	0.261 (0.023)
Local Heuristic	**0.751 (0.004)**	$9,137,485 ($955,828)	0.354 (0.018)

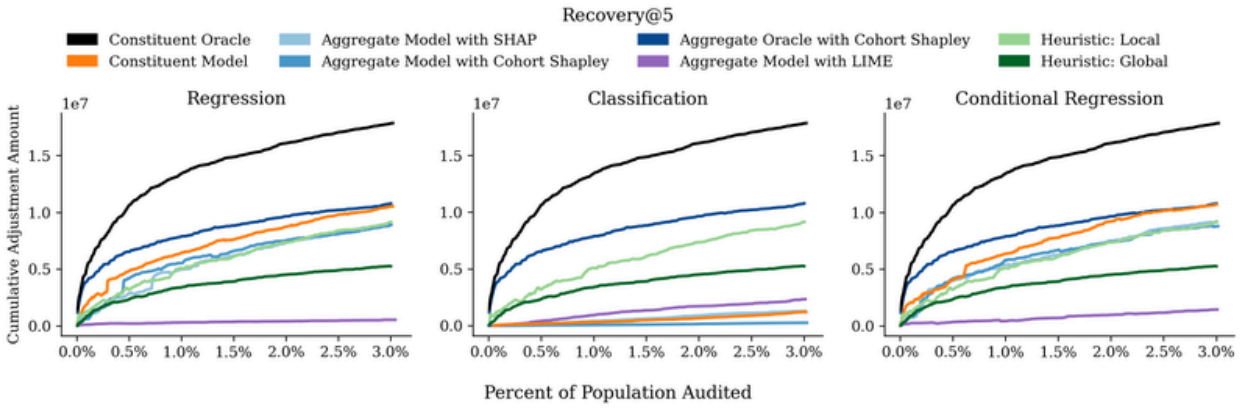


Figure C.2. Recovery@k as the audit budget increases from 0.0% to 3.0%, for the regression, classification, and positive reported conditional models. SHAP = SHapley Additive explanation; LIME = Local Interpretable Model-Agnostic Explanations.

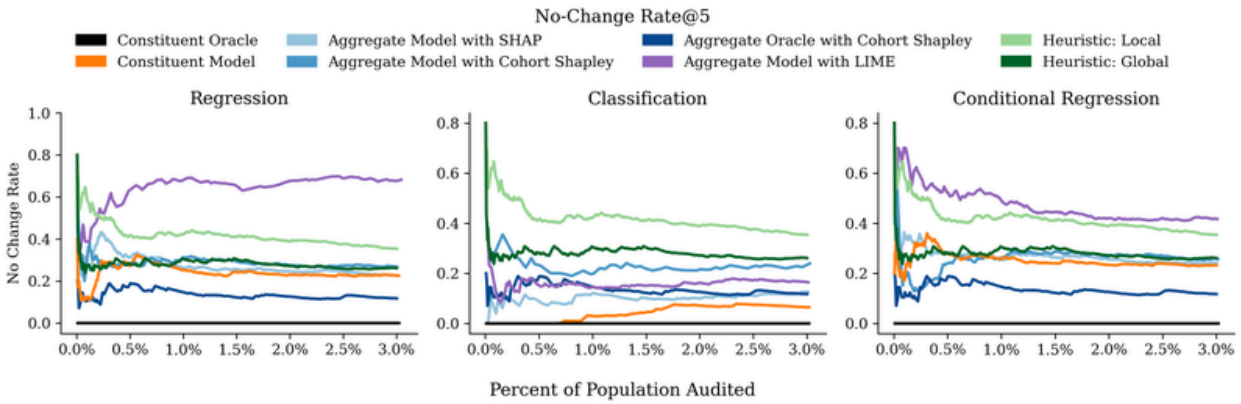


Figure C.3. No-change rate@k as the audit budget increases from 0.0% to 3.0%, for the regression, classification, and positive reported conditional models. SHAP = SHapley Additive explanation; LIME = Local Interpretable Model-Agnostic Explanations.

©2026 Isabel O. Gallegos, Benjamin B. Seiler, Jacob Goldin, Carlos Guestrin, and Daniel E. Ho. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. In this work, we do not analyze any operational models used by the IRS. Instead, we build our own models for analysis. Our findings apply only to these constructed models and may not directly map to real IRS operational models, which could differ in their levels of local explainability. ↩

2. For example, different auditors who conduct NRP audits may find different adjustments on the same return. The IRS exploits that variation when estimating the tax gap (targeting the adjustment of the hypothetical 'best auditor'). In addition, even NRP audits may miss certain forms of misreporting, and NRP auditors may accept certain nonsuspicious line items as valid without a full investigation, even though it is possible that some of those line items would prove to be erroneous. Because our focus is on explaining what specific auditors actually detect and adjust, we assume—consistent with historical risk modeling efforts by the IRS and other academic work (Henderson et al., 2023; Holtzblatt & Engler, 2022)—the average NRP auditor's adjustment is the ground truth misreporting amount. ↩

3. Adverse action notices do include a disclosure of factors that contributed to the decision. ↩

References

- Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a transparent evaluation of post hoc model explanations. In S. Koyejo, S.

Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 15784–15799). Curran Associates. ↩

- Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. *Proceedings of Machine Learning Research*, *97*, 161–170. ↩
- Arras, L., Osman, A., & Samek, W. (2022). CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, *81*, 14–40. https://doi.org/10.1016/j.inffus.2021.11.008 ↩
- Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1194–1206). ACM. https://doi.org/10.1145/3531146.3533179 ↩
- Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1479–1503). ACM. https://doi.org/10.1145/3531146.3533204 ↩
- Braithwaite, J., & Braithwaite, V. (1995). The politics of legalism: Rules versus standards in nursing-home regulation. *Social & Legal Studies*, *4*(3), 307–341. https://doi.org/10.1177/096466399500400301 ↩
- Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036 ↩
- Cockburn, I., Kortum, S., & Stern, S. (2003). Are all patent examiners equal? Examiners, patent characteristics, and litigation outcomes. In W. M. Cohen & S. A. Merrill (Eds.), *Patents in the knowledge-based economy* (pp. 19–53). The National Academies Press. ↩
- Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H., & Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 203–214). ACM. https://doi.org/10.1145/3514094.3534159 ↩
- Domingos, P., Lowd, D., Kok, S., Nath, A., Poon, H., Richardson, M., & Singla, P. (2016). Unifying logical and statistical AI. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science*

(pp. 1–11). ACM. https://doi.org/10.1145/2933575.2935321

↩

- Elzayn, H., Smith, E., Hertz, T., Guage, C., Ramesh, A., Fisher, R., Ho, D. E., & Goldin, J. (2024). Measuring and mitigating racial disparities in tax audits. *Quarterly Journal of Economics*, *140*(1), 113–163. https://doi.org/10.1093/qje/qjae027

↩

- Equal Credit Opportunity Act, Pub. L. No. 94-239, 90 Stat. 251 (1974). https://www.govinfo.gov/content/pkg/USCODE- 2011-title15/pdf/USCODE-2011-title15-chap41-subchapIV.pdf

↩

- Fair Credit Reporting Act , Pub. L. No. 91-508, 84 Stat. 1114 (1970). https://www.govinfo.gov/content/pkg/USCODE-2011-title15/pdf/USCODE-2011-title15-chap41-subchapIII.pdf

↩

- Feinstein, J. S. (1989). The safety regulation of US nuclear power plants: Violations, inspections, and abnormal occurrences. *Journal of Political Economy*, *97*(1), 115–154. https://doi.org/10.1086/261595 ↩
- Feng, C., Li, L., & Xu, C. (2023). Advancements in predicting and modeling rare event outcomes for enhanced decision-making. *BMC Medical Research Methodology*, *23*(1), Article 243. https://doi.org/10.1186/s12874-023-02060-x ↩
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In F. Bonchi, F. Provost, T. Eliassi-Rad, W. Wang, C. Cattito, & R. Ghani (Eds.), *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (pp. 80–89). IEEE. https://doi.org/10.1109/DSAA.2018.00018

↩

- Guidotti, R. (2021). Evaluating local explanation methods on ground truth. *Artificial Intelligence*, *291*, Article 103428. https://doi.org/10.1016/j.artint.2020.103428 ↩
- Henderson, P., Chugg, B., Anderson, B., Altenburger, K., Turk, A., Guyton, J., Goldin, J., & Ho, D. E. (2023). Integrating reward maximization and population estimation: Sequential decision-making for Internal Revenue Service audit selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(4), 5087–5095. https://doi.org/10.1609/aaai.v37i4.25637

↩

- Internal Revenue Service. (2023a). *Inflation Reduction Act strategic operating plan, FY 2023-2031*. https://www.irs.gov/pub/irs-pdf/p3744.pdf ↩
- Internal Revenue Service. (2023b). *Internal Revenue manuals: Classification of NRP returns*. https://www.irs.gov/irm/part4/irm_04-022-003 ↩
- Internal Revenue Service. (2023c). *IRS Newsroom: The Tax Gap*. https://www.irs.gov/statistics/irs-the-tax-gap ↩
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868 ↩
- Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 275–284). ACM. https://doi.org/10.1145/3097983.3098066 ↩
- Leavitt, M. L., & Morcos, A. (2020). *Towards falsifiable interpretability research*. ArXiv. https://doi.org/10.48550/arXiv.2010.12016 ↩
- Liu, Y., Khandagale, S., White, C., & Neiswanger, W. (2021). *Synthetic benchmarks for scientific research in explainable machine learning*. ArXiv. https://doi.org/10.48550/arXiv.2106.12543 ↩
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Curran Associates. ↩
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9 ↩

- Mase, M., Owen, A. B., & Seiler, B. B. (2019). *Explaining black box decisions by Shapley cohort refinement*. ArXiv. https://doi.org/10.48550/arXiv.1911.00467 ↩
- McCarthy, J. (1959). Programs with common sense. *Proceedings of the Symposium on the Mechanization of Thought Processes* (Vol. 10, pp. 75–92). National Physics Laboratory. ↩
- Miró-Nicolau, M., Jaume-i Capó, A., & Moyà-Alcover, G. (2024). Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence*, *335*, Article 104179. https://doi.org/10.1016/j.artint.2024.104179 ↩
- Morgan, T. (2002). *Business rules and information systems: Aligning IT with business goals*. Addison-Wesley Professional. ↩
- Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1198–1212). ACM. https://doi.org/10.1145/3593013.3594074 ↩
- Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), Article 295. https://doi.org/10.1145/3583558 ↩
- Office of the Inspector General, U.S. Nuclear Regulatory Commission. (1995). *Factors contributing to inconsistency in the operating reactor inspection program* (No. OIG-95A-04). ↩
- Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1139–1150). ACM. https://doi.org/10.1145/3593013.3594069 ↩

- Regulation 2024/1689. *Laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* European Parliament, Council of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng ↩
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778 ↩
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x ↩
- Shamsabadi, A. S., Yaghini, M., Dullerud, N., Wyllie, S., Aıvodji, U., Alaagib, A., Gambs, S., & Papernot, N. (2022). Washing the unwashable: On the (im)possibility of fairwashing detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 14170–14182). Curran Associates. ↩
- Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton University Press. ↩
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). ACM. https://doi.org/10.1145/3375627.3375830 ↩
- U.S. Government Accountability Office. (2005). *Nursing homes: Despite increased oversight, challenges remain in ensuring high-quality care and resident safety* (No. GAO-06-117). https://www.gao.gov/assets/gao-06-117.pdf ↩
- von Halle, B. (2001). *Business rules applied: Building better systems using the business rules approach*. Wiley Publishing. ↩

- Yang, F., Du, M., & Hu, X. (2019). *Evaluating explanation without ground truth in interpretable machine learning*. ArXiv. https://doi.org/10.48550/arXiv.1907.06831

  ↩