Tulane Economics Working Paper Series

Optimal Audit Targeting with Machine Learning: Evidence from Pakistan

Nicholas Lacoste
Tulane University
nlacoste1@tulane.edu

Zehra Farooq
Federal Board of Revenue
Pakistan
zehra.farooq@gmail.com

Working Paper 2603
February 2026

Abstract

This paper bridges welfare economics and machine learning econometrics to develop empirically implementable algorithms for optimal audit targeting. We derive a sufficient statistic-based targeting algorithm that depends on three individualized causal effects: the immediate revenue recovered from an audit, the causal effect of an audit on long-run tax revenue, and the marginal administrative cost of an audit. We estimate these effects with a variety of machine learners comparing causal forests, LASSO, gradient boosted trees, and neural networks using the universe of Pakistani income tax returns, exploiting years in which audits were assigned completely at random. We implement our targeting algorithms in out-of-bag years, comparing them to the real-world policy when audits were partially or entirely targeted. We show that the real world audit program in Pakistan lost almost 173,000 Rs ($\approx \$1,700$) in net revenue per-audit, while our optimal policy generates 285,000 Rs ($\approx \$2,800$) in expected net revenue per-audit. We also find that targeting audits based on immediate recoup is sub-optimal to targeting on long-run deterrence in this setting. Moving forward, our framework offers a general approach to empirical welfare maximization using machine learning in resource-constrained policy settings.

Keywords: optimal audit policy, tax enforcement, machine learning, sufficient statistics
JEL codes: H21, H26, C14, C45

Optimal Audit Targeting with Machine Learning: Evidence from Pakistan*

Nicholas Lacoste[1] and Zehra Farooq[2]

[1]Tulane University
[2]Federal Board of Revenue, Pakistan

December 10, 2025

Abstract

This paper bridges welfare economics and machine learning econometrics to develop empirically implementable algorithms for optimal audit targeting. We derive a sufficient statistic-based targeting algorithm that depends on three individualized causal effects: the immediate revenue recovered from an audit, the causal effect of an audit on long-run tax revenue, and the marginal administrative cost of an audit. We estimate these effects with a variety of machine learners comparing causal forests, LASSO, gradient boosted trees, and neural networks using the universe of Pakistani income tax returns, exploiting years in which audits were assigned completely at random. We implement our targeting algorithms in out-of-bag years, comparing them to the real-world policy when audits were partially or entirely targeted. We show that the real-world audit program in Pakistan *lost* almost 173,000 Rs ($\approx \$1,700$) in net revenue per-audit, while our optimal policy *generates* 285,000 Rs ($\approx \$2,800$) in expected net revenue per-audit. We also find that targeting audits based on immediate recoup is sub-optimal to targeting on long-run deterrence in this setting. Moving forward, our framework offers a general approach to empirical welfare maximization using machine learning in resource-constrained policy settings.

Keywords: optimal audit policy, tax enforcement, machine learning, sufficient statistics
JEL: H21, H26, C14, C45

*We'd like to thank the Pakistani Federal Board of Revenue for providing support. We'd also like to thank Nicholas' dissertation committee members: William Dodds, Katy Bergstrom, and Augustine Denteh, for invaluable guidance and suggestions. Lastly, we'd like to thank Juan Rios for incredibly helpful comments, as well as seminar participants at Tulane University, CAIDS, and SEA. Email addresses: Nicholas Lacoste (Corresponding Author) nlacoste1@tulane.edu; Zehra Farooq zehra.farooq@gmail.com

1 Introduction

There is a strong association between the fiscal capacity of a state and its economic development. Significant tax collection gaps can impact the government's ability to adhere to the optimal tax schedule and provide public goods. This has been re-emphasized recently by Besley and Persson (2014) (among others), who highlight that developing countries collect between 10-20% of their GDP in taxes, as opposed to their OECD counterparts which collect around 40% on-average, despite having similar marginal tax rates on income. Economists have pointed to revenue mobilization constraints as significant barriers to economic development for over 60 years (e.g. Kaldor (1963)), yet the issue remains salient even today (Benitez et al. (2023); Dom et al. (2022)). It is well-established that a significant portion of this issue stems from high rates of tax evasion in developing countries due to pervasive informal sectors and small enforcement budgets (e.g. Jensen (2022)).[1]

Modern tax authorities target audits through algorithms designed to predict evasion likelihood in order to better allocate their resources (Best et al., 2021). However, this strategy fails to consider important realities of optimal tax enforcement policies. First, heterogeneity in an audit's effect on individuals' behavior may affect the government's long-run tax base equally to (or more than) the revenue recouped today from an audit. If audits shift peoples' perceptions of the riskiness of their tax filing behavior, audits may induce future evasion deterrence or, conversely, exits to the informal sector (if the barriers to exit are low). Targeting audits on evasion likelihood also ignores heterogeneity in the administrative costs to conduct audits across taxpayers and any compliance costs imposed on audited individuals. In theory, audit selection policy (and any policy, for that matter) should be guided by a measure of social welfare that encompasses the full set of revenue and cost channels to the government and society.

In this paper, we provide a framework for deriving optimal audit selection rules using causal machine learning (ML). We begin by developing a simple model which we use to formulate the optimal audit targeting policy. Importantly, we illustrate that the optimal audit targeting policy can be expressed in terms of three individualized causal effects ("sufficient statistics"): (1) the expected upfront revenue recoup of an audit, (2) the expected effect on the net present value of long-run tax revenue collected – which can broadly be attributed to evasion deterrence, and (3) the expected administrative cost of an audit. In this manner, we contribute to the large literature in public economics on using sufficient statistics to compare the welfare incidence of alternative policies. For example, sufficient statistics like the Marginal Excess Burden – MEB – (e.g. Eissa et al. (2008); Eissa and Hoynes (2011)), the Net Social Benefit – NSB – (e.g. Olken (2007); Heckman et al. (2010)), and the Marginal Value of Public Funds – MVPF – (e.g. Mayshar (1990); Slemrod and Yitzhaki (1996); Hendren (2016); Hendren and Sprung-Keyser (2020)) have all been used to compare the welfare impacts *across* policy domains using *average* causal effects. Additionally, Bergstrom et al. (2025) have recently shown that welfare-weighted MVPFs and/or welfare-weighted NSB's can be used to determine optimal policy reforms across alternative policies. Here, we show that these concepts naturally extend to deriving the optimal allocation of treatment *within* policies by leveraging recent econometric advances for estimating *individualized* causal effects with machine learning.

We then estimate these sufficient statistics by exploiting a national program of randomized audits in

[1] Serious tax collection gaps are not limited to developing countries. For example, Sarin and Summers (2019) estimate that the US can generate an additional $1 trillion in tax revenue over 10 years by investing in the IRS's audit capacity.

Pakistan. Specifically, we use the universe of Pakistani individual income tax returns and audit results during two years where audit selection was determined completely at-random as well as three years after where audits were targeted towards those with a greater suspected evasion risk. This large scale quasi-experiment allows us to employ a set of causal machine learning estimators and obtain unbiased estimates of conditional average treatment effect functions (CATEs) for (1) the expected revenue recouped from an audit and (2) the long-run deterrence effects of an audit, and it allows us to test these estimates out-of-bag against real-world audit policy. We employ two complementary approaches to estimate these causal functions. First, we directly target the CATE with Causal Forests (Wager and Athey, 2018). Second, we estimate proxy predictors for the CATE function using LASSO, gradient-boosted trees (XGBoost), and Neural Networks. We then refine these proxies following Chernozhukov et al. (2018) to derive the best linear predictors of the CATE function on each proxy. In this sense, our work also fits into a burgeoning literature on using machine learning within the context of randomized policy interventions to improve selection (e.g. Ash et al. (2024); Knittel and Stolper (2025), etc.).

In addition to tax returns and audit results, we also observe the annual budgets of each of Pakistan's 16 regional tax offices. While these annual budgets are granular with respect to yearly spending on each line item of labor/overhead/etc., they do not allow us to directly estimate the marginal cost of any particular audit. In order to generate heterogeneity in audit costs, we administer a survey to tax officials in the Pakistani Federal Board of Revenue (FBR). This survey elicited auditor time-use and had tax officers approximate a mapping between tax return line items and audit duration. We document that the average audit in Pakistan takes 17.4 hours to complete, but there is substantial heterogeneity in this duration which varies according to several taxpayer characteristics. For instance, we find that taxpayers with higher incomes, numerous income sources, large business losses, and large tax credits can take up to 80% longer to audit on-average. We use the results of this survey to estimate individualized audit costs. We find that in the first year of semi-random (i.e. targeted) audits in Pakistan, the average audit cost the FBR 219,341 Rs to conduct ($\approx \$2,128$ USD), but ranged from 135,000 Rs ($\approx \$1,300$ USD) in the bottom 50% of cases to as much as 1.4M Rs ($\approx 14,000$ USD) and 2M Rs ($\approx \$19,500$ USD) in the top 1% and top 0.1% cases, respectively.[2]

After imputing marginal audit costs and estimating the necessary causal effects on revenue, we apply our targeting algorithm to three out-of-bag years where audits were either partially randomized or completely based on a selection algorithm. We find that all of the machine learners can predict the initial revenue recoup with a high level of accuracy and that this accuracy does not deteriorate over time. This is suggestive that the CATE functions estimated in years of randomized audits do not structurally change (at least, not quickly) over time. Our estimates also appear to be well-calibrated to the causal function of the deterrence effect. We estimate small, positive average deterrence effects during the randomized years, but with a substantial amount of heterogeneity. Specifically, audits increased discounted future tax liability by $\approx 163,109$ Rs ($\approx \$1,560$ USD) over three years on average according to the best calibrated proxy learner. However, the top 1% of deterrence impacts are as high as $\approx 526,000$ Rs ($\approx \$5,000$ USD) while the bottom 1% are as low as $\approx -186,000$ Rs ($\approx -\$1,780$). This is consistent with the canonical literature in development economics

[2]Boning et al. (2025) report similar estimates in the US, but look across the income distribution rather than the range of costs themselves. They find an audit in the bottom 50% of the income distribution costs the IRS about $5,200, while an audit of the top 1% and 0.1% of income costs about 2.1x and 2.8x more to conduct, respectively. We estimate qualitatively similar patterns when considering how Pakistan's costs vary across the income distribution for comparison. Audits of individuals in the bottom 50% of income cost about $2,000 on average, while audits in the top 1% and top 0.1% of income cost about 2.5x and 3.8x more, respectively.

that suggests some "marginal filers" may be induced to exit into the informal sector entirely or learn better evasion practices if audited (see e.g. Loayza (1996)).

We estimate that Pakistan's audit program was highly costly relative to the small tax base improvements it generated. In the first out-of-bag year, the average audit of an individual taxpayer generated a net loss of almost $-173,000$ Rs ($\approx -\$1,700$ USD) to the government. In contrast, the policies we derive from our selection algorithm generate about 285,000 Rs ($\approx \$2,800$) in expected net revenue per-audit. Combining these estimates of revenue effects with our estimates of administrative cost effects, we construct MVPF estimates that suggest audits under our derived policy impose a \$1.58 expected welfare cost per-dollar of net revenue raised through audits from our preferred machine learners. This is comparable to Boning et al. (2025)'s figure in the US of a \$1.30 average welfare cost per-dollar of revenue raised, suggesting that it is perhaps achievable for the enforcement policy of a developing economy to emulate the welfare incidence of a more developed country.

The final part of our analysis focuses on the characteristics of audited individuals under the real-world policies vs. the policies derived from our targeting algorithms. The first key empirical takeaway from this exercise is that the targeting algorithms based on machine learners heavily favor auditing those with large expected deterrence impacts over those with large expected initial recoup. Somewhat surprisingly, we find that there is a tradeoff between the two impacts. This leads to the implication that, at least in this developing economy setting, the most efficient approach to expanding the tax base is to (counterintuitively) *not* target those with the largest expected recoup, but rather to target those who will more broadly contribute to future tax revenue streams. Individuals audited under the ML-derived policies have greater profits for their businesses (all audited taxpayers in Pakistan are business owners), greater non-salary incomes, declare more business expenses and tax credits, and declare more turnover costs. They are also more likely to declare large losses from capital assets.

This paper expands upon and coalesces recent theoretical and empirical work on using predictions to improve the tax audit process. Caspi et al. (2024), for example, build upon Kaplow (1990), Mayshar (1991), Keen and Slemrod (2017), and others to conduct a thorough theoretical exercise exploring the welfare-optimal perturbation to the audit probability regime when one has accurate predictions of evasion. Empirical exercises include several papers in the computer science literature which present algorithms to detect tax evasion (e.g. Wu et al. (2020); Ruan et al. (2019); and several others). These papers are primarily concerned with out-of-bag prediction quality despite the fact that audits are endogenously selected in practice, and are therefore of limited policy use despite their mathematical design interest. Recently, Battaglini et al. (2024) aimed to address this endogenous selection problem by presenting ML algorithms that leveraged the longitudinal nature of audits in Italy. They find evidence that one my improve evasion detection rates rather significantly by re-prioritizing and replacing the audits with the lowest success likelihood with audits with greater success likelihood. The primary limitation of the existing research is that it has attempted to address only part of the problem when considering optimal treatment allocations with machine learning. That is, the distribution of any treatment intervention should be determined in such a way that maximizes social welfare. The papers which have considered the welfare impacts of audits have either lacked an empirical/applied component or lacked an optimization component. The papers which have considered using ML in the context of audits have focused on the ability of governments to raise more upfront revenue, not necessarily on the govern-

ment's ability to induce deterrence in a cost-effective manner – and never in highly resource-constrained environments such as developing countries. Our work is the only study (to our knowledge) that attempts to bridge this gap and tackle the entirety of the issue: presenting a structural welfare model that is simple, estimable, and realistically implementable with a set of predictions obtainable through causal ML.

Our work also contributes to the small literature on understanding the administrative costs to conduct audits. While, to the best of our knowledge, we are the first paper that attempts to estimate individualized audit costs, Boning et al. (2025) recently expanded on the work of Holtzblatt and McGuire (2020) and estimated average audit costs with heterogeneity across the income distribution in the US. Lastly, we contribute to the literature on understanding the deterrence effects of audits in both developed and developing countries. Our findings are generally in-line with what most papers estimate in that audits are typically associated with positive deterrence effects. For example, Kleven et al. (2011), Advani et al. (2023), and DeBacker et al. (2018) each estimate positive and persistent deterrence effects of about 30% of the initial audit recoup amount in Denmark, the UK, and the US, respectively. However, our estimates contrast with similar work in Pakistan (Best et al., 2021) which did not find notable deterrence impacts among Pakistani VAT firms.

The rest of this paper proceeds as follows. Section 2 outlines our baseline model and derives our primary audit selection algorithm. Section 3 discusses our empirical policy environment and the administrative tax return data we employ. We also discuss the data on annualized tax office budgets and our auditor survey. Section 4 discusses our machine learning approaches to estimating CATE functions. Section 5 reviews the results of our prediction exercises and compares the real-world policies with our derived policies. Finally, we conclude in Section 6 with some closing advice to governments who may wish to employ our framework.

2 Theory: Deriving Audit Targeting Rules

We begin with a general model which derives the necessary sufficient statistics and accompanying solution algorithms for determining optimal audit policy under various policy objectives and constraints. Our model builds upon earlier work on optimal tax enforcement (Keen and Slemrod (2017); Caspi et al. (2024)) and audit MVPFs (Boning et al., 2025). We also leverage the insights regarding optimal policy reforms of Bergstrom et al. (2025) in several cases. We will apply this model empirically to the audit policy of Pakistan from 2014 – 2018. However, we believe this model to be a general framework for audit selection policy and beyond.

2.1 Setup

We observe tax filers indexed $i \in [1, N]$ over time periods $t \in [0, \infty)$. We assume individuals differ in terms of a vector of characteristics, $\boldsymbol{x} \in \boldsymbol{X}$, which are determined according to an unobserved density function. Notably, the government can only observe a subset of the characteristic vector, $\widetilde{\boldsymbol{x}} \subset \boldsymbol{x}$. In each period, the government must select whom to audit among the taxpaying population. Denote $\boldsymbol{\alpha}_t = \{\alpha_1, \alpha_2, ..., \alpha_N\}_t$, where $\alpha_i \in \{0, 1\}$, as the vector of binary audit selections in year t.

Individuals have a utility function, $U_i(y_{it}, z_{it}; \boldsymbol{\alpha}_t, \boldsymbol{x})$, where y_{it} indicates consumption and z_{it} indicates true labor income, both of which are endogenous to personal characteristics and the distribution of audits

(which includes their own audit status and possibly the audit status of others).[3] Individuals are taxed according to a tax liability function of their true income, $T(z_{it})$, but people can misreport their income at $\hat{z}_{it}$ in order to evade $T(z_{it}) - T(\hat{z}_{it}) = e_{it}$ dollars (or Pakistani Rupees in our empirical setting) of this tax liability which they keep if not audited.[4] If $\hat{z} > z$, the excess liability is refunded to the taxpayer and no audit will occur, so $e_{it} \geq 0$ always. If the person is audited, we assume all evasion is detected and they must pay this e_{it} dollar amount plus a financial penalty, $\phi(e_{it})$, which increases in e_{it}. Additionally, we assume that individuals make evasion decisions by considering their ex-ante perceived risk of an audit. Denote $\hat{\boldsymbol{a}}_t = \{\hat{a}_1, ..., \hat{a}_N\}_t$ as the vector of these perceived audit probabilities, where $\hat{a}_{it}(e_{it}; \boldsymbol{\alpha}_{t-k}, \widetilde{\boldsymbol{x}}) \in [0, 1]$ is individual i's perceived audit risk which evolves as a function of their prior audit history and the subset of their characteristics observable to the government.

In our setup, we model the government's problem by considering real-world tax enforcement policy dynamics. It is critical to distinguish between the federal government, who sets agency budgets and spends tax revenue, with the tax enforcement agency, who ultimately decides who is audited and conducts the audits using their assigned administrative budget. We therefore assume the enforcement agency's fixed internal budget for tax audits is exogenous from their perspective. There are three mechanisms through which audit policy affects the net budget of the federal government: (1) through the upfront recouped tax liability which is absorbed into the federal government's general revenue fund, (2) through the net present value (NPV) of future declared tax liability, and (3) through the upfront administrative cost to conduct audits. Formally, we define these components as:

$$\begin{aligned}
&\underbrace{R^m(\boldsymbol{\alpha}_t) = \sum_{i=1}^{N} \alpha_{it}\Big(\phi(e_{it}) + e_{it}\Big)}_{\text{Upfront mechanical revenue recouped}} \\
&\underbrace{R^f(\hat{\boldsymbol{z}}; \boldsymbol{\alpha}_t) = \sum_{t=1}^{\infty} \beta^{t-1} \sum_{i=1}^{N} T\Big(\hat{z}_{it}(\hat{a}_{it})\Big)}_{\text{NPV of future tax revenue streams}} \\
&\underbrace{\mathcal{C}(\boldsymbol{\alpha}_t) = \sum_{i=1}^{N} \alpha_{it} C_i(\boldsymbol{x})}_{\text{Upfront mechanical administrative costs}}
\end{aligned} \tag{1}$$

Where $C_i(\boldsymbol{x})$ is the administrative cost of auditing person i, β is an exogenous discount rate (assumed to be equivalent between the federal government and its agencies), and $T\Big(\hat{z}_{it}(\hat{a}_{it})\Big)$ represents the total taxes paid by person i in year t. $R^f(\cdot)$ in period $t = 1$ is therefore the discounted sum of the declared taxable income of individuals over time, which is a function of their perceived audit risk if they declare $\hat{z}_{it}$. This perceived audit risk is in-turn affected by the distribution of audits in period $t = 1$.

Finally, we assume that social welfare under audit regime $\boldsymbol{\alpha}$ is a weighted sum of individual utilities plus

[3] The tax evasion literature typically distinguishes between "specific deterrence," which refers to behavioral responses resulting from one's own audit status, and "general deterrence," which refers to behavioral responses stemming from changes to the audit environment itself (e.g. Allingham and Sandmo (1972); Boning et al. (2025)). Our setup allows conceptually for both types of responses to enter, but we will ultimately focus on specific deterrence in order to maintain computational tractability.

[4] This notational convention follows the setup of Boning et al. (2025) in order to keep the evasion decision as a dollar figure rather than as a share of true income. This notation will ultimately simplify the transition to the empirical strategy (discussed in the upcoming sections) which will be to estimate dollar values of evasion.

the additional tax revenue generated from regime $\boldsymbol{\alpha}$ multiplied by the shadow value of public funds, λ:

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \psi(\boldsymbol{x}) U_i(y, z; \boldsymbol{\alpha}, \boldsymbol{x}) + \lambda \big[R^m(\boldsymbol{\alpha}) + R^f(\boldsymbol{\alpha}) \big] \tag{2}$$

Where $\psi(\boldsymbol{x})$ is the welfare weight the government places on type-$\boldsymbol{x}$ individuals. Intuitively, λ captures the welfare gain to society when the federal government raises \$1 of additional revenue and spends it on whichever policy option they choose.

2.2 The Government's Problem

The tax agency's goal is to select the set of audits that maximizes welfare under the knowledge that (1) they must adhere to a strict internal budget constraint and (2) that the federal government controls the revenue raised by their audits – assuming the long term budget of the government balances by spending all additional revenue on the federal government's chosen numeraire policy (or set of policies, without loss of generality). The tax agency must therefore solve (dropping the t subscript to reduce clutter):

Tax Agency's Problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{N} \psi(\boldsymbol{x}) U_i(y, z; \boldsymbol{\alpha}, \boldsymbol{x}) + \lambda \Big[E_{\widetilde{\boldsymbol{x}}} [R_i^m(\boldsymbol{\alpha}) + R_i^f(\boldsymbol{\alpha})] \Big] \\ \text{s.t.} \quad & E_{\widetilde{\boldsymbol{x}}}[\alpha_i C_i(\boldsymbol{x})] \leq \bar{\mathcal{C}} \\ & \alpha_i \in \{0, 1\} \quad \forall i \end{aligned}$$

The expectations with respect to $\widetilde{\boldsymbol{x}}$ reflect that the agency must obtain asymptotically unbiased predictions of R^m, R^f, and C using the observable subset of taxpayer characteristics.[5] $\bar{\mathcal{C}}$ is the exogenous (to the tax agency) maximum upfront expenditure the agency can incur (in expectation) when conducting audits. If we allow the last "marginally audited" individual to be audited with some probability $\alpha_i \in [0, 1]$, this stochastic optimization problem becomes an instance of the fractional knapsack problem (Dantzig, 1957). The optimal audit policy can therefore be solved with an appealing treatment prioritization algorithm akin to greedily selecting audits according to a revenue-cost ratio (Sun et al., 2024).

Before outlining the solution algorithm, it is useful to note that $\psi(\boldsymbol{x})\Delta U_i = \eta_i \frac{1}{\Upsilon(\boldsymbol{x})} \Delta U_i = \eta_i WTP_i$, where $\Upsilon(\boldsymbol{x})$ is the marginal utility of consumption for type-$\boldsymbol{x}$ individuals and $\Delta U_i = U_i(\cdot; \alpha_i = 1, \boldsymbol{x}) - U_i(\cdot; \alpha_i = 0, \boldsymbol{x})$. So the object $\eta_i = \psi(\boldsymbol{x})\Upsilon(\boldsymbol{x})$ is the *social* marginal value of consumption for type-$\boldsymbol{x}$ individuals and WTP_i is person i's willingness-to-pay for an audit, which will be weakly negative.

Proposition 1. *The solution to the **Tax Agency's Problem** is defined by an algorithm which greedily ranks and sequentially audits individuals. This greedy ranking is based on a sufficient statistic representing the expected welfare gain of auditing person i, then (externally) spending all revenue raised on the chosen numeraire policy. This is outlined as Algorithm 1.*

[5] There is an implication here that the observable subset $\widetilde{\boldsymbol{x}}$ must be the full set of confounders, $\boldsymbol{x}$, in the absence of experimental/quasi-experimental variation in audit probabilities. In our empirical setting, we leverage experimental variation in $\boldsymbol{\alpha}$, and therefore the observable subset of heterogeneity dimensions serves only for more efficient CATE estimation. We will discuss this in greater detail in Section 4.

Algorithm 1 Greedy Audit Selection by sufficient statistic θ_i

1: **Input:** Predictions of the treatment effects $\Delta\hat{R}_i^m$, $\Delta\hat{R}_i^f$, and $\Delta\hat{C}_i$; predictions of $\widehat{WTP}_i$ for all i; estimated value of λ

2: Compute θ_i for each unit i:

$$\theta_i = \eta_i \cdot \frac{\widehat{WTP}_i}{\Delta\hat{C}_i} + \lambda \cdot \frac{\Delta\hat{R}_i^m + \Delta\hat{R}_i^f}{\Delta\hat{C}_i}$$

3: Sort individuals so that $\theta_{(1)} \geq \theta_{(2)} \geq \cdots \geq \theta_{(N)}$

4: Initialize: $\alpha_i = 0 \quad \forall i$

5: **while** $\sum_i^N \alpha_i \Delta\hat{C}_i(\boldsymbol{x}) < \bar{\mathcal{C}}$ **do**

6: Set $\alpha_i = 1$ for highest-ranked i where $\alpha_i = 0$

7: **end while**

8: For marginal unit j, set $\alpha_j = \frac{\bar{\mathcal{C}} - \sum_i^N \alpha_i \Delta\hat{C}_i(\boldsymbol{x})}{\Delta\hat{C}_j}$

9: The resulting audit allocation will implicitly solve the following, where μ is the θ_i value of marginal unit j:

$$\alpha_i = \begin{cases} 0 & \text{for all} \quad \frac{\eta_i \widehat{WTP}_i + \lambda[\Delta\hat{R}_i^m + \Delta\hat{R}_i^f]}{\Delta\hat{C}_i} < \mu \\ [0,1] & \text{for } j \text{ where} \quad \frac{\eta_i \widehat{WTP}_i + \lambda[\Delta\hat{R}_i^m + \Delta\hat{R}_i^f]}{\Delta\hat{C}_i} = \mu \\ 1 & \text{for all} \quad \frac{\eta_i \widehat{WTP}_i + \lambda[\Delta\hat{R}_i^m + \Delta\hat{R}_i^f]}{\Delta\hat{C}_i} > \mu \end{cases}$$

10: **Output:** Final audit selections $\boldsymbol{\alpha}$

Proof. See B.1 □

What Algorithm 1 describes is a greedy algorithm that yields the optimal distribution of audits based on a set of sufficient statistics. Practically, the agency may follow this algorithm by obtaining predictions of $\Delta\hat{R}_i^m$, $\Delta\hat{R}_i^f$, $\Delta\hat{C}_i$, $\widehat{WTP}_i$, and λ in order to compute θ_i for each individual, where $\Delta\hat{R}_i^m$, $\Delta\hat{R}_i^f$, and $\Delta\hat{C}_i$ are the causal effects of an audit on R_i^m, R_i^f, and C_i, respectively. Audits are then chosen by sequentially auditing those with the largest θ_i values until the budget is met in expectation.

Comment (Zero welfare weights assumption). The above solution algorithm reveals that θ_i is the relevant sufficient statistic which must be estimated/predicted for each individual in order to derive the optimal selection policy. However, our definition of θ_i reveals that three additional parameters are needed beyond $\Delta\hat{R}_i^m$, $\Delta\hat{R}_i^f$, and $\Delta\hat{C}_i$ in order to derive θ_i. In particular, we require (1) the shadow value of public funds (λ), (2) the social marginal value of consumption for person i (η_i), and (3) person i's willingness-to-pay to avoid being audited. We can avoid estimating these terms under the simplifying assumption that welfare weights are zero among the audited population of Pakistan. Under zero welfare weights, the WTP_i term drops out for the entire sample, and the λ term is now irrelevant to the selection regime as it acts as a constant scalar for all individuals. We believe this assumption to be reasonable based on our discussion upcoming in Section 3: the taxpaying population in Pakistan consists only of business owners with personal incomes high enough to tax, and whatever policy λ may be attributed to is likely targeting very poor individuals with much higher social weights in this context. For these reasons, we will conduct our main analysis under this assumption, and the θ_i statistic which we need to predict is proportional to a simple revenue-cost ratio: $\hat{\theta}_i = \frac{\Delta\hat{R}_i^m + \Delta\hat{R}_i^f}{\Delta\hat{C}_i}$. Nonetheless, in Appendix A, we consider the addition of non-zero welfare weights, and show that an extension model based on Boning et al. (2025) can derive the WTP_i in terms of most of the same causal effects as above, with one additional causal effect being the compliance cost of the audit on individual i. We refer to this term as the "taxpayer burden" of the audit which is the monetary and psychic costs of audit compliance.

Comment (The Net Social Benefit and the Marginal Value of Public Funds). Our sufficient statistic, θ_i, is equivalent to an individualized estimate of a welfare-weighted version of the Net Social Benefit (WNSB) welfare criterion (Hendren and Sprung-Keyser (2022); García and Heckman (2022)) and is in-line with Bergstrom et al. (2025)'s definition. In order to interpret θ_i as an individualized $WNSB$, the shadow value of public funds parameter, λ, should be explicitly interpreted as a welfare-weighted Marginal Value of Public Funds (WMVPF) of the numeraire policy. Under the assumption that welfare weights are zero in the target population, this parameter is irrelevant because it is simply a scalar to the revenue/cost ratio defined in Algorithm 1. However, as we discuss in Appendix A, if non-zero welfare weights are included, it becomes necessary to estimate λ. This can be a difficult task as it requires knowledge of what the numeraire policy is *and* reliable estimates of its WMVPF according to Saez and Stantcheva (2016), Hendren and Sprung-Keyser (2020), etc.

3 Institutional Background and Data

We now move to our empirical exercise, which applies the results from Section 2 to Pakistan. As we will discuss, Pakistan poses a unique opportunity to estimate these individualized causal effects and evaluate if the framework from Section 2 can improve upon the observed policy.

3.1 Pakistan's Audit Program: 2010 – 2018

Before 2010, tax enforcement in Pakistan was decentralized: local Federal Board of Revenue (FBR) offices determined audit cases independently. In 2010, the FBR decided to roll-out a centralized auditing program where audits were determined based on a randomized or semi-randomized lottery ("semi-random" in this case means that audit selection was random conditional on a set of tax return characteristics). The FBR attempted to conduct the first semi-randomized audit lottery in 2012, but the selection process was challenged in court. As such, the FBR conducted audits completely at-random among the entire tax filing population using public computer ballots until the litigation process ended at the end of 2015. In 2016 and 2017, the FBR conducted their first set of semi-targeted audits by randomly selecting audits among a sample of high-risk taxpayers, where audit eligibility was again determined through a confidential set of parameters. The FBR refers to this selection method as their "Parametric" approach. In 2018, audits continued to be selected semi-randomly, but the eligible sample of taxpayers was selected by an updated algorithm they called the Risk-Based Audit Management System (RAMS). RAMS was a refinement of the Parametric selection method as it employed data-driven weights to the eligibility parameters and cross-referenced tax returns with third party data to flag inconsistencies. After 2018, the RAMS system no longer conducted random audits and instead directly selected them.

After each audit selection drawing, the FBR released the criteria for exclusion from the ballot and the number of audits that were selected. This information could be found on their portal called the Taxpayer's Audit Monitoring System (TAMS). In all years other than 2018, individuals already under audit from the previous tax year were excluded from the draw. It was also the case that individuals making $> 50\%$ of their income from salary/wages were excluded in all years. Drawings took place in public with taxpayer representatives present, and the list of selected cases was posted on the FBR portal with names and addresses

redacted. See FBR (2015) for full details.

Table 1 presents summary statistics regarding the timeline and overarching results of Pakistan's audit policy for individual taxpayers during the years we observe in our data. We document substantial tax evasion rates in our sample: between 35 – 39% of filers were caught evading some tax liability in the years of random audits,[6]. Additionally, of the audited taxpayers caught evading, they were estimated to have evaded between 30 – 40% of their true tax liability.

Our empirical analysis will rely primarily on the years 2014 – 2018. Despite initially very high rates of tax evasion, voluntary tax compliance steadily improved over these years: the FBR registered an additional 300,000 – 400,000 new filers each year while audits detected fewer evaders each year (and interestingly, this trend continued even after 2015 when tax audits were partially targeted towards higher risk taxpayers). There were, however, no additional FBR compliance initiatives taking place from 2014 – 2017 other than audits.[7] After 2017, the dramatic increase in the number of new filers corresponds with several federal initiatives aimed at increasing compliance, including the development of an FBR branch dedicated to encouraging compliance and educating taxpayers on the benefits of it (called the Directorate for Broadening the Tax Base) and various amnesty schemes. These reforms took place outside of our primary sample.

Table 1: Audit and Evasion Statistics by Year

Year	(1) Total Filers	(2) First Time Filers	(3) Audit Regime	(4) Number Audited	(5) Share Audited	(6) Share Evaders	(7) % Liability Evaded (Avg.)
2014	1,108,177	–	Random	52,683	4.8%	38.9%	39.8%
2015	1,350,325	367,732	Random	69,732	5.2%	34.9%	30.4%
2016	1,599,220	330,375	Parametric	30,777	1.9%	28.9%	25.5%
2017	1,930,069	398,529	Parametric	9,140	0.5%	29.6%	21.0%
2018	2,830,049	949,199	RAMS	9,211	0.3%	14.2%	11.5%
2019	3,142,855	560,352	RAMS	Unobserved	–	–	–
2020	3,286,852	306,589	RAMS	Unobserved	–	–	–
2021	2,871,756	74,579	RAMS	Unobserved	–	–	–

Notes: This table shows audit and evasion statistics by year. Column (1) gives the total number of tax filers. Column (2) shows the number of newly registered (first-time) filers. Column (3) describes the audit selection mechanism used in each year. Column (4) gives the number of taxpayers audited. Column (5) shows the share of filers audited. Column (6) reports the share of audited taxpayers found to have underreported income. Column (7) shows the average share of true tax liability (as determined by the audit) that was evaded among those found to have evaded.

In Appendix Table 11, we verify the balance of the randomization in 2014 and 2015. We compare 17 individual characteristics between audited vs. non-audited individuals. Like Best et al. (2021) who used the same randomization years with Pakistani VAT firms, we generally find that the two groups are very similar in both randomization waves.

[6]These figures coincide tightly with other papers using Pakistani tax return data. For example, Waseem (2023) estimates an evasion rate of 35 – 40% among VAT business filers during these years, as does Best et al. (2021).

[7]In Pakistan, the only true enforcement mechanism during these years was tax audits. If a taxpayer chooses not to file a return, there are no penalties as they will not be caught by the FBR. However, there are still strong incentives to pay income taxes and file a return. Namely, taxpayers who file returns are subject to much lower taxes on VAT and other point-of-sale transactions. Much of the compliance growth from 2014 – 2017 was due to natural growth in understanding these incentives.

3.2 Revenue Data

We have administrative data from the Pakistani FBR which contains the universe of individual tax returns in Pakistan from 2014 – 2021. The individual income tax return in Pakistan is structured similarly to itemized returns in most other countries. It begins with a section on income sources, where filers declare their salary income and net investment income (from stocks, property, business, etc., declared separately). Next, there are sections allowing filers to declare deductible business expenses, withholdings, and tax credits. Taxpayers must also specify the sources of their business income (e.g. foreign source, agriculture, property, etc.). We use each of these itemized declarations as primary covariates in our machine learning models described in the next section.[8]

Finally, we also observe detailed information on audits from 2014 – 2018 from the TAMS and the FBR portal. This includes a comprehensive list of audited individuals, associated audit dates, the amount of detected tax evasion, and ultimate tax recoup amounts. We merge this data with tax returns via unique taxpayer IDs.

3.3 Audit Cost Survey

On the administrative cost side, we observe annual budgets from each of Pakistan's 16 regional FBR tax offices and the Islamabad headquarters. Appendix Figure 21 displays an example (values redacted) annual budget from one of the regional tax offices in the 2017 – 2018 fiscal year. This budget contains the total annual expenditures of the tax office broken down into 13 primary line items (and further breaks down those primary line items into additionally granular line items) including tax officer wages; fringe benefits such as medical benefits, performance pay, continuing education, and travel/contingency reimbursements (this category is called "Allowances" in the sample budget); office overhead expenses; etc. Despite the granularity of these expense reports, they do not allow us to directly back-out the marginal cost of any particular audit on their own, which is ultimately what we require in order to estimate θ_i. In order to supplement our budget data with the domain knowledge of tax officers, we administer a survey to all tax officers in the FBR regarding time-use and audit duration. Our survey consists of two main sections: (1) a section with questions regarding the share of tax officer time spent auditing individual taxpayers vs. other aspects of their job, and (2) a section detailing questions regarding audit duration and heterogeneity in audit duration based on the observable covariates contained in tax returns. We received responses from 108 tax officers representing 15 of the 16 regional tax offices. We will describe the survey further as we progress through this section.

In order to generate accurate estimates of marginal audit costs, we draw on the insights of Boning et al. (2025) and Holtzblatt and McGuire (2020) to develop an approach that generates an accurate imputation of *average* audit costs. We then combine this with our survey responses to generate more accurate heterogeneity in audit costs based on observables. In this section, we will describe both aspects of our approach.

[8]In addition to tax returns, we also observe the universe of wealth statements over the same time period. Pakistan, unlike most OECD countries, requires a declaration of wealth in addition to the standard income tax return (though filers are not taxed on wealth as they are in the handful of OECD countries which require wealth statements). Wealth statements include items such as agricultural/residential/business property, mechanical assets, financial assets, livestock, etc. We do not incorporate wealth statement variables in our algorithms because this type of information is not typically collected by tax authorities, and we would like our results to be representative of what can be derived using a typical set of covariates. In future work however, we plan to investigate incorporating these covariates.

3.3.1 Estimating Average Cost per-Audit

Boning et al. (2025) estimate the average cost of an audit by including expenses such as direct tax officer wages, indirect labor costs (e.g. fringe benefits, training costs, and manager or support staff labor costs), organization-wide costs (e.g. building rents, software, and information technology costs), and central management overhead (e.g. IRS expenses conducted outside of the central audit business unit) in their computation. They first calculate the direct wage costs to conduct an in-person audit, followed by multipliers for each line item by taking the ratio of the total per-audit line item cost (i.e. the total line item cost divided by the number of audits conducted) over the direct per-audit labor cost. Applying these multipliers to the direct wage cost yields estimates of the average per-audit expenditure stemming from each line item.

Our approach to estimating the average cost to conduct an audit draws largely on this logic, but we also account for the fact that not all tax office expenses are attributable to audits of individual income tax returns. Specifically, we want to difference out variable costs associated with revenue-generating activities other than audits of individual income tax returns. Formally, denote the total cost for tax office q as:

$$TC_q = \sum_p C_{p,q} = \sum_p \Big[TFC_{p,q} + TVC_{p,q} \Big] \tag{3}$$

Where $C_{p,q}$ represents the expenditure under budget line item p, which may include fixed and variable components, denoted $TFC_{p,q}$ as total fixed costs and $TVC_{p,q}$ as total variable costs, respectively. Each tax office conducts a set of K activities indexed by k, where each activity represents either a revenue-generating task (e.g. audits of individuals or businesses) or a support function (e.g. information technology, secretary services, overhead expenses). We treat support activities as fixed costs that must exist for any revenue-generating activity to occur. These fixed costs are therefore allocated proportionally across all revenue-generating activities according to the share of officer time devoted to each.[9]

Denote $\tilde{K} \subset K$ as the subset of activities which are revenue-generating. Also denote s_k as the share of tax officer labor hours which are spent conducting activity k, where $\sum_k s_k = 1$. We formally define the total cost of any given revenue-generating activity k in office q as:

$$TC_{k,q} = \frac{s_k}{\sum_{j \in \tilde{K}} s_j} \sum_p C_{p,q} \tag{4}$$

The above formulation is true under the assumption that audit costs are proportional to the number of labor hours it takes to conduct them. If k represents audits of individual income tax returns, the average cost per-audit in office q can then be expressed as:

$$\bar{C}_q = \frac{TC_{k,q}}{A_q} \tag{5}$$

Where A_q is the number of individual audits conducted in office q. In our survey, we asked tax officers the following question: *"About what percentage (%) of your time is spent on the following activities: (1) audits of individual income tax returns, (2) audits of businesses (AOPs)*[10]*, (3) audits of corporations or other*

[9] For example, we assume that if the FBR hypothetically stopped auditing corporations, they would still be able to audit individuals without issue. Whereas if, say, the FBR stopped paying IT employees, this would negatively impact their ability to conduct all other activities. But if a given revenue-generating activity was removed, there would be a lower need for support activity expenses and those would fall proportionally.

[10] AOPs refer generally to businesses which are not corporations.

entities, or (4) non-audit activities." The mean response was that tax officers spend about 24% of their time auditing individuals, 22% of their time auditing AOPs, 22% of their time auditing corporations or other entities, and 32% of their time on non-audit related activities. Therefore, we estimate the total share of annual expenses attributable to audits of individual income tax returns to be $24/68 = 35\%$ of the total budget of each tax office. We calculate the mean cost per-audit separately for each tax office as $\bar{C}_q = 0.35(\sum_p C_{p,q})/A_q$.

Comment (Population representativeness of the sample). Above, we mention that we extrapolate the average share of time attributable to audits of individuals relative to other revenue-generating activities (35%) across all tax offices. One may be concerned that there is heterogeneity in tax officer time-use that varies across tax offices and that this heterogeneity may be correlated with whether a tax officer decided to complete the survey. In other words, there could be an issue that the survey is not representative of the population of tax officers in Pakistan. In order to investigate this, we perform z-tests comparing the expected number of respondents to the observed number of respondents from each tax office. Appendix Table 8 displays the results of this exercise using the share of tax officers from each office in 2021 as our benchmark. We find that, while two tax offices are under represented (Gujranwala and Karachi), and two are over represented (Hyderabad and Sargodha), the distribution of responses approximately matches the observed distribution of officers.

3.3.2 Estimating Audit Cost Heterogeneity

After computing the average cost of an audit within each tax office, we aim to leverage the domain expertise of tax officers to obtain better estimates of costs to audit different types of taxpayers. As we outlined above, we assume the variable cost of an audit lies solely with the number of hours an auditor spends conducting that audit. The second set of questions is therefore designed to elicit how different taxpayer characteristics affect audit duration. Denote $\bar{h}$ as the average number of hours it takes to conduct an audit. We estimate the number of hours to audit individual i as $\bar{h} + \sum_j \delta'_j x_{ij}$, where δ_j is the incremental time required to audit someone with characteristic x_j. We assume that the cost to audit individual i is proportional to the time it takes to audit them. Therefore, we model the expected cost to audit any given individual as:

$$C_i = \bar{C}_q(1 + \frac{\sum_j \delta'_j x_{ij}}{\bar{h}})$$

In order to elicit $\bar{h}$, we ask tax officers the following question: *"On average, how many hours does it take to complete your share of a single audit of an individual income tax return?"* We then share an example which states: *"For example, if you spend 2 hours per-month on a single audit case, and 7 months to complete the case, then it takes you 14 hours to complete the audit case."* The average response was that it takes 17.4 hours to complete the average audit.

Next, we asked tax officers several follow-up questions aimed at deriving δ_j for each of several taxpayer characteristics. We considered 12 variables that met the following criteria: (1) they can be found on income tax returns, and (2) they were deemed "important" when predicting revenue according to VIF score (see section 4), or (3) they were deemed "important" after internal discussions with FBR officials. These variables were: declared business income, declared total income, declared taxable income, turnover, management/administrative expenses, income from property, income from foreign sources, income from agriculture, income from other sources, claimed refunds, brought forward losses declared, and total tax credits claimed.

We first asked respondents the following set of questions: *"We will present you with a series of variables that may be associated with the duration of audits of individual income tax returns to complete. For each variable, please select if, in your experience, this variable is a good predictor of longer/shorter audits. NOTE: we are only considering audits of individual income tax returns here, NOT audits of businesses, corporations, or other entities."* Then, for each of the 12 variables, if the tax official selected "Yes, this is an important predictor of longer-than-average audits," they were given a follow-up question asking for them to select how many more/fewer hours the average audit would take over different reasonable thresholds of that variable. For example, if Declared Taxable Income was selected as an important predictor, they were asked to select how many additional hours the audit would take if Declared Taxable Income took on a value within certain ranges. Appendix Figure 22 displays what the respondent saw regarding this example. If the official selected "No, this is not an important predictor of longer-than-average audits," then they did not receive the follow-up question and the assumed adjustment value is zero. Appendix Figure 23 displays the average values of δ_j for each covariate from these responses. On average, every variable we asked about was assessed by tax officials to be important as a predictor of audit duration.[11] For every variable, declaring larger values tends to lead to relatively longer audits. That is, higher income taxpayers, taxpayers with various sources of substantial income, taxpayers claiming large business losses, and taxpayers claiming large tax credits will all take longer to audit. Several respondents indicated this was (perhaps unsurprisingly) due to the additional verification steps required to validate more income/credits/losses/etc.

4 Machine Learning CATEs and Policy Optimization

This section will discuss our empirical approach to estimating individualized treatment effects and using those estimates to derive optimal audit selection policies following Algorithm 1. In Section 2, we established three conditional average treatment effects (CATEs) that must be estimated in order to calculate θ_i.[12]

$$\begin{aligned}
\tau_1(\boldsymbol{x}) &= E_{\boldsymbol{x}}[R^m \mid \alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[R^m \mid \alpha_i = 0, \boldsymbol{x}] = \Delta R^m(\boldsymbol{x}, \alpha) \\
\tau_2(\boldsymbol{x}) &= E_{\boldsymbol{x}}[R^f \mid \alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[R^f \mid \alpha_i = 0, \boldsymbol{x}] = \Delta R^f(\boldsymbol{x}, \alpha) \\
\tau_3(\boldsymbol{x}) &= E_{\boldsymbol{x}}[C \mid \alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[C \mid \alpha_i = 0, \boldsymbol{x}] = \Delta C(\boldsymbol{x}, \alpha)
\end{aligned} \tag{6}$$

Recall that α_i indicates the audit status of individual i who is of type-$\boldsymbol{x}$. In the plausible scenario where non-audited individuals generate no additional upfront revenue or administrative costs, we can credibly simplify $\tau_1(\boldsymbol{x})$ and $\tau_3(\boldsymbol{x})$ to conditional mean functions (i.e. $E_{\boldsymbol{x}}[\,\cdot \mid \alpha_i = 1, \boldsymbol{x}]$) because the counterfactual potential outcome is known and equal to zero for all audited individuals. Identification of these three causal functions relies on the assumption that the stated potential outcomes are conditionally independent of audit assignment, i.e. $\big(R^m(\alpha_i) \perp \alpha_i\big)|\boldsymbol{x}$. In our study, this assumption is satisfied by construction in 2014 and 2015 because audits were assigned completely at-random. However, in the absence of experimental/quasi-

[11] The share of officials voting that each predictor was important are: Declared Business Income (86%), Declared Total Income (91%), Declared Taxable Income (89%), Turnover (91%), Administrative Costs (90%), Other Income (89%), Property Income (75%), Foreign Source Income (80%), Agricultural Income (72%), Declared Refunds (83%), Declared Loss (89%), and Declared Tax Credits (86%)

[12] In the extension model outlined in Appendix A, there is a fourth causal effect: the taxpayer burden (i.e. the compliance cost of an audit), represented as $\tau_4(\boldsymbol{x}) = E_{\boldsymbol{x}}[B \mid \alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[B \mid \alpha_i = 0, \boldsymbol{x}] = \Delta B(\boldsymbol{x}, \alpha)$. We omit this parameter in our main analysis, but estimating it follows exactly the same process as $\tau_1(\boldsymbol{x})$ and $\tau_3(\boldsymbol{x})$ and is described in Appendix Section A.2.

experimental variation in audit assignment, the observable subset of covariates must be equivalent to the full set of confounding covariates (i.e. $\widetilde{\boldsymbol{x}} = \boldsymbol{x}$) in order to satisfy this assumption.

Our treatment effects of interest are each interpreted as the difference in discounted long-run values of the outcome if audited in the current year vs. if not audited in the current year. To illustrate, let us first consider the CATE on future tax revenue, $\tau_2(\boldsymbol{x})$. In the year 2015 for individual i, discounted long-run tax revenue can be constructed as: $R^f_{i,2015} = T(\hat{z}_{i,2015}) + \beta T(\hat{z}_{i,2016}) + \beta^2 T(\hat{z}_{i,2017}) + ...\beta^k T(\hat{z}_{i,2015+k})$, where $T(\hat{z}_{it}) = T(z_{it}) - e_{it}$ is the collected tax revenue in year t from individual i and β is the government's exogenous discount rate, which we assume to be 5%.[13] Under randomized audits in 2015, the ATE would be identified as τ_2 in the following regression:

$$R^f_{i,2015} = \alpha + \tau_2 \alpha_{i,2015} + \boldsymbol{\gamma'} \boldsymbol{x}_{i,t} + \varepsilon_{i,2015} \tag{7}$$

Where $\boldsymbol{x}_{i,t}$ is a vector of (optional – in the case of Equation 7) individual controls. However, in our application, we do not seek the average impact, τ_2, but rather we seek the conditional average treatment effect (CATE) function which can be expressed as $\tau_2(\boldsymbol{x}_{i,t})$ in the following relaxation of equation 7:

$$R^f_{i,2015} = \alpha + \tau_2(\boldsymbol{x}_{i,t})\alpha_{i,2015} + f(\boldsymbol{x}_{i,t}) + \varepsilon_{i,2015} \tag{8}$$

Here, the treatment effect is now allowed to vary flexibly with $\boldsymbol{x}_{i,t}$. Constructing the other outcomes of interest is the same in principal, and a semi-parametric regression akin to Equation 8 identifies the CATE of interest. However, the process is simplified by the assumption that all recoups, administrative costs, and (in the extension case) compliance costs occur in the current period and are not staggered throughout time (which, for practical purposes, is generally correct in the case of Pakistan). In this case, R^m_i and C_i are simply constructed as the observed values in the current year. Notably, we estimate τ_3 in the manner described in Section 3, not via Equation 8.

A few notes are worth mentioning regarding our practical estimation of Equations 7 and 8. First, we estimate CATE functions using audits from 2014 and 2015 as these are the only observed years in which audits were completely randomized. We use two separate training samples to conduct this estimation. The first sample uses the combined returns and audits of 2014 and 2015 to estimate CATEs. Therefore, the covariate matrix $\boldsymbol{X}_{i,2014+2015}$ contains only variables included in current year tax returns and individuals may be observed twice in training (we cluster standard errors at the individual level in this case). The second sample uses only 2015 audits but with 2014 returns and audit outcomes as additional covariates in $\boldsymbol{X}_{i,2015}$. These two samples have distinct advantages and disadvantages. For the first sample, the possibility of capturing path dependency of CATE predictions is traded off in favor of a larger sample size. The second sample makes the opposite choice: any structural changes to the data generating process between 2014 and 2015 are captured in the estimates, but the statistical power is lower.

We consider four years of tax revenue to construct $R^f_{i,t}$. This is simply due to the fact that we only observe returns up to 2021 and audit outcomes up to 2018, so when we ultimately extrapolate the CATE estimates derived from 2014 and 2015 to future years in our empirical exercise, we can do this through

[13] We include $T(\hat{z}_{i,2015})$ in this measure to allow empirically for deterrence effects to exist in the current period, though in practice this value is expected to difference out in the treatment effect estimate.

2018 and with at most four years of tax payments. Our estimates of long-run revenue impacts will be understated if deterrence persists longer than four years. Prior work has generally found deterrence to be persistent nearly indefinitely (e.g. DeBacker et al. (2018); Boning et al. (2025)). The consequences of this understatement could be substantive for determinance of the optimal audit allocation in two ways. First, by understating the impact of an audit on long-run tax revenue, the optimal policy will mechanically favor (relatively) auditing individuals with a larger initial recoup prediction over those with possibly larger deterrence effects. If the government's discount rate on future tax revenue is sufficiently low, this could lead to a sub-optimal audit allocation relative to the government's preferences. However, if one is willing to extrapolate the four-year post-audit deterrence effect to a greater number of future periods, as prior work would suggest is reasonable, this problem may easily be remedied.[14] Second, if the rate of "deterrence decay" (i.e. the speed at which declared tax revenue reverts to its' prior level in period 1) differs substantially across taxpayers in unobserved years (i.e. in years $(t+3)$ onward), this would not be captured in our CATE estimates and any future-year extrapolation of deterrence effects would be problematic. The same prior research has generally not found substantial heterogeneity in the persistence of deterrence across the income distribution, which helps us alleviate that concern – though further work is warranted on this topic.[15]

Comment (Predictors in $\boldsymbol{X}$). We include a rich set of covariates in the predictive models described in the rest of this section. In the models trained on 2014+2015 data, we use 18 line items on current year income tax returns including total income and taxable income; income from salary, property, capital assets, business, agriculture or other; income from foreign sources; exempted income (e.g. pensions); turnover; business expenditures (e.g. wages, overhead) and profits (gross profits and accounting profits); tax liability; tax credits; and how many days late (or early) the return was filed.[16] For models trained on 2015 data only, we use the same line items on 2015 returns, but we also include the same covariates from the taxpayer's past year return plus variables indicating if they were audited in the past year and how much remittance was paid if they were, giving us 38 covariates in those models.

4.1 CATE Estimation with Causal Forests

Our first approach to mapping the CATE functions of interest is to directly target them via Causal Forests (Athey et al., 2019). Causal Forests are an extension of the Causal Tree (Athey and Imbens, 2016) method developed to non-parametrically search for local regions of treatment effect homogeneity by recursively partitioning the data. In this subsection, we provide a brief primer on the Causal Tree and Causal Forest within our context.

Consider the data generating processes over any of our relevant outcomes (we will use R_i^f for illustration) as a function of audit status:

$$R_i^f = f(\boldsymbol{x}_i) + \tau_2(\boldsymbol{x}_i)\alpha_i + \boldsymbol{\varepsilon}_i$$

[14] Ultimately, we do not conduct this extrapolation in favor of using observed years only in this analysis. However, as we will discuss in Section 5.2, the derived policies will ultimately favor auditing those with large deterrence effects regardless. And therefore, we do not expect this exercise to have a substantial impact on the resulting distribution of audits in this setting.

[15] Another note worth mentioning is that we remove individuals audited more than once from our estimating and out-of-bag samples. This is because we are interested in the causal effect of a single audit rather than a sequence of audits. Doing this drops a negligible 102 total cases from 2014 – 2018 because individuals already under audit were generally excluded from selection as discussed in Section 3.

[16] Most audited taxpayers in Pakistan are business owners as individuals making more than 50% of their income from salary were generally exempt from selection during these years. Business owners must declare their business activities, so many of our variables are applicable to business owners.

Where $f(\boldsymbol{x}_i)$ is a flexible function of the confounders and $\tau(\boldsymbol{x}_i)$ is the CATE function, defined exactly as in Equation 6 as the difference in potential outcomes for observation i:

$$\tau_2(\boldsymbol{x}_i) = E[R_i^f(1) - R_i^f(0)|\boldsymbol{X} = \boldsymbol{x}_i]$$

The high-level idea of the Causal Tree is to leverage matching estimators to identify subgroups with common treatment effects when identification of the ATE itself comes from randomized treatment assignment. To illustrate, consider expressing the ATE of the above data generating process as a matching estimator (we drop the subscript on $\hat{\tau}_2$ going forward to reduce clutter):

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}[R^f(\alpha_i = 1|\boldsymbol{x}_i) - R^f(\alpha_i = 0|\boldsymbol{x}_i)] \\ &= \underbrace{E[R_i^f(1) - R_i^f(0)|\alpha_i = 1]}_{\text{ATE}} + \underbrace{E[R_i^f(0)|\alpha_i = 1] - E[R_i^f(0)|\alpha_i = 0]}_{\text{Selection/Omitted Variable Bias}}\end{aligned}$$

If conditional independence holds, this implies that $E[R_i^f(0)|\alpha_i = 1, \boldsymbol{x}_i] = E[R_i^f(0)|\alpha_i = 0, \boldsymbol{x}_i]$ and $\hat{\tau}$ is unbiased. Intuitively, constructing a matching estimator with $\boldsymbol{x}_i$ works because we are comparing the outcomes of *similar enough* treatment and control units such that treatment assignment is conditionally random among that subset. However, one need not employ matching for the purpose of constructing valid control groups if valid control groups already exist, as in the case of randomized treatment assignment. Instead, the goal of the Causal Tree is to construct alternate subgroups where the covariates in $\boldsymbol{x}_i$ are used to match on treatment effect rather than treatment status probability.

Causal Trees group observations by greedily and recursively splitting the data along covariates, $x_k \in \boldsymbol{x}$, and thresholds, c, into finer and finer sub-samples. Splits occur at "nodes." The resulting partitions from any given split are known as "children nodes" whereas the larger sample which was split is called a "parent node." Further splits are made along children nodes into additional children nodes until stopping criteria are met, in which case we have a "leaf" (or sometimes called a "terminal node").[17]

The optimal split at any given node is determined through a gradient descent process. The first step is to estimate the ATE within the "root" (i.e. first) node, $\hat{\tau}_R$, which is the GMM solution to the identifying local moment condition. In our case, this corresponds to the OLS estimate of the ATE because the conditional independence assumption is the relevant moment condition and we have not made any splits yet:

$$\begin{aligned} E[\alpha_i\varepsilon_i|\boldsymbol{x}_i] &= 0 \\ \implies E[\alpha_i(R_i^f - \alpha_i\tau(\boldsymbol{x}_i))|\boldsymbol{x}_i] &= 0 \\ \implies \hat{\tau}_R &= \frac{\sum_i \alpha_i R_i^f}{\sum_i \alpha_i^2}\end{aligned}$$

To avoid confusion, from here we consider the root node as a generic parent node because the splitting process is the same on all future parent nodes. So we replace the R notation with P to denote that we refer to any parent node.

[17] "Greedy" splitting refers to the process of selecting the optimal split relative to the current node only: never selecting suboptimal splits or reverting back to prior splits. "Recursive" splitting refers to continuous splitting along the parent node, then subsequently along child nodes, and so on until stopping criteria are met.

The second step is to search over covariates and thresholds for the split that maximizes the squared-difference in estimated ATEs between the resulting partitions, C_1 and C_2, which can be described by the objective function:

$$\max_{x_k,c} \quad n_{C_1} n_{C_2} (\hat{\tau}_{C_1} - \hat{\tau}_{C_2})^2 = \sum_{j=1}^{2} \frac{1}{n_{C_j}} \Big(\sum_{i \in C_j} \rho_{i,P} \Big)^2$$

Where $\rho_{i,P} = \frac{\alpha_i (R_i^f - \alpha_i \hat{\tau}_P)}{\frac{1}{n_P} \sum_{i \in P} \alpha_i}$ is the influence function for observation i on the ATE within parent node P. Expressing the objective function as each unit's contribution to the ATE let's us see more clearly the gradient descent approach: maximizing the difference in ATEs across resulting child nodes is equivalent to maximizing the sum of average influence functions across the child nodes.

A key implementation detail is a process Athey and Imbens (2016) refer to as "honest estimation." It is well-known that the consistency and asymptotic normality of non-parametric estimators relies on cross-validation (e.g. see Zheng and van der Laan (2011); Chernozhukov et al. (2018); Yadlowsky et al. (2023); among others). Honest estimation is a method to conduct internal cross-validation by creating two disjoint sub-samples when training a single Causal Tree. The first subsample is used for splitting, while the second subsample populates the nodes/leaves with CATE estimates. Another implementation point is regarding hard-coded splitting rules used to control balance, known as "hyperparameters" in the machine learning literature. A potential problem arising from greedy splitting is that unbalanced splits may occur (i.e. too few treatment or control units in a terminal leaf). These hyperparameters essentially define stopping rules such that a tree knows when to cease further splitting along a branch. Some of the important ones include (1) the minimum number of treatment/control observations in a child node (2) the minimum share of the parent node that must be contained in a child node (3) the share of the training sample used as hold-out for honest estimation (4) an optional imbalance penalty parameter. The default hyperparameter settings in the *grf* package in R (a stable CRAN package written by the original authors of the algorithm) are generally well-optimized for performance and we follow most of these.[18]

Wager and Athey (2018) note that, while point estimates of Causal Trees have beneficial asymptotic qualities, their structure remains quite sensitive to the initial split used to create the training data and they do not produce "smooth" estimates of CATE functions. The Causal Forest addresses these issues by growing B Causal Trees, where each tree is grown on a random sub-sample of the training data[19] and a random sub-sample of the covariate vector $\boldsymbol{X}$. Athey et al. (2019) show that one can describe the CATE function as a weighted-average of the ATE, where the weight on training point i at a given out-of-bag test point $\boldsymbol{x}$, denoted $\omega_i(\boldsymbol{x})$, is the share of total trees where training point i falls into the same terminal leaf as $\boldsymbol{x}$ – a process they refer to as "Generalized Random Forests" (GRF):

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_i^n \omega_i(\boldsymbol{x}) \alpha_i R_i^f}{\sum_i^n \omega_i(\boldsymbol{x}) \alpha_i^2}$$

A key innovation of Athey et al. (2019) is to show that the resulting GRF estimates of the CATE are

[18] Moderate performance improvements may be possible by conducting k-fold cross-validation on various hyperparameter combinations. However, for computational tractability, we do not conduct this exercise.

[19] Readers more familiar with random forests will note that this is slightly different from the bootstrapped sub-samples of the training data used in traditional random forests.

asymptotically normal and consistent, and that one can obtain an estimator of the asymptotic variance through this process. This is notably equivalent to deriving a nonparametric kernel function of the training points that produces consistent CATE estimates on out-of-bag data. To build further intuition of this process, Figure 4 in Appendix C shows a simple, hypothetical example of the forest-training process around a test point $\mathbf{x}$. Panel (a) shows the raw data, where treated units are displayed as blue triangles and control units are displayed as red circles. Three sample trees are trained along the two dimensions of $\boldsymbol{X} = \{X_1, X_2\}$ and are aggregated to a nonparametric kernel shown in Panel (e).

Comment (Residual Learning). It is worth mentioning a technical detail we omitted for simplicity regarding re-centering operations. In practice, Causal Forests actually use residualized observations in training to improve prediction quality. That is, two auxiliary regression trees are typically estimated: one which predicts the conditional mean function, $m(\boldsymbol{x}) = E[R_i^f | \boldsymbol{X}_i]$, and another to predict the propensity score function, $p(\boldsymbol{x}) = E[\alpha_i | \boldsymbol{X}_i]$. These auxiliary forests are trained via leave-one-out cross-validation (i.e. estimates on unit i do not come from trees trained on unit i), denoted with the notation $(-i)$. All subsequent training operations are conducted on Robinson (1988) residualized outcomes and treatment indicators: $\widetilde{R}_i^f = R_i^f - \hat{m}^{(-i)}(\boldsymbol{x})$ and $\widetilde{\alpha}_i = \alpha_i - \hat{p}^{(-i)}(\boldsymbol{x})$.

For each of our outcomes, R^m and R^f, we estimate two Causal Forests. The first uses both 2014 + 2015 tax returns as training data, the second uses 2015 returns only. In the 2015-only model, we use one year of prior tax returns plus the current year as covariates for prediction. However, in the combined 2014 + 2015 model, we use only the current year returns. All forests are trained with 2,000 trees using 50% of the training sample to determine splits and the remaining 50% to populate CATE estimates within each tree. Because propensity scores are known, we provide them to the model rather than estimating an auxiliary forest for $\hat{p}(\boldsymbol{x})$. All covariates are scaled to be between 0 and 1.

4.2 CATE Estimation with Machine Proxy Learners

Our second approach to mapping each CATE of interest is to follow the method established in Chernozhukov et al. (2018). Under randomized audits, generic machine learners can be used to target CATE functions consistently if the CATE function exhibits sufficient sparsity[20] and the models are trained via cross-validation. Unfortunately, sufficient sparsity is a strong and untestable assumption that, if violated, may lead to inconsistent identification from generic machine learning approximations of conditional mean functions. Chernozhukov et al. (2018) provide methods for refining these possibly biased machine learners under mild assumptions so that they may be re-centered along the true CATE function and therefore made unbiased by construction.

To summarize their approach, assume there exists a true CATE function defined as the difference in expected potential outcomes conditional on $\boldsymbol{X}$ for either of our outcomes of interest, $Y \in \{R^m, R^f\}$:

$$s_0(\boldsymbol{X}) := E[Y(1) - Y(0) | \boldsymbol{X}] = m(1, \boldsymbol{X}) - m(0, \boldsymbol{X})$$

Where $m_1 = m(1, \boldsymbol{X})$ and $m_0 = m(0, \boldsymbol{X})$ are conditional mean functions. Assume that propensity scores

[20] Sparsity is an assumption that requires the CATE function to depend meaningfully on a sufficiently small number of dimensions in $\boldsymbol{X}$ despite $\boldsymbol{X}$ being of possibly large dimensions. It essentially requires that the function be well approximated by a relatively simple function of $\boldsymbol{X}$.

are known, bounded away from 0 or 1, and are given by the following for our binary audit indicator, α:

$$p(\boldsymbol{X}) := E[\alpha|\boldsymbol{X}] = P(\alpha = 1|\boldsymbol{X})$$

Predictive ML methods can, in principal, be used to estimate m_1 and m_0, then the CATE is therefore estimated by taking their difference. Define a generic ML estimator of the CATE, for example one from a LASSO, Neural Network, or other, as:

$$S(\boldsymbol{X}) = \hat{m}_1 - \hat{m}_0$$

Chernozhukov et al. (2018) show that, while $S(\boldsymbol{X})$ may not be an unbiased predictor of $s_0(\boldsymbol{X})$, one can obtain an unbiased predictor of $s_0(\boldsymbol{X})$ on the ML proxy predictor by solving the following theoretical equation which defines the best linear predictor of $s_0(\boldsymbol{X})$ from $S(\boldsymbol{X})$:

$$BLP[s_0(\boldsymbol{X})|S(\boldsymbol{X})] := \beta_1 + \beta_2\Big[S(\boldsymbol{X}) - E[S(\boldsymbol{X})]\Big]$$

Where $\beta_1 = E[s_0(\boldsymbol{X})]$ is the ATE and $\beta_2 = \frac{Cov[s_0(\boldsymbol{X}),S(\boldsymbol{X})]}{Var[S(\boldsymbol{X})]}$ is the linear relationship between the ML proxy predictor of the CATE and the true CATE function. Therefore if $\beta_2 \approx 1$, the ML proxy is an approximately unbiased estimator of the CATE on its own. If $\beta_2 \approx 0$, then there is either no heterogeneity or the heterogeneity is poorly identified by the ML proxy. The BLP is therefore a refinement of the generic ML proxy prediction of the CATE function such that it is constructed to be unbiased.[21]

The above theoretical equation can be solved by training ML proxy predictors and estimating the following weighted linear projection:

$$Y = \gamma_0'\boldsymbol{Z}_1 + \gamma_1\Big(\alpha - p(\boldsymbol{X})\Big) + \gamma_2\Big(\alpha - p(\boldsymbol{X})\Big)\Big(S(\boldsymbol{X}) - E[S(\boldsymbol{X})]\Big) + \varepsilon, \quad E[w(\boldsymbol{X})\varepsilon\boldsymbol{Z}] = 0 \tag{9}$$

Where $\boldsymbol{Z}_1 = [1, m_0(\boldsymbol{X}), p(\boldsymbol{X}), p(\boldsymbol{X})S(\boldsymbol{X})']$ and $w(\boldsymbol{X}) = [p(\boldsymbol{X})(1 - p(\boldsymbol{X}))]^{-1}$. Under mild assumptions, $\gamma_1 = \beta_1$ and $\gamma_2 = \beta_2$.

We train three ML proxy predictors for each CATE function of interest. These include LASSO, Gradient-Boosted Trees (XGBoost), and a Neural Network with two hidden layers and L2 regularization.[22] The workhorse ML package is the *mlr3* R package (for the neural network we use the *mlr3torch* extension package), and the *GenericML* R package assists with the construction of the BLP. For all outcomes, we Winsorize the top 5% of values to help guard against the algorithms chasing extreme idiosyncratic observations.[23] For all ML methods, the models are trained to minimize the weighted-residual objective function corresponding to **Definition 5.1 (A)** in Chernozhukov et al. (2018), which they show better centers the ML proxy around

[21] Note that each of these empirical objects and parameters (i.e. $S(\boldsymbol{X})$ and the BLP itself) are dependent on the random split used to define the training sample. There is therefore some randomness induced by the sample itself. This randomness may, however, be eliminated by repeated partitioning and re-training followed by aggregation of the resulting models. We will follow this approach.

[22] For each outcome and each year, we train a feed-forward Neural Network (Multilayer Perceptron). It contains two hidden layers with 64 neurons in the first hidden layer and 32 in the second. This is intended to allow the model to capture a wide array of functional complexities while keeping it somewhat tractable. We use the Rectified Linear Unit (ReLU) activation function, and process 128 mini-batches each pass over 100 epochs. We use the adam optimizer with a 0.01 learning rate, an MSE loss function, and L2 regularization.

[23] We also followed the Battaglini et al. (2024) approach to Winsorizing by estimating the parameters of a Pareto distribution on the top 25% of values, then imputing the conditional mean of the top 5 percentile based on that distribution to all units in the top 5 percentile. Our data is, however, extremely long-tailed and produced degenerate conditional means. We therefore favor traditional Winsorization.

the true CATE function (before post-processing into the BLP), where A denotes the training sample:[24]

$$(\hat{m}_0, S) \in \arg\min_{m_o,s} \sum_{i \in A} w(\boldsymbol{X}_i)[Y_i - m_0(\boldsymbol{X}_i) - \{\alpha_i - p(\boldsymbol{X}_i)s(\boldsymbol{X}_i)\}]^2$$

To perform the BLP exercise, we train each proxy predictor separately for 2014 + 2015 and 2015 only (and of course, separately for each outcome) with repeated 2-fold cross-validation over 100 splits of the training data for hyperparameter tuning. We then estimate the BLP parameters following the above discussion for the best model in each class (in terms of out-of-bag MSE), and apply the refined predictions to the out-of-bag years (2016 – 2018). We one-hot encode factor variables, allowing missingness to be a valid category for encoding. For continuous variables, we impute 0 for missing values and create a second indicator column for missingness.[25] We also scale all variables to be between 0 and 1.

5 Analysis of Socially Optimal Audit Policies

We will now discuss the results of the empirical exercises outlined in Section 4. The first part of this section, Section 5.1, evaluates the quality of each ML learner on out-of-bag data in order to assess how well each algorithm identifies the ΔR^m and ΔR^f causal functions and how well these estimates hold-up in future years. We also aim to establish which learners are generally best at identifying each causal function. Because we can directly observe the ΔR_i^m treatment effects for the sample of taxpayers who were audited each year, we display binscatter plots of the true ΔR_i^m value against each ML learner's prediction of $\Delta \hat{R}_i^m$ in out-of-bag years. For the Causal Forest learners, we also estimate rank-average treatment effects (RATEs) following Yadlowsky et al. (2023) as an omnibus measure of prediction quality on held-out data as well as various calibration statistics. For the other ML learners, we display the results of Equation 9, which helps evaluate how well each proxy predictor identifies the ATE and the CATE. We also estimate Λ and $\bar{\Lambda}$ as goodness-of-fit measures from Equations (3.11) and (3.12) of Chernozhukov et al. (2018) in order to establish the best proxy learners for ΔR^m and ΔR^f, respectively.

The next part of this section, Section 5.2, presents the main results of this paper. We first compare the welfare incidence of the real-world audit regime in each year vs. the expected welfare incidence of each optimal policy according to each ML learner. Next, we dig deeper into the targeted subsamples to understand the characteristics of audited individuals under the observed policy vs. the optimal policies.

5.1 Evaluation of Prediction Quality

5.1.1 RATE and Calibration Estimates

We report rank-weighted average treatment effect (RATE) estimates and Causal Forest calibration results. The RATE serves as a method for evaluating out-of-bag prediction quality of the Causal Forest. It comprises of two components. First, we compute a "Targeting Operator Characteristic (TOC)" – a function

[24] Another option is to solve a similar objective function after conducting a Horvitz-Thompson transform which in some cases may perform slightly better. For our purposes we do not do this. We refer the reader to Definition 5.1 (B) in Chernozhukov et al. (2018) for a detailed discussion.

[25] Tree-based methods (e.g. Causal Forests, XGBoost, and traditional random forests) can handle missing data elegantly by splitting on missingness itself. Most other ML regression learners cannot do this. So in order to avoid dropping incomplete cases or removing variables when training LASSOs or Neural Networks, we allow missing values to take on their own encoding with respect to each variable.

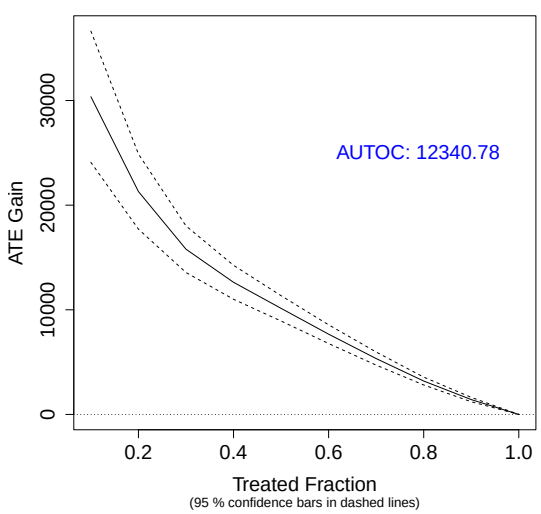
representing the gain over the ATE from ranking units by their CATE prediction and treating sequentially. The TOC function for either outcome $Y \in \{R^m, R^f\}$ can be expressed as the following:

$$TOC(q) = \underbrace{E\left[Y_i(1) - Y_i(0) | \tau(\boldsymbol{x}_i) \geq F^{-1}_{\tau(\boldsymbol{x}_i)}(1-q)\right]}_{\text{Subsample ATE}} - \underbrace{E\left[Y_i(1) - Y_i(0)\right]}_{\text{ATE}} \tag{10}$$

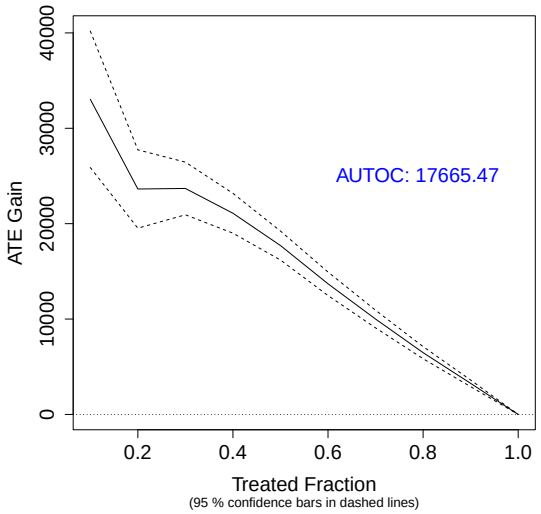
Where q represents a given percentile of the CATE distribution ($F_{\tau(\boldsymbol{x}_i)}$). To illustrate, at $q = 0.1$, the TOC value will be the difference in the ATE from treating the top 10% of the CATE distribution vs. randomly allocating treatment (i.e. the ATE of the full target sample). The second component of the RATE is the total cumulative gain from ranking known as the "Area Under the TOC (AUTOC)," which is simply calculated as $AUTOC = \int_0^1 TOC(q)\, dq$. The best ranking algorithm would maximize this value among all possible ranking algorithms, but any algorithm which is able to capture heterogeneity will at least yield a positive value for this metric because it will generate gains from ranking.

To estimate the RATE, we train two auxiliary causal forests using the training data (i.e. 2014 + 2015 or 2015 only). The first is known as a "test forest," and it is trained on a random 70% subsample of the data. The second is called an "evaluation forest," and it is trained on the remaining 30%. The idea is that the evaluation forest generates fitted values of doubly robust CATEs for the in-sample 30%, which can be aggregated efficiently into a consistent estimate of the ATE in that sample. We then apply the test forest to that sample, ranking the out-of-bag predictions and counterfactually treating sequentially. Figure 1 displays the resulting TOC results for ΔR^m and ΔR^f. These results suggest that both Causal Forests do a good job at identifying heterogeneity in ΔR^m and ΔR^f within the evaluation data. To the extent that the CATE estimates from the evaluation forest serve as a source-of-truth for revenue effects, the test forests generate consistent gains to targeting over the entirety of the treatment effect distribution. This is suggestive evidence that the Causal Forest is able to successfully rank individuals by their individualized CATEs.

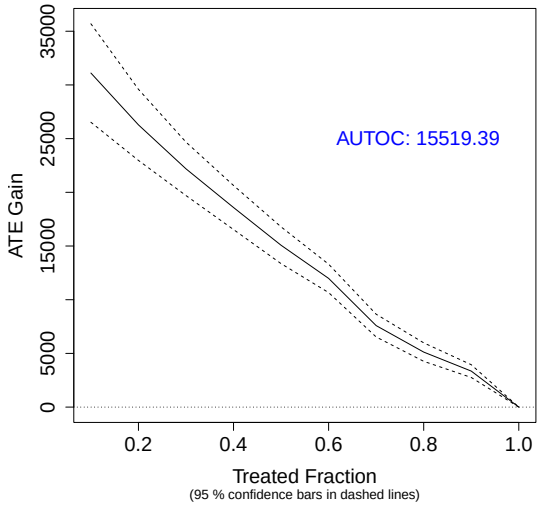
We also display the calibration of each Causal Forest to out-of-bag data in Appendix Table 12. These results are generated with the test_calibration() function of the *grf* R package. The function computes a linear fit between the CATE and the forest predictions plus the mean forest prediction on held-out data. Generally, a coefficient of $\beta_1 \approx 1$ suggests that the forest is able to identify the ATE effectively, and a coefficient of $\beta_2 \approx 1$ additionally suggests that the forest can effectively identify the heterogeneity. It is also a good sign that there *is* heterogeneity in the first place if β_2 is significantly different from 0. We find that the forests are typically well-calibrated, with β_1 and β_2 coefficients very close to 1 for all models except for the Causal Forest of ΔR^f trained on 2015 data only, which has somewhat lower calibration coefficients.



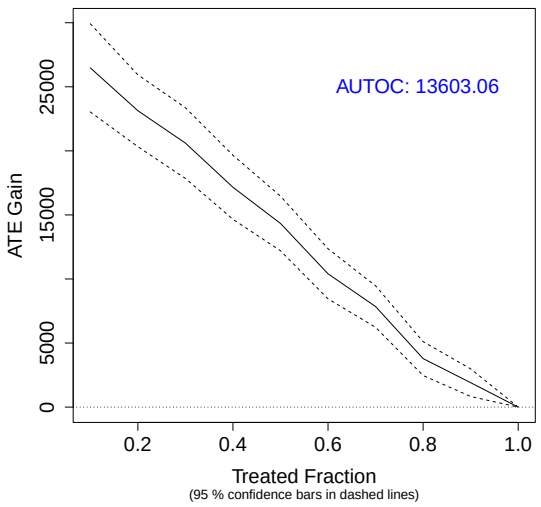
(a) R^m (2014 + 2015)



(b) R^m (2015 only)



(c) R^f (2014 + 2015)



(d) R^f (2015 only)

Figure 1: RATE: Upfront and Long-Run Revenue

Notes: This figure shows the results of the RATE exercise for the revenue-based outcomes. The first Causal Forest (the "test forest") trains on 70% of the training data (e.g. 2014 + 2015 or 2015 only) and the second Causal Forest (the "evaluation forest") trains on the remaining 30%. The test forest makes predictions on the units in the evaluation forest, then ranks them and estimates Equation 10 over each prediction percentile. The solid black line plots the gain over the ATE by treating only the top q% of the distribution. The dashed lines represent confidence intervals of that gain. The AUTOC value represents the RATE metric (the "Area Under the TOC"), which is the percentile-weighed sum of gains.

5.1.2 Best Proxy Learners

Next, we turn to evaluating how well the other proxy predictors fit the data and which proxy predictors are best for each outcome. Table 2 displays the results of Equation 9 as well as estimates of Λ and $\bar{\Lambda}$ which serve as goodness-of-fit measures proposed in Chernozhukov et al. (2018). The coefficient β_1 represents the

debiased ATE on the proxy learner, and a coefficient $\beta_2 \approx 1$ indicates that the proxy learner is close to a consistent estimator of the CATE function before debiasing. Λ is proportional to the R^2 in a regression of $s_0(\boldsymbol{X})$ on $S(\boldsymbol{X})$. Therefore, larger values of Λ indicate a given proxy learner is better for BLP analysis. $\bar{\Lambda}$ is computed by dividing the CATE estimates into quartiles and is proportional to the R^2 of a regression of $s_0(\boldsymbol{X})$ on $\bar{S}(\boldsymbol{X})$ within each quartile (i.e. $lm\{s_0(\boldsymbol{X}) \sim \sum_{k=1}^{4} \bar{S}(\boldsymbol{X})\mathbb{1}(S(\boldsymbol{X}) \in I_k)\}$, where I_k indicates membership of quartile k). Therefore larger values of $\bar{\Lambda}$ are similarly indicative that the proxy learner captures heterogeneity over different segments of the CATE distribution. For a more complete discussion of these parameters, see Chernozhukov et al. (2018). We sort the learners in Table 2 by Λ for each outcome, and our results suggest that the Neural Network is consistently the best predictor of ΔR^f while the XGBoost is consistently the best predictor of ΔR^m.

Table 2: Best ML Proxy Predictors

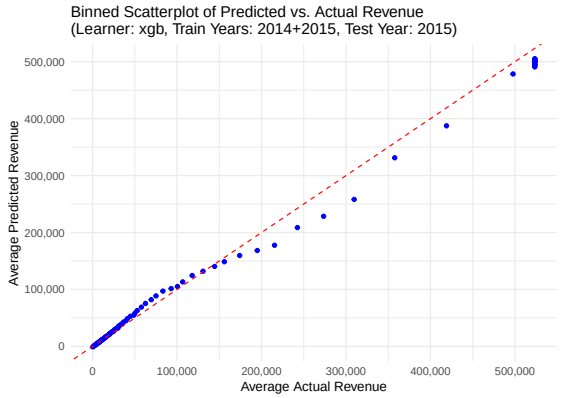
Learner	Train Years	β_1	β_2	SE (β_1)	SE (β_2)	Λ	$\bar{\Lambda}$
R^f							
nnet	2015	162,773.71	1.60	638.42	0.00	37,844.76	45,794.97
nnet	2014 - 2015	163,109.39	1.44	688.97	0.01	27,581.68	121,007.36
xgb	2014 - 2015	43,080.23	0.74	367.89	0.00	9,168.31	9,717.44
lasso	2015	97,436.98	0.77	689.22	0.01	5,006.36	32,440.61
xgb	2015	37,979.33	0.63	370.55	0.00	4,809.80	4,175.77
lasso	2014 - 2015	114,919.02	0.86	767.68	0.01	4,542.47	41,679.86
R^m							
xgb	2014 - 2015	29,732.06	0.88	24.05	0.00	3,018.00	4,040.05
xgb	2015	33,074.20	0.74	27.95	0.00	2,212.75	2,747.55
nnet	2014 - 2015	33,507.78	0.84	35.95	0.00	82.98	1,237.10
lasso	2014 - 2015	33,350.78	0.77	36.02	0.00	59.04	1,239.83
nnet	2015	34,596.95	0.52	36.62	0.00	56.46	1,371.53
lasso	2015	34,178.50	0.63	36.79	0.00	14.54	1,297.15

Notes: This table evaluates the best ML proxy learners for BLP and for heterogeneity. β_1 represents the ATE for the given outcome as estimated by the given proxy learner. β_2 evaluates how well the proxy learner (before debiasing) estimates the CATE function. A perfect proxy predictor would have a $\beta_2 = 1$. Λ and $\bar{\Lambda}$ represent how well the de-biased proxy learner estimates the ATE and the heterogeneity in the treatment effects, respectively. Larger values are considered better for both Λ and $\bar{\Lambda}$.

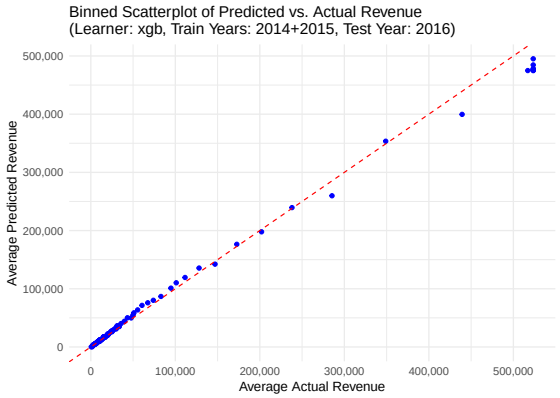
5.1.3 Out-of-Bag Predictions of ΔR^m

Recall from Section 4 that the causal function for the effect of an audit on upfront revenue recoup can be represented as $\tau_1(\boldsymbol{X}) = \Delta R^m = E_{\widetilde{\boldsymbol{x}}}[R^m|\alpha = 1, \boldsymbol{X}] - \underbrace{E_{\widetilde{\boldsymbol{x}}}[R^m|\alpha = 0, \boldsymbol{X}]}_{=0} = E_{\widetilde{\boldsymbol{x}}}[R^m|\alpha = 1, \boldsymbol{X}]$. This means that we can effectively observe the CATE for one of the two key revenue outcomes. Therefore, we can directly evaluate how well each machine learner predicts ΔR^m on out-of-bag data by plotting the observed revenue recoup of audited individuals vs. the predicted recoup. Figure 2 displays a binscatter plot over each percentile of ΔR^m among those observed as audited in each year from 2015 – 2018 vs. the mean prediction of $\Delta \hat{R}^m$ for those individuals from the XGBoost learner (the best ΔR^m learner from Table 2). The first year (2015) shows the in-sample fit of these predictions, while the years 2016 – 2018 are out-of-bag predictions. We find that the XGBoost learner predicts extremely well over the entire distribution of audits, and the quality of these predictions does not decay over time. This is also true for each of the other ML learners. We present similar figures in the Appendix: Figures 9, 10, and 11 for LASSO, Neural Networks, and Causal Forests, respectively. Each learner captures heterogeneity consistently over the distribution of

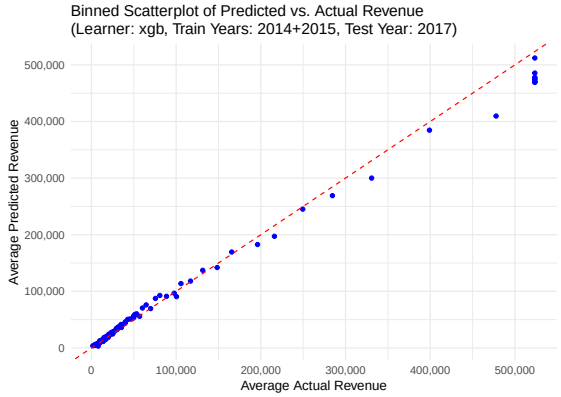
audit magnitudes and over time. Though it is worth noting that the Causal Forest seems to systematically over-predict ΔR^m, despite the fact that it does a good job capturing the heterogeneity pattern.



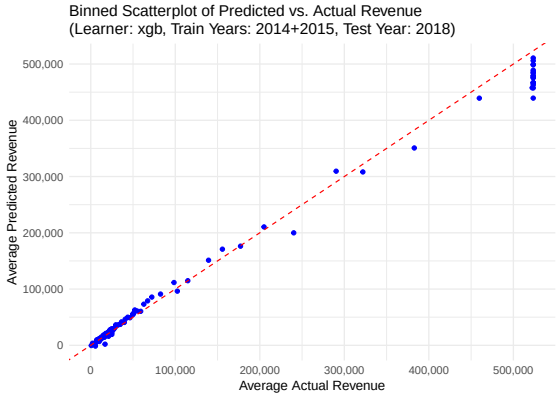
(a) R^m (Year = 2015)



(b) R^m (Year = 2016)



(c) R^m (Year = 2017)



(d) R^m (Year = 2018)

Figure 2: Predicted vs. Observed R^m (XGBoost)

Notes: Each figure shows a binscatter of the mean value of ΔR^m from the observed audit samples in 2015 – 2018 over each percentile, plotted against the predicted values of $\Delta \hat{R}^m$ for units in that percentile according to the de-biased XGBoost proxy predictor algorithm. The dashed red line is a 45-degree line.

The binscatter plots of Figures 2, 9, 10, and 11 only display results for those observed as having a positive value of ΔR^m conditional on them being audited. However, the majority of audits (approximately 60%) in our sample yield no upfront revenue at all. Figures 5, 6, and 7 plot the full distribution of CATE estimates for ΔR^m among those audited for the LASSO, XGBoost, and Neural Network, respectively. We color-code individuals by their audit result: orange for if the audit yielded positive upfront revenue ($\Delta R^m > 0$), blue if it did not ($\Delta R^m = 0$). We find that nearly all individuals where no evasion was detected were placed in the lowest histogram bucket for all three learners. We also plot the histograms of ΔR^f in these figures for the full sample along with (pre-debiased) ATE estimates from each learner. We find that each learner estimates a small, positive effect on the net present value of tax revenue collected over the next three years (i.e. ΔR^f) between $\approx 25,000$ – $140,000$ Rupees ($242 – 1,348 USD). Notably, this finding is slightly larger than the deterrence effect estimates of audits on Pakistani VAT firms from Best et al. (2021), who found that audits led to no tangible deterrence. This suggests that audits may impact firm beliefs and/or behavior differently

than those of individuals.

5.2 Comparison of the Optimal Policies vs. the Observed Policy

Now that we have established that the ML learners are well-calibrated and can each predict well out-of-bag, we follow Algorithm 1 and compute individualized values of θ_i, then derive the optimal audit selection rules. Table 3 displays the result of this exercise for the zero welfare weight scenario. The first column represents various statistics of interest from the real-world policies implemented from 2016 – 2018 in Pakistan. In 2016 (as an illustrative example), audits in Pakistan yielded an average upfront revenue recoup of 23,234 Rupees ($\approx$ \$222 in 2025 USD) and an average effect on discounted future revenue of about the same amount (as estimated by the Neural Network learner). However, we estimate the average administrative cost of audits to be 219,309 Rupees ($\approx$ \$2,096 in 2025 USD) in that year, indicating that the FBR lost about \$1,700 per-audit in their individual income tax return program in 2016.

We also compute the marginal value of public funds (MVPF) and the net social benefit (NSB) for each policy in each year in order to evaluate their welfare incidences. In the zero welfare weight scenario displayed in Table 3, the NSB is just the aggregate revenue/cost ratio and is interpreted as the revenue gain per-dollar of government funds spent on audits under each policy. Larger values of the NSB are therefore strictly superior, and an estimate of λ (discussed in detail in Appendix A) would allow one to directly translate this net-revenue measure to a welfare measure by re-interpreting the NSB value as the welfare gain per-dollar of government expenditure on audits.

We compute the MVPFs in the manner discussed in Appendix A in order to produce a measure comparable to the Boning et al. (2025) MVPF measures in the US. This computation involves setting $\eta_i = 1$ uniformly across all individuals and computing the taxpayer's willingness-to-pay to avoid an audit. This measure is interpreted as the welfare *cost* imposed on society per-dollar of government net-revenue gained through audits. An ideal MVPF in this case is the non-distortionary policy where $MVPF = 1$, however anything less than one indicates a policy which loses money for the government or otherwise costs more to society than is raised in net present value. The MVPF estimate from the real-world policy of -0.904 is therefore akin to an "infinitely poor" MVPF policy, where each dollar (Rupee) of net revenue *lost* imposes a \$0.90 additional welfare cost on taxpayers. We refer the reader to Appendix A for a detailed discussion on MVPF construction and robustness exercises where welfare weights are allowed to vary with income – where we find very similar results.

Table 3: Optimal vs. Observed Policy (zero welfare weights, train years = 2014 – 2015)

		Machine Learners				
Metric	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
2016						
Mean ΔR_i^m	23,233.92	1,232.59	158,762.60	6,410.33	2,616.49	3,805.91
Mean ΔR_i^f	23,308.52	393,949.90	64,394.88	254,429.00	283,183.25	392,868.28
Mean ΔC_i	219,341.60	110,012.21	129,071.16	112,488.92	132,375.99	110,124.21
MVPF	-0.904	1.579	3.058	2.137	2.294	1.576
NSB	0.212	3.592	1.729	2.319	2.159	3.602
Number Audited	30,747	58,822	3,632	57,527	48,885	58,763
2017						
Mean ΔR_i^m	30,366.79	1,184.55	522,179.81	18,132.00	3,247.93	2,166.37
Mean ΔR_i^f	20,061.96	437,748.71	13,424.82	336,834.28	427,906.61	437,159.67
Mean ΔC_i	762,177.17	303,409.71	388,732.67	297,603.01	338,891.57	303,449.44
MVPF	-0.606	4.358	4.970	8.782	6.510	4.350
NSB	0.066	1.447	1.378	1.193	1.272	1.448
Number Audited	9,140	22,251	140	3,829	10,056	22,248
2018						
Mean ΔR_i^m	19,475.66	783.71	538,138.90	38,103.51	1,253.14	2,065.34
Mean ΔR_i^f	6,447.41	443,213.10	22,669.80	409,152.92	439,228.45	442,809.69
Mean ΔC_i	824,331.32	313,884.42	412,450.16	335,681.24	348,945.38	314,041.22
MVPF	-0.549	4.619	5.170	5.513	6.718	4.600
NSB	0.031	1.415	1.360	1.332	1.262	1.417
Number Audited	9,211	23,347	84	1,394	9,926	23,334

Notes: This table computes the results of the real-world observed policy in Pakistan for 2016 – 2018 vs. the predicted values from the optimal derived policies according to each machine learner. For each year, we report the mean revenue recoup from an audit under the observed policy and the mean predicted recoup under the derived policies (ΔR_i^m), the mean estimated deterrence effect under each policy (ΔR_i^f), and the mean estimated administrative cost under each policy (ΔC_i). We also report the computed $WMVPF$ of each policy, the $WNSB$, and the number of audits conducted under each policy as well as the mean welfare weight (uniformly = 1 here) as η_i. The column titled "Optimal Policy (Best Learners)" combines the XGBoost learner for predicting ΔR_i^m with the Neural Network learner for predicting ΔR_i^f.

Our results indicate that every machine learner is able to generate a policy that produces significant welfare gains in expectation over the observed policy, as well as substantially more net revenue for Pakistan. The first "Machine Learners" column displays the expected welfare incidence of a policy which combines the XGBoost learner for ΔR^m and the Neural Network learner for ΔR^f, which serves as our preferred estimate of the optimal audit policy. The optimal policy generates \$3.59 in expected revenue per-dollar spent on audits. Conversely, the observed policy in Pakistan generated only \$0.21 per-dollar spent on audits. In the MVPF framework, this policy imposes \$1.58 in expected welfare costs per-dollar of revenue raised through audits, only slightly more than the comparable estimate in the US of \$1.30 by Boning et al. (2025). Interestingly, the optimal policies designed by nearly every learner (save the Causal Forest, which as demonstrated above seems to over-predict ΔR^m out-of-bag) strongly favor auditing those with large predicted values of $\Delta \hat{R}^f$ over those with large initial revenue recoup. This can be seen by examining the first two rows of each year, where the predicted recoup per-audit is actually quite small in most optimal policies even relative to the observed policy, but the effect on future tax revenue is very large.

It turns out that there is somewhat of a tradeoff between targeting audits based on upfront recoup vs. long run deterrence. Appendix Figure 12 plots binscatters of predicted $\Delta \hat{R}^f$ vs. $\Delta \hat{R}^m$ over percentiles of $\Delta \hat{R}^f$. Every learner estimates an inverse relationship between the two causal functions. While we cannot say for certain why this would be the case, we speculate that there are a large number of audits that miss

or ignore evaded tax revenue in practice. Therefore, it is very possible that individuals who are able to hide some of their evasion got off with informal warnings, recognize that they got lucky, or otherwise did not enjoy the audit process, which results in positive deterrence in future tax revenue despite the audit revealing a small amount of evaded revenue.

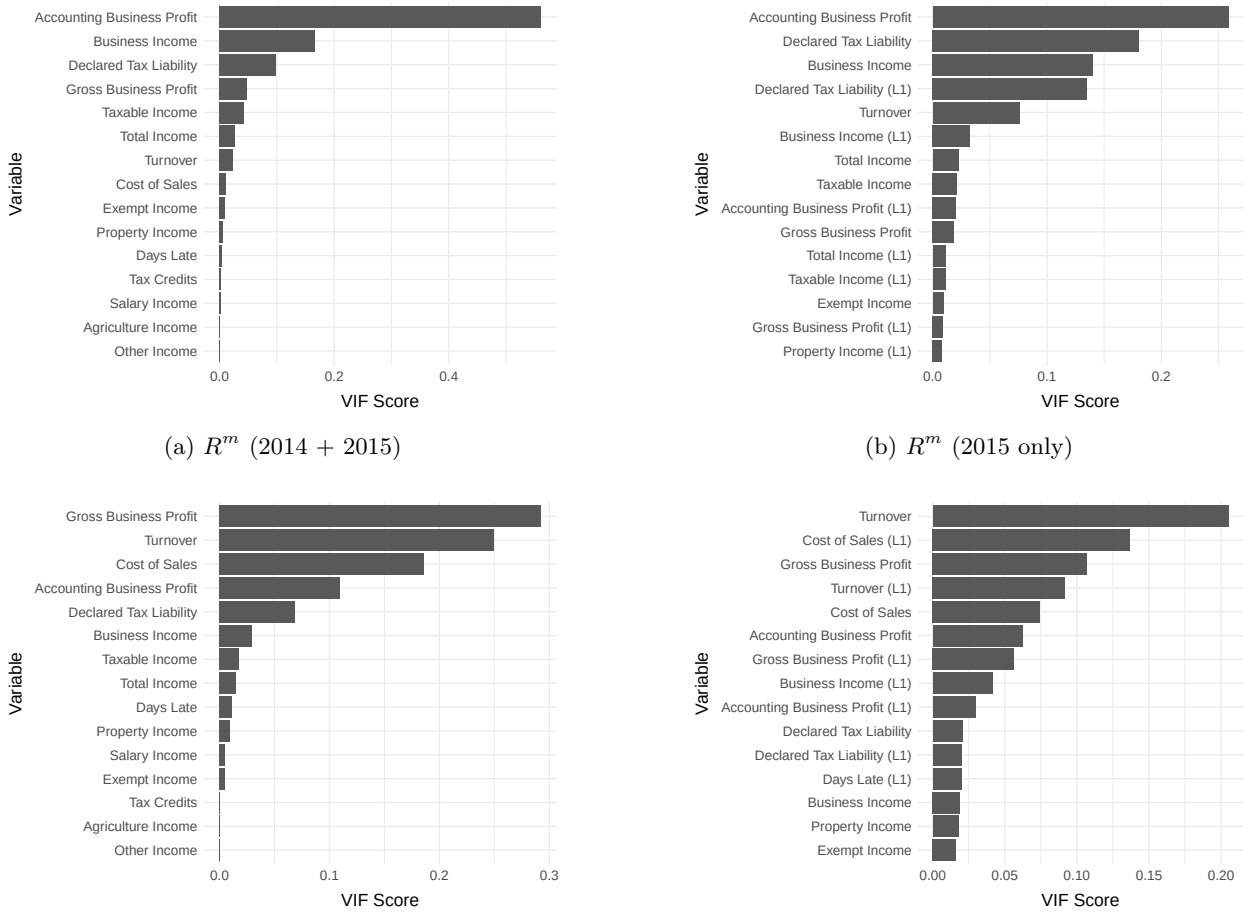
It is important to note that the welfare figures (e.g. the MVPF and NSB) as well as the net-revenue figures are sensitive to our estimates of the mean administrative cost per-audit. As discussed in Section 3, the number of individual audits conducted each year began to fall after 2015, but tax office budgets did not fall proportionally. This was likely due to the FBR diverting resources into other programs, which would imply that our average cost estimates are likely too high in later years (or too low in earlier years). This has no effect on the ML learners' ability to rank individuals by θ_i, but it does mean that one should not compare the revenue or welfare measures across years. These measures may, however, be compared *within* years because these time-varying factors are constant within year.

5.3 Describing the Optimal Audit Samples

Because we have estimated a variety of individualized causal effects along the two main revenue dimensions, R^m and R^f, there is potential to examine how these effects vary along observable covariates. In this subsection, we try to understand *who* is audited under the optimal derived policies and which characteristics are important in predicting these causal effects. First, we present "Variable Importance Factors" (VIF) for the Causal Forest learners in Figure 3. A rudimentary measure of "importance" can be computed when training forest-based algorithms by calculating the share of trees which split along a given covariate, weighted by the depth at which the split occurred so that earlier splits are weighted more heavily. Figure 3 displays these scores for the top 15 most important covariates for each Causal Forest learner. The models trained on only 2015 returns additionally allow us to explicitly examine if prior year tax return covariates were considered important when splitting.

For ΔR^m, the same three variables were by far the most important regardless on if the training sample was 2014+2015 or 2015 only: current year accounting profit/loss, current year business income, and current year tax liability. The 2015-only model also identified the tax liability from the previous year as important. After these income variables, most other tax return line items had relatively low VIF scores in the Causal Forests. Conversely, the CATE function estimates for ΔR^f were sensitive to a somewhat different set of variables. These important covariates included current and prior year turnover, current year gross profit, current and prior year cost of sales, and current year accounting profit.

We examine how $\Delta\hat{R}^m$ and $\Delta\hat{R}^f$ estimates vary with some of these covariates according to each of the machine learners. Appendix Figures 13, 14, 15, and 16 plot the LASSO, XGBoost, Neural Network, and Causal Forest estimates of the average $\Delta\hat{R}^m$ within each percentile of 4 covariates: accounting profit/loss, cost of sales, declared tax credits, and total income. We similarly plot how the estimates of the average $\Delta\hat{R}^f$ change with the same four covariates in Appendix Figures 17, 18, 19, and 20. While all four learners estimate positive relationships between these covariates and $\Delta\hat{R}^m$, the relationship is much more mixed when considering $\Delta\hat{R}^f$. In particular, our preferred $\Delta\hat{R}^f$ learner (the Neural Network) estimates a negative relationship between these covariates and the causal effect of an audit on long-run tax revenue, highlighting the tradeoff between targeting audits based on upfront recoup vs. long-run deterrence.



(a) R^m (2014 + 2015)

(b) R^m (2015 only)

(c) R^f (2014 + 2015)

(d) R^f (2015 only)

Figure 3: Variable Importance: Upfront and Long-Run Revenue

Notes: Each panel shows variable importance scores from the causal forest models trained on each revenue outcome. R^m denotes upfront revenue collected in the year of audit, while R^f denotes the net present value of future revenue collected in subsequent years. The years in parentheses denote the training sample for that forest. For example, (2014 + 2015) indicates the model was trained on both 2014 and 2015 audits. Higher scores indicate greater importance in the model's splitting decisions. Variable names ending with (L1) represent the 1-year lag of that variable. For example, Turnover (L1) is the 2014 tax return value of Turnover for tax returns in 2015. These are applicable only to the 2015 models because we can explicitly include lagged returns as covariates.

We now consider the question of who is audited under alternative audit regimes. Table 4 displays the covariate values of the average audited individual under the real-world policy vs. each optimal policy using 2016 as the out-of-bag illustration year. Keeping in mind the high-VIF covariates, audited individuals under the optimal policies generally have substantially higher values of business profits (both gross and accounting profit), declared tax liability, cost of sales, and turnover. Individuals audited under the optimal policies are also are more likely to declare losses from capital assets and make a smaller share of their income from salary. Covariate missingness also appears to be an important differentiator of target status (and notably, also explains some of the apparent discrepancies in Table 4 such as taxable income being larger than total income among targeted individuals for the "Best Learners" policy). Table 5 shows the share of individuals targeted under each policy who reported missing values for each covariate. Notably, individuals targeted under the ML policies were far more likely to not report their income, turnover, cost of sales, and profits,

and far more likely to report a value of exempt income.

These empirical results are unaffected by the addition of non-zero welfare weights. Appendix Tables 9 and 10 display the same results under non-zero welfare weights, and the same conclusions can be drawn.

Table 4: Targeted Sample Means (Zero welfare weights, train years = 2014 – 2015)

Variable	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
Year: 2016						
Total Income	989,081	1,000,728	1,074,431	844,010	785,852	1,003,779
Salary Income	1,293,377	869,257	952,605	800,278	614,524	871,778
Taxable Income	1,054,493	22,185,538	1,050,327	836,039	702,247	22,019,938
Property Income	806,022	279,450	493,268	367,288	417,001	282,417
Business Income	548,630	361,851	622,482	580,319	411,061	361,351
Capital Asset Income	1,410,654	-283,765	1,114,884	743,575	-197,899	-284,007
Other Income	525,967	193,172	460,045	284,132	421,547	194,680
Foreign Income	943,436	1,512,461	872,082	464,940	1,093,671	1,508,615
Agriculture Income	323,042	179,500	250,142	165,461	396,580	179,506
Exempt Income	4,381,363	137,478,822	5,278,303	1,763,652	1,987,223	138,103,683
Turnover	8,485,991	24,559,268	11,475,814	14,158,423	4,613,971	23,448,177
Cost of Sales	12,060,681	121,212,378	16,154,743	48,732,794	8,675,775	107,473,423
Gross Business Profit	1,508,841	8,920,171	1,941,692	4,466,936	965,564	8,046,468
Accounting Business Profit	787,097	1,072,353	854,281	1,334,840	508,063	1,058,455
Declared Tax Liability	151,114	240,600	153,899	452,552	25,968	242,072
Tax Credits	874,598	45,606	916,342	1,171,528	542,736	50,743
Days Late	72	617	170	1,212	536	616

Notes: This table displays a comparison of covariate means across targeted subsamples with zero welfare weights. The first column reports the average values of various tax return line items among individuals selected for audit under the observed policy. The remaining columns show the same averages for people selected for audit under each derived policy. Note that our data have substantial missingness (these are pre-audit tax returns) even among seemingly important line-items. This accounts for the discrepancy between some of these averages such as Taxable Income vs. Total Income in some cases.

Table 5: Share Missing by Covariate — Targeted (zero welfare weights, train years = 2014–2015)

Variable	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
Year: 2016						
Total Income	9.7%	66.3%	40.0%	63.7%	52.6%	66.0%
Salary Income	95.3%	73.2%	94.4%	83.5%	73.9%	73.3%
Taxable Income	11.3%	66.7%	39.4%	63.6%	53.6%	66.4%
Property Income	88.2%	86.7%	76.6%	91.4%	95.4%	86.6%
Business Income	25.1%	93.3%	54.6%	83.9%	82.2%	92.9%
Capital Asset Income	99.8%	99.7%	99.8%	99.9%	99.9%	99.7%
Other Income	96.2%	96.3%	95.6%	97.0%	96.9%	96.2%
Foreign Income	99.9%	99.8%	99.8%	99.8%	100.0%	99.8%
Agriculture Income	94.4%	90.5%	93.3%	89.6%	90.6%	90.5%
Exempt Income	87.9%	56.1%	90.7%	49.4%	83.4%	56.3%
Turnover	33.3%	94.0%	55.7%	84.5%	83.6%	93.6%
Cost of Sales	59.3%	98.7%	74.7%	95.2%	92.9%	98.5%
Gross Business Profit	50.3%	98.3%	71.0%	94.0%	91.3%	98.1%
Accounting Business Profit	23.7%	93.2%	53.1%	83.1%	82.1%	92.8%
Declared Tax Liability	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Tax Credits	94.0%	93.0%	95.9%	99.4%	97.3%	92.9%
Days Late	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Notes: This table displays a comparison of the share of individuals with missing covariate data across targeted subsamples in the zero welfare weight scenario. The first column reports the share of individuals with missing values over various tax return line items among individuals selected for audit under the observed policy. The remaining columns show the same shares for people selected for audit under each derived policy.

6 Conclusion: How can Governments Incorporate this Framework?

This paper provides a framework for governments to leverage machine learning and welfare theory to derive socially optimal audit allocations. We have shown that audit regimes created under our targeting algorithm generate substantially more net revenue for the government and have much improved welfare incidence in expectation than the real-world targeting policies in Pakistan. A key empirical result of our analysis is that the optimal policies favored auditing those with larger expected deterrence impacts over those with larger immediate recoup predictions – and in-fact there is a tradeoff between the two in this setting. We also show that the optimal policies targeted individuals with greater business profits, cost of sales, tax liability, and turnover, while targeting those declaring capital asset losses and those making less income from salary.

We document that, according to tax officers, audit duration (and therefore administrative expense) generally increases with additional income, business expenses, tax credits, etc. This is consistent with anecdotal evidence that as audit complexity increases, audit duration increases and therefore more resources are dedicated to individual audit cases (i.e. more income sources, expenses, credits, etc. requires additional third party validation). These relationships, while expected, are imprecise with survey data. Going forward, we urge tax enforcement agencies to collect detailed information on per-audit administrative costs. By collecting granular data on individual audit duration through time-sheets or logging, even more refined audit selection policies can be derived with machine learning than those presented here.

We would like to close with a brief discussion on how this framework may be rolled out by governments in a number of audit settings – especially tax enforcement agencies in developing economies. Beginning with the case in question (audit policy in Pakistan), there are a few things to keep in mind. First, audit policy provides a convenient setting for machine learning because for several of the key causal functions, the treatment effect predictions are equivalent to conditional mean function predictions under treatment. As we discussed in the main paper, the causal effect on government costs and short-run revenue from auditing any given individual is equivalent to $E[C_i|W_i = 1, \boldsymbol{X}_i]$ and $E[R_i^m|W_i = 1, \boldsymbol{X}_i]$ because we know that $E[C_i|W_i = 0, \boldsymbol{X}_i] = E[R_i^m|W_i = 0, \boldsymbol{X}_i] = 0$ (i.e. it is costless to *not* audit someone and this will yield no recouped revenue). The same can be said for taxpayer compliance cost in the non-zero welfare weight scenarios presented in Appendix A. The only "treatment effect" which relies on randomization to identify is the aspect of revenue which comes from deterrence. Therefore, this is the only aspect of the optimal policy which cannot be estimated with a general-purpose predictive algorithm and requires quasi-experimental or experimental variation.

The simplest solution is to extrapolate the $\Delta\hat{R}^f$ causal functions estimated in randomized audit years, under the assumption that the CATE function identified is relatively unchanged over time. This is, of course, a strong assumption to maintain on its own despite the fact that we present evidence that the CATE functions estimated on $\Delta\hat{R}^m$ do not lose efficacy over the 3 years of out-of-bag data we test. We recommend that Pakistan (and any other tax authority wishing to replicate these results) maintain a subset of audits conducted at-random among a representative sample of individuals so that these causal functions can be recalibrated over time. There is precedent for these types of programs. For example, several papers investigate deterrence leveraging the random audit programs conducted by the IRS (DeBacker et al., 2018)

and the HMRC (Gemmell and Ratto (2012), Advani et al. (2023)) in the US and UK, respectively.

References

Advani, A., W. Elming, and J. Shaw (2023, May). The Dynamic Effects of Tax Audits. *The Review of Economics and Statistics 105*(3), 545–561.

Allingham, M. G. and A. Sandmo (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics 1*.

Ash, E., S. Galletta, and T. Giommoni (2024). A Machine Learning Approach to Analyze and Support Anti-corruption Policy. *American Economic Journal: Economic Policy*.

Athey, S. and G. Imbens (2016, July). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360. Publisher: Proceedings of the National Academy of Sciences.

Athey, S., J. Tibshirani, and S. Wager (2019, April). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178. Publisher: Institute of Mathematical Statistics.

Battaglini, M., L. Guiso, C. Lacava, D. L. Miller, and E. Patacchini (2024, September). Refining public policies with machine learning: The case of tax auditing. *Journal of Econometrics*, 105847.

Benitez, J. C., M. Mansour, M. Pecho, and C. Vellutini (2023, September). Building Tax Capacity in Developing Countries.

Bergstrom, K., W. Dodds, and J. Rios (2025, July). Optimal Policy Reforms. *Working Paper*.

Besley, T. and T. Persson (2014, November). Why Do Developing Countries Tax So Little? *Journal of Economic Perspectives 28*(4), 99–120.

Best, M., J. Shah, and M. Waseem (2021). Detection Without Deterrence: Long-Run Effects of Tax Audit on Firm Behavior. *Working Paper*.

Boning, W. C., N. Hendren, B. Sprung-Keyser, and E. Stuart (2025, February). A Welfare Analysis of Tax Audits Across the Income Distribution. *The Quarterly Journal of Economics 140*(1), 63–112.

Caspi, A., J. Goldin, D. Reck, and D. E. Ho (2024). Optimal Tax Audits Using Predictions. *Working Paper*.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018, June). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. *NBER Working Paper Series No. 24678*.

Chetty, R. (2006, December). A New Method of Estimating Risk Aversion. *American Economic Review 96*(5), 1821–1834.

Dantzig, G. B. (1957). Discrete-Variable Extremum Problems. *Operations Research 5*(2), 266–277. Publisher: INFORMS.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2018, February). Once Bitten, Twice Shy? The Lasting Impact of Enforcement on Tax Compliance. *The Journal of Law and Economics 61*(1), 1–35.

Dom, R., A. Custers, S. Davenport, and W. Prichard (2022, February). *Innovations in Tax Compliance: Building Trust, Navigating Politics, and Tailoring Reform*. The World Bank.

Eissa, N. and H. Hoynes (2011, June). Redistribution and tax expenditures: the earned income tax credit. *National Tax Journal 64*(2.2), 689–729. Publisher: The University of Chicago Press.

Eissa, N., H. J. Kleven, and C. T. Kreiner (2008, April). Evaluation of four tax reforms in the United States: Labor supply and welfare effects for single mothers. *Journal of Public Economics 92*(3), 795–816.

FBR (2015). Audit Policy 2015. Federal Board of Revenue, Pakistan, Taxpayer's Audit Wing.

García, J. L. and J. J. Heckman (2022, March). On Criteria for Evaluating Social Programs. *Institute of Labor Economics (IZA)*.

Gemmell, N. and M. Ratto (2012, March). Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal 65*(1), 33–57. Publisher: The University of Chicago Press.

Guyton, J. and R. H. Ii (2021). The Compliance Costs of IRS Post-Filing Processes. *An IRS-TPC Research Conference: Advancing Tax Administration*.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010, February). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1), 114–128.

Hendren, N. (2016). The Policy Elasticity. *Tax Policy and the Economy 30*, 51–89. Publisher: [The University of Chicago Press, The National Bureau of Economic Research].

Hendren, N. and B. Sprung-Keyser (2020). A Unifed Welfare Analysis of Government Policies. *Quarterly Journal of Economics 135*(3), 1209–1318.

Hendren, N. and B. Sprung-Keyser (2022, May). The Case for Using the MVPF in Empirical Welfare Analysis. *NBER Working Paper Series*.

Holtzblatt, J. and J. McGuire (2020). Effects of Recent Reductions in the Internal Revenue Service's Appropriations on Returns on Investment. *IRS Research Bulletin: Proceedings of the 2020 IRS/TPC Research Conference*.

Jensen, A. (2022, January). Employment Structure and the Rise of the Modern Tax System. *American Economic Review 112*(1), 213–234.

Kaldor, N. (1963). Taxation for Economic Development. *The Journal of Modern African Studies 1*(1), 7–23. Publisher: Cambridge University Press.

Kaplow, L. (1990, November). Optimal taxation with costly enforcement and evasion. *Journal of Public Economics 43*(2), 221–236.

Keen, M. and J. Slemrod (2017, August). Optimal tax administration. *Journal of Public Economics 152*, 133–142.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark. *Econometrica 79*(3), 651–692. Publisher: [Wiley, Econometric Society].

Knittel, C. R. and S. Stolper (2025, May). Using Machine Learning to Target Treatment: The Case of Household Energy Use. *The Economic Journal*, ueaf028.

Loayza, N. V. (1996, December). The economics of the informal sector: a simple model and some empirical evidence from Latin America. *Carnegie-Rochester Conference Series on Public Policy 45*, 129–162.

Mayshar, J. (1990, December). On measures of excess burden and their application. *Journal of Public Economics 43*(3), 263–289.

Mayshar, J. (1991). Taxation with Costly Administration. *The Scandinavian Journal of Economics 93*(1), 75–88. Publisher: [Wiley, The Scandinavian Journal of Economics].

Olken, B. (2007, April). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy 115*(2), 200–249. Publisher: The University of Chicago Press.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica 56*(4), 931–954. Publisher: [Wiley, Econometric Society].

Ruan, J., Z. Yan, B. Dong, Q. Zheng, and B. Qian (2019, March). Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. *Information Sciences 477*, 508–532.

Saez, E. and S. Stantcheva (2016, January). Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review 106*(1), 24–45.

Sarin, N. and L. H. Summers (2019, November). Shrinking the Tax Gap: Approaches and Revenue Potential.

Slemrod, J. and S. Yitzhaki (1996). The Costs of Taxation and the Marginal Efficiency Cost of Funds. *Staff Papers (International Monetary Fund) 43*(1), 172–198. Publisher: Palgrave Macmillan Journals.

Sun, H., E. Munro, G. Kalashnov, S. Du, and S. Wager (2024, March). Treatment Allocation under Uncertain Costs. arXiv:2103.11066 [stat].

Wager, S. and S. Athey (2018, July). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association 113*(523), 1228–1242. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1319839.

Waseem, M. (2023, February). Overclaimed refunds, undeclared sales, and invoice mills: Nature and extent of noncompliance in a value-added tax. *Journal of Public Economics 218*, 104783.

Wu, Y., B. Dong, Q. Zheng, R. Wei, Z. Wang, and X. Li (2020, July). A Novel Tax Evasion Detection Framework via Fused Transaction Network Representation. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 235–244. ISSN: 0730-3157.

Yadlowsky, S., S. Fleming, N. Shah, E. Brunskill, and S. Wager (2023, November). Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects. arXiv:2111.07966 [stat].

Zheng, W. and M. J. van der Laan (2011). Cross-Validated Targeted Minimum-Loss-Based Estimation. In M. J. van der Laan and S. Rose (Eds.), *Targeted Learning: Causal Inference for Observational and Experimental Data*, pp. 459–474. New York, NY: Springer.

A Appendix: Extension – Non-zero welfare weights

In this Appendix, we provide an extension to the baseline model defined in Section 2 and allow for welfare weights to be non-zero. Allowing for non-zero welfare weights introduces complexity to the analysis by requiring estimates of additional parameters in order to construct individualized θ_i values and follow Algorithm 1. Namely, we require two additional non-trivial parameter estimates. First, we need the willingness-to-pay term, denoted WTP_i, which represents individual i's willingness-to-pay for an audit (or conversely, the negative of the WTP_i is their willingness-to-pay to *avoid* an audit). Second, we require an estimate of the welfare-weighted Marginal Value of Public Funds ($WMVPF$) of the numeraire policy (denoted λ) which closes the budget. This extension will focus mainly on deriving the parameters making up the WTP_i term. But let us begin with a brief discussion of the $WMVPF_\lambda$ term. The $WMVPF_\lambda$ is defined as the ratio of society's willingness-to-pay for a marginal increase to λ from the status quo level of λ (denoted $\tilde{\lambda}$) over the net effect on the government's budget which results from that change (Saez and Stantcheva, 2016):

$$\begin{aligned} WMVPF_\lambda &\equiv \frac{dW/d\lambda}{dNC/d\lambda}\Big|_{\tilde{\lambda}} \\ &= \frac{\sum_{i=1}^{N} \psi(\boldsymbol{x}) dU_i(\cdot)/d\lambda}{dNC/d\lambda}\Big|_{\tilde{\lambda}} \\ &= \bar{\eta}\frac{\sum_{i=1}^{N} WTP_i(\boldsymbol{x})}{dNC/d\lambda}\Big|_{\tilde{\lambda}} \\ &= \bar{\eta} MVPF_\lambda \end{aligned} \tag{11}$$

Where $\frac{dNC}{d\lambda}|_{\tilde{\lambda}}$ refers to the change in net cost (i.e. additional upfront mechanical spending less any fiscal externalities which may affect tax revenue collected) which results from increasing the numeraire policy by $d\lambda$ from $\tilde{\lambda}$. Of course, without knowledge of what λ actually is, we cannot obtain an estimate of its $WMVPF$. In developed countries, the literature typically assumes this policy to be linear income taxation (Bergstrom et al., 2025). However, this is likely not the case in a developing country. Going forward, we will think of this policy as a non-distortionary lump sum transfer: for example, a Universal Basic Income (UBI) with no income effects, in that revenue generated from audits is shared roughly equally across the population. We acknowledge however, that the selection of this policy (or vector of policies) is non-trivial to the ultimate estimate of $WNSB_i$ values and that better knowledge of the ongoing transfer policies of the Pakistani federal government is needed to obtain better audit selection policies in the non-zero welfare weight case.

Allowing for λ to be a non-distortionary transfer implies that the $WMVPF_\lambda = 1$ if the η_i values are normalized to integrate to 1. We assume that the η_i values are given by $\eta_i = z_i^{-\gamma}$, where γ is a parameter governing income inequality aversion which we will allow to equal 1 or 2.[26] Loosely, a value of $\gamma = 1$ corresponds to a utilitarian policymaker and utility over consumption equal to $\log(y)$, with larger values of γ indicating greater inequality aversion (Chetty, 2006). z_i corresponds to our estimate of true labor income. In our sample, we observe each taxpayer's declared taxable income and declared tax liability: $T(\hat{z}_i)$. Recognizing that their true tax liability is $T(z_i) = T(\hat{z}_i) + e_i$, we can plug in our estimate of e_i to obtain an estimate of $T(z_i)$. From here, we can invert the Pakistani income tax formula (which is a standard piecewise

[26] We normalize by selecting the normalization constant, c, which is the mean true income. This is found by setting $\int \eta(z) f(z) dz = 1$, where $f(z)$ is the distribution of income. This implies $\int c z^{-\gamma} f(z) dz = 1 \implies c \int z^{-\gamma} f(z) dz = 1 \implies c = (E[z^{-\gamma}])^{-1}$. So $\eta(z) = \frac{z^{-\gamma}}{E[z^{-\gamma}]}$ in practice.

marginal tax rate schedule) to obtain an estimate of their true labor income: $z_i = \frac{T(z_i) - Base_b}{MTR_b} + LB_b$. Where $Base_b$ is the cumulative tax rate paid until reaching the bracket b that the taxpayer should be in based on $T(z_i)$, MTR_b is the marginal tax rate of bracket b, and LB_b is the minimum income (lower bound) of tax bracket b. For example, if $T(\hat{z}_i) = 20,000$ and $e = 165,000$, then $T(z) = 185,000$. This income tax burden of 185k falls into the 4th tax bracket, where the cumulative tax paid from brackets 1-3 is $Base_4 = 180,000$. The $MTR_4 = 0.25$, and the lower bound of income in bracket 4 is 2.2M Rupees. So our estimate of true income is $z = \frac{185k - 180k}{0.25} + 2,200,000 = 2,220,000$. All of this allows us to abstract away from computing a WMVPF. However, we recognize that this involves some fairly strong assumptions. Firstly, welfare weights can (and probably should) be based on a deeper set of personal characteristics than income. We do not see this as a major limitation in our case because, like previously mentioned, if they are normalized to integrate to 1 then the WMVPF of a non-distortionary lump-sum transfer will equal 1. Secondly, the assumption that the numeraire policy is a UBI policy with no income effects is strong. In practice, a policymaker with more knowledge of Pakistan's federal redistribution programs can improve our procedure by inserting a WMVPF estimate more reflective of reality.

A.1 The Taxpayer's Problem

Now we will move to uncovering the WTP_i components needed to construct the rest of the θ_i. That is, we need to understand what individualized causal functions comprise the WTP_i. For this, we turn to the taxpayer's problem. Our model will proceed building upon the optimal tax administration literature of Keen and Slemrod (2017), Caspi et al. (2024), and especially Boning et al. (2025).

Taxpayers aim to maximize their expected net present value of utility by selecting true labor income (z_{it}), and their evaded tax liability (e_{it}) in each period $t \in [0, \infty)$. Recall that y_{it} represents consumption. Also recall that $\hat{a}_{it}(e_{it}; \boldsymbol{\alpha}_{t-k}, \widetilde{\boldsymbol{x}})$ represents person i's ex-ante perceived audit risk if they evade e_{it} dollars of tax liability and have experienced $\boldsymbol{\alpha}_{t-k}$ as the historical audit environment over the past k periods. Denote $\boldsymbol{\alpha}_i = (\alpha_{i0}, ..., \alpha_{i\infty})$ as the dynamic set of audit realizations the taxpayer may face in the future. In order to make progress, we assume quasi-linear utility. We also assume taxpayers are risk neutral and there are no psychic costs to evading (though we do allow for psychic costs to being audited). Under these assumptions, the utility function can be stated as $U_i(\cdot) = y_{it} - \xi(z_{it}) - (B_{it} + \phi(e_{it})) \cdot \alpha_{it}$, where $\xi(\cdot)$ represents dis-utility of labor, $\phi(e_{it})$ represents the monetary penalty from (being caught) evading e_{it}, and B_{it} represents the monetary and psychic cost of compliance with an audit, also known as "taxpayer burden." B_{it} includes expenses such as opportunity cost of lost labor hours, cost of hiring accountants/lawyers, etc. With this, we can state the Taxpayer's Problem as:

Taxpayer's Problem

$$\begin{aligned} \max_{z_{it}, e_{it}} \quad & E_{\hat{a}}\left[\textstyle\sum_{t=1}^{\infty} \beta^{t-1} U_i(y_{it}, z_{it}, e_{it}; \hat{a}_{it})\right] \\ \text{s.t.} \quad & U_{it} = z_{it} - T(z_{it}) + e_{it} - \xi(z_{it}) \qquad \text{if } \alpha_{it} = 0 \\ & U_{it} = z_{it} - T(z_{it}) - \phi(e_{it}) - \xi(z_{it}) - B_{it} \qquad \text{if } \alpha_{it} = 1 \end{aligned} \tag{12}$$

Because of the linear separability of the problem, the choice of labor income z_{it} and evaded tax liability e_{it} are independent, and can therefore be considered individually. Plugging the constraints into the objective function in Equation 12 yields the taxpayer's problem for the choice of evaded tax liability:

$$\max_{e_{it}} \quad \sum_{t=1}^{\infty} \beta^{t-1} \Big[\hat{a}_{it} \underbrace{\Big(z_{it} - T(z_{it}) - \phi(e_{it}) - \xi(z_{it}) - B_{it}) \Big)}_{\text{Utility if audited}} + (1 - \hat{a}_{it}) \underbrace{\Big(z_{it} - T(z_{it}) + e_{it} - \xi(z_{it}) \Big)}_{\text{Utility if not audited}} \Big] \tag{13}$$

The optimal evasion amount in each period, e^*_{it}, will ignore all terms where e_{it} does not enter, and therefore will solve in each period:

$$\max_{e_{it}} \quad (1 - \hat{a}_{it}) e_{it} - \hat{a}_{it} \phi(e_{it}) \tag{14}$$

The solution follows as a general result of first-order conditions for quasi-linear utility:

$$(1 - \hat{a}_{it}) = \hat{a}_{it} \frac{d\phi}{de_{it}}\Big|_{e^*_{it}} \implies e^*_{it} = \Big(\phi'(e_{it}) \Big)^{-1} \Big(\frac{1 - \hat{a}_{it}}{\hat{a}_{it}} \Big) \tag{15}$$

Plugging in the Marshallian values of evasion and labor income into utility, we may obtain the indirect utility function for any given period t:

$$\widetilde{U}_{it}(\alpha_{it}) = z^*_{it} - T(z^*_{it}) - \xi(z^*_{it}) + (1 - \alpha_{it}) e^*_{it} - \alpha_{it}(\phi(e^*_{it}) + B_{it}) \tag{16}$$

Denote V^1_i and V^0_i as the net-present value of utility in *future* periods if $\alpha_{it} = 1$ or 0, respectively. The full expected indirect utility function can be stated as:

$$V_i(\alpha_{it}) = \alpha_{it}[\widetilde{U}_{it}(1) + V^1_i] + (1 - \alpha_{it})[\widetilde{U}_{it}(0) + V^0_i] \tag{17}$$

A.1.1 Willingness to Pay to Avoid an Audit

We are now able to derive taxpayer i's willingness-to-pay to avoid an audit in the first period, $t = 1$. Individuals are willing to pay the value of their compensating variation: the amount of money from their own income such that their utility is equated between the event of an audit vs. no audit. This term is implicitly defined as $V_i(1) + WTP_i = V_i(0) \implies WTP_i = V_i(0) - V_i(1)$. This can be written as:

$$\begin{aligned} WTP_i = V_i(0) - V_i(1) &= \widetilde{U}_{i1}(0) + V^0_i - \big(\widetilde{U}_{i1}(1) + V^1_i \big) \\ &= \underbrace{e^*_{i1} + \phi(e^*_{i1})}_{\text{Mechanical Recoup: } \Delta R^m_i} + \underbrace{(V^0_i - V^1_i)}_{\text{Long-run tax revenue: } \Delta R^f_i} + B_{i1} \end{aligned} \tag{18}$$

To see that $V^0_i - V^1_i = \Delta R^f_i$, let's expand the indirect utility of individual i in period $t = 2$ as a function of their audit status in period $t = 1$:

$$\widetilde{U}^{\alpha_1}_{i2} = z^*_{i2} - T(z^*_{i2}) - \xi(z^*_{i2}) + (1 - \alpha_{i2}) e^{*,\alpha_1}_{i2} - \alpha_{i2}(\phi(e^{*,\alpha_1}_{i2}) + B_{i2}) \tag{19}$$

The effect on their indirect utility in period $t = 2$ as a result of an audit is simply the difference in total tax revenue collected by the government from reduced evasion:

$$\begin{aligned} \widetilde{U}^0_{i2} - \widetilde{U}^1_{i2} &= (1 - \alpha_{i2}) e^{*,0}_{i2} - \alpha_{i2}\Big(\phi(e^{*,0}_{i2}) + B_{it} \Big) - (1 - \alpha_{i2}) e^{*,1}_{i2} - \alpha_{i2}\Big(\phi(e^{*,1}_{i2}) + B_{it} \Big) \\ &= e^{*,0}_{i2} - e^{*,1}_{i2} - \alpha_{i2}\Big(e^{*,0}_{i2} - e^{*,1}_{i2} + \phi(e^{*,0}_{i2}) - \phi(e^{*,1}_{i2}) \Big) \end{aligned} \tag{20}$$

Notably, in the above equation, $e_{i2}^{*,0} - e_{i2}^{*,1} = T(z_{i2}^{*,0}) - T(\hat{z}_{i2}^{*,0}) - T(z_{i2}^{*,1}) + T(\hat{z}_{i2}^{*,1})$. Because we know that the labor income decision is independent of the evasion decision, we know that $T(z_{i2}^{*,0}) = T(z_{i2}^{*,1})$ and therefore the difference in evasion is captured by $e_{i2}^{*,0} - e_{i2}^{*,1} = T(\hat{z}_{i2}^{*,1}) - T(\hat{z}_{i2}^{*,0})$.

Also, in our empirical setting as we discuss in the main paper body, it was incredibly rare that individuals were audited again in periods after their original audit. Because we are interested in the optimal audit selection regime in the current period only rather than in future periods, and because in our empirical setting it was typically the case that audited individuals were not selected again, we treat α_{i2} as if it were equal to 0. What this means is that we can entirely attribute the long-run effect of an audit in period $t = 1$ on the net-present value of utility as simply the causal effect on declared tax revenue over time:

$$V_i^0 - V_i^1 = \sum_{t=2}^{\infty} \beta^{t-1} \left[T\left(\hat{z}_{it}^{\alpha_1=1}\right) - T\left(\hat{z}_{it}^{\alpha_1=0}\right) \right] = \Delta R_i^f \tag{21}$$

A.2 Estimating Taxpayer Burden

Now that we have derived the relevant causal functions in the willingness-to-pay term, we see that there is one additional CATE that needs to be estimated: a value we refer to as the "taxpayer burden." Taxpayer burden is interpreted as the cost of time and effort (in dollars) imposed on the taxpayer by means of producing paperwork or other documents related to the audit itself, hiring lawyers and/or accountants, or any other monetary expenses associated with complying to the audit itself (as well as the dollar-valued psychic costs of audit compliance). This value is not observable in our data, and will likely vary primarily according to how imposing the assessor is on the taxpayer. In order to impute this value, we translate estimates from a survey conducted in the US by Guyton and Ii (2021) to real Pakistani Rupees (PKR). They estimate an average taxpayer burden on audited individuals of \$3,198. They also estimate a regression akin to $B_i = \alpha + \boldsymbol{\gamma}' \boldsymbol{X_i} + \epsilon$, where $\boldsymbol{X}_i$ is a vector of survey variables. They estimated a coefficient on log(income) to be 0.18, indicating the total burden increases by 0.18% on-average with a 1% increase in income. We extrapolate this estimate across the income distribution to generate (very moderate) heterogeneity in taxpayer burden estimates as a function of income. We also follow the approach taken by Boning et al. (2025), who also refer to this survey. They estimated an average B/C ratio of \$0.50, and chose to hold this constant for all individuals. That is, for an individual with an estimated audit cost of \$1,000, they impose a burden estimate of \$500. We follow their approach and assume a B/C ratio of \$0.50, which generates most of the heterogeneity in taxpayer burden. We elect to maintain the Boning et al. (2025) B/C ratio estimate because, in our view, it requires stronger assumptions to extrapolate the average burden estimated in the U.S. to taxpayers in Pakistan. The Guyton and Ii (2021) estimate of \$3,198 mean burden translates to $520,085$ PKR in 2021. If adjusted for inflation, this value drops to $336,389$ PKR in 2014. Our estimate of the average cost to conduct an audit in 2014 is only $140,800$ PKR. If we were to extrapolate the US survey results to Pakistan, we would estimate a B/C ratio of 2.39 by dividing the average burden estimate by the average audit cost. We feel this value likely overestimates the true compliance cost for the average Pakistani taxpayer and elect to follow the Boning et al. (2025) approach. However, we recognize that further research is warranted to better understand this value in developing countries.

A.3 Optimal Policy Results with Non-Zero Welfare Weights

In this section we display results akin to the main paper results in Table 3 allowing for non-zero welfare weights. Table 6 shows the optimal policies when $\gamma = 1$, which recall is approximately equivalent to a utilitarian policymaker where individuals have utility over consumption equal to $\log(y)$. Table 7 presents the same but introduces a greater degree of inequality aversion with $\gamma = 2$. Notably, the "best" values of the WMVPF welfare criterion changes slightly when welfare weights can vary. Specifically, we seek values close to (but not below) 0 in this case. Most ML learners unsurprisingly favor auditing those with very small welfare weights, and many of the same trends emerge. For instance, it is generally still preferable to audit those with larger impacts on future tax revenue than those with larger predicted initial recoups. All learners are able to generate substantial welfare gains over the observed policies.

Table 6: Optimal vs. Observed Policy ($\gamma = 1$, train years = 2014 – 2015)

		Machine Learners				
Metric	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
2016						
Mean ΔR_i^m	23,233.92	1,571.95	254,737.25	7,892.82	3,200.44	5,157.78
Mean ΔR_i^f	23,308.52	438,314.39	41,067.78	254,376.10	277,395.66	436,040.54
Mean ΔC_i	219,341.60	145,691.60	151,590.17	137,187.69	160,218.66	145,626.25
WMVPF	-0.019	0.100	0.013	0.017	2.350	0.100
WNSB	0.243	3.221	1.963	1.927	3.517	3.232
Number Audited	30,747	44,417	2,094	47,170	40,390	44,437
Mean η_i (targeted)	0.091	0.057	0.005	0.006	0.784	0.057
2017						
Mean ΔR_i^m	30,366.79	1,610.41	524,015.19	100,074.65	6,609.90	3,208.05
Mean ΔR_i^f	20,061.96	462,161.79	13,792.33	336,713.42	515,386.55	460,853.71
Mean ΔC_i	762,177.17	393,883.58	391,578.10	333,005.16	410,964.97	393,840.05
WMVPF	-0.507	0.048	0.015	0.028	0.038	0.048
WNSB	-0.050	1.186	1.379	1.320	1.280	1.187
Number Audited	9,140	17,036	128	624	4,127	17,098
Mean η_i (targeted)	0.269	0.005	0.003	0.005	0.006	0.005
2018						
Mean ΔR_i^m	19,475.66	697.65	541,881.15	90,635.47	2,461.31	1,301.49
Mean ΔR_i^f	6,447.41	471,361.78	18,607.41	534,229.57	506,982.16	471,028.95
Mean ΔC_i	824,331.32	394,059.29	419,335.38	420,702.43	423,405.24	394,088.25
WMVPF	-0.459	4.911	0.020	0.017	6.301	4.898
WNSB	-0.281	2.170	1.343	1.493	2.484	2.171
Number Audited	9,211	18,596	44	313	2,894	18,595
Mean η_i (targeted)	0.644	0.573	0.004	0.004	0.752	0.573

Notes: This table computes the results of the real-world observed policy in Pakistan for 2016 – 2018 vs. the predicted values from the optimal derived policies according to each machine learner in the scenario when welfare weights are generated according to $\eta_i = z_i^{-\gamma}$ where $\gamma = 1$. For each year, we report the mean revenue recoup from an audit under the observed policy and the mean predicted recoup under the derived policies (ΔR_i^m), the mean estimated deterrence effect under each policy (ΔR_i^f), and the mean estimated administrative cost under each policy (ΔC_i). We also report the welfare-weighted $MVPF$ ($WMVPF$) and welfare-weighted NSB ($WNSB$) of each policy, and the number of audits conducted under each policy. This table also reports the mean welfare weight of targeted individuals under each policy as η_i. The column titled "Optimal Policy (Best Learners)" combines the XGBoost learner for predicting ΔR_i^m with the Neural Network learner for predicting ΔR_i^f.

Table 7: Optimal vs. Observed Policy ($\gamma = 2$, train years = 2014 – 2015)

		Machine Learners				
Metric	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
2016						
Mean ΔR_i^m	23,233.92	1,549.63	246,521.20	7,860.66	3,193.33	5,159.52
Mean ΔR_i^f	23,308.52	437,899.05	42,964.21	254,426.00	277,621.33	435,623.38
Mean ΔC_i	219,341.60	145,477.98	150,646.28	137,139.25	160,252.39	145,419.31
WMVPF	-0.001	0.085	0.000	0.000	2.392	0.085
WNSB	0.254	3.193	1.922	1.913	3.552	3.204
Number Audited	30,747	44,482	2,176	47,187	40,381	44,500
Mean η_i (targeted)	0.078	0.049	0.000	0.000	0.799	0.049
2017						
Mean ΔR_i^m	30,366.79	1,580.02	522,446.19	95,360.10	6,544.36	3,201.02
Mean ΔR_i^f	20,061.96	462,179.58	13,794.95	336,023.38	512,794.91	460,887.49
Mean ΔC_i	762,177.17	394,097.10	391,131.15	332,272.85	410,377.41	393,886.75
WMVPF	-0.518	0.000	0.000	0.000	0.000	0.000
WNSB	-0.051	1.177	1.371	1.298	1.266	1.178
Number Audited	9,140	17,131	129	655	4,210	17,140
Mean η_i (targeted)	0.272	0.000	0.000	0.000	0.000	0.000
2018						
Mean ΔR_i^m	19,475.66	697.30	541,881.15	87,054.41	2,317.61	1,301.05
Mean ΔR_i^f	6,447.41	471,422.53	18,607.41	516,192.36	503,395.18	471,079.61
Mean ΔC_i	824,331.32	393,953.12	419,335.38	410,203.69	424,893.01	393,972.46
WMVPF	-0.443	4.949	0.000	0.000	6.119	4.936
WNSB	-0.270	2.180	1.337	1.471	2.354	2.181
Number Audited	9,211	18,601	44	332	3,124	18,600
Mean η_i (targeted)	0.626	0.578	0.000	0.000	0.689	0.578

Notes: This table computes the results of the real-world observed policy in Pakistan for 2016 – 2018 vs. the predicted values from the optimal derived policies according to each machine learner in the scenario when welfare weights are generated according to $\eta_i = z_i^{-\gamma}$ where $\gamma = 2$. For each year, we report the mean revenue recoup from an audit under the observed policy and the mean predicted recoup under the derived policies (ΔR_i^m), the mean estimated deterrence effect under each policy (ΔR_i^f), and the mean estimated administrative cost under each policy (ΔC_i). We also report the welfare-weighted $MVPF$ ($WMVPF$) and welfare-weighted NSB ($WNSB$) of each policy, and the number of audits conducted under each policy. This table also reports the mean welfare weight of targeted individuals under each policy as η_i. The column titled "Optimal Policy (Best Learners)" combines the XGBoost learner for predicting ΔR_i^m with the Neural Network learner for predicting ΔR_i^f.

B Appendix: Proofs

B.1 Proof of Proposition 1

Proof. 1. Define the following sets of conditional average treatment effects within the potential outcomes framework:

(a) $\Delta R_i^m \equiv E_{\boldsymbol{x}}[R_i^m|\alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[R_i^m|\alpha_i = 0, \boldsymbol{x}]$

(b) $\Delta R_i^f \equiv E_{\boldsymbol{x}}[R_i^f|\alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[R_i^f|\alpha_i = 0, \boldsymbol{x}]$

(c) $\Delta C_i \equiv E_{\boldsymbol{x}}[C_i|\alpha_i = 1, \boldsymbol{x}] - E_{\boldsymbol{x}}[C_i|\alpha_i = 0, \boldsymbol{x}]$

For ease of reading, denote $\Delta\mathcal{R}_i \equiv \Delta R_i^m + \Delta R_i^f$. Also denote $\Delta U_i \equiv U_i(\boldsymbol{y}, \boldsymbol{z}, \alpha_i = 1, \boldsymbol{\alpha_{-i}}; \boldsymbol{x}) - U_i(\boldsymbol{y}, \boldsymbol{z}, \alpha_i = 0, \boldsymbol{\alpha_{-i}}; \boldsymbol{x})$

2. Denote $v_i \equiv \psi(\boldsymbol{x})\Delta U_i + \lambda \cdot \Delta\mathcal{R}_i$. The tax agency's problem in any given period t is proportional to:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i}^{N} \alpha_i v_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i \cdot \Delta C_i \leq \bar{\mathcal{C}} \\ & 0 \leq \alpha_i \leq 1 \quad \forall i \end{aligned}$$

3. Formulate the agency's Lagrangian, where μ, π_i, and γ_i are Lagrange multipliers:

$$\mathcal{L} = \sum_{i}^{N} \alpha_i v_i - \mu(\sum_{i}^{N} \alpha_i \cdot \Delta C_i - \bar{\mathcal{C}}) + \sum_{i}^{N} \pi_i \alpha_i - \sum_{i}^{N} \gamma_i(\alpha_i - 1)$$

4. Solving this yields a system of KKT first-order conditions:

   (a) Stationarity:
   $$\frac{\partial \mathcal{L}}{\partial \alpha_i} = v_i - \mu \cdot \Delta C_i + \pi_i - \gamma_i = 0$$

   (b) Primal feasibility:
   $$\sum_{i}^{N} \alpha_i \cdot \Delta C_i \leq \bar{\mathcal{C}}; \quad 0 \leq \alpha_i \leq 1 \quad \forall i$$

   (c) Dual feasibility:
   $$\mu \geq 0, \quad \pi_i \geq 0, \quad \gamma_i \geq 0$$

   (d) Complementary slackness:
   $$\mu(\sum_{i}^{N} \alpha_i \cdot \Delta C_i - \bar{\mathcal{C}}) = 0, \quad \pi_i \alpha_i = 0, \quad \gamma_i(\alpha_i - 1) = 0$$

5. These imply the optimal allocation of audits adheres to a threshold rule:

$$\alpha_i = \begin{cases} 0 & \text{if} \quad \frac{v_i}{\Delta C_i} < \mu \\ [0,1] & \text{if} \quad \frac{v_i}{\Delta C_i} = \mu \\ 1 & \text{if} \quad \frac{v_i}{\Delta C_i} > \mu \end{cases}$$

   (a) For the marginal audit individual, $0 \leq \alpha_i \leq 1$. Therefore $\pi_i = \gamma_i = 0$ and $v_i - \mu \cdot \Delta C_i = 0 \implies \frac{v_i}{\Delta C_i} = \mu$ by stationarity

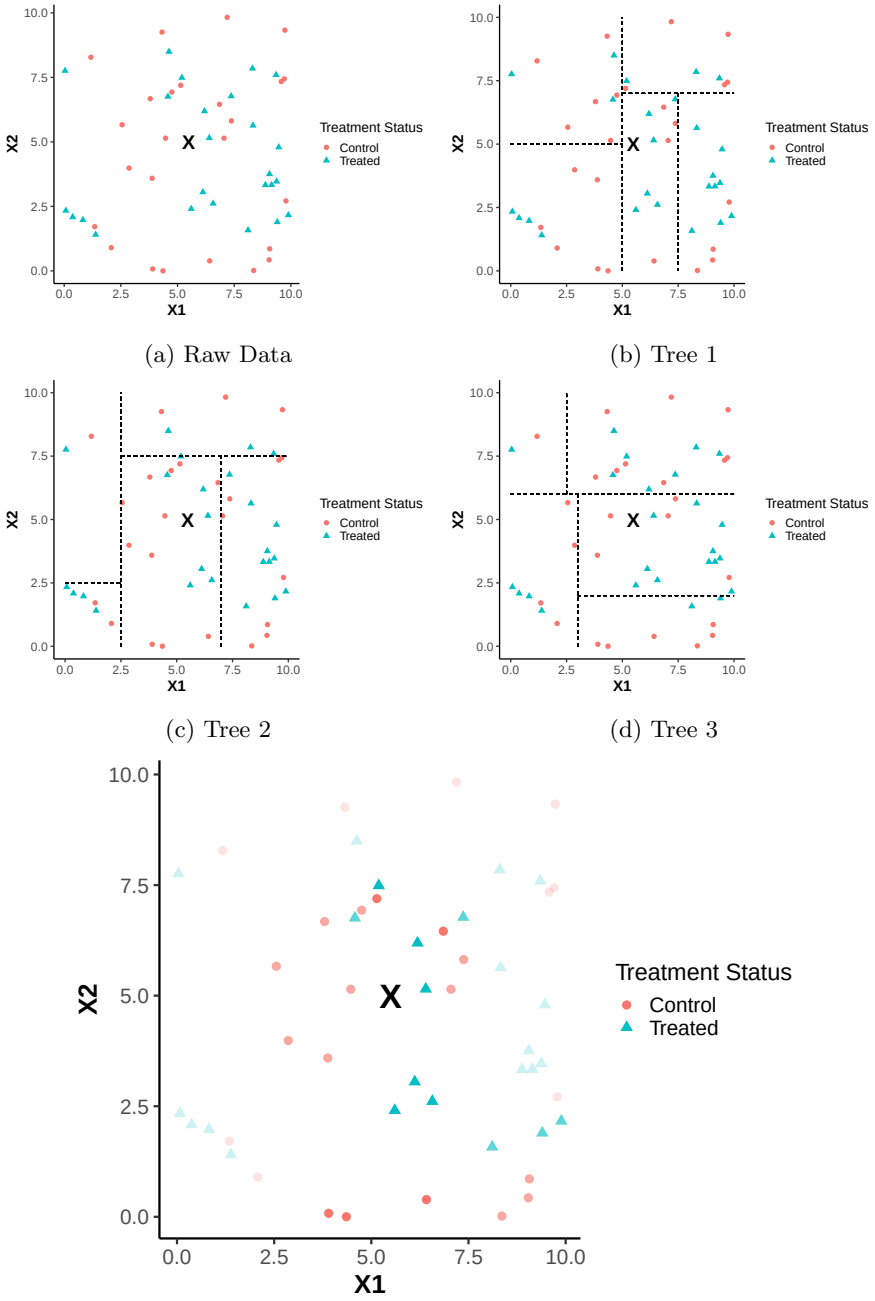
   (b) For non-audited individuals, $\alpha_i = 0$. Therefore $\pi_i > 0$, $\gamma_i = 0$, and $v_i - \mu \cdot \Delta C_i < 0 \implies \frac{v_i}{\Delta C_i} < \mu$ by stationarity

   (c) For audited individuals, $\alpha_i = 1$. Therefore $\pi_i = 0$, $\gamma_i > 0$ and $v_i - \mu \cdot \Delta C_i > 0 \implies \frac{v_i}{\Delta C_i} > \mu$ by stationarity

6. By the primal feasibility constraint, it must be true that a greedy algorithm that ranks individuals by $\theta_i = \frac{v_i}{\Delta C_i}$ and sequentially audits until the budget is met in expectation will implicitly satisfy the threshold rule at optimality.

□

C Appendix: Additional Figures



(a) Raw Data

(b) Tree 1

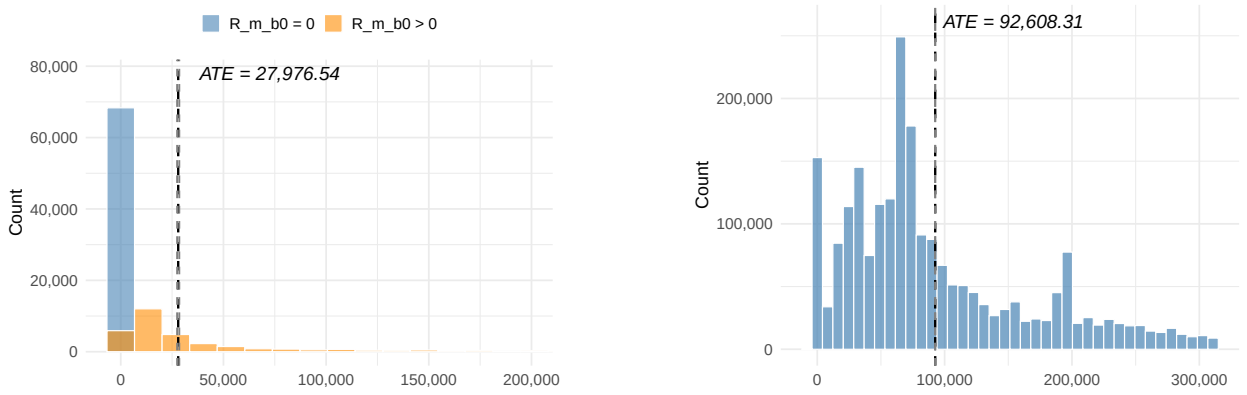
(c) Tree 2

(d) Tree 3

(e) Forest-Defined Kernel

Figure 4: Illustration of Forest-Based Aggregation

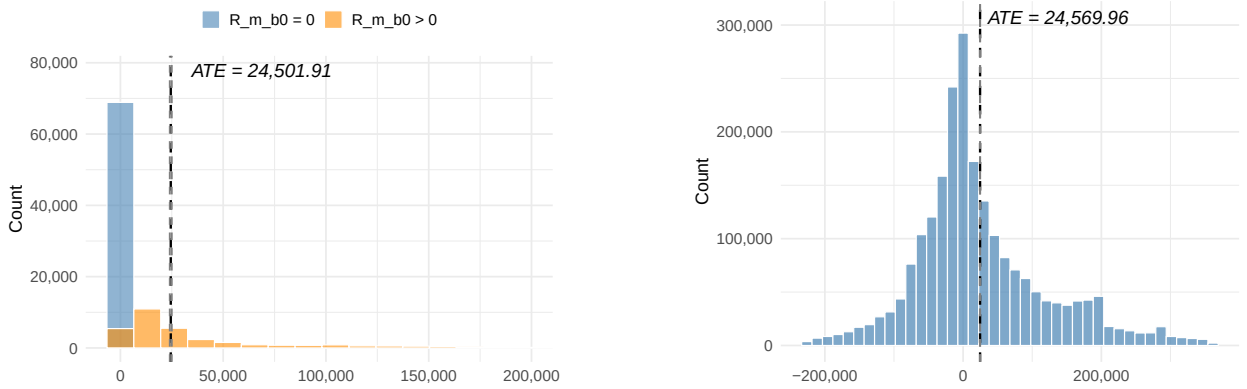
Notes: These figures show an example derivation of forest-defined weights for a given test point, $\boldsymbol{x}$, in a hypothetical causal forest with 3 trees. Panels (b), (c), and (d) illustrate partitions created by each of the three trees, and panel (e) shows the aggregation and how points are re-weighted based on how frequently they fall in the same terminal leaf as $\boldsymbol{x}$.



(a) R^m (2014 + 2015) (b) R^f (2014 + 2015)

Figure 5: CATE Distributions: Upfront and Long-Run Revenue (LASSO)

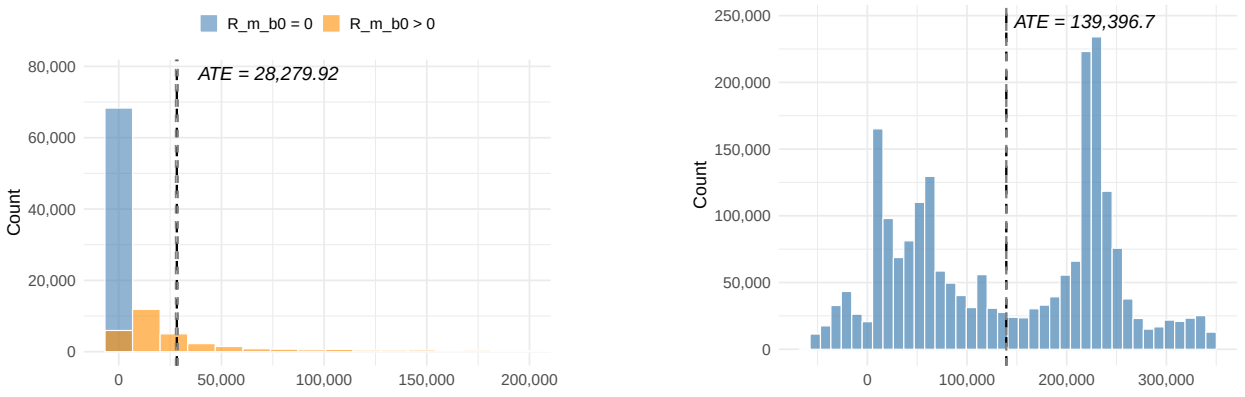
Notes: Each panel shows the distribution of in-bag (i.e. leave-one-out estimates for observations in the training year) Conditional Average Treatment Effect (CATE) estimates from LASSO models trained on the revenue outcomes. R^m denotes upfront revenue collected in the year of audit, while R^f denotes the treatment effect on the net present value of future revenue collected in subsequent years. For R^m, we plot the distribution of CATEs only for audited individuals so that we can also inspect prediction quality on $R^m = 0$ individuals. The blue color-code indicates individuals who were observed as $R^m = 0$ given that they were audited, and the orange color-code indicates individuals who were observed as $R^m > 0$. For R^f we plot the full distribution of CATE estimates, The dashed lines are the LASSO estimate of the average treatment effects (before BLP-debiasing). These are measured in Pakistani Rupees.



(a) R^m (2014 + 2015) (b) R^f (2014 + 2015)

Figure 6: CATE Distributions: Upfront and Long-Run Revenue (XGBoost)

Notes: Each panel shows the distribution of in-bag (i.e. leave-one-out estimates for observations in the training year) Conditional Average Treatment Effect (CATE) estimates from XGBoost models trained on the revenue outcomes. R^m denotes upfront revenue collected in the year of audit, while R^f denotes the treatment effect on the net present value of future revenue collected in subsequent years. For R^m, we plot the distribution of CATEs only for audited individuals so that we can also inspect prediction quality on $R^m = 0$ individuals. The blue color-code indicates individuals who were observed as $R^m = 0$ given that they were audited, and the orange color-code indicates individuals who were observed as $R^m > 0$. For R^f we plot the full distribution of CATE estimates, The dashed lines are the XGBoost estimate of the average treatment effects (before BLP-debiasing). These are measured in Pakistani Rupees.

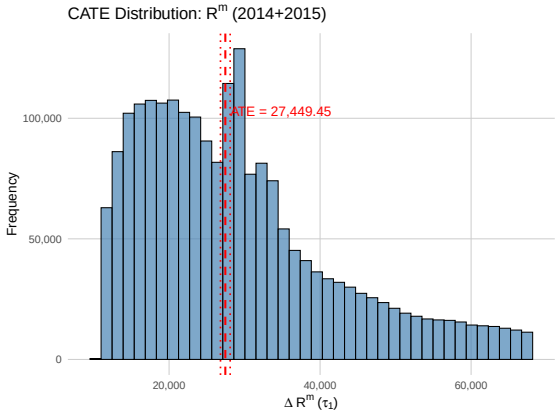


(a) R^m (2014 + 2015)

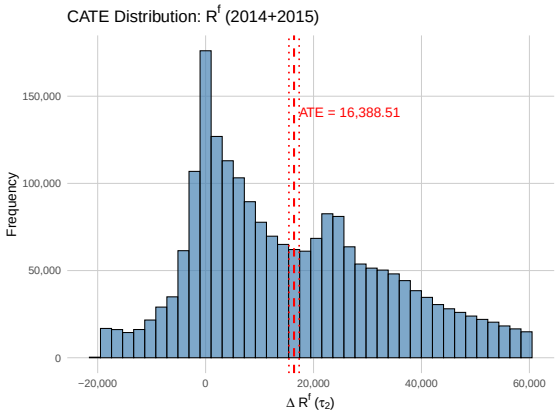
(b) R^f (2014 + 2015)

Figure 7: CATE Distributions: Upfront and Long-Run Revenue (Neural Network)

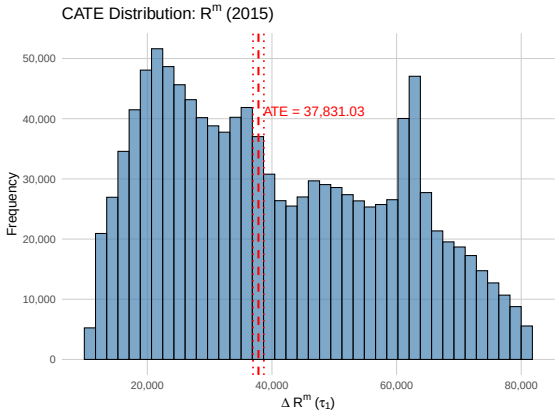
Notes: Each panel shows the distribution of in-bag (i.e. leave-one-out estimates for observations in the training year) Conditional Average Treatment Effect (CATE) estimates from Neural Network models trained on the revenue outcomes. R^m denotes upfront revenue collected in the year of audit, while R^f denotes the treatment effect on the net present value of future revenue collected in subsequent years. For R^m, we plot the distribution of CATEs only for audited individuals so that we can also inspect prediction quality on $R^m = 0$ individuals. The blue color-code indicates individuals who were observed as $R^m = 0$ given that they were audited, and the orange color-code indicates individuals who were observed as $R^m > 0$. For R^f we plot the full distribution of CATE estimates, The dashed lines are the Neural Network estimate of the average treatment effects (before BLP-debiasing). These are measured in Pakistani Rupees.



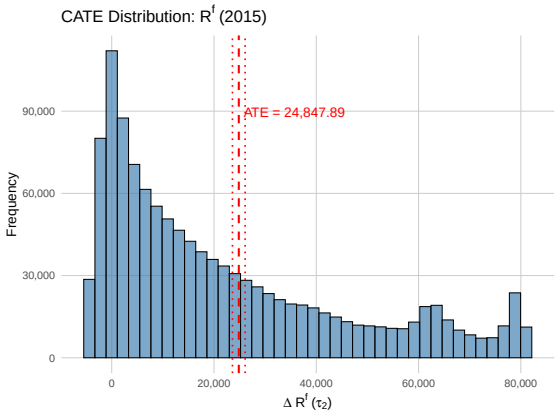
(a) R^m (2014 + 2015)



(b) R^f (2014 + 2015)



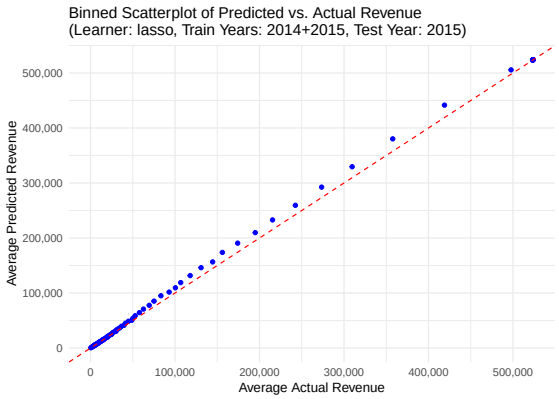
(c) R^m (2015 only)



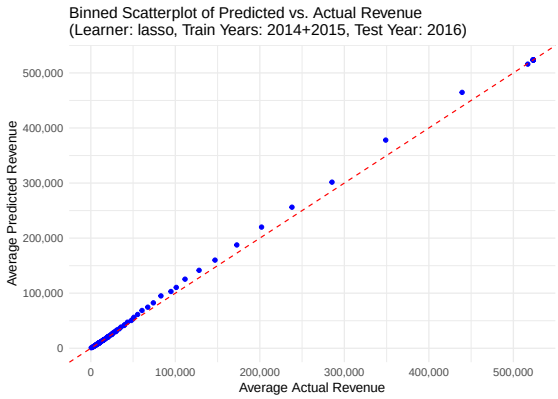
(d) R^f (2015 only)

Figure 8: CATE Distributions: Upfront and Long-Run Revenue (Causal Forest)

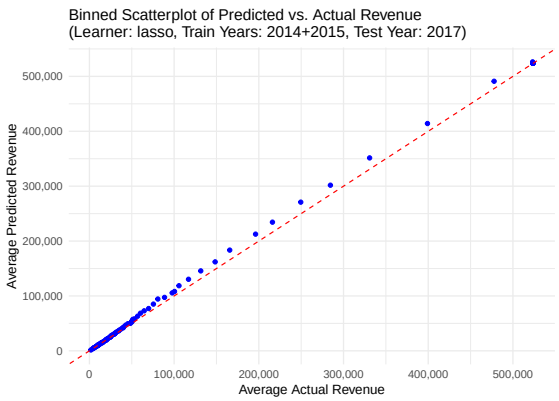
Notes: Each panel shows the distribution of in-bag (i.e. leave-one-out estimates for observations in the training year) Conditional Average Treatment Effect (CATE) estimates from causal forest models trained on the revenue outcomes. R^m denotes upfront revenue collected in the year of audit, while R^f denotes the net present value of future revenue collected in subsequent years. The dashed red line is the causal forest's estimate of the average treatment effect in that year with accompanying confidence intervals. These are measured in Pakistani Rupees.



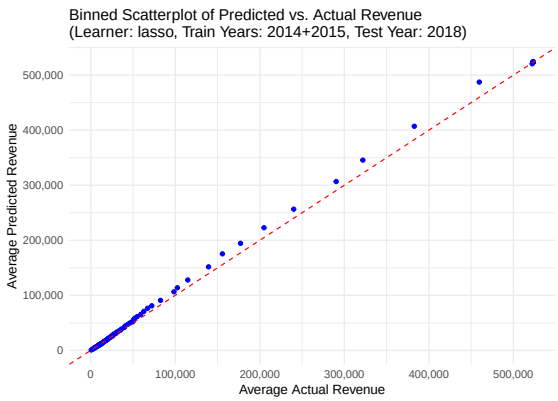
(a) R^m (Year = 2015)



(b) R^m (Year = 2016)



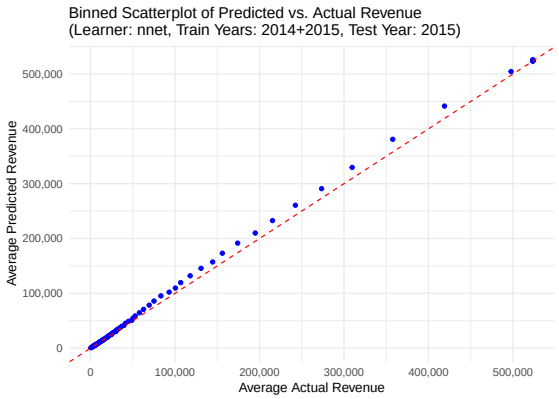
(c) R^m (Year = 2017)



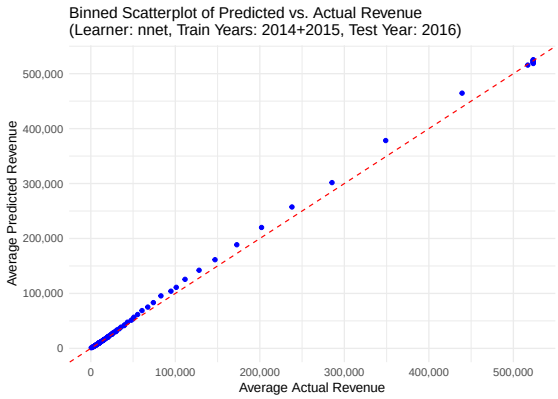
(d) R^m (Year = 2018)

Figure 9: Predicted vs. Observed R^m (LASSO)

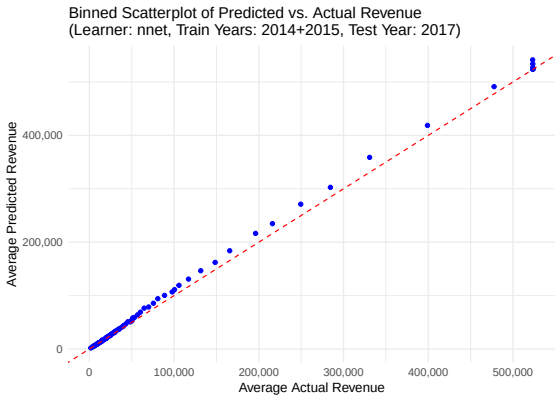
Notes: Each figure shows a binscatter of the mean value of ΔR^m from the observed audit samples in 2015 – 2018 over each percentile, plotted against the predicted values of $\Delta \hat{R}^m$ for units in that percentile according to the LASSO.



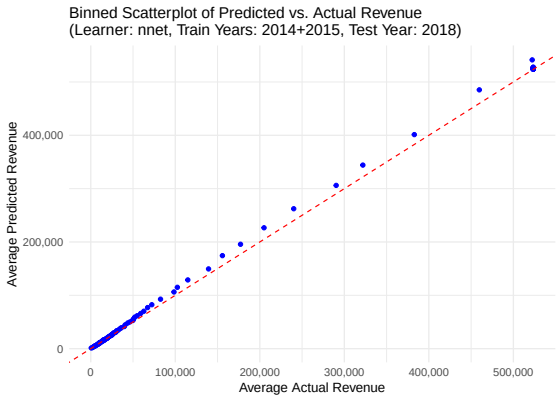
(a) R^m (Year = 2015)



(b) R^m (Year = 2016)



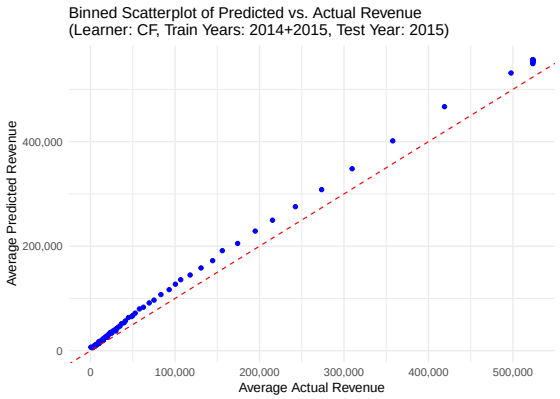
(c) R^m (Year = 2017)



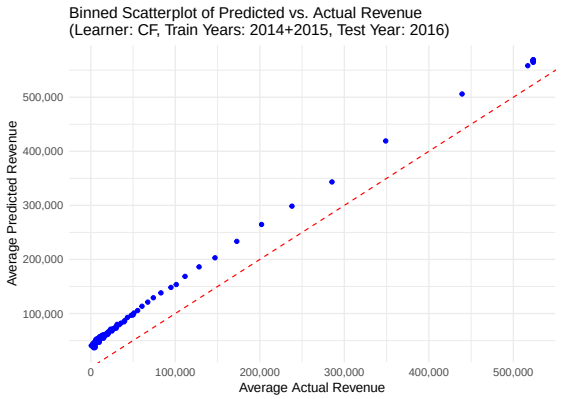
(d) R^m (Year = 2018)

Figure 10: Predicted vs. Observed R^m (Neural Network)

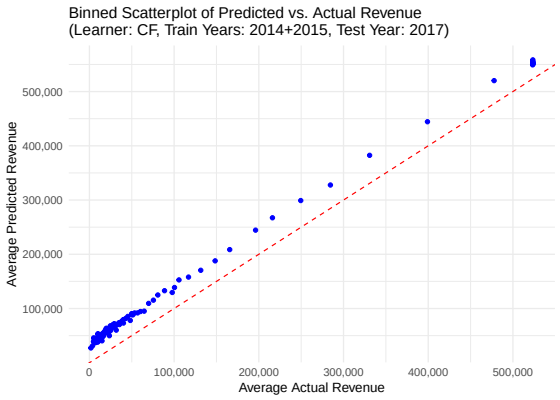
Notes: Each figure shows a binscatter of the mean value of ΔR^m from the observed audit samples in 2015 – 2018 over each percentile, plotted against the predicted values of $\Delta \hat{R}^m$ for units in that percentile according to the de-biased Neural Network proxy predictor algorithm. The dashed red line is a 45-degree line.



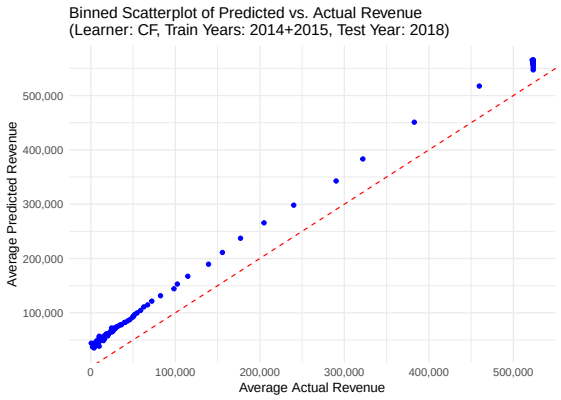
(a) R^m (Year = 2015)



(b) R^m (Year = 2016)



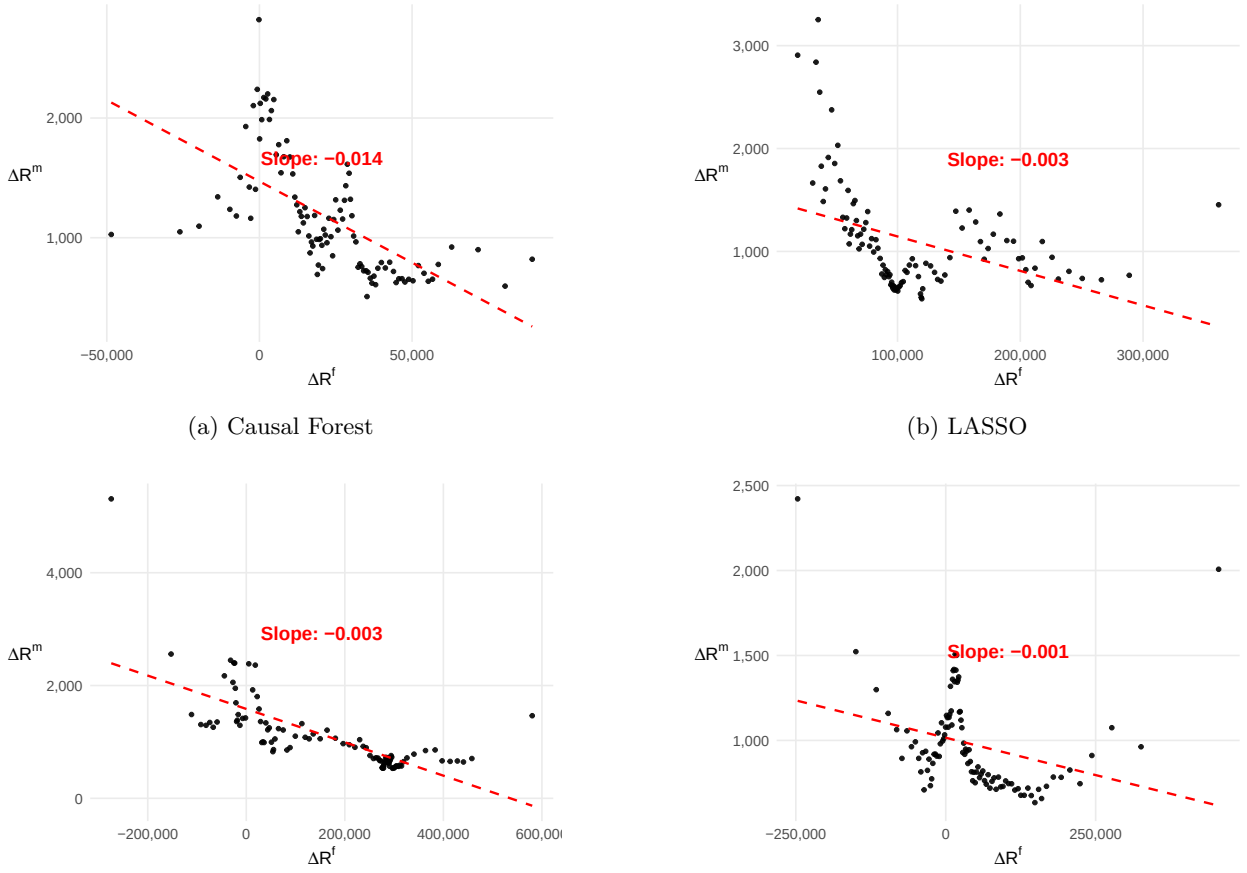
(c) R^m (Year = 2017)



(d) R^m (Year = 2018)

Figure 11: Predicted vs. Observed R^m (Causal Forest)

Notes: Each figure shows a binscatter of the mean value of ΔR^m from the observed audit samples in 2015 – 2018 over each percentile, plotted against the predicted values of $\Delta \hat{R}^m$ for units in that percentile according to the Causal Forest.



(a) Causal Forest

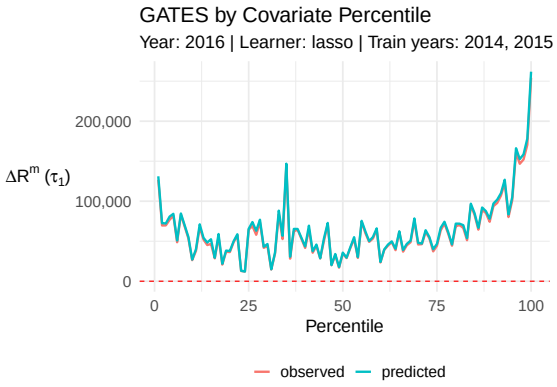
(b) LASSO

(c) Neural Network

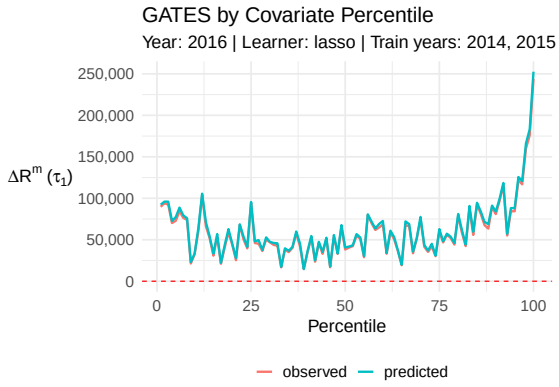
(d) XGBoost

Figure 12: $\Delta\hat{R}^m$ vs. $\Delta\hat{R}^f$ Tradeoff

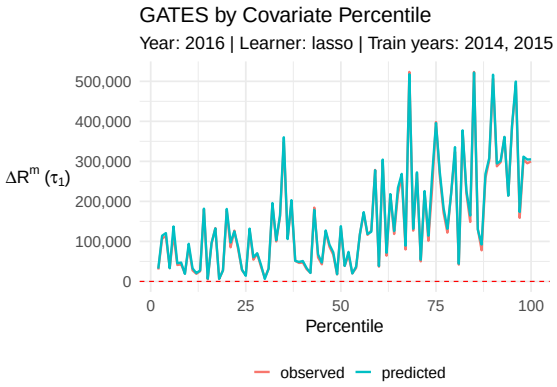
Notes: Each figure shows a binscatter of the mean value of $\Delta\hat{R}^m$ vs. the mean value of $\Delta\hat{R}^f$ over each percentile of $\Delta\hat{R}^f$. We also include a red dashed fit line with the corresponding slope.



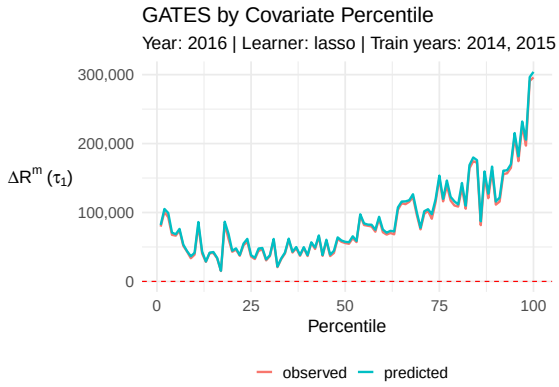
(a) Accounting Profits



(b) Cost of Sales



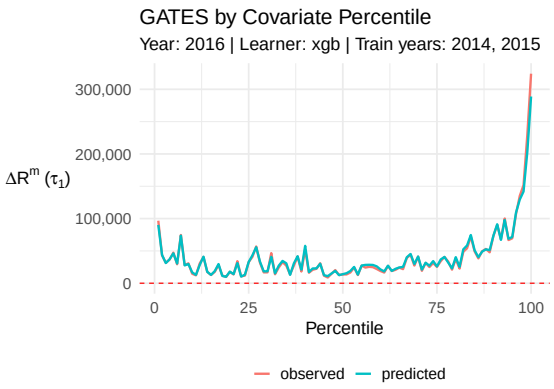
(c) Tax Credits



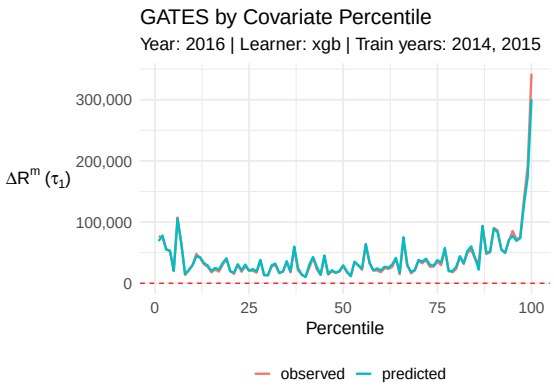
(d) Declared Total Income

Figure 13: ΔR^m Distribution over Covariates: LASSO

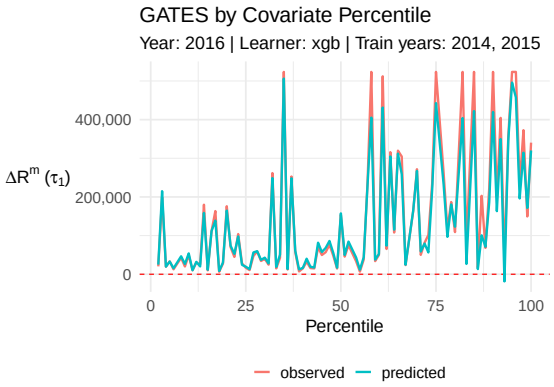
Notes: Each figure shows a binscatter of the mean value of R^m from the observed audit samples in 2015 – 2018 over each percentile of various important covariates, plotted against the predicted values of $\hat{R}^m$ for units in that percentile according to the de-biased LASSO proxy predictor algorithm. The dashed red line is a 45-degree line.



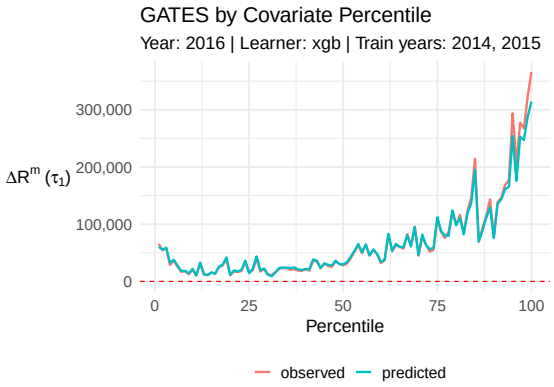
(a) Accounting Profits



(b) Cost of Sales



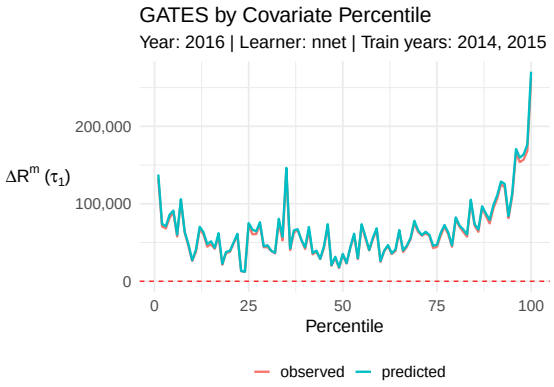
(c) Tax Credits



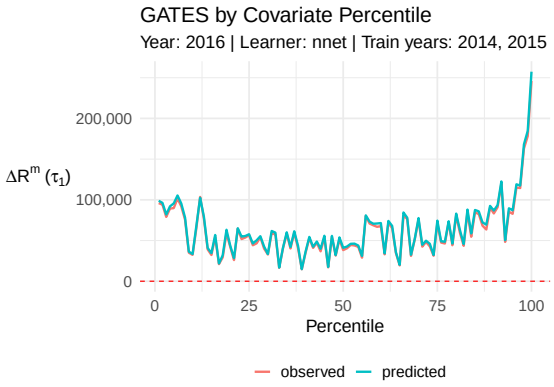
(d) Declared Total Income

Figure 14: ΔR^m Distribution over Covariates: XGBoost

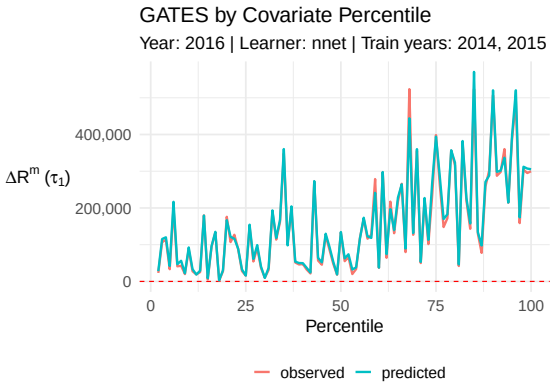
Notes: Each figure shows a binscatter of the mean value of R^m from the observed audit samples in 2015 – 2018 over each percentile of various important covariates, plotted against the predicted values of $\hat{R}^m$ for units in that percentile according to the de-biased XGBoost proxy predictor algorithm.



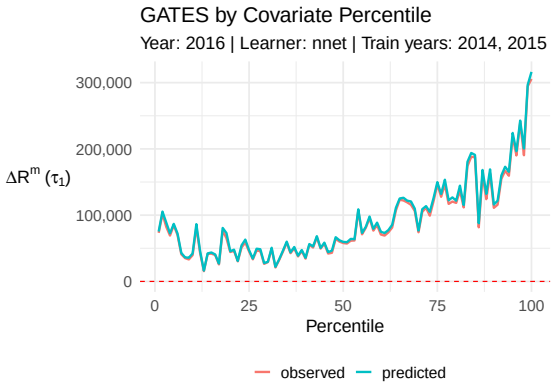
(a) Accounting Profits



(b) Cost of Sales



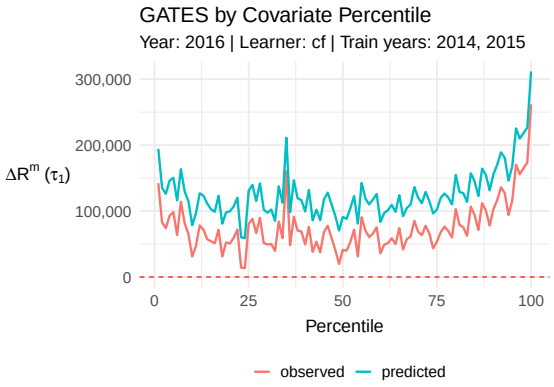
(c) Tax Credits



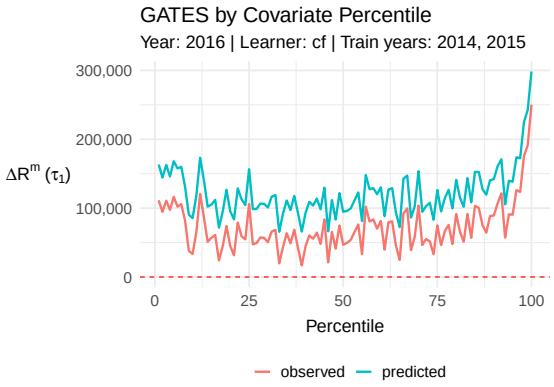
(d) Declared Total Income

Figure 15: ΔR^m Distribution over Covariates: Neural Network

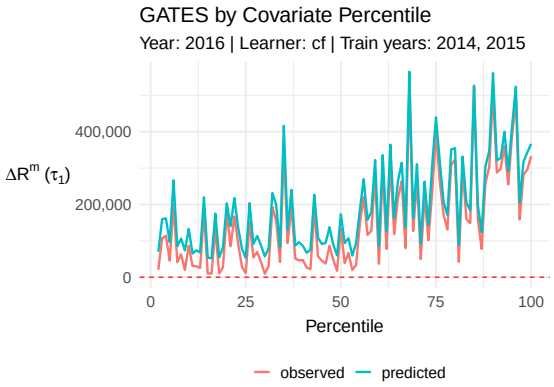
Notes: Each figure shows a binscatter of the mean value of R^m from the observed audit samples in 2015 – 2018 over each percentile of various important covariates, plotted against the predicted values of $\hat{R}^m$ for units in that percentile according to the de-biased Neural Network proxy predictor algorithm.



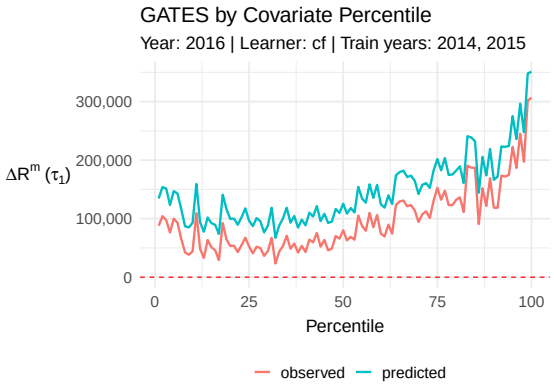
(a) Accounting Profits



(b) Cost of Sales



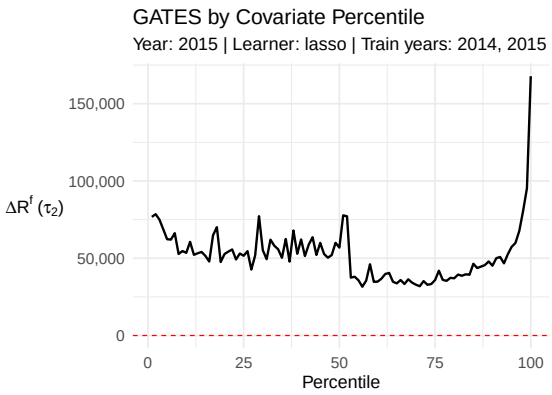
(c) Tax Credits



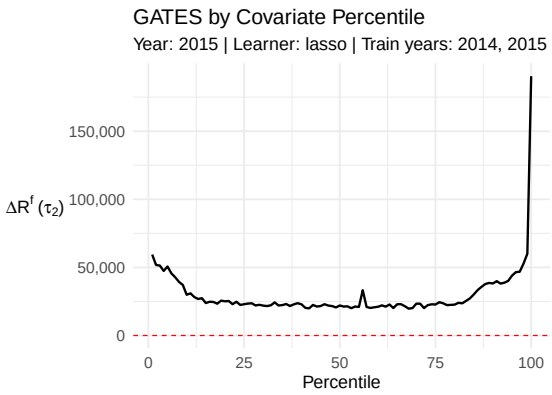
(d) Declared Total Income

Figure 16: ΔR^m Distribution over Covariates: Causal Forest

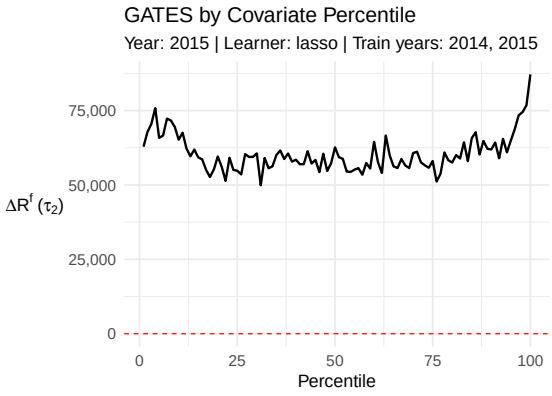
Notes: Each figure shows a binscatter of the mean value of R^m from the observed audit samples in 2015 – 2018 over each percentile of various important covariates, plotted against the predicted values of $\hat{R}^m$ for units in that percentile according to the Causal Forest.



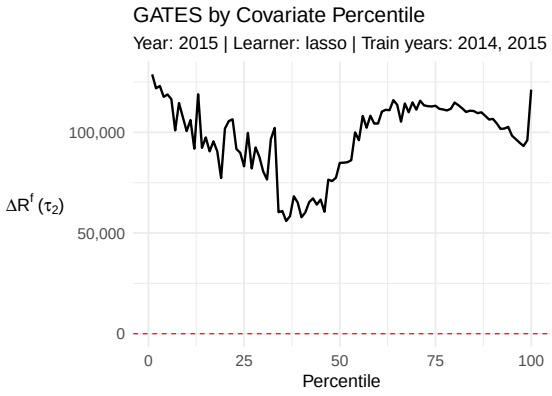
(a) Accounting Profits



(b) Cost of Sales



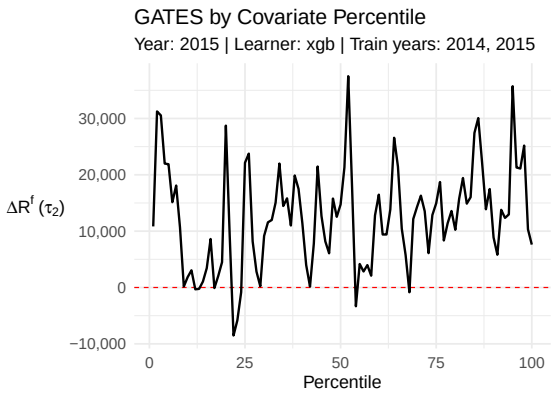
(c) Tax Credits



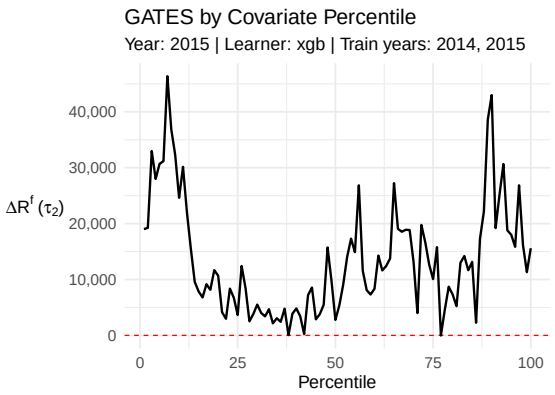
(d) Declared Total Income

Figure 17: ΔR^f Distribution over Covariates: LASSO

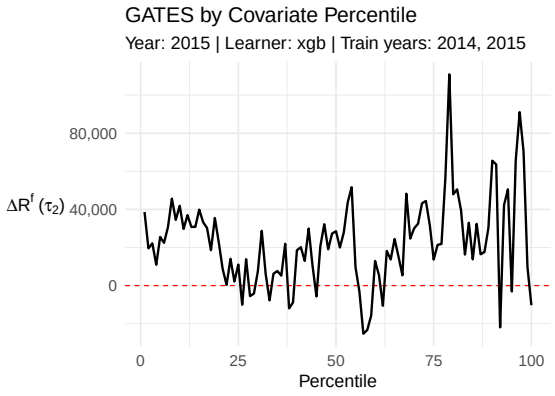
Notes: Each figure shows the average treatment effect for each percentile of various important covariates according to the de-biased LASSO proxy predictor algorithm.



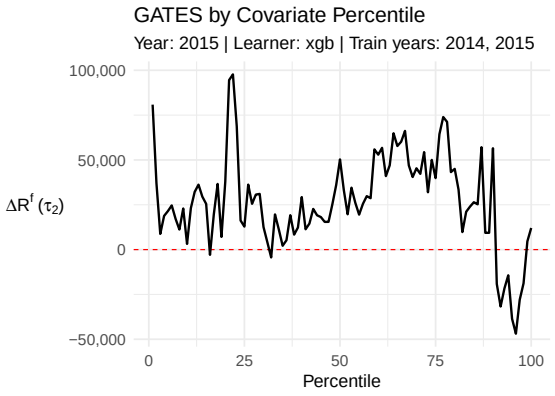
(a) Accounting Profits



(b) Cost of Sales



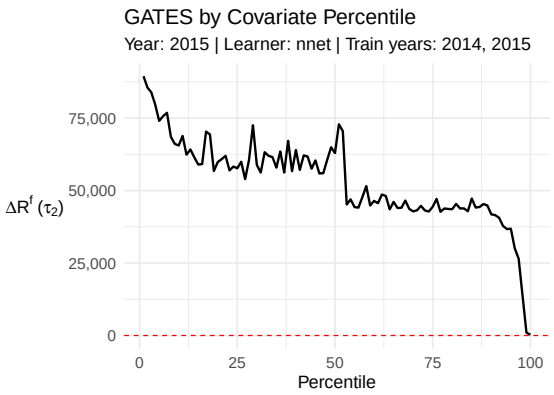
(c) Tax Credits



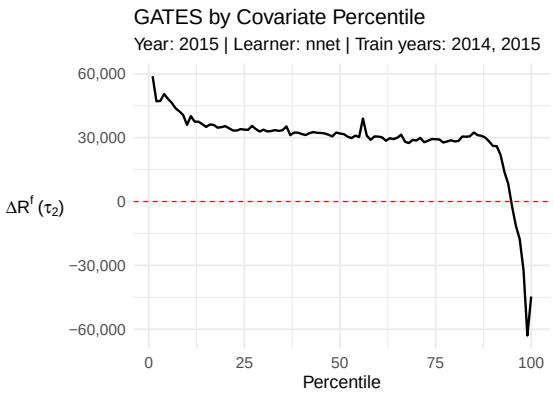
(d) Declared Total Income

Figure 18: ΔR^f Distribution over Covariates: XGBoost

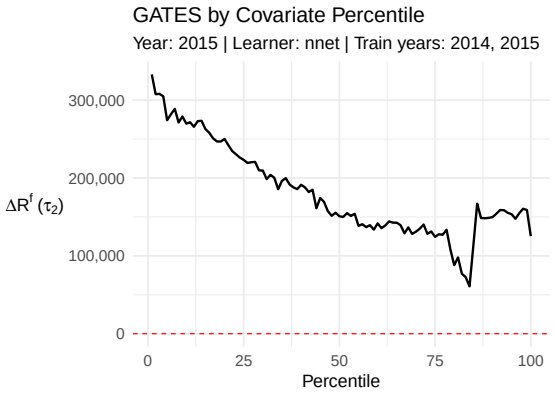
Notes: Each figure shows the average treatment effect for each percentile of various important covariates according to the de-biased XGBoost proxy predictor algorithm.



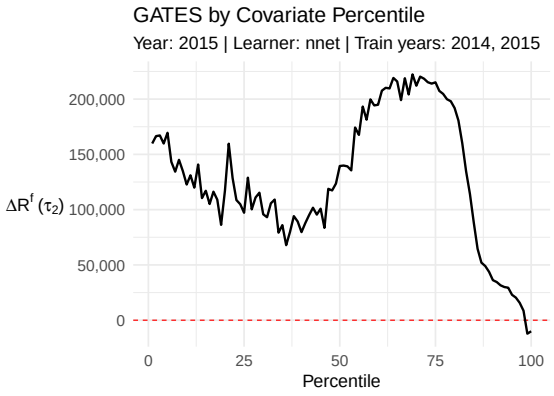
(a) Accounting Profits



(b) Cost of Sales



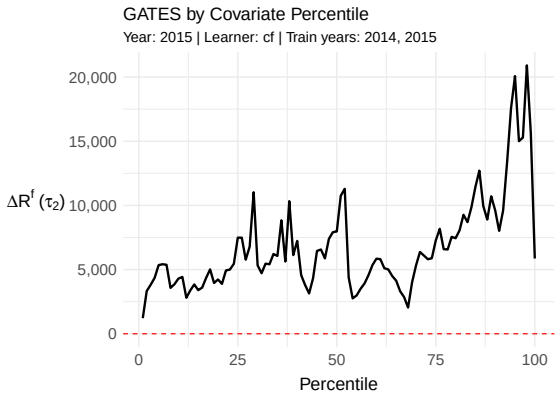
(c) Tax Credits



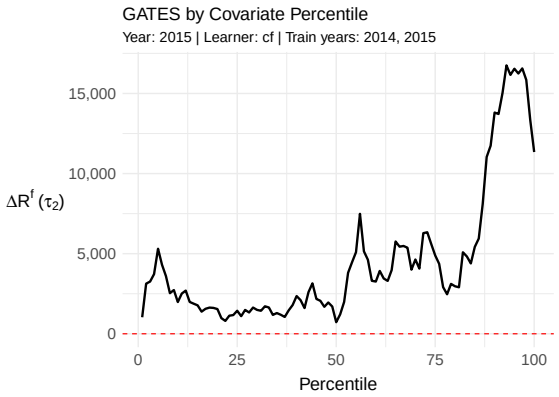
(d) Declared Total Income

Figure 19: ΔR^f Distribution over Covariates: Neural Network

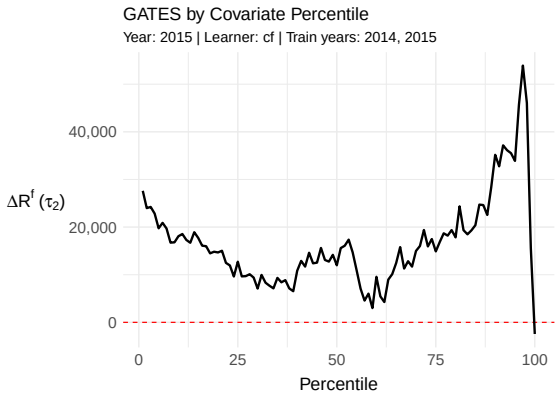
Notes: Each figure shows the average treatment effect for each percentile of various important covariates according to the de-biased Neural Network proxy predictor algorithm.



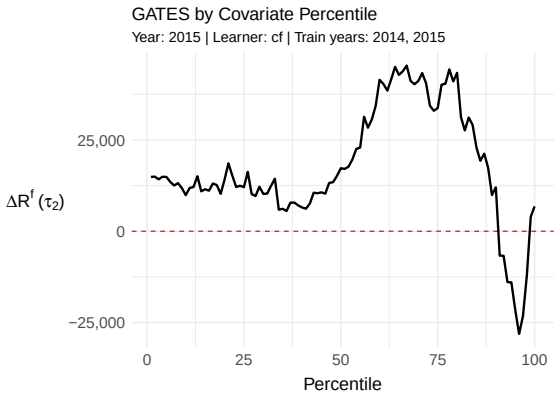
(a) Accounting Profits



(b) Cost of Sales



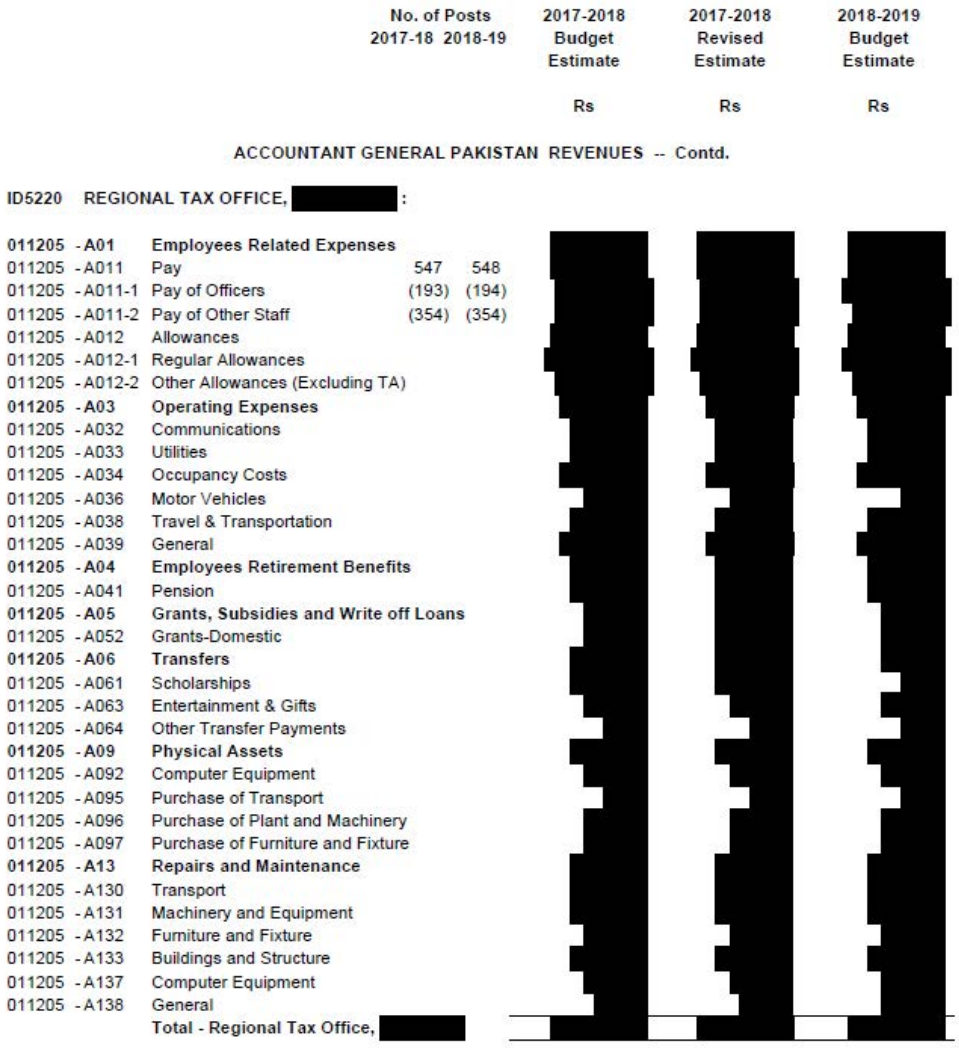
(c) Tax Credits



(d) Declared Total Income

Figure 20: ΔR^f Distribution over Covariates: Causal Forest

Notes: Each figure shows the average treatment effect for each percentile of various important covariates according to the Causal Forest.



		No. of Posts 2017-18	2018-19	2017-2018 Budget Estimate	2017-2018 Revised Estimate	2018-2019 Budget Estimate
				Rs	Rs	Rs
	ACCOUNTANT GENERAL PAKISTAN REVENUES -- Contd.					
ID5220	REGIONAL TAX OFFICE, [redacted] :					
011205 - A01	**Employees Related Expenses**					
011205 - A011	Pay	547	548			
011205 - A011-1	Pay of Officers	(193)	(194)			
011205 - A011-2	Pay of Other Staff	(354)	(354)			
011205 - A012	Allowances					
011205 - A012-1	Regular Allowances					
011205 - A012-2	Other Allowances (Excluding TA)					
011205 - A03	**Operating Expenses**					
011205 - A032	Communications					
011205 - A033	Utilities					
011205 - A034	Occupancy Costs					
011205 - A036	Motor Vehicles					
011205 - A038	Travel & Transportation					
011205 - A039	General					
011205 - A04	**Employees Retirement Benefits**					
011205 - A041	Pension					
011205 - A05	**Grants, Subsidies and Write off Loans**					
011205 - A052	Grants-Domestic					
011205 - A06	**Transfers**					
011205 - A061	Scholarships					
011205 - A063	Entertainment & Gifts					
011205 - A064	Other Transfer Payments					
011205 - A09	**Physical Assets**					
011205 - A092	Computer Equipment					
011205 - A095	Purchase of Transport					
011205 - A096	Purchase of Plant and Machinery					
011205 - A097	Purchase of Furniture and Fixture					
011205 - A13	**Repairs and Maintenance**					
011205 - A130	Transport					
011205 - A131	Machinery and Equipment					
011205 - A132	Furniture and Fixture					
011205 - A133	Buildings and Structure					
011205 - A137	Computer Equipment					
011205 - A138	General					
	Total - Regional Tax Office, [redacted]					

Figure 21: Sample Tax Office Budget

Notes: This figure is an example annual budget for a given tax office. We have redacted the name of the tax office and the budget line item values.

Table 8: Comparison of Tax Officer Distribution and Survey Respondents by Office

Tax Office	Officers	Respondents	Officer Share (%)	Respondent Share (%)	Expected Respondents	p-value
Abbottabad	80	3	2.27	3.23	2.11	0.533
Bahawalpur	140	3	3.96	3.23	3.69	0.717
Faisalabad	318	8	9.01	8.60	8.38	0.891
Gujranwala	190	1	5.38	1.08	5.01	0.065*
Hyderabad	189	15	5.35	16.13	4.98	<0.001***
Islamabad	225	9	6.37	9.68	5.93	0.193
Karachi	311	3	8.81	3.23	8.19	0.058*
Karachi II	350	4	9.91	4.30	9.22	0.027**
Karachi III	40	1	1.13	1.08	1.05	0.803
Lahore	355	8	10.05	8.60	9.35	0.642
Lahore II	40	1	1.13	1.08	1.05	0.803
Multan	251	6	7.10	6.45	6.61	0.528
Peshawar	313	9	8.86	9.68	8.24	0.782
Pindi	240	5	6.79	5.38	6.32	0.386
Quetta	96	0	2.72	0.00	2.53	0.108
Sargodha	136	13	3.85	13.98	3.58	<0.001***
Sialkot	146	3	4.13	3.23	3.85	0.659
Sukkur	111	1	3.14	1.08	2.93	0.253
Total	3531	103	100.00	100.00	103.00	—

Notes: This table displays the geographic distribution of tax officers in the FBR vs. the geographic distribution of responses to our survey. The first column displays the number of tax officers in each regional tax office in 2021. The second column displays the number of respondents from each tax office. The third and fourth columns show the share of tax officers from each office vs. the share of survey respondents from each office. The last two columns show the expected number of respondents from each office with p-values of their difference computed from z-tests.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

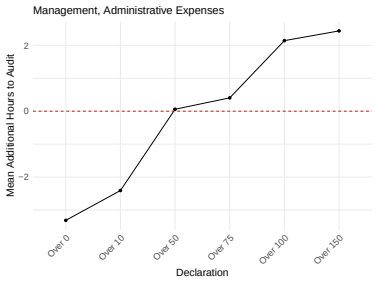
28. Above what threshold does **Declared Taxable Income** become a predictor of audit duration? And about how many hours does it usually add/subtract from the average? Please select 1 bo per-row.

Check all that apply.

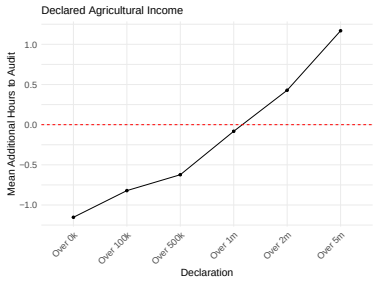
	15+ fewer	10 fewer	5 fewer	2 fewer	1 fewer	0	1 extra	2 extra	5 extra
0k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
100k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
200k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
300k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
500k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
750k+	☐	☐	☐	☐	☐	☐	☐	☐	☐
1 million+	☐	☐	☐	☐	☐	☐	☐	☐	☐
2 million+	☐	☐	☐	☐	☐	☐	☐	☐	☐
5 million+	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 22: Sample Audit Duration Grid

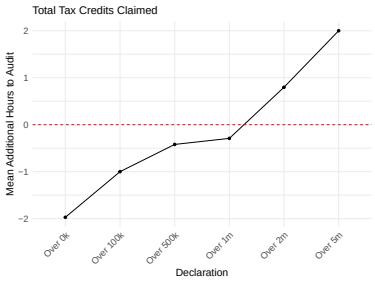
Notes: This image is an example of one of the questions that tax officers would see if they selected a tax return variable as an important predictor of audit duration. That is, over possible ranges of the variable's value, they were asked to map out the pattern of heterogeneity by selecting how many more/fewer hours they estimate the audit would take.



Admin Cost

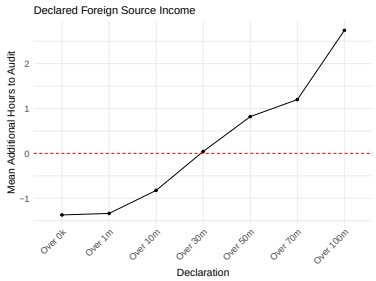


Agricultural Income

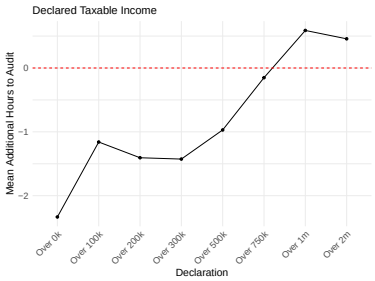


Credits

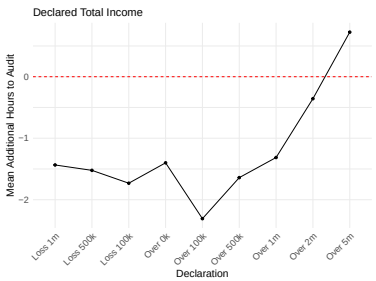
Business Income



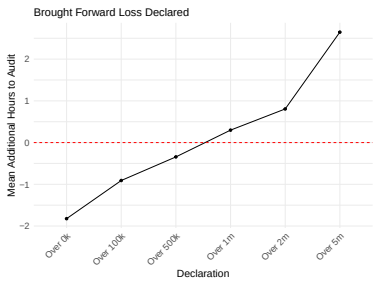
Foreign Income



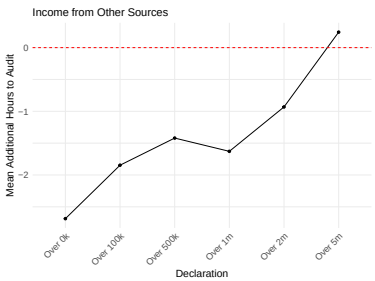
Taxable Income



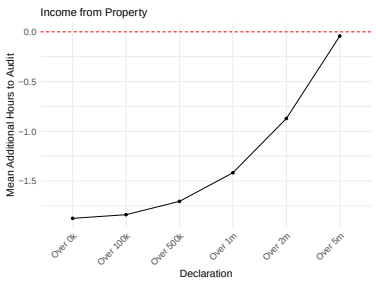
Total Income



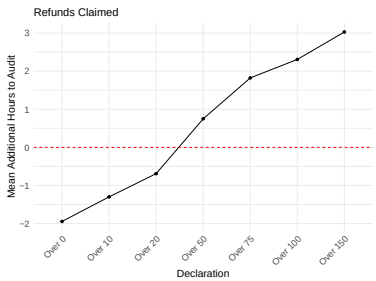
Losses



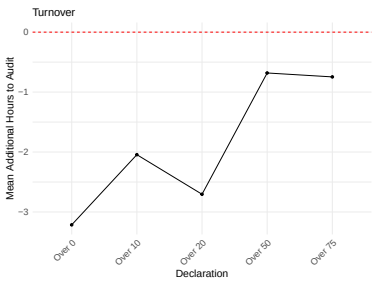
Other Income



Property Income



Refunds



Turnover

Figure 23: Audit Duration Distributions by Income Type

Notes: This displays the mean adjustment to the average audit duration in hours over various ranges of 12 taxpayer covariates according to the tax officer survey.

D Appendix: Additional Tables

Table 9: Targeted Sample Means ($\gamma = 1$, train years = 2014 – 2015)

Variable	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
Year: 2016						
Total Income	989,081	827,285	1,111,966	704,524	906,492	830,781
Salary Income	1,293,377	745,390	1,030,684	649,844	701,902	745,600
Taxable Income	1,054,493	819,459	1,093,049	698,269	752,493	823,155
Property Income	806,022	154,156	537,275	390,166	405,753	158,087
Business Income	548,630	586,664	675,907	601,910	457,802	581,118
Capital Asset Income	1,410,654	-26,327	1,337,360	540,450	193,124	-27,625
Other Income	525,967	222,071	480,794	331,499	405,443	231,324
Foreign Income	943,436	1,326,937	1,016,077	511,390	1,101,984	1,278,882
Agriculture Income	323,042	159,866	263,276	150,257	348,497	159,826
Exempt Income	4,381,363	115,877,750	4,320,820	1,045,501	1,396,120	116,327,307
Turnover	8,485,991	17,093,422	11,639,811	10,941,696	4,638,271	16,561,408
Cost of Sales	12,060,681	78,640,381	16,971,140	35,870,444	8,528,716	72,616,176
Gross Business Profit	1,508,841	6,024,638	2,012,991	2,938,699	972,586	5,646,569
Accounting Business Profit	787,097	828,029	868,051	846,160	504,020	812,287
Declared Tax Liability	151,114	217,630	193,849	251,845	28,820	218,111
Tax Credits	874,598	48,838	928,673	837,044	537,702	54,314
Days Late	72	458	214	1,144	320	455

Notes: This table displays a comparison of covariate means across targeted subsamples with welfare weights generated from $\eta_i = z_i^{-\gamma}$, where $\gamma = 1$. The first column reports the average values of various tax return line items among individuals selected for audit under the observed policy. The remaining columns show the same averages for people selected for audit under each derived policy.

Table 10: Targeted Sample Means ($\gamma = 2$, train years = 2014 – 2015)

Variable	Observed Policy	Optimal Policy (Best Learners)	Optimal Policy (Causal Forest)	Optimal Policy (LASSO)	Optimal Policy (XGBoost)	Optimal Policy (Neural Network)
Year: 2016						
Total Income	989,081	817,742	1,081,131	692,743	894,631	821,463
Salary Income	1,293,377	737,305	958,047	642,482	691,065	737,335
Taxable Income	1,054,493	810,127	1,062,773	686,652	740,976	813,840
Property Income	806,022	153,007	510,645	375,758	403,036	156,560
Business Income	548,630	586,224	669,376	597,504	455,908	580,557
Capital Asset Income	1,410,654	-25,527	1,114,884	494,623	133,004	-27,150
Other Income	525,967	220,901	465,765	321,293	400,801	229,801
Foreign Income	943,436	1,280,399	889,218	491,068	1,101,984	1,279,017
Agriculture Income	323,042	159,479	258,372	150,970	346,800	159,353
Exempt Income	4,381,363	116,203,508	4,285,048	1,046,391	1,444,786	116,668,022
Turnover	8,485,991	17,107,321	11,510,814	10,850,022	4,620,591	16,583,369
Cost of Sales	12,060,681	78,866,906	16,665,133	35,597,680	8,512,021	72,799,541
Gross Business Profit	1,508,841	6,035,589	1,980,809	2,930,015	955,955	5,648,524
Accounting Business Profit	787,097	826,638	860,359	845,624	495,471	811,567
Declared Tax Liability	151,114	216,577	186,887	249,783	28,344	216,959
Tax Credits	874,598	47,592	922,470	849,911	527,970	54,334
Days Late	72	462	231	1,147	326	458

Notes: This table displays a comparison of covariate means across targeted subsamples with welfare weights generated from $\eta_i = z_i^{-\gamma}$, where $\gamma = 2$. The first column reports the average values of various tax return line items among individuals selected for audit under the observed policy. The remaining columns show the same averages for people selected for audit under each derived policy.

Table 11: Balance: Audited vs. Not Audited

	Year: 2014				Year: 2015			
Variable	Mean (Audit = 0)	Mean (Audit = 1)	Diff. in Means	p-value	Mean (Audit = 0)	Mean (Audit = 1)	Diff. in Means	p-value
Total Income	948,053	928,566	-19,487	0.974	1,300,269	1,358,848	58,579	0.969
Salary Income	1,220,079	1,356,898	136,820	0.318	2,204,641	2,277,395	72,754	0.902
Taxable Income	912,695	915,983	3,288	0.984	985,146	1,354,243	369,097	0.085
Property Income	662,823	498,191	-164,633	0.782	891,050	949,782	58,732	0.556
Business Income	1,611,500	612,584	-998,917	0.764	561,891	603,295	41,405	0.371
Capital Asset Income	669,224	1,183,473	514,249	0.697	682,988	507,989	-174,999	0.843
Other Income	365,512	320,139	-45,373	0.876	357,011	485,765	128,753	0.021
Foreign Income	740,387	444,847	-295,540	0.402	704,635	683,447	21,188	0.639
Agriculture Income	399,860	225,923	-173,937	0.000	371,514	466,350	94,836	0.289
Exempt Income	2,055,579	3,596,589	1,541,010	0.743	5,240,522	5,647,039	406,517	0.313
Turnover	10,735,866	11,975,164	1,239,298	0.694	10,430,982	10,866,649	435,667	0.582
Cost of Sales	10,812,791	10,385,510	427,281	0.799	13,236,766	12,248,231	-988,535	0.125
Gross Business Profit	795,991	1,211,478	415,487	0.000	1,940,208	1,669,360	-270,848	0.466
Accounting Business Profit	8,569,216	8,829,294	260,078	0.910	831,389	813,483	17,906	0.856
Declared Tax Liability	107,187	99,159	-8,028	0.264	319,795	395,973	76,178	0.615
Tax Credits	279,360	196,715	-82,645	0.236	1,298,902	1,269,345	-29,557	0.825
Days Late	25	15	-10	0.437	29	34	5	0.569

Notes: This table displays balance test results for each randomization wave over various tax return covariates. For each year we restrict the sample to tax returns corresponding to that tax year and estimate the model $y_i = \alpha + \beta \cdot \mathbb{1}(\text{Audited}_i) + \epsilon_i$, where y_i corresponds to each covariate in the first column. The first two columns under each year report the means of the non-audited individuals vs. the audited individuals, respectively. The third and fourth columns under each year report the $\hat{\beta}$ coefficient from the regression and the corresponding p-value of the difference.

Table 12: GRF Calibration Results

Causal Forest Model	Mean Forest Prediction (β_1)	Differential Forest Prediction (β_2)	SE (β_1)	SE (β_2)	p-value (β_1)	p-value (β_2)
R^m (train = 2014–2015)	0.917	0.825	0.010	0.021	0	0
R^f (train = 2014–2015)	0.889	0.727	0.040	0.039	0	0
R^m (train = 2015)	0.888	0.880	0.012	0.026	0	0
R^f (train = 2015)	0.661	0.527	0.064	0.065	0	0

Notes: This table evaluates the calibration of the causal forest learners. We estimate the best linear fit of causal forest predictions on held-out data using the test_calibration() function in the *grf* R package. Values of "Mean Forest Prediction" close to 1 indicate that the causal forest is doing a good job of capturing the average causal effect, and values of "Differential Forest Prediction" close to 1 additionally suggest that the forest is well-calibrated to predicting heterogeneous causal effects. Also, a β_2 coefficient significantly different from zero is evidence that there is significant heterogeneity in the data. We use the HC1 covariance type for the standard errors.