Statistical Methods & Applications (2023) 32:237–270
https://doi.org/10.1007/s10260-022-00643-4

ORIGINAL PAPER

2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy

Pierfrancesco Alaimo Di Loro[1,2] · Daria Scacciatelli[3] · Giovanna Tagliaferri[1,4]

Accepted: 2 May 2022 / Published online: 6 June 2022

Abstract

The revenue loss from tax avoidance can undermine the effectiveness and equity of the government policies. A standard measure of its magnitude is known as the *tax gap*, that is defined as the difference between the total taxes theoretically collectable and the total taxes actually collected in a given period. Estimation from a *micro* perspective is usually tackled in the context of *bottom-up* approaches, where data regularly collected through *fiscal audits* are analyzed in order to provide inference on the general population. However, the sampling scheme of fiscal audits performed by revenue agencies is not random but characterized by a selection bias toward *risky* taxpayers. The current standard adopted by the Italian Revenue Agency (IRA) for overcoming this issue in the Tax audit context is the *Heckman model*, based on linear models for modeling both the selection and the outcome mechanisms. Here we propose the adoption of the CART-based Gradient Boosting in place of standard linear models to account for the complex patterns often arising in the relationships between covariates and outcome. Selection bias is corrected by considering a re-weighting scheme based on propensity scores, attained through the sequential application of a classifier and a regressor. In short we refer to the method as *2-step Gradient Boosting*. We argue how this scheme fits the sampling mechanism of the IRA fiscal audits, and it is applied to a sample of VAT declarations from Italian individual firms in the fiscal year 2011. Results show a marked dominance of the proposed method over the currently adopted *Heckman model* in terms of predictive performances.

Keywords Gradient Boosting · Selection bias · Tax gap · Machine learning

Extended author information available on the last page of the article

1 Introduction

Tax evasion is the illegal activity of a person or an entity who deliberately avoids paying a true tax liability. This yields a revenue loss for the public budget and can undermine the effectiveness and equity of government policies. It represents one of the main problems in modern economies, where government budget is constantly under control and is required to match strict standards (Santoro 2010). The magnitude of the national tax evasion is usually quantified in terms of the *tax gap*, i.e. the difference between the total amounts of tax theoretically collectable and the total amounts of tax actually collected in a given period.

The techniques used for its estimation can be divided into two broad methodological approaches: *macro* and *micro*. Methodologies based on a *macro* perspective (*top-down*) usually employ macroeconomic indicators or national accounts data to directly derive a gross estimated of the revenue loss. Methodologies based on a *micro* perspective (*bottom-up*) instead consider individual data derived from administrative tax records (usually provided by internal fiscal agencies) and audit data[1] (Dangerfield and Morris (1992), Pisani and Pansini (2017), OECD (2017)). Unlike the former, the bottom-up perspective is able to derive estimates of single components of different taxpayers, and can be used to investigate individual factors that affect the individual inclination to evade taxes.

Several tax evasion studies are regularly conducted in many states by the corresponding revenue agencies. The tax gap research in Italy can be traced back to around 2000, since when the Italian Revenue Agency (IRA) provides yearly estimates of the tax gap for each tax typology using an in-house developed top-down methodology. Over the last few years, the IRA and Sogei[2] started to work on preliminary estimates of the tax gap related to self-employed taxpayers and small firms. The final objective is to evaluate, with different methods and different timing, the correspondence between the declared and the due amounts at the individual level (D'Agosto et al. 2016). The study is based on a bottom-up approach that integrates data from different sources, such as the Tax Register (Anagrafe Tributaria) and the Tax Audits Database (FISCALIS Tax Gap Project Group 2018). The Tax Register contains the "Declared Income Tax Base" (BID) for different entries of the individually filed tax returns, among which the Value Added Tax (VAT). The Tax Audits Database contains a correction of the declared taxable base for the same entries (namely the the "Potential Tax Base" - BIT), but it contains only a *small* and *non-random* sample of *audited* taxpayers. Size and sampling scheme of the individual Tax audits are determined by two main factors. First, they are quite expensive and time-consuming and only a small portion of the whole population of taxpayers (less than 10%) can be involved. Second, audits are performed with an administrative (and not statistical) end: recover as much revenue loss as possible given the limited time and resources available to perform the fiscal checks. Therefore, the auditing activity is not (and cannot be) based on random

[1] Data derived from ad-hoc tax assessments/controls on the taxpayer.

[2] Società Generale d'Informatica S.p.A. Information technology company 100% owned by the Italian Ministry of Economics and Finance.

audits, but relies on risk-based criteria decided and verified by specialized tax-auditors. As the tax-auditors do not have prior knowledge about the evaded amounts, their selection is based on the assumption that non-compliant taxpayers behave differently from compliant ones in terms of auxiliary information available on public registers (e.g. Tax Register). This induces an indirect *selection bias* on the observed response variable, which invalidates the inference of standard regression methods when generalizing from the observed sample to the whole population of taxpayers (Särndal et al. (2003), Särndal and Lundström (2005)).

The most widespread method to correct for such selection bias in the Tax gap estimation area is the Tobit-like model known as the "Heckman model" (Tobin (1952), Amemiya (1986), Heckman (1976), Heckman (1979)), currently adopted by the IRA and other Revenue Agencies worldwide (Braiotta et al. (2015), Kumar et al. (2015)). It verifies and accounts for the existence of a direct correlation between the selection scheme and the analyzed outcome (evaded amount), somehow taking for granted the effectiveness of the tax-auditors risk assessment. One major limitation of the classical Heckman model is its reliance on linear models for describing both the selection and the outcome processes. Real data are instead often characterized by more complex patterns, where linearity is usually a very coarse approximation that works fine for explanatory purposes, but lacks the necessary flexibility to produce accurate predictions. For this reason, various semiparametric estimation techniques have been increasingly used in combination with the Heckman model (Li and Racine 2007; Newey 2009; Marra and Radice 2013; Wojtyś and Marra 2015).

We here propose the adoption of a non-parametric Machine Learning tool to improve on the prediction accuracy of individual tax bases in for the IRA bottom-up studies. Algorithms from the Machine Learning literature are natural candidates for the analysis and detection of complex non-linear patterns in the data. However, their naive application in the context of tax audits would neglect the selection bias stemming from the non-random sampling scheme. We draw from Zadrozny (2004), Cortes et al. (2008) and consider a re-weighting scheme based on the sequential application of a classifier—for estimating the propensity scores (selection probabilities)—and a regressor—for predicting the true potential tax base. The propensity scores obtained at the first step are used to correct for the selection bias in the second. In particular, we use of the Gradient Boosting (GB) algorithm (Friedman 2001) with Classification And Regression Trees (CART) as base learners in both steps (i.e. *2-step Gradient Boosting*). The use of CARTs make this algorithm apt for data of any nature (without requiring much pre-processing) and exhibiting non-linear patterns. Their ensembling through the Gradient Boosting accrues an efficient and robust estimation of the boosted prediction function that circumvents over-fitting. The CART GB has been chosen in place of other alternative methods as it is extremely flexible and it is not a fully black-box as other Machine Learning methods. Indeed, it allows to compute measures of variable importance for interpreting the mechanisms behind its decisions and behavior (Deng and Yan 2019). It has already proved to perform well in estimating propensity scores in previous works (Lee et al. 2010) and, in the last decade, it has been successfully applied in a great variety of fields: economics (Liu et al. (2019), Yang et al. (2020)), biology (Moisen et al. (2006), clinical studies (Teramoto (2009), Zhang et al. (2019), Yin et al. (2020)), sociology (Kriegler and Berk (2010)). Let us point out that the selection bias correction considered in Zadrozny (2004), Cortes et al. (2008) is valid for any pair of classifier and regressor. As far as uncertainty quantification and prediction intervals are concerned, the 2-step GB must rely on non-parametric approaches. We here exploit a particular version of the bootstrap method that is able to quantify both bias and variability of the final estimates (Stine 1985; Efron and Tibshirani 1993; Efron et al. 2003; Kumar and Srivistava 2012).

The Heckman model and the proposed method are used to estimate the potential tax base on a representative sample of the Italian *individual firms*[3] for the fiscal year 2011. We exploit the information available in the Italian *Tax Register* (including auxiliary information and declared VAT) and collected on the audited taxpayers (*Tax Audits Database*, including the potential VAT) to infer the potential tax base on the whole sample. This is then used to build estimates of the individual and aggregate undeclared tax bases (VAT gaps) and evasion intensity scores.

The final aim of this work is to compare the predictive performances (both at the aggregate and individual level) of the proposed approach and of the currently adopted Heckman model. The results show that the 2-step GB yields more accurate individual potential VAT predictions, that ensure an improvement on the derived estimates of the overall tax gap and of the *VAT evasion intensity scores* (ratio between VAT gap and the declared part). In particular, the more reliable evasion intensity scores can be profitably used to support the tax-auditors in the detection of high risk taxpayers, drive for future tax audit selection, and recover a larger part of the revenue loss without increasing compliance controls and incurring in larger costs.

The paper is further organized as follows. Section 2 provides a general discussion on the considered working framework. In particular, Sect. 2.1 provides a brief description of the Heckman Model, its hypotheses and limitations, and Sect. 2.2 describes the proposed method. Section 2.3 includes a description of the bootstrap method used to derive the prediction intervals in the considered non-parametric setting. Finally, Sect. 3 envisions the application of both models on the above-mentioned data and the discussion of the results.

2 Modeling approach: working framework and methodology

As an effect of *tax non-compliance*, the amount of tax revenue actually collected by the state (declared tax base - BID) is generally lower than the true potential tax base (BIT). The difference is due to the presence of an undeclared part (BIND). These three quantities are connected through a basic accounting equation, known as the *main formula of gap computation*:

$$\mathrm{BIND} = \mathrm{BIT} - \mathrm{BID}.$$

[3] Individuals liable for tax on income as self-employees persons and small individual companies (ownership, board of directors and management are totally controlled by one person).

The amount of the BIND does not only directly affect revenue sufficiency through reduced tax revenues, but also impacts on income distribution and equity, limiting development and sustainable economic growth. Hence, its proper quantification is key in order support the decisions of policy makers on the allocation of the National budget.

Here we pursue the modeling and estimation of the individual *VAT potential tax base* of Italian *individual firms* BIT_i, from which the individual undeclared parts can be derived as $\mathrm{BIND}_i = \mathrm{BIT}_i - \mathrm{BID}_i$, $i = 1, \ldots, N$. This is in line with the micro perspective taken by *bottom-up* approaches to tax-gap estimation. It allows obtaining aggregate estimates of the overall *true potential tax base* $\mathrm{BIT} = \sum_{i=1}^{N} \mathrm{BIT}_i$ and undeclared parts $\mathrm{BIND} = \sum_{i=1}^{N} \mathrm{BIND}_i$, but also highlighting possible risk factors associated with a larger evasion intensity (non-compliance ratio) of single taxpayers (Braiotta et al. 2015). Whilst challenging from different point of views, the great advantage of this perspective is that the detected patterns may be used to identify possible high-risk groups of taxpayers and drive future selections for tax audits

In Italy, individual level information on tax compliance are either available from administrative sources such as the Tax Register, or collected via IRA Tax Audits. The Tax Register contains data on the *declared* income taxes $\{\mathrm{BID}_i\}_{i=1}^{N}$ for the whole population of taxpayers $\mathcal{P}$, while the *undeclared* tax base is available only on a small non-random sample of audited taxpayers $\mathcal{S}$: $\{\mathrm{BIT}_i\}_{i \in \mathcal{S}}$. The selection of taxpayers to audit is performed according to undisclosed criteria established by the Director of the Revenue Agency (D'Agosto et al. 2016), with the aim to recover as much loss revenue as possible given the limited resources for performing the audits. The tax-auditors do not have prior knowledge on the potential tax base, and hence cannot base their selection directly on it. Therefore, their selection is based on the auxiliary information available on the tax register. Units manifesting a suspicious behavior may be flagged as at risk of non-compliance and selected for an audit. If the selection process works as desired, it must be marginally correlated with the corresponding potential and undeclared tax base. Hence, it is non-negligible with respect to the outcome of interest and provides a sample affected by *selection bias*. If not taken into proper account, this can invalidate any inference and negatively affect the estimation accuracy (Särndal and Lundström 2005).

Let us denote the variable of interest Y, and let it be observed only on a non-random sample of units $\mathcal{S}$ selected according to a certain sampling design. The expected value of the outcome on each unit can be decomposed as:

$$\mathbb{E}[Y_i] = \mathbb{E}[Y_i \,|\, i \in \mathcal{S}] \cdot P(i \in \mathcal{S}) + \mathbb{E}[Y_i \,|\, i \in \mathcal{S}^c] \cdot P(i \in \mathcal{S}^c). \tag{1}$$

If the sampling design is non-ignorable, then:

$$\mathbb{E}[Y_i | i \in \mathcal{S}] \neq \mathbb{E}[Y_i \,|\, i \in \mathcal{S}^c],$$

and any inferential conclusion based only on observations $i \in \mathcal{S}$ cannot be directly generalized to estimate $\mathbb{E}[Y_i \,|\, i \in \mathcal{S}^c]$. In the specific context of tax audits, this implies that it is not possible to directly get any estimate for the potential tax base of not-audited units by using only information on the audited ones. By classical sampling theory (Lohr 2019), a correct estimation based on $\mathcal{S}$ could be obtained if the selection probabilities of each unit $i = 1, \ldots, N$ in the sample $\pi_i = P(i \in \mathcal{S})$ were known. These would allow to correct for the *non-rapresentativeness* of the selected sample but, in general, are never available a-priori. Nevertheless, while it is true that the outcome of interest is available only on units $i \in \mathcal{S}$, other covariates are usually known on all units. These can be used to model the selection mechanism and estimate the unknown inclusion probabilities $\hat{\pi}_i$, $i = 1, \ldots, N$, paving the way to *selection-bias correction* techniques.

2.1 The Heckman model: a brief review

The Heckman two-stage estimation procedure was initially introduced by Heckman (1979) as an econometric tool to estimate behavioral relationships from non-randomly selected samples in a regression setting. It is the standard method used by the Italian Revenue Agency to correct for the selection bias in the tax gap bottom-up estimates (Bordignon and Zanardi 1997; Pisani 2014), as it is for many other fiscal authorities (Werding 2005; Toder 2007; Imes 2013; Kumar et al. 2015). It models the effect of the selection bias as an ordinary specification error problem, implying a direct correlation between the outcome variable and the sampling mechanism. The observed outcome $\{y_i\}_{i=1}^N$ is assumed to be the realization of a bi-variate latent process $\{(z_{i1}, z_{i2})\}_{i=1}^N$. The first latent component z_{i1} is the actual variable of interest, while the z_{i2} represents the propensity to be selected of unit i. Given two *disjoint* sets of covariates $\boldsymbol{X}^1 = \left[\mathbf{x}_i^1\right]_{i=1}^N$ and $\boldsymbol{X}^2 = \left[\mathbf{x}_i^2\right]_{i=1}^N$, the two latent components are then expressed as:

$$\begin{aligned} z_{i1} &= \mathbf{x}_{i1}^T \boldsymbol{\beta_1} + u_{i1} \\ z_{i2} &= \mathbf{x}_{i2}^T \boldsymbol{\beta_2} + u_{i2} \end{aligned} \qquad \forall i = 1, \ldots, N.$$

Their relationship with the outcome of interest is the following:

$$y_i = \begin{cases} z_{i1} & z_{i2} > 0 \\ \text{Unobserved} & z_{i2} \leq 0, \end{cases}$$

where it becomes apparent how z_{i2} is the random quantity regulating the selection mechanism. Given that the expression of the two latent components does not share any covariates, if the two errors u_{i1} and u_{i2} are independent, then the selection mechanism would be negligible with respect to the estimation of y_i. However, if the two errors are dependent, selection bias would affect any estimates based on the known values of y_i. In particular, Heckman assumes that u_{i1} and u_{i1} have a bi-variate Normal distribution:

$$\begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}\right) \qquad \forall\, i = 1, \ldots, N, \tag{2}$$

where $\sigma_{12} = \sigma_{21}$ regulates the extent of the selection bias effect on the outcome. All parameters can be estimated in the likelihood framework, where an expression of the full likelihood can be found in Amemiya (1986). However, such inference was initially discarded in the seminal paper by Heckman (1979) due to the too long computing time it would have required. Heckman's original proposal is instead based on a *limited information maximum likelihood* criteria, where the sample selection is characterised as a special case of omitted variable problem on the observed sample $\{y_i\}_{z_{i2}>0}$. It can be proved that the omitted variable is the so-called *Inverse Mills-Ratio*:

$$\lambda(\mathbf{x}_{i2}^T \boldsymbol{\beta}_2/\sigma_2) = \frac{\Phi(-\mathbf{x}_{i2}^T \boldsymbol{\beta}_2/\sigma_2)}{1 - \Phi(-\mathbf{x}_{i2}^T \boldsymbol{\beta}_2/\sigma_2)}, \tag{3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the *standard Normal*. Let s_i be indicator variable defined as:

$$s_i = \begin{cases} 1 & z_{i2} > 0 \\ 0 & z_{i2} \leq 0 \end{cases}$$

Then, the value of the *Inverse Mills-Ratio* can be estimated fitting a *Probit* model for s_i on $\boldsymbol{X}^2$. We can then get an unbiased estimation of the model on the $\{y_i\}_{z_{i2}>0}$ by using the following regression:

$$y_i = \mathbf{x}_{i1}^T \boldsymbol{\beta_1} + \beta_\lambda \lambda(\mathbf{x}_{i2}^T \hat{\boldsymbol{\beta}}_2/\hat{\sigma_2}) + \epsilon_i \tag{4}$$

where $\beta_\lambda = \frac{\sigma_{12}}{\sigma 2}$ and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Sign and magnitude of β_λ summarize direction and intensity of the relationship between the outcome variable and the selection process. This estimation method can be proved to be consistent as long as the normality of u_2 holds and it is currently the standard way to obtain final estimates for the Heckman model.

However, even if this model looks elegant and can provide an effective solution in a lot of real world applications, it is not devoid of criticism (Stolzenberg and Relles 1997; Puhani 2000). For instance, it is generally not possible to distinguish a priori which covariates should affect the selection process and which ones the outcome. In these cases, $\mathbf{X}^1$ and $\mathbf{X}^2$ may have a large set of variables in common (or even be *identical*), breaking the theorethical foundations of the Heckman model. In practice, estimation is often pursued neglecting this issue, but two main practical complications may arise. First of all, Eq. 4 is only identified through the non-linearity of the *Inverse Mills Ratio* in Eq. (3). Since $\lambda(\cdot)$ is an approximately linear function over a wide range of its arguments, collinearity problems are likely to affect stability and reliability of the final estimates. Secondly, if the selection depends on covariates that also affect the outcome, then the observed sample $\{(y_i, \mathbf{x}_{i1})\}_{z_{i2}>0}$ will not be representative of the whole population with respect to the covariates in common. If the same linear relationship between the covariates and the response variable holds over the whole domain, then this would not be an issue. When instead the linearity is just an approximation, this may lead to wrong estimation of the corresponding slope coefficients. Indeed, the estimated slope will be based only on the behavior over the observed range of the predictors, while it neglects the possibly different behavior over the unobserved portion of their domain.

In addition, one could discuss whether the hypotheses of the Heckman model actually suits our application. First of all, Heckman's correction relies on the linear regression setting for the modeling of both the selection mechanism and the outcome. As highlighted in the introduction, many real-world application find linearity to be a too restricting assumption for such relationship. Secondly, we must recall that the audit sample is not self-selected, but the *Italian Revenue Agency* selects it through the set of globally available covariates $\mathbf{X}$. This means that the selection mechanism and potential tax base are marginally correlated, but this correlation vanishes when conditioning on the covariate values. Hence, if the propensity weights are correctly estimated and the estimated conditional mean function is able to follow the true behavior of the outcome, the two processes shall be independent. These last considerations exacerbate the flaws of the classic Heckman model implementation, emphasising the need for alternative solutions. There are many works that explore possible semi or non parametric expression of the mean functions within the Heckman setting, for instance considering *Polynomial and Spline regression* (Newey 2009; Marra and Radice 2013; Wojtyś et al. 2016) and *Generalized Additive Models* (Wojtyś and Marra 2015; Hastie and Tibshirani 2017). These methods can be positioned in between the interpretability of linear models and the flexibility of machine learning approaches. They can be seen as natural competitors to the method we propose, but their investigation is out of the scope of this work.

2.2 The 2-step Gradient Boosting approach

The approach considered in this paper is based on cost-sensitive learning (Elkan 2001; Breiman et al. 2017) as it is discussed in Zadrozny et al. (2003), Zadrozny (2004), Cortes et al. (2008). It consists of a re-weighting scheme that switches from the paradigm of *Empirical Risk Minimization* (Devroye et al. 2013) to *Weighted Empirical Risk Minimization* (Cortes et al. 2010; Du Plessis et al. 2014; Bekker et al. 2019). Observations less likely to be selected are given a larger weight to correct for their under-representation in the selected sample. In particular, such weights shall be equal to the inverse of the selection probability of each unit and must be estimated through an auxiliary model. Once the weights are known, the primary model (i.e. the prediction model) can be fit onto the re-weighted sample and provide correct predictions - as long as the outcome variable and the selection mechanism are marginally correlated, but conditionally independent given $\boldsymbol{X}$. Solutions of this kind are very common in the correction for bias deriving from non-negligible sampling designs, and is usually referred to as *propensity score weighting* (Rosenbaum and Rubin 1985; DuGoff et al. 2014). It is applied in different contexts, e.g.: when incorporating the response probability to correct for the non-response bias (Bethlehem (1988), Alho (1990), Bethlehem (2010)), in the case of the inverse probability of treatment weighting (IWTP) (Hirano et al. (2003), Austin and Stuart (2015)). It relies on the accuracy of the predictions of both the auxiliary classifier in the first step and the primary (substantial) predictive model in the second. Here we propose using the Gradient Boosting algorithm (described in Appendix A) on *Classification And Regression Trees* (CART) in both steps. The CART GB can easily model non-linear relationships, avoid over-fitting through regularization and it does not rely on any target distributional assumption. It is known to provide good predictive performances in many application fields, and already proved to perform well in the estimation of propensity scores in previous works (Lee et al. 2010; Deng and Yan 2019). The comparative merits of the proposed method and the Heckman model on the audited units are verified in Sect. 3. The same is not possible on un-selected units (not audited) as their true outcome is unknown. Hence, we propose a toy example in Appendix B where the selection mechanism tries mimicking the one of IRA audits.

2.2.1 The method

Let the complete population list $i = 1, \ldots, N$ be available and accompanied by a common set of covariates $\{\boldsymbol{x}_i\}_{i=1}^N$ and an indicator variable:

$$s_i = \begin{cases} 1 & y_i \text{ is observed} \\ 0 & y_i \text{ is not observed} \end{cases}$$

The theoretical foundations of the proposed methods rely on the conditional writing:

$$p(Y_i, S_i \,|\, \boldsymbol{x}_i) = p(Y_i \,|\, s_i, \boldsymbol{x}_i) \cdot p(s_i \,|\, \boldsymbol{x}_i),$$

and the following set of hypotheses:

H1 the probability to be included in the sample for unit i, $P(s_i = 1|\boldsymbol{x}_i)$, can be explained by the observed set of covariates $\boldsymbol{x}_i$;

H2 the response variable of unit i, Y_i, is conditionally independent from the sampling design given the vector of covariates $\boldsymbol{x}_i$:

$$P(Y_i|s_i, \boldsymbol{x}_i) = P(Y_i|\boldsymbol{x}_i) \quad \forall\, i \in \{1, \ldots, N\}.$$

H1 is an hypothesis that must hold also in the Heckman setting. The most relevant assumption is instead **H2**, which implies that the marginal dependence of the outcome variable on the sample design $\mathbb{E}[Y_i|s_i] \neq \mathbb{E}[Y_i]$ is fully explained by the available covariates:

$$\mathbb{E}[Y_i|s_i, \boldsymbol{x}_i] = \mathbb{E}[Y_i|\boldsymbol{x}_i] \quad \forall\, i \in \{1, \ldots, N\}. \tag{5}$$

As mentioned before, the tax auditors perform the risk-assessment using only the known covariates and hence **H2** can be reasonably assumed to hold in this context.

From a purely theorethical point of view, Eq. 5 could be estimated by simply fitting the chosen model on the set of selected units. However, since selected units differ from the un-selected ones in terms of the covariates values, the selected sample is not fully representative of the whole population. The optimization procedure favor good fits on the over-represented set of units and disregard the performances on under-represented ones. Hence, if the selection scheme is sharp in the covariates space, predictions on the un-selected units will rely on a certain dose of extrapolation. Propensity score weighting is useful to correct for this kind of bias, and its implementation is sketched here below.

1. *Classification*. A chosen classifier is trained on the whole sample with the binary variable *selected in the sample* s_i as the target. It must detect patterns and regularities between the selection mechanism of the units and the corresponding covariate values $\boldsymbol{x}_i$, and then provide accurate estimates of the selection probabilities $\{\hat{\pi}_i\}_{i=1}^N$. The estimated selection probabilities are nothing else but approximations to the first order inclusion probabilities:
$$\hat{\pi}_i \approx \pi_i = P(S_i = 1 \,|\, \boldsymbol{x}_i), \qquad i = 1, ..., N. \tag{6}$$
This practically is an attempt to reverse-engineer the selection criteria adopted by the tax-auditors to select units to audit.
2. *Prediction*. A prediction model is trained only on the selected sample $\{i : s_i = 1\}$ with the *response* variable y_i as a target. It is now possible to incorporate the inclusion probability resulting from Eq. 6 as individual weights in order to correct for the imbalanced representation of different units. These can be used to produce the inverse weights defined as:
$$v_i = \frac{P(S_i = 1)}{\hat{\pi}_i} \propto \frac{1}{\hat{\pi}_i}, \qquad i = 1, \ldots, N, \tag{7}$$
where $P(S_i = 1)$ is the probability to be selected notwithstanding the set of covariates, hence constant across individuals. The formula of the inverse weights in Eq. 7 derives from the *Bias Correction Theorem* (Zadrozny 2004; Cortes et al. 2008), which states the following.

Theorem 1 *For all distributions D, for all classifiers h and for any loss function l(h(X), y), if we assume that* $P(s|\boldsymbol{x}, y) = P(s|\boldsymbol{x})$ *(that is, S and Y are independent given* $\boldsymbol{x}$*), then*:
$$\mathbb{E}_{X,Y\sim D}[l(h(X), Y)] = \mathbb{E}_{X,Y\sim \widetilde{D}}[l(h(X), Y)|s = 1],$$
where $\widetilde{D} \equiv P(s = 1)\frac{D(X,Y,s)}{P(S=1|\boldsymbol{x})}$

2.3 Estimation of uncertainty

Machine Learning models often cannot rely on parametric/distributional assumptions on the data generation mechanism for directly deriving interval estimates of their predictions. Different approaches to the construction of non-parametric intervals have been proposed in the literature. One very common technique directly estimates the conditional quantiles of the response variable at the desired levels $\alpha/2$ and $1-\alpha/2$ by performing two separate quantile regressions[4]. While this should provide intervals that do guarantee the nominal coverage of $1-\alpha$, this method assumes unbiasedness of the estimates and neglects the propagation of model uncertainty at all stages of the procedure (i.e. the mean regression and the quantile regressions). Other alternatives rely onto empirical approaches to derive the error distribution of the predictions, such as: bootstrap on the observed set (Efron et al. (2003), Heskes (1997)); bootstrap on the prediction error (Coulston et al. 2016); fit of additional predictive models that target the observed predictions error (Shrestha and Solomatine 2006). Here, we consider a technique proposed in Kumar and Srivistava (2012), which can be seen as a direct extension to the non-parametric framework of the work originally proposed by Stine (1985) in the linear regression setting.

Let us assume the observed outcome Y_i can be expressed as:

$$Y_i = \phi(\boldsymbol{x}_i) + \epsilon_i,$$

where $\phi(\cdot)$ is the *true* and *unknown* function which links covariates and outcomes and ϵ_i is the sampling noise (or error term, independent of Y_i and possibly of $\boldsymbol{x}_i$). Training a predictive learner to provide predictions for the Y's given $\boldsymbol{x}$ means building an approximation $F^*(\cdot)$ to $\phi(\cdot)$ which is optimal in some terms (i.e. according to the chosen loss function). The obtained $F^*(\cdot)$ is just one realization of all possible approximations of the true underlying function $\phi(\cdot)$ we could get training the same learner on different sets of data $(Y, \boldsymbol{X})$. Intrinsically, it must be characterised by some *variability* and it is also likely to be *biased* (Hastie et al. 2009). The generic outcome Y_i can then be expressed as:

$$Y_i = F^*(\boldsymbol{x}_i) + \mathcal{S}_{F^*}(\boldsymbol{x}_i) + \mathcal{B}_{F^*,\phi}(\boldsymbol{x}_i) + \epsilon_i,$$

where $\mathcal{S}_F(\boldsymbol{x}_i)$ is the *model variance noise term* (or *model uncertainty*, arising from the variability of the estimated function) and $\mathcal{B}_{F,\phi}(\boldsymbol{x}_i)$ is the *model bias term.* In order to build meaningful and reliable intervals for the predictions, all these sources of error shall be properly quantified.

Following Kumar and Srivistava (2012), let us randomly build B *bootstrap* samples of arbitrary size n from the original train set. All samples are approximately distributed according the empirical (observed) joint distribution of Y and $\boldsymbol{X}$. The same algorithm shall be trained separately and independently on the B bootstrapped samples, yielding B different approximations of the function relating covariates and the expected value of the response variable $\left\{F_b^*(\cdot)\right\}_{b=1}^B$. Using each of these functions, we can get B sets of predictions on all the data. The whole procedure is outlined in Fig. 1.

Quantifying model uncertainty. The model uncertainty distribution can be trivially derived using only the B bootstrapped sets of predictions, independently

[4] This can be trivially achieved by setting the proper loss function.

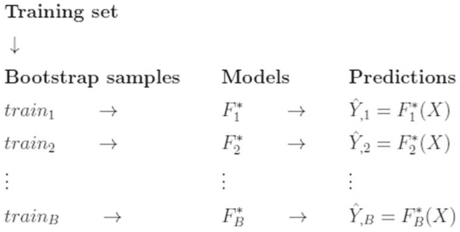
Training set
↓
Bootstrap samples
Models
Predictions
$train_1$ → F_1^* → $\hat{Y}_{,1} = F_1^*(X)$
$train_2$ → F_2^* → $\hat{Y}_{,2} = F_2^*(X)$
$train_B$ → F_B^* → $\hat{Y}_{,B} = F_B^*(X)$

Fig. 1 Bootstrapping procedure scheme. $\hat{Y}_{,i}$ represent the N-sized vector of predictions on the whole observed set of covariates X

from the observed values y_i. For B large enough, the average $\overline{F_B^*(\boldsymbol{x}_i)}$ is a good approximation of the expected value $\mathbb{E}[F^*(\boldsymbol{x}_i)]$. Centering the bootstrapped predictions about their mean, we then get B realizations from the model noise distribution (i.e. zero mean and variance $\mathbb{V}[F^*(\boldsymbol{x}_i)]$) for each observed covariate vector $\boldsymbol{x}_i$:

$$F_b^*(\boldsymbol{x}_i) - \overline{F_B^*(\boldsymbol{x}_i)} = \hat{\mathcal{S}}_{F,b}(\boldsymbol{x}_i) \sim \mathcal{S}_{F^*}(\boldsymbol{x}_i), \quad i = 1,\ldots,N, \; b = 1,\ldots,B.$$

These realizations can then be used to compute separately for each $\{\boldsymbol{x}_i\}_{i=1}^N$ the empirical quantiles at levels $\alpha/2$ and $1-\alpha/2$, and derive the desired intervals for each $F^*(\boldsymbol{x}_i), i = 1,\ldots,N$[5]. If $F^*(\boldsymbol{x}_i)$ were guaranteed to be unbiased, this would be sufficient to get intervals for $\phi(\boldsymbol{x}_i) = \mathbb{E}[Y_i|\boldsymbol{x}_i]$. However, we also need to quantify the bias $\mathcal{B}_{F,\phi}(\boldsymbol{x}_i)$ in order to get intervals for $\phi(\boldsymbol{x}_i)$ and the sampling noise to extend the intervals to Y_i.

Quantifying sampling noise and bias. Disentangling sampling error and bias is an impossible task. However, for the sole purpose of building prediction intervals, there is no counter-indication in quantifying their sum altogether. Let us denote the sum as $\upsilon(\boldsymbol{x}_i) = \mathcal{B}_{F,\phi}(\boldsymbol{x}_i) + \epsilon_i$. Considering for each $\boldsymbol{x}_i$ the bootstrapped averages $\overline{F_B^*(\boldsymbol{x}_i)}$, we get that:

$$\mathbb{V}\left[\overline{F_B^*(\boldsymbol{x}_i)}\right] = \frac{\mathbb{V}[F^*(\boldsymbol{x}_i)]}{B} \overset{B\to\infty}{\rightarrow} 0.$$

Therefore, for B large enough, the obtained bootstrap averages $\left\{\overline{F_B^*(\boldsymbol{x}_i)}\right\}_{i=1}^N$ are affected by a negligible amount of the model variance. Hence, the empirical errors $\hat{\upsilon}(\boldsymbol{x}_i) = y_i - \overline{F_B^*(\boldsymbol{x}_i)}$ are N realizations (i.e. one for each observation) from the distribution of $\upsilon(\boldsymbol{x}_i)$. One single realization for each $\boldsymbol{x}_i$ does not allow to get an approximation for the observation-specific distribution of $\{\upsilon(\boldsymbol{x}_i)\}_{i=1}^N$, but resorting to some simplifying assumptions it becomes possible to pool the realizations of $\upsilon(\cdot)$ at the different $\boldsymbol{x}_i$'s to get an overall estimate of bias and sampling noise.

- If the distribution of $\upsilon(\boldsymbol{x}_i)$ is independent from $\boldsymbol{x}_i$, which is to say $\upsilon_i \overset{iid}{\sim} f_\upsilon(\cdot)$, $\forall i = 1,\ldots,N$, all N errors $\{\upsilon(\boldsymbol{x}_i)\}_{i=1}^N$ can be used to approximate the common $f_\upsilon(\cdot)$ and compute the $\alpha/2$ and $1-\alpha/2$ quantiles.

[5] Note that this quantification is able to account also for possible heteroskedasticity of the outcome.

- If the independence assumptions of $\upsilon(\boldsymbol{x}_i)$ on $\boldsymbol{x}_i$ does not hold, which is to say $\upsilon(\boldsymbol{x}_i) = \upsilon_i|\boldsymbol{x}_i \overset{ind}{\sim} f_\upsilon(\boldsymbol{x}_i)$, $\forall i = 1,\ldots,N$, we still need to assume there exists a predictable pattern between them. In this case, we can still partially pool all the $\upsilon(\boldsymbol{x}_i)$'s together by fitting an additional regression model $f^*(\boldsymbol{x}_i) = \hat{\upsilon}(\boldsymbol{x}_i)$.

Combining the two sources of error. Finally, under the working assumption of independence between the *model variance noise* estimators and *model bias plus sampling error* estimators $\hat{\upsilon}_i$, we can obtain the quantile at any level p of the overall noise distribution as:

$$q_p(\boldsymbol{x}_i) = \hat{S}_{F,B}(\boldsymbol{x}_i)_p + \upsilon(\boldsymbol{x}_i)_p.$$

The confidence interval at level $1-\alpha$ for the predictions can be computed as:

$$C_{1-\alpha}(Y_i) = \left[F^*(\boldsymbol{x}_i) + q_{\alpha/2}(\boldsymbol{x}_i),\ F^*(\boldsymbol{x}_i) + q_{1-\alpha/2}(\boldsymbol{x}_i)\right].$$

Performances of this interval-building procedure are assessed in practice in the application of Sect. 3 and on the *toy-examples* of Appendix B.

3 Estimation of the Italian Value-Added Tax (VAT) gap

The reference population under analysis is composed of all the Italian *individual firms* that were included in Tax Register for the fiscal period 2007-2014. Tax auditing is performed yearly within every fiscal period on the same set of audited units, which usually spans a 7 years time. Since the analysis here performed dates back to 2018, we will be referring to the most recent year available at the time, which was 2011[6].

The whole population consists of $N = 2.3$ millions individual firms, where only the 0.82% have been audited (see Table 2). About 159 different features are available in the Tax Register, all concerning various area of information about the owner and its firm: personal data; economic sector of operation; taxable income and tax by type; revenues, expenses incurred, taxable base, gross and net tax; presumptive turnover provided by *Business Sector Studies*. A summary of the areas and categories of the available covariates organized by source is reported in Table 1. Unfortunately, a more detailed list cannot be provided for under the signed non-disclosure agreement.

The whole analysis is carried out using the open-source software R, exploiting the *tidyverse* logic for an easy and efficient data wrangling and management (Wickham et al. (2015), Wickham (2016)). Data in the Tax Register Database are extremely 'sensitive' and were analyzed under strict processing conditions. Only one personal computer was allowed to interface with the database and download the data, which

[6] An unavoidable delay occurs between the availability of the audit data and the fiscal year of reference. There is a lag between the fiscal (audited) year and the year in which the control is performed and the auditing process performed by the IRA requires (on average) two years to produce final data from the raw ones. Estimates on a specific tax year are usually available within six to seven years from the fiscal year of reference.

could not be moved in any case to external virtual machines. Therefore, the analysis has been performed on a single laptop with particularly limited computing capabilities[7]. This has no direct effect on the predictive performances, but has significantly slowed down the fitting time and, most importantly, has not allowed the contemporary consideration of the whole dataset of individual firms[8]. Therefore we consider only a sub-sample of all the not-audited units, stratified controlling for three main demographic variables: fiscal regime, regions and branch of economic activity (ATECO). All the audited units (low in numbers) have been included in the analysis. The final composition of the sample is reported in the second column of Table 2.

Sub-sampling on the not-audited taxpayers shall directly affect only first step of the modeling approach (Sect. 2.2), while impacts the second one only indirectly (i.e. estimation of inclusion probability could be less accurate and affect the selection bias correction). Nevertheless, the *under-sampling of the majority class* (Chawla 2009; He and Garcia 2009; Fernández et al. 2018) is one of the most common strategies to handle imbalanced classification tasks. Given the great imbalance between audited and not-audited taxpayers in our population, the necessary sub-sampling could be beneficial to the overall classification performances (Lee 2014; More 2016).

The aim of this paper is to verify the comparative performances of the *2-step GB* (introduced in Sect. 2) and the *standard Heckman model* (currently adopted by the IRA) for the estimation of individual potential tax bases BIT. Conclusions will naturally extend to the evaded tax bases (BIND) and the overall VAT gap. This straight-up comparison does not allow to single out whether it is the *Gradient Boosting* or the *2-step* strategy to mostly affect the final performances. Let us stress how our work is mainly motivated by the adoption of the *Gradient Boosting*. The 2-step correction is adopted as a tool to account for the sample selection bias in this alternative setting, and we expect it to have a minor impact. One of the reviewers keenly observed that performing further comparisons with semi-parametric versions of the Heckman model (Li and Racine 2007; Newey 2009; Wojtyś and Marra 2015; Wojtyś et al. 2016) would allow dissolving this doubt. This point is indeed relevant and deserves further investigation.

The flow of the analysis for both models is resumed in Fig. 2.

The raw data have undergone a common pre-processing step. This encompassed joining the information derived from different sources, cleaning the uninformative variables (i.e. variables with unary values or high proportion (more than 80%) of missingness), and imputing the missing values for the remaining ones (using mean-subsitution). Further data manipulation (e.g. standardization, log-transforms, etc.), currently adopted by the IRA to favor the performances of the Heckman model, has been considered also in our setting. Let us point out that it has not sensibly affected the performances of our model[9].

[7] Processor: Intel Pentium dual-core E1040; RAM: 4gb.

[8] RAM was too limited.

[9] The Gradient Boosting is not highly affected by the same data criticism that plagues the standard linear regression setting, e.g. multicollinearity, skewness, deviations from Normality assumptions, etc. (Efron and Hastie 2016).

This version of the dataset is then provided as an input to our method. Observations have been split in a 70%/30% proportion between train set and test set, controlling for the *audited* and *not-audited* variable (see Table 3).

On the other hand, the cleaned version of the dataset provided as an input to the Heckman model contains only a subset of all the available features as selected by the IRA). Unfortunately, we are not allowed to disclose what and how many variables are actually considered, but we can share the IRA guidance on the feature-selection process. First, a comprehensive set of relevant variables is selected by considering different criteria: prior knowledge on the phenomenon (i.e. some variables are kept because presumed to be important for economic reasons); bi-variate dependence and correlation scores between the feature and the target (chi-square test, t-test, F-test, etc.); multivariate selection based on variable importance (decision trees). Second, the *Variance Inflation Factors* (VIF) of a probit and a linear regression coefficients are checked in order to control for multi-collinearity and drop additional features.

At this point, observations are split in train and test according to the same seed, so that the Heckman model is trained and tested on the same units of the 2-steps GB and results are comparable.

3.1 2-steps GB fitting and uncertainty assessment

The procedure outlined in Sect. 2.2 involves the subsequent application of two different CART-based gradient boosting algorithms: a classification model targeting the auditing probabilities π_i and a weighted regression model targeting the potential tax base BIT_i. These have been implemented in the software `R` using functions from the package `gbm` (Ridgeway 2007). In particular, the function `gbm` from the homonym package fits the model to a train set and gives control over a variety of arguments.

A very important role is played by the `bag.fraction`, i.e. the fraction of randomly selected train set observations used to propose the next tree in the expansion. A value lower than 1 introduces randomnesses into the model fitting and prevents overfitting or getting stuck in local maxima[10]. The user can pick different regularization strategy to evaluate the loss function at each iteration. If `bag.fraction` is lower than 1, then *out-of-bag* validation (*bagging* (Breiman 1996)) gets feasible; alternatively, one may select a `train.fraction` lower than 1 (proportion of units in the train set) and perform *hold-out* validation; finally, one may set a number of `cv.fold` > 2 to split the train set in different *folds* and perform a complete cross-validation. As by default, our implementation exploits the *out-of-bag* validation as a good compromise between speed and accuracy. The `bag.fraction` value has been set to 0.5 for both steps.

Finally, the fit of any CART-based GB depends on some major tuning parameters. The most relevant are: the number of iterations *n.iter*, the depth of each single tree d, the minimum number of observations in the final nodes of each tree, and the learning parameter λ. They are not directly estimated in the fitting process,

[10] This is a non-parametric implementation of the *stochastic gradient descent*.

Table 1 Summary of the main covariates obtained from the Tax register and the Tax audits, organized by category

Source	Category	Features
Tax register	Personal data	Age (individual)
		Gender (individual)
		Fiscal regime (individual)
		Branch of economic activity (firm)
		Region (firm)
	IRAP	Amortization
		Operating costs
		IRAP total revenues
		IRAP total tax
		Net production value
		Revenues
	PIT	Personal Income Tax
		Labour cost (amount of)
		Negative components of income (costs and expenses)
		Revenues from activities
		Gross income
		Income
		Total remuneration
		Tax deductions
		Input/Output Tax
		Other incomes
		Total expenses
		Profit
	VAT	Operations generating VAT
		Value Added Tax
		Purchases and imports
		Total VAT credit
		Volume of business (BID)
		Input/Output Tax
		Taxable transactions
Tax audits	Audit	Potential volume of business (BIT)
		Undeclared volume of business (BIND)
		Assessment indicator variable

Table 2 Total and sampled population of individual firms

Fiscal audits	Total population		Sub-sample	
	Frequency	Percentage	Frequency	Percentage
Not-audited	2′275′219	99.18%	45′489	70.85%
Audited	18′718	0.82%	18′718	29.15%
	2′293′937	100%	64′207	100%

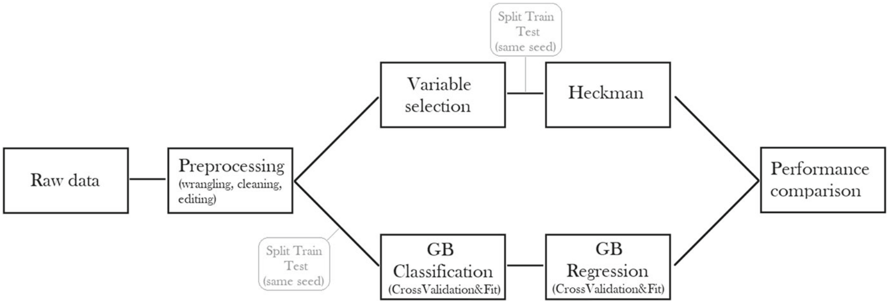

Fig. 2 Different steps of the analysis starting from the raw data up to the implementation and performance comparison between the two models

Table 3 Train and Test set sizes

	Train set	Test set	Total
Not-audited	$31'836$	$13'653$	$45'489$
Audited	$13'064$	$5'654$	$18'718$
	$44'900$	$19'307$	$64'207$

but must be fixed before running the algorithm. An accurate choice of these parameters is key to avoid over-fitting and get good out-of-sample predictive performances. Typically, a fixed grid of different parameters combinations is built, and the combination returning the best performances (according to some metric) is chosen. The refinement and extension of the grid must take into account the computational time to fit and validate all the alternatives. Given the computational limitations, we had to limit the tuning to a very rough grid in both steps. The performances for different combinations have been validated using a naive (but efficient) *hold-out* logic.

First step. The classification model of the first step consider the whole sub-sampled population of $64'207$ units, with the audited/not-audited variable as target ($45'489$ not-audited and $18'718$ audited). The final outcomes are approximations $\{\hat{\pi}_i\}_{i=1}^N$ to the inclusion probabilities $\{P(i \in s)\}_{i=1}^N$ of each unit. The number of minimum observations in each final node has been fixed to the default value of 10, while the tested *n.iter*, d and λ belonged to the following sets:

$$n.iter \in \{30, 40, \ldots, 1000\}, \qquad d \in \{2, 3\}, \qquad \lambda \in \{0.01, 0.02, \ldots, 0.1\}$$

The metric chosen to evaluate the model performance in this step is the *Area Under the Curve* (AUC, Fawcett (2006)) score on the test set . The optimal choice returned an AUC value of 0.8 and was associated to the set of parameters:

$$\{n.iter_{opt} = 1000,\ d_{opt} = 2,\ \lambda_{opt} = 0.1\}\ .$$

The gbm function automatically returns variables scores of their importance in the fitting process (roughly speaking, the percentage of splits they determined). The variables detected as *most discriminating* are the declared tax base (BID), the activity branch, the dimension, and the incomes of the firm. This pattern is evident also by performing some basic descriptive analysis on the available data. For instance, Table 4 reports some distribution summaries of the BID separated in audited and not-audited units. The difference between two groups is really significant (t-test $185'295.75$ vs $92'834.80$; $p<0,001$). The tax-audit selection criteria is skewed toward taxpayers with high value of the BID and highlights the inclination of the IRA to audit firms with larger business volumes.

Second step. The regression model of the second step considers the $18'718$ audited units, with the BIT_i as target variable. Each observation in the train set is weighted by the inverse of the predicted inclusion probability resulting from the previous step. This can be done by utilizing the argument `weights` in the `gbm` function. In practice, when computing the loss function, the error on each unit is weighted by:

$$v_i = \frac{1}{\hat{\pi}_i}, \quad \forall\, i \in s.$$

The optimal parameters have been chosen by minimazing the *Mean Squared Error* (*MSE*) index on the test set. For a clearer interpretation of the results, the MSE can be divided by the variance of the test-set and subtracted to 1. This yields a pseudo-R^2 measure of goodness of fit:

$$\widetilde{R}^2 = 1 - \frac{MSE}{\sigma^2_{y_{te}}},$$

where $\widetilde{R}^2 \approx 1$ implies a perfect fit, $\widetilde{R}^2 \approx 0$ (or <0) denotes a bad fit. The best value obtained for the $\widetilde{R}^2$ on the testing set is 0.828, with tuning parameters:

$$\{n.iter_{opt} = 380,\ depth_{opt} = 2,\ \lambda_{opt} = 0.1\}$$

The predictive performances of the 2-step GB and the Heckman model (fitted on the same train set) are visually compared in Fig. 3. The largest errors committed by the 2-step GB are lie in the bottom-right section of the left plot in Fig. 3. These are characterized by a substantial underestimation of the true potential tax base for firms with unexpectedly large BIT. These errors are in common with the standard Heckman model, as the same points are clearly visible also in the bottom right section of the right plot in Fig. 3. Apparently, the large BIT of those units cannot be explained in terms of the available covariates by neither of the two models. At the same time, the Heckman model shows an evident overestimation pattern that is absent in the 2-step GB. This behavior is typical of when the linearity assumption is violated or the trend changes at different sections of the covariates space. This impacts the bias and variance of the final prediction function, that lacks the sufficient flexibility to accommodate any of the two behaviors.

The accuracy of the predictions of both models are compared also (and most especially) on the test set (see Table 5). While the aggregate estimates of the total BIT are very close to each other, it is possible to notice strong differences in terms

Table 4 Descriptive statistics on the distribution of the BID for audited and not-audited units

		Not-audited	Audited
Mean		92′834.80	185′295.75
Median		42′176	73′362
Standard deviation		230′980.72	387′869.98
Percentiles	25	19′736	30′532.75
	50	42′176	73′362
	75	88′589	169′942.50

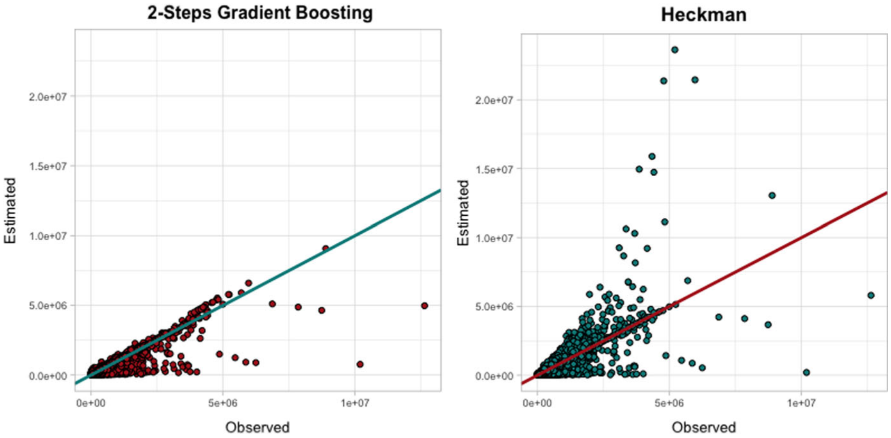

Fig. 3 Comparison of observed and predicted values of the BIT for the 2-step GB (left) and the Heckman model (right), on all the audited units

of individual accuracy. Indeed, the $\widetilde{R}^2$ of the Heckman model is equal to 0.657, which is sensibly lower value than the one 0.828 achieved by the 2-step GB.

It is thus evident how the non-linearity of the CART-based GB is extremely beneficial to better capture the complex patterns relating the covariates and the outcome. A possible halfway solution would be to consider an Heckman model with non-linear effects, such as *local polynomials*, *smoothing splines*, or *Generalized Additive Models* (Li and Racine 2007). These have already been successfully considered in other contexts (Newey 2009; Marra and Radice 2013; Wojtyś and Marra 2015), and future research efforts shall be devoted to investigate their merits and to compare them with our proposal.

Interval estimates. Interval estimates for observations in the train and test sets have been produced following Sect. 2.3 on the second step of the procedure, while keeping the first step fixed. This choice has been driven by the need to reduce the overall computational cost, which was especially high for the classification task. The number of bootstrap samples has been set to $B = 100$, each with size $n_B = 13'064$. Two alternative models for the bias and sampling error $\upsilon(\boldsymbol{x}_i) = \mathcal{B}_{F^*}(\boldsymbol{x}_i) + \epsilon_i$ have been considered. The first assumes that the distribution of $\upsilon(\boldsymbol{x}_i) = \upsilon_i$ is independent of $\boldsymbol{x}_i$, implying constant bias over the whole set of units. The

Table 5 Observed and estimated values of the total undeclared tax base BIND and the potential tax base BIT (in millions) on the test set (1st and 2nd rows); $\widetilde{R}^2$ (3rd row)

	Observed	2-step GB	Heckman
BIND	314.364	292.400	340.375
BIT	1′315.864	1′293.901	1′341.875
$\widetilde{R}^2$		0.828	0.657

second allows the conditional mean of υ_i on $\boldsymbol{x}_i$ to depend on the predicted value $F^*(\boldsymbol{x}_i)$:

$$\mathbb{E}[\upsilon_i \,|\, \boldsymbol{x}_i] = \beta_0 + \beta_1 \cdot F^*(\boldsymbol{x}_i),$$

where β_0 and β_1 can be estimated through *Ordinary Least Square* (OLS). Confidence intervals for the conditional mean $\mathbb{E}[\upsilon_i \,|\, \boldsymbol{x}_i]$ can then be directly derived in the OLS setting.

Let us denote the prediction intervals at level $\alpha = 0.95$ as $C_{0.95}(\boldsymbol{x}_i) = \left(\hat{l}_{0.95}(\boldsymbol{x}_i),\, \hat{u}_{0.95}(\boldsymbol{x}_i)\right)$. Comparative performances of the resulting intervals are compared on the whole set of audited units s, on the train set s_{tr}, and on the test set s_{te}. The relevant metrics are the *coverage* and the *average width*, defined respectively as:

$$Cov_{0.95} = \sum_i \mathbb{I}_{\left(\hat{l}_{0.95}(\boldsymbol{x}_i),\, \hat{u}_{0.95}(\boldsymbol{x}_i)\right)}(y_i)$$

$$\bar{W}_{0.95} = \frac{1}{N}\sum_i \left(\hat{u}_{0.95}(\boldsymbol{x}_i) - \hat{l}_{0.95}(\boldsymbol{x}_i)\right)$$

Results are summarized in Table 6. Both approaches provide a satisfying coverage, which is really close to the nominal level of 0.95. The linear model on υ_i presents slightly lower coverage in all sets, but this is more than compensated by the greatly improved performances in terms of average width. It is also interesting to notice how performances are very similar across all the sets considered (all the audited, only the train set, only the test set). This is reassuring in terms of the transferability of the same performances on the non-audited set (which cannot be tested because of the unobserved true outcome). While the results are more than satisfactory, highlighting a good robustness of this technique, it is strongly suggested to consider both steps in the interval building procedure. This would propagate the uncertainty up to the final estimates, and potentially recover the 2 nominal coverage points lost in the bias linear model.

It may be of primary interest obtaining an interval estimate for the overall BIND, verifying if the observed value on the test set is contained within the interval bounds. However, there is no way to quantify the aggregate *model bias and sampling error* on one single observation. We thus account only for the model variance, and obtain the intervals as they are reported in Table 7. The observed

Table 6 Intervals evaluation metrics for the two proposed approaches: constant bias VS bias increasing linearly with the prediction value

Metric	Bias model	Set		
		Audited	Train set	Test set
$Cov_{0.95}$	Constant	0.960	0.961	0.957
	Linear	0.93	0.931	0.928
$\bar{W}_{0.95}$	Constant	224’961	224’977	224’924
	Linear	163’091	220’081	162’120

Table 7 Intervals for BIND and respective true observed value, in Millions

Parameter	Set		
	Audited	Train set	Test set
BIND	1’041.415	727.051	314.364
$C_{0.95}$	(932.371, 1’064.755)	(655.418, 757.978)	(278.110, 317.081)

value of the BIND, either in the train, test and complete set, is within the estimated bounds.

In terms of computational burden, the complete bootstrapping procedure took approximately 18 hours (with $B = 100$) on the considered hardware. Greater values of B can only improve the interval approximations, with theoretical justification on the optimal choice of B in Hall (1986), but require more computational power (or time). Generally speaking, increasing the size of B is not a big deal: the computational time increases linearly with B and it could be drastically reduced by using a better performing processor, or by parallelizing the procedure on a reasonable number of cores.

3.2 An estimate of the revenue loss and VAT evasion intensity

The two models are then used to produce predictions for all the units in the not-audited population.

We here focus on the results from the 2-step GB, being the one with better performances in all metrics. Given the stratified structure of the sample, we trust that these estimates will reflect properly the behaviour in the general population. The predicted *VAT gap* turnover $\widehat{\mathrm{BIND}} = \widehat{\mathrm{BIT}} - \mathrm{BID}$ is of about 3.36*Bln* of euro (3′360′930′741€), with an interval of [3.03*Bln* €, 3.56*Bln* €].

Another quantity of interest is the so-called *VAT evasion intensity*, here defined on the line of the evasion intensity used in Braiotta et al. (2015). A synthetic measure p of it is defined as the ratio between the undeclared tax base and potential tax base:

$$p = \frac{\sum_{i=1}^{N} \mathrm{BIND}_i}{\sum_{i=1}^{N} \mathrm{BIT}_i}.$$

If values at the individual level are available, it is also possible to compute the unit specific evasion intensity as:

$$p_i = \frac{\mathrm{BIND}_i}{\mathrm{BIT}_i}, \qquad i = 1, \ldots, N.$$

Low values of the ratio indicate a compliant behaviour, and viceversa.

The observed VAT evasion intensity on the audited taxpayers is of $p = 23.09\%$. The estimated $\hat{p}$ on the audited set obtained from the 2-step GB and the Heckman models are 22.12% and 25.89%, respectively. The former provides a result way closer to reality than the latter Using the predictions on the entire sub-sample of taxpayers, the estimated intensity $\hat{p}$ is 30.40% for the 2-step GB and 29.77% for the Heckman model. These estimates are in line with the quantification of the Tax Gap built using top-down approaches by the Italian Ministry of Economics and Finance for the same years, as reported in MEF (2014), OECD (2016). It is very interesting to notice how the evasion intensity is estimated to be larger on the whole sample than on the set of audited units. At the same time, the observed average evaded by the audited taxpayers amounts to $55'637.09$, while the average evaded in the whole sample estimated by the GB is equal to $52'345.24$. This suggests that the Italian Revenue Agency is not selecting the individuals with the larger evasion intensity. Apparently, the Italian Revenue Agency favors the auditing of individual firms with large business volume and hence the larger potential to evade large amounts (in absolute terms). It is less interested in individual firms with small business volume, even supposing they present a large evasion intensity. This is probably because each single small individual firm, even if non-compliant, represent only a small percentage of the total revenue loss.

Individual level predictions allow also the estimation of the evasion intensity of specific group of taxpayers p_c. They may be used to identify classes of individuals at high-risk of evasion, and may be of help in the selection procedure of future fiscal audit. The intensity related to a specific class of individuals c is straightforwardly estimated as:

$$\hat{p}_c = \frac{\sum_{i \in c} \mathrm{BI\hat{N}D}_i}{\sum_{i \in c} \hat{BIT}_i}, \quad c \in \mathcal{P}(\{1, \ldots, N\}),$$

where $\mathcal{P}(\{1, \ldots, N\})$ is the power set of the the population indices.

Unfortunately, description of the results is limited by the confidential nature of the information. We are allowed to show the results related to the intensity by age class, just as an example. The observed intensity on the audited taxpayers highlights a decreasing trend by age. The 2-step GB returns results that are coherent with the observed values, both in terms of magnitude and direction. The Heckman model over-estimates the intensity in all classes and, most importantly, does not recover the true age trend in the older classes (see Table 8). Table 9 shows how the estimated pattern on the whole sub-sample of taxpayers mimics the one on the

Table 8 Observed and estimated evasion intensity by age classes on the audited taxpayers

Age	Size	Observed			Gradient Boosting			Heckman		
		BIT Bln	BIND Bln	Prop	BIT Bln	BIND Bln	Prop	BIT Bln	BIND Bln	Prop
[18 – 25)	270	0.06	0.02	25.24%	0.06	0.01	23.97%	0.07	0.02	28.55%
[25 – 45)	7876	1.71	0.43	24.91%	1.69	0.41	23.92%	1.76	0.48	26.98%
[45 – 65)	9275	2.35	0.52	22.25%	2.32	0.49	21.28%	2.42	0.59	24.71%
over 65	1297	0.39	0.08	19.84%	0.38	0.07	18.69%	0.43	0.12	27.67%
Total	18718	4.51	1.05	23.09%	4.45	0.98	22.12%	4.68	1.21	25.89%

Table 9 Estimated evasion intensity by age classes in the not-audited taxpayers

Age	Size	Gradient Boosting			Heckman		
		BIT Bln	BIND Bln	Prop	BIT Bln	BIND Bln	Prop
[18 – 25)	976	0.13	0.05	39.07%	0.13	0.04	36.71%
[25 – 45)	28250	4.26	1.45	34.10%	4.11	1.30	31.61%
[45 – 65)	30496	5.64	1.60	28.45%	5.64	1.60	28.37%
Over 65	4485	1.02	0.25	24.65%	1.08	0.32	29.28%
Total	64207	11.05	3.36	30.40%	10.95	3.26	29.77%

audited set. In particular, the 2-step GB emphasizes differences between classes: the 7 points gap between the youngest and the oldest class estimated by the Heckman model becomes a 15 points gap using the 2-step GB.

4 Concluding remarks

This work presents a technique for the bottom-up estimation of the Italian Value-added tax (VAT) gap. It draws from the class of non-parametric Machine Learning models and embeds a solution to correct for the selection bias due to the not-random auditing strategy. The method most widely used in tax evasion estimates with risk-based audit data is the standard Heckman model, currently implemented in lot of countries all around the world (including Italy). The form currently adopted by the IRA relies on linear models to describe both the selection mechanism of audits and the individual potential tax base. This limits the covariates' basin that this model can consider, but most especially hinders its predictive performances. Real data collected for administrative purposes often exhibit complex behaviors and dependence patterns, which are hardly matched by the linearity assumptions or Gaussianity. National agencies could leverage more recent techniques and a larger portion of their data to detect high value tax evasion behaviors, define new tax evasion patterns, and identify organized tax evasion networks. The OECD guidelines of these last few years (see http://www.oecd.org/going-digital/ai/ and Chapter 5 of OECD (2018)) continuously encourages the use of innovative machine learning and AI methods to face these challenges. This does not necessarily mean replacing the traditional approaches, but integrating the two in order to get a wider horizon of possibilities to improve on policy-making.

The 2-step GB we are proposing drives toward this direction. It is based on the CART-based Gradient Boosting algorithm, a powerful ensemble machine learning technique that proved to perform very well over a wide variety of datasets. It can detect non-linear and complex patterns in the data, it does not rely on strict distributional assumptions, and it can exploit covariates of any size and nature. We combined it with a re-weighting strategy to non-parametrically correct for the selection bias of tax-audits, and we showed how it can substantially improve on the predictive performances of a standard Heckman model (despite our computing limitations).

As far as our application is concerned, the 2-step GB provided a better account of variability in the observed potential tax base. The current standard (i.e. standard Heckman model), on the contrary, flattened out most of the individual differences. The potential predictive ability on the non-audited set could not be verified on the real set of data, as the outcome on not-audited units is unknown. Therefore, we proposed a toy example with a simulated non-random selection in Appendix B. The proposed method was able to retrieve information from the artificially biased train set and sensibly outplayed the Heckman model in terms of predictions on the un-selected (unobserved) set of data.

The improved predictive power of the proposed method is extremely relevant from many perspectives. The more accurate individual predictions can be used to efficiently target potential evaders. For instance, deriving accurate evasion intensity scores for all the taxpayers can help in the detection of units who are likely to hide a large part of their incomes. This can be used to drive future selection audits, simplifying the recover of the TAX revenue loss. Furthermore, it can help to frame the significant challenges and opportunities facing tax administrations to better manage compliance. As an example, the diagnostic of the 2-step Gradient Boosting decision rules highlighted how the audit selection process performed by the Italian Revenue Agency seems to not favor the selection of the *less-compliant* individuals in relative terms, but in absolute terms. This choice looks reasonable in light of the limited resources of tax administration offices and the consequent possibility to check only a small portion of the whole population. However, it consistently neglects the (potentially very large) portion of population of small-evaders that do not report the most of their incomes.

Section 2.3 built upon established literature to propose a strategy for building confidence intervals in the considered setting. This provided more than satisfying results in terms of coverage (on the audited units) and interval width, which is sufficiently small as compared to the magnitude of the predicted values. Reassurance about the goodness of performances on the not-audited units are obtained in the toy example of Appendix B.

The possible further developments of this work are various and promising. As pointed out by one of the reviewers, National institutions like Bank of Italy or Istat are recently leveraging Machine Learning to exploit the rich sources of unstructured data (e.g. free text) available in the *Big Data* era. These can be profitably used to draw relevant insights on various social and economics aspects (i.e. consumers behavior, population in general, macro-economic figures, etc.), but their use has been so far hindered by the limitations of classical modeling in many application fields. The adoption of machine learning techniques can favor the employment of such unstructured data in this context, where useful patterns characterizing firms at higher risk of evasion could be found in the auditors written report, on the web, social media, etc.

Let us recall that the analysis exposed in this work has been performed only on a small subset of all the available observations because of hardware limitations. We are confident that way better results may be achieved by analyzing the whole set of data. The improved computational power would allow the application of a complete k-fold cross-validation, possibly on finer grids of parameters, and find a better combination.

This work shows how linearity is a too strict assumption for achieving good predictive performances in complex data such as those resulting from tax-audits. We proposed one possible method to overcome this issue, but there are many others that could achieve the same objective. Future studies may consider alternative non-parametric learners for one or both steges of the 2-step method (such as *Random Forests*, *Extreme Gradient Boosting*, *Support Vector Machines*, *Neural Networks* etc.). Or, as mentioned in the main text, they could investigate the potential of the semi-parametric Heckman model, i.e. with a mean term modeled as a *Polynomial and Spline regression* (Newey 2009; Marra and Radice 2013; Wojtyś et al. 2016) or *Generalized Additive Models* (Wojtyś and Marra 2015; Hastie and Tibshirani 2017). It would be extremely interesting to compare the performances of the 2-step method with these latter alternatives, as they parametrically embed a Heckman-style correction. This could shed light on the comparative merits of the alternative sample bias correction methods in this specific context.

A The Gradient boosting algorithm

The *Gradient Boosting* is a very powerful algorithm that allows building predictive models for both the classification and regression tasks. It is an ensemble algorithm that relies on the concept of boosting, which is a technique for reducing bias and variance in supervised learning, firstly introduced in the seminal paper of Schapire (1990). The *Gradient* in front of the term *Boosting* refers to a very flexible formulation of the boosting, firstly proposed by Friedman (2001). This particular version exploits the *Gradient Descent* in order to robustify and hasten the optimization procedure on the loss function.

Let us consider the usual set of covariates $X = \{x_1, ..., x_N\} \in \mathcal{X}$ and the response variable $Y \in \mathcal{Y}$. The final aim of any supervised learning algorithm is to train itself on a set of data $\{X_i, Y_i\}_{i=1}^{N}$ whose covariates and response variables are known, and then produce an approximation $F^*(x)$ to the function $F(X) : \mathcal{X} \rightarrow \mathcal{Y}$ that generally relates X and the expected value of $Y|X$. The approximation is obtained in such a way that the expected value of a pre-specified loss function $L(Y, F(X))$ is minimized with respect to the joint distribution of all the observed pairs (X, Y). In practice, the algorithm learns from the examples provided to it in the form of a train set and it looks for that approximation F^* such that:

$$F^*(X) = \arg\min_F E_{Y,X} L(Y, F(X)) = \arg\min_F E_X[E_y(L(Y, F(X)))|X].$$

The choice of the loss function depends on the nature of the problem and of the outcome variable. For instance, in the case of the regression task, the usually adopted loss function $L(Y, F(X))$ is the *Mean Squared Error*. The peculiarity of the boosting procedure is that it approximates $F(X)$ using a function of the form:

$$F^*(X) = \sum_{m=0}^{M} \beta_m h_m(X)$$

where $h_m(X)$ are functions known as *Base Learners* and $\{\beta_m\}_0^M$ are real coefficients. The base learners are functions of X derived from another, simple, learning algorithm and the β's are expansion coefficients used to combine the base learners outcomes. Either the base learners and the expansion coefficients are estimated using the data from the train set using a *forward-stagewise* procedure. As any recursive algorithm, it starts from an initial guess $F_0(X)$ and then the new set of coefficients and learner are derived as:

$$(\beta_m, h_m) = \arg\min_{\beta,h} \sum_{i=1}^{N} L(Y_i, F_{m-1}(X_i) + \beta h(X_i)) \quad \forall\, m = 1, ..., M \tag{8}$$

and

$$F_m(X) = F_{m-1}(X) + \beta_m h_m(X_i, a))$$

Algorithm 1 Gradient Boosting pseudo-code example

```
$F_0(X) = \arg\min_{\rho,\beta} \sum_{i=1}^{N} L(y_i, \beta)$
for $m = 1$ up to $M$ do
    $\widetilde{Y}_{i,m} = -\left[\frac{\partial L(Y_i, F(X_i))}{\partial F(X_i)}\right]_{F(X)=F_{m-1}(X)}, \forall\, i \in \{1, ..., N\};$
    $h_m = \arg\min_h \sum_{i=1}^{N} L\left(\widetilde{Y}_{i,m} - h(X_i)\right)$
    $\beta_m = \arg\min_\beta \sum_{i=1}^{N} L(Y_i, F_{m-1}(X_i) + \beta h_m(X_i))$
    $F_m(\mathbf{x}) = F_{m-1}(X) + \lambda\beta_m h_m(X_i))$
end for
```

Unfortunately, choosing the best pair (β_m, h_m) at each step for an arbitrary loss function is a computationally infeasible optimization problem in general. This is where the gradient descent plays a key role, leading to the *Gradient Boosting* algorithm. It solves the optimization problem in Eq. 8 through an approximation that is legitimate whenever the loss function $L(Y, X)$ is differentiable. At each step

$m = 1, \ldots, M$, the base learner $h(X)$ is chosen according to the best fit on the *pseudo-residuals* $\{\widetilde{Y}_{i,m}\}_{i=1}^{N}$, deriving from the previous step:

$$h_m = \arg\min_h \sum_{i=1}^{N} L\big(\widetilde{Y}_{i,m} - h(X_i)\big),$$

where:

$$\widetilde{Y}_{i,m} = Y_i - \rho F_{m-1}(X), \qquad i = 1, \ldots, N, \ \rho \in \mathbb{R}^+.$$

The pseudo-residual values $\{\widetilde{Y}_{i,m}\}_{i=1}^{N}$ play the role of the gradient, driving the optimization procedure towards the right direction step after step. In this simplified framework, given the base learner $h_m(X)$, the best value β for β_m can be obtained as:

$$\beta_m = \arg\min_\beta \sum_{i=1}^{N} L(Y_i, F_{m-1}(X_i) + \beta h(X_i)).$$

A very common modification to the standard gradient boosting algorithm includes a *shrinkage parameter* λ, which modifies the update rule in the following way:

$$F_m(X) = F_{m-1}(X) + \lambda \beta_m h_m(X_i, a)), \qquad m = 1, \ldots, M, \ \lambda \in [0, 1].$$

This parameter controls the learning rate of the algorithm and while introducing some bias, allows for the regularization of the procedure and avoids over-fitting (Efron and Hastie 2016). The whole algorithm is resumed in the pseudo-code Algorithm 1.

The most common version of the Gradient Boosting uses fixed-size CART (usually small, with low number of branches and/or splits) as base learners, whose predictive ability is strongly enhanced by their boosting combination (Efron and Hastie 2016). Either the shrinkage parameter and the parameters that define each single random tree (number of splits, number of branches, etc.) are not estimated during the procedure. In the Machine Learning context they are known as *tuning parameters* and they need to be chosen in advance and stay fixed. Typically, they are selected via searching procedure based on the *cross-validation* in order to avoid over-fitting (Hastie et al. 2009).

B Toy example: houses to rent

We want to verify comparative performances of the Heckman model and the proposed 2-step GB on a set of data in which we artificially induce a *selection bias*. The considered dataset has been downloadeded from Kaggle[11] and it contain various information about houses to rent in different cities in Brazil (Rubens 2020). It contains 10′962 records, each with thirteen variables: two about the location of the house (city and area), six house-specific features (number of rooms, bathrooms, ...),

[11] https://www.kaggle.com.

and five economic amounts (homeowners association tax, rent amount, property tax).

We will consider the property tax, that is to say the annual tax the landlord should pay on the considered property, as our outcome variable of interest. It has been converted to the log-scale so that the Heckman linear predictor could respect its natural domain. The same log-transform has been considered also for the 2-step GB in order to guarantee the fair comparison of the two methods.

A preliminary exploratory data analysis and a brief data cleaning step is required in order to favor proper fitting of both the models. The variable *area* has been excluded from the analysis because it presents too many modalities (exactly 514), of which many only have 1 or 2 observations. The variable *floor* is not available in the 23% of the records and its value has been imputed using by mean substitution (complete average). The *property tax* is not available for 1596 observations and, being this the outcome variable, such records have been omitted from the analysis. Finally, all the economic features present some very extreme outliers, which can be probably attributed to recording errors. In order to detect in an automatic way such values, all these variables have been converted to the log-scale and all instances trespassing the $q_{0.75} + 1.5 \times \text{IQR}$ threshold (17 records) have been deleted[12].

The resulting version of the dataset counts 9079 records. Now, we need to select a sub-sample of observations to play the role of the *audited taxpayers* (both covariates and outcome known), while the remaining the role of *not-audited taxpayers*. The latter can be used to test the predictive abilities of the two models on the set where the outcome is indeed unknown. This selection must be performed in a way as similar as possible to how the Italian Revenue Agency selects the taxpayer to audit, i.e. by trying to maximize the selection of units with the highest outcome according to the available information. To this end, we realized that the *Homeowner Association Tax* (HOA) is a good proxy for the *property tax* (PT). This association is visible in the data, which show a correlation of ≈ 0.55 on a linear scale and of ≈ 0.67 on the log-scale. We decided to pick a cut-off point, let's say the 90th percentile of the HOA, and audit all the units with HOA greater than the selected cut-off point plus a small sample of units selected at random (5% of the total). The HOA is then eliminated from the dataset. The two final groups are respectively composed of $1'304$ (audited) and $7'775$ (not audited) records. Figure 4 shows how our covariate-based selection mechanism induced a strong selection bias on the outcome of interest.

In order to test the performance of the final predictive models also on the "audited" units, only the 70% of those are used in the training process while the remaining 30% are kept out as testing set.

The Heckman model is fitted to the data through the function `heckit` from the `sampleSelection` package available on the CRAN, which provides estimation routines for a variety of tobit-like models. The *2-step GB* is fitted through the function `gbm` from the `gbm` package available on the CRAN. 5-folds cross-validation (argument `cv.folds` of the `gbm` function) has been performed on the same grid of the original application introduced in Sect. 3.

[12] IQR is the *interquartile range*: $q_{0.75} - q_{0.25}$.

The final prediction accuracy has been measured in terms of *Mean Squared Error* (MSE) on the test portion of the "audited" units and on all the "not-audited" units. In order to re-scale by the *prediction difficulty* of each set, also a measure of Relative Mean Squared Error (RelMSE) is provided, where the MSE is divided by the variance of the observations in the test-set:

$$RelMSE = \frac{MSE}{\sigma^2_{y_{te}}}.$$

Results are summarized in Table 10.

We can clearly notice how the Gradient Boosting largely outperforms the Heckman model, producing sensibly lower errors (especially on the set of not-audited units). A graphical comparison of the prediction accuracy on the audited test set and not-audited units is provided in Figs. 5 and 6, respectively. The Heckman model predictions on the not-audited set seem to be strongly biased. It estimated completely flawed coefficients for some variables, as they provided good fit on the audited set but not on the not-audited set. Table 11 shows some of the most flawed coefficients as they are compared with the ones we would have got with a standard linear model on the whole set of units (audited and not-audited together). The 2-step GB, thanks to its more flexible structure, is instead able to adapt to the behavior at the borders of the observed covariates domain and keep more robust performances in the two sets.

We also verify the ability of the proposed confidence interval construction to guarantee the nominal coverage for both the "audited" and the "not-audited" units. In particular, intervals are built assuming constant bias and their performances are resumed in Table 12. Metrics are slightly better for the audited group, but coverage is close to the nominal level in both cases. The average interval width is satisfactory as compared to the magnitude of the outcome ($\bar{y}_{nsel} = 4.95$ and $\bar{y}_{nsel} = 6.3$).

We can then conclude that the non-linearity and distribution free-ness of the Gradient Boosting, combined with proper weighting to account for the non-random selection of the training units, provides a significant improvement in predictive performances. This is especially true whenever it can be reasonably assumed an

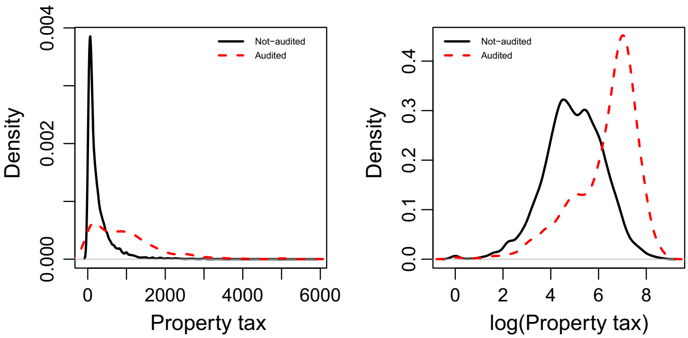

Fig. 4 Distribution of the outcome variable in the "audited" and "not-audited" groups on the linear (left) and log (right) scales

Table 10 Error metrics of the two considered models on the not-audited set (all out-of-sample) and the *audited test-set* units, in terms of MSE and RelMSE

Model	Not-audited		Audited test set	
	MSE	RelMSE	MSE	RelMSE
Heckman	9.696	6.32	0.759	0.379
2-step GB	0.794	0.439	0.730	0.365

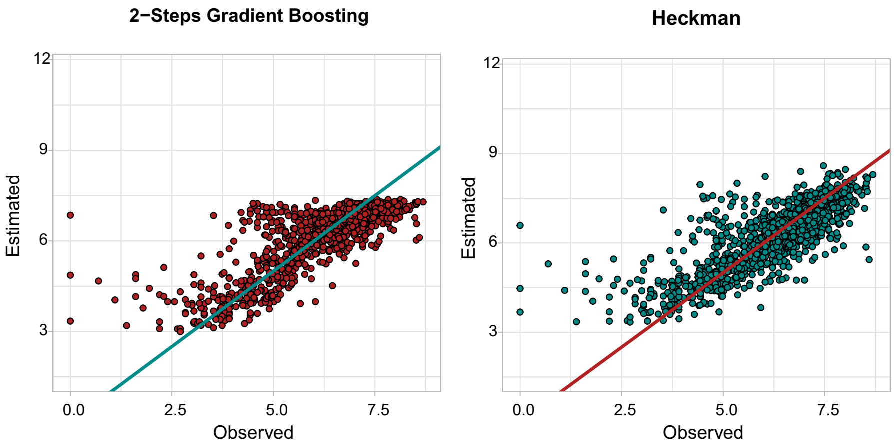

Fig. 5 Comparison of observed and predicted values for the 2-step GB (left) and the Heckman model (right) on the audited units test set

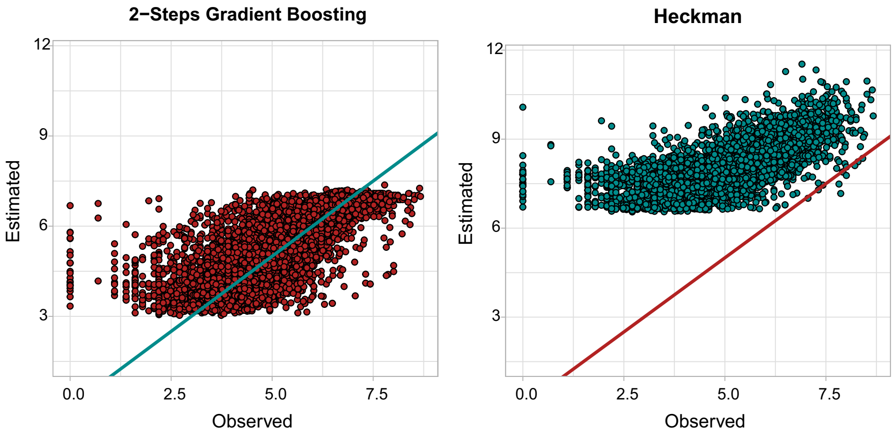

Fig. 6 Comparison of observed and predicted values for the 2-step GB (left) and the Heckman model (right) on the not-audited units

Table 11 Comparison between some linear coefficients estimated by the outcome part of the Heckman model and the linear model fitted on all the observations (audited and not audited)

Coefficient	Heckman model	Complete linear model
Intercept	6.825	3.361
Campinas	– 0.401	– 0.27
Porto Alegre	– 0.53	– 0.234
N. of Bathrooms	0.027	0.21
Floor	0.015	0.005
Fire inscurance	0.029	– 0.017

Table 12 Intervals evaluation metrics in the not-audited and audited groups

Metric	Set	
	Not-Audited	Audited
$Cov_{0.95}$	0.942	0.978
$\bar{W}_{0.95}$	4.1	3.9

indirect correlation between the outcome of interest and the selection mechanism, which is driven by (unknown) criteria depending on the available covariates.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s10260-022-00643-4.

References

Alho JM (1990) Adjusting for nonresponse bias using logistic regression. Biometrika 77(3):617–624

Amemiya T (1986) Advanced econometrics (1985 ed.)

Austin PC, Stuart EA (2015) Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 34(28):3661–3679

Bekker J, Robberechts P, Davis J (2019) Beyond the selected completely at random assumption for learning from positive and unlabeled data. In: joint European conference on machine learning and knowledge discovery in databases, Springer, pp 71–85

Bethlehem J (2010) Selection bias in web surveys. Int Stat Rev 78(2):161–188

Bethlehem JG (1988) Reduction of nonresponse bias through regression estimation. J Off Stat 4(3):251

Bordignon M, Zanardi A (1997) Tax evasion in italy. Giornale degli economisti e annali di economia pp 169–210

Braiotta A, Carfora A, Pansini RV, Pisani S (2015) Tax gap and redistributive aspects across italy. Ital Revenue Agency Discuss Topics 2:1–27

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees. Routledge, Milton Park

Chawla NV (2009) Data mining for imbalanced datasets: an overview. In: Data mining and knowledge discovery handbook, Springer, pp 875–886

Cortes C, Mohri M, Riley M, Rostamizadeh A (2008) Sample selection bias correction theory. In: international conference on algorithmic learning theory, Springer, pp 38–53

Cortes C, Mansour Y, Mohri M (2010) Learning bounds for importance weighting. Nips Citeseer 10:442–450

Coulston JW, Blinn CE, Thomas VA, Wynne RH (2016) Approximating prediction uncertainty for random forest regression models. Photogramm Eng Remote Sens 82(3):189–197

Dangerfield BJ, Morris JS (1992) Top-down or bottom-up: aggregate versus disaggregate extrapolations. Int J Forecast 8(2):233–241

Deng Y, Yan Y (2019) Propensity score weighting with generalized boosted models to explore the effects of the built environment and residential self-selection on travel behavior. Transp Res Rec 2673(4):373–383

Devroye L, Györfi L, Lugosi G (2013) A probabilistic theory of pattern recognition, vol 31. Springer Science & Business Media, Berlin

Du Plessis MC, Niu G, Sugiyama M (2014) Analysis of learning from positive and unlabeled data. Adv Neural Inf Process Syst 27:703–711

DuGoff EH, Schuler M, Stuart EA (2014) Generalizing observational study results: applying propensity score methods to complex surveys. Health Serv Res 49(1):284–303

D'Agosto E, Marigliani M, Pisani S (2016) A general framework for measuring vat compliance in italy. Argomenti di discussione dell'Agenzia delle Entrate 2016(2)

Efron B, Hastie T (2016) Computer age statistical inference, vol 5. Cambridge University Press, Cambridge

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & hall, London, p 436

Efron B et al (2003) Second thoughts on the bootstrap. Stat Sci 18(2):135–140

Elkan C (2001) The foundations of cost-sensitive learning. International joint conference on artificial intelligence, Lawrence Erlbaum Associates Ltd 17:973–978

Fawcett T (2006) An introduction to roc analysis. Pattern Recognit Lett 27(8):861–874

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets, vol 10. Springer, Berlin

FISCALIS Tax Gap Project Group (2018) The concept of tax gaps - Corporate Income Tax Gap Estimation Methodologies. Taxation Papers 73, Directorate General Taxation and Customs Union, European Commission, https://ideas.repec.org/p/tax/taxpap/0073.html

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat, pp 1189–1232

Hall P (1986) On the number of bootstrap simulations required to construct a confidence interval. Ann Stat, pp 1453–1462

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, vol 1. Springer series in statistics

Hastie TJ, Tibshirani RJ (2017) Generalized additive models. Routledge, Milton Park

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284

Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In: Annals of Economic and Social Measurement, Volume 5, number 4, NBER, pp 475–492

Heckman JJ (1979) Sample selection bias as a specification error. Econom J Econom Soc, pp 153–161

Heskes T (1997) Practical confidence and prediction intervals. In: Advances in neural information processing systems, pp 176–182

Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4):1161–1189

Imes AJT (2013) An examination of the sales and use tax gap based on minnesota audit experience. Tech. rep

Kriegler B, Berk R (2010) Small area estimation of the homeless in los angeles: an application of cost-sensitive stochastic gradient boosting. Ann Appl Stat, pp 1234–1255

Kumar S, Srivistava AN (2012) Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. NASA Technical Reports Server

Kumar S, Rao RK et al (2015) Minimising selection failure and measuring tax gap: an empirical model. National Institute of Public Finance and Policy, New Delhi, Tech. rep

Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Stat Med 29(3):337–346

Lee PH (2014) Resampling methods improve the predictive power of modeling in class-imbalanced datasets. Int J Environ Res Public Health 11(9):9776–9789

Li Q, Racine JS (2007) Nonparametric econometrics: theory and practice. Princeton University Press, Princeton

Liu J, Wu C, Li Y (2019) Improving financial distress prediction using financial network-based information and ga-based gradient boosting method. Comput Econ 53(2):851–872

Lohr SL (2019) Sampling: design and analysis. Chapman and Hall/CRC, London

Marra G, Radice R (2013) Estimation of a regression spline sample selection model. Comput Stat Data Anal 61:158–173

MEF (2014) Rapporto sulla realizzazione delle strategie di contrasto all'evasione fiscale, sui risultati conseguiti nel 2013 e nell'anno in corso, nonché su quelli attesi, con riferimento sia al recupero di gettito derivante da accertamento all'evasione che a quello attribuibile alla maggiore propensione all'adempimento da parte dei contribuenti. Tech. rep., https://www.mef.gov.it/documenti-allegati/2014/Rapporto_art6_dl66_13_luglio.pdf

Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards TC Jr (2006) Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol Modell 199(2):176–187

More A (2016) Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048

Newey WK (2009) Two-step series estimation of sample selection models. Econom J 12:S217–S229

OECD (2016) Italy's tax administration. Tech. rep., https://www.mef.gov.it/inevidenza/documenti/Rapporto_OCSE_Eng.pdf

OECD (2017) Tax Administration 2017: comparative information on OECD and other advanced and emerging economies. OECD Publishing, https://www.oecd-ilibrary.org/taxation/tax-administration-2017_tax_admin-2017-en

OECD (2018) OECD Science, Technology and Innovation Outlook 2018. OECD, https://doi.org/10.1787/sti_in_outlook-2018-en, https://www.oecd-ilibrary.org/content/publication/sti_in_outlook-2018-en

Pisani S (2014) Tax gap and the performance of italian revenue agency. an ongoing project. Discussion Topics, Italian Revenue Agency (1)

Pisani S, Pansini Rv (2017) Bottom-up estimates of tax gap by the italian revenue agency. http://docplayer.net/139929090-Bottom-up-estimates-of-tax-gap-by-the-italian-revenue-agency.html

Puhani P (2000) The Heckman correction for sample selection and its critique. J Econ Surv 14(1):53–68

Ridgeway G (2007) Generalized boosted models: a guide to the gbm package. Update 1(1):2007

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33–38

Rubens J (2020) Brazilian Houses To Rent [Version 2]. https://www.kaggle.com/rubenssjr/brasilian-houses-to-rent#houses_to_rent_v2.csv, data retrieved from Kaggle

Santoro A (2010) L'evasione fiscale: quanto, come e perché. Il mulino

Särndal CE, Lundström S (2005) Estimation in surveys with nonresponse. John Wiley & Sons, Hoboken

Särndal CE, Swensson B, Wretman J (2003) Model assisted survey sampling. Springer Science & Business Media, Berlin

Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227

Shrestha DL, Solomatine DP (2006) Machine learning approaches for estimation of prediction interval for the model output. Neural Netw 19(2):225–235

Stine RA (1985) Bootstrap prediction intervals for regression. J Am Stat Assoc 80(392):1026–1031

Stolzenberg RM, Relles DA (1997) Tools for intuition about sample selection bias and its correction. Am Sociol Rev, pp 494–507

Teramoto R (2009) Balanced gradient boosting from imbalanced data for clinical outcome prediction. Statistical applications in genetics and molecular biology 8(1)

Tobin J (1952) A survey of the theory of rationing. Econom J Econom Soc, pp 521–553

Toder E et al (2007) What is the tax gap? Tax Notes 117(4):367–378

Werding M (2005) Survivor benefits and the gender tax gap in public pension schemes: observations from germany. CESifo Working Paper Series (1569)

Wickham H (2016) Package 'tidyr'

Wickham H, Francois R, Henry L, Müller K et al (2015) dplyr: a grammar of data manipulation. R package version 04:3

Wojtyś M, Marra G (2015) Copula based generalized additive models with non-random sample selection. arXiv preprint arXiv:1508.04070

Wojtyś M, Marra G, Radice R (2016) Copula regression spline sample selection models: the r package semiparsamplesel. J Stat Softw 71(1):1–66

Yang JC, Chuang HC, Kuan CM (2020) Double machine learning with gradient boosting and its application to the big n audit quality effect. J Econom

Yin C, Cao J, Sun B (2020) Examining non-linear associations between population density and waist-hip ratio: an application of gradient boosting decision trees. Cities 107:102899

Zadrozny B (2004) Learning and evaluating classifiers under sample selection bias. In: proceedings of the twenty-first international conference on Machine learning, ACM, p 114

Zadrozny B, Langford J, Abe N (2003) Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE international conference on data mining, IEEE, pp 435–442

Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, et al. (2019) Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med, 7(7)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pierfrancesco Alaimo Di Loro[1,2] · Daria Scacciatelli[3] · Giovanna Tagliaferri[1,4]

✉ Pierfrancesco Alaimo Di Loro
p.alaimodiloro@lumsa.it; pierfrancesco.alaimodiloro@gmail.com

Daria Scacciatelli
dscacciatelli@sogei.it

Giovanna Tagliaferri
ggtagliaferri@sogei.it; giovanna.tagliaferri19@gmail.com

[1] Department of Statistics, La Sapienza, Piazzale Aldo Moro, 5, Rome 00185, Italy

[2] Present Address: Department of Law, Economics, Politics and Modern Languages (GEPLI), LUMSA, Borgo S. Angelo, 13, Rome 00193, Italy

[3] Present Address: Forecasting models and statistical analysis, SOGEI, Via M. Carucci, 99, Rome 00143, Italy

[4] Present Address: Business Intelligence and Data Warehouse, SOGEI, Via M. Carucci, 99, Rome 00193, Italy