Predicting the Deterrence Effect of Tax Audits. A Machine Learning Approach *

Michele Rabasco†, Pietro Battiston‡

December 26, 2022

Abstract

We apply machine learning methods to the prediction of deterrence effects of tax audits. Based on tax declarations data, we predict the increase in future income declarations after being targeted by an audit. We find that flexible models, such as classification trees and ensemble methods based on them, outperform penalized linear models such as Lasso and ridge regression in predicting taxpayers more likely to increase their declarations after an audit. We show that despite the non-randomness of audits, their specific time structure and the distribution of changes in declared amounts suggest a causal interpretation of our results; that is, our approach detects a heterogeneity in the reaction to a tax audit, rather than just forecasting an unconditional future increase. We find that taxpayers identified by our model will on average increase their declared income by €14 461 – the average among all audited taxpayers being €-205. Our approach allows

*This research is part of a research agreement between DEMS, University of Milan-Bicocca, and the Italian Revenue Agency, that we warmly thank for providing the data as well as further assistance. The views and opinions expressed are those of the authors and do not necessarily reflect the views or positions of the Agency. This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy. We thank Alessandro Santoro for precious discussions, and two anonymous referees for their insightful comments.

†University of Bologna, University of Milan-Bicocca

‡University of Pisa, University of Parma. Email: me@pietrobattiston.it

the tax agency to yield significantly larger revenues by appropriately targeting tax audit.
Keywords: deterrence effect, machine learning, tax evasion, prediction
JEL classification: H26, D90, C53, C55.

1 Introduction

The literature on tax compliance identifies different effects of tax audits on audited taxpayers (Plumley, 1997; Dubin, 2007). Beyond the immediate effect represented by recovered unpaid tax as the result of a tax audit process, audits can result in a change in *subsequent* tax behavior by taxpayers (Andreoni et al., 1998; Alm et al., 2009; Ratto et al., 2013; Mazzolini et al., 2022). The focus of the present study is the *heterogeneity* of the deterrence effect of tax audits on audited taxpayers.

There are at least two reasons why audits might carry a deterrence effect. The first is known as "target effect": audited taxpayers can be led to believe that they have become a target of enforcement initiatives. Hence, they are likely to increase their perceived probability of being audited again if they interpret an audit as a sign of specific attention devoted to them by the tax authority (Hashimzade et al., 2013). [1] This was also reproduced in a laboratory experiment by Kasper and Rablen (2022), in which audits targeting taxpayers with the lowest levels of compliance result in them raising their compliance.

The second reason is related to the ability of the tax authority to successfully uncover tax evasion: an audit may lead a taxpayer to re–evaluate this ability, both upwards and downwards. Since fines are roughly proportional to discovered evasion, there is an obvious negative relationship between the perceived ability of the tax authority and the propensity of the taxpayer to conceal taxable earnings. For example, experiments have shown that to be effective in increasing a subsequent tax declaration, audits need to be able to detect all undeclared income (Kasper and Alm, 2022) — but not more, as shown by Lancee et al. (2022) in the case of *unfair* audits.

The literature on tax compliance posits that reported income is positively related both to perceived probability of being audited (Hashimzade et al.,

[1] If audits are actually known to be random, and subject to budget constraints, the opposite phenomenon of "bomb crater effect" (Mittone, 2006) could actually be expected.

2013) and to the effectiveness and severity of the punishment (Allingham and Sandmo, 1972; Yitzhaki, 1974): hence, if either of these factors increases as a result of an audit — and the literature presented above suggests this is the case — we can expect audited taxpayers to increase their reported income as a result.

Whether indirect effects are relevant to tax compliance has been the subject of different studies. Kleven et al. (2011), DeBacker et al. (2015), and Advani et al. (2018) find a positive effect of audits on total subsequent income declaration, driven mostly by taxpayers not subject to third-party information. Beer et al. (2020) and Gemmell and Ratto (2012) show that the effect of an audit on tax declarations in subsequent years markedly depends on whether taxpayers experienced an additional tax assessment as a result of the audit. Hence, tax audits might, in principle, have a *negative* effect on yields, to the extent that many of the selected taxpayers are found to be compliant.

Our contribution to the literature consists in *identifying* the individual taxpayers who are more likely to increase the income declared in their next tax declaration, conditional to be audited. This is an important prediction problem because tax auditing is costly, and hence poses a resource allocation problem. While not being the only goal of a tax authority, the possibility to increase future tax revenues is certainly an important element in determining which taxpayers will be audited. Best et al. (2021), who study a similar problem, approach the heterogeneity of changes in future declared income by focusing on one variable at a time and studying how the audit effect changes along its range. Compared to them, we partly sacrifice the interpretability of our analysis in exchange for a more flexible approach that involves the interaction of a large number of variables.

To conduct our analysis we employ an administrative dataset provided by the Italian Revenue Agency:[2] tax audits in the data set are hence not random. This on the one hand adds value to the analysis because its setup does not depart from the typical approach of tax authorities to audits; on the other hand, absent the randomness of treatment assignment, care must be taken as to the causal interpretation of our results – specifically, the causal link between an audit and a subsequent increase in income declaration. In what follows, we first focus on the prediction of which taxpayers will increase

[2]The dataset, anonymized and confidential, is made available by the agency for research purposes.

their declarations following an audit. We later show, based on the time structure of tax audits and on the distribution of income, why this evidence can be interpreted causally.

For the prediction, we employ Machine Learning (ML) algorithms on the aforementioned administrative panel data set of tax-returns and audits, containing information about a large number of Italian taxpayers. ML algorithms are designed to predict a target variable (supervised) or finalized to pattern recognition (unsupervised). In this paper, we employ an approach of the first kind. The family of ML supervised algorithms includes methods that are very different among them, including regularized regression (Lasso and ridge), regression and classification trees, random forest, support vector machines, neural networks. [3] Although developed in computer science and statistical literature, these methodologies are receiving increasing interest from economic scholars, as more and more economic applications are identified (Varian, 2014). In her review, Athey (2018) shows some of the most promising applications of ML in economics. One such application consists in identifying ex-ante, among a basin of potential beneficiaries of a policy, those who would likely behave in such a way as to ensure the higher effectiveness of the intervention. Andini et al. (2017), for instance, exploit ML algorithms to select the target of a tax bonus intended to stimulate consumption. Similarly, Andini et al. (2022) design a policy-assignment rule driven by ML to increase the effectiveness of public guarantee programs. Both contributions represent an example of how ML can be used to better allocate resources, by targeting beneficiaries who respond more to a policy as compared to the current beneficiaries.

In this paper, we focus on the prediction of future tax behavior of audited taxpayers, comparing the results of different supervised classification methods. These algorithms are first *trained* on an already classified dataset, and then used to classify new data. Specifically, we employ classification trees and two ensemble methods based on of them: random forest and AdaBoost (see section 3). As is customary in the ML literature, we test the prediction ability of such models in out-of-sample predictions. For reference, we also compare the performance of these models with that of penalized linear methods Lasso (Least Absolute Shrinkage and Selection Operator) and ridge regression, which, while designed for applications with many variables, share the limited flexibility of widely used linear models such as OLS. These are

[3] See Varian (2014) and Mullainathan and Spiess (2017) for more details.

methods intended to cope with applications where a large number of features are available; they are still very different from the other ML methods mentioned, as they do not identify and exploit nonlinearities and interactions – which is the key feature of decision trees, and ML methods in general.

For the problem under analysis, we observe the best prediction ability with random forest algorithms. We use such best performing model to simulate a prediction exercise on behalf of the tax administration and quantify the benefits in terms of revenues. Specifically, we show that by selectively targeting taxpayers that are predicted to positively respond to an audit, the collected income for the subsequent tax year increases by around €14 461, on average, per audited taxpayer.

The outline of the paper is as follows. In the next section, we describe our data. In section 3, we introduce our decision tree model and the methodologies adopted to measure its prediction ability, tune its hyperparameters and test ensemble approaches. Then we compare the results of our decision trees based models with results from Lasso and ridge. In section 4, a cost-benefit analysis is carried out. The final section concludes.

2 Data

We use an administrative data set released by the Italian tax authority which contains tax reports of Italian taxpayers. The data cover the entire population of self-employed and sole proprietorships from three large Italian regions: Lombardy, Lazio and Sicily, located in the North, Center and South of the country, respectively.

The data set is a balanced panel including data from 2007 to 2011 for 662,241 taxpayers each year, resulting in 3,311,205 observations (although, as discussed below, our analysis only exploits a subset of this data). In fact, the sample is defined as including taxpayers having filed tax declarations in the given regions of operation and types of employment in each of the five years considered. The population analyzed is of particular interest: since these taxpayers are typically not subject to third-party reporting, they have a high opportunity to report a level of income lower than the true one.

Available data comprise 146 variables, including information on taxpayers' demographic characteristics (gender, age, province of residence etc.), information on taxpayers' economic activity (sector, city where they carry out their activity, number of dependent workers, years of operation...), tax-

related variables (gross and net income, paid taxes, value added tax declarations), information on compliance to *Studi di Settore* (SDS), an auditing scheme applying to Italian enterprises below a given threshold of turnover.[4] Finally, the data include audit-related variables (whether a taxpayer was subject to an audit, its date, the fiscal year it refers to...). All categorical variables are transformed into dummies corresponding to each value, so that the number of variables actually employed in the analysis increases to 1496.

In Italy, tax audits can only be carried within five years from the year to which a given tax declaration refers: Table 1 shows that audits carried out in a given year are largely concentrated on tax returns referring to the last two tax periods before the legal expiration. Since the last fiscal year we have data for is 2011, our analysis of audited taxpayers' behavior can only be run on taxpayers audited before this year, therefore leading to a reduction in our sample.[5] Audits can take different forms: most are conducted via mail, but some may include on–site visits: our data does not include information on which. Not included in our data are also systematic cross–check of different sources of information (e.g. on expenditures, or ownership of luxury goods) that the agency performs and that can result in audits being triggered, but of which individual taxpayers are not otherwise made aware.

Table 1: Distribution of audits across years

Year of audit Tax year	2007	2008	2009	2010	2011	2012
2007	0	104	764	2016	4761	10127
2008	1	0	54	669	3463	6547
2009	0	0	0	30	387	1686
2010	0	0	0	0	52	554
2011	0	0	0	0	0	41
All	1	104	818	2715	8663	18955

[4]We refer to (Santoro and Fiorio, 2011) for a more detailed description of the *Studi di Settore*.

[5]The data set includes audits carried out during 2012, but not declarations made in that year.

In order to properly conduct our analysis, it is of critical importance to consider the Italian tax reporting schedule, as well as some norms regulating auditing activities. In particular, since we want to analyze the effect of an audit, we are interested in the moment in which a taxpayer is informed of it. The Italian tax schedule envisages that taxpayers can submit their declaration (concerning the previous year's income) until the end of September. Hence, in order for it to affect the declaration made in year t (and referring to the income from year $t-1$), the taxpayer must be made aware of an audit before September 30 of year t. Since our data set provides us with the exact date on which a taxpayer was informed of an audit, we define the variable "audited in year t" as indicating the presence of an audit communicated between October 1 of year $t-1$ and September 30 of year t.[6] Once we account for this timing, the sample of taxpayers audited "in 2010" – that is, those for which the effects of an audit are expected to influence 2011 declarations – is composed of 2935 observations, rather than the 2715 shown in Table 1).

2.1 Descriptive Statistics

The aim of our prediction exercise is to identify taxpayers who increase their reported income after an audit. Specifically, we focus on net income and look at its change from one year to the other, normalized by taking its ratio over the net income in the first year.[7] To the extent that yearly changes in income are expected to be roughly proportional to income itself, larger increases *in absolute value* would be presumably found among taxpayers who declare larger incomes. We focus on the relative change precisely to disregard this scale effect, which is already obvious to the tax administration.

As highlighted in Table 1, the number of recorded audits per year varies largely, due to the typical delay between a tax declaration and a tax audit referring to it. In order for our exercise to be feasible, we need to observe declared income in two consecutive years *and* to know which taxpayers were

[6] We cannot exclude that some taxpayers make delayed tax declarations – which result in mild fines: luckily, we observe that months from October to December account for a very little number of audits; hence, our results are unlikely to be affected by such late declarations. Vice–versa, taxpayers can submit their declaration well in advance, and we do not observe the exact date in which a declaration was filed, but they have time until the deadline to freely make amendments.

[7] In the rare case in which a *negative* net income was reported in the first year, we normalize by dividing by its absolute value.

subject to audits in the first of the two years. Since the last year for which we observe tax declarations is 2011, the last useful audit year is 2010. Furthermore, given that audits run in 2010 encompass 75% of all recorded audits, we decide to restrict to such year, sacrificing part of the sample in exchange for a more homogeneous set of data and a more directly implementable approach (as the tax agency typically plans each year which taxpayers to target with an audit). Hence, the 2935 observations described at the end of the previous session form our main sample for training prediction models. Later, we will run a cost–benefit analysis also involving declarations for year 2010 of non–audited taxpayers. Because of some missing variables used for the analysis, this will involve 589 633 observations (down from the 662 241 in the data set).

Figure 1 compares audited and non-audited taxpayers in terms of percentage increase in reported net income. It shows that audited taxpayers are relatively abundant in the tails of the distribution – in particular in the right one – and underrepresented in the center. The presence of audited taxpayers in the left tail is a clear sign that audits do not *systematically* result in an increase of declared income – on the contrary – and can be explained by considering that audits are not randomly assigned. Specifically, taxpayers who are underperforming – and hence declare a lower and lower income over the years – might have a higher probability to become the subject of an audit: if such audit does not succeed in reversing the trend (for instance because no evasion was involved), they will remain on the left tail of the distribution. We observe that, while the presence of audited taxpayers in the left tail can also be related to the period of economic crisis which we study, the presence of audited taxpayers in the right tail is unlikely to be. We devote more attention to the specificity of year 2010 at the end of Section 4.2.

That audited taxpayers are more overrepresented in the right tail, however, is encouraging evidence. First, it suggests that there is scope for audits to increase future revenues as compared to the *status quo*. Second, it suggests that such an endeavor will be more successful the more we are able to discriminate in advance which taxpayers will lie in the left or in the right tail of the distribution of income percentage increase. The peak at -100 refers to taxpayers reporting a positive income in 2010 but zero income in 2011. The smaller peak at 100 refers to taxpayers reporting a *negative* income in 2010 and zero income in 2011.

Given the aim of the paper of employing flexible algorithms to identify taxpayers on the right tail of Figure 1, as a first step we want to ascertain whether these tails are not directly related by some simple demographic char-

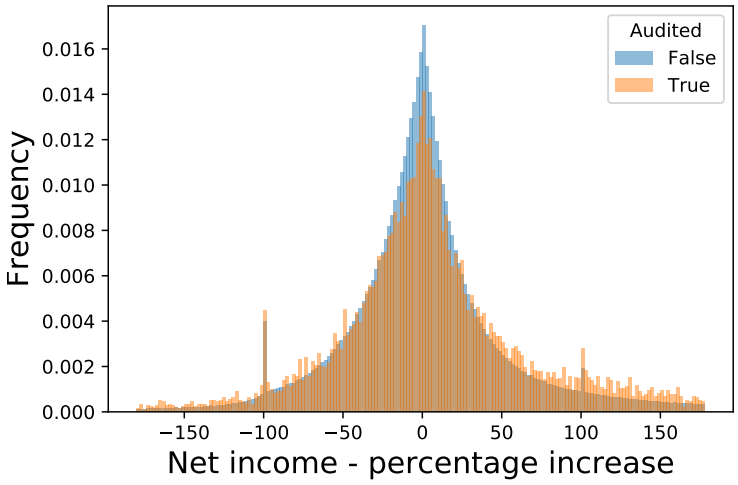

Figure 1: Distribution of changes in reported net income between 2010 and 2011

acteristic. If, for instance, taxpayers from a given region were systematically more prone to increasing their declared income after an audit, increasing overall declared income would be as simple as concentrating more audits in that region (possibly reducing the intensity of audits in other regions). Hence, despite regional or sectorial differences not being the focus of our analysis, we provide some evidence on the distribution of changes in reported net income between 2010 and 2011, conditional on being subject to an audit during 2010, and disaggregated on some characteristics of taxpayers. Namely, taxpayers are grouped by region (Lazio, Lombardy, and Sicily) in Table 2 and Figure 2, by sector (agriculture, trade, industry, private services, and public services) in Table 3 and Figure 3, and by and class of revenues in Table 4 and Figure 4. Since the mean can be disproportionately influenced by outliers, we focus on quartiles of the distribution of changes in reported net income, which represent the top, center and bottom of the blue rectangles in the boxplots. Overall, the resulting evidence is in line with the analysis of Figure 1. Indeed, the distribution of declared incomes by audited activities is more spread than that of non-audited activities for all categories with the exception of small activities (revenues below €10 000 per year), for which no visible difference is present for the quartiles. Hence, the fact that audited taxpayers

are relatively overrepresented in the tails of the distribution of year-to-year percentage changes in declared incomes is not driven by any specific subset of the population, but rather is a generalized phenomenon. In fact, it is again consistent with the distribution of audited taxpayers being determined by their selection: those who are audited because they appear to be declaring lower income than expected might either be irremediably underperforming, or vice–versa concealing part of the income, waiting to be brought to light.

Table 2: Changes in reported net income between 2010 and 2011 (%) - Descriptive statistics (Regions)

Region	Audited	N	25%	50%	75%
Lazio	False	152372	-26.54	0.05	28.99
	True	874	-27.83	0.18	37.14
Lombardy	False	304755	-20.08	0.99	24.88
	True	853	-22.62	4.94	43.62
Sicily	False	129571	-30.41	-0.85	31.36
	True	1208	-34.65	-1.39	39.58

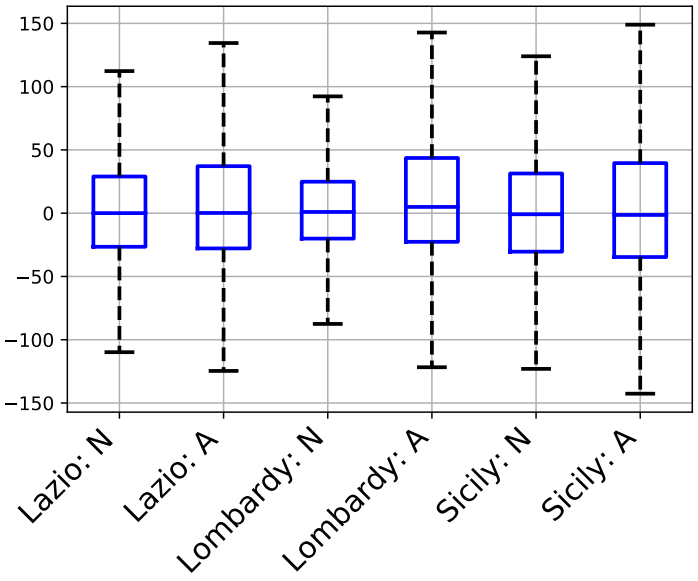

Figure 2: Distribution of percentage change in reported net income between 2010 and 2011 (Regions) / (A)udited - (N)on audited

Table 3: Changes in reported net income between 2010 and 2011 (%) - Descriptive statistics (Sectors)

Region	Audited	N	25%	50%	75%
Agriculture	False	5964	-56.84	1.63	83.87
	True	37	-145.84	-45.99	71.01
Trade	False	174372	-24.17	-0.42	23.05
	True	1281	-32.11	-1.23	34.59
Manufacturing	False	115006	-22.95	0.90	27.23
	True	521	-30.37	1.81	49.71
Priv. services	False	265577	-25.04	1.04	30.93
	True	1026	-25.40	4.60	45.64
Pub. services	False	25779	-12.32	-1.30	13.93
	True	70	-12.20	0.66	34.91

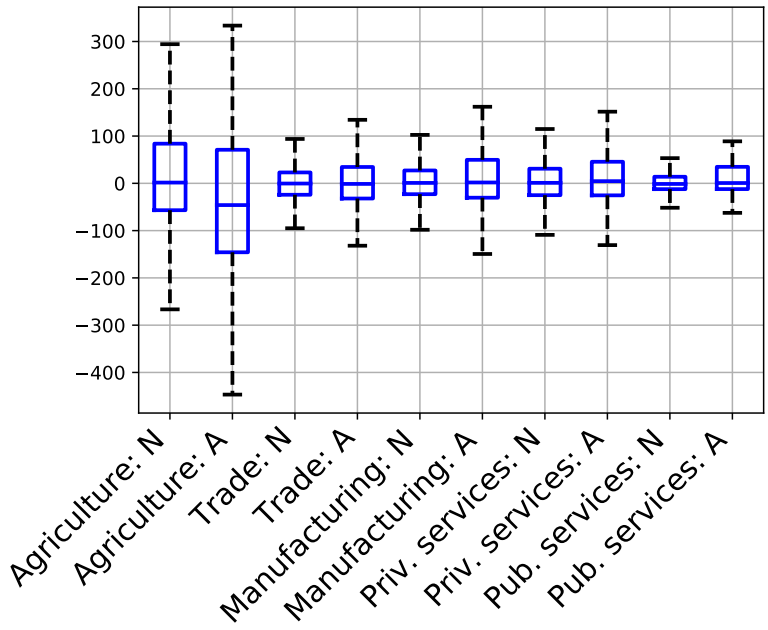

Figure 3: Distribution of percentage change in reported net income between 2010 and 2011 (Sectors) / (A)udited - (N)on audited

Table 4: Changes in reported net income between 2010 and 2011 (%) - Descriptive statistics (Classes of revenues)

Revenues	Audited	N	25%	50%	75%
0-10K	False	85506	-30.02	3.61	66.26
	True	628	-32.05	3.58	63.54
10K-100K	False	357759	-22.83	0.72	25.96
	True	1185	-27.51	4.25	46.49
>100K	False	143433	-23.38	-1.61	19.39
	True	1122	-29.17	-2.02	25.32

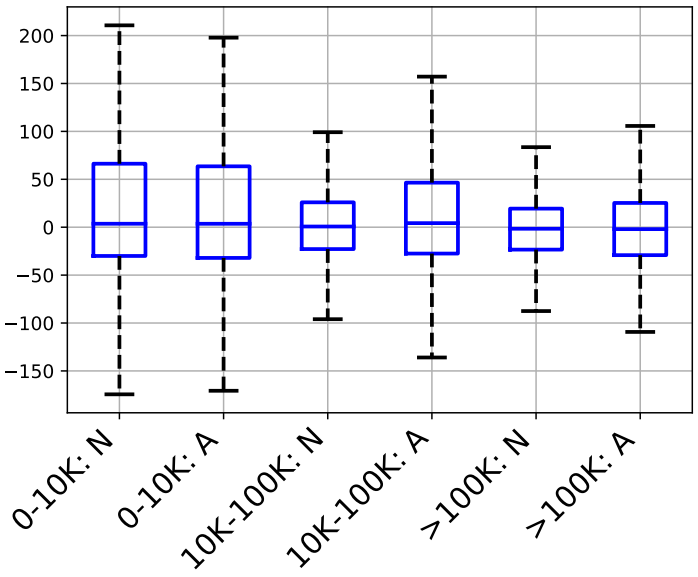

Figure 4: Distribution of percentage change in reported net income between 2010 and 2011 (Class of revenue) / (A)udited - (N)on audited

3 Methods

We classify audited taxpayers as *increasing* or *decreasing* based on their position in the distribution of relative change, i.e., of whether their normalized change in reported income is above or below the median. This results in a dichotomic variable which, by definition, is perfectly balanced (50% of taxpayers are increasing and 50% are decreasing), and which we will consider as target variable for our prediction exercise. The dichotomization of our target variable is dictated by the fact that we are trying to solve a classification problem (Smola and Vishwanathan, 2008), with the ultimate goal to provide a revenue agency a possible approach to the selection ot taxpayers to be audited. The decision to split the sample at the median is relatively arbitrary, but it is motivated by the fact that the percentage change in declared income is larger than 0 for 50.06% of our sample, so it is already very close to being perfectly balanced. In other words, roughly half of taxpayers exhibit an increase in reported income, and roughly half exhibit a decrease (the share of those who feature no change being negligible, at 0.14%). While methods are available to correct unbalanced samples in order to, for instance, obtain unbiased results from decision trees (Krawczyk et al., 2014; García et al., 2009), we consider that setting the threshold at the median is a small price to pay for a simpler implementation and a simpler interpretation of results.

Being an *increasing* taxpayer will be the target variable for our prediction methods. The prediction will be based on characteristics of the taxpayer as observed in the last declaration filed before the audit. Notice that this is typically *not* the audit that the declaration refers to, but rather a later one – since as shown in Table 1, more than one year typically passes between a tax declaration and the enactment of an audit referring to it.

The performance of supervised ML methods is typically evaluated out of sample – that is, by splitting the data set into a training and a test sample. To evaluate our models, we adopt the related method of *cross-validation*: the sample is split into K subsamples of equal size, and the training-test procedure is applied repeatedly by taking each of these subsamples as a test, and the rest of the sample as a training sample. This results in K different predictions, of which we take the average performance as the performance of the model itself. Choosing the value of K is not trivial: increasing it, indeed, reduces the bias but increases the variance of the expected out-of-sample error. We will hence test different values. Having reduced our analysis to a single year (2010) has the added benefit of guaranteeing that in the cross-

validation exercise, the same taxpayer is never observed in both the training and the test subsample.

In order to evaluate the quality of the single prediction – when run on the test sample, i.e., out of sample – we then adopt the widely used F_1 *score*. Our prediction models are classifiers which yield a boolean variable with 1 denoting that an audited taxpayer is expected to be increasing her reported income, and 0 denoting that she is expected to decrease her reported income. The F_1 *score* is defined as the harmonic mean of a prediction's *Precision* – the share, among taxpayers predicted increasing, of taxpayers who end up increasing their reported income – and its *Recall* – the ratio of correctly predicted *increasing* taxpayers over all *increasing* taxpayers:[8]

$$F_1 = 2\,\frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

We will look for models with an F_1 *score* as high as possible. In our setting, this metric reflects the fact that both targeting individuals who will not respond to an audit (i.e., low precision) and missing an individual with a potentially positive response to audits (i.e., low recall) represent a cost to the tax administration. Specifically, auditing taxpayer not inclined to increase their net income declaration as a result of an audit may mean wasting resources.

3.1 Classification trees

We start our prediction exercise by studying the performance of different classification trees. While such models are not considered as top performers in the ML literature, two advantages they have are their ease of interpretation, and the robustness to heterogeneous data, making data cleaning mostly

[8]The most common alternative metric for estimating prediction performance, the ROC AUC, is sensitive to the prediction ability of a method for any given *threshold* chosen: in other terms, it is not dependent on any assumption on the number of positive predictions desired/expected. In our exercise, given both the evidence from Figure 1 and the natural subdivision of taxpayers between increasing and decreasing, we focus on a threshold of 0.5, and hence disregard the others. In other terms, we are not interested in a general evaluation of prediction models, but in using them to identify a specific subsample of taxpayers, which is observed to include roughly half of the sample. Based on this reasoning, we use the F_1 *score* because it includes a balance between type-I errors (which decrease prediction) and type-II errors (which decrease recall).

unnecessary. Classification trees can attain different levels of complexity depending on their number of levels/leaves: while a higher complexity typically leads to a higher in-sample prediction power, this might be at the cost of a decrease in power out-sample – this is the essence of the *overfitting problem*. Moreover, the process of tuning the parameters of such algorithms can also be unstable with respect to the data used: i.e., the specific classification trees we will build will change based on changes to the training set.

In practice, we use the CART (Classification And Regression Trees) procedure introduced by Breiman et al. (1984). This procedure starts by splitting the dataset into two subsamples based on a given variable and a given threshold over it, so as to decrease as much as possible the heterogeneity (impurity) of the outcome variable. As a measure of node impurity, we use the Gini index, defined as

$$GiniIndex = \sum_{h=1}^{H} p_{mh}(1 - p_{mh}), \tag{2}$$

where $h = 1, 2, \ldots, H$ are the values the target variable can take, and p_{mh} is the proportion of class h observations in node m: in our case, $H = 2$, and hence this becomes

$$\begin{aligned} GiniIndex =& \sum_{h=1}^{2} p_{mh}(1 - p_{mh}) \\ =& p_{m1}(1 - p_{m1}) + p_{m2}(1 - p_{m2}) \\ =& 2 \cdot p_{m1}(1 - p_{m1}) \end{aligned}$$

(where the last passage exploits the fact that $p_{m2} = 1 - p_{m1}$).

The best split in our data is hence represented by the combination of a variable and a split point that will result in the smallest weighted average of the Gini index computed in the resulting two subsamples. After the first split, the two obtained regions can be each split into two more regions. This process continues recursively by induction, following a top-down greedy approach.[9] Such procedure continues until a stopping rule is reached, and naturally results in a prediction model, as nodes are identified by regions in the space of

[9]Basically, a greedy procedure applies a problem-solving heuristic to obtain a locally optimal choice at each split.

explanatory variables that tend to be relatively homogeneous in the outcome variable. Hence, as a result, we obtain a tree where the final nodes (leaves) give a classification for the variable we want to predict.

As already mentioned, we are looking for decision trees which are complex enough to capture useful patterns, but not so much as to incur in overfitting. In order to do that, we look for the highest value of the out-of-sample prediction F_1 *score* across trees of different complexity. Specifically, for each tree, we fix the maximum allowed number of final nodes and we repeat the training-testing procedure with such maximum number set to all values between 2 to 100. The result of this procedure is shown in Figure 5. Such figure is based on cross-validation estimation with $K = 100$ – which, based on experimentation, represents a good trade-off between bias and variance of the various subsamples.

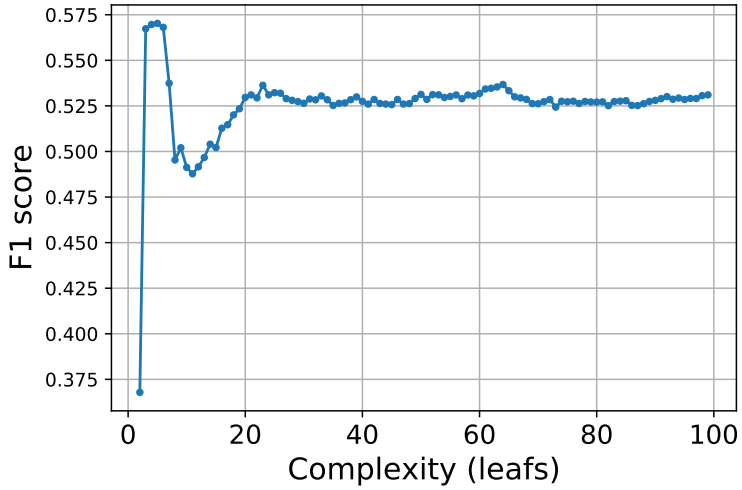

Figure 5: Decision tree: accuracy-complexity trade-off, maximum number of leaves changing from 2 to 100, K =100

The result of this exercise is that the best decision tree has 5 final leaves, corresponding to a F_1 *score* value of 0.57. Given that both our sample and our prediction are roughly balanced, precision and recall are increasing to the same extent: in other terms, the value of 0.57 implies that we are able to identify taxpayers (*increasing* or *non increasing*), 7% of the time more than randomly (that is, when F_1 *score* $= 0.50$).[10]

[10]Indeed, precision and recall are two ratios which have the same numerator and, in our

Two interesting phenomena can be observed from Figure 5. The first is that if we increase the number of allowed leaves beyond the optimal value of 5, we get a sharp decrease in prediction ability, soon reaching the F_1 *score* of a completely random prediction model, which is 0.5. This is the essence of overfitting. The second phenomenon is that as we increase the complexity of the model further, the prediction ability roughly stabilizes. The best performing tree is shown in Figure 6. We observe that of the four branching nodes, the first two are responsible for most of the discriminatory power (as Figure 5 shows that a tree with three leafs is only slightly suboptimal).

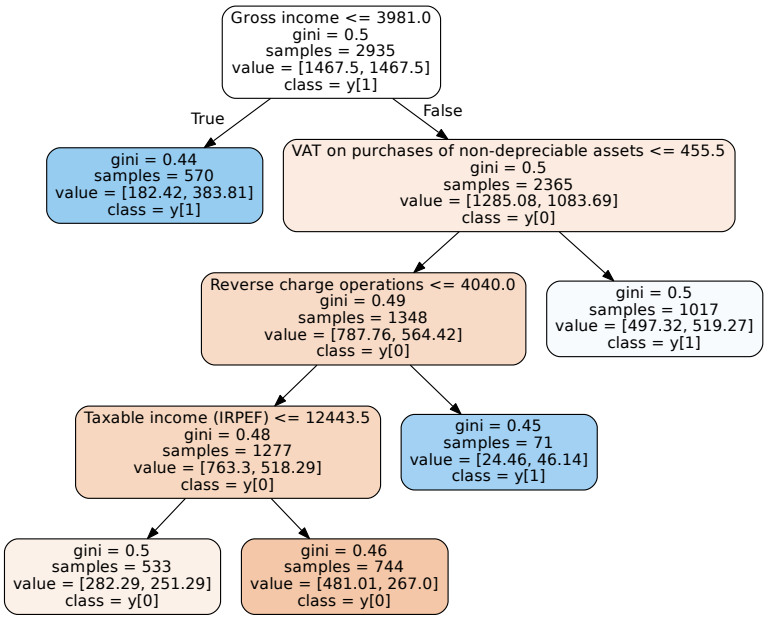

Figure 6: Classification Tree, stopping rule: 5 leaves. The *class* of each node denotes a relative predicted propensity to increase (y[1]) or decrease (y[0]) reported income following an audit. For instance, the first node indicates that having a gross income lower than 3 981 € (condition "True") denotes a relatively high predicted propensity to increase (left node on the second level).

The Gini index of our tree's leaves is an indication of how well they iden-

specific exercise, both have half the sample size as denominator.

tify taxpayers which are homogeneous with regards to their target variable – the closer to 0.5 it is, the less a leaf is successful. Hence, the two leaves that perform best in this respect are the first and fourth from left in Figure 6, colored in blue (with Gini index 0.44 and 0.45), and the third, colored in orange (with Gini index 0.46). The former two both identify taxpayers who are expected to be increasing (class=y[1]); the orange one taxpayers who are expected to be decreasing (class=y[0]). The fact that the optimal model is reached after only a few interactions (involving 1 to 4 variables, depending on the leaf) suggests that our target variable is difficult to predict, despite our large number of variables. Still, this representation of the tree allows us to gain some insights on the factors that relate with a higher propensity to increase reported income as result of an audit: for instance, taxpayers who declare gross income lower than €3 981 or taxpayers who declare more than €3 981 and purchases of non-depreciable assets lower than €455 and operations by application of value added tax (VAT) reverse charge higher than €4 040 have a higher probability to be *increasing* in the year subsequent to an audit (the probability is given by the ratio of positive to total cases, that is $\frac{383.81}{182.42+383.81} = 0.68\%$ and $\frac{46.14}{24.46+46.14} = 0.65\%$, respectively).[11]

In what follows, we improve our analysis by applying two ensemble techniques that are known to often improve the prediction ability of decision trees: random forest and AdaBoost.

3.2 Random Forests

Random forests (Breiman, 2001) work by producing a large number of decision trees, each trained by employing only a subset of input variables as candidates for the splitting procedure, and only a subsample of the initial sample. The prediction is then computed as the mean of each tree's prediction. This approach allows to increase the accuracy of the prediction while at the same time reducing the impact of overfitting – as splitting rules that depend on a small number of observations will lose importance. By building trees with limited correlation among them, the variance of the ensemble estimator is reduced.

[11]The number of subjects in each subsample is not an integer because of the weights applied to the sample. Positive and negative observations are attributed weights such that the total mass of the two categories is the same (1 467.5 in our case – see the first node in Figure 6) and that the sum is equal to the true number of subjects (2 935).

As is common in literature, we set the number of selected features for each tree to $m = \sqrt{p}$ where p is the total number of explanatory variables. While a higher number of trees improves the robustness of a forest, there is a threshold beyond which improvements become negligible – and come at a computational cost (Oshiro et al., 2012). Table 5 displays the result of some experiments with random forests for different choices of the number of trees and the number of cross-validation folds K.

Based on experimentation, we observe that 100 trees represent a good trade-off between bias and variance of the prediction of the various subsamples. This choice allows us to increase the F_1 *score* of our prediction of about 0.02 with respect to the prediction with our best performing classification tree.

Table 5: Random forest: F_1 *score* mean (std) for different numbers of trees and cross-validation splits (K)

F_1 *score* N.of Trees/K	mean	std
50/50	0.56	(+/- 0.14)
50/100	0.58	(+/- 0.20)
100/50	0.60	(+/- 0.12)
100/100	0.58	(+/- 0.20)
150/50	0.60	(+/- 0.14)
150/100	0.58	(+/- 0.19)

Concerning the interpretability of the model, random forests have an obvious disadvantage compared to decision trees, as they are composed by a large number of different trees. This said, it is still possible to gauge the relative importance of the different variables: for a random forest of 100 trees, this is done in Table 6. Indeed, when training a tree, the importance of any given variable can be established by looking at the extent to which it contributes in decreasing the weighted impurity. Averaging such decrease in impurity over trees, we can obtain the importance of a variable for the entire random forest. Table 6 shows the first 20 most important variables sorted by the relative importance with respect to the most important one; as could be expected, the most important variables for our prediction are

features related to gross income ("Gross income" and "Total revenue") and to IRPEF (Italian income tax) taxable base ("Taxable base") for the year before the audit. Immediately after in the ranking, we find variables related to VAT declaration. We can observe that among the first 20 most important variables, the only one not directly linked to declared income is the taxpayer's age.

Table 6: Relative importance of variables for a random forest of 100 trees (first 20 variables)

Feature	Importance
Total revenue (new mode)	1.000
Taxable income (IRPEF)	0.989
Taxable base	0.920
Taxpayer's age	0.901
Gross income	0.842
Total VAT tax base	0.805
Purchases and imports within bracket - base	0.778
Total liabilities	0.776
Purchases and imports within bracket - tax	0.770
Total purchases and import	0.767
Sum of purchases and imports	0.757
Deducted VAT	0.751
Added value	0.746
Other purchases and imports - tax base	0.734
Presumptive turnover from Studi di settore	0.734
Total tax on taxable operations - old version	0.731
VAT Taxed operations from sales to consumers	0.729
VAT amount	0.724
Total revenues	0.716
Total assets	0.704

3.3 Adaptive Boosting (AdaBoost)

Boosting is another ensemble approach allowing to improve the predictive accuracy of ML models. The guiding principle is to sequentially create simple predicting models (*weak* learners), that iteratively improve the fit, and

then combine them: the resulting prediction is obtained by averaging the prediction of the individual learners.

A popular boosting classifier is AdaBoost (Freund and Schapire, 1997). The main idea behind it is to focus on observations that are difficult to correctly classify: in practice, this means, when training further learners, attributing higher weights to observations on which previous learners failed more.

Boosting is a general meta-algorithm – unlike random forest, it can be applied to virtually any prediction method. In our case, we use as base classifier the decision tree. Hence, the algorithm will work by creating many simple decision trees, where later trees are trained on samples where higher weights are attributed to observations which were incorrectly trained by previous trees. Table 7 shows some experiments of the application of AdaBoost, when varying the maximum number of possible trees and the cross-validation K.

Again, we do not find that increasing the number of trees beyond 100 provides substantial improvements in prediction. Results are in general better than with the classification tree, and comparable to the random forest.

Table 7: AdaBoost - F_1 *score* mean (std) for different numbers of trees and cross-validation splits (K)

F_1 *score* N.of Trees/K	mean	std
50/50	0.58	(+/- 0.14)
50/100	0.59	(+/- 0.17)
100/50	0.60	(+/- 0.13)
100/100	0.58	(+/- 0.19)
150/50	0.60	(+/- 0.14)
150/100	0.59	(+/- 0.20)

As with the random forest method, we investigate the average decrease in impurity contribution of each variable (Table 8). Confirming the results from the random forest, Adaboost attributes the largest importance to the gross income and in general, to variables related to costs, revenues and VAT.

Table 8: Relative importance of the variables for Adaboost of 100 steps (first 20 variables)

Feature	Importance
Gross income	1.000
Taxable income (IRPEF)	0.575
Depreciation of tangible capital goods	0.538
Total liabilities	0.505
VAT due	0.494
Initial goods existence	0.467
VAT on purchases of non-depreciable assets	0.457
Duration of the activity	0.434
Total revenues	0.428
Surplus from multi-year works	0.352
Total assets	0.334
Taxpayer's age	0.333
Taxable base	0.302
Trade of clothes, shoes and accessories	0.296
Presumptive turnover from Studi di settore	0.295
Operational costs	0.281
Total IRAP tax	0.278
Services costs	0.245
VAT credit from previous year declaration	0.243
Costs for production of services	0.238

3.4 Penalized linear methods

In order to compare our results with prediction models that are more similar to the linear models typically employed in the economics literature, we now proceed to analyzing the results obtained with two penalized linear models: ridge regression ad Lasso. The aim of these methods is to improve the prediction ability of the linear regression model by decreasing its variance. This is done by penalizing the size of the regressors set (Tibshirani, 1996).[12]

Both models can be expressed as the minimization of the following loss function:

$$L(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \alpha\sum_{j=1}^{p} g(\beta_j). \tag{3}$$

The two models differ only for the specific *penalty* term $g(\beta_j)$. In both cases, g is a strictly positive function, so that the term is minimized for $\beta_j = 0$. The form of this function is $g(\beta) = \beta^2$ in ridge regression and $g(\beta) = |\beta|$ in Lasso. The amount of shrinkage – that is, the penalization applied to positive coefficients – depends on the parameter $\alpha \geq 0$. In particular, when $\alpha = 0$ both models coincide with OLS (the last term in Equation (3) disappears).

Penalized linear models naturally assign a real number – rather than a Boolean value – to each observation, where a higher number corresponds to a higher probability of being increasing. For comparison with tree-based models, we dichotomize such prediction, selecting a threshold such that the resulting Boolean prediction is balanced. In other terms, a taxpayer is expected to be increasing if her score is larger than the median score. It should be observed that Lasso and ridge regressions operate towards a reduction of the coefficients associated with each variable. However, this coefficient will depend on the magnitude of each variable (e.g. also depending on its measurement unit). Due to that, we proceed to a preliminary standardization of variables.

Figures 7 and 8 show, respectively, the F_1 *score* obtained by the ridge and Lasso models for different values of α, tested by cross-validation with $K = 50$.

[12]More specifically, since the ordinary least squares (OLS) estimates often show high variability, we could reduce the variance *shrinking* the coefficients of the regression toward zero or setting them to zero, hence reaching a smaller bias.

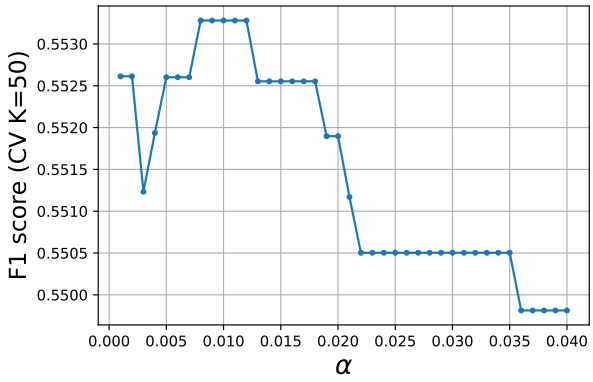

Figure 7: Ridge - F_1 *score* for different values of α and $K = 50$ folds cross-validation

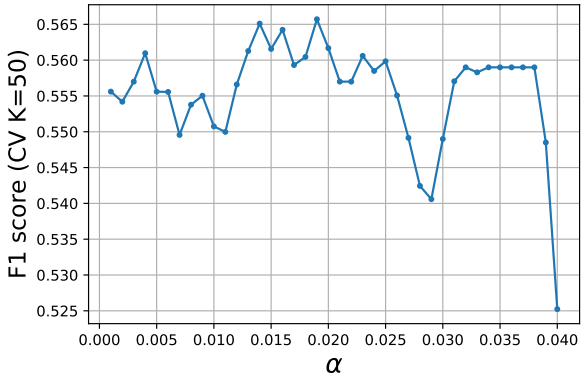

Figure 8: Lasso - F_1 *score* for different values of α and $K = 50$ folds cross-validation

As we can see in the two figures, the maximum value for the F_1 *score* is between 0.55 and 0.56, lower than that found employing random forests and AdaBoost. In correspondence of the optimal value $\alpha = 0.19$, the Lasso only attributes positive coefficients to 20 variables out of 1496. For larger values of α, both models penalize the coefficients more than would be optimal: for instance, the largest value of α considered, 0.035, corresponds to a Lasso model with *only* the "Gross income" variable. Vice-versa, for smaller values of α, we can clearly see the effect of overfitting, as the inclusion of further variables results in a decrease in out-of-sample F_1 *score*.

To complete the discussion about linear methods, we provide the list of variables selected by the Lasso regression (Table 9). The variables with a positive sign of the coefficients positively influence the probability that a taxpayer will increase her net income the year after an audit. The opposite happens for the variables with a negative value. In Lasso regression, gross income is again is the most important feature in explaining the taxpayer's future change in declared income.

Table 9: Lasso - values of non-zero coefficients ($\alpha = 0.19$)

Variable	Coefficient
Gross income	-0.0197
Congruity with Studi di settore	-0.0105
SDS: adaptation (additional revenues)	-0.0102
SDS: full adaptation (additional revenues)	-0.0052
Total VAT due	-0.0048
Taxpayer's age	-0.0034
Trade and automobile mechanics	-0.0027
Wholesale of other covering materials	-0.0020
Wholesale and retail of automobiles	-0.0020
Trade of non-domestic electric appliances	-0.0019
Wholesale of watches and jewelry	-0.0013
Retail of watches and jewelry	-0.0000
Law firms	0.0001
Placement of hydraulic and heating/cooling systems	0.0006
Water supply, sewages, waste disposal	0.0017
Representatives of heterogeneous products	0.0021
Province of Bergamo	0.0023
Veterinary services	0.0048
Gap between declared and presumptive revenues (%)	0.0070
Law firms	0.0118

3.5 Methods comparison

In this section, we provide a comparison of the methods employed, both in terms of predicting performance and factors considered important to detect an *increasing* income reporting behavior after an audit. Table 10 compares the performance of the employed methods. While the differences in performance might seem minor, it should be kept in mind that a completely random prediction mechanism will have F_1 *score* equal to 0.5; so the difference in performance between the best performing methods (the two ensemble methods) and the worst performing methods (the two penalized linear methods) is analogous to the difference between the latter methods and a coin toss. In other words, more flexible methods do show a much better prediction ability in our sample, suggesting that machine learning methods, as compared to for instance linear prediction models, are worth the added complexity and reduced interpretability. Indeed, methods based on classifications trees not only make a selection among variables but also exploit the presence of non-linear relations between them, while ridge and Lasso are unable to capture interactions in the data. This explains the better performance of the former methods.

Table 10: Comparison of the prediction performance among ML methods employed - F_1 *score*

Method	F_1 *score*
Classification tree	0.57
Random forest	0.60
Adaboost	0.60
Lasso	0.56
Ridge	0.55

It is important to emphasize that the main contribution of ML algorithms consists in discovering potentially complex relationships between available variables – relationships which might not have been hypothesized in advance – and not to investigate or confirm a causal model. In other terms, ML algorithms are designed to obtain a good prediction out-of-sample – hence without overfitting – but not to infer causal relationships between variables,

and even algorithms that produce coefficients do not necessarily result in consistent estimates Mullainathan and Spiess (2017). Therefore, the analysis of trained ML models can only provide suggestive evidence as to which individual features affect tax compliance, or response to an audit. So from the point of view of the causal interpretation – which we discuss later – all prediction methods considered are equivalent.

A common denominator of tables 6, 8 and 9 is that gross income and taxpayers' age are relatively important factors to predict whether a taxpayer will increase income declaration after an audit. The importance of reported gross income in detecting which taxpayers are more likely to increase income declaration the year after an audit has a relatively straightforward interpretation: a larger economic activity easily goes hand in hand with a more organized tax reporting, hence more difficult to manipulate and adapt after a tax audit (always in relative terms). An interpretation of the importance of taxpayers' age is less obvious, but it can be argued that this is related to the subjective probability of being audited again Hashimzade et al. (2013). The subjective estimated probability of audits is likely to be more strongly affected for younger taxpayers, who may overestimate the probability to be inspected again – differently from old taxpayers, who may have gathered from their experience a more robust estimate of such probability.

Finally, differently from classification models, Lasso regression attributes explanatory power to belonging to specific sectors, such as "trade of clothes, shoes and accessories", "wholesale of other covering materials", "general trade and automobile mechanics", "trade of non-domestic electric appliances" and "wholesale of watches and jewelry". This result could be associated to the fact that in different sectors it can be more or less easy to conceal income during a tax control; hence taxpayers could have less incentive to increase subsequent income reporting, expecting a limited sanction even in the eventuality of another inspection (Pomeranz, 2015). Finally, and quite surprisingly given the known geographic heterogeneity of tax evasion in Italy (D'Agosto et al., 2014), geographical features do not seem to bear much importance.

In what follows, we quantify the potential benefits of employing such ML methods in the identification of taxpayers to target with audits through a cost-benefit analysis. To this aim, we will employ the taxpayers selection provided by the random forest algorithm, because it is the method reach-

ing the largest value of F_1 *score*[13], hence the best model to predict which taxpayers will increase declarations after an audit among those employed.

4 Discussion

4.1 Causal interpretation

In order to transform the insights presented so far into a normative result for the targeting of tax audits on behalf of a tax authority, the causal interpretation of our results is crucial.

We start this causal analysis by emphasizing that our machine learning analysis does not guarantee, *ex ante*, the identification of any causal effect. That is, it is generally possible that an approach such as the one described above identifies a spurious correlation between a specific taxpayer profile and the increase of declared income after an audit.

So in order to rule out this eventuality *ex post*, we classify possible alternative explanations of our results by regrouping them in three main categories:

1. audited taxpayers *generally* increase their declared incomes, and our prediction model has no merit,

2. our model identifies taxpayers that increase their declared incomes, *regardless of being audited*,

3. our model specifically identifies audited taxpayers that increase their declared income, but the increase is *not caused* by the audit.

These three possible explanations — which are not mutually incompatible — are a sensible way to frame our identification problem because they highlight that its structure is akin to a differences—in—differences—in—differences approach, with the particularity that the group of interest is a *subset* of the treated group (audited taxpayers), and our objective is to identify not a general average treatment effect, but a subset on which it is positive.[14] The first two explanations would imply that there is actually no

[13]Adaboost reaches an F_1 *score* comparable to that of random forest: we prefer the latter because it is associated to a lower variance.

[14]Equivalently, we could define our effect of interest as an Average Treatment Effect on the Treated (Angrist and Pischke, 2008) for a hypothetical audit campaign guided by our prediction model, rather than for the actual campaign ran by the tax agency.

difference when we compare our "treated" sample (audited, selected taxpayers) with the appropriate category — be it audited taxpayers, or taxpayers identified by our model. The third explanation would imply that the *common trends hypothesis* is not satisfied.

Alternative explanation 1. is easy to refute based on observable data: we can observe for instance in Figure 1 that while a significant share of taxpayers audited in 2010 did increase their declared income in 2011, another large part of them *decreased* their declared income, as compared to the wider population of non–audited taxpayers. In fact, we can rule out this explanation entirely by simply computing the average change in declared income from 2010 to 2011 for *all* audited taxpayers: it is €-205. So among audited taxpayers the general trend is one of (slightly) *decreasing* declared income after the tax audit. Notice that this is not necessarily in contrast with theories of perceived probability of sanctions, according to which tax audits should generally increase future compliance (see Section 1). Indeed, since tax audits are not random, they may be relatively likely to target taxpayers whose declared income would be *anyway* decreasing over time (e.g. businesses that are facing generalized difficulties).

Alternative explanation 2. is similarly easy to check against observable data, by analyzing changes in declared income for all subjects selected by our model, be them audited or not. Figure 9 displays, analogously to Figure 1, the distribution of changes in reported income, but this time distinguishing based on whether the individual is predicted to increase declared income in the following year. We can observe that the distribution of subjects predicted to increase does appear more skewed to the right than the distribution of subjects predicted to decrease, but this reflects an average increase in income of only €1 505 (upwards triangles in Figure 10) — as compared to the €14 461 of our effect of interest (stars in Figure 10), that is, of the average increase in declared income for audited taxpayers who are predicted as increasing. Hence, the general propensity to increase declared income by (non–audited) subjects that our model predicts to positively react to an audit can explain at most 11% of the observed effect.

Finally, we turn to the more subtle alternative explanation 3. We start by observing that our models were not specifically trained to discriminate whether a taxpayer is audited — they were trained *only* on audited taxpayers. In essence, this is why, as already mentioned, our empirical approach cannot guarantee, ex ante, the identification of a causal result: we *could* have ended up characterizing a population of taxpayers that simply tend to

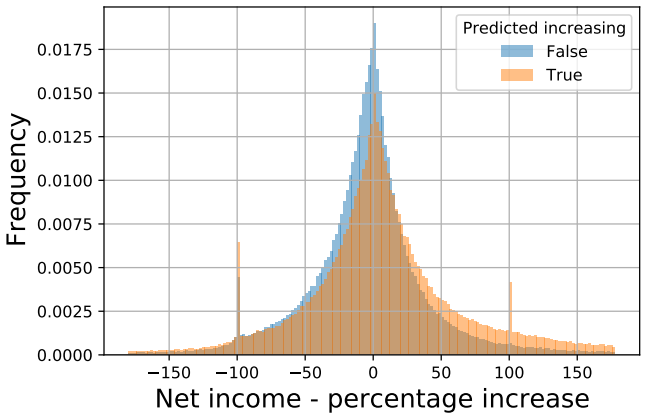

Figure 9: Changes in reported net income from 2010 to 2011, conditional on prediction

increase their income from one year to the other, regardless of audits. However, since this is false (see discussion of alternative explanation 2. above), there is no reason to expect that the identified population has an intrinsically larger likelihood of increasing declared income *and* of attracting audits. That is, having provided our model *no data whatsoever* on what determines selection into audit, we have even more so avoided to systematically aim for the intersection of "selection into audit" and "future increase in declared income". Clearly, the only exception to this line of reasoning is the case in which the *same* features that determine selection into audits also determine the intrinsic propensity to increase future declared income. We know (see discussion of alternative explanation 1 above) that audited taxpayers do not generally increase their income in the following declaration. Still, it could be that *some* taxpayers who are selected for an audit are intrinsically likely to increase their future income, and that our models end up identifying those. The absence of such a coincidence is ultimately our identifying assumption.

> **Independence assumption:** taxpayers who increases future income *regardless* of audits are not systematically more targeted by audits than other taxpayers.

By definition, given that audits are non–random, we cannot prove this assumption based on observables. It is the counterpart of the *conditional independence* assumption that is used by matching approaches such as that

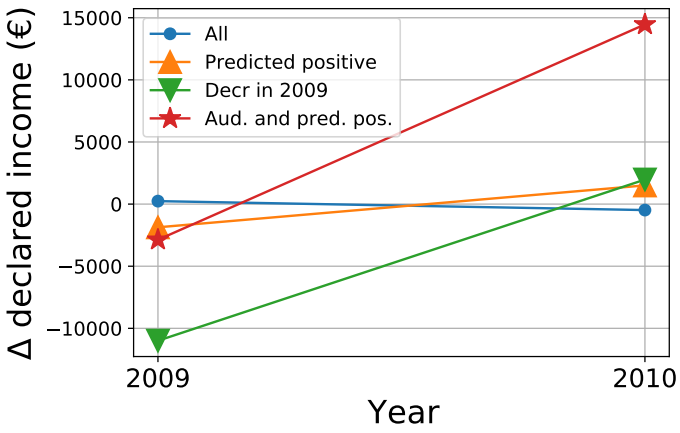

Figure 10: Changes in reported net income in 2010 and in 2011, for different subsets of data

by Mazzolini et al. (2022) (or of the assumption of randomness of audits in the work of Beer et al., 2020). We can, however, analyze the plausibility of a couple of conceivable mechanisms.

First, one could expect the tax agency to actively target taxpayers who are expected to increase their future declared income. We already know this is false on average, but it could hold for a subset of audits. However, there is no point for the agency in auditing taxpayer based on their future increase of declared income, *unless* this increase is caused specifically by the audits — that is, unless the effect we are capturing has a causal interpretation. We note, for the records, that to the best of our knowledge the Italian revenue agency has no program in place to direct tax audits based on the effect on future revenues — a cost–benefit analysis of such a program is sketched in the following section. But still, our assumption does not exclude that the revenue agency might be already focusing (part of) its audits to taxpayers expected to *react positively* to these audits. The assumption only excludes that the agency is doing something far less sensible — systematically targeting taxpayers who for some reason happen to be likely to increase future revenues, but on which the audits have no effect.

Second, one could expect the tax agency to target taxpayers featuring declarations that are suspicious because of lower than usual declared income. There could be a mechanical rebound effect according to which taxpayers who in a given year feature unusually low declared income tend to both attract tax

audits, and follow a reversion to the mean according to which they increase declared income next year. Once more, we know this is not true in general for audited taxpayers, but it could be true for a subset of them. The main reason why such a spurious effect is not plausible is related to the timing of audits. Table 1 shows that only very few (30) of the audits under analysis applied to the year just before the increase (2010), while the vast majority applied to previous years. Hence, for such effect to be explaining our results, it should manifest itself two, or more, years later, which is unlikely. In any case, we can also provide a more empirical argument against this mechanism, by looking in our data for such a rebound effect. Figure 10 compares the change in declared income in 2009 and in 2010 for different subsets of taxpayers. The sample of all taxpayers and the sample of all taxpayers predicted to increase their declared income after being audited (but audited only in minimal part) feature only minimal fluctuations around 0. The sample of taxpayers who in 2009 reported lower income than in 2008 does feature a possible rebound effect, but despite their strong average decrease (due to the definition of the sample) of €-10 998 in 2009, their increase in 2010 (€1 973) is much lower than that observed for our sample of interest. At the same time, our sample of interest features a much lower decrease in declared income from 2008 to 2009: €-2 873. In other terms, assuming that the observed pattern for the sample "decreasing in 2009" indeed represents a rebound effect, it consists — in its most extreme form, given that these are the taxpayers more likely to feature it — in an increase in 2010 that is around 18% of the decrease in the previous year. That is, for our sample of interest it can explain at most $-2\,873 \times 18\% \approx$ €517 — orders of magnitude lower than what we observe.

It is important to highlight that our approach differs from that of Beer et al. (2020) or Mazzolini et al. (2022) first and foremost in terms of aims. That is, we do not claim the ability to identify an average treatment effect (ATE). First, our model is only trained on audited subjects, so if anything, our result is more similar to a local average treatment effect (ATT): to the extent that audits are non–random, our predictions are not informative about non–audited subjects. Second, and most importantly, the sample on which we identify a causal effect is *defined* by our prediction model by splitting in half the sample of treated subjects and implicitly comparing the two sub-samples; it does not allow for a general assessment of audit effects even just on treated subjects. It only implies that, by appropriately *restricting* the sample of audited subjects, it is possible to *increase* the efficacy of audit campaigns in terms of future revenues.

In principle, a prediction model that identifies taxpayers who increase their declared income after an audit is at the same time identifying taxpayers who *decrease* their income after an audit, and whom the revenue agency may want to avoid targeting.[15] The so–called *bomb crater effect* (Mittone, 2006; Mittone et al., 2017) is a possible theoretical explanation (also observed in laboratory experiments) for why some taxpayers could *reduce* their declared income after an audit. While such an effect cannot be excluded entirely, it is of limited practical importance, to the extent to which the revenue agency can, in a perspective of future revenue maximization, avoid directing audits at them. It is also worth noting that a single tax audit can have potentially serious consequences on small businesses — as confirmed by anecdotal evidence — because either of fines levied or of the mere cost of compliance, or litigation. Hence, even if a causal effect was present, it could not be unambiguously identified as a *reduction* of tax compliance — an actual reduction of income could be at play. To this, we add that the sample of taxpayers who are predicted to decrease their income in 2010 happens to have increased their income in 2009 by €6 446 on average. This makes it more likely that a reversion to the mean might explain at least part of the effect in the left tail.

4.2 Cost-Benefit Analysis

Given that our prediction model predicts whether a taxpayer is likely to increase her income declaration after an audit, we are now able to quantify the potential benefits of employing such methods in the identification of taxpayers to target with audits. We do this first by analyzing taxpayers predicted to be increasing from 2010 to 2011.

The random forest model with 100 trees predicts 1 495 *increasing* taxpayers among the 2 935 audited ones (close to 50%, a natural consequence of the output variable being balanced by definition). Looking at the net income reported by these taxpayers, we find, as already mentioned, that their average increase is around €14 461. Conversely, the average increase of taxpayers who are predicted to be decreasing is around €-15 431. Meanwhile, the general average increase for audited taxpayers in the years under study was €-205. This can be interpreted as saying that if the agency had audited *only* taxpayers predicted to increase their declared income in reaction to an audit, it would have obtained a reaction in subsequent tax declarations €14 461 -

[15] We thank an anonymous referee for this observation.

(-205) = €14 666 higher per audit, on average. This represents a sizeable amount, because the average declared income before the audit is €18 304, so that the estimated aggregated effect of the audit on this sample is an increase of revenues by 80%. Clearly, taxpayers with different declared income may be contributing differently to this effect: for reference, the median predicted percentage increase in declared income is 40%.[16]

In principle, given our focus on the *relative* change in declared income, the best performing prediction algorithm is not guaranteed to present the largest estimated increase in absolute declared income. However, we find that this happens to be the case: the best performing penalized linear model, a Lasso with $\alpha = 0.019$, results in an effect of €4 028 (instead of €14 666) the best decision tree (Figure 6) in an effect of €2978, and even the best AdaBoost configuration, despite reaching an F_1 *score* very similar to that of our preferred random forest model (see Table 10), obtains an estimated effect of only €6 460. So while a ranking based on the relative change is not necessarily consistent with one based on the size of the effect (as the Lasso and the decision tree would swap places), the random forest model happens to dominate in both cases.

In order to assess the external validity of our results, we again consider the results of the same prediction, run on the entire population of tax declarations for the same year (589 633 observations), hence including taxpayers not subject to tax audits. We first observe that our method predicts 0.55% of taxpayers to be increasing – that is, a *larger* share than among taxpayers who were indeed audited. In other terms, the tax agency followed criteria which are sharply different from what our prediction would suggest. Moreover, the presence of a large basin of taxpayers who are predicted to be increasing but are not audited according to current selection rules of the tax agency suggests that the tax agency could easily replace – i.e., stop auditing – audited taxpayers who are predicted as decreasing, and in doing so it would increase the effect of audits in terms of future revenues.

We acknowledge that increasing future revenues is not necessarily the main objective of a revenue agency. And in fact, had this been the unique goal, the natural approach would have been to focus on the median of the *absolute* increase in income, rather than the relative one, in the definition of the outcome variable. We chose to adopt the latter of these two measures

[16]The *average* percentage increase is 353%, but this figure is clearly influenced by a few very large relative increases, typically starting from very small declared incomes.

because while the magnitude of absolute changes is trivially related to the size of the economic activity of a taxpayer, this on the one hand does not provide any new insight on the heterogeneity of reactions to audits, and on the other hand, if adopted as the driving criterion, it would result in the agency directing *all* its efforts on the largest taxpayers, ultimately making tax audits extremely predictable for taxpayers, and possibly also raising fairness concerns.

Still, our results show that there is scope for the agency to increase revenues by considering the predicted increase at the taxpayer level as one of the determinants in the targeting of tax audits. Furthermore, while our approach is unable to discriminate the effect of targeting actual tax evaders from that of targeting taxpayers *who as a reaction reduce their future tax evasion*, it is likely that the features that predict an increase in declared income do at least in part identify concealed tax evaders. While further research should investigate to which extent this relationship holds, it represents a further argument why our approach can be a useful instrument in the hands of a tax authority.

It is worth noting that, since our data set is balanced at the origin, any taxpayers who were audited in 2010 but did not subsequently file a tax declaration are not present. This would apply for instance to professionals who happened to cease operations in the year of a tax audit. We believe the proportion of such taxpayers to be too small to affect our estimates. In any case, we have no reason to expect that taxpayers selected for an audit according to our scheme would have a larger propensity to cease operation than taxpayers selected according to any other criterion.

In addition, we have observed in the previous section that a limited portion of the measured effect might be due to a generalized propensity to increase declared income even in absence of an audit (€1 505), and an even smaller portion to a possible reversion to the mean (€517). Although these may erode the predicted benefits, they altogether represent less than 14% of the effect of interest.

Our analysis considers a year, 2010, during which Italy was enduring the harsh effects of the 2007-2008 financial crisis. This is also reflected in the overall average variation in declared income being substantially unchanged between 2009 and 2010 (see Section 3 and Figure 1), a clear sign of stagnant growth. In principle, this could limit the applicability of our findings to periods of normal growth. However, the presence, among audited taxpayers, of taxpayers who will increase their income in reaction to an audit is hardly

a specificity of a stagnant economy. On the contrary, it is likely that our selection method could result in a higher increase in accrued revenues in years of more sustained growth. Still, analyzing how relative distribution of post–audit changes in declared income relates to the contingent conditions of the economy is an important endeavour for future research.

5 Conclusions

We employ ML models to identify those taxpayers who are more likely to increase their tax return after an audit, and show that this prediction exercise can be attributed a causal interpretation – that is, it identifies a heterogeneity in the *reaction* to tax audits.

This has important implications for the ability of the revenue agency to indirectly increase revenues, by means of targeted auditing. We find that audited taxpayers tend to feature, throughout the year in which they are audited, more extreme relative variations as compared to the general population of taxpayers: that is, they tend to either increase more, or decrease more, their declared income from one year to the other, and so discriminating the two behaviors in advance is particularly important.

Our prediction exercise focuses on a target variable denoting whether a taxpayer increased its declared income from 2010 to 2011 more than the population median, in relative terms. We find that classification trees are able to predict a substantial component of the taxpayers' reaction, obtaining an out-of-sample F_1 *score* of 0.58, and that random forests and AdaBoost further improve the prediction, both reaching an F_1 *score* of 0.60. On the other hand, penalized linear models such as Lasso and ridge regression reach a maximum F_1 *score* between 0.55 and 0.56.

In most of the methods employed, the gross income is the most important variable to predict if a taxpayer will increase his net income declaration after an audit: in particular, a larger declared income seems to be a predictor of a *lower* probability to increase one's declared income in reaction to an audit. In general, variables related to revenues and costs have also an important role in the prediction. Among demographic variables, only the taxpayer's age seems to be relevant for our prediction. On the other hand, neither geographical and sector variables seem to have an important role.

We run an analysis of the potential benefits, in terms of subsequent tax revenues, from using the results of such prediction to target audits. We

find that taxpayers who are predicted to react positively to an audit would on average increase their declarations by €14 461 relative to an average audited taxpayer. We further show that the profile of a taxpayer who reacts positively to an audit does not simply correspond to the profile of a taxpayer who increases its income from a year to the following one, supporting our causal interpretation. We can hence conclude that predicting the reaction to audits, despite not necessarily be the first objective of a revenue agency, can still be an important instrument in its hands in order to increase revenues.

References

Advani, A., W. Elming, and J. Shaw (2018). The dynamic effects of tax audits. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, Volume 111, pp. 1–30. JSTOR.

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics 1*(3-4), 323–338.

Alm, J., B. R. Jackson, and M. McKee (2009). Getting the word out: Enforcement information dissemination and compliance behavior. *Journal of Public Economics 93*(3-4), 392–402.

Andini, M., M. Boldrini, E. Ciani, G. De Blasio, A. D'Ignazio, and A. Paladini (2022). Machine learning in the service of policy targeting: the case of public credit guarantees. *Journal of Economic Behavior & Organization 198*, 434–475.

Andini, M., E. Ciani, G. De Blasio, A. D'Ignazio, and V. Salvestrini (2017). Targeting policy-compliers with machine learning: an application to a tax rebate programme in italy. *Bank of Italy Temi di Discussione (Working Paper) No 1158*.

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of Economic Literature 36*(2), 818–860.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics*. Princeton university press.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pp. 507–547. University of Chicago Press.

Beer, S., M. Kasper, E. Kirchler, and B. Erard (2020). Do audits deter or provoke future tax noncompliance? evidence on self-employed taxpayers. *CESifo Economic Studies 66*(3), 248–264.

Best, M., J. Shah, and M. Waseem (2021). The deterrence value of tax audit: Estimates from a randomized audit program. *Working Paper*.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2015). Once bitten, twice shy? the lasting impact of irs audits on individual tax reporting. *Journal of Financial Economics 117*(1), 122–138.

Dubin, J. A. (2007). Criminal investigation enforcement activities and taxpayer noncompliance. *Public Finance Review 35*(4), 500–529.

D'Agosto, E., M. Marigliani, and S. Pisani (2014). Asymmetries in the territorial vat gap. *Argomenti di Discussione of Italian Revenue Agency 2*.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

García, V., R. A. Mollineda, and J. S. Sánchez (2009). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pp. 441–448. Springer.

Gemmell, N. and M. Ratto (2012). Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal 65*(1), 33.

Hashimzade, N., G. D. Myles, and B. Tran-Nam (2013). Applications of behavioural economics to tax evasion. *Journal of Economic Surveys 27*(5), 941–977.

Kasper, M. and J. Alm (2022). Audits, audit effectiveness, and post-audit tax compliance. *Journal of Economic Behavior & Organization 195*, 87–102.

Kasper, M. and M. Rablen (2022). Tax compliance after an audit: higher or lower? *Journal of Economic Behavior & Organization*.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica 79*(3), 651–692.

Krawczyk, B., M. Woźniak, and G. Schaefer (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing 14*, 554–562.

Lancee, B., L. Rossel, and M. Kasper (2022). When the agency wants too much: Experimental evidence on unfair audits and tax compliance. *Mimeo*.

Mazzolini, G., L. Pagani, and A. Santoro (2022). The deterrence effect of real-world operational tax audits on self-employed taxpayers: evidence from italy. *International Tax and Public Finance 29*(4), 1014–1046.

Mittone, L. (2006). Dynamic behaviour in tax evasion: An experimental approach. *The Journal of Socio-Economics 35*(5), 813–835.

Mittone, L., F. Panebianco, and A. Santoro (2017). The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology 61*, 225–243.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Oshiro, T. M., P. S. Perez, and J. A. Baranauskas (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition*, pp. 154–168. Springer.

Plumley, A. H. (1997). The determinants of individual income tax compliance: Estimating the impacts of tax policy, enforcement, and IRS responsiveness. *Department of the Treasury Internal Revenue Service Publication 1916*.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review 105*(8), 2539–69.

Ratto, M., R. Thomas, and D. Ulph (2013). The indirect effects of auditing taxpayers. *Public Finance Review 41*(3), 317–333.

Santoro, A. and C. V. Fiorio (2011). Taxpayer behavior when audit rules are known: Evidence from italy. *Public Finance Review 39*(1), 103–123.

Smola, A. and S. Vishwanathan (2008). Introduction to machine learning. *Cambridge University, UK 32*(34), 2008.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Yitzhaki, S. (1974). A note on income tax evasion: A theoretical analysis. *Journal of Public Economics 3*, 201–202.