DEMS
Department of Economics, Management and Statistics

DEMS WORKING PAPER SERIES

Optimizing Tax Administration Policies with Machine Learning

Pietro Battiston, Simona Gamba and Alessandro Santoro

No. 436 – March 2020

Department of Economics, Management and Statistics
University of Milano – Bicocca
Piazza Ateneo Nuovo 1 – 2016 Milan, Italy
http://dems.unimib.it/

Optimizing Tax Administration Policies with Machine Learning

Pietro Battiston[*], Simona Gamba[†], Alessandro Santoro[‡]

March 5, 2020

Abstract

Tax authorities around the world are increasingly employing data mining and machine learning algorithms to predict individual behaviours. Although the traditional literature on optimal tax administration provides useful tools for ex-post evaluation of policies, it disregards the problem of which taxpayers to target. This study identifies and characterises a loss function that assigns a social cost to any prediction-based policy. We define such measure as the difference between the social welfare of a given policy and that of an ideal policy unaffected by prediction errors. We show how this loss function shares a relationship with the receiver operating characteristic curve, a standard statistical tool used to evaluate prediction performance. Subsequently, we apply our measure to predict inaccurate tax returns issued by self-employed and sole proprietorships in Italy. In our application, a random forest model provides the best prediction: we show how it can be interpreted using measures of variable importance developed in the machine learning literature.

Keywords: policy prediction problems, tax behaviour, big data, machine learning
JEL classification: H26, H32, C53.

[*]University of Parma, Italy, me@pietrobattiston.it
[†]Catholic University of Milan, Italy, gamba.simona@gmail.com
[‡]DEMS, University of Milan-Bicocca, Italy, alessandro.santoro@unimib.it. This research is part of a research agreement between DEMS and the Italian Revenue Agency, that we warmly thank for providing the data as well as further assistance.

1 Introduction

In recent years, economic research has increasingly shown interest in prediction policy problems (Kleinberg et al., 2015). For example, the efficiency of the teacher selection process can be improved by predicting which teacher will add greater value (Rockoff et al., 2011); similarly, to design a policy that can prevent shooting at school, it is essential to make a prior correct identification of youths most at risk of violence (Chandler et al., 2011).

Prediction is crucial for the design of tax administration policies, but the literature on this aspect is limited. In Belgium, tax authorities have developed different predictive models that helped in reducing the tax debt (OECD, 2019). These models are used to predict if a person or a company will pay any withstanding debts within 14 days of receiving a phone call, or after a later visit by a bailiff. In Canada, a comprehensive dataset was formed linking taxpayers' filing and assessment information, risk profiles, historical audits, collections and appeal information (OECD, 2019). Based on such data, machine learning algorithms are used to predict income and sales taxes that small and medium enterprises should pay, or for attributing a risk score to each taxpayer.

While prediction can improve the efficiency of any tax policy that which is administered individually to taxpayers, it is particularly interesting in the context of *proactive* policies. Designed to *ex ante* promote voluntary tax compliance, these policies are being increasingly adopted by tax authorities around the world (OECD, 2017, p. 54). These policies increase tax revenues and may improve the relationship with taxpayers, also by avoiding some of the costs typically associated to traditional *reactive* policies (such as tax audits).

The characterization of an optimal prediction-based policy is different from that of a standard tax administration policy. Keen and Slemrod (2017) show that optimal tax administration policies, ranging from desk and field audits to customer services, are those which balance their revenue effect with their social cost. In particular, audits are correctly designed when, at the margin, the additional revenue they provide (with respect to the case where no audit is conducted) is exactly equal to the sum of public (e.g. administrative) and private (e.g. compliance) costs.[1] Hence, standard tax administration policies are typically evaluated on an ex-post basis.

When prediction-based policies are taken into consideration, it is essential to conduct an ex-ante evaluation of their efficiency. This efficiency depends

[1] Administrative costs typically include the costs to finance tax authorities, while compliance costs are associated with the requirements of tax rules and procedures.

not only on the expected frequency of both *false negatives* and *false positives*, but also on the estimated loss of revenue and additional costs associated with false negatives and false positives. Unlike a standard policy, a predictive policy may be inefficient even if, for every action correctly undertaken, the revenue exceeds the social costs. This can happen if a large cost is paid for wrongly undertaken actions.

In the last few years, prediction methods have witnessed a substantial boost in their performance, differentiation, and range of applications. Today, machine learning approaches have become integral to the *modus operandi* of firms globally, and their adoption by public bodies and researcher, including economists (see Varian, 2014 and Kleinberg et al., 2015), has seen a surge. In the machine learning literature, a general scheme of training and validation allows the implementation and comparison of different predictive algorithms, with penalised linear methods, decision trees, and neural networks being the most common examples. Several different approaches to the measurement of prediction errors have also been developed, but they typically compare different machine learning methods from a methodological point of view, instead than comparing the costs and benefits of specific policies based on these methods.

This study makes two contributions. First, we identify a loss function that assigns a social cost to each prediction-based policy; such cost is identified as the distance from an ideal policy which optimally targets taxpayers, and its minimization allows us to identify the optimal policy. Second, we apply the loss function to the design of a policy based on the prediction of inaccurate tax returns issued by self-employed and sole proprietorship taxpayers in Italy.

The paper is organised as follows. In Section 2, we derive the loss function and we relate its interpretation to the theory of optimal tax administration policies. In Section 3, we apply our approach to an Italian data set of tax declarations. Although the best model (a random forest) is a 'black box', in Section 4, we provide some insights on the interpretation of our results. Section 5 concludes the study.

2 Optimal prediction-based policies

A predictive tax administration policy is activated towards a taxpayer i if and only if $p_{it} \geq \tau$, where $p_{it} \in [0, 1]$ is the estimated probability that a given administrative intervention will be effective, and the threshold τ is chosen by the tax authority. Therefore, a policy is characterised by a prediction model to calculate p_{it} and a threshold τ discriminating between targeted and non-targeted taxpayers.

Selecting a prediction model involves choosing a method (i.e. an algorithm) and a vector of (hyper)parameters, which are used to tune the algorithm (examples will be provided in Section 3.2). In the increasingly important field of *supervised machine learning*, the accuracy of prediction models is typically assessed by analysing their performance in out-of-sample prediction (Varian, 2014). Specifically, in a *cross-validation* procedure (Kleinberg et al., 2015), the sample is iteratively split into a training and a testing sample. Figure 1 represents the *bias-variance trade-off*: as the complexity of the prediction model increases, it becomes easy to reduce the in-sample bias. However, the variance increases simultaneously; that is, the model specialises on the specific sample and performs worse out-of-sample (overfitting). The *total* prediction error includes both the bias and variance components, and hence presents a U-shaped form with respect to the model complexity. The optimal model is characterised by an intermediate level of complexity, for which the total prediction error out-of-sample has been minimised.

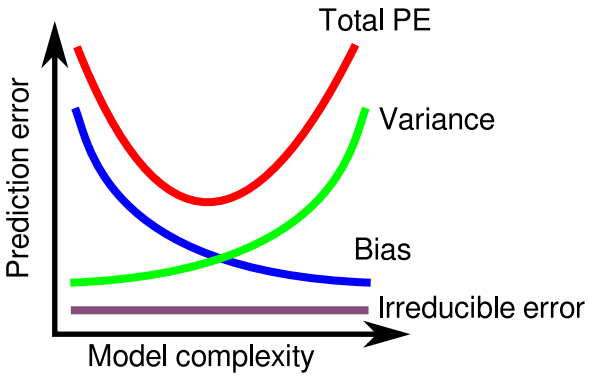


Figure 1: Model complexity and prediction error.

The goal of the tax authority is to exclusively target taxpayers who, in absence of the policy, would be non-compliant. Every prediction model can generate both type I (false positive) and type II (false negative) prediction errors. The false positive rate (FPR) denotes the share of compliant taxpayers wrongly predicted as non-compliant, and it is equal to 1-TNR, where TNR is the true negative rate (or 'specificity' in the machine learning literature). The false negative rate (FNR) is the share of non-compliant taxpayers wrongly predicted as compliant, and it is equal to 1-TPR, where TPR is the true positive rate (or 'sensitivity' in the machine learning literature).

Suppose that every administrative policy targeting a future non-compliant taxpayer yields, on an average, an increase in social utility β, for example

due to additional revenues.[2] However, the policy entails, on an average, a social cost for taxpayers (opportunity cost of using private resources, i.e. the time devoted by taxpayers to deal with the tax authority's requests) equal to δ and an average administrative cost (e.g. wages paid to tax officers involved in planning and implementing the policy) equal to γ. Let λ be the shadow cost of raising a dollar of budget (i.e. the unitary cost of distortionary taxation). Hence, a completely error-free intervention would generate social welfare equal to $a = \beta - \delta - (1+\lambda)\gamma$ per targeted taxpayer. We assume that $a > 0$, so that every correctly targeted intervention produces an increase in social welfare.[3] We then define $b = \delta + (1+\lambda)\gamma$ as the total cost of the intervention, so that $a = \beta - b = \beta - (\delta + (1+\lambda))\gamma$.

Given a prediction model that assigns each taxpayer a probability of tax evasion, we want to find a threshold τ that minimises the loss in social welfare with respect to an error-free policy. Denoting the size of the population and the number of non-compliant (*positive*) taxpayers by N and P, respectively, this loss can be written as:

$$\mathcal{L}(\tau) = aFNR(\tau)P + bFPR(\tau)(N - P) \tag{1}$$

where the first term is the social loss derived from not targeting a share FNR of non compliant taxpayers and the second is the loss derived from targeting a share FPR of compliant ones. A necessary condition for the optimality of τ is that $\mathcal{L}'(\tau) = 0$, that is:

$$\frac{FNR'(\tau)}{FPR'(\tau)} = -\frac{b}{a} \cdot \frac{N - P}{P} \tag{2}$$

The right-hand side of Equation (2) is strictly negative given that $P < N$. As τ increases, that is, as the rule becomes 'stricter' and the policy involves less taxpayers, type I errors become less frequent (FPR decreases) while type II errors become more frequent (FNR increases). Therefore, the condition can be interpreted as providing a marginal rate of substitution between the two types of errors. That is, for a given population – with a given share of non-compliant taxpayers – τ is optimal if the ratio between the marginal rate of change in the two errors is equal to the ratio of their weighted marginal costs.

[2] Assuming $\beta > 0$ is equivalent to assuming, in the terminology of Keen and Slemrod (2017), that the marginal social utility generated by every dollar of tax revenue is equal to $v'(g) > 1$, where $v'(g)$ is the social marginal utility of public goods and 1 is the social marginal utility of private consumption.

[3] We assume away the issue of *initial* budget, i.e. of administrative costs that need to be covered in advance.

2.1 Interpretation of the optimality condition

Equation (2) can be compared to the optimality condition for an administrative policy described by Keen and Slemrod (2017),[4] where the elasticity of reported income with respect to the policy equals a weighted sum of administrative and compliance costs. To see the similarity, consider the case of $\frac{FNR'(\tau)}{FPR'(\tau)} = -1$, where a marginal increase in τ generates an increase in false negatives which exactly offsets the decrease in false positives. In this case, Equation (2) becomes:

$$\frac{P}{N}\beta = b \tag{3}$$

which compares the social benefit of the policy (left-hand side) to its costs (right-hand side). The discount factor $\frac{P}{N}$ emerges from the degree of heterogeneity among taxpayers introduced by our model (with the policy ideally focusing only on the P positive taxpayers).

When we drop the assumption that $\frac{FNR'(\tau)}{FPR'(\tau)} = -1$; instead, we introduce a novel dimension of policy (in)efficiency, that is, we consider a trade-off between false positives and false negatives in the policy implementation, which represents this study's main focus.

From the point of view of the machine learning literature, equation (2) can be directly interpreted as a tangency condition on the receiver operating characteristic (ROC) curve, a graphical plot that shows the performance of a binary classifier system in repsonse to an increase in the discrimination threshold τ (each point of the curve corresponds to a different value of τ). This curve can be constructed as a plot of TPR versus FPR (Cali and Longobardi, 2015). Indeed, the slope of the ROC curve is $\frac{TPR'(\tau)}{FPR'(\tau)}$: by replacing $FNR(\tau)$ with 1-$TPR(\tau)$ in equation (2), this can be rewritten in terms of the derivative of the ROC (see Figure 2) as

$$ROC' = \frac{b}{a}\left(\frac{N}{P} - 1\right). \tag{4}$$

The ROC curve always has a positive slope (as both TPR and FPR are decreasing functions of τ) and it is usually represented as a concave function. While it is not necessarily true that any given prediction method would yield a concave ROC, this is a harmless assumption. Indeed, if there is an interval $[\tau_1, \tau_2]$ on which the ROC has a positive curvature, then the prediction for any intermediate value, $\tau \in [\tau_1, \tau_2]$, can be improved by replacing it with a convex combination of predictions calculated in τ_1 and τ_2. A consequence of

[4] See their Equation (27).

the concavity of the ROC is that the loss function $\mathcal{L}$ is single-peaked, that is, there cannot be multiple minima, and the above stated condition (Equation 4) is necessary and sufficient.

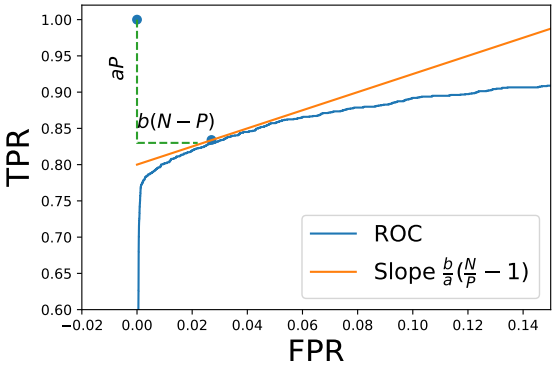


Figure 2: The optimal threshold as a tangency point

It must be noted that an error free prediction would be such that $TPR = 1$ and $FPR = 0$, so it would correspond to the point $(0, 1)$ in the ROC curve plot. Therefore, minimizing the loss function (Equation (2)) corresponds to minimizing a weighted distance from the ROC to $(0, 1)$, with weights aP and $b(N - P)$.

In practice, in order to minimise the loss function, the first step will be to estimate the ratio $\frac{b}{a}$. For example, Keen and Slemrod (2017) argue that, in the United States, every dollar of revenue raised (which corresponds in our model to $\beta = 1$), entails compliance costs of $\delta = 0.11$ and administrative costs of $\gamma = 0.006$. In presence of a shadow cost of $\lambda = 0.2$, we have $b = 0.117$ and $\frac{b}{a} = 0.133$. The second step will be to choose the model m and the threshold $\tau \in [0, 1]$ that minimise the loss function:

$$\min_{m} \min_{\tau \in [0,1]} \mathcal{L}_m(\tau). \tag{5}$$

3 Application: prediction of tax reports

We apply the framework described above in the study of taxpayers' response to the Italian business sector studies (*Studi di Settore*, SDS). Within SDS, small self-employed workers and sole proprietorships know the value of revenues (i.e. the total value of sales) the tax authority presumes that these businesses should report. Taxpayers also know that the probability to be

audited is lower if their reported revenues are at least as high as this presumptive value.[5] Taxpayers declaring at least the presumptive revenues are defined as *congruous*, while those declaring less than the presumptive revenue are *non-congruous*.

This institutional framework creates a strong incentive for taxpayers to report exactly the presumptive revenues (*'bunching'*). This may happen in two cases. First, taxpayers whose true revenues are above the presumptive ones can decide to report the latter to maximise post-tax income (bunching *from above*). This is similar to what happens when taxpayers' reports bunch at kink points of the marginal tax schedule. A taxpayer may also decide to report the presumptive revenues even if it is higher than the true one to avoid the increased risk of audit, or to avoid the cost of providing evidence of the true revenues (bunching *from below*). Both types of bunching are undesirable – while the former represents a form of tax evasion, the latter goes against the fairness of the tax system.

The presence of bunching calls for proactive, prediction-based policies that can steer in advance tax reporting behaviour in order to reduce administrative costs associated with ex-post audits and potentially avoid a subsequent litigation process. To see why, consider that the individual thresholds in the Italian SDS provide a legal weak presumption – for either the tax authority or the taxpayer. On the one hand, an audited non-congruous taxpayer will have to prove that the presumption does not apply to her case, for example because her input productivity is lower than that presumed by SDS. On the other hand, to audit a bunching (or any congruous) taxpayer, the Italian tax authority adopts a different and costlier enforcement technology, based on the traces generated by the paper trail, such as invoices issued and payments made within the production process. Obtaining these traces can be highly expensive for the tax authority, especially for firms selling goods and services to final consumers rather than to other businesses.[6] Hence, targeting bunchers in advance allows the tax agency to direct actions at them preemptively, thus increasing compliance without using costly audits.

[5] See Appendix A for a detailed description of how presumptive revenues are calculated. It must be noted that the presumptive revenues can, to some extent, be manipulated by the taxpayer. The presumptive revenues are the product of input productivities, as calculated by the tax authority, and input quantities reported by the taxpayer (see Santoro and Fiorio, 2011).

[6] 'In practice, it is relatively easy to cross-check tax returns to detect misreported intermediate input sales because the buying firm has an incentive to record its expenses to claim tax credits. In the case of final sales, the consumer has no incentive to keep a receipt and therefore it is significantly harder to cross-check those transactions against other information sources, especially when they are made in cash.' (Almunia and Lopez-Rodriguez, 2018).

Italian SDS present only one example of a threshold-based tax auditing strategy that results in strategic bunching. Almunia and Lopez-Rodriguez (2018), for instance, consider Spanish firms that strategically bunch below the eligibility threshold of €6 millions of reported revenues; this allows them to avoid stricter tax enforcement implemented by a 'Large Taxpayers Unit', with more – and highly experienced – auditors per taxpayer. One important difference between the Italian SDS and the Spanish example is that, in the former, the threshold can change across taxpayers, while, in the latter, it is fixed at a same level for all firms. Additionally, in the Spanish case, there is no discontinuity at the threshold in the enforcement technology – there is no change in the legal reporting requirements and the legal procedures available to process the information generated by business transactions.

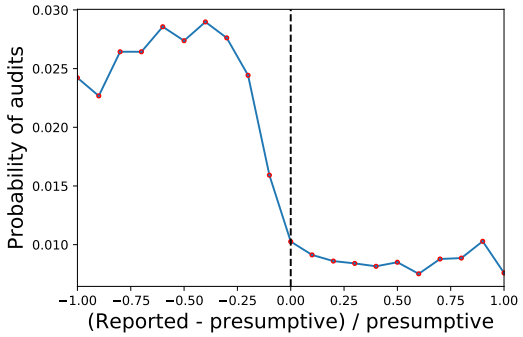
In the next section, we will use several observable characteristics of taxpayers at year $t-1$ to predict the bunching behaviour in year t. This approach can be used by the revenue agency to implement proactive actions aimed at fostering compliance.

3.1 Data

We analyse a dataset provided by the Italian Revenue Agency, which contains information on the reported income of the entire population of self-employed and sole proprietorships residing in the regions of Lombardy (North), Lazio (Centre), and Sicily (South), for the period between 2007 and 2011, included. The dataset has a perfectly balanced panel structure, where each of the 662 241 individuals is observed in each of the 5 years, for a total of 3 311 205 observations. For each observation, many pieces of information are available, summarised in 460 variables. These include

- demographic characteristics, such as age, gender, city and province of residence, and number of open VAT positions;
- detailed content of the tax reports, including main revenues, costs, tax bases, and the amount of tax due for three taxes – personal income tax (IRPEF), value-added local tax (IRAP), and VAT;
- information about possible audits, including whether a taxpayer was audited, the year and the amount ascertained and the outcome of the audit – whether the audit discovered evaded amounts and whether the taxpayer accepted to pay or a litigation was initiated;
- compliance with respect to SDS – congruity/non-congruity status.

Figure 3: Observed audit probability for non-congruous and congruous taxpayers.



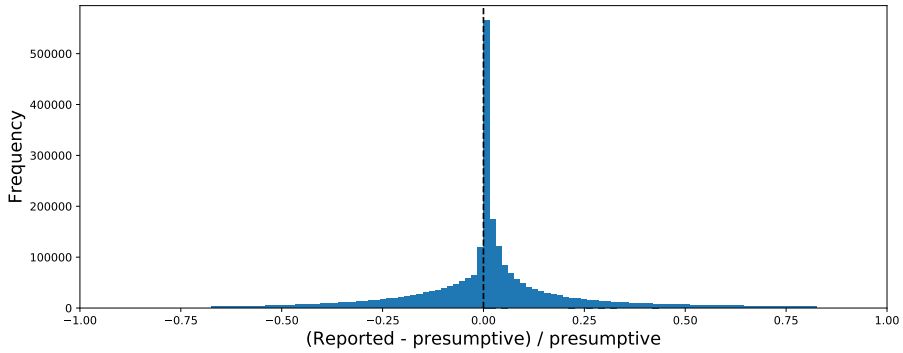
Note: The vertical line indicates a reported revenue equal to the presumptive one.

Two things emerge immediately from the data. First, Figure 3 shows that, in line with the policy design, the objective probability to be audited is higher for non-congruous than for congruous taxpayers. Moreover, at least on a left neighbourhood of the threshold (non-congruous taxpayers), audit probability is increasing in the percentage difference between reported and presumptive revenues.[7]

Second, taxpayers bunch at the presumptive revenues (see Figure 4), as expected. This is in line with previous evidence on the application of SDS (Santoro and Fiorio, 2011). Figure 4 gives an idea of the frequency of the bunching behaviour over the entire period: it shows a high frequency of taxpayers who report revenues in the neighbourhood of their presumptive value, for a specific bin size of 2%. Given that any bin size would be somewhat arbitrary, in the subsequent section we focus only on the prediction of exact bunching – cases where the reported revenue is exactly equal to its presumptive value. Table 1 displays the number and frequency of these cases. Moreover, we will label as bunchers only the (24 101) exact bunchers at *strictly positive* presumptive revenues, thus not considering taxpayers having zero presumptive *and* reported revenues.

[7]The instability of dots further away from the threshold reflects the reduced numerosity of taxpayers. The relatively low probabilities of audits at the left extremum might reflect cases from which the agency has limited hope of getting revenues. Overall, evidence seems to suggest that audit selection is indeed based, among other things, on a combination of congruity status and other factors that presumably reflect the predicted profitability of an audit.

Figure 4: Deviation of tax returns from presumptive value



Note: Given the histogram bin size of 0.02, the tallest bar represents all taxpayers whose declared revenues are larger than the presumptive amount by no more than 2%. The definition of 'bunchers' in our analysis is instead that the two amounts coincide.

Table 1: Bunching statistics

Year	Total	Bunchers	Share
2007	660019	6131	0.009
2008	657912	6335	0.010
2009	655954	4048	0.006
2010	656692	3891	0.006
2011	657696	3696	0.006

3.2 Prediction methods

We briefly describe the prediction methods we use in our application. Each method is characterised by different parameters that tune the prediction algorithm, typically referred to as *hyperparameters*. The choice of a prediction method and that of the values for hyperparameters determines the *prediction model*.

We start with two types of penalised linear models. This class of methods extends OLS by reducing the number of relevant regressors. Specifically, Lasso minimises $\sum_i (y_i - \beta X_i)^2 + \alpha \sum_b |\beta_b|$, while Ridge minimises $\sum_i (y_i - \beta X_i)^2 + \alpha \sum_b \beta_b^2$. A larger value of α – the only hyperparameter in this case – corresponds to a stronger penalization, and both methods reduce to OLS if $\alpha = 0$. In our empirical exercise below, we test such models on different

values of α (0.1, 0.01, 0.001), thereby attributing a widely different weight to the penalization term.

Decision (or classification) trees are nonlinear models used to predict a binary variable. In such models, a branching tree is constructed by automatically and iteratively splitting the sample between observations with a given variable larger or lower than a given value. The sample is ultimately partitioned into a number of subsamples such that the *impurity*[8] of each (i.e, the dispersion of the outcome variable within the subsample) is minimised. The depth (number of levels) of the tree is, in this case, our hyperparameter of interest. For example, consider Figure 5 – the splitting rule is reported at the top of each branching (non-leaf) node. The blue (red) colour represents a higher (lower) ratio of bunchers within the node. In the following section, we experiment with depths in the range from 1 to 10 (as we find that larger depths result in overfitting).

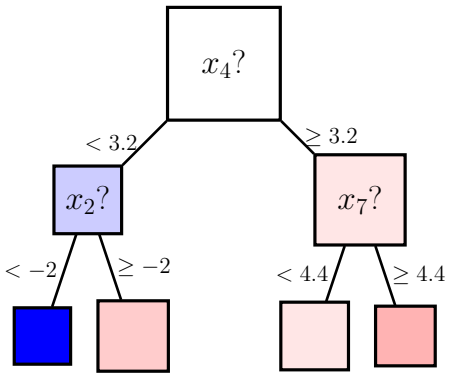


Figure 5: An example of decision tree.

Subsequently, we experiment with random forests. The principle behind such algorithms is exactly the same as decision trees; however, whereas in the discussion above one tree only was constructed and used to predict elements of the test sample, random forests comprise a pool of trees. Specifically, the random forest algorithm can be described as follows (see Figure 6). First, n_{tree} samples of equal size are randomly drawn from original data. Second, a decision tree is trained on each of these samples, with the following modification: at each node, the best split is chosen among a random subsample

[8]A typical impurity measure is the Gini impurity. Given a set of n observations subdivided into K classes, its Gini impurity can be calculated as $1 - \sum_{k=1}^{K} p_k^2$, where p_k is the frequency of class k (hence $\sum_{k=1}^{K} p_k = 1$). Thus, the impurity is equal to zero when all observations are of a same class, and it is maximised (taking a value of 0.5 in the case K=2) when all classes are equally frequent.

of m_{try} predictors. Third, a prediction is made for new data by aggregating the predictions of the n_{tree} trees using the majority vote criterion. It was observed empirically that this strategy makes random forest more robust than decision trees, in particular to overfitting (Breiman, 2001; Friedman et al., 2001). m_{try} and n_{tree} (together with the number of levels) are hyperparameters. As highlighted by Friedman et al. (2001) (p. 596), random forests suffer little from overfitting when employing larger trees; while we still find that limiting the depth slightly improves performance, we will test far larger values than those with decision trees, up to 40.

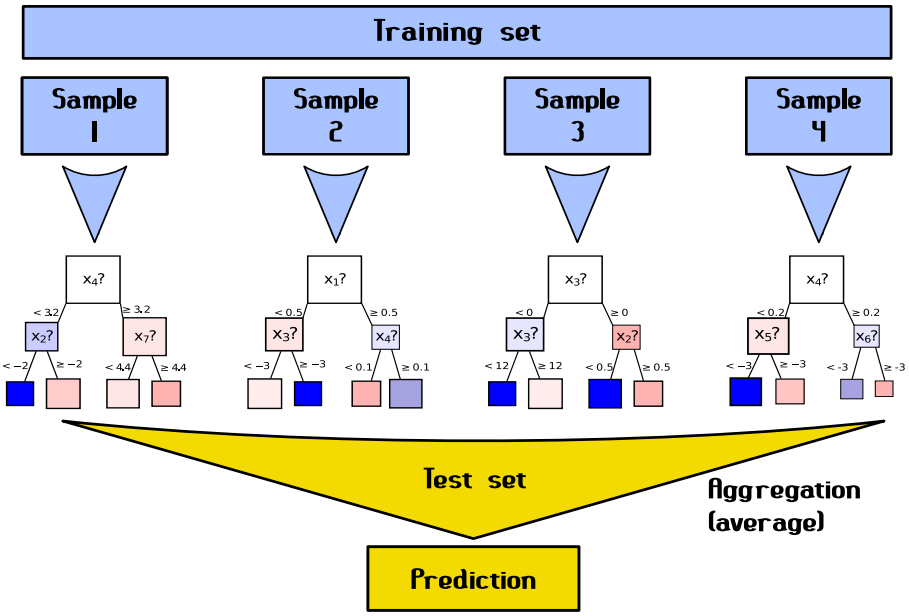


Figure 6: An example of random forest.

With both decision trees and random forests, the most relevant parameters are the number of nodes or branches. Given a tree with L levels $l = 0, \ldots, L-1$, each of them includes as many as 2^l nodes, and hence the number of terminal nodes will be at most 2^{L-1}. The predicted bunching probability for every taxpayer is the frequency of bunchers within the terminal node to which the taxpayer belongs.

Finally, we consider *neural networks*. Currently, this is arguably considered the most successful machine learning method, both in academia and in the industry, owing to its versatility. A neural network can be seen as a series of layers of regressions (logistic regressions, for instance), in which the output of a layer is the input of the following layer. The advantage of neural networks is their extreme flexibility – weights between the different neurons

(nodes) of the network vary by adapting to the training. However, their disadvantage lies in the unfeasibility to interpret their content, whereas decision trees and related methods, as we shall see, give us an intuition about the basis of the selection rule. Neural network can be made more or less complex by changing the structure of the hidden layers. For example, in our exercise below, the label $(7, 7, 7, 7, 7)$ will denote a neural network with five hidden layers, each composed of 7 nodes. With neural networks, the predicted probability of bunching is given by the output of the last level. In a neural network, many different aspects can be tuned; however, in the next section, we mostly focus on the interplay between the number and size of the hidden levels and the prediction ability of the model.[9] Hence, the hyperparameters of the model will consist of the sequence of the sizes of its hidden layers, such as $(7, 7, 7, 7, 7)$ in the exercise below. We will test neural networks with varying numbers of levels – from 1 to 7 – of varying size – from 1 to 7. Indeed, given the available data, larger models tend to easily result in overfitting.

3.3 Optimal prediction

We now proceed to analysing the performance of the models described in the previous section when applied to our empirical exercise.

First, we show how the in-sample ('training') and out-of-sample ('prediction') errors change across models, as measured with the area under the ROC (AUC) – a standard approach in the machine learning literature. In Figure 7, each marker corresponds to a different model – the resulting curve is the empirical counterparty of the 'Total error' displayed in Figure 1.[10] Importantly, such measure aggregates the prediction error across all possible thresholds.[11]

In order to apply the method presented in Section 2.1 to identify the optimal policy, we first need to select a proper value for the ratio $\frac{b}{a}$ (as P and N are instead determined empirically: in our case, P is the average number of bunchers over the observed time span). Recalling that $a = \beta - b$, this ratio depends ultimately on the comparison between β, i.e. the additional social

[9] In particular, all the neural networks we train are composed of dense layers (given the heterogeneity of the data, and the absence of space or time dimensions), and all use the standard ReLU activation function.

[10] Recall that variables referring to year $t-1$ are used to predict the bunching behaviour in year t. Hence, we cannot predict the bunching behaviour for the first year included in our sample (2007).

[11] See Appendix B for another possible measure of prediction quality, the pseudo-R^2 statistics, that is, the ratio between the mean squared error and the variance in the data.

Figure 7: Prediction error as measured by AUC

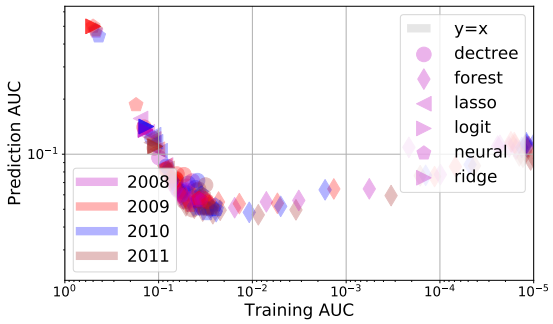
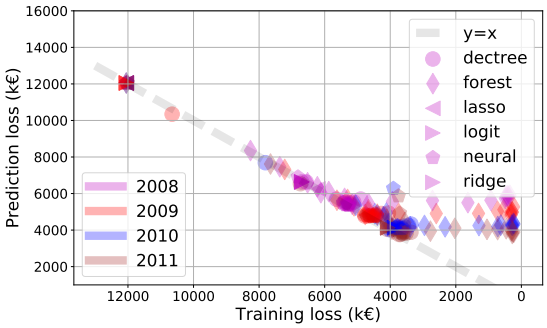


Figure 8: Prediction error as measured by $\mathcal{L}$.



Note: equivalent of Figure 7 obtained when evaluating models based on $\mathcal{L}$ and $\frac{b}{a} = 0.5$.

utility that every administrative action can generate, and b, i.e. the social cost (administrative + private) of that action. Some back-of-the envelope calculations suggest that a plausible value for b is around €1,000,[12] and a plausible value of β is in the range between €2,000 and €5,000[13], so that

[12]Proactive policies are delivered through customer services and often consist of a one-to-one relationship between a tax officer and the taxpayer. Suppose that this involves 10 hours of working time for each of the two parties, and that each hour is worth 50 €: under these hypotheses, $b = 50 \times 10 \times 2 = 1000$ €.

[13]Consider bunchers from above (tax evaders); since, in Italy, 60% to 70% of self-reported income is evaded (MEF, 2019) and bunchers report €20,000 of taxable income, on an average, if their propensity to evade matches the average, then their income evasion would total to approximately €30,000. With an average tax rate of 33%, this roughly corresponds to €10,000 of evaded taxes. Given that the prediction can be used to implement 'soft'

the ratio $\frac{b}{a}$ should be in the interval $(0.25, 1)$. Second, for each model m, we calculate τ_m^* finding the minimum value of the loss function.

Figure 8 is the equivalent of Figure 7 when adopting our loss function $\mathcal{L}$ as a measure of fit; it shows the prediction error, as measured by $\mathcal{L}$ for the best performing model (i.e. calibration of hyperparameters), and threshold within each method. Table 2 summarises the results of our exercise. It shows, for each year in our sample and each prediction method analyzed, which hyperparameters characterise the best performing model. It also shows how each of these models performs with respect to the other methods, displaying the resulting value for the loss function $\mathcal{L}$, in the case of $a = 2,000$, $b = 1,000$. For reference, the welfare benefits of an ideal policy with perfect prediction are in the range between 6 and 12 M€, depending on the number of actual bunchers in that year. As highlighted by the results reported in Table 2, different methods result in markedly different policy effectiveness levels. The last two columns provide the threshold τ characterizing each model and the resulting share of targeted subjects. For instance, for a random forest trained to predict bunching in 2011, the loss function is minimised when considering 25 levels and adopting a threshold of $\tau = 0.4$; this implies that all taxpayers with an individually estimated probability of bunching larger or equal than this value should be targeted. This model results in 0.67% of taxpayers being targeted.

Results are reasonably stable across years. Random forests are the best performing models in all years; in particular, they significantly outperform neural networks. Decision trees, while ranking behind random forests, as expected, perform relatively well – for relatively shallow configurations. Linear models are quite clearly outperformed by the previously mentioned models; this holds to a limited extent for Lasso with low α (the best performing Lasso), and to a larger extent for logit and Ridge (see also the significant difference between the value of $\mathcal{L}$ of these models and that of the others).

It is worth observing that the share of targeted subjects is low (consistently with the low observed frequency of bunching) and heterogeneous across models. In particular, best performing models tend to result in a larger share of targeted taxpayers, while some models (e.g. logit and ridge in 2009) do not lead to targeted taxpayers. In a context characterised by the rarity of the bunching behaviour, if the predictive algorithm is not efficient, then best option might not even be to target any taxpayer.

In general, there is no mechanical relationship between the threshold and

policies (e.g. nudging) rather than actual audits, it seems reasonable to consider a recovery rate between 20% and 50% of evaded taxes, that is, a β in the range between €2,000 and €5,000.

the share across models. Within the same model, increasing the threshold will necessarily reduce the share of targeted taxpayers; however, different models will result in different distributions of individual predicted probabilities, of which the share represents the cumulative density function computed at the given threshold.

Table 2: Best hyperparameters and prediction results, by year and method

year	method	alpha	depth	levels	$\mathcal{L}$ (k€)	Threshold	Share (%)
2008	forest		30		350	0.300	1.169
2008	dectree		8		4900	0.571	1.016
2008	neural			7	5251	0.364	1.047
2008	lasso	0.001			6202	0.576	1.066
2008	logit				6701	1.000	1.255
2008	ridge	0.100			6701	1.000	1.255
2009	forest		40		237	0.400	0.745
2009	dectree		8		4056	0.333	0.583
2009	neural			7,7,7,7,7	4410	0.485	0.574
2009	lasso	0.001			4727	0.281	0.560
2009	logit				12050	1.000	0.000
2009	ridge	0.001			12050	1.000	0.000
2010	forest		35		224	0.400	0.698
2010	dectree		8		3369	0.333	0.592
2010	neural			7,7,7,7,7	3608	0.296	0.590
2010	lasso	0.001			3843	0.370	0.569
2010	logit				3866	1.000	0.572
2010	ridge	0.010			3866	1.000	0.572
2011	forest		25		230	0.400	0.672
2011	dectree		8		3370	0.321	0.620
2011	neural			7,7,7,7,7	3553	0.334	0.604
2011	lasso	0.001			3915	0.400	0.539
2011	logit				4078	1.000	0.687
2011	ridge	0.010			4078	1.000	0.687

Note: 'alpha', 'depth', 'levels': method–specific hyperparameters (see Section 3.2). 'Threshold': threshold value for being targeted by the policy, resulting in the lowest value of loss function. $\mathcal{L}$: corresponding value of the loss function (thousands of euros). 'Share': corresponding share of targeted taxpayers.

4 Interpretation

One might be tempted to believe that correlations between the variable to be predicted and the predictors, as well as the measures of variable importance provided by statistical packages, might lead to inferring something about the model underlying the data (Mullainathan and Spiess, 2017).

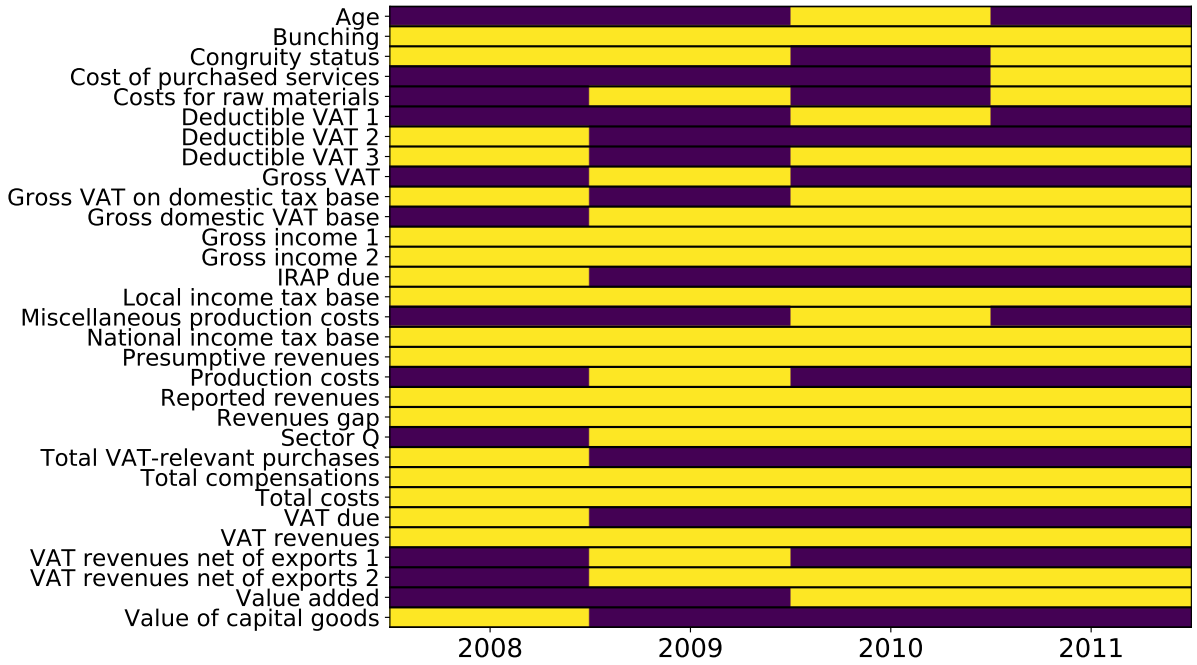
However, Mullainathan and Spiess (2017) argue that one of the problems in finding an underlying model is the intrinsic *instability* of the prediction, which they see as an Achilles heel. Indeed, on the one hand, 'the very appeal of these algorithms is that they can fit many different functions'; on the other hand, it often occurs that a same variable assumes a different importance in different partitions of the data, so that there are few stable patterns of predictors. The authors illustrate instability by showing that predictors selected by a LASSO regression change significantly from one random sample to another in the same data set mainly because the predictors are correlated to each other.

There are good reasons to believe that this instability problem is less severe with random forest models. In such models, variable importance can be measured with the Mean Decrease Impurity (MDI) proposed by Louppe et al. (2013). This is calculated in two steps. First, at the node level, the impurity decrease is computed as the difference between the impurity of the node and the weighted sum of the impurities of the two children nodes (weights are the frequencies of observations within each child node). Second, this measure is aggregated at the tree level by summing across all nodes employing the same variable and at the forest level by averaging over all trees. The resulting number is a measure of the variable's importance for prediction. Hence, a variable is more important if (i) it is used in many nodes, (ii) such nodes are located close to the root, (iii) in their children nodes, the impurity decreases significantly. The impurity of each node can be measured in different ways, but the Gini impurity index is typically used for the purpose. While we expect variable importance computed on single trees to have a high variability across trees of a random forest, MDI is averaged over all trees, and hence is likely to be more stable.

Concerning our application to Italian data, Figure 9 shows, in yellow cells, the 20 most important predictors (those with a higher MDI) for the best performing model (see Table 2) in each year. We can note that 11 variables in our sample are among the best predictors across all 4 years analysed; these variables include bunching, the level of presumptive and that of reported revenues, the level of reported gross income and tax base, total

compensations, and total costs.[14] While the remaining variables appear for some years only, there is still a strong overlap (only 31 variables appear in the top twenty at least once).

Figure 9: Most important variables per year



Note: yellow cells denote the 20 most important features in the best performing model for each year. Importance is measured according to MDI, employing the Gini impurity index.

Essentially, importance is a measure that does not provide information about the profile of the targeted population. To start with, one would like to know not only whether a variable is important for prediction, but also the *sign* of the correlation (if any) with the outcome variable. Some more insights can be gained by comparing descriptive statistics for most important predictors between the whole population and the targeted one.

In Table 3, in addition to the MDI, we report, for the subset of the 11 stable predictors emerging from Figure 9, the average values within the entire population (*All*), within the population of *actual* bunchers, and within the population of bunchers as *predicted* by the best performing model for year 2011.

[14]Recall that all predictors refer to year $y-1$, where y is the year for which bunching is predicted.

Table 3: Analysis of prediction results – important variables

	Importance	2011		
		All	Actual	Predicted
Bunch	0.24	0.01	0.79	0.84
Gross income 1	0.02	25537.54	63950.45	67053.76
Gross income 2	0.02	27052.80	64027.06	67120.25
Local income tax base	0.03	29944.97	66728.58	69478.17
National income tax base	0.02	29945.73	66729.46	69479.10
Presumptive revenues	0.02	85384.00	83383.00	86373.98
Reported revenues	0.02	100794.06	83140.63	86076.72
Revenues gap	0.14	-15410.06	242.37	297.27
Total compensations	0.02	76306.73	83863.35	86705.21
Total costs	0.02	52202.77	19961.34	19670.24
VAT revenues	0.01	104335.57	74135.38	76593.53

Note: data for year 2011. Features importance: MDI employing the Gini impurity index, with the best performing model. Figures are computed within the entire population (*All*), within the population of *actual* bunchers, and within the population of bunchers as *predicted* by the best performing model.

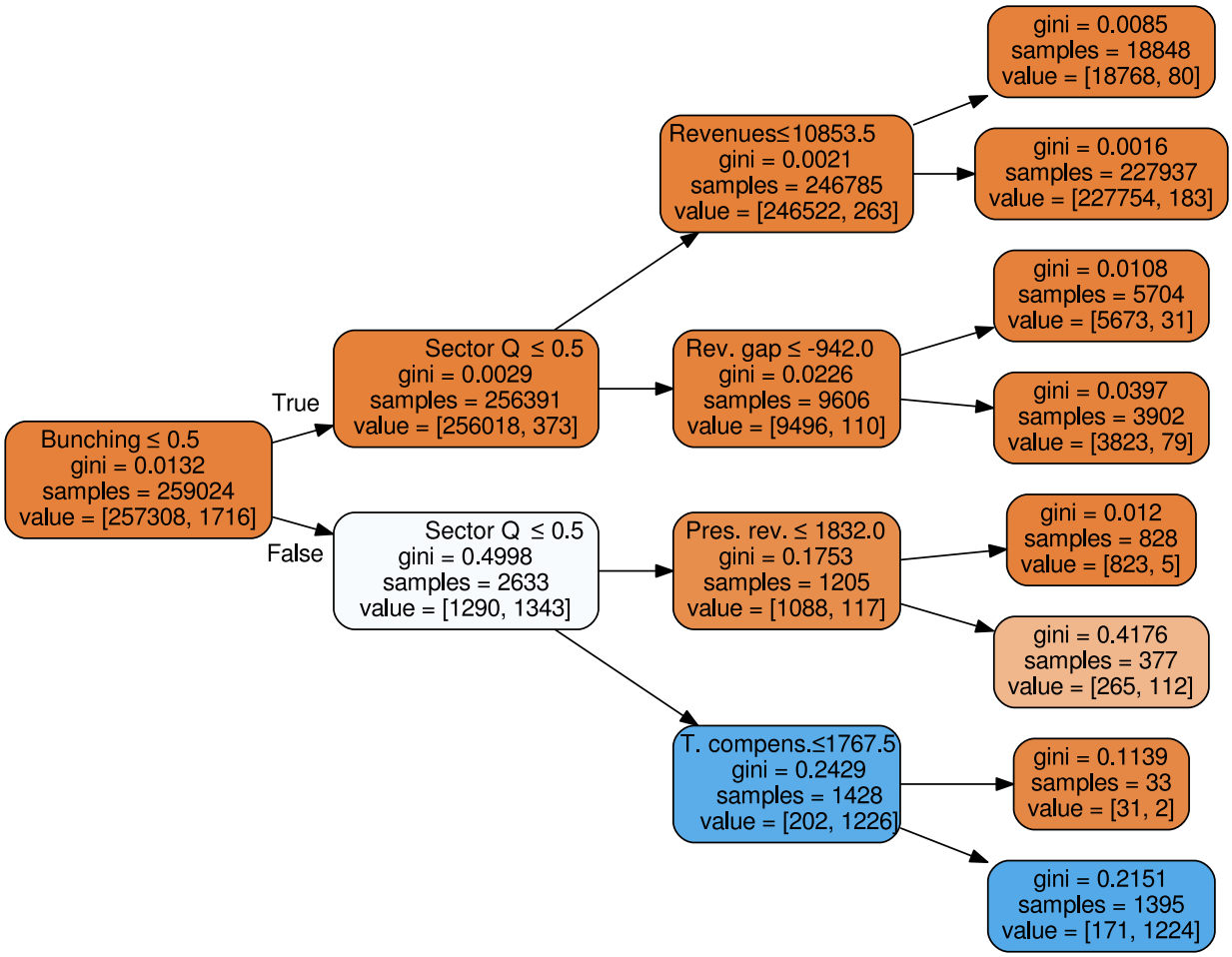
The profile of actual bunchers, as opposed to the entire population, is mainly characterised by a higher level of bunching, lower reported costs and revenues, and higher income, while having similar values of presumptive revenues. A comparison with the *Predicted* column shows that each of these characteristics is clearly captured by the prediction model. It is tempting to provide a comprehensive interpretation of this evidence. Recall that presumptive revenues can be reduced by underreporting costs. Hence, taxpayers with higher revenues and inputs (costs) can manipulate their declaration of the latter, thus decreasing presumptive revenues.[15] By bunching, they will also underreport revenues. Since, for each euro of underreported costs, presumptive revenues decrease by more than one Euro (see Appendix A), this strategy will allow bunchers to reduce the tax base with respect to the true one.

The complexity of interactions defining the profile of predicted bunchers using a random forest cannot be represented graphically. Still, interactions highly influencing the prediction are likely to be featured in the first levels of a decision tree trained on the same data as the random forest. Figure 10 shows the first three levels of the best performing decision tree for year 2011 (see Table 2). The persistence of bunching is confirmed; for taxpayers who were not bunching in the previous year and operate in the public services sector,

[15]This can explain the similarity of *Presumptive revenues* between bunchers (83 383€) and the whole population (85 384€).

a positive or mildly negative revenue gap also appears to be a predictor of bunching; the public services sector is also a relevant predictor of bunching for taxpayers who were bunching in the previous year. This is in line with the literature stressing the relevance of the traceability of operations for compliance decisions (Almunia and Lopez-Rodriguez, 2018) – public services are purchased by final consumers, and therefore are not traceable.

Figure 10: First 3 levels of the decision tree for year 2011



Clearly, machine learning algorithms are developed, and compared, with

the aim of delivering the best possible prediction using the available data, regardless of the interpretability of results or their consistency with an a priori model. In the Italian case considered, for instance, the outcome of the prediction, and the ultimate goal of the tax authority, is first and foremost a list of taxpayers predicted to be bunchers. Still, our example shows how some machine learning techniques also provide interesting insights about the prediction process, which can also be considered to ehance our knowledge on the economics of tax evasion.

5 Concluding remarks

The framework proposed in this study integrates the traditional theory of optimal tax administration with an approach to individual prediction, which is increasingly important for tax authorities around the world. This approach aims at maximizing social welfare, which is measured via the loss function $\mathcal{L}$, and results in an optimal prediction-based policy that is characterised by a profiling algorithm and an expected welfare impact. This approach is agnostic as to which supervised machine learning algorithm is used and allows a comparison between the prediction ability of different methods and hyperparameters values.

The tax evasion literature typically focuses on models where causal links between a set of independent variables and a dependent variable are hypothesised and, sometimes, validated empirically. Even when empirical evidence exists in support of these links, the use of these models for prediction can become ineffective. A characterisation of the entire *underlying model* of tax evasion to a reasonable degree of realism requires reliable measures of behavioural aspects such as 'limited computation abilities, misperceptions, hyperbolic discounting, non-standard preferences' (Alm, 2018). It would be a very complex model to define, and empirically calibrating its components would be a daunting task, especially in a field in which detailed empirical evidence is scarce (tax evasion is most often hidden from authorities). Hence, it is almost unfeasible to base predictions on theoretical and empirical insights coming from the tax evasion literature. Machine learning approaches employ data-driven methods that are less sensitive to the availability of variables; if there is a stable relationship between observable variables and the outcome of interest, then such a relationship can be identified and exploited for prediction.

The acceptance of an inductive data-driven approach does not imply giving up the interpretability of predictions. In this study, we have presented examples of how insights can be derived from trained machine learning mod-

els. These models implicitly define a profile of targeted individuals that can be further investigated using the existing literature as a benchmark.

We believe that our approach can be applied to the design and implementation of the entire range of tax administration policies. Our empirical application focuses on a *proactive* policy aimed at incentivizing the taxpayer to adopt a given tax behaviour in the future. However, the approach we describe also applies to standard *reactive* policies, implemented after the taxpayer has exhibited a given tax behaviour. Tax verification activities are far from being mechanisms that smoothly and efficiently uncover tax evasion, as typically postulated in theoretical models. Instead, they are socially costly activities, which include administrative costs as well as monetary and opportunity costs for involved taxpayers. Predicting the reaction of taxpayers targeted by reactive policies can improve the efficiency of these policies. For example, prediction can allow the tax authority to focus on audits yielding higher additional revenues (Beer et al., 2019) and on more effective debt collection policies.

References

Alm, J. (2018). What motivates tax compliance? *Journal of Economic Surveys 33*(2), 1–1.

Almunia, M. and D. Lopez-Rodriguez (2018). Under the radar: The effects of monitoring firms on tax compliance. *American Economic Journal: Economic Policy 10*(1), 1–38.

Beer, S., M. Kasper, E. Kirchler, and B. Erard (2019). Do Audits Deter or Provoke Future Tax Noncompliance? Evidence on Self-employed Taxpayers? IMF Working Papers, Fiscal Affairs Department 223, International Monetary Fund.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Cali, C. and M. Longobardi (2015). Some mathematical properties of the roc curve and their applications. *Ricerche matematiche*.

Chandler, D., S. D. Levitt, and J. A. List (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review 101*(3), 288–92.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Keen, M. and J. Slemrod (2017). Optimal tax administration. *Journal of Public Economics 152*(C), 133–42.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review, Papers and Proceedings 105*(5), 491–95.

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts (2013). Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 431–439. Curran Associates, Inc.

MEF (2019). Relazione sull'economia non osservata e sull'evasione fiscale e contributiva - anno 2019. Technical report, Ministry of Economy and Finance.

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

OECD (2017). Tax Administration 2017: Comparative Information on OECD and Other Advanced and Emerging Economies. Technical report, OECD Publishing.

OECD (2019). Tax Administration 2019: Comparative Information on OECD and Other Advanced and Emerging Economies. Technical report, OECD Publishing.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy 6*(1), 43–74.

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

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

A Business sector studies

Since 1998, Italy has adopted a system called 'Studi di settore' (SDS), which literally translates to '*Business Sector Studies*', to analyze the tax behaviour of taxpayers reporting annual revenues not higher than €5,000,000 in their economic activity. The Revenue Agency (RA) collects information on structural variables (e.g., size of offices and warehouses, location, sector of activity, main characteristics of customers and providers, etc.) and on accounting variables (inputs and costs). On the basis of structural variables only, the RA divides taxpayers into C clusters. In any given year t, and within each cluster I_c, with $c = \{1, 2, ..., C\}$, it then selects a group of taxpayers that are believed to be reliable. Then, for each cluster, the RA estimates on such taxpayers the relationship:

$$R_{i,t-3} = \beta'_{c,t-3}\mathbf{x}_{i,t-3} + \epsilon_{i,t-3} \tag{6}$$

where $R_{i,t-3}$ is the value of revenues reported by taxpayer i at time $t-3$, $\mathbf{x}_{i,t-3}$ is the $J \times |N_c|$ matrix of inputs at time $t-3$, and $\epsilon_{i,t-3}$ is an idiosyncratic error. $\beta_{c,t-3}$ is a $J \times 1$ vector of unknown productivity parameters for cluster c,: its estimation – using standard linear regression techniques – is denoted as $\mathbf{b}_{c,t}$.

Taxpayers are provided with a freely downloadable software, called Gerico, where the value of each element of $\mathbf{b}_{c,t}$ is reported. In other words, although the productivity vector is exogenous to the taxpayer, she is allowed to analyse it while reporting her own vector of inputs to declare, $\mathbf{x}_{i,t}$. This means that the presumptive revenues can be manipulated, although a normality analysis is performed so that too low values of $\mathbf{x}_{i,t}$ cannot be costlessly reported.

Hence, presumptive revenues for the taxpayer i belonging to the population of active taxpayers in cluster c and tax year t are calculated as:

$$\overline{R}_{i,t} = \mathbf{b}'_{c,t}\mathbf{x}_{i,t}.$$

A taxpayer is defined as *congruous* if she reports revenues which are at least equal to the presumptive one, and *non congruous* in the opposite case. The main difference between these two statuses is the following:

- a non congruous taxpayer has a higher probability to be audited, since the tax authority can ask her to justify why she has reported lower–than presumptive revenues – the burden of proof being onto the taxpayer;
- a congruous taxpayer has a lower but still non-zero probability to be audited: for instance, she can be audited if the tax authority has other

evidence on her possible evasion coming from cross-examination of accounting books, from on-site audits, from the analysis of her bank or credit card accounts, and so on.

Thus, whilst congruous taxpayers can be audited by the tax authority using only the evidence provided by other methods, non congruous taxpayers can be audited using either business sector studies or other methods, resulting in a higher audit objective probability for non-congruous taxpayers. These rules are well known, which suggests that the subjective probability to be audited is also likely to decrease if (at least) presumptive revenues are reported.

B Additional material

Since the prediction error varies across thresholds, also the R^2 varies. In Figure 11, we plot the prediction error using the R^2 when the threshold is fixed at a level such that the share of predicted and actual bunchers are the same.

Figure 11: Prediction error using R^2