Power-law Mixtures of Bayesian Forests for Value Added Tax Audit Case Selection

Christos Kleanthous
cs.kleanthous@edu.cut.ac.cy
Cyprus University of Technology
Limassol, Cyprus

Theodoros Christophides
theodoros.christophides@eecei.cut.ac.cy
Cyprus University of Technology
Limassol, Cyprus

Sotirios Chatzis
sotirios.chatzis@cut.ac.cy
Cyprus University of Technology
Limassol, Cyprus

ABSTRACT

Tax authorities need to maximize the yield of the limited tax audits they afford to perform each year. Thus, they need to predict the likelihood of a candidate audit resulting in a satisfactory yield; this predictive process is usually referred to as audit case selection. Random Forests (RFs) constitute a standard method for Value Added Tax (VAT) audit case selection. Despite, though, their success, their predictive performance is still below the expectations of tax authorities, that need to timely detect cases of significant audit yield potential. This lackluster performance is mainly attributed to the fact that RFs cannot deal with data that entail non-stationary nature, multiple modalities, or discontinuities. These are common characteristics of real-world datasets; thus, the incapacity to properly address them is a major suspect for undermining their performance. This work addresses these issues by considering a generative non-parametric Bayesian model with power-law behavior, capable of generating distinct (Bayesian) RFs over the observations space of the modeled data. This way, our approach enables capturing an indefinite number of distinct classification patterns, while being able to effectively handle outliers. The latter advantage is of paramount importance for the effectiveness of the modeling procedure in cases where few large parts of the observations space can be modeled by few RF classifiers, yet there is a large number of small parts of the observations space that require distinct RFs to be properly modeled (power-law nature). We provide an efficient algorithm for model inference, based on the variational Bayesian framework, and prove its efficacy using real-world datasets.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Value Added Tax, audit selection, random forests, non-parametric Bayesian mixture model, variational inference.

ACM Reference Format:
Christos Kleanthous, Theodoros Christophides, and Sotirios Chatzis. 2020. Power-law Mixtures of Bayesian Forests for Value Added Tax Audit Case Selection. In *ACM International Conference on AI in Finance (ICAIF '20), October 15–16, 2020, New York, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383455.3422515

ICAIF '20, October 15–16, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7584-9/20/10…$15.00
https://doi.org/10.1145/3383455.3422515

1 INTRODUCTION

The Value Added Tax (VAT) represents an excess of 30% of EU tax revenue [13]. VAT is a consumption tax collected on behalf of the state by the businesses, for almost all the services and goods consumed in the EU. According to the latest annual published European Commission report [12], the difference between the estimated and actual VAT collected, typically referred to as the VAT-gap, is estimated for all EU Member States at 12.3% (147 billions of Euros).

The Tax Authorities have the responsibility to the state for achieving the maximum taxpayer compliance with the tax legislation so as to maximize tax revenue. To achieve this goal the limited resources available need to be allocated carefully in order to achieve the highest possible taxpayer compliance with the tax laws and the best tax collections. One critical aspect of this issue is the prioritisation of the compliance action to be taken against different taxpayers. The taxpayer behaviour is very difficult to analyze due to diversity of compliance behaviour, absence of taxpayer motives for tax legislation compliance and the tax legislation complexity.

In general taxpayer compliance can be classified in four categories: i) Decided not to comply ii) Do not want to comply iii) Try to comply and iv) willing to comply. Therefore a systematic approach is usually adopted to identify the major risks in respect not only to the number of taxpayers not being tax compliant but also to the amount of tax and how it will be addressed.

The Legislation provides the tax departments with a plethora of different measures that can be employed to minimize non compliance from the taxpayers with different consumption of resources. Easy to navigate web pages, taxpayer education, publications, written communication require negligible resources from the tax authorities and can address thousands of taxpayers at the same time. The result is maximum voluntary compliance from taxpayers who are willing to comply or try to comply with the tax legislation with minimum effort. For example taxpayers who have not filed the latest tax return one week before the deadline can be send a reminder letter.

On the other hand in the case of taxpayers who decided not to comply or do not want to comply the opposite is true. Control audit visits are labor intensive and require limited and expensive resources. For an audit of a single taxpayer with a non compliant past a team of expert tax auditors is required Fig.1.

Therefore, it is key that we come up with effective audit case selection processes, with the aim of selecting the audit cases that are expected to yield significant tax. This procedure has to be performed by leveraging some sort of automation. To this end, a variety of

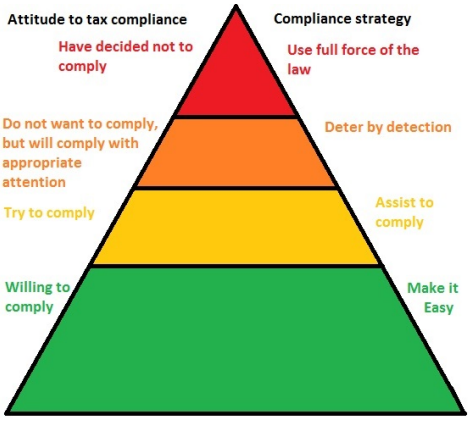

Figure 1: Attitude to tax compliance and compliance action

automation methods are currently employed. Traditionally, expert auditors resort to rules-based risk modelling systems for selecting audit cases [29]. However, this method of risk modelling is highly biased as it reflects the experts' understanding of taxpayer behavior, which may be partial and incomplete. In addition, as it is needed to write hundreds of rules, this is an expensive and time consuming procedure. Finally, it suffers from the limited capability of human experts to express their acquired experience in the form of rules. Besides, under this paradigm, selection is performed on the basis of the number of satisfied rules exceeding some heuristically-set benchmark. Apparently, this threshold value has to change each time we need to accommodate more or updated rules; this renders system maintenance almost prohibitive.

Due to these disadvantages, many EU tax authorities are currently considering advanced data analytics and machine learning as a promising alternative towards automated audit case selection [11]. Indeed, tax authorities collect and process millions of tax returns that contain a plethora of diverse information. Thus, it is reasonable to assume that tax return corpora naturally lend themselves to the formation of appropriate training datasets for successful machine learning-based audit case selection systems. A prevalent characteristic of the existing developments in the field is the lack of tailor-made machine learning models that can make the most out of the vastly available tax return data. For instance, Random Forests (RFs) constitute one of the most commonly used machine learning approaches in the context of tax audit case selection [16]. The main reasons behind their prominence can be traced to: (i) their capacity to effectively deal with high-dimensional data; (ii) their computational efficiency, both when it comes to training and when it comes to prediction generation; (iii) their notable robustness to outliers and non-linear features in the training data; (iv) their capability to effectively learn from unbalanced classification data, which are typical in real-world datasets stemming from tax audits.

Motivated from the importance of addressing the VAT-gap problem, this paper offers a state-of-the-art solution to VAT audit case selection. We posit that the development of tailor-made machine learning models that take into account the special characteristics of the VAT return data will allow for unleashing the untapped potential in the field. To this end, we build upon the existing state-of-the-art in the field, which is based on RFs. Our approach is inspired from the fact that a single RF model cannot sufficiently model data that entail non-stationary covariance functions, multi-modal output, or discontinuities. On the contrary, to model data with such properties via RFs, we need to fit different RFs, each on only a part of the data with coherent statistical behavior.

Our approach attacks these issues by considering a generative nonparametric Bayesian model with power-law properties, capable of generating distinct (Bayesian) RFs over the observations space of the modeled data. Our method comprises the postulation of a nonparametric Bayesian model based on the introduction of a Pitman-Yor [21] process prior over the space of possible RFs, defined in the whole space of input variables. Our approach is designed for better dealing with tasks entailing non-stationary covariance functions, multi-modal output, or discontinuities, which simple RF models may fail to successfully handle. Thus, it enables effectively capturing an indefinite number of distinct classification patterns, while being able to effectively handle outliers. The latter advantage is of paramount importance for the effectiveness of the modeling procedure in cases where few large parts of the observations space can be modeled by few RF classifiers, yet there is a large number of small parts of the observations space that require distinct RFs to be properly modeled (power-law nature). We dub our approach the Pitman-Yor process mixture of Bayesian forests (PYP-RF) for VAT audit case selection.

The remainder of this paper is organized as follows: In the following Section, we provide a brief overview of the related work. Then, we introduce our approach, elaborate on its rationale, and devise its training and inference algorithms, based on a truncated variational inference paradigm. Further, we perform an extensive experimental evaluation of our approach using real-world data from Cyprus tax authority. Finally, in the concluding Section, we summarize our contribution and outline open issues for further research.

2 METHODOLOGICAL BACKGROUND

2.1 Rules-based and data mining systems

The software SAS Enterprise Miner[1] has been used in the past to select audit cases for VAT purposes. Before use, an extensive feature engineering is a must to cater for missing data and categorize data in different groups.

The Irish Revenue Office has addressed the issue of selecting high tax yield non compliant taxpayers with the employment of data mining [10] using the SAS software (SAS Enterprise Miner and SAS Enterprise Guide). The experience gained from the banks and insurance companies is utilized against the non compliant taxpayers. According to the IDC [27] SAS and IBM have a 43% (revenue volume) combined market share of tools for predictive analytics.

The task of selecting few taxpayers with the highest expected audit yield accurately is almost impossible to be performed manually. The allocation of the experienced tax auditors needs high degree of accuracy to be translated in to maximum revenue as the number of audits that can be performed is limited. This cannot be achieved manually consistently.

[1] https://www.sas.com/en_us/software/enterprise-miner.html

Since the process of audit case selection cannot be facilitated manually the automation process of identifying high risk taxpayers and audit case selection was spearheaded by heuristic rules set by expert auditors [29]. A decision to audit a taxpayer depends if the number of rules that apply exceed a threshold, for example if fifty rules "fire" an audit is performed. This process is highly subjective since is based on the opinion of the experts and requires time and effort from a team of experts who create hundreds of rules for different areas like number of late returns, sales fluctuations, inconsistencies in the amounts declared etc.

After relying to the rules-based systems for many years the tax departments moved away from these complicated, subjective rules and thresholds and considered more reliable automated alternatives like advanced analytics and machine learning models [11]. Tax departments collect vast amounts of data for each and every taxpayer like filed returns which are not utilized. If exploited successfully this data can open the road for an automated process for selecting accurately taxpayers with high audit yields.

2.2 Bayesian Forests

RF models constitute one of the most popular methods for both regression and classification. Their functionality revolves around the concept of decision trees (DTs) [7]. As DTs are formulated by means of a random partition procedure, they constitute weak learners the performance of which may become underwhelming when dealing with difficult classification or regression tasks. To compensate for this weakness, RFs resort to the ensemble learning rationale: They fit multiple DTs on the same dataset, each performing different hierarchical random splits, $\boldsymbol{\theta}$, of the input space. Then, prediction is performed on the basis of an appropriate voting mechanism.

Recently, [25] introduced an alternative view towards RFs: Their empirical Bayesian forest (EBF) algorithm replaces the Poisson distribution, from which the tree parameters $\boldsymbol{\theta}$ are drawn, with an Exponential (or Dirichlet, when normalized) posterior, $\mathcal{T}(\boldsymbol{\theta})$. In this context, a suggested sample (random partition), $\boldsymbol{\theta}$, is retained if it facilitates the minimization of the Gini impurity index on the training dataset. This inferential treatment has been shown to induce a reliable performance gain across diverse application areas.

2.3 Bayesian Nonparametrics

Nonparametric Bayesian modeling techniques, especially Dirichlet process mixture (DPM) models, have become very popular for performing nonparametric density estimation [18, 19, 28]. Briefly, a realization of a DPM can be seen as an infinite mixture of distributions with given parametric shape (e.g., Gaussian). This theory is based on the observation that an infinite number of component distributions in an ordinary finite mixture model tends on the limit to a Dirichlet process (DP) prior [1, 19]. Eventually, as a part of the model fitting procedure, the nonparametric Bayesian inference scheme induced by a DPM model yields a posterior distribution on the proper number of model component densities (inferred clusters) [5], rather than selecting a fixed number of mixture components. Hence, the obtained nonparametric Bayesian formulation eliminates the need of doing inference (or making arbitrary choices) on the number of mixture components (clusters) necessary to represent the modeled data.

2.4 The Pitman-Yor (PY) process

DP models were first introduced by Ferguson [14]. A DP is characterized by a base distribution G_0 and a positive scalar α, usually referred to as the innovation parameter, and is denoted as $\mathrm{DP}(\alpha, G_0)$. Essentially, a DP is a distribution placed over a distribution. Let us suppose we randomly draw a sample distribution G from a DP, and, subsequently, we independently draw M random variables $\{\Theta_m^*\}_{m=1}^M$ from G:

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{1}$$

$$\Theta_m^*|G \sim G, \quad m = 1, \dots M \tag{2}$$

Integrating out G, the joint distribution of the variables $\{\Theta_m^*\}_{m=1}^M$ can be shown to exhibit a clustering effect. Specifically, given the first $M-1$ samples of G, $\{\Theta_m^*\}_{m=1}^{M-1}$, it can be shown that a new sample Θ_M^* is either (a) drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+M-1}$, or (b) is selected from the existing draws, according to a multinomial allocation, with probabilities proportional to the number of the previous draws with the same allocation [4]. Let $\{\Theta_c\}_{c=1}^C$ be the set of distinct values taken by the variables $\{\Theta_m^*\}_{m=1}^{M-1}$.

The PY process functions similar to the DP. Let us suppose we randomly draw a sample distribution G from a PY process, and, subsequently, we independently draw M random variables $\{\Theta_m^*\}_{m=1}^M$ from G:

$$G|\delta, \alpha, G_0 \sim \mathrm{PY}(\delta, \alpha, G_0) \tag{3}$$

with

$$p(\Theta_M^*|\{\Theta_m^*\}_{m=1}^{M-1}, \delta, \alpha, G_0) = \frac{\alpha + \delta C}{\alpha + M - 1} G_0 + \sum_{c=1}^{C} \frac{v_c^{M-1} - \delta}{\alpha + M - 1} \delta_{\Theta_c} \tag{4}$$

where v_c^{M-1} is the number of values in $\{\Theta_m^*\}_{m=1}^{M-1}$ that equal to Θ_c, $\delta \in [0, 1)$ is the discount parameter, $\alpha > -\delta$ is its innovation parameter, and G_0 the base distribution.

This way, the PY process gives rise to a rich-gets-richer clustering property, i.e., the more samples have been assigned to a draw from G_0, the more likely subsequent samples will be assigned to the same draw. Further, the more we draw from G_0, the more likely a new sample will again be assigned to a new draw from G_0. These two effects together produce a *power-law distribution* where many unique Θ_m^* values are observed, most of them rarely [21], thus allowing for better modeling observations with heavy-tailed distributions. In particular, for $\delta > 0$, the number of unique values scales as $O(\alpha M^\delta)$, where M is the total number of draws. Note also that, for $\delta = 0$, the PY process reduces to the DP.

A characterization of the (unconditional) distribution of the random variable G drawn from a PY process, $\mathrm{PY}(\delta, \alpha, G_0)$, is provided by the stick-breaking construction of Sethuraman [23]. Consider two infinite collections of independent random variables $\boldsymbol{v} = (v_c)_{c=1}^\infty$, $\{\Theta_c\}_{c=1}^\infty$, where the v_c are drawn from a Beta distribution, and the Θ_c are independently drawn from the base distribution G_0. The stick-breaking representation of G is then given by [26]

$$G = \sum_{c=1}^{\infty} \varpi_c(\boldsymbol{v}) \delta_{\Theta_c} \tag{5}$$

where

$$p(v_c) = \text{Beta}(1-\delta, \alpha+\delta c) \tag{6}$$

$$\varpi_c(\boldsymbol{v}) = v_c \prod_{j=1}^{c-1}(1-v_j) \quad \in [0,1] \tag{7}$$

and

$$\sum_{c=1}^{\infty} \varpi_c(\boldsymbol{v}) = 1 \tag{8}$$

The stick-breaking representation of the PY process makes clear that the random variable G is discrete. It shows explicitly that the support of G consists of a countably infinite sum of atoms located at Θ_c, drawn independently from G_0. Indeed, under the stick-breaking representation of the PY process, the atoms Θ_c, drawn independently from the base distribution G_0, can be seen as the parameters of the component distributions of a mixture model comprising an unbounded number of component densities, with mixing proportions $\varpi_c(\boldsymbol{v})$.

3 PROPOSED APPROACH

Let us consider a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, where $y_n \in \{0,1\}$ is the result of the nth audit, with the value of 1 corresponding to an audit yield deemed satisfactory by the tax authority (i.e., exceeding some threshold). The specific selection of this threshold value used in the development of our model will be discussed in the experimental section. In our work, the observed data points $\boldsymbol{x}_n$ are obtained from the VAT records of the taxpayers selected for audit. These are 47-dimensional vectors that comprise the following attributes: (i) economic activity type, classified according to the Eurostat NACE classification[2]; (ii) district codes; (iii) type of person (physical, legal); (iv) declared amounts, including VAT due/local sales, VAT due/EU purchases, VAT refundable (purchases), VAT payable, net value of sales, net value of purchases, value of zero-rated sales, value of purchases from EU (goods and services), and value of sales to EU (goods and services).

On this basis, the audit case selection task can be framed as a binary classification task. Since RFs currently constitute the most popular machine learning approach used for audit case selection, we initiate the formulation of our model considering that the random decision variables y_n can be expressed via a function $f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$, where the latent variables $\boldsymbol{\theta}$ are drawn from an EBF, $\mathcal{T}(\boldsymbol{\theta})$. Further, we postulate that the classification mechanism encoded into the distribution of the random variables y_n cannot be uniquely described by a single latent function $f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$, but $f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ is only an instance of the (possibly infinite) set of possible latent functions $f_{\boldsymbol{\theta}_c}(\boldsymbol{x}_n)$, $c = 1, \ldots, \infty$, parameterized from different EBFs, $\boldsymbol{\theta}_c \sim \mathcal{T}_c(\boldsymbol{\theta})$. Then, to determine the association between observations, $\boldsymbol{x}_n$, and latent functions, $f_{\boldsymbol{\theta}}(\cdot)$, we impose a PY process prior over this set of functions. The power-law nature of the PY process prior distribution allows for effectively handling cases of heavy-tailed observable data, which are prevalent in VAT audit case selection processes, as discussed previously.

Let us introduce the set of variables $\{z_{nc}\}_{n,c=1}^{N,\infty}$, with $z_{nc} = 1$ if the function modeling the correlation pattern between the observation $\boldsymbol{x}_n$ and the corresponding classification decision y_n is captured by the cth inferred (component) EBF, otherwise $z_{nc} = 0$. Based on this assumption, and the descriptions of the EBF model as well as the PY process prior, the prior configuration of the proposed PYP-EBF model is defined as follows:

$$p\big(y_n|\boldsymbol{x}_n, z_{nc}=1\big) = \text{Bernoulli}(y_n|f_{\boldsymbol{\theta}_c}(\boldsymbol{x}_n)) \tag{9}$$

$$p(z_{nc}=1|\boldsymbol{v}) = \varpi_c(\boldsymbol{v}) \tag{10}$$

$$\varpi_c(\boldsymbol{v}) = v_c \prod_{j=1}^{c-1}(1-v_j) \in [0,1] \tag{11}$$

with

$$\sum_{c=1}^{\infty} \varpi_c(\boldsymbol{v}) = 1 \tag{12}$$

$$p(v_c) = \text{Beta}(1-\delta, \alpha+\delta c) \tag{13}$$

and the cth inferred EBF is:

$$p(\boldsymbol{\theta}_c) = \mathcal{T}_c(\boldsymbol{\theta}) \tag{14}$$

while the functional form of the decision probability $f_{\boldsymbol{\theta}_c}(\boldsymbol{x}_n)$ is the mean of the probabilities pertaining to the $y=1$ class over the trees consisting the cth EBF (as encoded into the inferred vector $\boldsymbol{\theta}_c$).

3.1 Inference algorithm

Inference for nonparametric models can be conducted under a Bayesian setting, typically by means of variational Bayes (e.g., [6]), or Monte Carlo techniques (e.g., [22]). Here, we prefer a variational Bayesian approach, due to its considerably better scalability in terms of computational costs. Our variational Bayesian inference algorithm for the PYP-EBF model comprises derivation of a family of variational posterior distributions $q(.)$ which approximate the true posterior distribution over the infinite sets $Z, \boldsymbol{v} = (v_c)_{c=1}^{\infty}$ and $\{\boldsymbol{\theta}_c\}_{c=1}^{\infty}$, and the innovation parameter α. Apparently, Bayesian inference is not tractable under this setting, since we are dealing with an infinite number of parameters.

For this reason, we employ a common strategy in the literature of Bayesian nonparametrics, formulated on the basis of a truncated stick-breaking representation of the PY process [6]. That is, we fix a value C and we let the variational posterior over the v_i have the property $q(v_C = 1) = 1$. In other words, we set $\varpi_c(\boldsymbol{v})$ equal to zero for $c > C$. Note that, under this setting, the treated PYP-EBF model involves a full PY process prior; truncation is not imposed on the model itself, but only on the variational distribution to allow for tractable inference procedure. Hence, the truncation level C is a variational parameter which can be freely set, and not part of the prior model specification.

Let $W \triangleq \{\boldsymbol{v}, \alpha, Z, \{\boldsymbol{\theta}_c\}_{c=1}^{C}\}$ be the set of all the parameters of the PYP-EBF model the (posterior) distributions of which we need to train w.r.t. the available dataset $\mathcal{D}$. Variational Bayesian inference introduces an arbitrary distribution $q(W)$ to approximate the actual posterior $p(W|X,Y)$ which is computationally intractable [3]. Under this assumption, the log marginal likelihood (log evidence), $\log p(X,Y)$ becomes [15]

$$\log p(X,Y) = \mathcal{L}(q) + \text{KL}(q||p) \tag{15}$$

where

$$\mathcal{L}(q) = \int \text{d}W q(W) \log \frac{p(X,Y,W)}{q(W)} \tag{16}$$

[2] https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF

and KL$(q||p)$ stands for the Kullback-Leibler (KL) divergence between the (approximate) variational posterior, $q(W)$, and the actual posterior, $p(W|X,Y)$. Since KL divergence is nonnegative, $\mathcal{L}(q)$ forms a strict lower bound of the log evidence, and would become exact if $q(W) = p(W|X,Y)$. Hence, by maximizing this lower bound $\mathcal{L}(q)$ (evidence lower bound, ELBO) so that it becomes as tight as possible, not only do we minimize the KL-divergence between the true and the variational posterior, but we also implicitly integrate out the unknowns W. For simplicity, we consider that the posterior $q(W)$ factorizes over each one of the parameters, similar to the imposed prior (mean-field assumption [8]). By construction, this iterative, consecutive updating of the variational posterior distribution is guaranteed to monotonically and maximally increase the ELBO $\mathcal{L}(q)$ [9].

Let us denote as $\langle . \rangle$ the posterior expectation of a quantity. Based on the previous discussion, ELBO maximization yields

$$q(v_c) = \text{Beta}(v_c|\beta_{c,1}, \beta_{c,2}) \tag{17}$$

where

$$\beta_{c,1} = 1 - \delta + \sum_{n=1}^{N} q(z_{nc} = 1) \tag{18}$$

$$\beta_{c,2} = \alpha + c\delta + \sum_{c'=c+1}^{C} \sum_{n=1}^{N} q(z_{nc'} = 1) \tag{19}$$

Similarly, regarding the posteriors over the latent variables Z that assign each data point to the inferred EBFs, we have

$$q(z_{nc} = 1|\boldsymbol{x}_n) \propto \exp\left(\langle \log \varpi_c(\boldsymbol{v}) \rangle\right) f_{\boldsymbol{\theta}_c}(\boldsymbol{x}_n) \tag{20}$$

where

$$\langle \log \varpi_c(\boldsymbol{v}) \rangle = \sum_{c'=1}^{c-1} \langle \log(1 - v_{c'}) \rangle + \langle \log v_c \rangle \tag{21}$$

with

$$\langle \log v_c \rangle = \psi(\beta_{c,1}) - \psi(\beta_{c,1} + \beta_{c,2}) \tag{22}$$

$$\langle \log(1 - v_c) \rangle = \psi(\beta_{c,2}) - \psi(\beta_{c,1} + \beta_{c,2}) \tag{23}$$

At this point, we emphasize that the mixture model is **driven by the observations space**, $\mathcal{X}$, as it also becomes apparent from the resulting posterior (20).

On the other hand, the sampled EBFs, $\boldsymbol{\theta}_c$ are inferred by resorting to the standard CART algorithm, as employed in the case of a single trained EBF [25], but presented with a subset of the available training dataset, $\mathcal{D}$. This subset is obtained by sampling from the posterior $q(\boldsymbol{z}_n)$ of each training example, and collecting the set of the data points, $\boldsymbol{x}_n$, with $z_{nc} = 1$.

The estimates of the posteriors prescribed above are updated consecutively and in an iterative fashion until convergence of the model ELBO. That is, on each training algorithm iteration, we update the expressions of the variational posteriors, resample assignments from the posteriors $q(z_{nc} = 1)$, and rerun the CART algorithm to obtain samples $\boldsymbol{\theta}_c$ from their corresponding posteriors. This concludes the derivation of the inference algorithm of our PYP-EBF model.

3.2 Prediction Generation

After training the proposed PYP-EBF model on a dataset pertaining to tax audits and their outcomes, $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, we end up with a set of inferred EBF's with trees encoded into the vectors $\{\boldsymbol{\theta}_c\}_{c=1}^{C}$. These can be used to generate predictions for unseen data $\boldsymbol{x}_*$. In the context of our addressed problem of tax audit case selection, this corresponds to deciding whether a tax payer with VAT records summarized into the vector $\boldsymbol{x}_*$ may result in an audit yield exceeding the set threshold.

To this end, we employ a maximum a posteriori (MAP) rationale. Specifically, we first compute the posterior probability of the test data point , $\boldsymbol{x}_*$, being assigned to the trained component EBF's. On this basis, we determine the *winner component* that maximizes $q(z_{*c})$. Then, we perform the classification task using the output probabilities of the winner EBF. We emphasize that this is in contrast to the more typically used mean-field approach, under which one would compute the average of the probabilities obtained from the inferred EBF's, weighted by the corresponding posteriors $q(z_{*c})$. However, we adopt this MAP approach as we have found it to perform consistently better.

4 EXPERIMENTAL EVALUATION

4.1 Dataset Collection

To evaluate our approach, we have managed to get access to an extensive real-world dataset of Cyprus tax authority. Specifically, we use a dataset comprising over 10,000 VAT returns audited in the last six years. The generated label information is set to 1 if the tax audit generated a yield which exceeded a set threshold. Following the instructions of the collaborating tax authority, we consider four *alternative thresholds*: (i) The yield value that the tax authority currently considers barely worth the required resources for audit *(Base scenario)*; (ii) this amount increased by 16%; (iii) the base amount increased by 32% and (iv) increased by 48%.[3]

4.2 Experimental Setup

The proposed approach was implemented in Python, using the scikit-learn library [20]. To allow for some comparative results, apart from our method we also evaluate the following competitors: (i) a baseline RF model [16]; (ii) the EBF algorithm that our novel PYP-EBF approach is inspired from [25]; and (iii) a popular kernel-based approach that relies on similarity criteria selected in an ad-hoc manner, namely the label propagation (LP) algorithm [2]. All the evaluated RF-type algorithms, i.e. PYP-EBF, EBF, and RF, comprised 500 samples (trees). The truncation threshold, C, of our approach is set to $C = 10$. This selection is reasonable, since we do not expect more than 10 distinct binary classification patterns in the limited available dataset. In all cases, the criterion used for retaining a proposed split in a sampled tree is the Gini impurity index, as suggested in [25]. The LP algorithm is evaluated using the RBF kernel, which is the default selection in the scikit-learn library. All the developed models are run on a Desktop PC, and do *not* require any specialized hardware, e.g. graphical processing units (GPUs).

[3]Note that, since actual VAT returns and VAT audit results are used, we are restricted from disclosure of the actual threshold values, as they constitute privileged information.

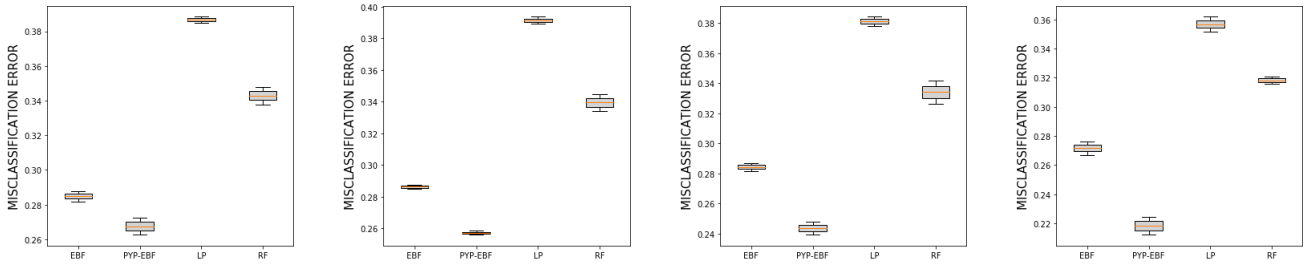

Figure 2: Left to right: Misclassification Errors: Base Scenario + 48%, 32%, 16%, 0%, respectively.

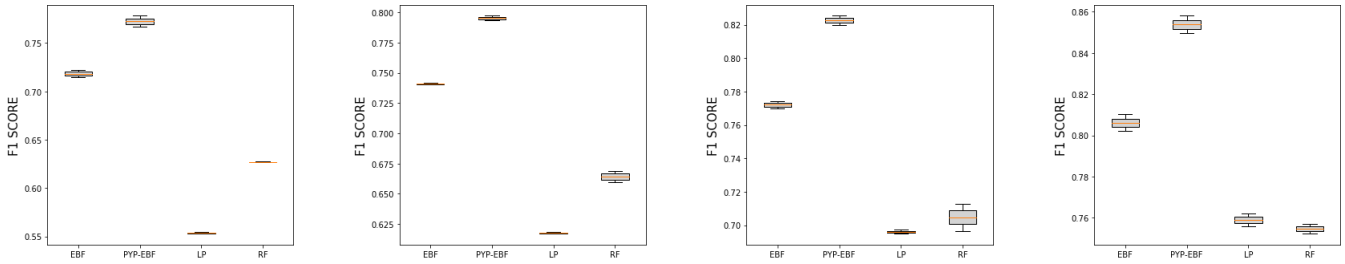

Figure 3: Left to right: F1 Scores: Base Scenario + 48%, 32%, 16%, 0%, respectively.

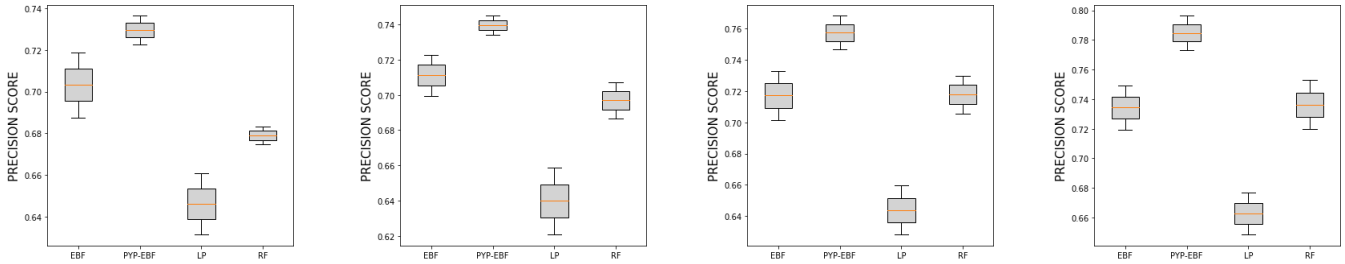

Figure 4: Left to right: Precision Scores: Base Scenario + 48%, 32%, 16%, 0%, respectively.

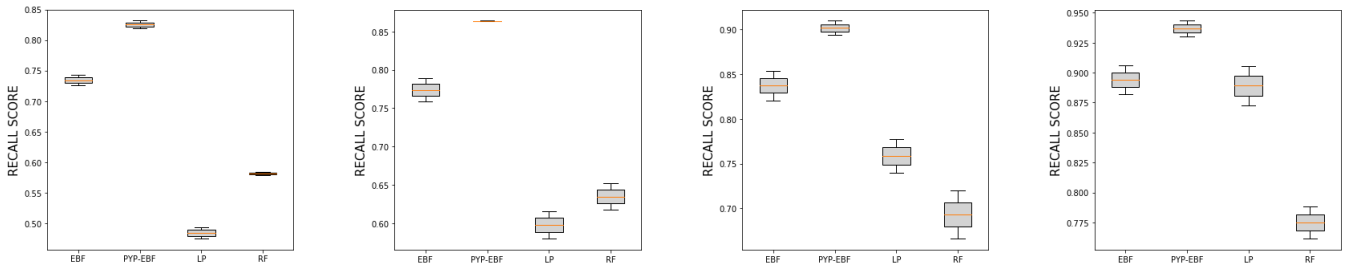

Figure 5: Left to right: Recall Scores: Base Scenario + 48%, 32%, 16%, 0%, respectively.

4.3 Results

Our quantitative evaluation is performed on an out-of-sample basis, that is on test data different from the training set. To this end, we perform 4-fold stratified cross-validation. In Fig.2 to Fig.5, we concisely illustrate the obtained performance of the evaluated algorithms. Specifically, we summarize the misclassification error, precision score, recall score, and F1 metrics obtained over the conducted four folds of cross-validation in the form of box-plots. We provide this illustration across all the four considered experimental scenarios (alternative tax yield thresholds). As we observe, our approach yields a significant performance improvement over the competition, which is consistent across all the employed evaluation metrics. Even more importantly, the obtained performance

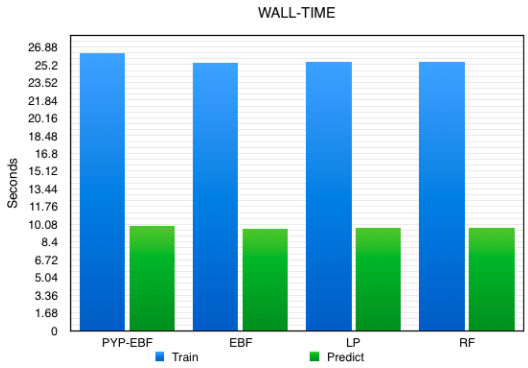

Figure 6: Evaluated Methods: Wall-Times of Model Training and Prediction Generation

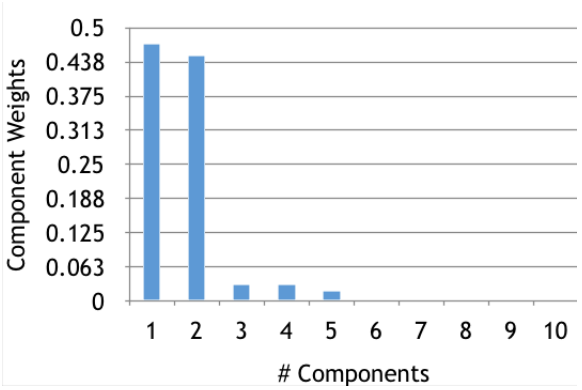

Figure 7: Component weight posterior expectations, $\langle \varpi_c(v) \rangle$, of the fitted PYP-EBF model.

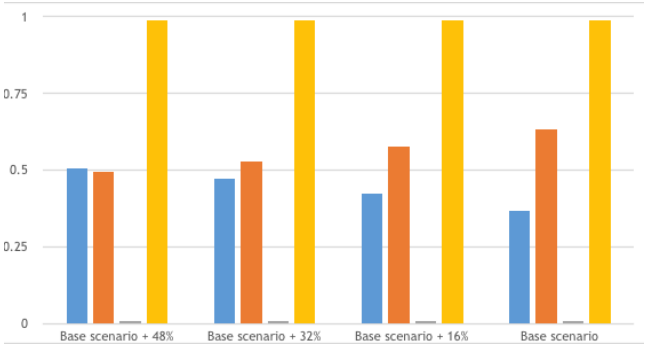

Figure 8: Deep Network: Confusion Plots (Blue-True Negatives %, Orange-False Negatives %, Grey- False Positives %, Yellow- True Positives %)

appears to be robust to an increase in the adopted audit threshold value. These outcomes provide overwhelming empirical evidence that our method offers a significantly more reliable outcome than the state-of-the-art in the field, thus better addressing the need of tax authorities to maximize the returns from the audits they can perform with their limited available resources.

Further, in Fig. 6 we demonstrate the computational times required for model training and testing, both in the case of our approach and the considered competitors. As we observe, the training time of PYP-EBF is increased over the alternatives, but only moderately so. This was expected, since our proposed approach entails fitting more parameters; this normally induces some computational overhead. On the other hand, the time required for generating predictions on our test set (which comprised almost 7,500 cases) exhibits only a barely notable increase over the competition. This finding vouches for the viability of our solution, which allows for a significant improvement in the quality of the audit case selection process, without compromising computational tractability. This is important for tax authorities, which need rapid development and response times, and cannot easily invest in high-performance computing facilities.

4.4 An Insight on the Power-Law Behavior

Further, we needed to examine how many mixture components remain effective after model training, and whether the fitted PYP-EBF model does actually yield a heavy-tailed distribution over the inferred components. To this end, in Fig. 7 we plot the component weight posterior expectations, $\langle \varpi_c(v) \rangle$, of the fitted PYP-EBF model, where we employed a truncation threshold $C = 10$. As we observe, our model yields two dominant components, and another three components with much lower weights. The remainder half of the initially postulated components effectively remain empty. This is an important outcome, as it corroborates both the usefulness of the power-law property of our model, as well as its capacity to infer how many components it actually needs, irrespectively of how big the truncation threshold is.

4.5 Comparison to deep learning

Finally, we examine the efficacy of a deep learning algorithm in the same setting. This is an important investigation, as deep learning solutions increasingly dominate the machine learning landscape. Specifically, we use the available data to train a conventional dense-layer deep network. Since the observations are 47-dimensional, the trained deep network comprises two dense hidden layers with 40 and 40 state-of-the-art ReLU units, respectively, regularized via the prominent method of Dropout [24] and trained via AdaGrad. We implemented this network in Tensorflow [17]. To obtain a statistically significant evaluation outcome, we perform 4-fold stratified cross-validation to select the hyperparameters (e.g., learning rate, batch size, etc.), as previously.

As we illustrate in Fig. 8 (confusion plot) the obtained performance is clearly disappointing. This lackluster outcomes apparently originates from the network completely failing to detect the negative class; indeed, network performance in the negative class is almost random in all scenarios. This corroborates our intuition that in the absence of large corpora of audited taxpayers, state-of-the-art techniques from the field of deep learning fail to yield any meaningful predictive outcome, as they become clearly unable to train. Note also that, in order to train this deep network, we needed specialised hardware, specifically a GPU, in contrast to the proposed approach.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we attacked a problem of major significance to European tax authorities, namely the effort to reduce the VAT-gap[12]. In this context, we devised a novel solution to the selection of taxpayers the potential audit of which may generate satisfactory yield. To this end, we relied and extended upon an algorithmic framework currently popular among tax authorities. i.e. RFs.

Specifically, our work was based on EBFs, which offer a principled (Bayesian inference) framework for obtaining a posterior distribution to draw the forest parameters (trees) from. Our method aimed to resolve a significant issue of EBFs, namely the fact that they are not designed for handling data that entail non-stationary nature, multiple modalities, or discontinuities; these are prevalent characteristics of tax return data, as is the case with other real-world applications as well. Our novel approach consists in postulating a (possibly) infinite mixture of EBFs, with each one meant to capture a single classification pattern that dominates a fraction of the observations space. To allow for the case of a power-law distribution of the available data over these patterns, we postulated a PY process prior over the data point assignment to component EBFs. We devised an efficient approximate inference algorithm for our model, based on the variational inference paradigm.

We performed a thorough experimental evaluation using a real-world dataset of filed VAT returns and corresponding audit outcomes, obtained from a EU tax authority. Our comparative results, considering both currently used RF-type approaches, as well as a deep learning baseline popular among machine learning practitioners, were extremely supporting of our solution. Specifically, PYP-EBF completely outperformed the alternatives, by a large margin, with respect to the obtained missclassification error, precision, recall, and F1 scores, over four different evaluation scenarios. This offers conspicuous empirical evidence supporting the efficacy and the usefulness of our approach. We also underline that these gains come for negligible computational overheads.

Currently, we aim to further and deepen our research work by examining how our methods can be leveraged to address other sources of tax evasion. Indeed, it is common practice for tax administrations to cross-validate and reconcile items declared in the tax returns of corporation tax and VAT, like revenue; a taxpayer who filed substantially different revenue amounts should expect an enquiry from the tax authorities. Therefore, taxpayers who under-declare revenue in their VAT returns are also expected to under-declare revenue for direct taxation purposes, and vice versa, so as to avoid attracting the scrutiny of tax authorities. Since VAT evasion and direct tax evasion are correlated, a model that combines raw data from both VAT returns and direct tax returns and performs joint audit case selection for both should yield higher accuracy compared to models addressing VAT and direct taxes separately. This remains to be confirmed in the context of our future research endeavors.

6 ACKNOWLEDGEMENTS

This work was supported by Cyprus Tax Department. We thank the Commissioner of the Department, Mr. Yiannis Tsangaris, for his invaluable support and investment in the pursued state-of-the-art technology.

REFERENCES

[1] C. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2, 6 (1974), 1152–1174.
[2] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, and Département D'informatique Et Recherche Opérationnelle. 2005. Efficient nonparametric function induction in semi-supervised learning. In *In AISTAT*. 96–103.
[3] C. M. Bishop. 1995. *Neural Networks for Pattern Recognition.* Clarendon Press.
[4] D. Blackwell and J. MacQueen. 1973. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* 1, 2 (1973), 353–355.
[5] D. Blei and M. Jordan. 2004. Variational methods for the Dirichlet process. In *21st Int. Conf. Machine Learning*. New York, NY, USA, 12–19.
[6] David M. Blei and Michael I. Jordan. 2006. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis* 1, 1 (2006), 121–144.
[7] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933404324
[8] D. Chandler. 1987. *Introduction to Modern Statistical Mechanics.* Oxford University Press, New York.
[9] S. Chatzis, D. Kosmopoulos, and T. Varvarigou. 2008. Signal modeling and classification using a robust latent space model based on t distributions. *IEEE Trans. Signal Processing* 56, 3 (March 2008).
[10] D. Cleary. 2011. Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit. *Electronic Journal of e-Government* 9, 2 (2011), 132–140.
[11] Daniel de Roux, Boris Pérez, Andrés Moreno, Maria del Pilar Villamil, and César Figueroa. 2018. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proc. ACM KDD*. ACM, New York, NY, USA, 215–222.
[12] European Commission Directorate-General for Taxation and Customs Union. 2018. Study and Reports on the VAT-gap in the EU-28 Member States:2018 FinalReport. *TAXUD/2015/CC/131* (2018).
[13] European Commission Directorate-General for Taxation and Customs Union. 2018. Taxation Trends.
[14] T. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1 (1973), 209–230.
[15] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. 1998. An introduction to variational methods for graphical models. In *Learning in Graphical Models*.
[16] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, and Hao Fu. 2015. Financial Fraud Detection Model: Based on Random Forest. *International Journal of Economics and Finance* 7 (2015), 178–188.
[17] Abadi Martın et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
[18] P. Muller and F. Quintana. 2004. Nonparametric Bayesian data analysis. *Statist. Sci.* 19, 1 (2004), 95–110.
[19] R. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* 9 (2000), 249–265.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[21] J. Pitman and M Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. In *Annals of Probability*, Vol. 25. 855–900.
[22] Yuting Qi, John William Paisley, and Lawrence Carin. 2007. Music Analysis Using Hidden Markov Mixture Models. *IEEE Transactions on Signal Processing* 55, 11 (2007), 5209–5224.
[23] J. Sethuraman. 1994. A constructive definition of the Dirichlet prior. *Statistica Sinica* 2 (1994), 639–650.
[24] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR* (2014).
[25] Matt Taddy, Chun-Sheng Chen, and Jun Yun. 2015. Bayesian and empirical Bayesian forests. (02 2015).
[26] Y. W. Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. Association for Computational Linguistics*. 985–992.
[27] Dan Vesset, Chandana Gopal, Carl W. Olofson, Stewart Bond, Maureen Fleming, and David Schubmehl. 2017. Worldwide Big Data and Analytics Software 2017 Market Shares: Healthy Growth Across the Board. *IDC's Worldwide Big Data and Analytics Software Taxonomy* US42353216 (2017).
[28] S. Walker, P. Damien, P. Laud, and A. Smith. 1999. Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Statist. Soc. B* 61, 3 (1999), 485–527.
[29] Roung-Shiunn Wu, Chin ou, Hui-ying Lin, She-I Chang, and David Yen. 2012. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications* 39, 10 (2012), 8769 – 8777.