Shapley value for tax audit data valuation

Michiel Van Roy[1], David Martens[1], Ann Jorissen[1], Anne Van de Vijver[1]

[1]University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium

Abstract: Tax authorities have access to unprecedented levels of information on taxpayers. However, these vast amounts of data lead to new challenges, with tax authorities running the risk of being overwhelmed by the enormous amount of incoming data. We present a valuation technique to quantify the value of features for tax audit prediction. Our approach, rooted in the game-theoretical Shapley value, effectively assigns importance to features derived from various Directives on Administrative Cooperation within the European Union and the OECD's automatic exchange of information. We show that our results can be used for global explanations of the predictive model, feature selection and determining which data should be acquired or cleaned with priority, similar to active feature acquisition. Our results can assist tax authorities in managing the large amounts of data they receive under the different disclosure regulations.

1 Introduction

Over the last decades, many different regulations and agreements concerning the exchange of information for tax purposes have emerged. The exchange of information has been increasingly important in the fight against tax avoidance and evasion ever since its introduction in the OECD Model Tax Convention from 1963. Milestone initiatives like the Global Forum on Transparency and Exchange of Information for Tax Purposes in 2000 promoted the exchange of information on request and the automated exchange of information. In light of this, the EU adopted Directive 2011/16/EU of 15 February 2011, which established the legal basis for administrative cooperation in the field of direct taxation in the EU. The scope of the original directive has been expanded multiple times with new types of data, with the aim of strengthening the administrative cooperation among tax authorities of Member States and is commonly known as the Directives on Administrative Cooperation (DACs).

Agreements on the exchange of information go well beyond the EU directive however. Many countries support bilateral agreements with one another to establish similar information flows. The OECD also supports exchange of information between tax authorities, with the automatic exchange of information as one of the main pillars. As a direct consequence however, tax authorities have witnessed an explosion in the amount information they receive from taxpayers. Tax authorities do not only receive data from taxpayers in their own jurisdiction, but they also receive related information from other jurisdictions under the exchange of information schedules.

The growth in information poses a challenge for tax authorities, who run the risk of being overwhelmed by the enormous quantities of information received. Indeed, a key finding from the European Commission in report COM/2017/0781 is that tax authorities' capacity to handle data has not increased at the same rate as the amount of data they receive. This could imply that tax authorities become less effective in exploiting and verifying this information, causing costs for tax authorities and taxpayers who are confronted with less efficient tax authorities.

In this regard, determining the value of data to predict tax audits could help tax authorities identifying relevant information, and allows them to make informed decisions on which data sources they should invest in. In this study, we quantify the value of several types of data received under the automated exchange of information for tax audits. Results can be used for a variety of purposes including global model explanations, determining which data should be acquired or quality-checked with priority, determining which data should get priority in the data cleaning process or should be disseminated to other departments of the tax authorities, and feature selection depending on the goal of the modeler. To the best of our knowledge, we are the first study to attribute value to data tax authorities receive under the automatic exchange of information regulations of the EU and OECD. This paper describes the result of a collaboration between the University of Antwerp and the department "Large enterprises" of the Federal Tax Authorities of Belgium.

2 Related work

The data mining literature on tax audits and value attribution is relatively scarce. We therefore also refer to research from the field of tax fraud, and other fields unrelated to tax but following similar approaches to ours. The notion of value is central in our research. We define value as the impact a feature has on model performance. "Impact" should be understood broadly, both in terms of magnitude of predictive power and presence of the feature in the dataset. It could namely be that a feature exhibits very high predictive power, but appears only a handful of times in the entire dataset. Consider the following fictitious example: a niche industry only appears three times in the entire dataset, yet membership to this industry is very indicative for the need of a tax audit. This feature would contain high predictive power. Because the feature appears only three times in the dataset, it can rarely be used to classify an instance in the right category however. A good valuation method should consider both predictive power and appearance, as it is desirable that a feature with high predictive power and high occurrence should be marked as more valuable than a feature with equal predictive power but lower occurrence (Moeyersoms, d'Alessandro, Provost, & Martens, 2017).

Features are thus evaluated on their impact on performance metrics of a predictive model. This impact can be positive as well as negative. The general use of value attribution methods in machine learning is widespread. Feature selection, data valuation and providing explanations for black-box models all rely on some form of valuation. In the explanations of black-box model, a distinction can be made between global and local importance methods. In this research, we will focus on the former.

2.1 Global importance

The relevance of global explanations originates from the need to trust model predictions and the need to gain insight in the problem domain. Knowledge extracted from data is often only useful when people understand and trust the model applied (Van Assche & Blockeel, 2007). The goal of global explanation methods is thus to provide users of predictive models with an overview of the predictive power of individual features. These methods can help a user to understand the role of the features across the entire model and dataset. Features can thus be ranked on their importance and contribution towards the model performance. Global explanations convey information about the features' general impact in the model, but do not explain why an individual instance received a certain class label like local importance methods do. Several studies in the tax fraud domain report classification rules and feature importance rankings for fraud detection (e.g. Basta et al., 2009; González & Velásquez, 2013; Gupta & Nagadevara, 2007). For example, González and Velásquez (2013) show by using global explanation methods on a neural network for false invoice detection that the main predictive power of the network stems from variables associated with the payment of VAT and to a lesser extent to income-related variables. Similarly, Vanhoeyveld, Martens, and Peeters (2020) calculate ratio's based on VAT declaration information and were able to rank these ratio's based on their predictive power for fraud detection for each company.

2.2 Feature selection

Related to global importance is feature selection. Feature selection considers the impact of features by attributing each feature an importance score so that low-impact features can be removed to make large-scale problems computationally efficient. Feature selection can also be used to improve classification accuracy, reduce computational burdens or to reduce the amount of training data needed to achieve a desired level of performance (Forman, 2003). Feature selection can thus be regarded as a way to value features according to their importance to the predictive model (Forman, 2003). In a tax setting, Hsu, Pathak, Srivastava, Tschida, and Bjorklund (2015) present a case study to examine the use of data mining techniques in audit selection for tax authorities. They use feature selection techniques to determine the predictive power of feature subsets, and subsequent consultations with tax domain experts were held to discuss whether low-scoring features should be kept. It is unclear however how the feature subsets were determined. Matos et al. (2020) develop a new feature selection algorithm specifically for a tax fraud detection context. After applying feature selection, they show significantly improved performance of fraud prediction algorithms.

2.3 Active information acquisition

Another setting in which determining the value of data can be useful, is active feature acquisition. The idea behind active feature acquisition is based on active learning, where the goal is to obtain new data that will improve model performance the most with a limited budget. Given that data acquisition can be costly, only acquiring the most valuable data will reduce the amount of resources needed to come to an accurate model. In the context of feature acquisition, modelers are usually confronted with missing features. The goal is to determine for which instances it is most interesting to acquire 'complete' feature information (Melville, Saar-Tsechansky, Provost, & Mooney, 2004; Saar-Tsechansky, Melville, & Provost, 2009; Zheng & Padmanabhan, 2002). Feature acquisition can be of importance for both model building as well as model usage (Provost, Melville, & Saar-Tsechansky, 2007). In the building phase, features can be acquired in an effort to improve model performance in general. For example, acquiring new feature information on instances that are misclassified can allow the model to learn new patterns to avoid such misclassification in the future (Melville et al., 2004; Saar-Tsechansky et al., 2009). In the model usage phase, active feature selection techniques have been successfully implemented to determine which features of a test case should be acquired, and in which order to minimize the cumulative cost of misclassifications and acquisition costs (Sheng & Ling, 2006). Similarly, Ghorbani and Zou (2019) develop an algorithm to estimate the value of datapoints, which can be used to determine which datapoints should be acquired with priority.

3 Data and methods

The Belgian tax authorities provided us with a unique and fully anonymized dataset containing reports which certain large companies must provide under two different reporting regulations. The first data source is the local file and forms part of transfer pricing documentation. The local file contains detailed information relating to specific material intercompany transactions.[1] The second data source is the information received under BEPS 5 and directive 201/2376/EU, better known DAC-3. Under these regulations, tax rulings and advanced pricing agreements are exchanged between different tax authorities. As a simplification for the sake of readability, we refer to the data received under both DAC-3 and BEPS 5 as ETR (Exchange of Tax Rulings). The third data source is the information received under directive 2018/822/EU, better known DAC-6. Under this directive, intermediaries and/or

[1] Art. 321/5, § 4 Belgian Income Tax Code and Royal Decree of 28.10.2016

taxpayers are required to report certain cross-border arrangements when they satisfy certain characteristics. Even though the data itself is private, the contents of all types of reports are public knowledge as XML-schemes to file these reports are available on the website of the Belgian Tax authorities. We will therefore briefly discuss the content, shape and characteristics of this data.

Table 1 Fictitious example of the different types of data occurring in this study for the Local file

Identity	Continuous	Discrete (<100 categories)					Discrete(>100 categories)	Label
Taxpayer pseudo-ID	Turnover cross border goods	Country cross border goods 1	Country cross border goods 2	Country cross border goods x	Activity	Transfer pricing method	Industry (NACEBEL code)	
ID$_1$	9,876,543	BE	CZ	NL	Limited risk distributor	CUP	64200	1
ID$_2$	1,000	BE	/	/	Fully Fledged	TNMM	46699	0
ID$_3$	/	/	/	/	Contract distributor	Cost plus	70100	1
⋮	⋮	⋮	⋮		⋮	⋮		⋮
ID	65,812	BE	IE	CA	Fully fledged	TNMM	14140	0

For each taxpayer in an Local file report, the data is characterized by the type of cross-border transaction, as well as several discrete attributes (e.g. the country of the relevant taxpayer and the country or countries involved parties, the transfer pricing method used, the activity of the taxpayer, …). For a subset of taxpayers, we know the label as they were audited by the tax authorities.

3.1 Local file

The local file is part of transfer pricing documentation that companies who are part of multinational enterprises must provide to tax authorities. Transfer pricing can roughly be understood as the prices divisions within a (multinational) company charge when selling goods or services to other divisions of the same company. It is easy to see that, when no restrictions on the pricing are imposed, multinational companies can let divisions in high-tax jurisdictions charge high prices to divisions in low-tax jurisdictions, reducing the profits in the high tax jurisdiction and increasing profits in the low-tax jurisdiction. The multinational would gain an advantage by letting the profits be taxed at lower rates, increasing profits after tax. To constrain such practices, transfer prices must adhere to the arm's length principle laid down in article 9 of the OECD model convention. The arm's length principle roughly states that the transfer prices should be set as if the divisions involved are not part of the same company, but rather are independent parties. The goal of the local file is thus to identify and report relevant related party transactions, the amounts involved in those transactions and the transfer pricing determinations made by the taxpayer with regard to those transactions for each country (OECD, 2014). The local file is used to ensure taxpayers' compliance with the arm's length principle in its material transfer pricing positions within a specific jurisdiction (OECD, 2014).

Table 2 Fictitious example of the different types of data occurring in this study for ETR

Identity	Continuous	Discrete (<100 categories)					Label
Taxpayer pseudo-ID	Transaction amount	Ruling type	Taxpayer country	Affected entity country 1	Affected entity country 2	Affected entity country x	
ID_1	1,234,567	ETR602	BE	BE	NL	…	1
ID_2	9,876	ETR601	GB	BE	/	…	0
ID_3	/	ETR606	CZ	AU	BE	…	1
⋮	⋮	⋮	⋮	⋮	⋮		⋮
ID	65,812	ETR609	IE	GB	IT	BE	0

For each taxpayer in an ETR report, the data is characterized by the amount of the ruling (if applicable), as well as several discrete attributes (e.g. the ruling type, the country of the relevant taxpayer and the country or countries of the affected entities). For a subset of taxpayers, we know the label as they were audited by the tax authorities.

The local file describes the management and shareholder structure of the local entity in the jurisdiction, as well as the activities and most principal competitors of the local entity. Additionally, summaries of the material controlled transactions (e.g. procurement of manufacturing services, purchase of goods, provision of services, loans, …) and the context in which such transactions take place must be reported. The method used to determine the transfer price must also be communicated. All such transactions must be broken down by involved tax jurisdiction (OECD, 2014). The full requirements and details of the local file can be found in the Annex II to Chapter V of the Guidance on

Transfer Pricing Documentation and Country-by-Country Reporting (OECD, 2014). A fictitious example of selected data from the local file can be found in Table 1.

3.2 ETR

DAC 3 and BEPS 5, taken together as "ETR" concerns the exchange of advance cross-border tax rulings and advance pricing agreements. An advance cross-border tax ruling is a confirmation or assurance that tax authorities give to taxpayers on how their tax will be calculated in a cross-border situation. Importantly, such confirmation must be given before the action on which the taxpayer wants assurance takes place. Similarly, an advance pricing arrangement determines the appropriate set of criteria between group companies for the determination of transfer prices. The most important information that needs to be exchanged is a summary of the transactions, a start date and period of validity of the tax ruling and the identification of the other involved jurisdiction(s) or persons in the other jurisdiction(s), other than natural persons, likely to be affected by the tax ruling. Full information can be found in Directive 2015/2376/EU and BEPS Action 5. A fictitious example of selected data from the ETR can be found in Table 2.

3.3 DAC 6

DAC 6 originates from a call of the European Parliament for tougher measures against intermediaries, such as lawyers and accountants who assist in arrangements that may lead to tax avoidance and evasion[2]. Under DAC 6, intermediaries and taxpayers must report details of cross-border arrangements that contain at least one of the hallmarks set out in Annex IV of Directive 2018/822/EU. These hallmarks can be understood as certain characteristics of the arrangement that present an indication of a potential risk of tax avoidance. The goal is thus to gather and exchange information on arrangements made by taxpayers which have a high perceived risk of tax avoidance. The main information that needs to be reported are the amount of the arrangement, the hallmark(s) it satisfies and the parties involved in the arrangement. Full information can be found in Directive 2018/822/EU. A fictitious example of selected data from DAC 6 can be found in Table 3.

Finally, the Belgian tax authorities also provide us with a dataset on which tax audits lead to an amendment in the taxpayers' declarations. Note that this does not mean that the taxpayer committed fraud. Any amendment made subsequent to a tax audit is included in our data, and no distinction can be made between an error correction or fraud. In addition, we only have data on companies that have been audited. This could introduce a possible selection bias in the data. The results of the analyses must thus be interpreted with this in mind.

[2] See the preamble of Council Directive 2018/822/EU

Local file and ETR data is observed as of 2017 on, and DAC-6 data is observed as of 2018. Data on tax audits is observed for 2021 and 2022. We only use data available preceding a tax audit, as data received past the audit should naturally not be predictive for the audit itself. In this research, the goal is to quantify the predictive value of the features used in the predictive model for each data source. The features in our models consist of the data taxpayers have to provide the tax authorities with in the ETR, DAC 6 and local file regulations, and the instances are the taxpayers. The data consists of discrete features and continuous features, as well as high-cardinality features[3]. As the main focus of our research is the attribution of value, we assume that the predictive model is given for the different data sources and that this is the best possible model the data science team could find. We thus do not focus on model building itself.

Table 3 Fictitious example of the different types of data occurring in this study for DAC 6

Identity	Continuous	Discrete (<100 categories)			Label
Taxpayer pseudo-ID	Hallmark amount	Hallmark type	Associated taxpayer country	Relevant taxpayer country	
ID$_1$	191,523	DAC6C1c	US	CH	1
ID$_2$	52,002,540	DAC6A3	NL	GB	0
ID$_3$	0	DAC6E3	/	CZ	0
⋮	⋮	⋮	⋮	⋮	⋮
ID	2,415,091	DAC6C1bii	AE	MT	1

For each taxpayer in a DAC-6 report, the data is characterized by the amount of the arrangement, as well as several discrete attributes (e.g. the type of hallmark, the country of the relevant taxpayer and the country of the associated taxpayer if one is present in the arrangement). For a subset of taxpayers, we know the label as they were audited by the tax authorities.

3.4 Shapley value for value distribution

A popular way to attribute importance to features or datapoints is based on the Shapley value from cooperative game theory (Cohen, Dror, & Ruppin, 2007; Covert, Lundberg, & Lee, 2020). As we will describe in this section, the Shapley value exhibits many desirable properties in a valuation setting. Originally, the goal of the Shapley value was to distribute the value of a cooperative game to all players

[3] Following Moeyersoms and Martens (2015), we deem a feature to be of high-cardinality when it has more than 100 different categories.

of this game in a fair and unique manner (Shapley, 1953). Linking game theory to machine learning, many machine learning problems can be understood as such cooperative games: A set of features or datapoints cooperate in a learning algorithm to achieve a certain outcome. This outcome can be overall model performance or an individual prediction, depending on the goal of the modeler. The Shapley value can thus be applied to distribute the value of the outcome back to the features or datapoints responsible for this outcome. Interestingly, many distribution schemes based on the Shapley value are model-agnostic and can thus be applied to every type of model. To understand the Shapley value and its properties, we follow Moeyersoms et al. (2017) by defining some core concepts, before we present the definition of the Shapley value:

1. N is the complete set of players or grand coalition, with cardinality $||N|| = n$
2. S is a subset of players, with $||S|| = s$ and $S \subset N$
3. $v(S)$ is a value function that represents the total utility (= predictive power) the set S generates when playing the game
4. $v(S \cup \{i\}) - v(S)$ is the marginal utility of adding player i to a set S.

The definition of the Shapley value for player i is the following (Shapley, 1953):

$$\varphi_i = \sum_{S \subset N\ i \notin S} \frac{(n-s-1)!\,s!}{n!} \big(v(S \cup \{i\}) - v(S)\big), \qquad i = 1, \dots, n.$$

The Shapley value can thus be understood as the average marginal utility a player contributes to every possible subset of players of the grand coalition. As stated before, the Shapley value can be used to allocate the total worth of the coalition back to each individual player in a "fair" and unique manner. "Fair" is defined as the satisfaction of several axioms which are desirable in a valuation setting, three of which are necessary to come to a unique solution (Shapley, 1953). These are the symmetry property (1), the efficiency property (2) and the additivity property (3). Let's again consider the complete set of players N and a subset these players $S \subset N$. The Shapley value is denoted by φ:

1. Symmetry property: If $v(S \cup i) = v(S \cup j)$ for every $S \subset N$ then $\varphi(i) = \varphi(j)$. In other words, if the contribution of adding player i to a subset S is always the same as adding player j to the same subset S, then i and j should receive the same value. If $v(S \cup i) = v(S \cup j)$ for every $S \subset N$ then $\varphi_i(v) = \varphi_j(v)$
2. Efficiency property: The Shapley value represents a distribution of the total value of the game $\sum_{i \in N} \varphi_i = v(N)$
3. Additivity property: When two independent games are combined, the values must be added player by player $\varphi_i(v) + \varphi_i(w) = \varphi_i(v + w)$

A fourth useful property but one which is not required to come to a unique solution is the dummy principle:

4. Dummy property: A player who does not contribute to any coalition should get a score of zero. If $v(S \cup i) - v(S) = 0$ for every S ⊂ N then $\varphi_i = 0$

Due to these properties, the Shapley value tends to outperform importance scores based on a single element like leave-one-out scores in correctly valuing data (Cohen et al., 2007; Keinan, Sandbank, Hilgetag, Meilijson, & Ruppin, 2006). Additionally, since one examines the marginal contribution of a player to every possible subset of players, the Shapley value takes interactions between players into account when attributing importance scores. This gives the Shapley value another edge over single element-based scores (Strumbelj & Kononenko, 2010).

It can be mathematically proven that the Shapley value is the only value that contains all these properties (see Shapley (1953) for more details). While the Shapley value clearly offers very interesting properties for valuation purposes, the major drawback is its computational complexity. To calculate the Shapley value exactly, it is necessary to calculate 2^n possibilities, which leads to computationally intractable solutions quickly when the number of players in N increases. Therefore, approximation methods are proposed.

3.5 Monte Carlo sampling

One way to approximate the Shapley value is to obtain an unbiased estimate of its value through Monte Carlo sampling (Castro, Gómez, & Tejada, 2009). To see how this works, it is useful to rewrite the Shapley value as the sum of adding player i to every possible order O of magnitude n:

$$\sum_{O \in \pi(N)} \frac{1}{n!} \left(v\left(Pre^i(O) \cup \{i\}\right) - v\left(Pre^i(O)\right) \right), \qquad i = 1, \ldots, n.$$

Where π is the set of all possible orders O of n players. The Shapley value can thus be obtained by listing all possible orders the players can enter the coalition in, and calculating the marginal contribution of adding the player of interest to the preceding elements in those orders. This formulation of the Shapley value lends itself well for a sampling-based approach. One can sample from a uniform distribution of orders in which the players participate to the coalition, and calculate the marginal contributions of the players in those orders to obtain unbiased estimates of the Shapley value (Castro et al., 2009). This Monte Carlo sampling approach can further be optimized by bounding the sampling error based on the theoretical variance (Maleki, Tran-Thanh, Hines, Rahwan, & Rogers, 2013) and by stratified sampling based on the position of the player in the orders (Castro, Gómez, Molina, &

Tejada, 2017; Maleki et al., 2013). We follow both the original sampling algorithm (Castro et al., 2009)[4] as well as the two-step approach presented in Castro et al. (2017). In this two-step approach, samples are taken with each feature appearing on every ordinal position in a random order, i.e. a stratum. When two or more samples per stratum are taken, variances of the estimation in the stratum can be calculated. The second step takes additional samples of the stratum based on the variance of the stratum determined in the previous step. The larger the variance in a stratum, the more sampling will be done for that stratum in an effort to reduce the variance, and thus come to a more precise estimate of the Shapley value. The original sampling algorithm and the two-step sampling approach are represented in Algorithm 1 and Algorithm 2 respectively.

Monte Carlo estimations can be calculated in polynomial time, assuming that the marginal contribution can be calculated in polynomial time as well (Castro et al., 2009). While Monte Carlo sampling can greatly reduce the computational burden, some problems are still too large to be computed efficiently. One solution is parallel processing, as the sampling procedure can easily be done on multiple cpu's. In addition, logically grouping features into higher-level meta-features to reduce the dimensionality of the feature matrix could also provide a solution (Chen, Zhang, Zhang, & Duan, 2016; Ghorbani & Zou, 2019; Kim et al., 2018). When groups are chosen logically and/or based on domain knowledge, calculations become feasible without sacrificing the interpretability of the results. For example, dummies belonging to a single categorical variable can be grouped together and added to the model all at once, which leads to a valuation of the entire categorical variable instead of a single dummy. In settings where determining the value of each single dummy is not required or even undesirable, such an approach can be preferable .

Several other approaches to estimate Shapley values also exist, such Shapley Additive Global Importance (SAGE) (Covert et al., 2020). This method is computationally even more efficient than Monte Carlo sampling, however it rests on the assumptions that all features are independent. In practice, this assumption is often violated. Variations of this technique taking into account feature dependence can partially solve this issue, however could violate sensitivity (i.e. attributing value to features that should not have received value) (Molnar et al., 2022)

[4] Note in Algorithm 1 we implement the efficiency improvements made by Song, Nelson, and Staum (2016) to the original algorithm

Algorithm 1: ApproShapley (Castro et al., 2009)

```
Inputs:
m = desired sample size
$\varphi_i \coloneqq 0, \forall\, i \in N$
Tracker:=0
While Tracker<m:
    Take $O \in \pi(N)$ with probability $1/n!$
    $v\big(Pre^i(O)\big)$:= base performance classifier
    For all $i \in N$
        Calculate $v\big(Pre^i(O) \cup \{i\}\big) \coloneqq performance\ of\ model\ with\ features\ 0, \ldots i$
        Calculate $x(O)_i = v\big(Pre^i(O) \cup \{i\}\big) - v\big(Pre^i(O)\big)$
        $\varphi_i \coloneqq \varphi_i + x(O)_i$
        $v\big(Pre^i(O)\big) \coloneqq x(O)_i$
    Tracker := Tracker + 1
$\varphi_i := \frac{\varphi_i}{m}, \forall\, i \in N$
```

Algorithm 2: Two-step ApproShapley (Castro et al., 2017)

```
Inputs:
m = desired sample size
$\varphi_i \coloneqq 0, \forall\, i \in N$
$P_l^i = stratum\ of\ feature\ i\ appearing\ on\ place\ l$
For all $l = 1, \dots, n$ and $i = 1, \dots, n$
    $m_{il}^{exp} := \frac{m}{2n^2}$
    $Tracker\_l \coloneqq 0$
    $Sum\_quad\_l \coloneqq 0$
    While $Tracker\_l < m_{il}^{exp}$:
        Take $O \in P_l^i$ with probability $1/(n-1)!$
        Calculate $v\left(Pre^i(O)\right) \coloneqq performance\ of\ model\ with\ features\ 0, \dots i-1$
        Calculate $v\left(Pre^i(O) \cup \{i\}\right) \coloneqq performance\ of\ model\ with\ features\ 0, \dots i$
        Calculate $x(O)_i = v\left(Pre^i(O) \cup \{i\}\right) - v\left(Pre^i(O)\right)$
        $\varphi_l^i \coloneqq \varphi_l^i + x(O)_i$
        $Sum\_quad\_l \coloneqq Sum\_quad\_l + {x(O)_i}^2$
    $Tracker\_l \coloneqq Tracker\_l + 1$
$s_{il}^2 := \frac{1}{(m_{il}{}^{exp} - 1)}(Sum_{quad_l} - \frac{(\varphi_l^i)^2}{m_{il}{}^{exp}})$
Calculate $m_{il}^{st}$
For all $l = 1, \dots, n$ and $i = 1, \dots, n$
    $Tracker\_l \coloneqq 0$
    While $Tracker\_l < m_{il}^{st}$:
        Take $O \in P_l^i$ with probability $1/(n-1)!$
        Calculate $v\left(Pre^i(O)\right) \coloneqq performance\ of\ model\ with\ features\ 0, \dots i-1$
        Calculate $v\left(Pre^i(O) \cup \{i\}\right) \coloneqq performance\ of\ model\ with\ features\ 0, \dots i$
        Calculate $x(O)_i = v\left(Pre^i(O) \cup \{i\}\right) - v\left(Pre^i(O)\right)$
        $\varphi_l^i \coloneqq \varphi_l^i + x(O)_i$
    $Tracker\_l \coloneqq Tracker\_l + 1$
$\varphi_l^i \coloneqq \frac{\varphi_l^i}{m_{il}^{exp} + m_{il}^{st}}$
$\varphi_i^{st,opt} := \frac{1}{n}\sum_{l=1}^{n} \varphi_l^i$ for all $i = 1, \dots, n$
```

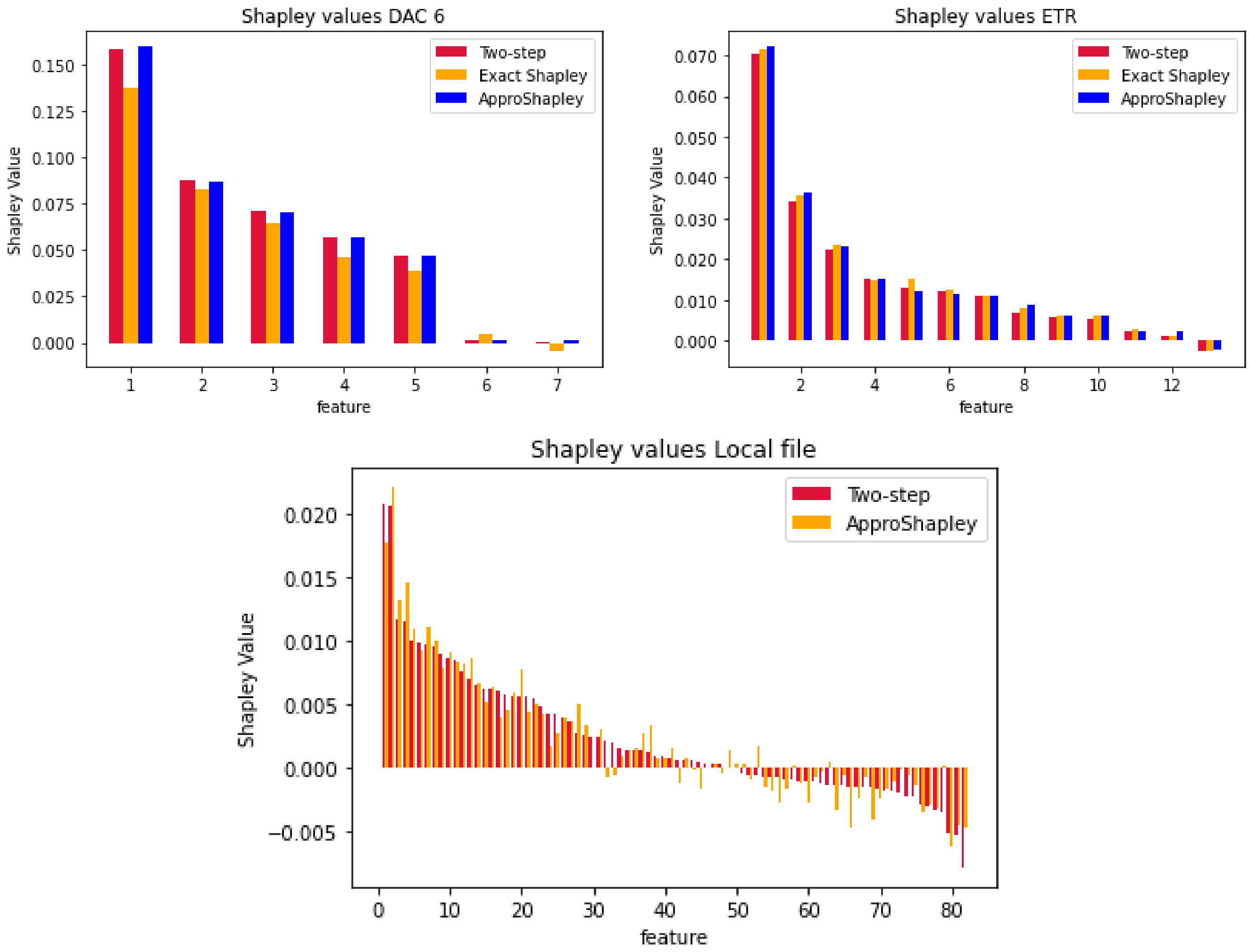
4 Experimental setup

By applying the techniques described in section 3, we aim to determine the predictive value of the different features in the ETR, DAC-6 and local file reports to predict tax audits. As stated before, we assume the predictive model is given. For the data on the local file, we train a random forest classifier. For the data on ETR, we use a non-linear support vector machine. Hyperparameters are tuned based on a held-out validation set. The minimum number of samples in a leaf for the random forest is 3, the regularization parameter C for the support vector machine is 32, the gamma parameter for the support vector machine is 1 and the chosen kernel is Radial Basis Function. Other hyperparameters are left on default values. For the data on DAC-6, we use a decision tree and a logistic model. We use a logistic model to compare the coefficients of the model to the Shapley values obtained by the estimation techniques. We use a 80%-20% train-test split for all datasets, and use 20% of the training set for validation purposes. To avoid obtaining results specific to the train-test split, we average our results over ten different train-test splits. We choose the Area-Under-Curve (AUC) as a performance metric, however any other performance metric can be used. We divide the difference between the observed AUC compared to the baseline value of 0.5 over the different features. For the ETR and DAC-6 reports, we can benchmark the approximations with the exact Shapley value since the dataset only consists of thirteen and seven features before splitting up the categorical variables respectively. The local file consists of several thousand features so we will only be able use approximation techniques. We are however able to group the features into logical meta-features based on the XML-categories and domain knowledge. This reduces the dimensionality substantially, and thus also the amount of samples needed to come to accurate estimations. Note that we still use the original features as input in the model for the estimation, but the Shapley value is calculated for the meta-feature (i.e. the group). The sampling approaches are based on 10,000 samples for ETR and DAC 6, and 40,000 for the local file.

5 Results

Results of the analyses are presented in Figure 1. For confidentiality reasons, we do not disclose the exact name of the features in this paper. However, we do provide this information to the tax authorities. In the case of DAC 6, estimation methods approximate the exact Shapley value very closely, with Two-step-ApproShapley generally approaching the true Shapley value better than the original sampling approach. For DAC 6, we see that one feature is the dominant contributor to the predictive value of the model, and that four other features also improve model performance substantially. Two other features receive a Shapley value of close to zero and thus contribute only little towards the predictive performance of the model. This can be explained by their low coverage. These two features do not have any value for most of the observations, which highlights the importance of taking coverage into account. For the ETR data, we see that the ApproShapley method tends to approximate the Shapley values better than the Two-step approach. One feature has a strong negative Shapley value, suggesting that this feature hurts the model performance. In the case of the local file, ten and nine features have negative Shapley values according to ApproShapley and Two-step-ApproShapley respectively, which means they hurt the performance of the model in general. Shapley values for the Local file tend to be smaller than those for DAC 6 and ETR, which makes sense given the larger number of features the local file contains. Nevertheless, the Shapley value allows us to rank these features based on their relative performance in the predictive model.

Figure 1 Shapley values per data source



To check whether the rankings are consistent between the different methods, we calculate Spearman rank coefficients. The Spearman coefficient indicates how well the relation between two variables can be described as monotonic. A perfect Spearman's value of 1 or -1 implies a perfect monotonic positive or negative relationship. Table 4 to Table 6 present the Spearman rank correlation coefficients between the different methods for all data sources. The strongly positive correlation coefficients between the Shapley methods mark that all methods come to very similar ranking of the features.

Table 4 Spearman rank coefficients local file

Local file data		
	ApproShapley	Two-step ApproShapley
ApproShapley	/	0.9512
Two-step ApproShapley	0.9512	/

Table 5 Spearman rank coefficients ETR

ETR data			
	Exact Shapley	ApproShapley	Two-step ApproShapley
Exact Shapley	/	0.9835	0.9890
ApproShapley	0.9835	/	0.9780
Two-step ApproShapley	0.9890	0.9780	/

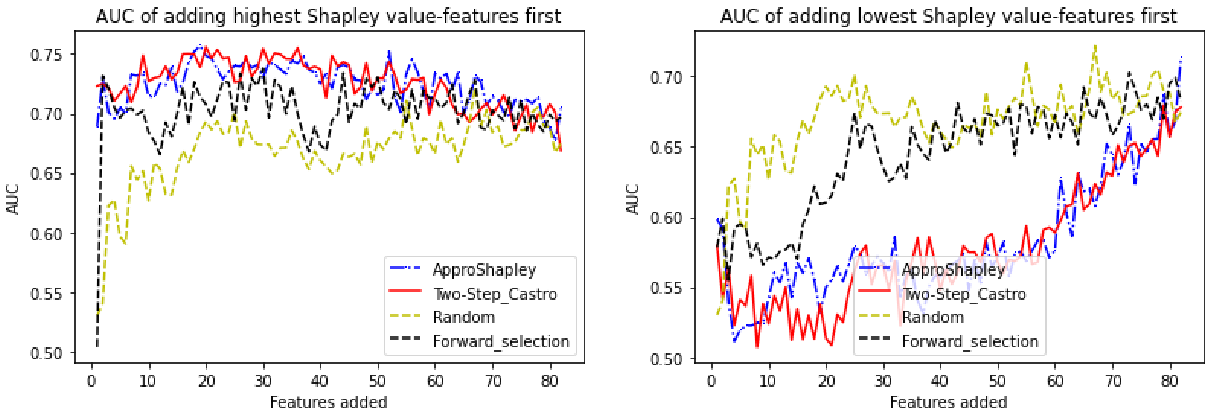
Table 6 Spearman rank coefficients DAC-6

DAC-6 data Decision Tree			
	Exact Shapley	ApproShapley	Two-step ApproShapley
Exact Shapley	/	0.8829	0.8829
ApproShapley	0.8829	/	0.8929
Two-step ApproShapley	0.8829	0.8929	/

DAC-6 data Logistic Regression			
	Beta coefficients	ApproShapley	Two-step ApproShapley
Beta coefficients	/	0.3871	0.3944
ApproShapley	0.3871	/	0.9855
Two-step ApproShapley	0.3944	0.9855	/

When comparing the coefficients of a linear model to the Shapley values, we see that lower levels of correlation exist. A possible explanation for this lower level of correlation can be the fact that a simple beta coefficient does not take into account the frequency of occurrence of a certain feature in the dataset, referring back to the fictitious example we provided in section 2. It is only a measure for the predictive value of the feature, and attributes great importance to features with good predictive power, but possibly a low occurrence in the dataset, which is exactly why we prefer the Shapley value to value the features.

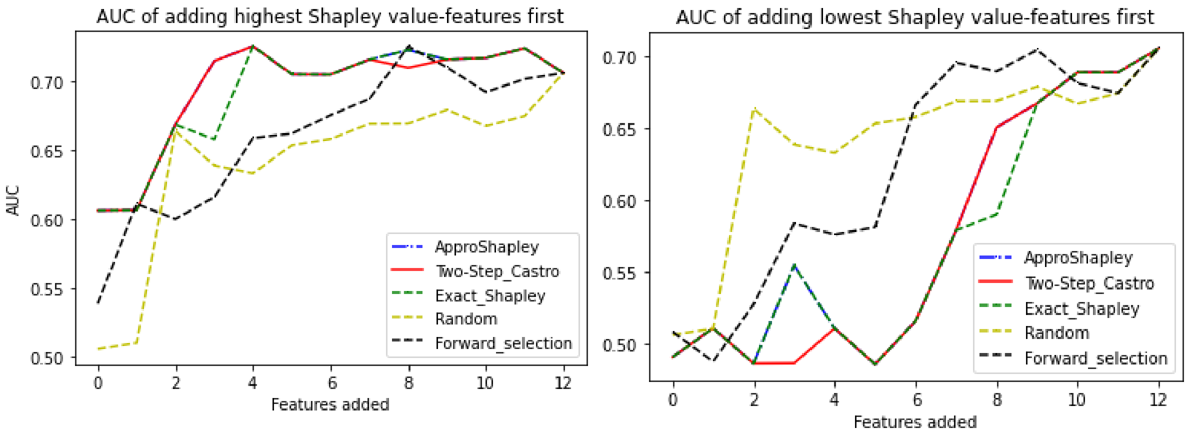
Figure 2 Local file Shapley values



An alternative way to evaluate the methods is to plot the change in performance of the model when we add the features to the model based on their respective rankings. Pane 1 and Pane 2 of Figure 2 plot the change in performance for the predictive model of the local file as we add features based on the highest-to-lowest Shapley values and vice versa respectively. The X-axis denotes the number of features added, the Y-axis denotes the performance for the corresponding number of features. We benchmark our methods against both random input selection and a greedy forward feature selection algorithm. The forward feature selection algorithm selects features based on the largest improvement in predictive performance after the previous feature is added at each step. Similar to the Shapley value estimation methods, results of the random and forward feature selection are also based on the same ten train-test splits.

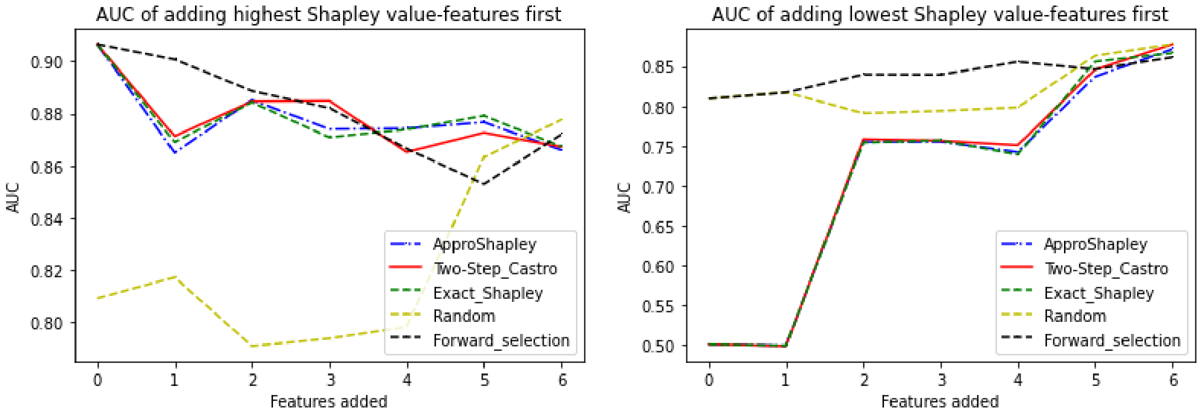
In the case of the local file, we see that adding high Shapley value-features to the model leads to an increase in performance for the first 30 to 35 features, which starts degrading when we add the remaining features. This is what we expect given that the features with negative Shapley values are added towards the end. The Shapley value outperforms random feature selection, and is also a clear improvement over forward feature selection. In the case of adding low Shapley value features first, we observe a stagnating performance when only the low Shapley value-features are added, followed by gradually increasing performance as we start adding the higher Shapley value-features. A real improvement in performance is noticeable after 60 features with the lowest values have been added. For both random and forward input selection, we see a much quicker rise in performance, indicating that already more valuable features are added sooner to the model than when we base our ranking on the Shapley value. The Shapley value is thus effective in ranking the features according to their predictive value.

Figure 3 ETR Shapley values



We present the results for the ETR dataset in Figure 3. When adding the high-Shapley value features first, an upward trend is observable when we add the first five features to the model. Afterwards, model performance stabilizes around the same performance level. The Shapley value clearly outperforms both random and forward feature selection. When we add the lowest Shapley values first, we observe that performance starts at level slightly below the base rate of 0.5 due to the strong negative Shapley value of the worst-ranked feature. Subsequently, performance increases notably after the five features with the lowest values are added, and features with higher Shapley values are starting to be included. In the case of random and forward input selection, performance rises much quicker, indicating that already more valuable features are added sooner to the model, thus showing again that the Shapley value is the most effective in ranking the features based on their contribution to the performance of the predictive model. Results for DAC-6 are presented in Figure 4. In the case of the DAC-6 data, the results are less pronounced due to the limited number of features available. In addition, note that the AUC for the DAC-6 data is incredibly high when only the feature with the largest Shapley value is used. As a result, adding features with high Shapley values still leads to a decrease in performance, as no model has better performance than a model based on this single feature. The forward input selection method also attributes the most value to this feature, and thus follows a similar trend to the Shapley value ranking methods. When we add the lowest Shapley values first, we again see that performance starts at a low level, and increases as the higher Shapley value points are added. In this case, the Shapley value is again better suited to identify the least valuable features compared to both random and forward feature selection.

Figure 4 DAC 6 Shapley values

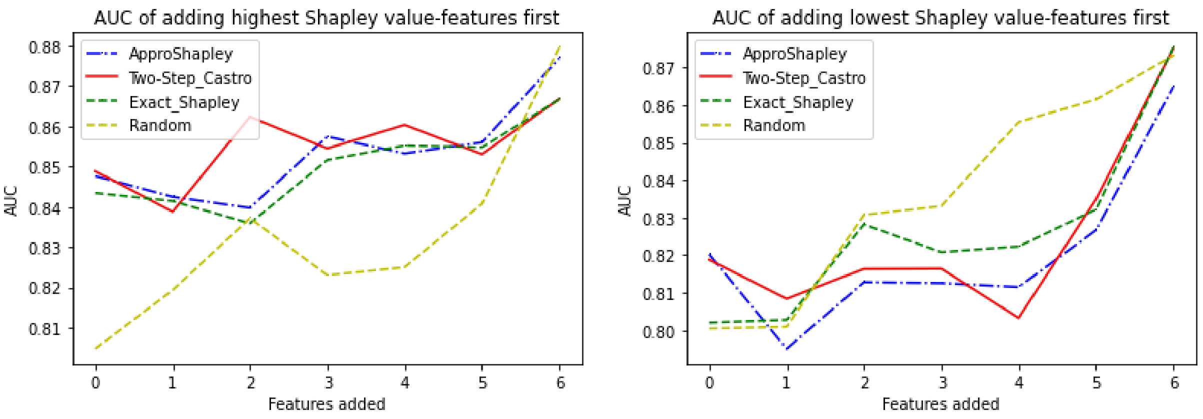


5.1 Prioritizing important features

As a second experiment, we use the Shapley value to determine which features should be acquired or cleaned with priority. Quite often, users of data can acquire more data at a certain cost, or users need to invest considerable resources to clean large sets of data before they can be useful. It is therefore important to know which data should be collected or cleaned first in order to allocate resources efficiently. To achieve this purpose, we run a simulation experiment to see whether the Shapley value can successfully determine for which features it is most interesting to acquire more observations. Unlike many previous active feature selection techniques, we do not consider which features we need to acquire for specific instances, but rather determine whether the feature itself should be considered for acquisition or cleaning.

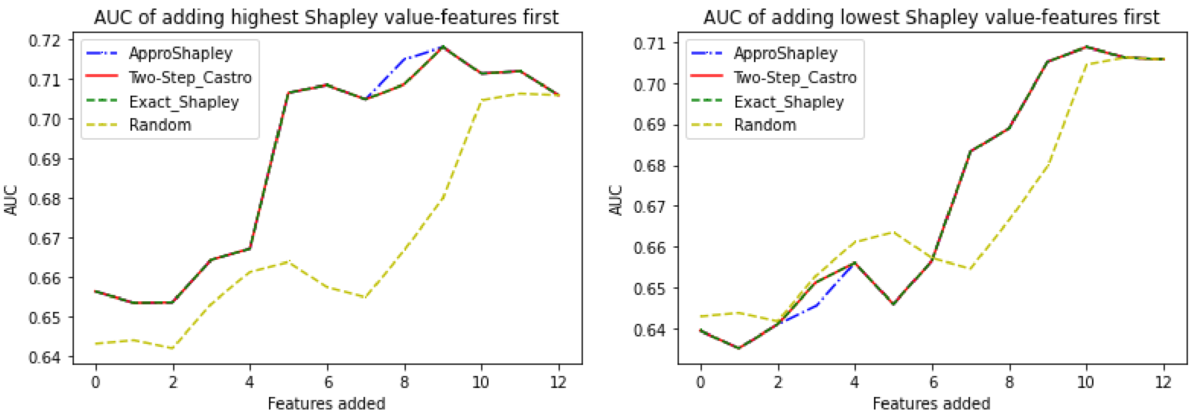
Specifically, we obtain a 'sample' of observations to perform the simulation. We delete approximately two-thirds of our training data and impute the deleted values based on the remaining third of datapoints to come to a new dataset. We chose to delete two-thirds of our data to ensure that the impact of newly added data will be large enough, while still ensuring that the remaining training dataset still contains enough useful information to train a predictive model. Subsequently, we recalculate the Shapley values of the features with ten different test-training splits for each data source, as it is more useful to examine how new acquisitions affect the distribution of estimations induced from different likely variations of the training set instead of examining the performance changes for a model based on one training set (Saar-Tsechansky et al., 2009). An obvious reason for this is that, in a real-world setting the training dataset could constantly change due to the acquisition of new information (Saar-Tsechansky et al., 2009). We thus want to reduce the risk that peculiarities in our test and/or training sets influence our results. We then add back the real datapoints to the dataset in a stepwise approach based on their Shapley values. We present both the case where we add the highest Shapley value features first, as well as the lowest Shapley value features.

Figure 5 Simulation experiment missing data DAC 6



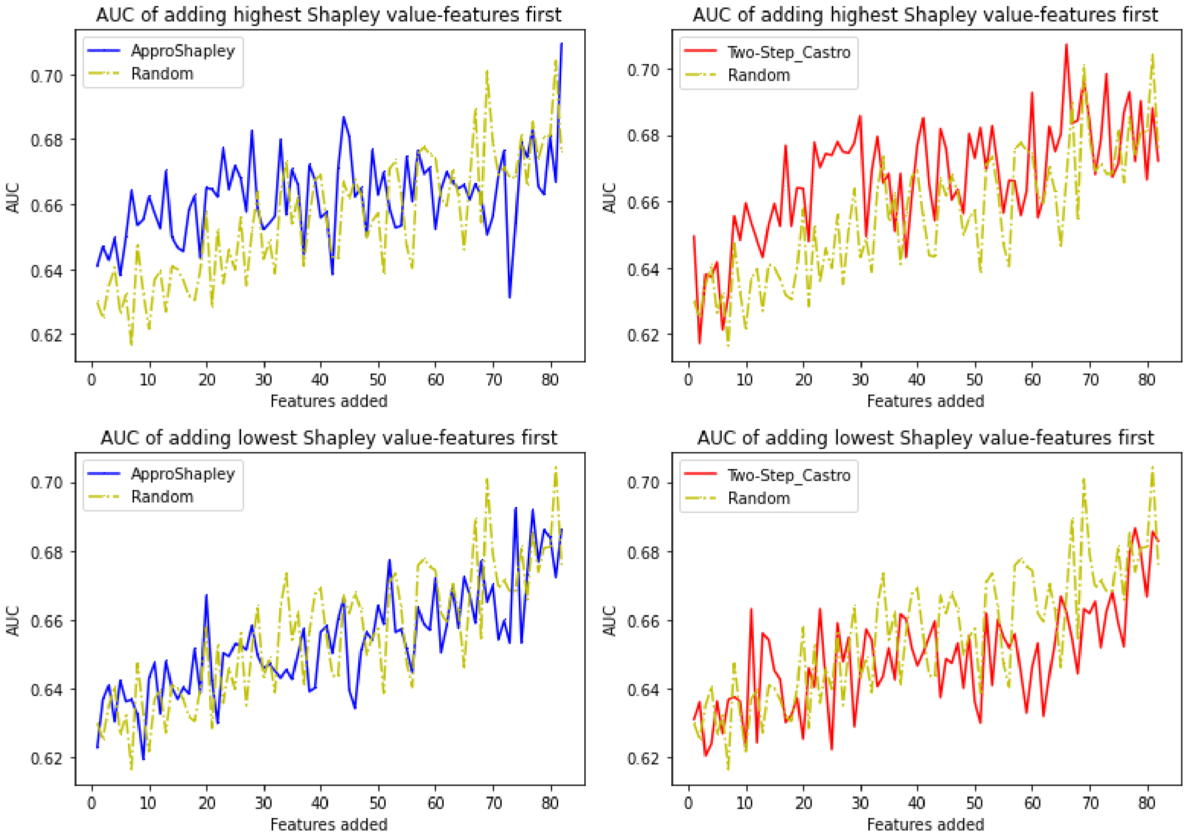
Results for DAC 6 data are presented in Figure 5. We see that performance rises much quicker when we prioritize adding the real values of the features with the highest Shapley values to the dataset, compared to adding the values of the lowest ranked features first. To obtain performance similar to the full model, we only need to add the real values of three to four of the most valuable features with priority compared to almost all features when we add the real values of the least valuable features first. In addition, we observe that adding the real values of the features with the highest/lowest Shapley values improves performance more quickly/slowly compared to adding real values of random features.

Figure 6 Simulation experiment missing data ETR



Results for ETR data are presented in Figure 6. We see again that the Shapley value successfully determines which features should receive priority when cleaning or acquiring the data. When we calculate performance based on datasets where we add the real values of the most important features first, we see that performance rises much quicker compared to the situation where we add the real values of features with the lowest Shapley values first. Specifically, we only need to add real values of six features when we prioritize the highest Shapley values to achieve performance similar to the complete model. In the case of adding real values of the lowest-ranked features first, we see that this level of performance is only achieved when adding real values of ten or more features, depending on the ranking method.

Figure 7 Simulation experiment missing data local file



Finally, results for the local file are presented in Figure 7. We benchmark the results for both sampling algorithms against random selection in separate plots for the sake of readability. Results are not as pronounced for this data source however, which is likely due to the larger amount of features and the small test set, which causes the curve to have an uneven course. Nevertheless, we see that performance of adding real values of the most important features according to both Shapley value sampling methods tends to be higher compared to adding the real values of random features. Using approximately 40 to 45 of the most valuable features in a model returns similar performance to the model on the full dataset, whereas we need to collect full data on almost all features when using the least valuable features first to notice a significant increase in performance. When we add the lowest ranked features first, we see that models with complete features based random feature selection tend to consistently outperform the models based on the Shapley value after the first 30 features are added, again indicating that the Shapley value is more effective in identifying the least valuable features. In summary, our valuation techniques can be used when practitioners are faced with budget constraints to prioritize the most valuable data for tax audit selection.

6 Discussion and conclusion

In this work, we examined how to attribute value to features in a tax setting for audit prediction. Our research is motivated by the need for tax authorities to keep enormous quantities of data manageable, as a result of recent disclosure regulations. We apply two different methods based on the theoretically sound Shapley value to attribute value to the different features of a predictive model. We find that the Shapley value successfully ranks features based on their value for predictive modeling. Our findings can be used for global interpretability of the model, feature selection purposes and to determine which data should be collected or cleaned with priority to achieve better performance. To the best of our knowledge, we are the first study to attribute value to confidential data tax authorities receive under the automatic exchange of information regulations of the EU and OECD.

One limitation of the study is that we assume the predictive model is given because we do not know which predictive model the tax authorities actually use in this setting. Even though both approximation methods are model agnostic and can thus be applied to any model, they both involve model retraining. The need for model retraining can cause long computational times when the predictive model is very complex. When the type of predictive model used by the tax authorities is known, model-specific applications of the Shapley value can be used if available to increase estimation efficiency. Another important limitation is possible selection bias. As mentioned before, we only have data on taxpayers that have been audited already in the first place. We do not have information on the exact reason these taxpayers were selected for audit. This bias could influence the importance of features in the sample, so caution is needed when generalizing these findings to the population.

Another limitation is the amount of available data. We perform supervised learning tasks, and thus need labeled instances. Obtaining class labels for instances in our context means that the tax authorities need to audit this instance, which turns labeling very costly. The limited availability of labeled instances in our datasets has implications especially for our sample experiment to prioritize certain features. Our valuation techniques are based on the value a feature currently has in a dataset. For smaller datasets like ours, it could be that a feature is valuable in truth, but that the information in our dataset is too limited to discover underlying predictive patterns. Expanding the dataset should relieve this problem.

7 Future research

As is the case in almost all applications of the Shapley value, calculations get challenging when datasets become larger. Future research could make use of ways to improve the efficiency of the estimation techniques further, such as only estimating the Shapley values of features in coalitions up until the intrinsic noise in performance of the model (Ghorbani & Zou, 2019) or using model-specific approaches when available. Future research could also try to integrate a forward-looking component in the estimation algorithm based on the expected improvement in predictive performance of a feature, in line with Saar-Tsechansky et al. (2009). This way, valuation would not only be based on the current value of the feature in the dataset, but would also contain an expectation of the value of newly acquired data.

In addition, several of the data sources used in this research have only been available for the few last years, which leads to a limited amount of labeled instances. Over the next few years, more labeled instances will become available which will improve the validity of our results. Another possible option is to combine data from different national tax authorities, as all our data sources are exchanged internationally. This would not only greatly increase the size of the dataset, but will also provide opportunities to research the data valuation problem in an international setting.

8 Bibliography

Basta, S., Fassetti, F., Guarascio, M., Manco, G., Giannotti, F., Pedreschi, D., . . . Pisani, S. (2009). *High quality true-positive prediction for fiscal fraud detection.* Paper presented at the 2009 IEEE International Conference on Data Mining Workshops.

Castro, J., Gómez, D., Molina, E., & Tejada, J. (2017). Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research, 82*, 180-188.

Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research, 36*(5), 1726-1730.

Chen, W., Zhang, M., Zhang, Y., & Duan, X. (2016). Exploiting meta features for dependency parsing and part-of-speech tagging. *Artificial Intelligence, 230*, 173-191.

Cohen, S., Dror, G., & Ruppin, E. (2007). Feature Selection via Coalitional Game Theory. *Neural Computation, 19*(7), 1939-1961.

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in neural information processing systems, 33*, 17212-17223.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res., 3*(Mar), 1289-1305.

Ghorbani, A., & Zou, J. (2019). *Data shapley: Equitable valuation of data for machine learning.* Paper presented at the International Conference on Machine Learning.

González, P. C., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications, 40*(5), 1427-1436.

Gupta, M., & Nagadevara, V. (2007). *Audit selection strategy for improving tax compliance: application of data mining techniques.* Paper presented at the Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December.

Hsu, K.-W., Pathak, N., Srivastava, J., Tschida, G., & Bjorklund, E. (2015). Data mining based tax audit selection: a case study of a pilot project at the minnesota department of revenue. *Real world data mining applications*, 221-245.

Keinan, A., Sandbank, B., Hilgetag, C. C., Meilijson, I., & Ruppin, E. (2006). Axiomatic scalable neurocontroller analysis via the Shapley value. *Artificial Life, 12*(3), 333-352.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).* Paper presented at the International conference on machine learning.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., & Rogers, A. (2013). Bounding the estimation error of sampling-based Shapley value approximation. *arXiv preprint arXiv:1306.4265*.

Matos, T., Macedo, J. A., Lettich, F., Monteiro, J. M., Renso, C., Perego, R., & Nardini, F. M. (2020). Leveraging feature selection to detect potential tax fraudsters. *Expert Systems with Applications, 145*, 113128.

Melville, P., Saar-Tsechansky, M., Provost, F., & Mooney, R. (2004). *Active feature-value acquisition for classifier induction.* Paper presented at the Fourth IEEE International Conference on Data Mining (ICDM'04).

Moeyersoms, J., d'Alessandro, B., Provost, F., & Martens, D. (2017). Attributing value in a data pooling setting for predictive modeling.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*, 72-81.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., . . . Bischl, B. (2022). *General pitfalls of model-agnostic interpretation methods for machine learning models.* Paper presented at the xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers.

OECD. (2014). *Guidance on Transfer Pricing Documentation and Country-by-Country Reporting*.

Provost, F., Melville, P., & Saar-Tsechansky, M. (2007). *Data acquisition and cost-effective predictive modeling: targeting offers for electronic commerce.* Paper presented at the Proceedings of the ninth international conference on Electronic commerce.

Saar-Tsechansky, M., Melville, P., & Provost, F. (2009). Active Feature-Value Acquisition. *Management Science, 55*(4), 664-684.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the theory of games, 2*(28), 307-317.

Sheng, V. S., & Ling, C. X. (2006). *Feature value acquisition in testing: a sequential batch test algorithm.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

Song, E., Nelson, B. L., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification, 4*(1), 1060-1083.

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research, 11*, 1-18.

Van Assche, A., & Blockeel, H. (2007). *Seeing the forest through the trees: Learning a comprehensible model from an ensemble.* Paper presented at the ECML.

Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing, 86*, 105895.

Zheng, Z., & Padmanabhan, B. (2002). *On active learning for data acquisition.* Paper presented at the 2002 IEEE International Conference on Data Mining, 2002. Proceedings.