Pattern Analysis and Applications (2020) 23:1457–1477
https://doi.org/10.1007/s10044-019-00852-w

INDUSTRIAL AND COMMERCIAL APPLICATION

Customs fraud detection

Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue

Jellis Vanhoeyveld[1] · **David Martens**[1] · **Bruno Peeters**[2]

Received: 11 December 2018 / Accepted: 11 October 2019 / Published online: 30 October 2019

Abstract
In this customs fraud detection application, we analyse a unique data set of 9,624,124 records resulting from a collaboration with the Belgian customs administration. They are faced with increasing levels of international trade, which pressurizes regulatory control. Governments therefore rely on data mining to focus their limited resources on the most likely fraud cases. The literature on data mining for customs fraud detection lacks in two main directions that are simultaneously addressed in this paper: (1) behavioural and high-cardinality data types are neglected due to a lack of methodology to include them. We demonstrate that such fine-grained features (e.g. the specific entities such as consignee, consignor and declarant and the commodities involved in a declaration) are very predictive. (2) Studies in the tax domain most often use standard learning algorithms on their fraud detection applications. However, customs data are highly imbalanced and this poses challenges for many inducers. We present a new EasyEnsemble method that integrates a support vector machine base learner in a confidence-rated boosting algorithm. This results in a fast and scalable learner that is able to drastically improve predictive performance over the base application of a support vector machine. The results of our proposed framework reveals high AUC and lift values that translate into an immediate impact on the customs fraud detection domain through an improved retrieval of tax losses and an enhanced deterrence.

Keywords Fraud detection · Behavioural data · High-cardinality attributes · Imbalanced learning · Support vector machines

1 Introduction

Recent decades have witnessed an explosion in the volume, variety and complexity of goods crossing the national borders as a direct effect of globalization, digitalization (e.g. e-commerce) and multi- or bilateral trade agreements [25, 27]. Customs are faced with the task of checking the legitimacy[1] of international trade transactions and impose[2] customs duties on imported or exported goods. The growth rate of such trade is disproportionate to the reinforcement of customs resources and this poses severe challenges for customs authorities to verify all import declarations and ensure compliance.

In general, customs duties are taxes levied on the import and export of goods. The tariff is either a percentage of the taxable amount[3] or a fixed amount calculated on the quantity of the goods (e.g. x Euro/kg), or a mixture thereof. In principle, a rate is set for each type of goods. Customs duties serve as an important source of revenue [48]. In Belgium, €2,55 billion[4] is collected related to import duties in 2016. Also, the customs rate can be adjusted to protect the domestic

✉ Jellis Vanhoeyveld
jellis.vanhoeyveld@uantwerpen.be;
vanhoeyveld.jellis@gmail.com

[1] Department of Engineering Management, University of Antwerp, Prinsstraat 13, Antwerp, Belgium

[2] Faculty of Law, University of Antwerp, Venusstraat 23, Antwerp, Belgium

[1] Legitimacy should be interpreted broadly and is not limited to verifying whether the imposed duties are paid and transport documents are filled in correctly. It also means protecting the environment and society against imported harmful/dangerous goods (e.g. counterfeit goods of low quality).

[2] Additionally, national taxes such as value-added tax (VAT) and excises related to international trade transactions can be collected by customs authorities.

[3] The taxable amount is the customs value or the amount on which the tax is levied.

[4] We refer to https://financien.belgium.be/nl/Statistieken_en_analysen/jaarverslag/cijfers/budget-ontvangsten/ontvangsten-aa-douane-en-accijnzen-1 for additional figures (also including VAT and excises).

industries from foreign exporters disrupting the local market with dumping prices.

The customs taxation system, like any tax regime, is prone to fraudulent abuse. There exist several kinds of fraud which are simultaneously addressed in this paper, such as [11, 25, 46, 48]: product misclassification, valuation fraud, smuggling, illegal drug traffic, importation of counterfeit goods and the manipulation of the origin of goods. An important type of fraud is valuation fraud, where the value of a good is under-declared to illegitimately enable a lower tax liability. Hence, the estimation of the correct value of goods is a vital task and is very challenging, as it depends on many factors [11, 48]: the type of good, the brand and serial number (depends on the supplier), the country of origin, the imported quantity, the specific trade agreements and, finally, global market conditions.

To harmonize regulatory control and trade facilitation[5] [25, 48], customs have adopted data mining as a key risk management tool that allows them to concentrate their limited resources on the high-risk targets without sacrificing any capacity to the compliant operators. This view is also acknowledged by the World Customs Organization (WCO) and examples of such risk management programmes include cargo selectivity systems, post-clearance audit programmes and Authorized Economic Operator (AEO) programmes [8].

Customs authorities worldwide have similar data repositories. Within the European Union, the Single Administrative Document (SAD)[6] [17] forms the basis for all customs declarations and covers all customs procedures such as import, export, transit and the placement of goods in customs warehouses. The European Commission issued guidance documents [14] to ensure a uniform implementation and a common understanding of the legislation (the Union Customs Code) concerning the SAD.

Efficient fraud detection leads to an immediate recovery of financial losses and enables an enhanced deterrence. Additionally, it allows for a smoother flow of goods for the compliant operators. Customs can ascertain the legitimacy of a trade transaction by (semi-automatic) analysis of trade documents (e.g. invoices and declarations), employing historic databases (e.g. AEO traders), gathering intelligence from informants and conducting physical cargo inspections [25]. With respect to valuation fraud, customs officers can also rely on their past experience to compare the value of an item under consideration with the value of similar goods (originating from the same country and/or supplier) [48].

There are a number of interesting domain challenges that require careful consideration. (1) The scoring applications require scalable data mining algorithms [27] as the volume of data is large and the declared goods need to be processed within a matter of seconds in an online environment.[7] (2) The cases of fraud are rare in comparison with the abundant legal cases (skewed data set). This imbalanced learning issue needs to be tackled during the construction of predictive models. (3) The dynamic nature of fraud requires a regular update of the models and the use of different strategies to target new fraud types (e.g. random targeting [25, 27] and outlier detection [11]—see Sect. 2). Finally, (4) customs authorities have limited resources (capacity), which means they can only verify a small fraction of incoming declarations.

In this study, we outline the development of a supervised customs fraud detection system taking into account the aforementioned domain challenges. This project describes the results of an ongoing collaboration between the Belgian Federal Public Service Finance, division Customs and Excise and the Antwerp Tax Academy and the Applied Data Mining research groups.

2 Related work

The data mining literature on customs fraud detection is relatively scarce due to tax administrations perceiving this kind of data as highly sensitive. Nevertheless, some studies have been published on the topic and they can be categorized into supervised versus unsupervised approaches[8] [11]. The former type of methods requires a set of labelled (i.e. fraud or compliant) training instances to construct a fraud detection model. Unsupervised approaches are the most flexible, as they do not require labelled data and can be directly applied to the entire population. Popular techniques in this class include expert rule-based systems and unsupervised anomaly detection methods (see the upcoming paragraphs).

The majority of studies on customs fraud detection employ supervised classification algorithms: Kumar and Nagadevara [27] apply decision trees and neural networks

[5] Trade facilitation means a rapid clearance of customs goods to have a minimal impact on economic commerce.

[6] Each member country can stipulate a number of additional national regulations.

[7] There are two types of inspections: (1) physical cargo checks when the goods enter the territory (e.g. inspecting containers). (2) Post-clearance audits which entail checking the books and verifying trade documents (e.g. invoices, SAD declarations) for irregularities. Regarding the former, Belgian customs impose a 6 second rule for the automated processing of an article involved in a SAD declaration (online environment). The post-clearance audit checks can be conducted up to 4 years after the date of declaration.

[8] There exists a third category of techniques, the semi-supervised approaches, that learn a discriminative boundary around the instances of a single class [5]. However, they do not seem to be applied in the area of customs fraud detection.

under various imbalanced learning settings on Indian customs data. Han and Ireland [25] evaluate the performance of three selection methods currently employed by the Korea Customs Service. They constitute a random selection, a manual selection and a supervised rule-based selection. Shao et al. [46] learn decision trees on the import declarations in China. The learned rules are subsequently amended by local customs officers to integrate expert knowledge. Yaqin and Yuming [55] establish a classification model on Chinese customs data based on association rules generated by the Apriori algorithm. However, all these studies lack in three directions (which we will cover in more detail in the upcoming paragraphs and Sect. 3): (1) they do not incorporate behavioural or high-cardinality attributes; (2) there is limited consideration for the class-imbalance problem as only a single paper deals with this issue; and (3) they rely on performance measures that do not take the limited capacity of customs into account.

Unsupervised approaches are infrequently studied in customs fraud detection. Singh et al. [48] developed an expert rule-based system for Indian customs data, where selection officers propose a set of rules that have an associated sensitivity (e.g. *if* importer is a trader *then* sensitivity is very high). The final model, a hybrid hierarchical fuzzy controller, combines the sensitivity of the rules into a final sensitivity score. These rules require regular updates in a dynamic fraud environment, which is a daunting task. Anomaly detection techniques constitute another type of unsupervised learning approach, where one wishes to find those entities that display a conduct that deviates from the common behaviour within the population under consideration [1, 5]. One assumes fraud to be rare and different from the behaviour of the compliant majority group. The studies of Digiampietri et al. [11] and Rad et al. [43] fall in this category and deal with customs fraud in Brazil and Iran, respectively. The use of outlier detection methods is rare, and this trend seems to be confirmed in the broader area of financial fraud detection [37, 53]. This is because supervised approaches typically outperform unsupervised techniques on the known set of fraudsters.

Customs fraud detection applications are confronted with the class-imbalance problem, as the number of defrauders is dominated by the number of compliant entities [11, 27]. Many supervised classification algorithms face difficulties when confronted with this imbalance [9, 29] (such as decision trees [31], neural networks [34] and support vector machines (SVMs) [2]). The learners will emphasize the majority class (e.g. compliant instances) and neglect the minority class (e.g. fraud cases), where the latter is the event of interest [52]. A complete literature overview on the imbalanced learning issue is beyond the scope of this paper and is addressed in the works of Chawla [7] and He and Garcia [26]. The imbalanced learning issue is rarely addressed in the area of tax fraud detection. In the domain of customs fraud, only a single paper proposed several solutions to circumvent the problems related to class imbalance [27]. They rely on simple duplication of minority class instances or undersampling of majority class instances at random or the inclusion of different weights associated with the fraud and legal class. However, more advanced techniques exist, as discussed next.

In general, methods designed to overcome the imbalanced learning issue fall in two directions [26, 27, 32, 52]. The first class of techniques operate at the data level and provide a more balanced distribution to the underlying inducer. They consist of oversampling of the minority class (by duplication of minority instances or generating artificial minority examples), undersampling of majority class instances (at random or using some informed approach that retains the most important majority examples) or a combination thereof. The second class of methods work at the algorithmic level[9] and integrate misclassification costs in the learning process (either directly during the design of algorithms or by including them via cost-sensitive boosting variants).

The EasyEnsemble (EE) technique was proposed by Liu et al. [32] as an imbalanced learning solution operating at the data level. The method samples a number of balanced subsets from the training data (bagging) and feeds these subsets to the AdaBoost (AB) boosting algorithm. More details are presented in Sect. 4.2.3. The EE version of Liu et al. [32] relies on a decision tree base classifier that is fed to a discrete version of AB. Their formulation has been adopted in several benchmark/application studies such as the works of Kumar et al. [28], Parvin et al. [38] and Yuan and Ma [56].

There are a limited number of published studies integrating a SVM base classifier in the EE formulation as conducted in this work. However, the SVM allows the strength of the base learner to be controlled by an appropriate choice of hyperparameters and this is an attractive feature for the AB algorithm [52]. Miguéis et al. [35] claim to test a SVM-based EE method, yet a detailed reading reveals that only the bagging component of EE is retained and the boosting part is ignored. Liu [30] integrates a standard nonlinear SVM with fixed hyperparameter settings in a discrete version of AB. Recently, Vanhoeyveld and Martens [52] made use of an instance-weighted linear SVM and subsequent logistic regression (LR) base classifier in the EE process, where the boosting component relies on an improved (with respect to a discrete version) confidence-rated AB algorithm [45]. We explain this approach in detail, together with a couple of advantages, in Sect. 4.2.3.

[9] Methods at algorithmic level are also called cost-sensitive learning techniques.

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

Identity	Structured			Behavioural	High-cardinality			Label
Article	Mass	Transport	Procedure	Commodity codes of articles in the same declaration	Commodity	Consignee	Country	
A_1	1500	Ship	4200	{'1702201090', '0710210010', '9025804090', '7202491011'}	'1702201090'	C_{2003}	CA	−1
A_2	2300	Train	0121	{'5209110000','9031809110', '0806101005'}	'5209110000'	$C_{19.600}$	RU	−1
A_3	450	Airplane	4000	{'1507109000','2008305510', ⋯ ,'1701111000'}	'1507109000'	$C_{120.000}$	AU	−1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$A_{\lvert\text{Train}\rvert}$	3000	Ship	5111	{'1507109000','1604160000', ⋯ ,'0406900100'}	'1517101000'	$C_{1.000.000}$	SO	+1

Each article A_i involved in a customs SAD declaration can be characterized by traditional attributes (e.g. net mass, transportation mode, requested procedure), behavioural features (e.g. commodity codes of all articles in the same declaration) and high-cardinality variables (e.g. commodity code, the recipient of the commodity (consignee), country of origin). For a subset of articles corresponding to checks conducted by customs officers, we know their label: compliant (−1) or fraud (+1)

Three kinds of data sources will be considered in this study that, due to their inherently different nature, require a different modelling approach: traditional, behavioural and high-cardinality data. We will introduce these types in the next paragraphs and illustrate their distinction by means of an example in the area of customs; see Table 1. Section 4.1 extends the discussion of the data set.

"Traditional" structured data are most often used in data mining studies and have a low-dimensional (around 10–100 features) and dense representation in feature space [20]. They constitute several continuous attributes (e.g. net mass, price, number of items) and discrete attributes with a limited number of categories (e.g. transportation mode, requested procedure—see guidance document [14]). Encoding the latter with dummy encoding still results in a low-dimensional feature space.

Behavioural data are *high-dimensional* (around 10^4-10^9 features) and *sparse* and are characterized by *m-to-n* relations [10, 47], as we explain next. They usually arise from capturing the fine-grained actions and/or interactions of persons or organisations (though this is not required when we refer to behavioural data in this study). They can be represented as a large and sparse matrix or, equivalently, as a bipartite graph [49] and require tailored techniques to deal with its high-dimensionality and sparseness. We refer to De Cnudde et al. [10] for an overview on the kind of classifiers that are suitable for such behavioural data sets. As an example, say that we want to predict whether an article (instance) appearing in a SAD declaration is fraudulent or compliant. Since there can be many articles involved in a single declaration, one could characterize each article (instance) by the commodity codes of all goods (articles) occurring in the same declaration; see Table 1. This can be represented as a large and sparse matrix, where an article appears in a row and each column (feature) corresponds to a specific commodity code. A value of 1 occurs at position (i, j) if article A_i has commodity code j appearing in the same declaration (a value of 0 otherwise). We refer to the top-left illustration of Fig. 3 to represent this situation. There can be multiple active features (columns with a 1) in each row, which we refer to as *m*-to-*n* relations (an instance can be involved in *m* relations and a column (feature) can be involved in *n* relations). This matrix is high-dimensional because of the more than ten thousand commodity codes appearing in the EU TARIC online database.[10] Furthermore, this matrix is very sparse because usually there are far fewer commodity codes occurring in a single declaration than the total number of possible commodity codes.

High-cardinality variables are discrete nominal attributes that take on a large number of distinct values,[11] ranging up to several millions. Continuing the running example of the previous paragraph, each article can be represented by its associated commodity code; see Table 1. A similar sparse matrix representation can be obtained, where each row corresponds to an article and each column to a specific commodity code. In this case, each row would only show a single 1 across all columns (e.g. article A_1 has a value of 1 only at the position of its associated commodity code '1702201090'). We refer to the top-right illustration of Fig. 3 to represent this situation. High-cardinality data are therefore characterized by 1-to-*n* relations (an instance can only be involved in 1 relation and a column (feature) can be involved in *n* relations). This is the main difference with behavioural kinds of data, where there can be more active features in

[10] TARIC extends the Combined Nomenclature (CN) and contains tariffs for each commodity according to its country of origin. The CN is a tool for the harmonized classification of goods within the EU and is a further development (with special EU-specific subdivisions) of the WCO's Harmonized System Nomenclature (HSN) [16].

[11] In this study, in accordance with Moeyersoms and Martens [36], an attribute is of high-cardinality in case it has more than 100 different categories.

Table 2 Comparison of supervised classification techniques adopted in the customs fraud detection literature (rule-based techniques (e.g. decision trees, association rules) and shallow neural networks) to the methods employed in this study (SVM_LIN, SVM_RBF and EE)

Technique	Applied in literature for Beh data	Scalability	Tackles IL issue	Comprehensibility	Predictive performance
Rule-based	No	No	No	Yes (trad data)	Low
Neural network	No	No	No	No	Avg.
SVM_LIN	Yes	Beh data: Yes Trad data: No	No	Yes	Low
SVM_RBF	Yes	No	No	No	High
EE	Yes	Yes	Yes	No	High

For a number of combinations in the table, it is appropriate to distinguish between behavioural (Beh) data and traditional (Trad) data. IL is short for imbalanced learning

each row. Moeyersoms and Martens [36] note that high-cardinality features are rarely used in predictive models because including them with dummy encoding would lead to an explosion in the dimensionality of the resulting data set, which most methods are currently unable to handle. Furthermore, rule-based methods would be incomprehensible if a rule is developed for each value of a high-cardinality attribute. They applied three transformation methods—weight of evidence (WOE) [57], supervised ratio (SR) and Perlich ratio (PR) [39]—that transform a high-cardinality variable into a single continuous feature whose value is correlated with the target label (i.e. fraud or legal).

We close this section with a high-level characterization of reported supervised classification algorithms used in the customs fraud detection literature and compare them with the techniques we propose in this paper; see Table 2. The description is based on the following characteristics: has the method been applied in the broader literature dealing with behavioural data? We refer to the work of De Cnudde et al. [10] and Vanhoeyveld and Martens [52] to answer this question. A second characteristic measures whether the algorithm scales well with the large data sizes that can be encountered. We indicated yes in case the method has a linear complexity in terms of the number of instances and features or in case the algorithm is parallelizable. We also indicate if the standard method in its basic form suffers from the imbalanced learning issue [2, 31, 34]. We should note that cost-sensitive versions of these algorithms have been proposed in the broader machine learning literature to deal with this problem, but they have not been examined in the area of customs fraud detection. Fourth, the comprehensibility of classification algorithms is indicated. In a customs fraud detection context, an explanation of why an instance is regarded as fraudulent can form the starting point of a tax investigation and can aid in the acceptance of the model by domain experts [33]. Finally, a general rough indication of the predictive performance of the envisioned techniques is outlined [3]. It is important to remark that the performance for a particular application depends on the specific data set under consideration, whether the hyper-parameters of the methods are well tuned and the choice of evaluation metric.

3 Contributions

The main contribution of this paper lies in the systematic investigation of the predictive value for customs fraud of each data source individually and the (added) value of combining them, taking into account the class-imbalance problem. Figure 1 presents the framework and is explained in more detail in Sect. 4.3.1. This framework has never been applied in the area of tax fraud detection yet proves to be very valuable. The high predictive performances revealed in Sect. 5 for high-cardinality and behavioural data, improving significantly when considering imbalanced learning solutions, highlight the importance of including such fine-grained features and are potentially valuable for customs administrations worldwide. Specific contributions in each

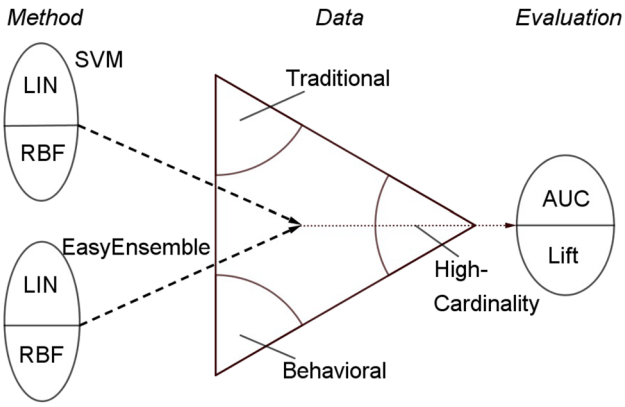


Fig. 1 Methodological framework. SVMs with linear (LIN) and non-linear (RBF) kernels and their integration with EasyEnsemble are applied on the three sources of data and evaluated with two performance metrics

part of the framework (data, methods, evaluation) will be outlined next.

The literature overview on customs fraud detection (see Sect. 2) reveals that only traditional data types have been used. The information available in high-cardinality or behavioural data is not used directly at identifier level. Most studies typically include derived low-dimensional attributes at higher abstraction levels to retain some of the information included in such features [11, 46, 48]. As an example, the importer can be represented by its type, which can take on five values [48]: government, public sector, manufacturer, trader or individual. However, such an approach loses the fine-grained information available at identifier level (e.g. who is the specific importer? Which specific commodity is imported?). The inclusion of such data is also very rare in the broader domain of tax fraud detection. One paper in the area of corporate residence fraud [21] assesses the value of behavioural and traditional data by analysing invoicing data. However, this kind of data differs from the type of high-dimensional and sparse data we consider in this study.

In this paper, the focus lies on identifying which data sources are highly predictive for customs fraud. Most studies in the tax domain typically examine a single data source (traditional data) and investigate which modelling technique is most suitable for their application. We argue that the *data perspective* could be considered as at least equally valuable than the *algorithmic perspective*. Knowing which data sources are predictive contains vital information and could influence certain policy making decisions (see the related discussion in Sect. 6). In Sect. 4.2.1, we will motivate our choice of modelling techniques. Having said this, there are two important remarks: (1) we note that the framework is open to different kinds of inducers and (2) it does not necessarily mean our proposed methods are optimal for the current application (though we observe high performances).

Even though this paper focuses on the data perspective, there are also a number of algorithmic contributions that we outline in the current and next paragraph. The first contribution is geared towards the imbalanced learning problem that is rarely addressed in the area of tax fraud detection. We make use of the EE technique, introduced in Sect. 2, as a state-of-the-art imbalanced learning solution. Vanhoeyveld and Martens [52] integrate an instance-weighted *linear* SVM and subsequent LR as a base learner in EE with a confidence-rated boosting algorithm. They investigate this technique on a large repository of imbalanced *behavioural* data sets. In our study, we develop a similar EE implementation in terms of the bagging and boosting components. The main difference (novelty) is that the base learner includes an instance-weighted linear or *nonlinear* SVM and subsequent LR that is suitable for *traditional* kinds of data.

The second algorithmic contribution focuses on the inclusion of high-cardinality attributes in predictive models, which presents a challenge (as discussed in Sect. 2). Moeyersoms and Martens [36] propose the SR technique. We argue that this method suffers from stability issues and propose a new smoothed version to overcome these concerns (see Sect. 4.2.4). Furthermore, a novel approach for dealing with high-cardinality data is presented, which pre-trains a predictive model on the sparse matrix representation of the attribute using tailored techniques. All methods employed transform a high-cardinality attribute into a single continuous feature that is correlated with the label; see the related discussion in Sect. 4.2.4. This is the main difference with creating aggregated (derived) attributes presented earlier in this section.

Studies in the tax domain commonly apply a standard learning algorithm with a default threshold, which does not take into account the limited resources available at customs (e.g. see [25, 27, 44]). In Sect. 4.3, suitable performance metrics, which are common in other data mining domains, are proposed that are either independent on a specific threshold or are evaluated in accordance with the available capacity. Note that these measures themselves are not new, but the novelty lies in the application thereof in the area of customs fraud detection.

4 Data, methods and evaluation

In this section, we outline the data used in this study, explain the different methods and highlight the employed evaluation metrics. Due to the highly sensitive and confidential nature of the data, we cannot disclose the relevant data sources. However, the specific implementations of the EE algorithm (see Sect. 4.2.3) and the results presented in Sect. 5 are accessible at the link: http://www.applieddatamining.com/cms/?q=software. Furthermore, additional results with respect to lift at different (arbitrarily chosen) capacity values are provided through this link.

4.1 Data

The Belgian customs administration provided us with a unique data set containing anonymized data regarding all SAD import declarations of the years 2015 and 2016 and involves an aggregated total of 9,624,124 declared records (an average of approximately 9.5 records every minute). The databases are constructed at article level; hence, each record corresponds to a single commodity involved in a declaration. Note that an identifier at declaration level is also provided with each article, so we can easily link articles belonging to the same declaration. Appendix 1 contains a SAD declaration form. Its numerous fields constitute raw input variables x_{raw} to characterize an article appearing in a customs declaration. We refer to Table 1 for an example thereof.

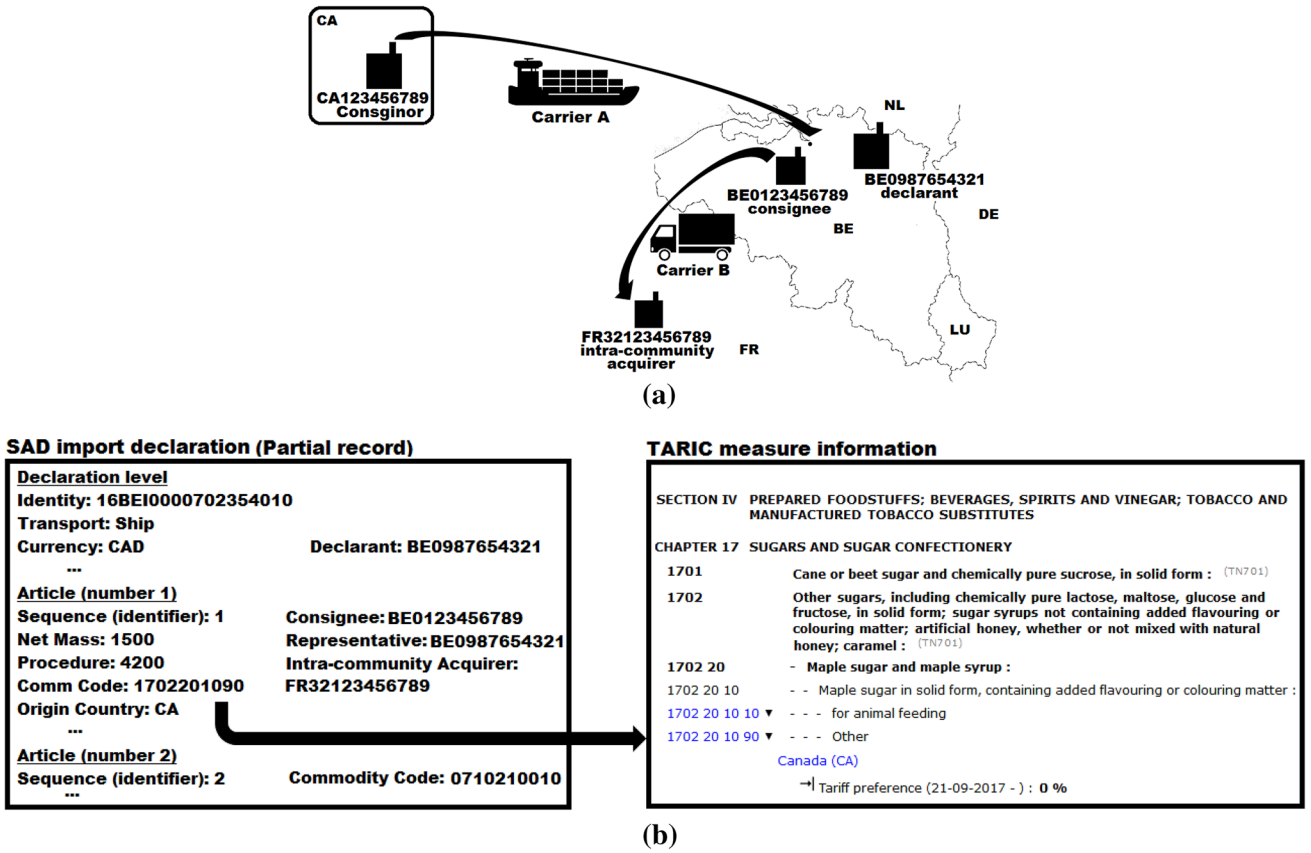


Fig. 2 Fictitious example **a** Maple syrup produced by a Canadian company CA123456789 (consignor) is imported by the Belgian company BE0123456789 (consignee) and forms the subject of a VAT exempt supply to the French company FR32123456789 (intra-community acquirer). The Belgian company BE0987654321 (declarant) is a forwarding agent that submits the SAD declaration. **b** Associated (partial) SAD declaration (left) with indication of the maple syrup import tariff retrieved from the TARIC database (right)

Let's look more closely to some of the specific fields of the SAD form by means of a realistic example, which is associated with the first row (Article A_1) of Table 1 and is illustrated in Fig. 2, where customs procedure 4200[12] (see guidance document [14]) is adopted. The Canadian company with VAT number CA123456789 is a foreign goods exporter (**consignor**) that manufactures maple syrup. The Belgian company BE0123456789 is the **consignee** that purchases these goods from the Canadian firm. Carrier A transports this commodity to the harbour of Antwerp, the second largest port in Europe in terms of volume of freight loaded or unloaded (224 million tonnes in 2017) [40]. Subsequently, carrier B delivers the maple syrup to the intended recipient (**intra-community acquirer**), the French company with VAT number FR32123456789 that purchased these goods from the Belgian firm BE0123456789. The **declarant**, the Belgian company with VAT number BE0987654321, is the entity that submits a SAD declaration to Belgian customs. In the example, the declarant is an external **customs representative** selected by the consignee (BE0123456789). A customs representative is *any person appointed by another person to carry out the acts and formalities required under the customs legislation in his or her dealings with customs authorities* (art. 5.6 UCC). This can for instance be a subsidiary, an accountant, a university professor (secondary occupation), etc. In practice, this is commonly a forwarding agent.[13]

[12] This means that imported goods are released for free circulation and their associated customs duties are levied in one member state (i.e. Belgium), yet payment of VAT (and where applicable excise duties) is suspended because the import is directly followed by an intra-community supply of the goods to another member state (i.e. France). VAT (and excises) are due in the member state of final destination (i.e. France).

[13] A forwarding agent (or freight forwarder) [51] is an entity that organizes the delivery of goods, without doing the actual transportation. He is responsible for choosing the carriers that deliver the goods in the most effective way in terms of transportation time and costs. Furthermore, he prepares the necessary documents (customs and insurance) and transport certificates. The forwarding agent acts as an intermediary in the logistics chain.

In Belgian customs declarations, the declarant can take on several 'values' depending on the procedure followed. It can be the consignee if this operator is established (has a VAT number) in Belgium. Alternatively, it can also be a customs representative as shown in the example. Note that the latter situation is typical in case the consignee is located abroad (e.g. a German company) and has no representation in Belgium. Figure 2b (left) shows the corresponding partial SAD declaration, where the identities of the involved operators are indicated together with a number of traditional attributes (e.g. net mass, currency, transport mode) and the declared commodities. The maple syrup constitutes the first article with the TARIC commodity code '1702201090'; see Fig. 2b (right[14]).

The numerous fields of the SAD declaration form contain a total of 90 raw variables. Seventy-six of them are of type traditional data (see Table 1 for examples) of which the vast majority are nominal (categorical) data. The remaining 14 variables are high-cardinality attributes, each containing more than 100 distinct values. Examples thereof include the specific entities[15] involved in a declaration as outlined in the previous paragraph (consignee, intra-community acquirer, representative, declarant), where these four types constitute four separate variables of type Operator Identity (OperID), the type of article (CommodityCode), the country of origin, the dispatch country, the identity of the customs warehouse, etc. As indicated in Sect. 3, such data are not included at identifier level in the prior literature. In our work, we include these data sources in the predictive modelling and consider this a key aspect because such data expose fine-grained information of the logistics chain.

We consider two types of behavioural data: (1) each article is represented by the four OperIDs (consignee, declarant, representative, intra-community acquirer) occurring in the declaration[16] and (2) each article is characterized by the commodity codes of all goods in the same declaration; see Table 1. We consider them as behavioural kinds of data because they can be represented as a large and sparse matrix and there is more than one active feature in each row (satisfies the behavioural data description outlined in Sect. 2).

The feedback results regarding checks conducted by customs officers were also available. Articles that were inspected are given a label (+1) in case of fraud and (−1) in case of compliance. We extracted a data set containing only these labelled instances that can be used for model construction and evaluation. Note that, due to capacity constraints faced by customs and the trade facilitation trade-off, the size of this labelled set is a small fraction of the total 9,624,124 records (it contains around 100,000 records). Its fraud occurrence is 3.74%. Due to confidentiality reasons, we cannot disclose the exact size of this subset nor further descriptive statistics. Note that all possible fraud types, such as (but not limited to) the ones described in Sect. 1, are included in the analysis.

4.2 Methods

4.2.1 Motivation

Regarding the choice of base learner, we opted for a SVM for a variety of reasons. First of all, the SVM is a state-of-the-art inducer that has achieved high predictive power across a large variety of application domains if its hyperparameters are tuned well. Secondly, it is the most commonly used learner for dealing with behavioural kinds of data as revealed in the literature overview of De Cnudde et al. [10]. Thirdly, learners that are traditionally being used for customs fraud detection (e.g. decision trees, neural networks) are not applicable for behavioural or high-cardinality data. By choosing for SVMs, we can consistently use the same kind of learner for dealing with all data types and hence the (added) predictive value of each data source can be more fairly compared. Fourth, an essential component of the EE algorithm is boosting, where too strong learners should not be used [22, 54]. The strength of a SVM can be controlled by changing its hyperparameters. Finally, the distinction between linear and nonlinear classifiers can be easily obtained by the choice of kernel. Linear methods are more comprehensible than a nonlinear variant, usually at the cost of a lower predictive power. This paper reveals the associated performance drop in opting for a linear method.

With respect to the choice of imbalanced learning technique, we propose a new variant of the EE algorithm, which integrates bagging with boosting and is a type of ensemble method tailored to the class-skew problem. Ensemble techniques usually occur amongst the top performers in data mining competitions. Vanhoeyveld and Martens [52] investigated the EE method on a benchmark repository of imbalanced behavioural data sets. They found EE to be superior to a collection of 10 imbalanced learning methods drawing from a variety of undersampling, oversampling and cost-sensitive learning variants in terms of AUC (see Sect. 4.3). Similarly, Dal Pozzolo et al. [41] demonstrate EE to outperform the techniques of random undersampling majority instances and the SMOTE synthetic minority oversampling approach in terms of AUC for a credit card fraud detection problem with traditional data. Furthermore, the aforementioned references demonstrate the scalability of the EE

[14] Extracted from http://ec.europa.eu/taxation_customs/dds2/taric/taric_consultation.jsp?Lang=en.

[15] The representative and intra-community acquirer occur far less frequently. Also note that the identity of the consignor is unknown in an import declaration.

[16] Each row therefore contains four ones. This time we consider the entities simultaneously which allows interaction effects to be revealed. In the case of high-cardinality variables, each attribute is treated separately. The main difference lies in the modelling.

algorithm. The method is very fast because each subset is only twice as large as the size of the minority class training data and each subset can be trained and evaluated in parallel. As outlined in Sect. 1, scalability is important.

4.2.2 Support vector machines (SVM)

The SVM is a popular state-of-the-art regularization-based technique that has been frequently applied in applications dealing with traditional kinds of data [2, 44] and sparse and high-dimensional data sets [10, 20, 21, 36]. Below, we briefly introduce this method. For a more detailed (theoretical) description, we refer to the work of Suykens et al. [50].

Given a collection of training instances $\{(x_i, y_i)\}_{i=1}^{N}$, with d-dimensional input data $x_i \in \mathbb{R}^d$ and their associated labels $y_i \in \{-1, 1\}$, the SVM is the solution of the following convex optimization problem (primal) [50]:

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{w^T w}{2} + C \sum_{i=1}^{N} \xi_i \\ &\text{s.t.} \quad y_i\left(w^T \varphi(x_i) + b\right) \geq 1 - \xi_i, i = 1, \ldots, N \\ &\qquad \xi_i \geq 0, i = 1, \ldots, N, \end{aligned} \tag{1}$$

where the learned classifier is represented as $y(x) = w^T \varphi(x) + b$ (the output (fraud) score assigned by the inducer corresponds to $y(x)$). Here, $\varphi(x)$ is a mapping from the input space to a (possibly) high-dimensional feature space. Hence, the SVM constructs a linear hyperplane in the new feature space $\varphi(x)$, which may result in a nonlinear classification in the original feature space x. w represents the model weights (coefficients) of the new features $\varphi(x)$ and b represents a bias term. Note that the mapping $\varphi(x)$ is usually not explicitly calculated (due to computational requirements), but one relies on the "kernel trick" as presented in the next paragraph. In the SVM formulation (1), ξ_i are the slack variables measuring classification errors for the training instances. The goal function represents a trade-off between maximizing the margin[17] and minimizing training set errors ($\sum \xi_i$). This is governed by the choice of regularization[18] parameter C.

One usually solves the dual formulation of primal problem (1), with solution $y(x) = \left[\sum_{i=1}^{N} \left(\alpha_i y_i K(x, x_i)\right) + b\right]$ with dual variables (support values) α_i. With respect to the choice of kernel $K(x, z) = \varphi(x)^T \varphi(z)$ (kernel trick), we opt for the linear kernel $K(x, z) = x^T z$ and the radial basis function (RBF) kernel $K(x, z) = \exp\left(-\gamma ||x - z||^2\right)$ (nonlinear, with γ a kernel parameter) in the case of traditional data. When dealing with the sparse matrix representations of behavioural or high-cardinality data, we only apply a linear SVM due to the fact that nonlinear methods are too computationally involved and do not provide better generalization performance than linear methods [10].

4.2.3 EasyEnsemble

The EE algorithm was proposed by Liu et al. [32] and benefits from a combination of bagging and boosting: the method randomly samples a number of balanced subsets (S in total) and feeds them to the AB boosting process (T boosting rounds). Bagging refers to the creation of balanced subsets, where in each subset majority class instances are randomly selected such that the number of majority class instances equals the number of minority class examples. Boosting is a sequential process, where the goal is to combine several base classifiers ($h_t(x)$, with t the boosting iteration number) to form an accurate ensemble model ($\sum_{t=1}^{T} \tilde{\alpha}_t h_t(x)$, with $\tilde{\alpha}_t$ a weight that reflects the contribution of classifier $h_t(x)$ to the ensemble model) [45]. In boosting, a weight distribution D_t is maintained (and updated) for all instances of its input data set. This distribution is used when classifier $h_t(x)$ is built (instances i having a high weight $D_t(i)$ are emphasized during model construction). Next, $h_t(x)$ is evaluated on the input data set and the weights D_t are increased for instances that are wrongfully classified and decreased for correct classifications, resulting in a new distribution D_{t+1} (that is used to build classifier $h_{t+1}(x)$). Hence, each classifier $h_{t+1}(x)$ focuses on the errors of the previous learner $h_t(x)$.

The learned classifier $x \rightarrow H(x)$ (the output (fraud) score assigned by the inducer corresponds to $H(x)$) of EE combines the base classifiers $h_{s,t}(x)$ of each subset s and boosting iteration t as follows:

$$H(x) = \left(\sum_{s=1}^{S} \sum_{t=1}^{T} \tilde{\alpha}_{s,t} h_{s,t}(x)\right) s = 1 \ldots S; t = 1 \ldots T, \tag{2}$$

where $\tilde{\alpha}_{s,t}$ represents the learned weights of classifier $h_{s,t}(x)$.

In the original EE version of Liu et al. [32], the base classifiers $h_{s,t}(x)$ are decision trees that are fed to a discrete AB boosting algorithm (where the classifiers $h_{s,t}(x)$ output binary values $\{-1, +1\}$). Vanhoeyveld and Martens [52] recently proposed a version where the base classifiers $h_{s,t}(x)$ consist

[17] The margin denotes the separation between the two classes (i.e. how far are the instances from both classes separated from the learned hyperplane?). Maximizing the margin coincides with minimizing the model complexity $w^T w/2$.

[18] Choosing a too large value for regularization parameter C results in a learner that is too sensitive on the training data (overfitting) and fails to generalize for unseen data. On the other hand, a too small value for C means that large errors can occur for the training data and a too simple model is obtained (underfitting) that is unable to distinguish between both classes.

of a weighted linear SVM and subsequent LR[19] that is integrated in a confidence-rated boosting algorithm [45] (where classifiers $h_{s,t}(x)$ output real-valued scores in $[-1,+1]$. A high score close to 1 means the classifier is fairly confident that the label should be +1). They tested this version on a benchmark repository of imbalanced behavioural data sets. We refer to Vanhoeyveld and Martens [52] for an in-depth discussion. We highlight a number of advantages of their formulation next. Schapire et al. [45] note a confidence-rated AB process to outperform a discrete version. Furthermore, an instance-weighted SVM (instead of a standard SVM or decision tree base classifier) enables the direct inclusion of the weight distribution D_t of the AB process during the learning stage of $h_{s,t}(x)$ [52].

In this study, we adopt the aforementioned linear SVM-based EE formulation [52] when dealing with behavioural or high-cardinality data types. The authors made use of the LIBLINEAR toolbox [18] that offers a linear SVM implementation tailored for high-dimensional and sparse data sets. In this work, a new EE implementation is developed suitable for traditional data sets. The bagging and boosting (see Appendix 2) component of EE is identical to the version of Vanhoeyveld and Martens [52]. The only difference is the choice of base classifier $h_{s,t}(x)$, where we rely on an instance-weighted linear or nonlinear SVM and subsequent LR. The SVM classifiers are constructed with the LIBSVM toolbox [6].

4.2.4 High-cardinality attributes

In this section, the transformation of a high-cardinality attribute into a single continuous feature whose values are correlated with the target label (i.e. fraud or legal) is outlined. These methods differ from the ones used in prior studies (see Sect. 2) in the sense that information on the label is integrated in the methodology.

Inspired from a social network perspective, the supervised ratio (SR) includes information on the proportion of fraud within each category of a high-cardinality variable. Let X be such a variable, then the supervised ratio for category (value) j of X is defined as follows [36]:

$$\mathrm{SR}_j^X = \frac{F_j^X}{F_j^X + L_j^X}, \tag{3}$$

where F_j^X and L_j^X represent the number of fraudsters and compliant entities that show a value of j for attribute X, respectively, in the training data set.[20] In case F_j^X and L_j^X both equal zero, then the SR score corresponds to the average fraud rate in the training set $F/(F+L)$. When an instance shows category j for high-cardinality attribute X, we assign its associated SR value SR_j^X that can be used in the construction and application of a final prediction model.

In retrospect, though the SR method is very intuitive, there are some stability concerns with the basic formulation. If a category is rarely observed, then its associated SR value can change drastically in case a new instance of that category is added to the training set. We therefore propose the following correction (smoothing) in the calculation of the supervised ratio:

$$\mathrm{SR\_Corr}_j^X = \frac{F_j^X + 1}{(F_j^X + 1) + \left(L_j^X + \frac{L}{F}\right)}. \tag{4}$$

If a category is frequently observed (i.e. $(L_j^X, F_j^X) \to \infty$), the $\mathrm{SR\_Corr}_j^X \to SR_j^X$. When there are few training observations in category j (i.e. $(L_j^X, F_j^X) \to 0$), then $\mathrm{SR\_Corr}_j^X \to F/(F+L)$. In other words, when there is much evidence for a particular category, the corrected SR approximates the standard SR. In case there is little evidence, we are more unsure and correct the SR in the direction of the fraud rate (base rate) $F/(F+L)$.

A final way of including a high-cardinality variable entails the pre-training of a SVM or EE SVM with linear kernel on the training data of the associated sparse matrix representation of the attribute. This model can be used to assign a score to the attribute under consideration for any instance. In other words, the value of a high-cardinality attribute is replaced by the score it receives under application of a pre-trained predictive model. Note that we use the validation set (see Sect. 4.3.1) for tuning the hyperparameters of the pre-trained model. In case the pre-training phase involves a linear SVM, then the aforementioned procedure corresponds to replacing the high-cardinality category j with the weight (coefficient) w_j of the solution vector w of the SVM [see Eq. (1)]. The latter does not apply in the case of EE with linear SVM, because the logistic regression component makes this learner nonlinear.

4.3 Evaluation

4.3.1 Methodology

The various fields of a SAD form constitute raw input variables x_{raw} to characterize an article involved in a customs

[19] The LR component transforms the real-valued SVM scores $w^T\varphi(x)+b$ to the range $[-1,+1]$ (as required for a confidence-rated boosting algorithm).

[20] In Sect. 5.1.2, we will detail which part of the training data is effectively used for calculating the SR values.

declaration. These features can be categorized into three types according to their definition set out in Sect. 2: traditional, high-cardinality and behavioural data. Obviously, the raw features x_{raw} as illustrated in Table 1 require pre-processing which results in the final feature representation x. Traditional data are processed by means of a dummy encoding for categorical variables and a standard statistical normalization for continuous attributes. High-cardinality features are transformed using any method outlined in Sect. 4.2.4. Behavioural data are pre-processed by representing them as a large and sparse matrix (see the related commodity codes example in Sect. 2). Hence, each article i is represented by its feature vector x_i that depends on the kind of data that are included. A classification model is trained based on the subset of labelled instances (articles with known fraud/legal indications). This model can subsequently be used to assign a fraud score $y(x)$ to any article based on its feature representation x.

The methodology is shown in Fig. 1. SVMs with linear (LIN) and nonlinear (RBF) kernels and their integration with EE are first applied to the three data types individually in Sect. 5.1. Next, these methods are used to build data ensemble models which combine the data sources through a stacking approach as explained in Sect. 5.2. Their performances are assessed with the evaluation metrics proposed in Sect. 4.3.2.

We rely on a standard stratified fivefold cross-validation to evaluate the proposed methodology, where we ensured that each test data point occurs only once across all folds. Within each fold, the collection of labelled instances is divided into training data (60%), validation data (20%) and test data (20%). The training data are used for building predictive models, the validation data are used for hyperparameter tuning, and the test data represent the generalization performance on hold-out data. The results shown in Sect. 5 are the average and standard deviation of the evaluation metrics applied to the test data of each fold.

With respect to design choices, the following parameter settings were chosen:

- $C = \left[10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}\right]$ (SVM)
- $\gamma = \left[2^{-9}, 2^{-6}, 2^{-3}, 2^{0}, 2^{3}, 2^{6}\right]$ (SVM)
- $S = 12$ & $T = 10$ (EE).

Note that the boosting iteration number $t \in [1, \ldots, T]$ is considered to be a tunable parameter.

4.3.2 Performance metrics

The area under the receiver operating characteristic (AUC) [19] measures the global ranking abilities of the model and is a suitable metric for performance evaluation of imbalanced data sets. Its easy statistical interpretation,[21] insensitivity on class skew and independence on the choice of threshold make this an appealing metric. AUC is also the preferred criterion in the assessment of unsupervised anomaly detection techniques [4, 23]. Imbalanced data and anomaly detection are related in the sense that outliers correspond to rare events of interest, just like the minority class.

Lift values [42] evaluate the occurrence of fraud cases in the top list (with the most suspicious cases) produced by a scoring mechanism and can be evaluated according to the available capacity. If we look at the $p\%$ most suspicious cases (the highest fraud scores), corresponding to a top N list of size $N = p\% \cdot T$, with T the total number of instances, then the lift can be defined as:

$$\text{lift}(p\%) = \frac{F_N}{F} \Big/ \frac{p}{100} = \frac{F_N}{F} \Big/ \frac{N}{T} = \frac{F_N}{N} \Big/ \frac{F}{T}. \tag{5}$$

In this equation, F_N represents the number of fraud cases in a top N list and F is the overall number of fraud cases in T. In Sect. 5, we have chosen to evaluate the predictive models at an arbitrary capacity value of $p = 1\%$ to mimic a real-life scenario where customs can only target a limited amount of incoming declarations.

Though lift values are appealing from a practical perspective, there is an important caveat. In a tax fraud detection context, customs officers can make mistakes and some fraud cases remain undetected (i.e. labelled compliant but in fact fraudulent). If the classification method is any good, this phenomenon is expected to occur precisely in the top list. Lift values are therefore less reliable than AUC, especially when evaluated for low values of p in Eq. (5). The AUC is more robust against wrongfully labelled instances because it ranges over all possible thresholds.

5 Results and discussion

5.1 Individual data sources

5.1.1 Traditional data

The results of the framework detailed in Sect. 4.3.1 are presented for traditional data in this section. From a data perspective this should be regarded as the baseline as the current literature is limited to this type of data. The input data contain a total of 76 variables, of which the vast majority are nominal variables (with fewer than 100 categories). After

[21] The AUC corresponds to the probability that a positive instance (fraud) is ranked higher than a negative instance (compliant). The ranking is obtained by sorting the instances according to the output scores produced by the classifier.

Table 3 Traditional data with feature selection (FS, top 20% of features)

Data	SVM_LIN	SVM_RBF	EE_LIN	EE_RBF
Trad_FS	62.3 (3.8)	74.1 (0.7)	76.4 (0.6)	**80.6 (0.4)**
Trad_FS	2.8 (0.9)	**15.7 (0.7)**	6.3 (0.6)	14.8 (1.4)

Results showing the average AUC (top) and lift ($p = 1\%$) (bottom) performance on test data across all fivefold. Standard deviations are included between brackets. Best performances are indicated in boldface

Table 4 Extending the results of Table 3 with a decision tree (Dec. Tree) and neural network (NN) base classifier

Data	Dec. Tree	NN
Trad_FS	74.0 (0.9)	**77.7 (0.8)**
Trad_FS	**15.9 (1.7)**	9.6 (1.0)

Best performances of average AUC (top) and lift ($p = 1\%$) (bottom) are indicated in boldface

applying dummy encoding on these variables, the final representation still has 1013 features. To reduce computational efforts with respect to training and prediction, we conduct a feature selection based on a simple t statistic on the training data. Ravisankar et al. [44] explain the details of this method and apply this technique in the related area of financial statement fraud detection. We should note, however, that more advanced methods do exist; see the work of Guyon and Elisseeff [24] for an introduction. Experiments with different feature percentages show that predictive performance (in terms of AUC) starts dropping when using less than 10% of all features. The AUC is approximately constant in the range of 10−100%. To allow some margin, we decided to include the top 20% of features in the final prediction model.[22] Note that many fields in the SAD form are optional; hence, they contain many empty values and are less informative. This explains why high levels of feature selection do not harm predictive performance.

Table 3 contains the results on traditional data with the top 20% of features. The EE technique can severely improve AUC performance when compared to the plain application of a SVM. This highlights the importance of dealing with imbalance. In terms of lift($p = 1\%$), the EE dominates in the case of a linear kernel, though it is slightly inferior when using a RBF kernel. The best lift at $p = 1\%$ has a value of 15.7. The base rate of fraud $\frac{F}{T}$ is 3.735% in the test set. Hence, according to Eq. (5), $\frac{F_N}{N} = 15.7 \times 3.735\% = 58.64\%$. This result shows the hit rate (precision) is very high in the top 1% list.

We provide additional results on traditional data in Table 4 with a decision tree[23] and neural network[24] base classifier, which are popular inducers in the customs fraud detection literature (see Sect. 2). The SVM_RBF achieves similar results in terms of AUC in comparison with a decision tree and a neural network. For the case of lift, it is similar to a decision tree and outperforms a neural network. It is clear that a SVM with nonlinear kernel is an attractive technique for this customs fraud detection problem. Also observe that the EE_RBF achieves the highest AUC amongst all methods tried and this again points to the importance of dealing with the imbalanced learning issue.

5.1.2 High-cardinality variables

The high-cardinality variables are included using the methods highlighted in Sect. 4.2.4. They can be regarded as a pre-processing phase that transforms the data into low-dimensional features (traditional data). Predictive models (e.g. SVM_LIN, SVM_RBF, EE_LIN, EE_RBF) can thereafter be trained on the resulting representation.

Moeyersoms and Martens [36] note that the supervised ratios should be calculated on a separate part of the training data to avoid overfitting. We included the option Split and divided the training data into two equally sized parts. 50% of the training data are used to calculate the SR values (or pre-train (EE) SVMs). The final prediction model is trained on the remaining 50% of the training data. However, learning curve analysis for behavioural or high-cardinality data sets reveals that including more data (in terms of the number of instances or the number of features) improves the predictive performance [10, 36], which is not the case when dealing with traditional kinds of data. In this high-dimensional and sparse setting, including more instances could overpower the negative effects related to overfitting. We therefore included

[22] Within each fold, the feature selection based on t statistic is computed on the training data. The set of 'optimal' features can therefore differ in each fold.

[23] We make use of the standard MATLAB function *fitctree*, see https://nl.mathworks.com/help/stats/fitctree.html, which fits a classification decision tree making binary splits. Default parameter settings are adopted. The split criterion is a hyperparameter that can take on Gini's diversity index or maximum deviance reduction (cross entropy). The MinLeafSize (minimum number of leaf node observations) is another hyperparameter that controls for overfitting. The following values were imposed: MinLeafSize= 2^z, with $z = [1;1.5;2;2.5;\ldots;5.5]$.

[24] The standard MATLAB function *patternnet*, see https://nl.mathworks.com/help/deeplearning/ref/patternnet.html, is used to construct a classification neural network with one hidden layer (sigmoid transfer function). Default parameter settings are adopted for the optimization algorithm (scaled conjugate gradient) and performance function (cross-entropy). The number of hidden neurons is a hyperparameter taking on values [5;10;15;20; … ;100]. We trained the neural network, with a given number of hidden neurons, for 10 times on the training data and selected the one with the best validation set performance (a neural network converges to a local optimum).

Table 5 High-cardinality data (OperID) under various pre-processing options, including the standard SR (see Eq. 3), the smoothed supervised ratio SR_Corr (see Eq. 4) and the pre-training of (EE)_SVMs on the sparse matrix representations of the attributes

Pre-process	Source	SVM_LIN	SVM_RBF	EE_LIN	EE_RBF
SR_F	OperID	77.9 (0.8)	76.5 (0.3)	79.1 (0.6)	79.1 (0.6)
SR_S	OperID	75.8 (2.0)	74.9 (2.0)	76.7 (1.3)	77.6 (0.9)
SR_Corr_F	OperID	80.1 (0.8)	78.5 (3.8)	80.3 (0.5)	**80.4(0.5)**
SR_Corr_S	OperID	78.2 (0.8)	76.8 (0.9)	78.2 (0.9)	78.9 (0.8)
(EE)_SVM_F	OperID	68.5 (0.9)	68.2 (0.9)	79.8 (0.6)	79.8 (0.6)
(EE)_SVM_S	OperID	66.9 (3.3)	65.4 (5.2)	78.0 (1.1)	78.5 (0.8)
SR_F	OperID	9.5 (1.0)	9.5 (1.4)	10.1 (0.9)	9.7 (1.5)
SR_S	OperID	7.3 (1.4)	9.2 (1.7)	8.9 (1.1)	10.7 (1.4)
SR_Corr_F	OperID	9.5 (1.4)	10.5 (1.3)	10.1 (0.8)	9.5 (0.9)
SR_Corr_S	OperID	8.6 (0.5)	8.7 (0.9)	8.7 (0.8)	**10.9(0.9)**
(EE)_SVM_F	OperID	3.3 (2.9)	8.0 (1.6)	10.1 (1.2)	10.2 (1.4)
(EE)_SVM_S	OperID	6.3 (1.0)	4.8 (1.1)	8.5 (1.2)	10.0 (1.1)
SR_F	All	81.2 (0.9)	79.2 (0.7)	81.9 (0.6)	81.8 (0.6)
SR_S	All	75.6 (1.4)	75.2 (1.9)	79.7 (0.9)	80.3 (1.1)
SR_Corr_F	All	79.9 (1.2)	79.3 (1.1)	**82.5 (0.6)**	**82.5 (0.6)**
SR_Corr_S	All	74.8 (0.6)	72.1 (2.4)	80.9 (1)	81.3 (0.9)
(EE)_SVM_F	All	69.4 (1.3)	68.9 (0.5)	82.1 (0.5)	82.0 (0.5)
(EE)_SVM_S	All	61.9 (4.6)	65.2 (3.7)	80.4 (1.0)	80.9 (0.8)
SR_F	All	9.8 (1.3)	13.0 (1.0)	10.8 (0.9)	12.7 (1.2)
SR_S	All	9.2 (2.0)	12.1 (1.2)	10.8 (1.2)	12.3 (1.4)
SR_Corr_F	All	10.3 (0.5)	**13.3 (1.2)**	11.4 (1.0)	12.7 (0.8)
SR_Corr_S	All	8.6 (1.2)	12.1 (1.1)	10.5 (1.3)	12.1 (1.6)
(EE)_SVM_F	All	6.8 (2.1)	8.9 (1.0)	10.5 (1.2)	12.3 (1.2)
(EE)_SVM_S	All	5.5 (0.6)	6.8 (1.3)	9.2 (0.9)	10.5 (1.0)

As a convention, the EE_SVM is used in the pre-training stage of the final prediction model EE_LIN or EE_RBF. The standard SVM is used in the pre-training stage when the final learner is SVM_LIN or SVM_RBF. S denotes the Split version (50−50% training data split) and F corresponds to the Full version (no training data split) as explained in Sect. 5.1.2. Results showing the average AUC and lift ($p = 1\%$) performance on test data across all fivefold with standard deviations in brackets. Best results per row are underlined, and the optimal performance over all pre-processing options is indicated in boldface

the option Full, where the entire training data set is used both in the calculation of the SR values [or pre-train (EE) SVMs] and in the construction of the final prediction model.

In Table 5, the predictive performance is shown for the four entities involved in a declaration (see Sect. 4.1, denoted by OperID) and all high-cardinality attributes (All). Note that these attributes appear to be very predictive as we are able to outperform traditional data in terms of AUC with much fewer variables (however, the lift values appear worse). Based on these findings, we can draw the following general[25] conclusions: (1) The EE version always outperforms the plain SVM application in terms of AUC. For the case of lift, this situation is confirmed in the vast majority of cases, though the highest lift on all high-cardinality data is achieved with the SVM_RBF. (2) The AUC of the Full version always dominates the AUC of the Split version. The lift of Full is almost always larger than the lift of Split. (3) The AUC of SR_Corr is dominating the AUC of the standard SR in the vast majority of cases.[26] The analysis of lift does not reveal a clear winner (tie). (4) In comparing the pre-trained EE_SVM with the SR, we observe that the final prediction model with EE version of the former always outperforms the EE version of the latter in terms of AUC. In terms of lift, there is a slight preference for the SR pre-processing. However, the standard pre-trained SVM version has a much lower AUC and lift than the SR pre-processing in case the final learner is a plain SVM. SVMs suffer more from the imbalanced learning issue than SR and are known to have a limited number of support vectors, which for high-cardinality data implies that many categories have zero weight (i.e. the solution vector w is sparse). This is not recommended because for this kind of data, each feature provides a small though relevant amount of additional information [10, 20].

5.1.3 Behavioural data

The two types of behavioural data occurring in this study are explained in Sect. 4.1. For reasons outlined in Sect. 4.2, we only apply a linear kernel to the sparse matrix representations of these types of data. The first row and second row of Table 6 show the results for the behavioural data CommCode and OperID, respectively. Again, the EE_LIN improves the predictive power (AUC and lift) when compared to a plain SVM_LIN. The results clearly indicate that there is a lot of value in these types of behavioural data. Apparently, *customs fraud seems to occur for certain kinds of goods and committed by certain operators.*

The last two rows of Table 6 represent the results when combining the two kinds of behavioural data. In a first pre-training stage, (EE)_SVMs with linear kernels are trained on the training data of each behavioural data source individually and are used to score all instances. In a second stage, a final predictive model is learned where the input corresponds to

[25] Based on counting the wins/losses/draws in comparing several methods. For example, in comparing the F with the S version, a pair (F,S) is formed with the same data set (2 types), the same type of pre-processing (3 types) and the same final model (4 types). This leads to checking a total of 24 pairs.

[26] In the case of OperID data, this is always the case. These are precisely the attributes with the highest cardinalities for which we expect stability issues to occur.

Table 6 Predictive performances for behavioural data considering the commodity codes of all articles (CommCode) and the four operators (OperID) involved in a declaration

Data	SVM_LIN	SVM_RBF	EE_LIN	EE_RBF
CommCode	67.2 (0.7)	X	78.5 (0.8)	X
OperID	67.9 (0.8)	X	80.5 (0.7)	X
All_F	71.8 (1.0)	71.7 (1.1)	**83.5 (0.3)**	83.4 (0.4)
All_S	70.3 (0.9)	70.1 (1.0)	80.5 (0.8)	80.5 (0.7)
CommCode	11.9 (1.2)	X	12.1 (1.5)	X
OperID	10.0 (0.8)	X	10.6 (0.6)	X
All_F	12.2 (1.9)	12.4 (2.1)	13.8 (1.0)	**14.0 (1.3)**
All_S	9.6 (2.3)	10.2 (2.4)	12.7 (0.9)	12.7 (0.6)

The combination of these behavioural data sources (All) is obtained by pre-training (EE)_SVMs with linear kernels on their sparse matrix representations and including the two output scores in a final prediction model. As a convention, the EE_SVM is used in the pre-training stage of the final prediction model EE_LIN or EE_RBF. The standard SVM is used in the pre-training stage when the final learner is SVM_LIN or SVM_RBF. S denotes the Split version (50−50% training data split) and F corresponds to the Full version (no training data split) as explained in Sect. 5.1.2. Results showing the average AUC (top) and lift ($p = 1\%$) (bottom) performance on test data across all five-fold with standard deviations between brackets. Best performances per row are underlined and overall best performances are indicated in boldface

the two output scores generated in the previous step (i.e. one output score for each behavioural data source). In this set-up, we again investigate whether or not it is useful to conduct the first stage on a separate part of the training set. The results confirm once again that the Full version improves the performance over a Split version. Furthermore, the EE version outperforms a plain SVM. Also, the combination of the behavioural data sources is more valuable than each data source individually.

5.2 Combining data sources

The combination of the different data sources in a data ensemble model through stacking is illustrated in Fig. 3. High-cardinality data are included by replacing the value of a high-cardinality attribute by its associated smoothed SR value (SR_Corr, see Eq. 4) calculated on the training data set. Behavioural attributes are incorporated by pre-training a predictive model (SVM_LIN or EE_LIN) on the sparse matrix representation of the training data. The output scores of these pre-trained models serve as input features to a final predictive model that is constructed based on the training set. Note that we again select the 20% most predictive features prior to final model building.

Figure 4 shows the predictive performances for the linear SVM, which is the most comprehensible method, for each data source individually and their combination. We can clearly see that *including behavioural or high-cardinality data sources results in significant performance improvements compared to traditional data (the baseline)*. Figure 5 extends these results to the other inducers (SVM_RBF,

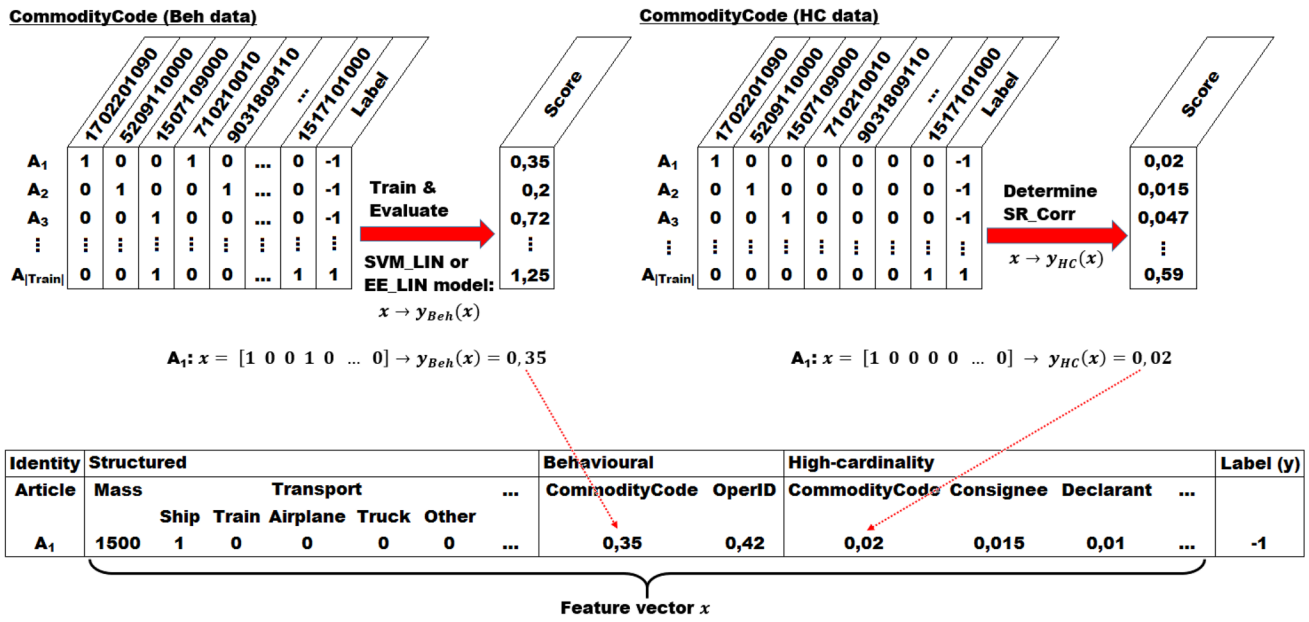


Fig. 3 Data ensemble stacking methodology. Behavioural data (i.e. commodity codes of all articles in the same declaration) are included by pre-training a predictive model on the sparse matrix representation of the training data and using this model to assign a score to any article (i.e. A_1) under consideration. High-cardinality data (i.e. the commodity code of an article) are included by replacing the value (i.e. '1702201090') of the attribute by its corresponding SR_Corr (see Eq. 4) value (i.e. 0, 02). The SR_Corr values are determined on the training data

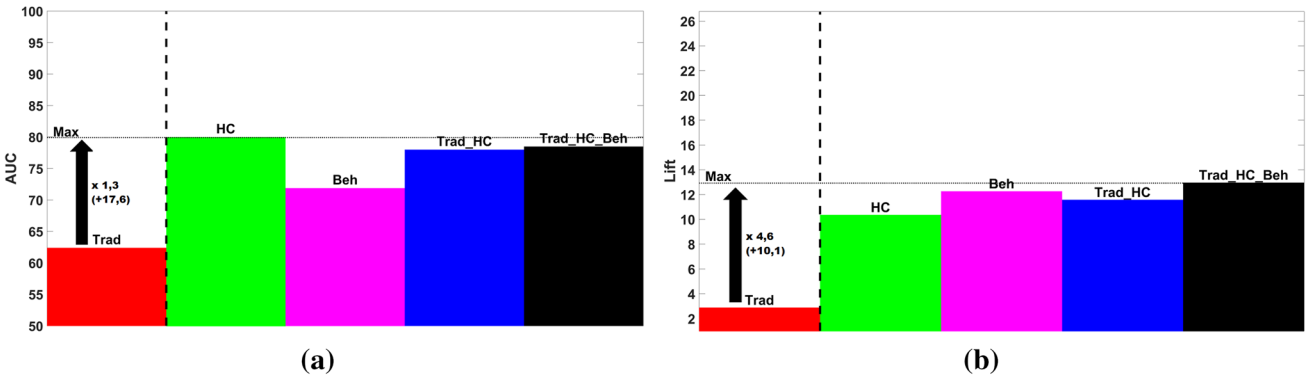


Fig. 4 Assessing the predictive power for customs fraud detection of traditional (Trad), High-Cardinality (HC), Behavioural (Beh) data individually as well as their combination in a stacked data ensemble model (Trad_HC and Trad_HC_Beh) under application of a linear SVM for **a** average AUC and **b** lift ($p = 1\%$) on test data across all fivefold

EE_LIN, EE_RBF). We refer to Appendix 3 for a tabular version thereof. Note that we show outcomes with AUC > 70 and lift > 9.5. This filters out the 'poor' results of traditional data under application of a linear SVM; see Table 3.

When comparing the plain SVM model to its associated integration with EE, we note the AUC performance of EE to be superior in all cases. In terms of lift, the EE almost always outperforms the SVM in case of a linear kernel. Only when using a RBF kernel, the EE version is inferior to the SVM model. The latter can possibly be explained as the boosting process emphasizes the hard to learn instances, which are not expected to occur in the top 1% list. Additionally, when some majority class subsets contain wrongfully labelled legal cases (i.e. presumed legal, but in fact fraudulent), then boosting attempts to make them appear lower in the list, dragging along neighbouring fraud cases. This effect should be minor as EE filters out majority class noise cases.

The high AUC revealed with high-cardinality or behavioural data, improving on the AUC of traditional data, point to their high predictive value even though the lift of traditional data can be larger in some cases. Also note that the number of high-cardinality variables (14) and behavioural features (2) is much smaller than the number of traditional attributes (76). As expected, the highest predictive performance in terms of both evaluation measures is obtained when combining all data sources in a final data ensemble model. This demonstrates the complementarity of the different data sources.

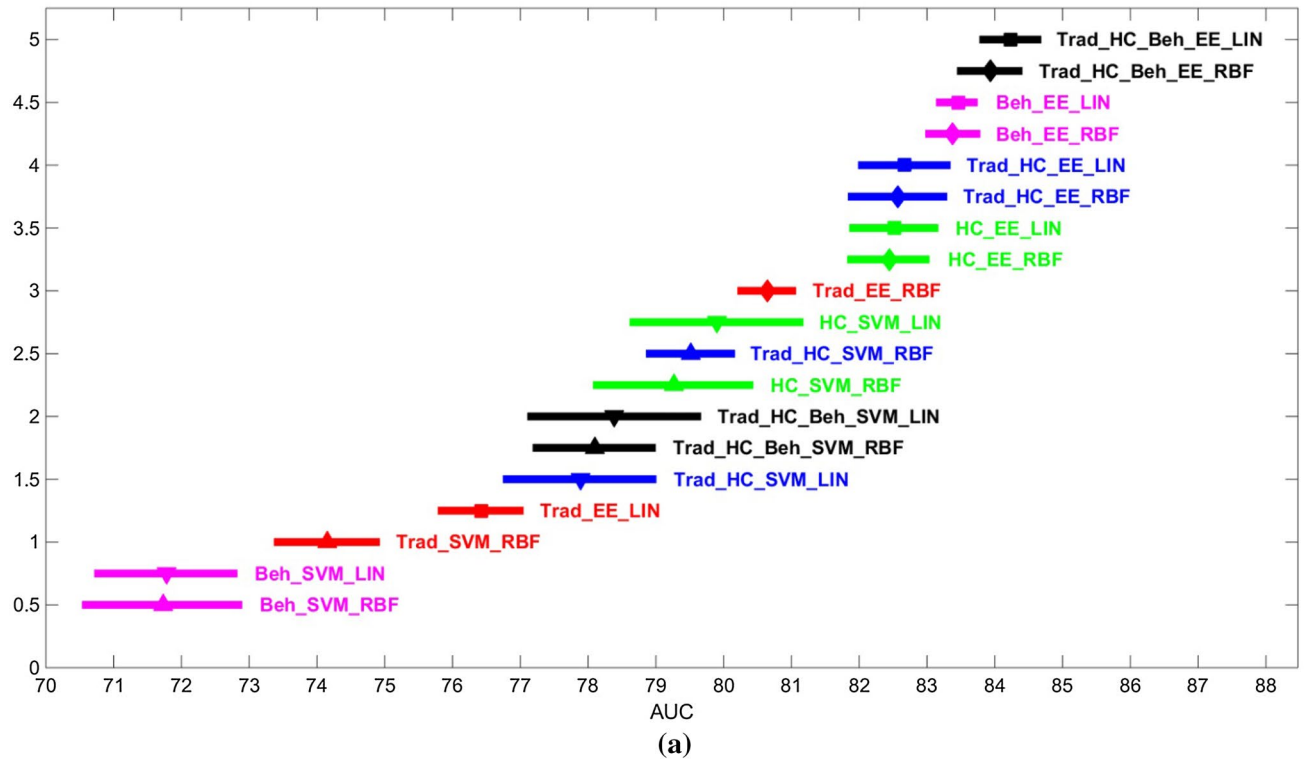
6 Conclusions

In this paper, a supervised customs fraud detection application was developed. We carefully assessed the predictive value of behavioural and high-cardinality data types in terms of AUC and lift performance and compared it to the kind of data that have traditionally been used, taking into account the class imbalance.

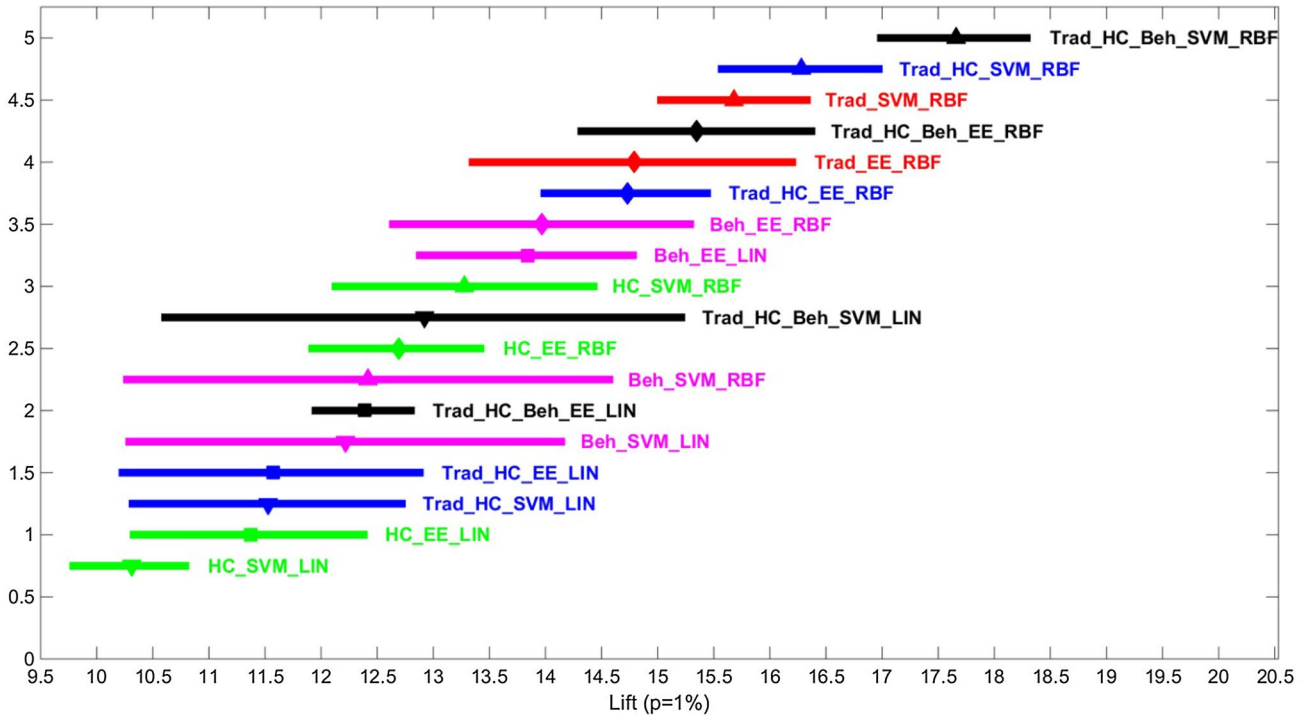
The class-skew problem requires careful consideration. We developed a confidence-rated support vector machine-based version of the EasyEnsemble algorithm. Results with respect to AUC show this method to outperform the plain application of a SVM. In terms of lift, the same conclusion can be reached when dealing with linear kernels.

Our experiments suggest that such fine-grained data at identifier level (e.g. consignee, declarant, type of commodity, country of origin, etc.) are very predictive for customs fraud and can even outperform classical kinds of data that contain many more variables. In hindsight, this comes as no surprise since, for instance, valuation fraud is an important fraud type. The value of a good precisely depends on such fine-grained information. We expect the impact on the customs fraud detection domain to be large as our analysis revealed high hit rates ($> 65\%$ in the top $p = 1\%$ list) that immediately translate into an improved recovery of financial losses and an enhanced deterrence. Furthermore, the proposed methodology could be adopted by customs administrations worldwide as they have access to similar data repositories as presented in this study.

Apart from an improved fraud detection capability, the finding that fine-grained features are highly predictive could have a number of operational and managerial implications. (1) Our proposed methodology (e.g. smoothed supervised ratio) enables customs to maintain a list of high-risk operators and commodities which can aid selection officers in the screening process. (2) This could lead to certain policy changes. In terms of high-risk goods, preventive measures could be undertaken (e.g. more thorough sealing requirements, imposing different transportation conditions, etc.). With respect to operators, EU customs currently adopt an AEO system [15] to certify (under strict conditions)



(a)



(b)

Fig. 5 Combining the data sources into data ensemble models. Additionally, the best performances of the individual data sources are repeated for convenience. Results showing the **a** AUC and **b** lift ($p = 1\%$) performance across fivefold of test data

Fig. 6 SAD declaration form retrieved from [13]

compliant operators, which has many benefits (e.g. less control of their goods). The automated procedure to assign a risk score to each operator can be helpful in checking whether an AEO is really complying with the formal standards. In a dynamic fraud environment, it is not excluded that AEOs change their behaviour to non-compliant at some point in future. Our methodology allows changing risk scores over time by updating the models.

In reference to this project, Kristian Vanderwaeren, administrator general of the Belgian Customs and Excise, reported: "The collaboration with the University of Antwerp enables the General Administration Customs and Excise to integrate certain data sources which were until now only being used in a far less automated manner. This should lead to more efficient fraud detection models and could additionally allow us to predict some relatively rare type of infringements such as on matters of (product) safety, environment and health. The cooperation allows the University of Antwerp to test their newest methods in a non-artificial and socially relevant context."

In terms of future research directions, we note that anomaly detection techniques are rarely applied. Supervised methods are limited to the discovery of known fraud types (at classification time). Outlier detection techniques are able to target new types of non-compliant behaviour since, by assumption, fraudsters engage in a behaviour that deviates from the norm [12, 43]. In a dynamic fraud environment, a methodology should be developed that integrates supervised classification with outlier detection, random targeting and active learning. Besides the data available in a customs declaration, other types of data could be included in the predictive modelling effort which poses a challenge, such as image data (e.g. container scanning) or sensor data (e.g. temperature measurements of goods stored in containers). Also, more advanced feature selection or dimensionality reduction methodologies [24] exist than the t-statistic-based approach [44] adopted in this work. One may consider to employ more useful features to deal with the fraud detection task, such as the works of Zhang et al. [58] and Zheng et al. [59].

Acknowledgements The authors would like to thank the Belgian Federal Public Service Finance division Customs and Excise for the provision of the data sets and their involvement throughout the project. The models described in this paper are not necessarily the ones used by the Belgian customs administration. Funding was provided by University of Antwerp (Grant No. DOCPRO4/Antigoon PS-IDnr. 29648).

Appendix 1: Single administrative document form

A blank SAD declaration form [13] is provided in Fig. 6. In Belgium, customs declarations are filed electronically by means of the PaperLess Douane en Accijnzen (PLDA) application.

Appendix 2: AdaBoost

Algorithm 1 presents the underlying AB boosting process for the EE technique that we have presented in Sect. 4.2.3.

Algorithm 1 AdaBoost with a SVM-LR combination as a base learner

Input: $(X,Y) = (x_1,y_1),...,(x_N,y_N); C, \gamma$ (optional) $, T$
Initialize distribution: $D_1(i) = 1/N,\ i = 1,\ldots,N$
for $t = 1$ **to** T **do**

- train base classifier $h_t(x)$ using distribution D_t. The base learner consists of instance weighted SVM and subsequent LR.

$$h_t \leftarrow Train\_WeakLearner(X,Y,D_t,C,(\gamma))$$

- compute the weighted confidence r_{AB} on the training data

$$r_{AB} \leftarrow \sum_{i=1}^{N} D_t(i) y_i h_t(x_i)$$

If $(r_{AB} = 1 \parallel r_{AB} \leq 0)$ **then** $\alpha_t \leftarrow 0$ and stop the boosting process

- choose $\alpha_t \in \mathbb{R}$

$$\alpha_t \leftarrow \frac{1}{2}\log\left(\frac{1+r_{AB}}{1-r_{AB}}\right)$$

- update distribution

$$D_{t+1}(i) \leftarrow \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

end for
Output the final hypothesis:

$$H(x) = \left(\sum_{t=1}^{T} \tilde{\alpha}_t h_t(x)\right) \text{ with } \tilde{\alpha}_t = \frac{\alpha_t}{\sum_{i=1}^{T}\alpha_i}$$

Appendix 3: Data ensembles

Table 7 presents the results shown in Fig. 5 in a tabular format.

Table 7 Predictive performances for each data source individually and their combination

Data	SVM_LIN	SVM_RBF	EE_LIN	EE_RBF
Trad_FS	62.29 (3.8)	74.15 (0.73)	76.42 (0.58)	80.64 (0.38)
HC	79.9 (1.23)	79.27 (1.14)	82.52 (0.61)	82.45 (0.57)
Beh	71.78 (1.01)	71.73 (1.14)	83.46 (0.27)	83.38 (0.35)
Trad_HC_FS	77.89 (1.09)	79.51 (0.6)	82.67 (0.63)	82.57 (0.68)
Trad_HC_Beh_FS	78.39 (1.23)	78.1 (0.86)	**84.23(0.4)**	83.93 (0.43)
Trad_FS	2.83 (0.86)	15.68 (0.65)	6.26 (0.63)	14.79 (1.44)
HC	10.31 (0.52)	13.28 (1.15)	11.37 (1.04)	12.69 (0.77)
Beh	12.22 (1.93)	12.42 (2.15)	13.84 (0.96)	13.97 (1.33)
Trad_HC_FS	11.53 (1.21)	16.28 (0.71)	11.57 (1.34)	14.73 (0.74)
Trad_HC_Beh_FS	12.92 (2.31)	**17.66(0.67)**	12.39 (0.44)	15.35 (1.03)

We refer to Tables 3, 5 (row SR_Corr_F, Source All) and 6 (row All_(EE)_SVM_F) for the outcomes on traditional, high-cardinality and behavioural data, respectively. Regarding data ensembles (Trad_HC and Trad_HC_Beh), high-cardinality data are included by assigning the SR_Corr (see Eq. 4) value to each category of this attribute. Behavioural data are included by pre-training predictive models on the sparse matrix representations and including the output scores as features in the final predictive model. The Full version (no training data split) is adopted. We also employ a feature selection (FS) phase, where the 20% most predictive features according to t test statistic are retained. Results showing the average AUC (top) and lift ($p = 1\%$) (bottom) performance on test data across all fivefold with standard deviations between brackets. Best performances per row are underlined, and overall best performances are indicated in boldface

References

1. Agyemang M, Barker K, Alhajj R (2006) A comprehensive survey of numeric and symbolic outlier mining techniques. Intell Data Anal 10(6):521–538
2. Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to imbalanced datasets. In: Machine learning: ECML 2004: 15th European conference on machine learning, Pisa, Italy, September 20–24, 2004. Proceedings. Springer, Berlin, pp 39–50. https://doi.org/10.1007/978-3-540-30115-8_7
3. Baesens B, Gestel TV, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. J Oper Res Soc 54(6):627–635. https://doi.org/10.1057/palgrave.jors.2601545
4. Campos GO, Zimek A, Sander J, Campello RJGB, Micenková B, Schubert E, Assent I, Houle ME (2016) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Discov 30(4):891–927. https://doi.org/10.1007/s10618-015-0444-8
5. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):15:1–15:58. https://doi.org/10.1145/1541880.1541882
6. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27::27
7. Chawla NV (2005) Data mining for imbalanced datasets: an overview. Data mining and knowledge discovery handbook. Springer, Boston, pp 853–867
8. Closs DJ, McGarrel EF (2004) Enhancing security throughout the supply chain. IBM Center for the Business of Government. http://www.businessofgovernment.org/sites/default/files/Enhancing%20Security.pdf. Retrieved 2 Feb 2018
9. Cruz R, Fernandes K, Costa JFP, Ortiz MP, Cardoso JS (2018) Binary ranking for ordinal class imbalance. Pattern Anal Appl 21(4):931–939. https://doi.org/10.1007/s10044-018-0705-4
10. De Cnudde S, Martens D, Evgeniou T, Provost F (2017) A benchmarking study of classification techniques for behavioral data. Working papers, University of Antwerp, Faculty of Applied Economics
11. Digiampietri LA, Roman NT, Meira LAA, Filho JJ, Ferreira CD, Kondo AA (2008) Uses of artificial intelligence in the Brazilian customs fraud detection system. In: Proceedings of the 2008 international conference on digital government research. Digital Government Society of North America, dg.o '08, pp 181–187
12. Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S (2002) A geometric framework for unsupervised anomaly detection. In: Barbará D, Jajodia S (eds) Applications of data mining in computer security. Springer, Boston, pp 77–101. https://doi.org/10.1007/978-1-4615-0953-0_4
13. European Commission (2003) Commission regulation (EC) no 2286/2003 of 18 December 2003 amending regulation (EEC) No 2454/93 laying down provisions for the implementation of council regulation (EEC) No 2913/92 establishing the community customs code. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02003R2286-20060101. Retrieved 14 Nov 2018
14. European Commission (2016) SAD guidance during the UCC transitional period. https://ec.europa.eu/taxation_customs/sites/taxation/files/guidance_transitional_sad_en.pdf. Retrieved 3 Feb 2018
15. European Commission (2018a) Athorised economic operator. https://ec.europa.eu/taxation_customs/general-information-customs/customs-security/authorised-economic-operator-aeo/authorised-economic-operator-aeo_en#what_is. Retrieved 7 July 2018
16. European Commission (2018b) The combined nomenclature. https://ec.europa.eu/taxation_customs/business/calculation-customs-duties/what-is-common-customs-tariff/combined-nomenclature_en. Retrieved 3 Feb 2018
17. European Commission (2018c) The single administrative document (SAD). https://ec.europa.eu/taxation_customs/business/customs-procedures/general-overview/single-administrative-document-sad_en. Retrieved 3 Feb 2018
18. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874
19. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

20. Junqué de Fortuny E, Martens D, Provost F (2014a) Predictive modeling with big data: is bigger really better? Big Data 1(4):215–226. https://doi.org/10.1089/big.2013.0037
21. Junqué de Fortuny E, Stankova M, Moeyersoms J, Minnaert B, Provost F, Martens D (2014b) Corporate residence fraud detection. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'14. ACM, New York, NY, USA, pp 1650–1659. https://doi.org/10.1145/2623330.2623333
22. García E, Lozano F (2007) Boosting support vector machines. In: 5th international conference machine learning and data mining in pattern recognition, MLDM 2007, Leipzig, Germany, July 18–20, post proceedings. IBaI Publishing, pp 153–167
23. Goldstein M, Uchida S (2016) A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE 11(4):1–31. https://doi.org/10.1371/journal.pone.0152173
24. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
25. Han CR, Ireland R (2014) Performance measurement of the KCS customs selectivity system. Risk Manag 16(1):25–43. https://doi.org/10.1057/rm.2014.2
26. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
27. Kumar A, Nagadevara V (2006) Development of hybrid classification methodology for mining skewed data sets—a case study of indian customs data. IEEE Int Conf Comput Syst Appl 2006:584–591. https://doi.org/10.1109/AICCSA.2006.205149
28. Kumar S, Biswas SK, Devi D (2018) TLUSBoost algorithm: a boosting solution for class imbalance problem. Soft Comput. https://doi.org/10.1007/s00500-018-3629-4
29. Li Q, Mao Y (2014) A review of boosting methods for imbalanced data classification. Pattern Anal Appl 17(4):679–693. https://doi.org/10.1007/s10044-014-0392-8
30. Liu T (2009) Easyensemble and feature selection for imbalance data sets. In: 2009 international joint conference on bioinformatics, systems biology and intelligent computing, pp 517–520. https://doi.org/10.1109/IJCBS.2009.22
31. Liu W, Chawla S, Cieslak DA, Chawla NV (2010) A robust decision tree algorithm for imbalanced data sets. In: Proceedings of the tenth SIAM international conference on data mining, SIAM, Philadelphia, vol 10, pp. 766–777
32. Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B Cybern 39(2):539–550. https://doi.org/10.1109/TSMCB.2008.2007853
33. Martens D, Provost F (2014) Explaining data-driven document classifications. MIS Q 38(1):73–100 10.25300/MISQ/2014/38.1.04
34. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD (2008) Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 21(2–3):427–436. https://doi.org/10.1016/j.neunet.2007.12.031
35. Miguéis VL, Camanho AS, Borges J (2017) Predicting direct marketing response in banking: comparison of class imbalance methods. Serv Bus 11(4):831–849. https://doi.org/10.1007/s11628-016-0332-3
36. Moeyersoms J, Martens D (2015) Including high-cardinality attributes in predictive models: a case study in churn prediction in the energy sector. Decis Support Syst 72:72–81. https://doi.org/10.1016/j.dss.2015.02.007
37. Ngai E, Hu Y, Wong Y, Chen Y, Sun X (2011) The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decis Support Syst 50(3):559–569. https://doi.org/10.1016/j.dss.2010.08.006 (on quantitative methods for detection of financial fraud)
38. Parvin H, Minaei-Bidgoli B, Alizadeh H (2011) Detection of cancer patients using an innovative method for learning at imbalanced datasets. In: Yao J, Ramanna S, Wang G, Suraj Z (eds) Rough sets and knowledge technology. Springer, Berlin, pp 376–381
39. Perlich C, Provost F (2006) Distribution-based aggregation for relational learning with identifier attributes. Mach Learn 62(1):65–105. https://doi.org/10.1007/s10994-006-6064-1
40. Port of Antwerp (2018) 2018 facts and figures. https://www.portofantwerp.com/sites/portofantwerp/files/POA_Facts_and_Figures_2018.pdf. Retrieved 14 Nov 2018
41. Pozzolo AD, Caelen O, Borgne YAL, Waterschoot S, Bontempi G (2014) Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst Appl 41(10):4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026
42. Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media Inc, Sebastopol
43. Rad HA, Arash S, Rahbar F, Rahmani R, Heshmati Z, Fard MM (2015) A novel unsupervised classification method for customs fraud detection. Indian. J Sci Technol 8(35):1–7. https://doi.org/10.17485/ijst/2015/v8i35/87306
44. Ravisankar P, Ravi V, Raghava Rao G, Bose I (2011) Detection of financial statement fraud and feature selection using data mining techniques. Decis Support Syst 50(2):491–500. https://doi.org/10.1016/j.dss.2010.11.006
45. Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336. https://doi.org/10.1023/A:1007614523901
46. Shao H, Zhao H, Chang GR (2002) Applying data mining to detect fraud behavior in customs declaration. In: Proceedings international conference on machine learning and cybernetics, vol 3, pp 1241–1244. https://doi.org/10.1109/ICMLC.2002.1167400
47. Shmueli G (2017) Analyzing behavioral big data: methodological, practical, ethical, and moral issues. Qual Eng 29(1):57–74. https://doi.org/10.1080/08982112.2016.1210979
48. Singh AK, Sahu R, Ujjwal K (2003) Decision support system in customs assessment to detect valuation frauds. In: Engineering management conference, 2003. IEMC '03. Managing technologically driven organizations: the human side of innovation and change, pp 546–550. https://doi.org/10.1109/IEMC.2003.1252333
49. Stankova M, Martens D, Provost F (2015) Classification over bipartite graphs through projection. Working papers 2015001, University of Antwerp, Faculty of Applied Economics
50. Suykens JA, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J, Suykens J, Van Gestel T (2002) Least squares support vector machines. World Scientific, Singapore
51. Universal Cargo (2018) What does a freight forwarder do and do you need one? https://www.universalcargo.com/what-does-a-freight-forwarder-do-do-you-need-one/. Retrieved 14 Nov 2018
52. Vanhoeyveld J, Martens D (2018) Imbalanced classification in sparse and large behaviour datasets. Data Min Knowl Discov 32(1):25–82. https://doi.org/10.1007/s10618-017-0517-y
53. West J, Bhattacharya M (2016) Intelligent financial fraud detection: a comprehensive review. Comput Secur 57(Supplement C):47–66. https://doi.org/10.1016/j.cose.2015.09.005
54. Wickramaratna J, Holden SB, Buxton BF (2001) Performance degradation in boosting. In: Proceedings of the second international workshop on multiple classifier systems, MCS '01. Springer, London, UK, pp 11–21
55. Yaqin W, Yuming S (2010) Classification model based on association rules in customs risk management application. In: 2010 international conference on intelligent system design and engineering application, vol 1, pp 436–439. https://doi.org/10.1109/ISDEA.2010.276

56. Yuan B, Ma X (2012) Sampling + reweighting: Boosting the performance of adaboost on imbalanced datasets. In: The 2012 international joint conference on neural networks (IJCNN), pp 1–6
57. Zdravevski E, Lameski P, Kulakov A (2011) Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms. In: The 2011 international joint conference on neural networks, pp 181–188. https://doi.org/10.1109/IJCNN.2011.6033219
58. Zhang L, Zhang Q, Zhang L, Tao D, Huang X, Du B (2015) Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. Pattern Recogn 48(10):3102–3112. https://doi.org/10.1016/j.patcog.2014.12.016 (discriminative Feature Learning from Big Data for Visual Recognition)
59. Zheng M, Zhou C, Wu J, Pan S, Shi J, Guo L (2018) Fraudne: a joint embedding approach for fraud detection. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–8