Corporate Residence Fraud Detection

Enric Junqué de Fortuny
University of Antwerp
Prinsstraat 13
Antwerp, Belgium
enric.junquedefortuny
@uantwerp.be

Marija Stankova[*]
University of Antwerp
Prinsstraat 13
Antwerp, Belgium
marija.stankova@uantwerp.be

Julie Moeyersoms
University of Antwerp
Prinsstraat 13
Antwerp, Belgium
julie.moeyersoms@uantwerp.be

Bart Minnaert
Ghent University
Henleykaai 84, G-building
Ghent, Belgium
bart.minnaert@ugent.be

Foster Provost
Stern School of Business
New York University
New York, NY 10012
fprovost@stern.nyu.edu

David Martens
University of Antwerp
Prinsstraat 13
Antwerp, Belgium
david.martens@uantwerp.be

ABSTRACT

With the globalisation of the world's economies and ever-evolving financial structures, fraud has become one of the main dissipaters of government wealth and perhaps even a major contributor in the slowing down of economies in general. Although corporate residence fraud is known to be a major factor, data availability and high sensitivity have caused this domain to be largely untouched by academia. The current Belgian government has pledged to tackle this issue at large by using a variety of in-house approaches and cooperations with institutions such as academia, the ultimate goal being a fair and efficient taxation system. This is the first data mining application specifically aimed at finding corporate residence fraud, where we show the predictive value of using both structured and fine-grained invoicing data. We further describe the problems involved in building such a fraud detection system, which are mainly data-related (e.g. data asymmetry, quality, volume, variety and velocity) and deployment-related (e.g. the need for explanations of the predictions made).

Keywords

fraud detection; corporate residence fraud; transactional data; structured data

1. INTRODUCTION

The social contract [30] between governments, citizens and corporations comprises the mutual agreement between these parties on how to allocate resources for common expenses such as road construction, hospitals and the environment. Most democratic societies have implemented this social contract in the form of a taxation system in which each party agrees to contribute to the total expenditure of the country. Needless to say, the success of such a system depends on the fairness and efficiency and thus the compliance of all actors to the system in place. Falsifying or withholding information in order to limit the amount of tax liability is therefore against the law and constitutes fiscal (or tax) fraud. This is a large-scale problem that affects a multitude of entities: the public sector, the private sector and citizens [25]. Fiscal fraud exists in several forms, which can broadly be categorized as evasion of direct (income and corporate tax) and indirect (VAT) taxes. Governments are a frequent target of fraudsters, who undermine the system and abuse its benefits, grants and tax programs.

In Belgium, fiscal fraud is acknowledged as a significant problem. The State Secretary for Fraud in Belgium even stated that *"Fraud is as Belgian as beer and fries"* [9]. Estimations by the European Commission show that the Belgian government loses about €30 billion annually due to fiscal fraud, which corresponds to 6% of its GDP [9], placing Belgium's among the highest fraud rates in Western Europe. On a larger level the overall European losses due to tax evasion and avoidance are estimated to be an astonishing €1 trillion [12]. These numbers show that the fight against fraud is a crucial aspect of any fiscal system. Not only does fraud cause serious damage to society, it also has a direct financial impact on individuals. The relevance of fraud detection in the current climate of severe fiscal consolidation and social hardship is motivated in the declaration of the G20 leaders of September 2013. In this statement, they emphasize the importance of ensuring that all taxpayers pay their fair share of taxes as well as the need to tackle tax avoidance, harmful practices and aggressive tax planning [27].

Since most tax systems use audits to ensure compliance with tax laws, an accurate selection of the most likely fraudulent cases is crucial to maintain an efficient tax inspection. Given this urgent need to identify specifically the most suspicious cases, the Belgian government joined forces with academia to work on automated data mining systems that look for fraud patterns in large amounts of data to detect corporate residence fraud. This type of fraud occurs when companies deceitfully attempt to place their residency in a

[*] Enric and Marija contributed equally to this work.

KDD'14, August 24–27, 2014, New York, New York, USA.
Copyright © 2014 ACM 978-1-4503-2956-9/14/08…$15.00.
http://dx.doi.org/10.1145/2623330.2623333

low-tax country in order to avoid paying the higher taxes of their real location. The data consists of two types of records: on the one hand we have structured data on the Belgian companies (sector, city, etc), on the other hand we have transactional data (invoicing logs) between Belgian and foreign companies. Although using this fine-grained transactional data can be tricky, the information that could be retrieved from it is very valuable in order to detect residence fraud. Consider the following (fictitious) example: let's say we see that a foreign company receives invoices from a golf club in Brussels. This could be an indication that the company and its owner(s) likely reside in Belgium. If this is indeed so, other foreign companies that also receive invoices from this specific golf club make for interesting suspects. Working at such a fine-grained identifier level makes available very informative data [28].

The potential of data mining techniques has also been acknowledged by governmental entities, including the Belgian government. In their action plan to strengthen the fight against tax fraud, the European Commission articulates it as follows: *"For tax administrations, the development and full use of automated tools and risk management techniques would release human and budgetary resources to concentrate on achieving targeted objectives."* [11]

The rest of the paper is structured as follows: In the next section, a literature overview is given on the importance of fraud detection, the different types of fraud, and the the main domain challenges. Section 3 looks deeper into the type and size of the data and Section 4 describes the specific methods that were used. Section 5 shows the results and the deployment, with concluding remarks in Section 6.

2. LITERATURE OVERVIEW

2.1 The Importance of Fraud Detection

As discussed above, the Belgian government is a frequent target of fraudsters. Abuse of the tax system is a very costly fraud type [25], with estimates of losses going into the billions of euros (dollars, pounds) for the EU, US and UK governments. Translating these numbers to impact on members of society is an easy exercise. For instance, Belgian estimates reveal that fraud against the public sector is estimated to be €30 billion per year and thus directly costs every adult in the country about €2,700 annually.

As mentioned before, the elementary form of damage from fraud in government-allotted resources is an immediate financial loss and thus the unfair redistribution of wealth. Note, however, that the consequences can be much broader. Fraud losses could result in cuts to thinly spread government-budgets, tax increases, less investment in the public sector (such as new roads, hospitals, schools, etc.) and eventually a slower economy altogether. Effective fraud detection, on the other hand, can lead to many benefits. Not only is there the direct impact of recovering parts of the loss of capital, increased effectiveness can also lead to enhanced deterrence [1]. That is, the increased likelihood of being captured, causes the net expected benefit from the fraudulent activities to be outweighed by their (proportionally increased) expected costs, thus decreasing the appeal of such fraud. Needless to say, governments try hard to cope with evermore creative fraud-schemes such as the ones addressed in this project.

2.2 Data Mining for Fraud Detection

In the literature, data mining has been applied to many domains for fraud detection. Some of them include the banking sector for discovering fraudulent credit card transactions or card applications [3, 6, 19, 33, 39], identifying fraudulent service subscriptions or calls in the telecommunications domain [8, 14, 15, 17], detecting false insurance claims [29], revealing websites with high level of non-intentional traffic for online advertising [35] or uncovering tax evasions in the public sector [2, 16, 41] and etc. A comprehensive overview of the complete fraud detection literature is beyond the scope of this paper (see [4, 26, 29]).

Many of the fraud detection studies need to deal, similarly to our work, with heterogeneous types of data and especially large amounts of transactional data. The applications are mainly in the banking and the telecommunications sectors, where the companies keep logs of card transactions and calls. Due to the high dimensionality of the transactional data, a very common approach in the literature is to perform some type of aggregation over the transactional data. One way to do so is to create transaction aggregates for each user account that characterize the typical legitimate behaviour of the user [5, 15]. Any new transaction that deviates from the typical behaviour of that user would be suspected as fraudulent. Other studies [3, 19, 39], take the approach of deriving RFM (Recency, Frequency, Monetary Value) attributes from the original features over a period of time. The RFM attributes are then used as inputs for a classification technique. Aggregating the transactions creates new structured data and loses the fine-grained information that is included in the transactions (cf., the golf club example from Section 1).

To our knowledge, there have been only few studies in the prior literature that take into account the information from very high-dimensional categorical attributes, especially the identifier attributes described by Perlich and Provost [28]. These attributes can represent particular identifiers as the companies accounts in our case, particular names of locations or persons and etc. The work of Fawcett and Provost [14, 15] incorporates such attributes by first searching for individual classification rules based on the transaction-level data (such as location in cell phone calls), and then building higher-level features based on these rules. The studies by Brause [6] and Sanchez [33] include these attributes by using classification based on association rules on transactional level applicable to smaller datasets. Cortes et al. [8] and Stitelman et al. [35] both employ a graph representation and apply relational inference on the networks defined among persons connected if they call each other [8] and among browsers connected if they visit the same website [35]. Our study explores and combines fraud data on both levels: we apply scalable algorithms to extract fine-grained knowledge from huge amount of transactional data and also consider the structured data. By doing so, we are able to harness the predictive power of both types of data, as well as the added value of combining them.

For the purpose of tax evasion, data mining has been applied to the problem of corporate fraud [7, 20], where companies falsify their financial statements, as well as Value Added Tax (VAT) evasion [2, 16, 41], solely on structured data.

2.3 Domain Challenges

Typical challenges encountered when applying data mining techniques in the domain of corporate residence fraud detection relate to positive label scarcity and quality. Additionally, due to the way in which the data is generated nowadays, we also encounter problems related to Big Data with respect to size (volume), type (variety) and speed of data generation and stationarity-violation (velocity). Furthermore, the acceptance by stakeholders of the resulting models is highly dependent on their comprehensibility, which needs to be taken into account both during and after the modeling phase.

Data scarcity: Fraud data are usually *highly unbalanced*, as there are many more non-fraudulent instances than the number of fraudulent ones. Furthermore, limited resources and the very expensive labeling procedure (auditing) further bias the class balance. Moreover, one often encounters pollution of the data labels: data instances can have wrong labels if a fraudulent instance has not yet been discovered and therefore is marked as a legitimate one. Additionally, very little structured data is available on the foreign companies (except for the country where they are located).

Volume, variety and velocity: Every quarter, the government receives millions of tax data entries containing hundreds or even thousands of transactions as well as structured data on each of the companies involved in these entries. As such, not only do the datasets have very large *volume*, the size also continues to increase. Even so, this is not the only issue related to *velocity*. Fraudsters are known to change the way in which they commit fraud in progressively more creative and covert ways to evade the detection systems in place. This adversary effect requires continuous back-testing and updating of the models because stationarity assumptions might be violated. Needless to say, when taken as a whole, the datasets coming from our domain need fast algorithms that can cope with these challenges.

As mentioned before, the government receives both tax declarations as well as transactional data. Furthermore, the government has a database with additional information on each of these companies. Ideally, one wants to connect all these various bits of information in order to obtain the best predictions. Unfortunately, it is not trivial to do so in a sensible way. For instance, how could one combine transactional logs (e.g., foreign company FC_1 transferred money to a golf club) with a geographical location? Possible answers include hierarchical modeling, ensemble methods and stacking; clearly, this situation opens up many possible paths of model combination and design. To the best of our knowledge, we are the first to propose a solution for this corporate governance problem.

Comprehensibility: The success of a tax fraud detection system depends on more than accurately flagging suspected cases. Each suspected case is sent to an investigator who determines whether it is indeed fraudulent and collects evidence. As each investigator develops his/her own expertise on tax fraud, this expertise can conflict with the predictions. If investigators receive many cases they see as clearly non-suspect, they might reject the prediction system altogether. When the system however explains why a case is flagged as suspect, investigators can quickly determine whether this is in line with their experience or not. Further, in a confirmed match situation, the explanation provided by the system can serve as a starting point for the actual investigation. Thus a model that is comprehensible at the instance/decision level is critical both to get user acceptance and to speed up the manual investigation.

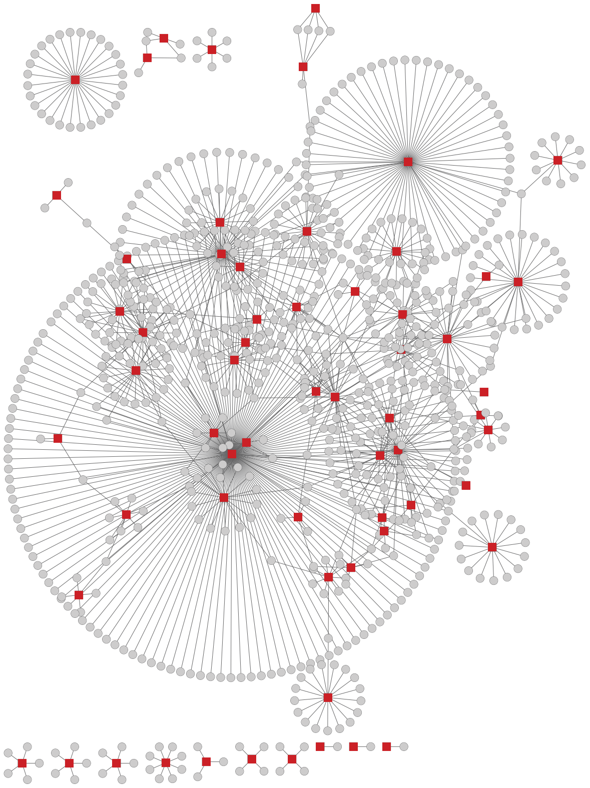


Figure 1: The structure of the invoicing network based on the incoming and outgoing transactions between the fraudulent foreign companies (red squares) and the Belgian companies they interact with (grey nodes). As can be seen from the big cluster, many of the fraudulent foreign companies are connected to the same Belgian companies.

3. DATA

Before we can dig into the modelling approaches for this domain, we must first discuss the exact data available to us. Although we received data from various sources, we can discern two main types of records. First we have invoicing records between 2,745,478 Belgian companies and 873,640 foreign companies (*transactional data, T*). Second, we also have structured information on each of the Belgian companies (*structured data, S*).

Transactional data: In terms of transactional invoicing data, we can distinguish between two types of invoices: incoming invoices from foreign companies to Belgian companies, and outgoing invoices from Belgian to foreign companies. We engineered three different datasets from these invoice logs: a dataset of incoming invoices, a dataset from outgoing invoices and a third dataset where we merged both the incoming and outgoing invoices. Additional statistics for the datasets are shown in Table 1. There can be multiple

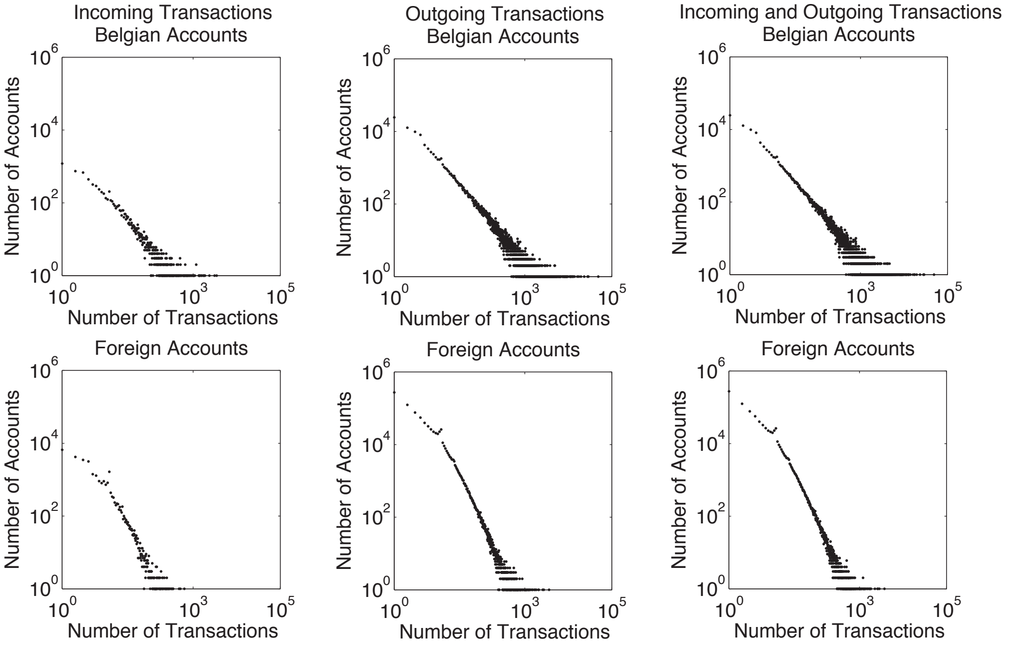


Figure 2: The number of unique transactions per account for the invoicing datasets, when considering the Belgian (top) and foreign (bottom) companies. Most of the Belgian companies typically send or receive invoices to only few foreign companies and, vice versa, most of the foreign companies interact with only few Belgian companies.

Table 1: Statistics for the three invoice datasets.

	Incoming	Outgoing	Incoming and Outgoing
Number of transactions	251,198	6,551,512	6,802,710
Number of unique transactions	73,753	1,955,912	2,029,641
Number of Belgian accounts	7,495	107,345	108,753
Number of Foreign accounts	30,541	858,131	858,703
Average number of transactions per Belgian account	9.84	18.22	18.66
Average number of transactions per foreign account	2.41	2.28	2.36

transactions between two companies, on different dates or with different amounts of money. Hence in Table 1, both the total number of transactions and the number of unique transactions between the companies are shown. The latter counts only the transactions where the sender/recipient pair is unique. Note that this transactional data can be represented both as a matrix and as a bipartite graph.

In the *matrix* representation each row i corresponds to a foreign company; column-values indicate whether the foreign company made a connection to resident company j, with entry $x_{i,j}$ equal to 1 and 0 otherwise. A *bipartite graph* (bigraph) is a graph that has two types of nodes and edges exist only between nodes of different type. A subset of the bigraph containing all of the fraudulent nodes is visualized in Figure 1 with red squares representing the fraudulent foreign nodes and the grey nodes representing Belgian companies they interact with. Figure 2 shows the degree distributions (number of transactions) of the Belgian-foreign bipartite company networks.

These graphs help us to understand the power of the fine-grained data in the modeling results presented below. Although keeping the full fine-grained data instead of aggregate values can be tricky to work with, previous studies have shown fine-grained transaction data to enhance the predictive power of models [18, 24, 28]. This is partly due to the fat tail in the degree distribution we see in Figure 2: many companies appear related to only very few other companies, but these low-connectivity companies make up the vast majority of the companies. Thus, it is relatively difficult to compress the company-related information into a small number of simple aggregate variables (that do not obscure the fine-grained connectivity information). Figure 1 in turn shows that the fraudulent foreign companies indeed do seem to interact with the same Belgian companies, as illustrated by the big cluster in which most fraudulent companies are found. Thus, it makes sense to intelligently—i.e., in a supervised fashion—examine the specific companies in the predictive modeling (note that it is informative both to be connected to one or more suspicion-inducing companies as well as to be connected to one or many suspicion-reducing companies; cf., [28]).

Structured data: Most of the available structured information is on residential companies because, to date, there is still no sharing of information between governments. This asymmetry in data is one of the challenges to overcome on the level of policy making. For each of the Belgian companies we have information on their geographical location, industry type, start-up date, etc. For foreign companies, we only know in which country they are located as well as the target label. As shown in Figure 3, we can infer certain aggregate characteristics for the foreign companies, based on what the average Belgian company that connects to it looks like.[1] For the particular foreign company shown in the figure, we can deduce that its average transaction value is a certain amount and that its usual geographical correspondence location is located in Brussels (median region in Belgium). These characteristics can be added into the input vector in order to augment the prediction information. This set-up leads to a total of 31 features per foreign company.

An important problem that arises in our scenario, due to limited resources and the very expensive labeling procedure, is skewness in the distribution of the target variable. Out of the total 873,640 available foreign companies, only 62 are marked as positive cases. Because of this skewness, we make use of AUC and lift curves and we repeat each of the experiments 10 times on different out-of-sample selections to ensure robustness and the external validity of the results.

4. METHODS

Given the variety and volume of the data, different feature engineering and modeling techniques are first applied and subsequently combined. In this section we first describe each of the different methods briefly after which we discuss their combination, displayed in Figure 4.

4.1 Structured Data

In the structured learning scenario, we are interested in predicting whether or not a foreign company is fraudulent, based on the aggregate, structured information of the associated resident companies. This turns out to be a classical predictive modeling set-up in which we predict target variable y based on vectors of structured data $\mathbf{x}$, one for each foreign company. To deal with the many-to-one variables, such as location, which appear in the transaction data, we encode them in the structured data via a "weight-of-evidence" encoding; this is a logarithmic transformation that allows one to transform a categorical variable into a variable that is monotonically related to the target variable [36, 37]. For example, if most of the Belgian companies connected to a foreign company are located in Brussels, this will be encoded as a one in the position of the dummy-encoded "Brussels" location variable. Examples of structured variables include the location of the linked company (up to town level), the main activity code of the linked company, and the legal construct type of the linked company (with a total of 31 such variables). Once the features have been engineered into a structured input vector, we train an SVM with a linear kernel. SVMs are known to work well with these kinds of data [32] and the choice of kernel is motivated by the need for comprehensibility of the model (more on this later).

4.2 Transactional Data

The transactional data can also be represented by vectors as follows: for each of the n foreign companies we look up its previous associations with companies in Belgium. Each of the m Belgian companies is represented by a feature and the value of this feature in the foreign company's m-dimensional vector $\mathbf{x}$ will be equal to one if such a connection was made and zero otherwise. By aggregating all of these vectors we end up with a very high-dimensional, but highly sparse, matrix. There are two main approaches of handling this kind of data: (a) applying propositional learners (such as SVMs and Naive Bayes) on the huge, sparse matrix representation and (b) using relational learning/inference on the graph representation.

Propositional learners.

A first approach is to gather all of the data in a big matrix and apply *SVM* (using the LibLinear package [13]). Clearly, due to the size of the data, this will take quite a while on a standard desktop computing set-up. Further, it likely will not perform very well due to class imbalance, as explained by Wallace et al. [38]. Indeed, poor performance is revealed in the very low AUC and lift values of this approach (SVM$_T$, Table 1). As a first improvement, we train the SVM on a balanced subset of the data. By equally weighing the number of positive and negative examples, the SVM learns to put equal importance on each of the classes and performs much better (SVM$_T$(50-50)). Other improvements toward this end, could be to directly optimize for a different loss function [31]. An in-depth discussion on this matter is beyond the scope of this application-focused paper.

In a similar vein, we also apply a *binary Bernoulli Naive Bayes* (NB), specifically tailored for massive, sparse, binary data [18]; let's call that "Big Bayes." This classifier uses the same input vectors $\mathbf{x}$, but makes an estimate based on the MAP likelihood estimation of a probability parameter for each of the features. These are gathered in a vector with elements $\theta_j = P(X_j = x_{i,j}|C = c)$ and used in a 'naive' model, where all features are assumed to be conditionally independent of each other, given the class, resulting in the following probability estimate for each class (i.e., fraudulent or not):

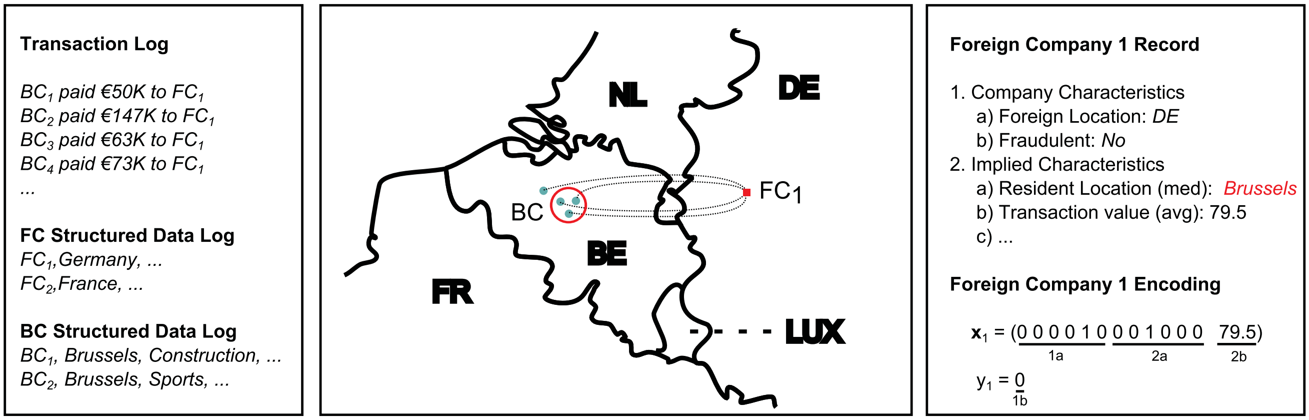
$$P(C = c|\mathbf{x_i}) \quad \propto \quad \prod_{j=1}^{m} (\theta_j)^{x_{i,j}} \, (1 - \theta_j)^{(1-x_{i,j})}$$

In this formulation, fraud is encoded by class label C, and the $x_{i,j}$ indicate whether a transaction was made from foreign company j to resident company i. A decision is made by comparing the probability estimate for the fraudulence of the company ($C = 1$) and the non-fraudulence ($C = 0$). The NB modeling procedure does not suffer from the class skew problems of the SVM. The Big Bayes modifications for massive, sparse data involve only having to process the non-zero elements of the huge matrix [18]; NB does not need any further modifications to be run efficiently on the fine-grained data.

Relational learners.

Intuitively, it makes sense to apply a learner that is specifically tailored for the networks resulting from transactions like the ones described in Section 3. In order to do so, we must first realize that such transactional logs between two

[1] Due to the sensitivity of the data, all of the examples given in the figures are only illustrative; aggregate results and statistics are of course computed on the true data.



Transaction Log

BC_1 paid €50K to FC_1
BC_2 paid €147K to FC_1
BC_3 paid €63K to FC_1
BC_4 paid €73K to FC_1
...

FC Structured Data Log
FC_1,Germany, ...
FC_2,France, ...

BC Structured Data Log
BC_1, Brussels, Construction, ...
BC_2, Brussels, Sports, ...

Foreign Company 1 Record

1. Company Characteristics
   a) Foreign Location: *DE*
   b) Fraudulent: *No*
2. Implied Characteristics
   a) Resident Location (med): *Brussels*
   b) Transaction value (avg): 79.5
   c) ...

Foreign Company 1 Encoding

$\mathbf{x}_1 = (\underbrace{0\,0\,0\,0\,1\,0}_{1a}\ \underbrace{0\,0\,1\,0\,0\,0}_{2a}\ \underbrace{79.5}_{2b})$

$y_1 = \underbrace{0}_{1b}$

Figure 3: Example of the feature engineering for structured data. The foreign company (FC_1) has many associated Belgian companies (BC_i). Original company characteristics such as the location are combined with implied characteristics such as the average transaction values and the median resident location.

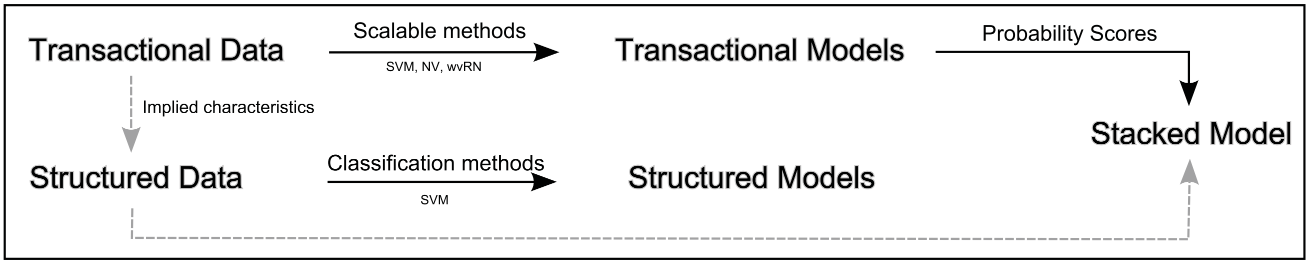


Figure 4: Overview of the system design for fraud detection. In a first step, the transactional data and structured data are engineered into features. Afterwards different modeling techniques are applied to generate both transactional and structured models. These models are evaluated separately at first, but combined in a final step as well into a stacked model.

entities (foreign and resident companies) can be represented as a bipartite graph. The visualization already suggests ("two-hop") assortativity in fraud status in the network of foreign companies.

Numerous relational learners exist for graphs with only one type of nodes. In order to make use of them, we can apply the three-step framework for classification within bigraphs proposed in [34]. The idea is to project the bigraph into a unigraph (graph with only one type of nodes) in which foreign companies are connected, based on shared Belgian company connections and then apply a relational learner. By additionally assigning weights to the edges in the projection, more information from the underlying bigraph can be preserved [34]. The resulting classification decision is then based on the posterior probability, defined as:

$$P(C = c|\mathbf{x}_i) \quad \propto \quad \sum_{j \in N(i)} w_{ij} \cdot P(C = c|N(\mathbf{x}_j)) \qquad (1)$$

$$w_{ij} \quad = \quad \sum_{k \in N(\mathbf{x}_i) \cap N(\mathbf{x}_j)} \tanh\left(\frac{1}{d_k}\right) \qquad (2)$$

Equation (1) presents the weighted-voted Relational Neighbor (wvRN) inference method [21]; with wvRN, the class probability of a node in the graph (a foreign company) is equal to the weighted average probabilities of all of its neighbors ($j \in N(\mathbf{x}_i)$). A neighbor is defined as a node that is linked to the node that is being investigated in the *projected graph* (in this case identified by a one in the $\mathbf{x}$ vector). As mentioned before, such a connection is made only if two foreign companies share a resident company. Equation (2) shows that the weighting (the similarity between two top-nodes) was chosen as a sum over the tanh of the inverse of the degrees of the shared nodes. That is, if say a Belgian company has connections with all of the foreign companies, this company will define a relatively low weight in the total probability calculation. If it does not, it will likely be more informative and thus should be weighted accordingly. These design choices are based on the results of the extensive experimental study on a wide variety of publicly available transactional datasets conducted by Stankova et al. [34].

4.3 Stacked model

Ideally, we want to build a model that incorporates all of the available information. As one can see from the previous sections, it is not trivial to combine these heterogeneous types of data because they require different sorts of models. One way to cope with this problem, while still preserving the variety of modeling approaches is to combine the models in a *stacked model.* The expected efficacy of such a model is explained by the fact that we are incorporating more information into one model than we did before, which should result in a lower modeling bias [40].

In our scenario, as the stacked model we use a linear SVM to produce a final model that is a linear combination of the output scores of the transactional classifiers and the structured model. An important reason for this particular design is that we do want to keep a maximum level of comprehensibility without sacrificing too much predictive performance. Specifically, the 31 variables of structured data are manageable to a human observer, but the millions of transactions clearly are not. It is much more informative to have the scores of these models encoded as variables—a human inspector can assess the contribution of the network-data component. Should this be high enough, specialized techniques can be used to inspect the underlying reasons for the predictions of the network-data component (as discussed in the next section).

Figure 4 summarizes all of the steps required to generate the complete, stacked model. First, the data is converted to (a) transactional (graph) data and (b) structured data. Next, predictive models are built on top of these data, each specifically tailored to cope with the particular aspects of the corresponding data (as explained in the previous section). Lastly, the scores of the graph models are combined with the structured data as input to the final stacked model.

5. RESULTS AND DISCUSSION

5.1 Results

The results of all of the previously explained methods in terms of predictive power (AUC) are shown in Table 2. The best performance for each dataset is denoted in boldface and underlined. Performances that are not significantly different at a 5% confidence level (according to a Wilcoxon signed rank test [10]) are tabulated in bold face. Significant differences at the 1% level are emphasized in italics, and differences at the 5% but not at the 1% level are reported in normal script. A first observation that one can make from this table is that our best models achieve very high AUC values (up to 96.22%). The somewhat high standard deviations on these percentages can be explained by the class imbalance (detecting one more or one fewer example can result in a percentage change of about 10%). Nevertheless, our results do show that our best model (the stacked combination of structured and relational models; SVM_{S+T}) performs significantly better than all of the transactional methods for the incoming and the outgoing data. Although it is still the best performing model for the combination of both data types, the variance is too large to conclude statistical significance at the 5% level using the Wilcoxon test.

Table 2: Results of different techniques in terms of AUC. Subscript S refers to models based on structured data. Subscript T refers to models based on fine-grained transaction data. Subscript $S+T$ refers to models based on both structured data and transaction data.

	Incoming data		Outgoing data		Combined data	
Technique used	AUC	std.dev.	AUC	std.dev.	AUC	std.dev.
wvRN$_T$	76.74	(±5.87)	77.32	(±6.21)	**94.55**	(±5.26)
Naive Bayes$_T$	76.64	(±5.94)	77.6	(±6.37)	**94.74**	(±5.45)
SVM$_T$	*46.88*	(±12.76)	*56*	(±9.27)	*70.26*	(±12.46)
SVM$_T$ (50-50)	*62.23*	(±21.03)	57.95	(±33.66)	74.85	(±19.97)
SVM$_S$	**82.71**	(±10.52)	**86.34**	(±7.74)	**91.77**	(±8.16)
SVM$_{S+T}$	**85.92**	(±7.48)	**86.44**	(±10.23)	**96.22**	(±4.8)

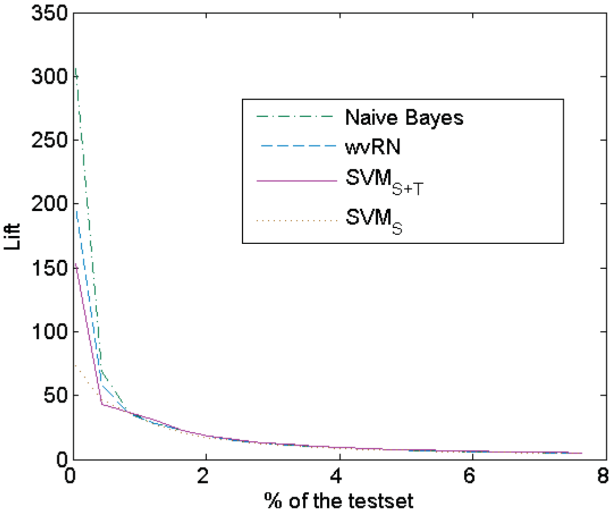


Figure 5: Lift curves of the combined dataset

Although these results are certainly interesting in terms of global predictive power and ranking ability, we should note that in the specific context of detecting fraudulent companies we are more interested in the lift (how much better than random) when targeting the highest ranking members of the dataset. This is because the fraud analysts investigate the companies deemed to be most suspicious. The lift curves (Figure 5) show the clear superiority at the highest percentiles of the models built on transactional data, where they are able to perform up to a few hundred times better as opposed to random targeting. The traditional, structured-data model and the stacked model deliver clear improvements as well, but at the highest percentiles the lifts are not nearly as strong as those resulting from using the fine-grained transaction-based models.

As we motivated previously, we can now observe empirically that the fine-grained information included in the transactional data provides substantial gains for detecting fraudulent companies. Referring back to our example, the other fraudulent companies that transact with the Brussels golf club receive high transactional fraud scores, and rightly so apparently—as demonstrated by the very high lifts. Once these other foreign companies that transact with these suspicion-conferring Belgian companies[2] are investigated, structured data may still help to find other suspects.

In conclusion, we can say that if one is interested in a global ranking method, the stacked model would be the best design choice in our scenario, whereas the models based on transactional data are better suited for detecting the most likely frauds. The latter result highlights the importance of keeping the fine-grained information as a whole as opposed to only aggregating it into summary variables.

5.2 Comprehensibility

In the actual deployment of our model, we have been made aware of the tremendous importance of being able to explain

[2] The Belgian companies themselves are not suspicious per se, but the foreign companies that transact with them are.

the decisions made. Specifically, the auditors need to understand the exact reasons why classification models make particular decisions. Cases (even if they be few) where the model makes an obvious wrong decision can create disillusionment with the system and reluctance to use it, unless the reasons behind the decision appear to be sound. Therefore, it is essential that the decisions made by the predictive model can be explained; the auditor can decide to over-rule a specific suggestion and confidently move on to the next one. Going back to our running example of a company that has received an invoice from a golf club in Brussels: Although it might be the case that most foreign companies that receive invoices from that entity are indeed fraudsters, a foreign company such as Rolex that has sponsored a golf tournament at this specific golf club (and therefore has also received an invoice) is likely not fraudulently located abroad. So if the explanation for the classification is given (i.e., receiving an invoice from the specific golf club), an auditor can quickly see why it is or is not valid in the context of the particular focal company.

To our knowledge, the distinction between different types of comprehensibility has received relatively little attention in the data mining literature, even though it often is a crucial criterion for final acceptance and increased use of the predictive models. At least two types of explanations exist. Global explanations provide improved understanding of the complete model, and its performance over the entire space of possible instances. Instance-level explanations on the other hand provide explanations for the model's prediction regarding a particular instance. When using transactional data, the total number of variables and/or data values considered by the model (in our case, millions) is much larger than for the typical structured data. Global explanation methods, such as examining the coefficients of a linear model or using a rule-based model, are simply not applicable in such a high-dimensional context. However, an instance-level approach used for document classification [23], which faces a similar challenge with a large vocabulary, can also be used in this transactional setting: an explanation is defined as the minimal set of entities one received/sent an invoice to, such that removing all the invoices to/from this set changes the predicted class from the class of interest. For our running example, an explanation could be: *'if this company did not receive an invoice from golf club XYZ in Brussels, the predicted class would change to non-fraudulent'*. As such, instance based explanations provide an excellent tool for models that use the fine-grained invoicing data. For more on how explanations can be used both to improve acceptance and also to improve the model itself, as well as further references to related work, we refer to [23].

Global explanations do still have value, but in a different way. Decision makers need insight into the general methods used by fraudsters and their evolution. One way to do so is to list all variables of the stacked model in order, ranked according to the size of the coefficients in the linear model. Then we could see for example that the country dummies for certain countries are very high on the list, as well as the scores from the transactional models, and certain activity codes. A rule-based model could provide similar insights. These insights may then lead to different sorts of cases being discovered, which then would prime the network models to find similar instances. We are not able to show the actual global explanations, as they involve confidential information.

5.3 Deployment

In reference to this project, State Secretary for Fraud John Crombez reported: *"The interaction between the two worlds [academia and government] has proven very valuable. Other countries are now visiting Belgium to see how the Social Intelligence and Investigation Service and the Special Tax Inspection service apply this technique. That is why we need to continue to invest in this technology."* Not only is the predictive performance of our models appreciated, but also considered to be important to success is the fact that in general use this data mining technique can operate on anonymized data, whereby each company is encoded as a "random" number. A company's identity only then needs to be revealed in the context of a particular investigation of a top-suspicion instance. Further, the emphasis on the comprehensibility of the results is deemed essential.

During deployment, the system has to deal with large volumes of heterogeneous data and with new data arriving every quarter, where the underlying data generating process is non-stationary due to the problem being adversarial. The stacked model approach specifically deals with the variety of the data by combining the transactional data from invoices with structural data from tax declarations. The need to retrain the model frequently is facilitated by the scalability of the underlying (naive Bayes and wvRN) methods. They can be run (on a desktop) on the complete data and produce results in a matter of minutes.

6. CONCLUSION

In this paper we have described what to our knowledge is the first data-mining-based method for building a system for detecting corporate residence fraud. The system is based on transactional and structured data, which is gathered by the Belgian government. The success of such a detection system in practice depends on a combination of factors, including efficiency, efficacy and comprehensibility. As such, an important part of our research was to evaluate how one can cope with these conflicting requirements. When used for targeting new fraudsters, a combination of the fine-grained transactional data model and instance-based explanations results in a good trade-off between the needs of an auditor. On the other hand, combining both structured data and fine-grained data in a stacked model is more suited when the main goal is to gain macro-level insights and policy guidance.

Given the success of this pilot study, we believe further research into this application to be a logical next step. There are still many opportunities for improvement. Besides simply improving the modeling methods, one particular aspect that we did not touch upon yet is the pro-active gathering of data with active data-acquisition techniques (see e.g., [22] for a suspicion-scoring application).

It is important to continue to stress the importance of deploying counter-fraud measures for the social good of countries. Although our experiments focus on data from the Belgian government, we hope that researchers from other countries are motivated by our results to apply such methods to or to find better methods for their own countries' data, and/or to convince their governments to do so. It is important for us to understand whether and how data mining indeed can improve government fraud detection efficacy and perhaps even policy making. Once we are convinced, then we can work to remove any lingering doubt or scepticism among decision makers.

7. ACKNOWLEDGMENTS

We are very grateful to the Belgian government for allowing us to write about our work on this problem and especially for their engagement in the project and their useful feedback. David Martens also thanks the Flemish Research Council (FWO) for financial support (Grant G.0827.12N) and Foster Provost thanks NEC for a Faculty Fellowship. Note that these exact models and techniques are not necessarily those used in production.

8. REFERENCES

[1] M. H. Baer. Linkage and the Deterrence of Corporate Fraud, 2008.

[2] S. Basta, F. Fassetti, M. Guarascio, G. Manco, F. Giannotti, D. Pedreschi, L. Spinsanti, G. Papi, and S. Pisani. High quality true-positive prediction for fiscal fraud detection. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 7–12. IEEE, 2009.

[3] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, 2011.

[4] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, pages 235–249, 2002.

[5] R. J. Bolton, D. J. Hand, et al. Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control VII*, pages 235–255, 2001.

[6] R. Brause, T. Langsdorf, and M. Hepp. Neural data mining for credit card fraud detection. In *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pages 103–106. IEEE, 1999.

[7] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak. Detecting management fraud in public companies. *Management Science*, 56(7):1146–1160, 2010.

[8] C. Cortes, D. Pregibon, and C. Volinsky. *Communities of interest*. Springer, 2001.

[9] J. Crombez. *Zwart en wit*. De Bezige Bij, 2013.

[10] J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[11] EUR-LEX. Communication from the commission to the european parliament and the council, 2012.

[12] European Commission. Fight against tax fraud and tax evasion: A huge problem, 2013.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[14] T. Fawcett and F. Provost. Combining data mining and machine learning for effective user profiling. In *Proceedings of the Third KDD International Conference on Knowledge Discovery and Data Mining*, pages 8–13, 1996.

[15] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

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

[17] C. S. Hilas and P. A. Mastorocostas. An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems*, 21(7):721–726, 2008.

[18] E. Junqué de Fortuny, D. Martens, and F. Provost. Predictive Modeling with Big Data: Is Bigger Really Better? *Big Data*, 1(4):215–226, Oct. 2013.

[19] P. Juszczak, N. M. Adams, D. J. Hand, C. Whitrow, and D. J. Weston. Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics & Data Analysis*, 52(9):4521–4532, 2008.

[20] E. Kirkos, C. Spathis, and Y. Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4):995–1003, 2007.

[21] S. A. Macskassy and F. Provost. A simple relational classifier. 2003.

[22] S. A. Macskassy and F. Provost. Suspicion scoring based on guilt-by-association, collective inference, and focused data access. In *International conference on intelligence analysis*, 2005.

[23] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(4), 2014.

[24] D. Martens, F. Provost, J. Clark, and E. Junqué de Fortuny. Mining fine-grained consumer payment data to improve targeted marketing. Technical report, Stern School of Business, New York University, 2013.

[25] National Fraud Authority. Annual fraud indicator 2013. 2013.

[26] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

[27] Organisation for Economic Co-operation and Development. Tax and development themes in recent G20 discussion, 2013.

[28] C. Perlich and F. Provost. Distribution-based aggregation for relational learning with identifier attributes. *Machine Learning*, 62(1-2):65–105, 2006.

[29] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[30] J.-J. Rousseau. *The Social Contract, Or Principles of Political Right (Du contrat social ou Principes du droit politique)*. 1762.

[31] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *The Journal of Machine Learning Research*, 10:2233–2271, 2009.

[32] Y. Sahin and E. Duman. Detecting credit card fraud by decision trees and support vector machines. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, 2011.

[33] D. Sánchez, M. Vila, L. Cerda, and J.-M. Serrano. Association rules applied to credit card fraud detection. *Expert Systems with Applications*, 36(2):3630–3640, 2009.

[34] M. Stankova, D. Martens, and F. Provost. Classification over bipartite graphs through projection. *University of Antwerp, working paper*, 2013.

[35] O. Stitelman, C. Perlich, B. Dalessandro, R. Hook, T. Raeder, and F. Provost. Using co-visitation networks for classifying non-intentional traffic. 2013.

[36] L. C. Thomas. *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*. Oxford University Press, 2009.

[37] L. C. Thomas, D. B. Edelman, and J. N. Crook. *Credit scoring and its applications*. Siam, 2002.

[38] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Class imbalance, redux. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 754–763. IEEE, 2011.

[39] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams. Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1):30–55, 2009.

[40] D. Wolpert. Stacked generalization. *Neural networks*, 1992.

[41] R.-S. Wu, C.-S. Ou, H.-Y. Lin, S.-I. Chang, and D. C. Yen. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, 39(10):8769–8777, 2012.