Forensic Science International 282 (2018) 24–34

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case

Elena Badal-Valero, José A. Alvarez-Jareño, Jose M. Pavía*

Department of Applied Economics, University of Valencia, Avenida de los Naranjos, s/n, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 15 June 2017
Received in revised form 13 October 2017
Accepted 6 November 2017
Available online 11 November 2017

Keywords:
Fabricated data
Fraud
Crime data
Neural networks
Random forests

ABSTRACT

Objectives: This paper is based on the analysis of the database of operations from a macro-case on money laundering orchestrated between a core company and a group of its suppliers, 26 of which had already been identified by the police as fraudulent companies. In the face of a well-founded suspicion that more companies have perpetrated criminal acts and in order to make better use of what are very limited police resources, we aim to construct a tool to detect money laundering criminals.
Methods: We combine Benford's Law and machine learning algorithms (logistic regression, decision trees, neural networks, and random forests) to find patterns of money laundering criminals in the context of a real Spanish court case.
Results: After mapping each supplier's set of accounting data into a 21-dimensional space using Benford's Law and applying machine learning algorithms, additional companies that could merit further scrutiny are flagged up.
Conclusions: A new tool to detect money laundering criminals is proposed in this paper. The tool is tested in the context of a real case.

1. Introduction

Practically on a daily basis, newspapers as well as radio and television news programs report on the occurrence of some or other economic crime: tax fraud, money laundering, corruption, embezzlement of public funds, etc. These are referred to as white collar crimes, crimes which call for more intelligence than brute force. Consequently, the tools for their detection and prosecution also have to be more sophisticated. In 1972, the American economist Hal Varian [1] proposed the use of Benford's Law as a prospective diagnostic tool for highlighting sets of economic and financial operations that require more in-depth scrutiny.

The Benford's Law was discovered by the astronomer and mathematician Simon Newcomb in 1881 [2], although its true value was not recognised until 57 years later when the physicist Frank Benford rediscovered it. Benford's Law affirms that the frequency distribution of leading digits in many real-life collections of numbers is not uniform. Benford's Law defines a biased distribution based on a logarithm law.

In the business and economics world, many data sets obey Benford's Law. Hence, if the economic data follow Benford's Law naturally, its non-compliance could be indicating the possible presence of irregularities in accounting or business-to-business transactions. Benford's Law can be used as a tool to direct us to an economic crime of money laundering or tax evasion [3].

Money laundering is a financial crime which has evolved over time and is implemented at different levels and to different degrees. According to Interpol, money laundering is defined as "any act or attempted act to conceal or disguise the identity of illegally obtained proceeds so that they appear to have originated from legitimate sources". The defrauded amounts range from the traditional laundering of small amounts of money from retail and local drug trafficking to large amounts (billions of euros) from business macro-structures emerging in recent decades and which operate on an international scale [4].

Predicate offences of money laundering are crimes against patrimony (e.g., robbery, theft, fraud or counterfeiting), public administration (tax fraud or evasion), corruption (bribery, influence peddling, embezzlement of public funds, disobedience of penal law, or prevarication), drug trafficking, people smuggling or corporate fraud, among others [5–7].

Money laundering foments unfair business competition, illegal money capital outflows, political and police corruption and social disaffection towards institutions. All agents involved in a criminal organization, with few exceptions, carry out illegal activities for the sole purpose of making a profit [8]. Hence, understanding

* Corresponding author.
E-mail address: pavia@uv.es (J.M. Pavía).

https://doi.org/10.1016/j.forsciint.2017.11.008
0379-0738/

money laundering as the "Achilles' heel" of any criminal organization is the key to combating illegal activities carried out by professional criminals and their enjoyment of illicit capital [9]. The main objective of the anti-money laundering and counter terrorism financing regime (AML/CFT) is to reduce crime rates related to professional crime, organized crime and terrorism, and in turn to protect society as a whole [10].

Failure to comply with Benford's Law is only evidence that the values of a set of numbers can be manipulated. It does not itself identify a crime. Benford's Law is not a universal law, like the law of gravity, and there will be data sets that do not conform to it. However, if the data appear manipulated, something must be behind this, and it would therefore be appropriate to investigate the reason for this anomalous behaviour.

On this basis, we analyse a database composed of the operations carried out between a company suspected of money laundering (parent or core company) and a group of more than 600 suppliers, some of which had previously been identified by police authorities as fraudulent or cooperative. The aim is to find patterns of behaviour in this set of companies which would then enable the identification of other companies that might deserve a more detailed scrutiny.

We use Benford's Law as a tool to characterize the accounting records of business operations between the core company and the suppliers and we apply four classification models (logistic regression, neural networks, decision trees and random forests) to identify other potential fraudulent suppliers. In the models, we incorporate the knowledge provided by the police on which companies have already been identified as collaborators. The ultimate aim is to uncover the largest number of fraudulent companies possible and, at the same time, reduce the likelihood of wrongly targeting companies who are operating correctly. Through the use of this methodology a group of companies have been identified that show a greater probability of fraudulent operations. This enables the scarce resources of the police investigators to be used more efficiently by focusing more on these companies.

This paper has been completed in the context of a police investigation from a Spanish case of money laundering in which the authors have collaborated as forensic data experts. As far as we know, this work represents the first step towards the use of machine learning for the detection of financial fraud in Spanish judicial cases.

The rest of the paper is organized as follows. Section 2 briefly reviews the use of Benford's Law in the literature. Section 3 focuses on methodological issues. In this section, we introduce Benford's Law, we detail the statistical tests implemented, describe the machine learning methods used and, after drawing attention to the challenge that entails handling clearly imbalanced data sets, we present the strategies used to deal with this. The data and the treatments to which they have been subjected are presented in Section 4. Section 5 shows the results obtained after applying the methods considered. The final section deals with discussions and conclusions.

2. A review of the literature

Outside the area of accounting and economics, Benford's Law has been applied to different fields of knowledge. In computing, Torres et al. [11] have verified that the size of the files stored in a personal computer follows Benford's Law. This knowledge can help to develop more effective data storage procedures, to carry out maintenance, or as a tool for detecting viruses or errors. In mathematics, Luque and Lacasa [12] have uncovered a statistical behaviour in the sequence of prime numbers and of zeros in the Riemann zeta function that coincides with the generalized Benford's Law. In election forensics, Benford's Law has been extensively used as a tool for detecting election fraud (e.g., [13–15]), although its effectiveness in this area has been called into question by Deckert et al. [16]. Crime statistics also follow Benford' Law, according to Hickman and Rice [17] official crime statistics in the US conform Benford distribution at the national and state levels. Furthermore, Benford's Law has also been applied to study the length of rivers [18], to detect scientific fraud [19], to assess quality of survey data [20] and to discover manipulation in self-reported toxic emissions data [21].

The use of Benford's Law in the financial and accounting area is more widespread. [3] suggests that it can be used to detect fraud in income tax returns and other accounting documents. This idea is reinforced in Nigrini [22], who states that "individuals, either through psychological habits or other constraints peculiar to the situation, will invent fraudulent numbers that will not adhere to the expected digital frequencies", and in Nigrini and Mittermaier [23], who proposed extending its use as a regular tool in auditing. Currently, a line of development in accounting applies Benford's Law to detect fraud, or the "manufacture" of data, in accounting and financial documents [24–26].

Quick and Wolz [27], Tam and Gaines [28], Rauch et al. [29], and Alali and Romero [30] constitute other examples. Quick and Wolz [27] examine data of income and from the balance sheets of various German companies for the years 1994–1998 and find that the series of numbers of the first and second digits are, in most cases, adjusted to Benford's Law. They find these patterns both when the analysis is performed on an annual basis and also when the inspection is carried out for the whole period. Tam and Gaines [28] scrutinize financial transactions undertaken in the context of election campaign finance and point to pockets of data that merit more careful inspection. Rauch et al. [29] inspect the macroeconomic data reported to Eurostat by the EU member states and find that Greece is, among all euro states, the country reporting the data with the greatest deviation from Benford's Law. In a more recent study, Alali and Romero [30] analyse the financial information corresponding to more than ten years of accounting data of a large sample of US public companies and find that the current assets (equipment, property, accounts receivable) do not conform to Benford's Law. This result leads them to conclude that during the period analysed there was an overestimation of the asset.

Günnel and Tödter [31] consider that Benford's Law is a simple, objective, powerful, and effective tool for identifying anomalies in big samples of data that require a detailed inspection; a vision shared by many authors. There are fewer consensuses on how the knowledge provided by Benford's Law should be used however. Günnel and Tödter [31] argue that controls over data manipulation should focus on the first digit. Whereas, Ramos [32] states that the analysis should focus on the first three digits. According to Ramos, the analysis of the first three digits offers a realistic electrocardiogram of the set of numbers, allowing a detailed observation of what happens at each point and which are the potentially fraudulent operations.

As has been shown, the use of Benford's Law in the field of accounting is prominent, having shown itself able to detect anomalies in accounting data. In accordance with this premise, in Subsection 3.2 we propose different measures based on Benford's Law. These measures are used, along with other variables, as indicators for the detection of patterns that conceal fraudulent operations with the ultimate aim of directing law enforcement authorities to companies that are more likely to have engaged in fraudulent operations.

3. Methodological issues

The aim of the current paper is to classify a set of suppliers as fraudulent or non-fraudulent based solely on the data available in the undisclosed accounting ledgers of a large company

investigated for laundering huge amounts of money. This is carried out by analysing the monetary payments from commercial operations carried out between the suppliers and the core company. In this research, we rely on machine learning techniques to find out, within a binary decision model, which patterns can detect those companies labelled by police experts as illegal (i.e., companies that are collaborators with the core company and are used as screens just to launder money). The aim is to determine whether there are additional companies that merit further scrutiny, other companies which can be classified as illegal, or likely to be illegal.

A major problem in the dataset lies in the fact that companies classified as fraudulent are very few in relation to the total number of suppliers contained in the dataset. That is, the response variable is very imbalanced. Therefore, we consider statistical techniques of balancing with the purpose of improving the predictive capacity of the models. This section details the methodological aspects related to the approaches and techniques used and their validation.

3.1. Benford's Law

Empirically, Benford [33] realized that, contrary to what might be expected, the frequency distribution of leading digits in many real-life collections of numbers is not uniform. He discovered that the frequency distribution that arises in many natural sets of numerical data is biased towards small numbers. Benford's Law sets down that the probability of occurrence of each first digit d_1 (=1, 2, . . . , 9) in many sets of numbers responds to the following probability mass function:

$$f_1(d_1) = P(X_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) d_1 = 1, 2, \ldots, 9$$

Its cumulative distribution function being:

$$F_1(d_1) = P(X_1 \leq d_1) = \log_{10}(1 + d_1)\, d_1 = 1, 2, \ldots, 9$$

From the first-digit distribution, it is not difficult to derive the frequency distribution for the second digit (e.g., [34]. The second-digit distribution can be given as:

$$f_2(d_2) = P(X_2 = d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k + d_2}\right) d_2 = 0, 1, 2, \ldots, 9$$

Notably, if a set of numbers follows Benford's Law, the percentages expected for the first and second digit are given in Table 1.

Benford's Law has as its main properties invariance in scale and in base [35,36,34]. The scale-invariant property implies that Benford's Law continues to be fulfilled even if the units of measurement are changed. That is, the level of fit of some data to Benford's Law is independent of the measurement system. In economic terms, the currency in which the variable is measured does not influence the result. The base-invariant property states that the logarithmic law remains independent of the base used. It is equally valid in base 10, in binary basis, or in any other base. Hill [34] proves that Benford's Law is the unique continuous distribution base-invariant and that scale-invariance (a property impossible for continuous variables) entails base-invariance, the reverse being untrue.

Table 1
Probability distributions of first- and second-digit Benford's Laws.

	0	1	2	3	4	5	6	7	8	9
First-digit	–	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Second-digit	12.0	11.4	10.9	10.4	10.0	9.7	9.3	9.0	8.8	8.5

Although Benford's Law is not universally applicable, it is more "robust" than one might guess. For instance, if we randomly select probability distributions and from each of them we take random samples, we see that the significant-digit frequencies of the combined set will converge to the Benford's distribution even if each particular distribution deviates from Benford's Law [37]. This last result is interesting, because it is a deep-rooted fact in psychology that people cannot behave truly randomly, even when it is to their benefit to do so [38]. Hence, due to human biases, when people manufacture data manually, the data rarely fits Benford's Law.

3.2. Statistical tests

A first step towards identifying fraudulent companies might be to calculate the frequencies of leading digits of the monetary amounts corresponding to each supplier and to study the degree of fit to Benford's Law that the data present, using any of the classic tests of goodness-of-fit (χ^2, Kolmogorov–Smirnov, Kuiper). However, this approach would not be adequate. On the one hand, as is well known, goodness-of-fit tests tend to reject the null hypothesis as soon as the sample size grows: they have too much statistical power. On the other hand, as stated by Giles [39] and Tam and Gaines [28], these tests can be excessively rigid for economic data. In real-life data sets, Benford's Law does not represent a true distribution but a distribution that we would expect to occur in the limit.

Under these conditions, we have opted for a more flexible alternative and we have mapped the distributions of monetary amounts of each supplier in a 20-dimensional space of p-values. This strategy allows us, on the one hand, to characterize the distribution of monetary amounts of each supplier and, on the other hand, to have a large battery of features that, along with other variables, can be included in a model of machine learning. Specifically, we have calculated the p-values of individual fit to Benford's Law of the frequencies of each of the first and second digits and computed the p-value for a fit of each data set to the first-digit Benford's Law using a statistical test based on the distance χ^2, but more flexible.

To measure the fit of first and second digit of each supplier to Benford's Law, we have computed the Z_i statistic (Nigrini [54]) and its associated p-value in a two-tailed test under the hypothesis that Z_i follows a standard Normal distribution.

$$Z_i = \frac{|n_{Oi} - n_{Ti}| - \frac{1}{2N}}{\sqrt{\frac{n_{Ti}(1 - n_{Ti})}{N}}}$$

where n_{Oi} is the frequency of either first or second digits equal to i in the subsample corresponding to the supplier for which the Z measurement is being computed, $n_{Ti} = Nf(i)$ is the associated expected frequency under Benford's Law (with f being equal to f_1 or f_2, as appropriate) and N is the number of operations corresponding to the supplier company in the accounts book.

Moreover, we have used an empirical test based on simulation to calculate the degree of global fit of the data for each supplier to the first-digit Benford's Law. The test, which we call the OverBenford test, eliminates the sample size effect. We obtain the p-value for the OverBenford test by simulation. Firstly, we draw B samples from f_1 with the same size N as our actual sample. Secondly, we compute for each of these new samples the χ^2-distance to the expected distribution. Finally, we calculate the p-value as the proportion of times the sample B distances computed in (ii) exceeds the χ^2-distance obtained from the observed sample. This approach mimics the approaches suggested for goodness-of-fit of continuous distributions implemented in Pavía [40].

3.3. Machine learning methods

Benford's Law provides the basis for classifying companies as legal or fraudulent. In a first phase, the p-values corresponding to the OverBenford statistic and the different Z_i are calculated. These p-values serve to characterize the behaviour of each of the companies in their daily operations and are used to perform the classification. In addition to the p-values, we have also used as classificatory variable the number of operations. These variables constitute the predictors (features) that are included in the machine learning models of classification to be used in this research. The automatic learning methodologies used have been: ridge logistic regression (LR), neural networks (NN), C4.5 decision trees (DT) and random forests (RF). In what follows in this subsection we outline the different procedures.

3.3.1. Ridge logistic regression

Logistic regression models are classic procedures widely used to model the relationship between a dichotomous variable and one or more features [41]. Logistic regression models are used to either (i) quantify the importance of the relationship between each of the features and the binary response variable or, (ii) classify instances between two categories. We use ridge logistic regression [42] for this second purpose. In ridge logistic regression, penalization is used to prevent overfitting occurring due to either collinearity among the predictors or high-dimensionality. The aim of shrinkage and penalization is to improve the predictive accuracy of the model.

3.3.2. Neural networks

Neural networks take their inspiration from the human brain. Neural networks as learning methods have been used for over 50 years and were developed separately in statistics and artificial intelligence to mirror the way human brains solve problems [43]. Artificial neural networks use concepts borrowed from our understanding of how the human brain responds to stimuli from model arbitrary functions. The neural networks are made up of a set of simpler elements, which we call neurons, that are interconnected in parallel in hierarchical form and that interact like the neural systems. A neural network has four basic elements: the number of layers, the number of neurons per layer, the degree of connectivity and the type of connections between neurons. The central idea of a neural network system is to model the response variable as a nonlinear function of the features by processing linear combinations of the inputs as derived features. In this sense, they are referred to as black-box algorithms because they use a complex and obscure mechanism to transform the inputs into a response. Neural networks are prediction tools, difficult to interpret, that can be used to predict both categorical and continuous variables. In this research, the neural network is trained within a supervised learning paradigm with the aim of identifying suspicious suppliers.

3.3.3. C4.5 decision trees

Decision trees represent another classical technique of machine learning. Decision tree learners build a model in the form of a tree structure. The decision tree classifies the instances according to an objective based on the available features, which can be quantitative or qualitative. A decision tree can be seen as a flowchart with decision nodes that can be interpreted as rules. Decision tree algorithms partition the data recursively until some condition is met, such as minimization of entropy or classification of all instances. Due to this procedure, the tendency is to generate trees with many nodes and nodes with many leaves, which leads to over-adjustment or overtraining. The tree will have great precision in the classification of the training data, but very little precision in classifying the instances of the test data. This problem is solved with a pruning procedure a posteriori. In this research, we used the algorithm C4.5 developed by Quinlan [55].

3.3.4. Random forests

A random forest is an ensemble-based method that uses decision trees as building blocks to construct more powerful prediction models. Breiman [44] develops random forests as an improvement of bagging by adding additional diversity to the decision tree models. In a random forest, each tree is built using a random subsample of the full feature set. Typically, in each split, the number of predictors considered is approximately equal to the square root of the total number of predictors. Compared to bagging, this has the effect of decorrelating the trees. After the ensemble of trees (the forest) is obtained, the model uses a vote to combine the trees' predictions. The instances are classified in the class that obtains the greatest number of votes from the trees that make up the forest. Random forests are viewed as one of the most popular machine learning algorithms due to their power, versatility, and ease of use.

3.4. Imbalanced data sets

Imbalanced data sets often occur in the real world, where the relevant category is usually the one that has a significantly lower percentage of instances [45]. According to López et al. [46], the machine learning community has addressed the issue of class imbalance using three basic strategies. The first consists in balancing the training set by undersampling the majority class, oversampling the minority class or generating synthetic data in the minority class. The second focuses on modifying the algorithm through an adjustment in the precision threshold or making a change to make it more sensitive to the minority class. The third seeks to make the learning cost-sensitive to errors arising especially in the minority class.

Undersampling can be used with very large datasets and applied to the majority class, reducing the number of instances of this class to balance it with the minority class. Since this method discards most of the instances of the majority class, information that could be relevant in the training set is lost. Oversampling works with the minority class, which is increased to balance it with the majority class. In this case no information is lost, but the training set is increased by copying and pasting observations of the minority class, which could lead to other problems.

Although these two techniques are easy to apply, given the scant presence of fraudulent companies in our base it is advisable to use other methods. In this work, we use a synthetic data generation method and a cost-sensitive learning method as alternatives to not balancing the sample, and we study the improvement that occurs with respect to using, directly with the original data, the machine learning techniques mentioned above. Recent examples with these approximations can be found in Rivera and Xanthopoulos [47] and Sahin et al. [48].

3.4.1. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE combines synthetic oversampling of the minority class with undersampling of the majority class as a tool to balance the sample [49]. This technique is based on a form of oversampling that provides new information relative to the minority class through the random generation of new instances in the minority class.

To generate the synthetic random set, SMOTE makes use of bootstrap and k-NN (k-Nearest Neighbours algorithm), combined. The minority class is oversampled by introducing synthetic examples by convex combining each minority class instance with a random sample drawn with replacement among its k minority class nearest neighbours in the feature space.

3.4.2. Cost matrix (cost sensitive learning)

This technique does not create distributions of balanced data, but seeks to balance learning by applying a cost matrix that accounts for the cost of an erroneous classification versus a correct one. This technique applies smaller costs (weights) to the instances of the majority class and larger costs to those of the minority class. The weights can be set to be inversely proportional to the fraction of instances of the corresponding class. In the case of binary classification the weight of a class can be adjusted by resampling to improve the predictive power of the model.

In fraud detection, identifying a company that has committed fraud as fraudulent or non-fraudulent to one that has not been fraudulent carries no cost. However, the cost associated with identifying a company that has committed fraud as negative (false negative) carries more costs than identifying a company that has not caused fraud as positive (false positive). The cost matrix is similar to the confusion matrix. The aim is to penalize the errors (false positives and false negatives) against the correct ones (true negatives and true positives). In our application, we do not penalize correct ones and we assign a higher cost to errors in the minority class than to errors in the majority class.

3.5. Assessing predictive accuracy

The performance of machine learning algorithms is typically assessed using global predictive accuracy. In other words, they are evaluated comparing aggregate proportions of successes (true positives and true negatives) and errors (false positive and false negatives). In our case, a true positive occurs when a fraudulent company is properly identified as fraudulent, a true negative is a case of a legal company correctly labelled, and false positives and false negatives are, respectively, cases of either legal and illegal companies wrongly identified, respectively, as illegal and legal.

Assessing accurateness using global predictive accuracy, however, is not advisable when the consequences of the different kinds of errors vary markedly and/or when the data is imbalanced [50]. In these scenarios, López et al. [46] consider that adequate figures of accuracy can be achieved by properly combining the rates of true positives (TP), true negatives, false positives (FP) and false negatives. A good approximation to synthesize these measures, centred on the group of interest, is the ROC (Receiver Operating Characteristic) curve. This curve shows the relationship between benefits (TP rate) and costs (FP rate) for different classification thresholds. No classifier can increase the number of TPs without, at the same time, increasing the FP rate. The area under the ROC curve (AUC) provides an overall measure of which model is best on average. The closer to 1 the AUC value is, the better the model.

In order to evaluate the predictive capacity of the different models, we have applied the traditional method of splitting the initial set into two subsets, one of training (70% of the data) and one of test (30%). If the model selected with the training data has a good predictive ability it will be able to correctly classify the instances of the test set. Since the number of original positives in the dependent variable is very low and the procedure of increasing the original data is revealed as the most advisable, we have chosen to repeat the procedure 10 times to assess the robustness of the conclusions. That is, 10 sets of training data with their corresponding sets of test data are randomly selected on the original data set. Cross-validation was used for the training sets to fit the models. The measures selected to evaluate the accuracy of the predictions are: (i) the area under the ROC curve, (ii) the Kappa statistic, and (iii) RMSE (Root Mean Squared Error).

These measures have been used to compare the different procedures (combination of automatic learning method and data balancing method) with respect to their predictive capacity in both training data (explanation) and in the test data (prediction). Some models may well explain the data with which they have been trained but may not be able to predict positives in the test set. Although a greater explanatory capacity than predictive capacity is to be expected, what is important, however, is the sensitivity of the model to the data of training and verification.

4. Data and methods

As in all research using real data, much of the work has been devoted to purification and treatment. The quality of any analysis relies heavily on the quality of the data used. This section describes the database, the treatment criteria implemented, the selection of variables and the characteristics of the learning groups and test group built.

4.1. The data set

The criminal case being analysed is one of the most voluminous cases of money laundering in Spanish history, both in terms of economic value and in the number of companies involved. The alleged criminal network consisted of a large number of suppliers, structured in clans and hierarchically organized, all of them coordinated by a single core company. This company, in turn, was part of an international business group which used obscured money trails to finance the core company. During the four years that the core company was operating, huge amounts of money circulated between the core company and its suppliers. In this period, the core company offered all its suppliers (legal and illegal) prices above the market price, always operating at a loss compared to a normal commercial activity. The aggressive pricing strategy allowed it (starting from zero) to take over 50% of the Spanish market share during its period of activity. In addition to funding, the origin of some of the goods being marketed was also investigated by the police. In some operations, there were even doubts that the merchandise actually existed.

The investigation collated a large database containing 285,774 commercial operations carried out by 643 suppliers with the core company. The variables available in the database (to which we do not have full access) include for each operation: which item was marketed, where it was marketed, who recorded the operation, the cost and quantity associated with the operation, the date of the operation and the supplier ID. This study is based on the analysis of the amounts of commercial operations, represented by analysis of the variable "Amount of the Operation". Table 2 provides a summary, in an undetermined unit of measure, of the distribution of this variable classified according to whether the supplier had been previously identified as fraudulent by the police.

Table 2
Distributions of the variable "Amount of the Operation" by type of supplier.

	# of operations	Min	P_{10}	P_{20}	P_{30}	P_{40}	Median	Mean	P_{60}	P_{70}	P_{80}	P_{90}	Max
Fraudulent	47381	0.01	0.54	1.39	2.69	4.34	6.44	13.54	9.20	13.17	18.90	31.17	496.00
Other	238393	0.00	0.11	0.28	0.64	1.36	2.58	10.51	4.51	7.14	11.53	20.20	>10000.00

Of the total number of suppliers, we are only certain that 26 of them are fraudulent, that is, they carry out operations that do not comply with the law. These are money laundering operations linked to economic gains made through the commission of other highly profitable crimes. This *a priori* information is provided by the police authorities based on the investigation of the associated judicial process.

A priori information only provides certainty that 4% of the suppliers are fraudulent; the fraudulent or non-fraudulent status of other companies not being known. When the classic goodness-of-fit tests are applied, the hypothesis that the data relating to commercial operations conform to Benford's Law is massively rejected. This result can be misleading since, as discussed in Section 3.2, the traditional tests do not serve as a measure to evaluate the fit in very large samples. In fact, if we compare graphically the fit of the whole sample to Benford's Law of the first two leading digits (see Fig. 1), we see that the data does seem to fit this distribution, despite the classic tests systematically reject the null hypothesis. This generalized rejection does not occur using the OverBenford test proposed in this paper. For instance, the p-value of testing the fit of the whole dataset to the first digit Benford's Law is 0.2303.

In any case, to evaluate if a company should be included or not in the group of defrauding companies, using exclusively the annotations corresponding to its operations with the core company—i.e., without using other external information (such as wiretaps or criminal records)—, it seems reasonable to have a minimum amount of data about the company: a minimum number of operations. Hence, we have screened the initial database to consider exclusively those companies for which a minimum number of operations (variable Frequency) exist. Specifically, companies that have performed at least 195 operations have been incorporated into the automatic detection analysis. This decision led to the exclusion of 3 suppliers from the sample of 26 initially identified as fraudulent by the police. So, the final analysis focused on 335 companies having the higher number of operations, which represent a total of 245,227 operations. Of the 335 companies studied, police experts identified 23 of them as fraudulent. That is, 6.87% of the instances belong to the minority class. This means that we have an imbalanced data set, and it is therefore appropriate to apply strategies to the sample such as those described in Section 3.4.

The decision to use 195 operations as the cut-off point was a compromise between keeping as many companies as possible in the final dataset and having a minimum number of operations for a statistically proper fit to the Benford's Law. Without this initial screening, the sample size for some leading digits would have been too small for some suppliers; an issue that would result in relative sample standard deviations too high for the larger digits and high volatilities in the p-values. For instance, in a sample of 195 operations, the number nine will appear on average 8.95 times (rounded up to 9) as leading digit, given that according to Benford's Law the probability of observing 9 as leading digit is just 4.59%.

In the final dataset, the average number of operations of a company identified as fraudulent is 2042.09 operations, while the average of other companies is 635.45 operations. This is consistent with what has been discussed in previous literature, which has revealed that companies that commit financial fraud and money laundering often show a certain predisposition to performing a large number of operations [51]. In this type of crime, criminals tend to perform as many operations as possible with the dual objective of hiding the fraud strategy and laundering as much money as possible. It has been decided, therefore, to include the frequency variable as predictor. As shown below, this variable has a high correlation with the fraud variable.

To sum up, for each of the companies analysed we have as predictors the number of operations registered (variable Frequency) and a set of 20 p-values—corresponding to Z-tests (where F1–F9 denotes the p-values associated with the first digits and S0 to S9 those linked to the second digits) and to the OverBenford test—, and as response variable the police consideration of each company as fraudulent or non-fraudulent (variable Invest). A detail of the marginal distributions of such variables is given in Fig. 2. The aim is to analyse if there is any type of behaviour pattern that distinguishes companies that commit fraud from those that do not in order to highlight other companies that might merit scrutiny by police and judicial authorities.

Finally, as one reviewer pointed out, it is interesting to note that we could also have considered two further variables as potential predictors. In addition to the above variables, we could have also computed a variable measuring the time of commercial relationship of each supplier with the core company (calculated as the difference between the date of the last operation and the date of the first operation) and another variable measuring the density of transactions over time. Indeed, as can be observed in Fig. 3-right, companies identified as fraudulent show, on average, shorter periods of activity. Furthermore, as is usual in money laundering cases and in our case too, it seems there are certain suppliers who are no more than a succession of ephemeral companies whose sole purpose is to obscure the audit trail (see Fig. 3-left).

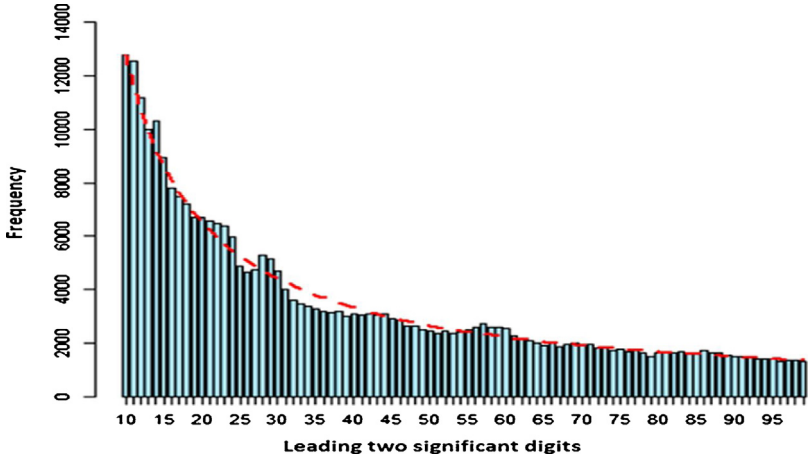

Fig. 1. Fit of the data to Benford's first two leading digit frequency distribution.

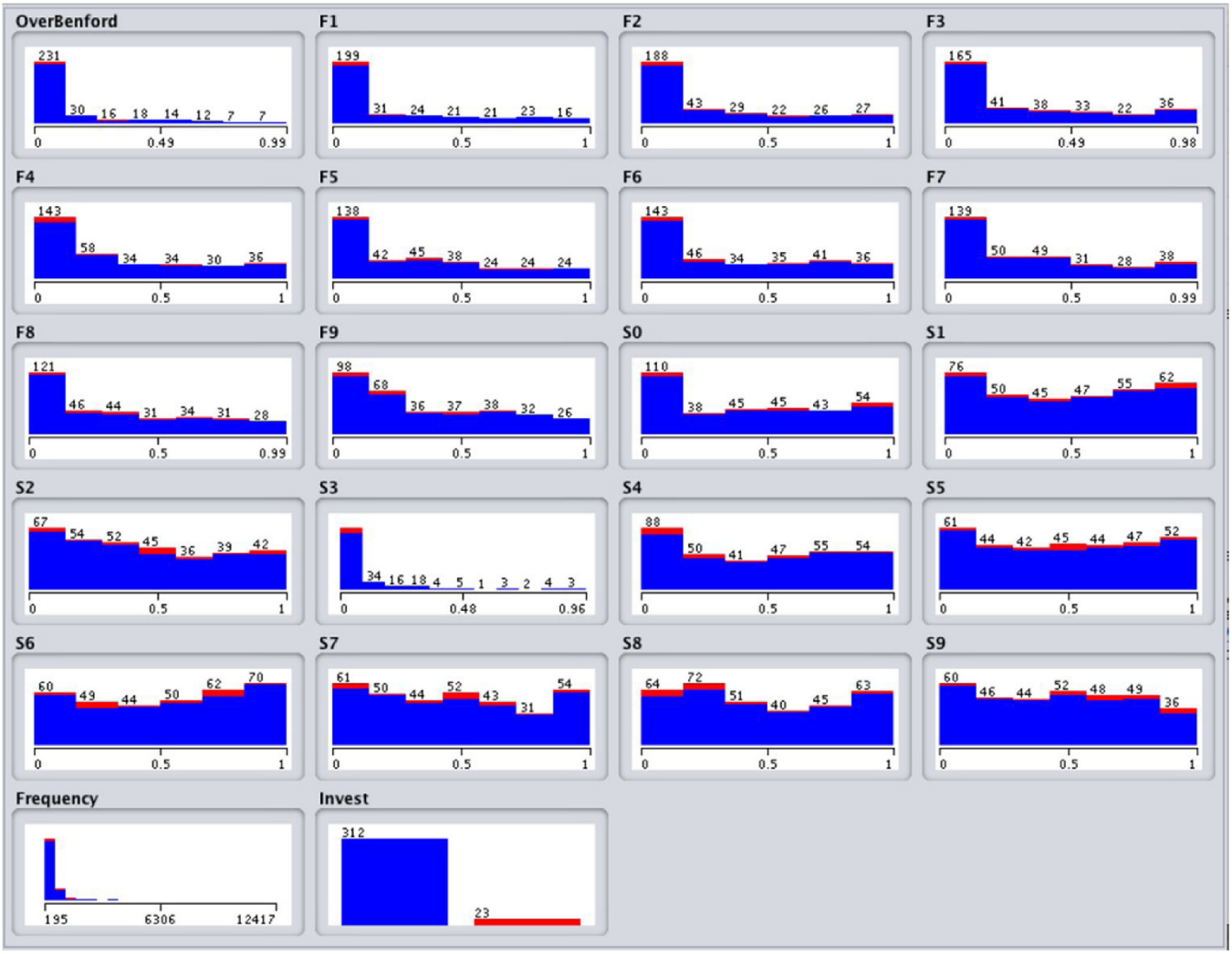

Fig. 2. Marginal frequency distribution of predictors and response variable. Weka output.

4.2. Selecting predictors

Although the machine learning procedures considered are designed to avoid the dangers of a high number of predictors in terms of multicollinearity and overparameterization, it is sometimes better to make a previous selection of predictors [52]. This is the case in this study, where we found that the models showed a lower predictive capacity without a previous selection of features. Hence, we have made a previous selection among the set of potential predictors using the Weka Ranker Search Method algorithm; which is based on correlations between predictors and response variable. Table 3 gives the predictors ordered by degree of relation to the response variable. From the total of predictors initially considered, we have included

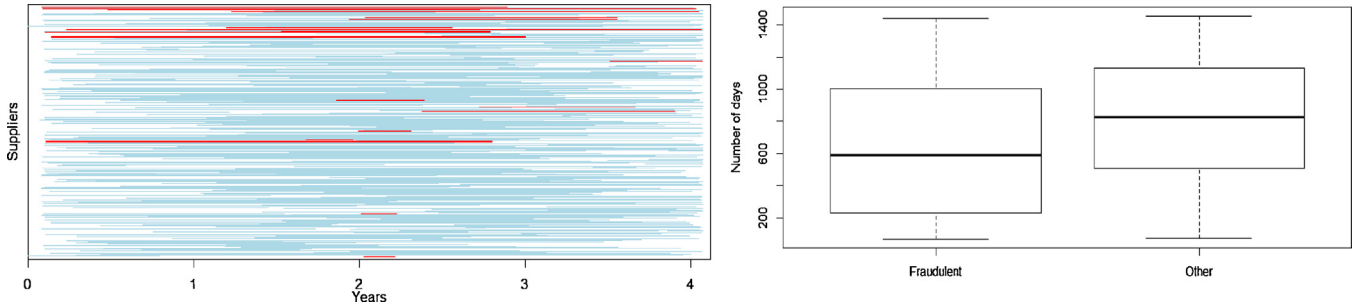

Fig. 3. Summaries of periods of activity between suppliers with more than 195 operations and the core company. Left-panel: Each line represents one supplier, in order of the supplier with the highest number of operations at the top and the lowest at the bottom. Each line starts with the date of the first operation of the corresponding supplier with the core company and ends with the date of the last operation. In red are those suppliers identified as fraudulent by the police. Right panel: boxplots of the distribution of the number of days of activity classified by type of supplier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Correlation ranking of the predictors. Output of Weka.

Frequency	F7	S4	S9	S8	OverBenford	S3	F3	F9	S2	F5
0.3157	0.1392	0.1348	0.1281	0.1060	0.1011	0.0911	0.0713	0.0678	0.0658	0.0573
S0	F2	F8	F6	S7	S5	S6	F1	F4	S1	
0.0424	0.0389	0.0360	0.0308	0.0274	0.0226	0.0198	0.0189	0.0166	0.0137	

in the models those that exceeded the value 0.05. In total, there are 11 predictors.

4.3. Splitting the sample

Once the predictors have been selected, the base is composed of 335 companies, 11 predictors and a response variable. These data are those on which the methodologies described in Section 3 have been applied.

Since the analysis (see Section 5) has been structured in two parts, we have also used two different strategies to divide the sample. On the one hand, the initial fit of the models is done with the complete base, using the machine learning approach based on cross-validation that allows determination of the tuning parameters that for each algorithm offer a greater predictive capacity in the training set. Specifically, we have randomly divided the learning set into 10 folds, each with 10% of the data. On the other hand, in order to analyse the impact that using the SMOTE strategy has on the predictive quality of the different models, we have also used the classic strategy of dividing the data into two groups, one of training (70% of the instances) and the other of test (30% of the instances). This last process has been repeated ten times (applying in all cases 10-fold cross-validation to fit the models with the training sets) to discount the possibility that the solutions could be dependent on the actual partition carried out. In all cases, instead of using the option of a stratified selection of the instances, with the categories as strata, the groups have been constructed by extracting instances randomly from the total data set. This has caused some variability in the internal composition of the groups. For example, training samples composed of 70% of the cases vary between a minimum of 14 positive cases (5.98%) and a maximum of 19 (8.12%).

5. Results

This section discusses the results. We have grouped the analyses into two subsections. The first subsection focuses on evaluating the explanatory/predictive capacity of the models and the impact of the different solutions implemented to deal with the challenge that entails working with such imbalanced data. From this analysis, we deduce that the SMOTE strategy, based on the generation of synthetic instances of the minority class, is the one that produces the best results. Thus, a second subsection is included, in which we perform an analysis of the sensitivity of the predictive power of the models when the original data are balanced using the SMOTE strategy.

In the first block of analysis, the different models are trained using cross-validation, after dividing the sample into 10 subsamples of equal size. In the second block of analysis, the initial division of the data set into two subsets (one of training, 70%, and another of test, 30%) was repeated 10 times, with the models again trained using cross-validation models on the training sets. On each of the training sets, we have applied the transformation of the data using the SMOTE algorithm to balance the target variable.

5.1. Assessing the models

The explanatory capacity of each of the four machine learning algorithms analysed has been evaluated under each of the three scenarios considered in terms of imbalanced data: (i) directly modelling the data, without considering the imbalance presented by the two categories of the response variable; (ii) using cost-sensitive learning (cost matrix); and (iii) applying a transformation to the data using the SMOTE algorithm to balance the dependent variable.

5.1.1. Imbalanced dataset

The results of training the models using the complete database without performing any transformation on the data (i.e., without taking into account the imbalance presented by the two categories of the response variable) are presented in Table 4.

The explanatory ability of the models is very high, but they present very low true positive rates in the target category, ranging from 13.04% for the logistic regression model to 34.78% for the neural network approach. By having such imbalanced data, the algorithms tend to favour classification in the dominant category, identifying very few fraudulent companies.

5.1.2. Cost matrix

In a cost-sensitive fit, we assume that the losses of an incorrect classification are asymmetric. The cost of checking the misclassification of the false positives would be minimal, while false negatives would entail a much higher cost, not only in terms of tax, but also in economic and security terms. The cost matrix allows the process to be balanced without having to perform any type of transformation on the data. Table 5 gives the results of training the models using the cost matrix.

On comparing the results of Tables 4 and 5, it is clear that there is a significant decrease in the overall accuracy of the models. The random forest model is the only algorithm that maintains 94% of instances correctly classified, while the rest of the models show a lower figure, with the logistic regression model showing a minimum of 73.73%. However, the rate of true positives improves substantially. The consequence of identifying a positive as a negative is that the algorithms correctly identify more companies as defrauders.

The cost-based approach has increased the detection of true positives at the expense of significantly raising false positives. The

Table 4
Confusion matrix of the models. Imbalanced dataset.

	LR		DT		NN		RF	
	No	Yes	No	Yes	No	Yes	No	Yes
No	311	1	302	10	301	11	312	0
Yes	20	3	17	6	15	8	19	4
Correctly classified	93.73%		91.94%		92.24%		94.33%	
Incorrectly classified	6.27%		8.06%		7.76%		5.67%	
TN rate (No)	99.68%		96.79%		96.47%		100.00%	
TP rate (Yes)	13.04%		26.09%		34.78%		17.39%	
FN rate (Yes)	86.96%		73.91%		65.22%		82.61%	
FP rate (No)	0.32%		3.21%		3.53%		0.00%	

Table 5
Confusion matrix of the models. Cost Matrix.

	LR		DT		NN		RF	
	No	Yes	No	Yes	No	Yes	No	Yes
No	234	78	290	22	285	27	309	3
Yes	10	13	16	7	15	8	17	6
Correctly classified	73.73%		88.66%		87.46%		94.03%	
Incorrectly classified	26.27%		11.34%		12.54%		5.97%	
TN rate (No)	75.00%		92.95%		91.35%		99.04%	
TP rate (Yes)	56.52%		30.43%		34.78%		26.09%	
FN rate (Yes)	43.48%		69.57%		65.22%		73.91%	
FP rate (No)	25.00%		7.05%		8.65%		0.96%	

only case where this does not happen is in the random forest model, which shows the least true positives but also less false positives.

5.1.3. SMOTE balance

One of the transformations of the data most used for balancing the categories is the algorithm SMOTE. Based on the above data, where there were 312 legal and 23 fraudulent companies, new instances of fraudulent companies had been synthetically generated, up to a total of 299. This gives a 51.06% "no" and a 49.94% of "yes". The models were subsequently trained on this new set. Table 6 provides a summary of outcomes.

The SMOTE approach does not substantially improve the overall explanatory ability of the models compared to the use of the cost matrix. However, the true positive rate has improved in all cases. With the original data, true positive rates ranged between 13.04% for logistic regression and 34.78% for neural network. With the cost matrix, results improved, reaching 26.09% for random forest and 56.52% for logistic regression. With the SMOTE transformation, true positive rates range from 82.27% for the logistic regression model to 94.98% for random forest.

The capacity of this third balancing strategy to identify fraudulent companies is greatly superior to the two previous ones; obtaining highly satisfactory results in the case of random forest. Of 611 instances, it only incorrectly classified 27 (4.42%), of which 15 were false negatives and 12 false positives.

5.1.4. Measurements of precision

Finally, the measurements of the ROC area, the Kappa statistic, and RMSE (Root Mean Squared Error) are taken in order to evaluate the different procedures as well as the joint action of models and techniques of balancing, as shown in Table 7.

From Table 7 it follows that the best results are obtained when the sample is balanced using SMOTE. When the original data is used with or without cost matrix the fit worsens. As for the classification algorithm, the best results are obtained for random forest, with a ROC area of 0.989 and a Kappa statistic of 0.912 (in both cases very close to 1) and the lowest RMSE compared to all the other models applied.

Table 6
Confusion matrix of the models. SMOTE.

	LR		DT		NN		RF	
	No	Yes	No	Yes	No	Yes	No	Yes
No	239	73	269	43	252	60	300	12
Yes	53	246	31	268	38	261	15	284
Correctly classified	79.38%		87.89%		83.96%		95.58%	
Incorrectly classified	20.62%		12.11%		16.04%		4.42%	
TN rate (No)	76.60%		86.22%		80.77%		96.15%	
TP rate (Yes)	82.27%		89.63%		87.29%		94.98%	
FN rate (Yes)	17.73%		10.37%		12.71%		5.02%	
FP rate (No)	23.40%		13.78%		19.23%		3.85%	

Table 7
Measurements of precision of the procedures.

	Imbalanced dataset			Cost matrix			SMOTE		
	ROC	Kappa	RMSE	ROC	Kappa	RMSE	ROC	Kappa	RMSE
LG	0.747	0.2061	0.2360	0.711	0.3243	0.4227	0.844	0.5675	0.4012
DT	0.635	0.2664	0.2702	0.615	0.2086	0.3320	0.894	0.7348	0.3499
NN	0.765	0.3400	0.2578	0.630	0.2104	0.3306	0.926	0.7252	0.3392
RF	0.740	0.2817	0.2268	0.773	0.3499	0.2415	0.989	0.9116	0.2088

5.2. Sensitivity analysis

The results of true positives detected when balancing the training data with the SMOTE algorithm are the ones with the most promising outputs. Therefore, in this subsection, we verify the predictive ability of the models based on a set of data independent of the one used for training.

With that purpose, we have performed 10 random divisions of the initial set of data, generating 10 pairs of subsets with respectively 70% of the cases (training) and 30% of the instances (test), and have applied the same methods as before with SMOTE. The number of fraudulent companies in training sets ranges from a minimum of 14 to a maximum of 19, while in the test set it is the inverse, ranging from a maximum of 9 to a minimum of 4.

Due to the different compositions of the different training and test sets, the subsets are not comparable in terms of the number of fraudulent and non-fraudulent instances and so Tables 8 and 9 do not have "yes/no" matrices but only show the means for cases classified correctly and incorrectly. Table 8 focuses on training sets and Table 9 is dedicated to test sets.

Although we have fewer instances to train the models, comparing the results of Tables 6 and 8 we observe that the explanatory ability (training set) of the models is very similar. If we now evaluate the predictive ability (test set) of the different models (see Table 9), we see that it has been reduced, especially in the true positives; particularly striking is the loss that occurs in the random forest model. The random forest goes from registering the best results with the training set to the worst ones with the validation set. In the average of the 10 divisions, neural network and logistic regression are the techniques that best predict in the test set.

Precision measurements draw the same results as the matrices of confusion. The fit for the training data is significant, random forest obtaining an average of the ROC area of 0.9873 with a range of 0.019. The Kappa statistic also draws the best results in random forest, with a very high value (0.89024) and small range (0.1103). The methodology that minimizes errors is again random forest with an RMSE mean of 0.22126 and a range of 0.0688.

When comparing the precision measurements of the training set with the test set (see Table 10), a significant loss of precision is observed, albeit different for each method. Taking the mean of the ROC area as a reference, the greatest loss is for random forest with a difference of 0.2331 between the training set and the test set. The smallest difference was observed in logistic regression, which was the one that obtained worse results in the training set. Random

Table 8
Average of confusion matrix summaries of the models (training set).

	LR	DT	NN	RF
Correctly classified	78.06%	88.31%	86.60%	95.16%
Incorrectly classified	21.94%	11.69%	13.40%	4.84%
TN rate (No)	77.01%	86.55%	83.60%	93.87%
TP rate (Yes)	79.09%	90.03%	89.54%	96.16%
FN rate (Yes)	20.91%	9.97%	10.46%	3.84%
FP rate (No)	22.99%	13.45%	16.40%	6.13%

Table 9
Average of confusion matrix summaries of the models (test set).

	LR	DT	NN	RF
Correctly classified	76.04%	83.66%	82.41%	90.40%
Incorrectly classified	23.96%	16.34%	17.59%	9.60%
TN rate (No)	77.41%	86.43%	84.13%	94.38%
TP rate (Yes)	56.72%	44.78%	58.21%	34.33%
FN rate (Yes)	43.28%	55.22%	41.79%	65.67%
FP rate (No)	22.59%	13.57%	15.87%	5.62%

forest is the one that has a greater loss in the Kappa statistic, the difference being 0.62105. Finally, the increase in the mean of the RSME is very similar for decision tree, neural network and random forest, being very small for logistic regression with only a slight increase of 0.0102.

As the ranges (the difference between the maximum and the minimum) of the precision measurements in the training sets are narrow, it seems the models are fairly stable and do not depend on the training data set selected. However, the ranges of precision measurements in the test set are much broader, depending on the predictive ability of the model in the set of data selected. This result can lead to question the robustness of the approaches. An alternative explanation is that the models are identifying (in the test sets) more actual fraudulent companies than initially identified, suggesting that as suspected even more companies may have perpetrated criminal acts. The implemented procedures provide a guide for law enforcement investigators to focus their inquiries and their resources on those companies that are systematically identified as fraudulent by a variety of methods.

6. Discussion and conclusions

Many real financial and economic datasets conform to Benford's Law, but this is not widely known. Hence, under the assumption that it is highly unlikely that the fit to the Benford distribution would be preserved when people fabricate data, Benford's Law has been used as a tool to detect accounting irregularities. In this work, we combine Benford's Law and machine learning algorithms as a tool to detect money laundering criminals in the context of a real Spanish court case.

To this end, we analyse the hidden accounting of a company investigated for money laundering and characterize, based on Benford's Law, the operations carried out by each of its suppliers by projecting all of its operations in a space of 21 dimensions. The information provided by the police about the fraudulent status of a subgroup of companies is used to detect, using machine learning models, other companies that could be potentially fraudulent. The proposed procedures have to consider the problem of working with highly imbalanced categories.

Regarding the machine learning algorithms analysed, Neural Networks produces the best performance when we work with the imbalanced dataset. This method shows the highest figures for the ROC area (0.765) and the Kappa statistic (0.340), but the smallest RMSE (0.227) is reached with Random Forest. Indeed, Random Forest is revealed as the best approach (among the ones tested) when balancing techniques are applied. When the cost matrix is employed, the best second option is Ridge Logistic Regression, despite showing the highest RMSE (0.423). However, using SMOTE, Neural Networks becomes the second best option, with Ridge Logistic Regression being the method that identifies fewer true positives.

The results show that, in general, the strategy used to balance the sample has more weight than the machine learning algorithm. The SMOTE strategy improves noticeably the results of all the algorithms. The SMOTE strategy achieves better results on the true positives than the cost matrix. However, the overall accuracy of the model is very similar, so the proportion of false positives increases with the SMOTE methodology. As a rule, the worst results with SMOTE are better than the best result with cost matrix. Only Random Forest detects more true positives when cost matrix is employed. The cost matrix identifies fewer positives, and consequently, fewer companies would be investigated. The explanatory capacity of Random Forest with SMOTE is very high (ROC, 0.989). However, its predictive capacity drops to a ROC of 0.789, although it still shows the best figures. Random Forest obtains the highest Kappa statistic and the smallest error, measured through RMSE, on the test group.

After applying the different procedures, we end up with a list of potential suppliers that were not initially identified as fraudulent but warrant further scrutiny. More specifically, after predicting the condition of fraudulent or non-fraudulent of each supplier by using SMOTE and the four machine learning algorithms, we find a total of 119 suppliers identified as potentially fraudulent by at least one of the strategies. Of these 119, however, only 44 are identified by at least two methods. These numbers drop to 12 and 4 when looking at companies identified as fraudulent by three or four methods, respectively. The next step is to decide which of these companies to investigate. Selecting a large number of companies to investigate would have two negative points. The first would be an increase in investigation costs: more companies investigated means more time, more personnel and more resources. The second is the disruption caused to companies deemed legal, who have to endure an investigation and the intrusion that accompanies it. In our opinion, the best solution is to focus inquiries and resources on those companies that are systematically identified as fraudulent by a variety of methods. Police experts could make the final decision as to which companies warrant more in depth scrutiny given that sometimes they have qualitative information that is not captured by the data available.

In this paper, we propose a strategy for detecting, in the context of a real court case, fraudulent suppliers of a big company investigated for money laundering. We do this by combining Benford's Law and machine learning algorithms using exclusively the data provide by the (amounts of the) operations performed between the core company and each of its suppliers. Although the case dealt with here is exceptional in the sense that we had access to the hidden accounting of the core company, we believe that some of the ideas implemented in this research would be transferable to other cases of money laundering and fraud given

Table 10
Summary of measurements of precision over the test sample.

	ROC			Kappa			RMSE		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
LG	0.7787	**0.708**	0.881	0.14947	0.0555	0.2914	0.40051	0.3493	0.4661
DT	0.6570	0.558	0.796	0.18590	**0.0901**	0.2930	0.39203	0.3226	0.4372
NN	0.7361	0.508	0.948	0.22728	0.0249	0.3671	0.39697	0.3397	0.4651
RF	**0.7886**	0.620	**1.000**	**0.26919**	0.0804	**0.5976**	**0.27975**	**0.2388**	**0.3224**

The best figures for each combination of precision and summary measures have been highlighted.

the eternal repetition of some patterns in criminal acts. In any case, as fraud is a dynamic process and this criminal activity is one of the most creative in the world (with some frauds only committed once or twice), whatever the strategy used, this must evolve to adapt to changes in criminal organizations and available data. This is indeed the strategy followed in Belgium to combat tax fraud, where since 2001 different agencies have been developing adaptive systems to fight against tax evasion and fraud. In particular, by integrating data mining and supervising machine learning tools, they use subsector specific statistical models that are periodically revised in an attempt to parameterize the evolution of the modus-operandi of criminals [53].

Acknowledgements

The authors wish to thank two anonymous referees for their valuable comments and suggestions and M. Hodkinson for translation of the paper into English. This work has been supported by the Spanish Ministry of Economics and Competitiveness under grant CSO2013-43054-R.

References

[1] H.R. Varian, Benford's Law, Am. Stat. 26 (3) (1972) 65–66.
[2] S. Newcomb, Note on the frequency of use of the different digits in natural numbers, Am. J. Math. 4 (1881) 39–40.
[3] M.J. Nigrini, The detection of income escape through an analysis of digital distributions, PhD Thesis, University of Cincinatti, 1992.
[4] N.L. Khac, M. Kechadi, Application of data mining for anti-money laundering detection: a case study, The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 2010.
[5] B. Unger, The Scale and Impact of Money Laundering, Edward Elgar, Cheltenham, UK, 2007.
[6] B. Unger, J. Hertog, Water always finds its way: identifying new forms of money laundering, Crime Law Soc. Change 57 (2012) 287–304.
[7] J. Walker, B. Unger, Measuring global money laundering: the Walker Gravity Model, Rev. Law Econ. 5 (2009) 821–853.
[8] M. Cardoso, Blanqueo de capitales: técnicas de blanqueo y relación con el sistema tributario, Agencia Estatal de Administración Tributaria, 2015 19/2015.
[9] W. Alhosani, Anti-money laundering, A Comparative and Critical Analysis of the UK and UAE's Financial Intelligence Units, Springer, 2016.
[10] IBA, ABA, CCBE, A Lawyer's Guide to Detecting and Preventing Money Laundering. International Bar Association, American Bar Association and Council of Bars and Law Societies of Europe, 2014.
[11] J. Torres, S. Fernández, A. Gamero, A. Sola, How do numbers begin? (The first digit law), Eur. J. Phys. 28 (3) (2007) 17–25.
[12] B. Luque, L. Lacasa, The first-digit frequencies of prime numbers and Riemann zeta zeros, Proc. R. Soc. Lond. 465 (2009) 2197–2216.
[13] W. Mebane, Election forensics: statistics, recounts and fraud, Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12–16, 2007.
[14] W. Mebane, R.M. Alvarez, T.E. Hall, S.D. Hyde, Election forensics: the Second Digit Benford's Law Test and recent American presidential elections, Election Fraud, Brookings, Washington, DC, 2008.
[15] L. Pericchi, D. Torres, Quick anomaly detection by the Newcomb—Benford Law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela, Stat. Sci. 26 (4) (2011) 502–516.
[16] J. Deckert, M. Myagkov, P.C. Ordeshook, Benford's Law and the detection of election fraud, Polit. Anal. 19 (3) (2011) 245–268.
[17] M.J. Hickman, S.K. Rice, Digital analysis of crime statistics: does crime conform to Benford's Law? J. Quant. Criminol. 26 (2010) 333–349.
[18] T. Revell, Man vs Maths: Understanding the Curious Mathematics That Power Our World, Quarto Knows, London, 2016.
[19] A. Diekmann, Not the first digit! Using Benford's Law to detect fraudulent scientific data, J. Appl. Stat. 34 (3) (2007) 321–329.
[20] G. Judge, L. Schechter, Detecting problems in survey data using Benford's Law, J. Hum. Resour. 44 (1) (2009) 1–24.
[21] S. de Marchi, J.T. Hamilton, Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory, J. Risk Uncertain. 32 (1) (2006) 57–76.
[22] M.J. Nigrini, Using digital frequency to detect fraud, The White Paper, April, 1–3, 1994.
[23] M.J. Nigrini, L.J. Mittermaier, The use of Benford's Law as an aid in analytical procedures, Auditing J. Pract. Theory 16 (2) (1997) 52–67.
[24] M.D. Beneish, The detection of earnings manipulation, Financ. Anal. J. 55 (5) (1999) 24–36.
[25] C. Durtschi, W. Hillison, C. Pacini, The effective use of Benford's law to assist in detecting fraud in accounting data, J. Forensic Account. 5 (1) (2004) 17–34.
[26] A. Asllani, M. Naco, Using Benford's Law for fraud detection in accounting practices, J. Soc. Sci. Stud. 2 (1) (2014) 129–143.
[27] R. Quick, M. Wolz, Benford's Law in deutschen Rechnungslegungsdaten, Betriebswirtschaftliche Forschung und Praxis 2 (2003) 208–224.
[28] W.K. Tam, B.J. Gaines, Breaking the (Benford) law: statistical fraud detection in campaign finance, Am. Stat. 61 (3) (2007) 218–223.
[29] B. Rauch, M. Göttsche, G. Brähler, S. Engel, Fact and fiction in EU-Governmental economic data, Ger. Econ. Rev. 12 (3) (2011) 243–255.
[30] F.A. Alali, S. Romero, Benford's Law: analyzing a decade of financial data, J. Emerg. Technol. Account. 10 (1) (2013) 1–39.
[31] S. Günnel, K.H. Tödter, Does Benford's Law hold in economic research and forecasting? Empirica 36 (3) (2009) 273–292.
[32] D. Ramos, Fraude: un nuevo enfoque para combatirlo, Auditoría Pública 38 (2006) 99–104.
[33] F. Benford, The law of anomalous numbers, Proc. Am. Philos. Soc. 78 (4) (1938) 551–572.
[34] T. Hill, A statistical derivation of the significant-digit law, Stat. Sci. 10 (1995) 354–363.
[35] R.S. Pinkham, On the distribution of first significant digits, Ann. Math. Stat. 32 (4) (1961) 1223–1230.
[36] T. Hill, The significant-digit phenomenon, Am. Math. Mon. 102 (4) (1995) 322–327.
[37] T. Hill, The first digit phenomenon, Am. Sci. 86 (1998) 358–363.
[38] W.A. Wagenaar, Generation of random sequences by human subjects: a critical survey of the literature, Psychol. Bull. 77 (1) (1972) 65–72.
[39] D.E. Giles, Benford's Law and naturally occurring prices in certain eBay auctions, Appl. Econ. Lett. 14 (3) (2007) 157–161.
[40] J.M. Pavía, Testing goodness-of-fit with the kernel density estimator: GoFKernel, J. Stat. Softw. 66 (1) (2015) 1–27.
[41] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd edn., Chapman and Hall, London, 1989.
[42] S. Le Cessie, J.C. van Houwekingen, Ridge estimators in logistic regression, Appl. Stat. 41 (1992) 191–201.
[43] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second edition, Springer, New York, 2009.
[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[45] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (1) (2006) 25–36.
[46] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[47] W.A. Rivera, P. Xanthopoulos, A priori synthetic oversampling methods for increasing classification sensitivity in imbalanced data sets, Expert Syst. Appl. 66 (2016) 124–135.
[48] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Syst. Appl. 40 (15) (2013) 5916–5923.
[49] N. Chawla, K.W. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: Synthetic Minority Oversampling Technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[50] I. Brown, C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Syst. Appl. 39 (3) (2012) 3446–3453.
[51] S.D. Demetis, Unfolding dimensions of an anti-money laundering/counter-terrorist financing complex system Lexis Nexis, Emerg. Issues 6019 (2011).
[52] J.H. Seo, D. Choi, Feature selection for chargeback fraud detection based on machine learning algorithms, Int. J. Appl. Eng. Res. 11 (22) (2016) 10960–10966.
[53] D.U. Potvin, Data Mining @ FPS Finance, 29, Tax Tribune, 2013, pp. 63–66.
[54] M.J. Nigrini, A taxpayer compliance application of Benford's Law, J. Am. Tax. Assoc. 18 (1) (1996) 72–91.
[55] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc., 1993, 1993.