WORKING PAPERS

UNIVERSITAT JAUME•I

Using machine learning for financial fraud detection in the accounts of companies investigated for money laundering

Economics Department

José A. Álvarez-Jareño
Elena Badal-Valero
José Manuel Pavía

2017 / 07

Castellón (Spain)

Using machine learning for financial fraud detection in the accounts of companies investigated for money laundering

José A. Álvarez-Jareño
Universitat Jaume I
Department of Economics
jareno@uji.es

Elena Badal-Valero
Universitat de València
Department of Applied Economics
Elena.Badal@uv.es

José Manuel Pavía
Universitat de València
Department of Applied Economics
Jose.M.Pavia@uv.es

2017 / 07

Abstract

Benford's Law is a well-known system use in accountancy for the analysis and detection of anomalies relating to money laundering and fraud. On that basis, and using real data from transactions undertaken by more than 600 companies from a particular sector, behavioral patterns can be analyzed using the latest machine learning procedures. The dataset is clearly unbalanced, for this reason we will apply cost matrix and SMOTE to different detecting patters methodologies: logistic regression, decision trees, neural networks and random forests.
The objective of the cost matrix and SMOTE is to improve the forecasting capabilities of the models to easily identify those companies committing some kind of fraud. The results obtained show that the SMOTE algorithm gets better true positive results, outperforming the cost matrix implementation. However, the general accuracy of the model is very similar, so the amount of a false positive result will increase with SMOTE methodology.
The aim is to detect the largest number of fraudulent companies, reducing, as far as possible, the number of false positives on companies operating correctly. The results obtained are quite revealing: Random forest gets better results with SMOTE transformation. It obtains 96.15% of true negative results and 94,98% of true positive results. Without any doubt, the listing ability of this methodology is very high.
This study has been developed from the investigation of a real Spanish money laundering case in which this expert team have been collaborating. This study is the first step to use machine learning to detect financial crime in Spanish judicial process cases.

Keywords: Benford's Law, unbalance dataset, random forest, fraud, anti-money laundering.

JEL classification: C14, C44, C53, M42

USING MACHINE LEARNING FOR FINANCIAL FRAUD DETECTION IN THE ACCOUNTS OF COMPANIES INVESTIGATED FOR MONEY LAUNDERING

JOSÉ A. ALVAREZ-JAREÑO
Departamento de Economía. Universitat Jaume I de Castelló
Avenida Vicent Sos Baynat, s/n, 12071 Castelló de la Plana
ELENA BADAL-VALERO
Departamento de Economía Aplicada. Universitat de València
Avenida de los Naranjos, s/n, 46022 Valencia
JOSE MANUEL PAVÍA-MIRALLES
Departamento de Economía Aplicada. Universitat de València
Avenida de los Naranjos, s/n, 46022 Valencia

Abstract

Benford's Law is a well-known system use in accountancy for the analysis and detection of anomalies relating to money laundering and fraud. On that basis, and using real data from transactions undertaken by more than 600 companies from a particular sector, behavioral patterns can be analyzed using the latest machine learning procedures. The dataset is clearly unbalanced, for this reason we will apply cost matrix and SMOTE to different detecting patters methodologies: logistic regression, decision trees, neural networks and random forests.

The objective of the cost matrix and SMOTE is to improve the forecasting capabilities of the models to easily identify those companies committing some kind of fraud. The results obtained show that the SMOTE algorithm gets better true positive results, outperforming the cost matrix implementation. However, the general accuracy of the model is very similar, so the amount of a false positive result will increase with SMOTE methodology.

The aim is to detect the largest number of fraudulent companies, reducing, as far as possible, the number of false positives on companies operating correctly. The results obtained are quite revealing: Random forest gets better results with SMOTE transformation. It obtains 96.15% of true negative results and 94,98% of true positive results. Without any doubt, the listing ability of this methodology is very high.

This study has been developed from the investigation of a real Spanish money laundering case in which this expert team have been collaborating. This study is the first step to use machine learning to detect financial crime in Spanish judicial process cases.

Key Words: Benford's Law, unbalanced dataset, random forest, fraud, anti-money laundering.

JEL Clasification: C14, C44, C53, M42

1. INTRODUCTION

Everyday, TV news displays economic crimes: tax evasion, money laundering, corruption, misappropriation of public funds, etc. All of them are known as *White Collar* crimes. In this kind of crimes, intelligence is more important than physical strength, and the tools used to detect and stop them are more sophisticated than what's usually used for other crimes. In 1972, American economist Hal Varian suggested to used Benford's law as a diagnostic tool in projective model results, especially to predict irregularities that require deeper inspections.
Noncompliance to Benford's law is just an evidence that shows that values could have been manipulated, not a crime in itself. Benford's law is not a Universal Law as Newton's law of universal gravitation and there are several instances when data doesn't comply with it. Despite of this, in the economic world Benford's law is present in many datasets and its abscence would be an evidence of irregularities in the accounting or transaction of certain companies. Benford's law can be a clue to discover an economic crime. If the data is manipulated, something can be hidden after this manoeuvre and it would be useful to investigate the reason of this behaviour.

In this study, we use Benford's law as a detection tool for anomalies in books and accounting records of companies investigated for financial crimes, applying four different models of pattern recognition (logistic regression, neural network, decision trees and random forests). In accordance with previous data provided by the police, these four models are used to classify the commercial activity made by the company, to obtain its suppliers and whoever was part of the financial engineering needed to perform the crime.

The follow study has been developed from the investigation of a real Spanish money laundering case in which this expert team have been collaborating. As far as we know, this study is the first step to use machine learning to detect financial crime in a Spanish judicial process case.

Thanks to the use of the methodology shown in this study, we were able to identify those companies showing a larger probability of completing fraudulent operations, focusing this way the limited police investigation resources to these companies.

The rest of the article is organized as follows: Section 2 reviews the use of Benford's law in the literature. Section 3 describes the methodology: (Benford's law, the tests used, the machine learning procedures employed and some issues about data balancing). All the data and transformations made will be explained in Section 4. Section 5 shows the results reached after applying the four abovementiones procedures using three different approaches. Sections 6 concludes.

2. LITERATURE REVIEW

The Benford´s Law has been applied to different fields of knowledge. In mathematics, Luque and Lacasa (2009) have revealed a statistical behaviour in the sequence of prime numbers and the Riemann Zeta Function. In computer engineering, Torres *et al.* (2007) have verified the fact that the size of the files stored in a personal computer follow the Law of Benford, and also that a better knowledge *a priori* on the data stored in a computer can facilitate the calculation and improve its speed; developing a more effective data storage as a tool for detecting viruses or errors. The law has been also employed in the study of the length of the rivers, Rauch et al. (2011), or employed in the detection of scientific fraud (Diekmann, 2007).

Specifically in economical field, Professor Mark Nigrini (1992), of the Cox School of Business, stated that it could be used to detect fraud in income tax returns and other accounting documents. A current development in the field of accounting has revealed the application of Benford's Law to detect fraud in the "manufacture" of data in financial documents.

Moreover, Quick and Wolz (2003) worked on data for the incomes and balances of several German companies for the years 1994-1998. Their results reveal that the first and second digit in most cases (both in a year by year analysis and in the whole period analysis) conforms the Benford's Law.

On the other hand, Günnel and Tödter (2009) suggest that controls for data manipulation should focus on the first digit. They consider that Benford's Law is a simple, objective and effective way to detect anomalies in large samples requiring a more detailed inspection tool. However, Ramos (2006) states that the best part of the analysis is the first three digits in which an electrocardiogram is actually obtained from the file and you can see in detail what happens at each point and what the possible fraud operations are.

In a more recent study, Alali and Romero (2013), which analyzed the financial information of more than ten years of accounting data from different companies, concluded that there is a significant error in the adjustment of the Benford´s Law on the current assets. That is, in capital goods, properties, accounts receivable, ..., which means that during the studied period there was an overestimation of the assets.

As evidenced, the use of Benford's Law in the field of accounting is large and thus has demonstrated its ability to detect anomalies in accounting data.

According to this premise this study proposes different measures of adjustment of the sample to the Benford´s Law, this measures are the indicators for the detection of patterns that conceal fraudulent operations, so this study is able to direct the police authorities towards companies that have greater fraudulent probability.

3. METHODOLOGY

3.1. Benford's Law

Empirically, Benford (1938) found that many data sets and mathematical sequences do not have a uniform distribution of the first digit, as one might expect, and yet it has a biased probability function as follows:

$$f(x_1) = \Pr(X = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) \quad d_1 = 1,2,\ldots,9$$

the distribution function,

$$F(x_1) = \Pr(X \leq d_1) = \log_{10}(1 + d_1) \quad d_1 = 1,2,\ldots,9$$

Starting from the first digit distribution, we can derive the second digit distribution, as follows:

$$f(x_2) = \Pr(X = d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10 \cdot k + d_2}\right) \quad d_2 = 0,1,2,\ldots,9$$

The most important properties of Benford's Law are both invariance in scale and invariance on the basis. If digits of X were randomly distributed, we would expect a uniform for the first digit value d_1 = 1, 2 distribution, ..., 9. However, a number of variables show a different distribution for the first digit and according with the demonstrations of Pinkham (1961) and Hill (1995) would comply with the two indicated properties:

- Invariance in scale. It has been empirically observed that when making changes in scale in those variables that conform to the logarithmic law the new transformed variable also fit well to this law. If the units of measurement are changed Benford's Law is still fulfilled, that is, it does not depend on the measurement system. In economic terms, the currency in which the variable object of study is measured is independent for the obtained results.
- Invariance on base. The logarithmic law is independent of the logarithmic base that is used, and is equally valid on base 10, on binary basis, or on any other basis. Hill (1995) showed that the logarithmic distribution is the only continuous distribution that is invariant on base and that invariance in scale implies invariance on base, but not vice versa.

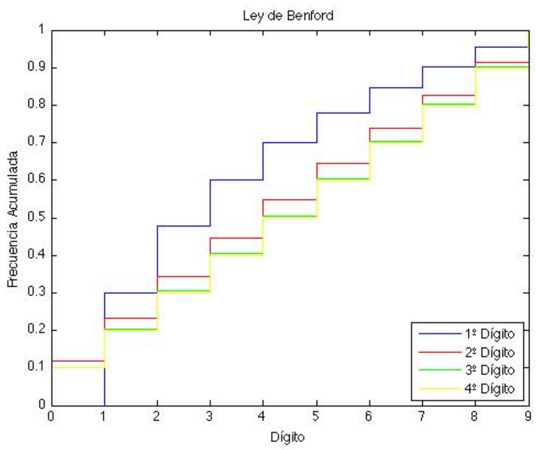
Table 1: Digits Probabilities.

Value	1º Digit	2º Digit	3º Digit	4º Digit
0	--	11.9679%	10.1784%	10.0176%
1	30.1030%	11.3890%	10.1376%	10.0137%
2	17.6091%	10.8821%	10.0972%	10.0098%
3	12.4939%	10.4330%	10.0573%	10.0059%
4	9.6910%	10.0308%	10.0178%	10.0019%
5	7.9181%	9.6677%	9.9788%	9.9980%
6	6.6947%	9.3375%	9.9401%	9.9941%
7	5.7992%	9.0352%	9.9019%	9.9902%
8	5.1153%	8.7570%	9.8641%	9.9863%
9	4.5757%	8.4997%	9.8267%	9.9824%
Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Source: Own elaboration.

Benford's Law is more robust than you can imagine. Not all numerical series follow a Benford distribution, however, if several distributions are selected randomly, and the samples taken from each of these distributions are random, then the frequency of the digits of the combined data set will converge to Benford´s Law Distribution. Figure 1 shows the cumulative distribution functions for the first, second, third and fourth digits.

Figure 1. Distribution function for first, second, third and fourth digit.



Source: Own elaboration.

As the occupied position is advanced, the probability tends to be uniform, and the probability of finding each of the different digits is 10%. Figure 1 shows that the representation of the distribution function converges to a uniform distribution for the 10 digits.

3.2. Test of hypothesis

The aim of this paper is to determine the fit of the sample under analysis[1] to Benford´s Law, in order to analyze patterns that could hide fraud operations.

Starting from the Chi-square test, an empirical test will be performed based on the simulation from which the adjustment measures of each company will be obtained.

Cho and Gaines (2007) and Giles (2007) state that the chi-square test is too rigid to evaluate goodness-of-fit, since the proportions of Benford's law do not represent a true distribution but their expected values in the limit. They propose using statistical or other methodologies that are less sensitive to sample size than Chi-square test.

The Kolgomorov-Smirnov and Kuiper tests have the same problem as the Chi-square tests, when they are applied to large databases, so we propose the next Z statistic to measure the adjustment of the first and second digits of each company.

The Z statistic is used to measure the conformity of a set of data to the Benford´s Law, the formula is as follows:

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

Where:

n_{oi}: Value observed in the sample.

n_{Ti}: Expected Value derived from Benford's Law

the term (1 / 2N) is a continuity correction term, it is only used when it is less than the first numerator term.

With the statistic Z we evaluate the proportion of the digits separately by determining which digits differ from the Benford distribution. This implies that for the first digit there are nine comparisons, and one can´t take the significance level of 5% to compare the p-values. The process of reducing the level of significance is based on Bonferroni's inequality (Hogg *et al.* (2005)). Each p-value is compared with $\alpha/9 = 0{,}05/9 = 0{,}0056$; obtaining an approximate probability of rejection of 0.05. If $P(|Z| > 2.77) = 0{,}0056$ any Z statistic greater than 2,77 absolute value implies the rejection of the null hypothesis.

As happens with the chi-square tests and other p-values based tests, the Z test, rejects the null hypothesis when we analyze the adjustment to the law of the whole set, this is due to the large amount of data analyzed (285,774 commercial operations of 643 supplier

[1] The total amount of commercial operations between the suppliers and the core Company.

companies), so we propose an empirical test based on the simulation that is not sensitive to the sample size, the OverBenford Test.

The OverBenford Test is based on the generation of 100 Benford´s distribution simulations of the same size as each of the companies, and then to perform a Chi-square contrast of each one, an alternative measure is obtained for each company that it is not so influenced by sample size.

The proposed contrast is as follows:

$$T = \sum_{i=1}^{m} \frac{(n_{oi} - n_{Ti})^2}{n_{Ti}}$$

Where:

n_{oi} : It is the value observed in the sample.

n_{Ti} : It is the synthetic value generated from a Benford´s distribution with the same size as the simple.

3.3. Machine Learning Methodology

This work uses the Benford´s Law as the basis for classifying businesses: as legal or fraudulent. In a first step the values corresponding to the statistical OverBenford and Z_i are calculated. These p-values will serve to determine the behaviour of each company in their daily operations, so they are used to perform the classification. If the number of operations is high, the use of the various digits can be observed, as well as if these companies follow the expected values and in what proportion. As indicated at the outset, a particular behaviour against the Benford´s Law may be an indicator of illegal operations, but it can´t be a crime himself or an evidence against a company.

The variables used to make the different classifications will be the p-values indicated, the frequency of operations and the dependent variable which has been generated by an expert, based on the knowledge of the operations carried out between the different companies.

The machine learning methodologies used are:

- Logistic regression.
- Neural network.
- Decision tree.
- Random forest.

The following part summarizes the different methodologies and their justification.

3.3.1. Logistic Regression

Logistic regression models are widely used to know the relationship between qualitative variable, dichotomous dependent variable (binary or binomial logistic regression) and one or more independent explanatory variables, or covariates, which can be qualitative or quantitative. Being the exponential type the initial equation, although it´s logarithmic transformation (logit) allows its use as a linear function.

Because the characteristics of the logistic regression models, two types of analysis can be performed:

- Quantify the importance of the relationship between each of the covariates and the dependent variable.
- Classify individuals within the two categories of the dependent variable, depending on the probability of belonging to one of them.

The second type of analysis is the one of interest for the study to be carried out here. Logistic regression is a widely used statistical tool to estimate an individual probability, as Salas (1996) used to determine the demand for university studies in Spain.

However, when the number of covariates is relatively high or when covariates have a high correlation the estimated parameters may be unstable. For this reason, we select the variables that will be used in the training of the model.

3.3.2. Neural Networks

The human brain inspires neural networks, and they try to reproduce the essential aspects of a real neuron. Neural networks (NN) are a set of simpler elements that are interconnected in form hierarchical and interacting like neural systems.

To be able to use them as represent systems of greater complexity NN can have feedback. A differential feature is that they can learn from experience through the generalization of cases.

Artificial neural networks constitute a technique of mass processing of information that emulates the essential characteristics of the neuronal structure in the biological brain.

As Sosa Sierra (2011) states, "a neural network is characterized by four basic elements: its topology, the learning mechanism, the type of association between input and output information and how this information is represented."

The neurons are distributed in the network forming layers of a certain number of basic elements. That is, there is an input layer that directly receives information from the external sources of the network, hidden layers that are internal to the network and do not have direct contact with the outside (from zero levels to a high number), being able to be interconnected in different ways, which determines the different topologies and an output layer that transfers information from the network to the outside.

Therefore, the fundamental parameters of the network will be: the number of layers, the number of neurons per layer, the degree of connectivity and the type of connections between neurons.

There are essentially two types of networks based on the learning paradigm: supervised and unsupervised. In supervised learning, the network is given the correct answer for each of the training instances, this type will be the one used in this work.

In this way the weights can be adjusted in order to approximate the response of the network to the response provided by the sample data.

In non-supervised learning, patterns and correlations are explored in the input data of the network, to be able to classify them.

3.3.3. Decision Tree

Decision Trees is another classic technique that is widely used in machine learning. Decision tree classifies the instances according to an objective based on the available variables, which can be qualitative or quantitative, and can be interpreted as a series of consecutive conditions.

Decision Tree algorithms split 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 is an over-adjustment or over-training. The tree will have a high accuracy in the classification of the training data, but very little precision to classify instances of the test data.

This problem must be solved with a posteriori pruning procedure of the tree construction. The idea is to measure the estimated error of each node, if the estimated error for a node is less than the estimated error for its sub-nodes then the sub-nodes are removed. The algorithm used is the C4.5, developed by Quinlam (1993), which allows pruning.

3.3.4. Random Forests

The Random Forest methodology, developed by Breiman (2001), is a variant of the bagging methodology that uses decision trees as classifiers. A random forest is a classifier consisting of a collection of tree classifiers that are generated by a randomly distributed vector identically and independently, and where each tree casts a vote for the most popular input class.

Each tree is constructed using a different bootstrap sample (random sampling with replacement) from the original training data set. To classify a new object, it is given the vector that describes it to each tree, which make its classification independent. The trees are

built without any pruning, letting it reach the highest possible height.
Instances are sorted with the class that gets the highest number of votes from the assembly trees. The results of random forests are difficult to interpret since they are the result of the aggregation of many decision trees.
Breiman (2001) states that the error of random forests depends on two fundamental factors: the correlation between the trees of the ensemble and the effectiveness of each individual tree.
Bagging method increases the stability of the decision tree that increases the robustness of the presence of redundant variables, making it very suitable in data sets with many variables.

3.4. Unbalanced Data

Unbalanced data sets are quite common in the scientific literature, and data sets with a low percentage of positive instances are the most relevant. Kotsiantis *et al.* (2006) indicate that these data sets are quite common in different fields. The methods used for the treatment of unbalanced data are:

- Balance the training set by:
    - Sub-sampling of the majority class.
    - Over-sampling of the minority class.
    - Generation of synthetic data in the minority class.
- Modify the algorithm by:
    - Adjusting class weight (Cost-Sensitive Learning).
    - Precision threshold adjustment.
    - Modify it to make it more sensitive to the minority class.

Subsampling can be used with large data sets and applied to the majority class by reducing the number of instances of this class. Since this method discards most of the instances of the majority class, information that could be relevant in the training set is lost.
Over-sampling, as opposed to sub-sampling, works with the minority class, which increases to balance it with the majority class. In this case, no information is lost, but the training set is increased by copying and pasting minority class observations, which could lead to other problems.
These two techniques are easy to apply but as indicated both have their own problems, so you have to use a more sophisticated approach, using any of the following methodologies.

3.4.1. Cost Sensitive Learning

This technique does not create balanced data distributions, but rather seeks to balance learning by applying a cost matrix describing the cost of misclassification versus the correct one.

This technique uses the cost associated with misclassification of observations by applying specific class weights as a function of loss (smaller weights for instances of the majority class and larger weights for those of the minority class). The weights can be set to be inversely proportional to the fraction of corresponding class instances.

In financial fraud example, there will be no cost associated with identifying a person who has committed fraud as positive and who has not made fraud as negative. However, the cost associated with identifying a person who has committed fraud as negative (false negative) is much more dangerous than identifying a person who did not cause fraud as positive (false positives).

The cost matrix is similar to the confusion matrix. The objective is to penalize the errors (false positives and false negatives) against the correct ones (real negatives and true positives).

3.4.2. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a technique that provides new information regarding the minority class as well as underrepresentation of the majority class (Chawla et al., 2002).

With this technique it is possible to balance the data set generating artificial data, so it would be a form of oversampling but with better conditions. This technique generates a random set of minority class observations to change the learning bias of the classifier towards the minority class.

This work uses bootstrapping and KNN (algorithm of the nearest K-neighbours) to generate the random set. Basically it takes the difference between the function in question and its nearest neighbour, then this difference is multiplied by a random number between 0 and 1, and adds it to the feature that helps in the selection.

4. Sample Description

The quality of the data will be the basis for a correct analysis and subsequent classification, so the majority part of the work has been devoted to cleaning and data transformation.

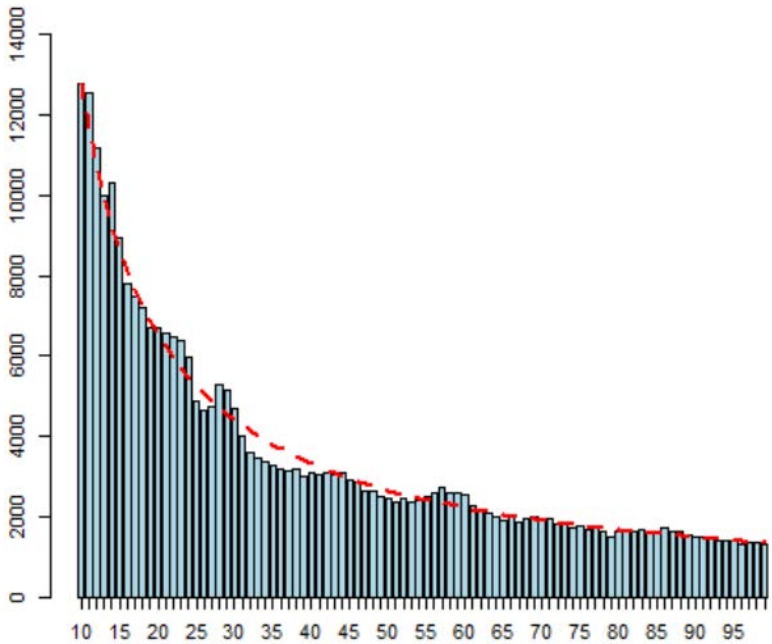
4.1. Sample and values generation

Given that the case analyzed is one of the most voluminous money-laundering cases in Spanish history, there is a large database containing 285,774 commercial transactions carried out by 643 suppliers with the company investigated. This study is based on the analysis of the amounts of each commercial operations.

It is only certain that 26 of the total number of companies are been identified previously as fraudulent, that is, the operations they carry out do not conform to legality. Therefore, we only have information a priori that 4% of companies are fraudulent, but it is unknown if the rest companies are or are not. The a priori information is provided by the police authorities based on a judicial process investigation.

If we compare the whole sample distribution with Benford's Law we can visually verify how perfect it fits.

Figure 2: 1º and 2º digits Benford´s Law Overall analysis.



Source: Own elaboration.

However, when the different tests are performed it is observed that the null hypothesis is rejected. The p-values obtained from the traditional tests would not serve as a measure for quantifying the fit in very large samples.

Table 2: Global Analysis of Benford's Adjustment to the First Digit.

Mean	Var	Ex.Kurtosis	Skewness
0.496	0.085	-1.224	0.026

Source: Own elaboration.

Table 3: First Digit Benford's Law Tests.

	X-Squared Test	Z-Test	OverBenford Test
P-Value	2.2e-16	2.3e-16	0.2303

Source: Own elaboration.

The OverBenford Test, which is the proposed test in this study, has the capacity to measure the adjustment to the Benford Law avoiding the sensitivity to sample size, a problem that the other tests analyzed have.

Finally we decide to include only 335 companies to carry out the analysis (just those with a minimum of operations), of which the experts have identified as fraudulent 23 of them, so only 6.87% of the instances belong to the minority class. Having clearly an unbalanced data set. It is opportune to apply balancing strategies to the sample.

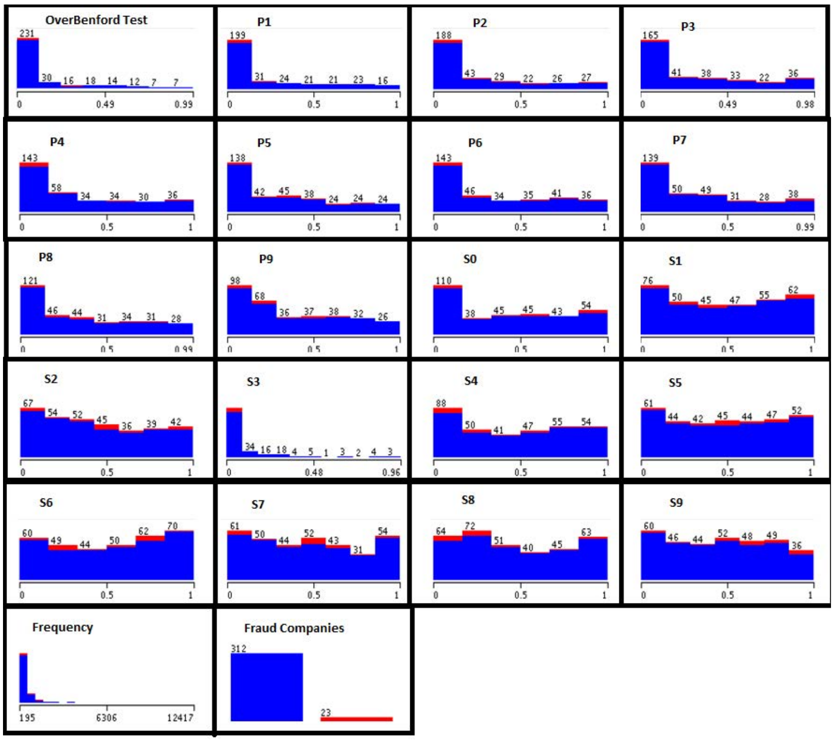
The minimum number of operations of each company (variable "frequency") for a company to be incorporated in the analysis is 195, which makes available 245,227 operations of 335 companies. The average number of operations of a fraudulent company is 2,042 operations, while the mean of the rest companies is 635.45 operations.

Given the predisposition of fraud companies (companies that commit financial fraud and money laundering) to generate the maximum number of possible operations with the aim of hiding the fraud strategy among them, this work decided to include frequency variable in the analysis, this is one of the most correlated with the variable fraud.

Therefore, we have for each of the companies: the measure of empirical test, OverBenford Test, the "frequency" or number of operations recorded and the Z-statistic p-values results of the first and second digits.

Thus, the objective is to analyze whether there is any pattern of behavior that differentiates companies that commit fraud which do not.

Figure 3: Variables´s Distribution.



Source**:** Own elaboration.

We have 21 independent variables: the OverBenford Test measure, 9 (P1 to P9) plus 10 (S0 to S9) Z statistic p-values of each first digit and the company's operations frequency.

4.2. Selection of Variables

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 (Seo and Choi, 2016). 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 2 gives the predictors ordered by degree of relation to the response variable. From the total of predictors initially considered, we have included in the models those that exceeded the value 0.05. In total, there are 11 predictors.

5. **Table 2**. 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	

6. Analysis of Results

For the evaluation of the models, three options have been proposed:

- Modelling data without applying any type of transformation.
- Use cost-sensitive learning (cost matrix).
- Apply SMOTE algorithm transformation to balance the dependent variable.

The results are in the following parts.

6.1. Data without Transformation

We perform the classification without any transformation performing on the data, and the following results are obtained through cross-validation:

Table 5: Non-Data Transformation´s Results.

	LG		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%	

Note: LG: Logistic Regression, DT: Decision Tree, NN: Neural Network, RF: Random Forest. Source**:** Own elaboration.

The predictive capacity of the models is very high, but they present very low positive real rates, between 13.04% of the logistic regression and 34.78% of the neural network. By having such unbalanced data, the algorithms tend to favour classification in the dominant category, identifying very few fraudulent companies.

6.2. Cost Matrix Application

It is assumed that the costs of incorrect classification are different. False positives would only have the cost of carrying out the corresponding investigation until determining their misclassification, however, false negatives would entail a much higher cost (taxes defrauded, etc.). The cost matrix will allow the target variable to be balanced without data transformation.

The results obtained are as follows:

Table 6: Cost Matrix Application´s Results.

	LG		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%	

Note: LG: Logistic Regression, DT: Decision Tree, NN: Neural Network, RF: Random Forest. Source**:** Own elaboration.

Comparing the results with those of the previous section, we detect an important precision model decreasing. The random forest is the only one that maintains 94% correctly classified instances, reducing in the rest cases to 73.73%. However, the true positive rate has been substantially improved. This improvement has been due to the fact that the algorithm identifies more companies as possible defrauders by the inclusion of the cost matrix.

This methodology has increased the detection of true positives in exchange for raising considerably the false positives. In the only case this does not happen is in the random forest, random forest is which identifies less true positives and less positives.

6.3. Application of SMOTE

Based on the above data, we had 312 legal companies and 23 non-legal companies, once applying SMOTE the new data set has up to 51.06% "no-fraudulent companies" and a 49.94% "rest companies". On this new training set it is created the new models with the techniques already exposed. The results obtained are shown below.

Table 7: Application of SMOTE Results

	LG		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%	

Note: LG: Logistic Regression, DT: Decision Tree, NN: Neural Network, RF: Random Forest. Source**:** Own elaboration.

This method does not substantially improve the predictive capacity compared to the use of the cost matrix. However, the true positive rate has improved in all cases. With the original data, real positive rates were between 13.04% of the logistic regression and 34.78% of the neural network. The cost matrix was improved to 26.09% of the random forest and 56.52% of the logistic regression (the one that improved the most). And finally, with the transformation of the data (SMOTE), true positive rates range from 82.27% of the logistic regression to 94.98% of the random forest.

Applying SMOTE, the capacity to identify illegal companies is much superior to the two previous ones, obtaining in the case of the random forest a very satisfactory result. Of out 611 instances, it only incorrectly classified 27 (4.42%), which 15 are false negatives and 12 false positives.

Finally, for the evaluation of the models we have taken the measurements of the ROC Area, the Kappa Statistic and RMSE (Root Mean Squared Error).

Table 8: Results Comparison

	No-Transformation			Cost-Matrix			SMOTE		
	ROC	Kappa	RMSE	ROC	Kappa	RMSE	ROC	Kappa	RMSE
LG	0.747	0.2061	0.236	0.711	0.3243	0.4227	0.844	0.5675	0.4012
DT	0.635	0.2664	0.2702	0.615	0.2086	0.332	0.894	0.7348	0.3499
NN	0.765	0.34	0.2578	0.63	0.2104	0.3306	0.926	0.7252	0.3392
RF	0.74	0.2817	0.2268	0.773	0.3499	0.2415	0.989	0.9116	0.2088

Note: LG: Logistic Regression, DT: Decision Tree, NN: Neural Network, RF: Random Forest. Source: Own elaboration.

From Table 8, it is deduced that the best results are obtained with the SMOTE algorithm against the untransformed data and the application of the cost matrix. As for the classification technique, the best is the random forest with an ROC area of 0.989 and a Kappa statistic of 0.9116, in both cases very close to 1.

7. CONCLUSIONS

In this kind of issues that depend on an expert judgement to determinate target variable label, it's not possible to be sure about the algorithm classification. Investigated companies would have been sorted as legal or illegal, but there are a significant number of un-investigated companies included inside the legal companies group. It is certain that fraudulent companies are fraudulent; however not all legal companies are legal. This is one of the principal purposes of this study: to be able to detect fraudulent companies based on similarities with other companies already investigated.

The obtained results show that the SMOTE algorithm gets better true positive results over the cost matrix implementation. However, the general accuracy of the model is very similar, so the amount of a false positive result will increase with SMOTE methodology. Cost matrix identifies fewer positives and, as a result, there would be less number of investigated companies.

Selecting a high number of investigated companies would have two negative points. First one is the increase of investigation costs. More investigated companies means more time, more staff and more resources. The second one is about annoyances caused to legal companies, because of the investigation process itselt and its consequences.

Random forest gets better results with SMOTE transformation. It obtains 96.15% of true negative results and 94,98% of true positive results. Without any doubt, the listing ability of this methodology is very high.

The best solution would be to use different methods and algorithms to evaluate different approaches. Data without transformation sorts true negatives very well and this could be used if we apply methods of assembling models for the final classification.

8. REFERENCES

- **Alali, F. A., & Romero, S. (2013).** Benford's Law: Analyzing a decade of financial data. Journal of Emerging Technologies in Accounting, 10(1), pp. 1-39. http://dx.doi.org/10.2308/jeta-50749
- **Benford, F. (1938).** The law of anomalous numbers. Proceedings of the American Journal of Social Science Studies ISSN 2329-9150 2017, Vol. 4, nº. 1 138 http://jsss.macrothink.org Philosophical Society, pp. 551-572.
- **Breiman, L. (2001).** Random Forests. Machine Learning. 45 (1), pp. 5-32.
- **Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002)**. Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research, 16, pp. 321–357.
- **Diekmann, A. (2007).** Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. Journal of Applied Statistics, Vol. 34(3), pp. 321-329. http://dx.doi.org/10.1080/02664760601004940
- **Giles, D. E. (2007).** Benford's law and naturally occurring prices in certain eBay auctions. Applied Economics Letters, Vol. 14(3), pp. 157-161. http://dx.doi.org/10.1080/13504850500425667
- **Günnel, S. y K. – H. Tödter (2009).** "*Does Benford's Law hold in economic research and forecasting?*" Empirica, Vol. 36, nº. 3, pp. 273-292.
- **Hill, T. (1995)**. "The Significant-Digit Phenomenon*".* The American Mathematical Monthly, Vol. 102, nº. 4, pp. 322-327

- **Hogg, R. V., McKean, J.W. y Craig, A. T. (2005).** Introduction to Mathematical Statistics, 6th ed., Pearson Prentice Hall, Upper Saddle River, New Jersey.
- **Kotsiantis, S., Kanellopoulos, D., y Pintelas, P. (2006).** Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, Vol. 30(1), pp. 25-36.
- **Luque, B. y L. Lacasa (2009)**. "*The first digit frequencies of primes numbers and Riemann zeta zeros*". Proceedings of the Royal Society A, Vol. 465, pp. 2197-2216.
- **Nigrini, M. J. (1992)**. The detection of income escape through an analysis of digital distributions. PhD Tesis University of Cincinatti.
- **Pinkham, R. S. (1961).** "On the distribution of first significant digits". The Annals of Mathematical Statistics, Vol. 32, nº. 4, pp. 1223–1230.
- **Quick R. y M. Wolz (2003)**. "*Benford's law in deutschen Rechnungslegungsdaten*". Betriebswirtschaftliche Forschung und Praxis, Vol. 55, pp. 208-224
- **Quinlan, R. (1993).** Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.
- **Ramos, D. (2006)**. "*Fraude: un nuevo enfoque para combatirlo*". Auditoria Pública, Vol. 38, pp. 99-104.
- **Rauch, B., Göttsche, M., Brähler, G., & Engel, S. (2011).** Fact and Fiction in EUGovernmental Economic Data. German Economic Review, Vol. 12(3), pp. 243-255. http://dx.doi.org/10.1111/j.1468-0475.2011.00542.x
- **Salas Velasco, M. (1996)** La regresión logística. Una aplicación a la demanda de estudios universitarios. Estadística Española, Vol. 38, nº. 141, pp. 193-217.
- **Sosa Sierra, M. D. C. (2011).** Inteligencia artificial en la gestión financiera empresarial. Revista Científica Pensamiento y Gestión, Vol. 23, pp. 153-186.
- **Tam Cho, W. K. y B. J. Gaines (2007).** "*Braking the (Benford) law: statistical fraud detection in campaign finance*". American Statician, Vol. 61, nº. 3, pp. 218–223
- **Torres, J.; S. Fernández, A. Gamero y A. Sola (2007)**. "*How do numbers begin? (The first digit law)*". European Journal of Physics, Vol. 28, pp. 17-25.
- **Varian, H. (1972).** "*Benford's law*". American Statistician, Vol. 23, pp. 65-66.