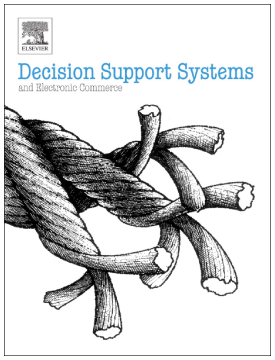
Accepted Manuscript



An investigation of bankruptcy prediction in imbalanced datasets

David Veganzones, Eric Séverin

PII: S0167-9236(18)30108-8
DOI: doi:10.1016/j.dss.2018.06.011
Reference: DECSUP 12969

To appear in: *Decision Support Systems*

Received date: 19 December 2017
Revised date: 30 May 2018
Accepted date: 29 June 2018

Please cite this article as: David Veganzones, Eric Séverin , An investigation of bankruptcy prediction in imbalanced datasets. Decsup (2018), doi:10.1016/j.dss.2018.06.011

An investigation of bankruptcy prediction in imbalanced datasets

David Veganzones[*a,b] • **Eric Séverin**[a]

a) Université de Lille, IAE Lille, 104 Avenue de Peuple Belge, 59000, Lille, France, Laboratoire Rime Lab. EA7396

b) Université de Paris Nanterre, IUT de Ville d'Avray, 200 Avenue de la République, 92001, Nanterre, France

Abstract

Previous studies of bankruptcy prediction in imbalanced datasets analyze either the loss of prediction due to data imbalance issues or treatment methods for dealing with this issue. The current article presents a combined investigation of the degree of imbalance, loss of performance, and treatment methods. It determines which imbalanced class distributions jeopardize the performance of bankruptcy prediction methods and identifies the recovery capacities of treatment methods. The results show that an imbalanced distribution, in which the minority class represents 20%, significantly disturbs prediction performance. Furthermore, the support vector machine method is less sensitive than other prediction methods to imbalanced distributions, and sampling methods can recover a satisfactory portion of performance losses. Accordingly, this study provides a better understanding of the data imbalance issue in the field of corporate failure and serves as a methodological guide for designing bankruptcy prediction methods in imbalanced datasets.

Keywords: bankruptcy prediction, imbalanced dataset, finance

JEL: C53, G33

[*] Authors' email addresses: david.veganzones@gmail.com; e.severin@wanadoo.fr
Corresponding Author: Veganzones: °+34 633 246 629; Séverin: +33 06 77 50 72 86

An investigation of bankruptcy prediction in imbalanced datasets

Abstract

Previous studies of bankruptcy prediction in imbalanced datasets analyze either the loss of prediction due to data imbalance issues or treatment methods for dealing with this issue. The current article presents a combined investigation of the degree of imbalance, loss of performance, and treatment methods. It determines which imbalanced class distributions jeopardize the performance of bankruptcy prediction methods and identifies the recovery capacities of treatment methods. The results show that an imbalanced distribution, in which the minority class represents 20%, significantly disturbs prediction performance. Furthermore, the support vector machine method is less sensitive than other prediction methods to imbalanced distributions, and sampling methods can recover a satisfactory portion of performance losses. Accordingly, this study provides a better understanding of the data imbalance issue in the field of corporate failure and serves as a methodological guide for designing bankruptcy prediction methods in imbalanced datasets.

Keywords: bankruptcy prediction, imbalanced dataset, finance

JEL: C53, G33

1. Introduction

The most recent financial crisis exposed the vulnerability of the financial system; more than ever before, firms of all sizes are suffering financial difficulties that sometimes lead to bankruptcy. Such difficulties affect financial institutions, shareholders, managers, employees, and governments alike, and it is crucial to be able to predict corporate bankruptcy. In turn, this critical corporate issue has become a major research area in the corporate finance field.

Although several corporate bankruptcy prediction models have been proposed, according to various prediction methods or variables (Balcaen and Ooghe, 2006), most have been designed using the classical paradigm of paired samples of available data (Chen et al., 2009; Olson et al., 2012). That is, the datasets contain the same number of bankrupt and non-bankrupt firms. Such a practice ignores real-world conditions, where bankruptcy is rare. Although the number of non-bankrupt firms is high, the proportion of bankrupt firms is very low, on an order ranging from 100:1 to 1,000:1. Therefore, in the real world, researchers face imbalanced datasets, in which bankrupt company observations are clearly outnumbered by non-bankrupt companies.

Therefore, we explore the predictive capacity of bankruptcy models in imbalanced datasets. Data and their characteristics are the most crucial elements of any prediction model (Anderson, 2007), so the imbalanced class distributions in datasets are relevant and demand analysis. The issue of data imbalance has been documented from two perspectives. The first acknowledges that when a bankruptcy prediction model uses a dataset that represents the real-world population - that is, an extremely low frequency of firm of firm bankruptcies- model's predictive performance is diminished, especially for bankrupt firms. The second offers a treatment technique for handling imbalanced datasets and improve the model's classification accuracy. Although these perspectives provide a foundation for understanding this issue, fundamental questions remain. Which imbalanced class distribution disturbs a model's predictive performance? What is the improvement capacity of treatment techniques? Datasets may present multiple imbalanced class levels that contain different proportions of bankrupt firms, because of the irregular bankruptcy rates in the population, the scarcity of bankrupt firms, and a lack of accessibility to these firms' information (Tian et al., 2015). To evaluate whether a bankruptcy prediction model's forecast capacity is jeopardized, it is essential to address the imbalanced proportion that significantly disturbs the performance of the model. Moreover, given that bankruptcy is a critical corporate issue that has social costs, it is important to predict it accurately. We therefore conduct an analysis of the capacity of treatment methods to predict bankruptcy in a scenario marked by imbalanced datasets.

Although the imbalanced datasets issue has received attention by the scientific community (Lane et al., 2012, Piri et al., 2018), to date, the bankruptcy prediction field has lacked insights into the relationships among the degree of imbalance, loss of performance, and the recovery capacity of treatment methods. Accordingly, we account comprehensively for the aspects of imbalanced datasets that significantly affect the performance of bankruptcy prediction models.

First, we explore the causal link between the data imbalance issue and bankruptcy prediction models' loss of performance. We compare the results achieved using balanced-distribution prediction models with those achieved using different degrees of imbalanced distributions. Second, we investigate the capacity of treatment methods to recover from the loss of performance caused by imbalanced datasets. We evaluate four sampling techniques, as treatment methods, that are widely used in literature to handle imbalanced datasets: random oversampling, random undersampling, the synthetic minority oversampling technique (SMOTE), and EasyEnsemble. Third, we explore the levels of sensitivity of prediction methods to imbalanced datasets, as well as the levels of sensitivity of sampling techniques to various training data sizes.

Our findings demonstrate the nature of the data imbalance issue and highlight its implications for the performance of bankruptcy prediction models. Most notably, we find that when a model uses a dataset that includes an imbalanced proportion of 4:1 (data that contains 80% non-bankrupt and 20% bankrupt firms) or higher, the model's ability to predict bankruptcy is jeopardized. However, we also find that support vector machines (SVMs) represent the method that is least influenced by imbalanced proportions, because any significant differences in their performance are apparent (except for the most imbalanced distribution, with data that contain an imbalanced proportion of 90% non-bankrupt firms and 10% bankrupt firms). Furthermore, we find that all sampling techniques achieve a similar recovery, even though the recovery represents only a satisfactory part of the performance loss. Finally, our results show that the SMOTE outperforms other sampling techniques when different imbalanced proportions and data size are taken into account.

Section 2 provides a review of literature review on the data imbalance issue and bankruptcy prediction models. Section 3 describes our research methodology. Section 4 presents and discusses results and Section 5 concludes.

2. Literature review

2.1. Data imbalance issue

Conceptually, a dataset presents a class imbalance if it contains unequal distributions between classes. However, literature generally accepts that a dataset is imbalanced when one class significantly outnumbers another (Kotsiantis et al., 2006).

Although imbalanced datasets appear frequently in the classification field (as bankruptcy predictions), classification models tend to expect equal misclassification costs, because it is the prevailing scenario. The models are designed to

optimize overall accuracy; they do not take into account the relative distribution of each class (Lopez et al., 2013). Therefore, they fail to represent properly the data characteristics of imbalanced datasets, which results in the development of suboptimal classification models that provide unfavorable predictions across data classes (Fernandez et al., 2010). Such degradation of prediction, caused by imbalanced distributions, occurs in the following way: During the learning phase (training set), classification models tend to concentrate on accurate classifications of the majority class, while ignoring the minority class, because classification rules maximize overall prediction accuracy. That is, the decision (classification) boundary of the majority class tends to invade the decision boundary of the minority class (Kim et al., 2015). As a result, in the prediction phase (test set), models often are biased toward the majority class; they accurately classify the majority classes but frequently misclassify the minority classes. Thus, the issue of imbalanced datasets originates in the learning phase, where the classifiers' prediction performance is disturbed, especially with regard to the minority class.

In the case of bankruptcy prediction, an imbalance scenario due to limited instances in the minority class is representative of the domain since bankruptcy firms are rare. Moreover, bankruptcy prediction domain presents two additional peculiarities that make it a rather challenging task. On the one hand, samples are described by financial attributes because they give a view of firms financial situation. However, even though such data presents major advantage to predict bankruptcy and their importance may not be neglected, the fact that they can be manipulated (Rosner, 2003; Charitou et al., 2007; Campa and Camacho, 2015) may lead to a distortion that can be detrimental for bankruptcy models performance. On the other hand, firms that follow similar paths in their deterioration may have different outputs. It is not uncommon that some firms acquire a sort of ability to allow them to survive more easily than others, while apparently their financial situation suggests no differences (D'Aveni, 1989). This fact might result in a problem of class overlapping, in which some data points may appear as (valid) examples for bankrupt and non-bankrupt firms. As it turns out, the dataset complexity in bankruptcy prediction field is a major factor for classification deterioration that may be amplified by the addition of the data imbalance issue. Thus, we proceed to explore jointly the complexity of bankruptcy prediction data in the context of imbalanced datasets. If we extrapolate the consequences of data imbalance issue, in which bankrupt firms represent the minority class, in combination with the complexity of bankruptcy prediction datasets, it has costly ramifications because the misclassification of bankruptcy cases produces loss in capital and "contagion-effects". The bankruptcy misclassification may not only have individual consequences but, it may cause a downward spiral for the whole economy with respect to employment, related firms and economic welfare (Balcaen and Ooghe, 2006). These rationales motivate our specific study of the performance of bankruptcy prediction models in imbalanced datasets.

2.2. Bankruptcy prediction models

Ever since Beaver (1966) and Altman (1968) first studied bankruptcy prediction, the classic paradigm of sample selection for bankruptcy models has been to choose balanced samples with available financial information, in which the proportions of bankrupt and non-bankrupt firms are equal. Balanced samples can be produced using a popular technique known as the paired sample, in which data containing firms that eventually failed are paired -usually according to size, industry, or age criteria- with firms that did not fail (Gordini, 2014; Kim and Han, 2003). This sample selection strategy provides a clear advantage, because it avoids class bias during the learning phase. The classifier maximizes the overall prediction regardless of the class distribution. However, the strategy also has a serious drawback: It does not represent the real-world proportion. Zmijewski (1984) demonstrates that if failed and non-failed proportions do not represent the real-world population, sample-selection bias still may occur, leading to underestimations of failed firms and overestimations of non-failed firms. Moreover, Ooghe and Joos (1990) claim that samples of failing and non-failing firms should be representative of the whole population of firms so that failure prediction models can be used in a predictive context.

Only a few studies explore bankruptcy prediction using real-world samples, that is, with datasets that contain imbalanced class distributions. Wilson and Sharda (1994) provide a primary example: They use three sample proportions on the training set: a balanced sample (composed of 50% failed and 50% non-failed firms) and two imbalanced proportions (20% failed and 80% non-failed firms; 10% failed and 90% non-failed firms) to analyze the prediction performance of discriminant analysis and neural network methods. These authors find that prediction methods achieve better results, especially in failed firms, when the training set presents a balanced sample. In addition, McKee and Greenstein (2000) investigate the capacity of three bankruptcy prediction methods in five highly imbalanced data sets and show that the imbalanced sample distribution in the learning phase causes poor classification performance, especially for bankrupt firms.

Therefore, several studies seek treatment methods to deal with imbalanced data sets. Most focus on sampling techniques (Chawla et al., 2004) that rely on mechanisms to balance sample distributions. Zhou (2013) applies several sampling techniques (two oversampling techniques and two undersampling techniques) to balance two highly imbalanced datasets; all of the techniques allow bankruptcy prediction methods to achieve better results than predictions according to the original imbalanced datasets. Kim and Ahn (2015) use a sampling technique on a training set that contains 620 bankrupt samples and 7,398 non-bankrupt firms to confirm its capacity to improve the performance of bankruptcy prediction methods.

Although these studies document the association between bankruptcy model performance and imbalanced datasets, they merely scratch the surface of the data imbalance issue. They examine the loss of performance caused by imbalanced

datasets but do not establish which degree of imbalance significantly affects the performance of various methods, even though datasets may present differently imbalanced distributions. Nor do they not evaluate the capacity recovery of sampling techniques according to different imbalanced proportions, even though several studies have used such techniques in the domain of bankruptcy prediction. This gap is paradoxical; these aspects are directly related. The capacity of sampling methods can be evaluated, once the loss of performance in imbalanced datasets has been established. Accordingly, we seek to provide a wider understanding of the performance of bankruptcy prediction methods in imbalanced datasets by examining the relationships among the degree of imbalance, loss of performance, and recovery of treatment methods.

The purpose of this study is two-fold: First, it makes possible to study the extent to which prediction methods can be influenced by imbalanced datasets in the context of bankruptcy prediction. This is of significant importance because if specific prediction methods significantly outperform others, then a rigorous study of such methods would provide assistance to assess firms' risk of default. Second, it makes possible to analyze the implication of sampling techniques on this issue. More precisely, it yields to how these techniques would be an efficient help so that they could materialize into a real solution to predict bankruptcy, and become the basis in which bankruptcy prediction models are made. Thus, the added-value of this paper has wide-ranging effects in the advancement of bankruptcy prediction field.

3. Research methodology

3.1. Data

We collected our data from the Altares database, which contains the balance sheets and income statements of French firms, which by law are required to file annual reports in the French commercial courts. We carried out three steps.

First, we selected four different samples of firms. We selected three according to industry sector criteria and particular sectors of activity. Because various types of firms have peculiar financial characteristics, their likelihoods of failure can differ depending on their industry sector. Therefore, we checked whether models shared similar results, regardless of firms' activity sectors, and chose three sectors that, on average, have the highest concentration of failed firms in France: service, construction, and retail. In a fourth sample, we selected firms that belong to any sector of activity, so we could examine the model's capacity to create good prediction rules.

Second, we collected two sets within each sample: a training set that estimates model parameters and a test set that estimates model accuracy. The balance sheets and income statements of collected firms in the training set were published in 2013, whereas for those in the test, annual accounts were published in 2014. Thus, we estimated an out-of-sample and

out-of time error with a one-year gap (Stein, 2007), in which none of the firms in the training set was included in the test set. Moreover, the selected bankrupt firms were those that proceeded to be liquidated or reorganized, and non-bankrupt firms were those that continued their activity over the studied period.

Third, to evaluate the performance of prediction methods in imbalanced datasets, we composed the training and test sets as follows: Because the data imbalance issue originates in the learning phase, we created six training sets of 1,500 firms, with ratios of non-bankrupt to bankrupt firms of 50/50, 60/40, 70/30, 80/20, 90/10, and 95/5 respectively[1].We also created a test set of 1,500 firms, in which the ratio of non-bankrupt to bankrupt firms was 95/5, that is, the same proportion as marked the period studied[2]. For the empirical study, we followed the procedure of Brown and Mues (2012), such that we started by randomly selecting 750 non-bankrupt and 750 bankrupt firms from the database[3]. Then, we randomly selected the number of bankrupt samples from the initial 750 bankrupt firms and randomly included non-bankrupt firms from the original database to create the next imbalanced proportion. By continuing this procedure, we created six imbalanced proportions for each sample. We repeated these steps 100 times to create 100 different training sets for each proportion and 100 different test sets. This procedure ensured the reliability of our results and avoided selection bias. Table 1 presents the configuration of these samples that we followed for each the four samples (services, construction, retail, and all sectors).

Table 1 Data by sample

Samples	Training set proportion						Test set proportion
	50/50	60/40	70/30	80/20	90/10	95/5	95/5
Bankrupt	750	600	450	300	150	75	75
Non-bankrupt	750	900	1050	1200	1350	1425	1425
Total	1500	1500	1500	1500	1500	1500	1500

3.2. Variables

[1] We established fixed total samples (1,500) so that we could test only the effect of different imbalanced proportions. In this way, we avoided the effect of different sample sizes on predictive performance.

[2] According to different sources, the bankruptcy average rate in France for this period ranges from 5%–2%. We tested a set that includes 5% and 2% of bankrupt firms and found no significant differences. Accordingly, we remained conservative and selected the 5% rate.

[3] We randomly selected firms from the database. The database includes about 900 bankrupt and 5,000 healthy firms for each sample; we did not include any of the firms twice.

Using the collected firms' balance sheets and income statements, we calculated 50 financial ratios to use as explanatory variables. We computed the same financial ratios as those used by du Jardin (2015), who provides five different ratios of firms' financial characteristics: liquidity, solvency, profitability, financial structure, activity, and turnover (Table 2). However, including all 50 financial ratios would have led to a very high-dimensional feature space that could have reduced the model's predictive ability. Thus, we performed a two-step variable selection process that allowed us to choose a reduced subset of the most relevant financial ratios.

First, we evaluated correlation values between each variable to measure information redundancy. We analyzed the correlation values within each sample and removed any highly correlated values, which in turn reduced potential model instabilities, such as the need to solve badly conditioned inverse matrices (Mensah, 1984). No extant theory specifies at which value a variable is highly correlated, so we empirically selected variables with correlation values lower than 0.65, whereas Atiya (2001) and Leshno and Spector (1996) selected a 0.7 value in the same context. Our approach is more conservative, to avoid any redundancy. All correlations between excluded variables were significant at the 1% threshold.

Table 2 Initial set of variables

Activity		Liquidity	
Cash Flow/Total Sales	CF/TS	(Cash + Mark. Sec.)/Current Liabilities	(C+MS)/CL
Cash Flow/Value Added	CF/VA	(Cash + Mark. Sec.)/Total Sales	(C+MS)/TS
EBIT/Value Added	EBIT/VA	Cash/Current Assets	C/CA
EBITDA/Total Sales	EBITDA/TS	Cash/Total Assets	C/TA
Gross Trading Profit/Total Sales	GTP/TS	Current Assets/ Current Liabilities	CA/CL
Net Income/Total Sales	NI/TS	Current Assets/ Total Assets	CA/TA
Net Income/Value Added	NI/VA	Current Liabilities/ Total Assets	CL/TA
Value Added/Fixed Assets	VA/FA	Current Liabilities/ Total Sales	CL/TS
Value Added/Total Assets	VA/TA	Inventories/Total Assets	I/TA
Value Added/Total Sales	VA/TS	Quick Assets/Current Liabilities	QA/CL
Profitability		Quick Assets/Total Assets	QA/TA
Cash Flow/Shareholder Funds	CF/SF	Working Capital/Total Assets	WC/TA
Cash Flow/Total Assets	CF/TA	Working Capital/Total Sales	WC/TS
EBIT/Shareholder Funds	EBIT/SF	Solvency	
EBIT/Total Assets	EBIT/TA	Financial Debts/Cash Flow	FD/CF
EBITDA/Permanent Equity	EBITDA/PE	Financial Expenses/EBITDA	FE/EBITDA
EBITDA/Total Assets	EBITDA/TA	Financial Expenses/Net Income	FE/NI
Net Income/Shareholder Funds	NI/SF	Financial Expenses/Total Assets	FE/TA
Net Income/Total Assets	NI/TA	Financial Expenses/Value Added	FE/VA
Profit before Tax/Shareholders Funds	PBT/SF	Turnover	
Financial Structure		Accounts Payable/Total Sales	AC/TS
Long Term Debt/Shareholders Funds	LTD/SF	Current Assets/Total Sales	CA/TS
Long Term Debt/Total Assets	LTD/TA	Inventories/Total Sales	I/TS
Net Op. Work Capital/Total Assets	NOWC/TA	Net Op.Work. Capital/Total Sales	NOWC/TS

Shareholder Funds/Permanent Equity	SF/PE	Receivables/Total Sales	R/TS
Shareholder Funds/Total Assets	SF/TA	Total Sales/Total Assets	TS/TA
Total Debt/Shareholder Funds	TD/SF		
Total Debts/Total Assets	TD/TA		

Notes: EBIT = earnings before interest and taxes. EBITDA = earnings before interest, taxes, depreciation and amortization. Mark. Sec. = marketable securities. Net Op. Work. Capital = net operating working capital. All ratios were calculated using the balance sheet or income statement amounts.

Source: du Jardin (2015).

Second, the variable selection method may influence the performance of prediction models (du Jardin, 2010), so we selected explanatory variables to design prediction models with four different selection techniques, belonging to either the filter or wrapper category. Filter methods use statistical techniques to select the best sets of variables; we used a stepwise search procedure with the Fisher F-test as a stopping criterion and a stepwise search procedure with a χ^2 test as a stopping criterion. In contrast, wrapper methods are based on a heuristic technique that selects variables according to their usefulness for a given classifier. In our case, the other two variable selection methods relied on a backward search procedure. Finally, we retained, for each sample, the variables selected by at least two of the four methods used. Table 3 lists the selected variables, by sector.

Table 3
Variables selected by sample

Service	Construction	Retail	All
C/TA	(C+MS)/CL	(C+MS)/TS	(C+MS)/CL
CL/TA	CA/CL	CA/CL	C/CA
WC/TA	FE/VA	CL/TA	QA/TA
FE/TA	EBITDA/TA	FE/TA	EBIT/TA
EBITDA/TA	LTD/SF	EBITDA/TA	LTD/SF
SF/PE	SF/TA	LTD/SF	TD/TA
TD/TA	EBIT/VA	SF/PE	CF/VA
NI/TS	NI/TS	NI/TS	NI/TS
AC/TS	TS/TA	NOWC/TS	NOWC/TS
NOWC/TS			

See Table 2

3.3. Classification methods

Many methods are available to predict corporate failure, though none significantly outperforms the others (Balcaen and Ooghe, 2006). Accordingly, we selected five classification methods with different characteristics, traditionally used in literature, to analyze the prediction capacities of competing methods. The first two methods, linear discriminant analysis (LDA) and logistic regression (LR), arise from well-known concepts of statistical decision theory; they provide robust results even though they rely on linear functions. The other three methods, neural network (NN), support vector machine

(SVM) and, random forest (RF) focus on learning; they make predictions directly from the data, which makes them reliable. Moreover, by relying on nonlinear approaches, we extend the possibilities for testing complex data.

3.3.1. Linear discriminant analysis

The LDA method is among the first used to predict bankruptcy (Altman, 1968). It assumes that class-conditional densities follow Gaussian distributions and that the distributions have a common covariance matrix (Wald, 1944). When LDA is employed to discriminate between failed and non-failed firms, it needs only to estimate the distributions means and their common covariance; LDA creates a discrimination score (z-score) to distinguish two classes by combining explanatory variables on a linear function. The z-score is computed as follows:

$$z = \sum_{i=1}^{n} (x_i w_i + c) ,$$

Where x_i represents explanatory variables, w_i indicates the discriminant weights, and *c* is a constant.

Although LDA assumes Gaussianity on the class-conditional distributions and equal covariance matrices, and though these assumptions do not hold in corporate failure, it has been widely used because of its robustness (Balcaen and Ooghe, 2006).

3.3.2. Logistic regression

Ohlson (1980) proposed LR to model the posterior probabilities of the classes, using linear functions of the independent variables, while ensuring that they sum to 1 and remain in [0,1] to provide a probabilistic interpretation. Similar to LDA, LR makes use of the log-likelihood ratio to assign a firm to either failed or non-failed classes; the log-ratio takes the form of a linear function. This method allows the use of non-linear maximum likelihood to estimate firms' probabilities of failure using a logistic function, based on dependent variables, in this case, financial ratios. The LR method takes the form of:

$$z = \frac{1}{1 + e^{-(w_0 + w_i x_i)}} ,$$

Where x_i are explanatory variables, w_i are the weights estimated using maximum likelihood estimation, and z is the score for a given firm.

Although both statistical methods, LR and LDA, have similar forms in their discriminant functions, the estimation of their parameters is quite different. The LR method makes fewer assumptions than the LDA method and is generally considered, in statistical literature, to be a safer method.

3.3.3. Neural networks

The NN technique is a mathematical model that emulates the function of a human brain. It is an efficient model for statistical pattern recognition (Bishop, 2006), providing a general framework for representing non-linear functional mapping between sets of input variables and output variables. It is designed by establishing an architecture that connects neurons among layers. In this study, we focus on the multilayer perceptron (MLP), composed of three layers: an input layer composed of *n* neurons for input variables, a hidden layer composed of *m* neurons, and an output layer. Every neuron in the hidden layer is connected to every neuron in the input and output layers. We estimate the connectivity weights -that is, the parameters of the NN representing the relevance of the connections between neurons-by a back-propagation learning method. An NN model computes a z-score that represents the failure probability of a given firm, as follows:

$$z = g\left(\sum_{J=0}^{M} w_{kj}\, g\left(\sum_{i=0}^{d} w_{ji} x_i\right)\right),$$

where g is the activation function, x_i are explanatory variables; w_{ji} corresponds to the weight matrix, including the bias term between the input node (i) and the hidden node (j); and w_{kj} corresponds to the weight matrix with bias connecting the hidden node to the output layer

Since Messier and Hansen (1988) introduced the NN method to the study of corporate failure, may authors have applied it because of its ability to learn complex nonlinear relationships and adapt well to data.

3.3.4. Support vector machines

The machine learning community has widely adapted the SVM, as proposed by Boser et al. (1992), for data classification. An SVM classifier maps training vectors into a higher dimensional space, where it finds a separating hyperplane with a maximal margin. The attractiveness of the SVM method arises largely because there is no need to know the form of the high-dimensional mapping function; it is necessary only to know its inner product, such that any dissimilarity function, even a non-linear function that holds some mild condition can be used. This feature is known as the "kernel trick." (Huang et al., 2004; Tay and Cao, 2001).

The classification capacity of the SVM relies on the ability to transform the input space into a more elaborate feature space in which the separability of the classes is enhanced, in a margin-maximization condition that increases generalization capability by constraining the structure of the model. An SVM is defined as follows:

$$\mathrm{MIN}_{\mathrm{w,b,e}}\ \frac{1}{2}\,\mathrm{w^t w} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{e_i}\,,$$

$$\text{subject to } y_i(w\varphi(x_i) + b) + e_i - 1 \geq 0 \qquad e_i \geq 0,$$

where $\varphi(x_i)$ maps training vectors to a high dimensional space; *w* is the weight vector; *b* is the bias term; *C* is the penalty for the error; and e_i is the slack variable (Vapnik, 1998). When the optimal hyperplane separation between classes is built, a classification decision is given as follows:

$$\mathrm{f(y)} = \mathrm{sign}\left(\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{y_i p_i K(x, x_i)} + \mathrm{b}\right),$$

where *sign* is the sign function; p_i is the parameter; *K* is the function; and in our study, $K(x, x_i) = \exp(-\delta|x_i - x_j|^2)$ is the kernel radial basis function.

3.3.5. Random Forest

The random forest classifier consists of an ensemble of decision trees, in which each classifier is generated using a random vector sampled independently from the input vector (Breiman, 2001). In RF, each tree is built from a bootstrap sample of the data and at each split; a random sample of predictors is examined. In the end, classification is determined by a majority vote for each case over the ensemble of classification trees. When constructing a tree, RF searches for a random subset of the input features (bands) at each splitting node and the tree is allowed to grow fully without pruning. Since only a portion of the input features is used and no pruning is required, random forest is computational fast and simple with a good performance.

3.4. Sampling methods

The original training set X is composed of X_{maj} (non-bankrupt firms) and X_{min} (bankrupt firms), in which the proportion of $\mathrm{X_{maj}}$ clearly outnumbers $\mathrm{X_{min}}$, that is, $\mathrm{X_{maj}} > \mathrm{X_{min}}$.

Sampling methods involve artificially re-sampling the training set, a procedure that is otherwise known as processing the data. Therefore, they amend the imbalanced distribution in the dataset by applying some mechanism that provides a balanced distribution, that is, modifies the original training set to obtain $X_{maj} = X_{min}$. The process of manipulating the distribution of the training samples thus allows a classifier to perform in the standard classification manner, in an effort to improve the methods' performance (Batista et al., 2004).

These methods have become an effective solution, because they provide better results than imbalanced distributions (Estabrooks et al., 2004). Thus, the concept of sampling techniques is to add or remove samples to reach the optimal

balanced distribution for prediction. There are two categories of sampling methods, depending on the data process applied: oversampling and undersampling.

3.4.1. Oversampling approach

The oversampling approach creates a balanced subset from the original dataset by duplicating samples of the minority class. We use two of the most common oversampling techniques: random oversampling and Synthetic Minority Oversampling Technique (SMOTE).

3.4.1.1. Random oversampling

Random oversampling, which is the most common technique and the easiest to implement, implies that the minority samples in the data are replicated randomly until the proportion of majority class is achieved. That is, given an X dataset where X_{min} represents the minority class, X_{maj} represents the majority class. This technique randomly copies one sample of X_{min} and adds it to the X dataset. This process repeats until a balanced proportion is obtained, that is, $X_{min} = X_{maj}$.

3.4.1.2. Synthetic Minority Oversampling Technique

The SMOTE, proposed by Chawla et al., (2002), is a powerful technique that has gained significant popularity because it has performed so well in various areas (Han et al., 2005; Saez et al., 2015). The technique generates artificial samples of a minority class by interpolating between several minority class examples that lie together (Kotsiantis et al., 2006). That is, for each minority sample, it introduces a synthetic sample along the line segment with any of its minority-class nearest neighbors. More precisely, it considers the data in which X_{min} represents the minority class and X_{maj} represents the majority class. First, for each $x \in X_{min}$, it finds the K-nearest neighbors, $X_{N_k} = \{X_k\}_{k=1}^{K}, X_k \in N_k(x) \subset X_{min}$. Second, randomly select $L \leq K$ samples of X_{N_k} (L depends on the amount of oversampling desired). Third, it generates for each of the L selected neighbors a synthetic sample along the line joining the minority sample.

3.4.2. Undersampling approach

The undersampling approach creates a balanced subset from the original data set by removing samples from the majority class. In this study, we used random undersampling and EasyEnsemble.

3.4.2.1. Random undersampling

The undersampling approach generates a balanced subset from the original dataset by removing samples from the majority class. Random undersampling is easy to visualize and understand. In contrast with random oversampling, it removes samples from the majority class to achieve the minority class proportion. Given the original dataset X, it

randomly eliminates samples from the majority class until the minority and majority classes have the same amounts, that is, $X_{min} = X_{maj}$.

3.4.2.2. EasyEnsemble

EasyEnsemble is a straightforward procedure. Given the data, we independently extract several subsets from the majority classes $X_{maj1}, X_{maj2} \ldots X_{majN}$, with the same amount of samples as the minority class. Then, we train a classifier using each subset of X_{maji} (i=1, ,N) and X_{min}. Finally, we combine all classifiers to make the final prediction by using a majority combiner.

3.5. Evaluation metrics

Unfortunately, some of the most common evaluation metrics used to measure classifiers' performance on balanced datasets are inappropriate for imbalanced datasets, as is the case for the most commonly used evaluation metric in bankruptcy prediction, the accuracy rate. It does not take into account sample distribution, which is crucial for imbalanced datasets. Moreover, accuracy rates can lead to erroneous conclusions. Imagine a dataset in which 95% of the observations are in one class and the remaining 5% are in the other class. A trivial prediction method can achieve a prediction accuracy of 95% if it focuses on predicting only the majority class, because the method will tend to choose only the majority, given that the results will be better. This rate suggests that the classifier is accurate, but in reality, it has ignored the prediction of the minority class, which is the main concern in cases of imbalanced datasets. Accordingly, we need to adjust the evaluations of model performance and rely on an evaluation metric that is not sensitive to sample distribution. We selected four evaluation metrics, widely used for imbalanced datasets: sensitivity, specificity, G-mean, and area under the receiver operating characteristic (ROC) curve (AUC) (He and Garcia, 2009; Kotsiantis et al., 2006). We selected sensitivity and specificity metrics, which are intuitive and practical, because each focuses on evaluating a type of firm sensitivity for failed firms and specificity for non-failed firms. G-mean evaluates the effectiveness of a classification in terms of ratio of sensitivity and specificity; it provides a glimpse of a method's overall performance. These metrics are calculated as follows:

$Sensitivity = \frac{TP}{TP+FN}$ is the percentage of bankrupt samples correctly classified.

$Specificity = \frac{TN}{TN+FP}$ is the percentage of non − bankrupt samples correctly classified.

$G - mean = \sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}}$,

where TP = bankrupt firm correctly classified, FN = bankrupt firm misclassified, TN = non-bankrupt firms correctly classified, and FP = non-bankrupt firm misclassified.

Finally, we used the AUC metric, which is adequate for assessing a method's overall performance in imbalanced datasets, because it is insensitive to misclassification costs and imbalanced distributions. Moreover, AUC provides a representation of the trade-off between a true positive (failed firms that have been correctly classified) and a false positive (failed firms that have been incorrectly classified) (He and Garcia, 2009). The AUC can be easily used to compare two classifiers, given that different ROC curves indicate their performances. For a classifier, the ROC curve needs to be as far to the top left corner as possible, where its value will be close to 1. In the example in Figure 1, the classifier with the solid line performs better than that with the dashed line.

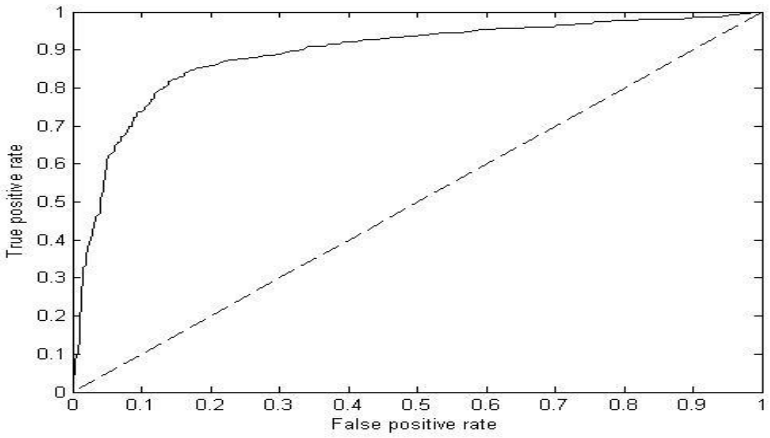


Fig. 1. Example of the ROC curves (X-axis represents 1-specifity and Y-axis represents sensitivity).

3. 6. Models setting

Although discriminant analysis and logistic regression require no specific settings, for the other compared methods a set of hyperparameters has to be tuned. To that effect, the training set is randomly divided in two subsets without repetitions. 80% of the training data is employed to estimate the parameters of the models, while the remaining 20% serves evaluate the performance of a model given some values of the hyperparameters. Here on, we will refer to the former subset as the training subset and to the latter subset as the validation subset. This division is done only for the model selection process, that is, to select the best hyperparameters values. Once the hyperparameters are set, the whole training set is employed to learn the actual model.

The model selection process uses a grid search process with 50 independent training and validation subsets. For each evaluated set of hyperparameters values, the geometric mean of the performance obtained in each of the 50 repetitions is used to identify the best ones. Next, we describe the hyperparameters of the competing methods.

The neural network is designed using a single hidden layer and one output neuron, in which the Levenberg-Marquardt algorithm was used as the optimization technique and the hyperbolic tangent as activation functions. Here, the model selection process is used to select the best performing number of hidden neurons in a predefined range of 5 to 20 neurons.

The radial-basis kernel SVM requires two parameters to optimize, the regularization parameter, C, is set as a value within 10-100 and, the parameter of radial basis function, *p*, is set up to be a value within 1-3.

The random forest requires to set just the number of trees in a predefined range of 100-250 trees.

Moreover, to compute the performance metrics for certain prediction methods, a cut-off value has to be determined so that it can be compared to the score estimated using such methods. Indeed, the assignment of an estimation to a determined class (bankrupt or non-bankrupt) implies that a cut-off has been a priori established, delimiting the separation between classes. Nonetheless, this delimitation varies according to the cut-off calculation strategy. In general, the cut-off is set so that maximizes the overall correct classification as it is the widespread way in the bankruptcy prediction literature. In an imbalanced dataset scenario though, this procedure is inappropriate as it is likely to lead to a high correct classification of the majority class; while producing poor result for the underrepresented class because the focus on correctly predict the majority instances results on a highest overall accuracy. Therefore, we consider an alternative strategy to compute the cut-off value. In line with du Jardin (2015) and Tang and Chi (2005), the cut-off value was determined based on the minimization of the expected cost of misclassification, which takes into consideration fundamental aspects concerning bankruptcy prediction such as misclassification cost and probability of failure, so that an efficient economical decision can be made. The expected cost of misclassification is represented as follows:

$$\text{Expected cost of misclassification} = c_1 \frac{e_1}{n_1} p_1 + c_2 \frac{e_2}{n_2} p_2$$

where c_1 and c_2 are the respective costs of misclassification for bankrupt and non-bankrupt firms; e_1 and e_2 are the type-I and type-II error respectively; n_1 represents the number of bankrupt firms, while n_2 is that of non-bankrupt firms; and p_1 and p_2 are the prior probabilities of bankrupt and non-bankrupt firms respectively.

In this case, the misclassification cost was kept to 1 for both bankrupt and non-bankrupt firms[4]. Besides, the prior probabilities were established according with the current bankruptcy rates when they were collected[5].

[4] Misclassified a failed firm (predict that a firm is healthy when it will fail) will cost a loss in capital, while misclassified a healthy firm (predict that a firm is failed when it is healthy) will only involve loss of commercial bargain. Nonetheless, financial institutions assess each firm in risk of default with a firm in condition to survive. It is for this reason that the cost was kept to 1 for both cases.

[5] Source: http://www.insee.fr.

4. Results

4.1. Results on different degrees of imbalance

This experimental study explores the effect of various degrees of imbalanced training sets on bankruptcy prediction models. In this regard, we first analyzed the sensitivity and specificity evaluation metrics that measure the impact of imbalanced datasets on the prediction capacity of bankrupt (sensitivity) and non-bankrupt firms (specificity). We determined these metrics by a fixed default threshold value that defines the boundary value to classify the sample into bankrupt and non-bankrupt firms. Table 4 indicates the results obtained by the bankruptcy models built on the different samples, which allow us to compare performance associated with imbalanced training sets, relative to that achieved with a balanced proportion.

Table 4
Sensitivity and specificity rates achieved with prediction models on balanced and imbalanced training sets by sample (Test set contains a 95/5 proportion)

Training set:		50% healthy and 50% bankrupt				60% healthy and 40% bankrupt			
Method		Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	Sens.	80.8%	80.8%	80.2%	80.5%	74.7%	75.0%	75.9%	74.4%
	Spec.	82.8%	83.5%	81.8%	84.5%	85.5%	86.3%	85.3%	89.9%
LR	Sens.	80.1%	80.5%	81.2%	82.5%	76.1%	76.2%	78.0%	71.0%
	Spec.	84.8%	82.6%	82.4%	83.3%	87.2%	89.2%	88.4%	88.6%
NN	Sens.	80.5%	80.0%	80.6%	84.4%	77.0%	75.2%	78.3%	76.3%
	Spec.	84.3%	84.6%	81.1%	85.5%	87.8%	87.7%	86.7%	88.8%
SVM	Sens.	82.5%	82.6%	80.0%	84.0%	80.5%	80.8%	78.0%	82.5%
	Spec.	84.1%	84.9%	82.9%	85.0%	87.7%	86.0%	85.8%	87.8%
RF	Sens.	81.2%	80.3%	81.6%	82.4%	79.4%	77.3%	78.7%	77.6%
	Spec.	83.7%	84.4%	82.3%	83.7%	85.7%	84.9%	85.0%	85.9%
Training set		70% healthy and 30% bankrupt				80% healthy and 20% bankrupt			
		Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	Sens.	65.7%	65.0%	63.5%	68.0%	45.7%	45.0%	43.5%	48.0%
	Spec.	94.5%	95.6%	94.6%	93.6%	96.1%	96.5%	95.8%	95.8%
LR	Sens.	66.8%	66.2%	63.3%	70.0%	46.8%	46.2%	43.3%	50.0%
	Spec.	93.4%	94.3%	95.2%	93.9%	95.7%	96.7%	96.0%	96.5%
NN	Sens.	61.0%	60.7%	69.2%	61.5%	51.0%	50.7%	49.2%	51.5%
	Spec.	93.3%	92.4%	93.3%	94.5%	95.6%	93.1%	94.6%	95.4%
SVM	Sens.	77.0%	76.5%	73.0%	79.5%	67.5%	65.5%	70.0%	72.5%
	Spec.	90.2%	91.3%	89.5%	90.2%	92.0%	91.8%	90.4%	91.3%
RF	Sens.	74.6%	70.1%	75.9%	72.6%	63.8%	56.5%	64.3%	57.1%
	Spec.	89.5%	88.7%	88.5%	90.6%	91.4%	92.1%	90.9%	91.8%
Training set		90% healthy and 10% bankrupt				95% healthy and 5% bankrupt			
		Serv	Reta	Cons	All	Serv	Reta	Cons	All

LDA	Sens.	29.5%	30.8%	27.0%	32.1%	11.0%	13.3%	10.0%	12.6%
	Spec.	97.5%	96.8%	97.3%	97.8%	98.6%	98.0%	98.8%	98.3%
LR	Sens.	33.2%	36.2%	35.0%	31.8%	11.2%	7.9%	12.5%	10.7%
	Spec.	96.5%	97.0%	97.0%	96.3%	97.6%	98.7%	97.9%	98.0%
NN	Sens.	32.2%	37.0%	35.4%	32.8%	13.3%	12.1%	16.5%	15.4%
	Spec.	96.7%	96.5%	95.8%	97.4%	98.9%	98.0%	98.3%	98.7%
SVM	Sens.	45.0%	49.6%	52.5%	45.8%	20.0%	19.2%	17.5%	22.0%
	Spec.	92.8%	93.4%	91.7%	92.4%	93.6%	94.4%	95.3%	93.2%
RF	Sens.	49.7%	41.5%	46.7%	38.5%	21.6%	22.3%	19.7%	19.4%
	Spec.	93.4%	92.9%	93.4%	92.0%	94.7%	94.4%	95.0%	94.2%

Notes: Sens. = sensitivity metric; Spec. = specificity metric. Serv = service sector; Cons = construction sector; Reta = retail sector; All = all sectors. LDA = linear discriminant analysis; LR = logistic regression; NN = neural network; SVM = support vector machine; RF = random forest.

Table 4 shows that on the whole, prediction performance decreases for failed firms (sensitivity) and increases for non-failed firms (specificity) when the training set presents an imbalanced distribution. It also confirms that prediction methods reward the classification of the majority class to the detriment of the minority class in imbalanced training sets. These results occur because in the presence of an imbalanced training set, the classification boundaries of the majority class tend to invade those of the minority class, thereby biasing the classification toward the majority class (Kim et al., 2015). In this regard, we contextualize the results by analyzing sensitivity and specificity metrics by the degree of imbalance. We find that sensitivity, on average, achieves a rate of 81.3% balanced proportion, whereas for a training set that is slightly imbalanced (60/40), the prediction already represents a rate of 77.1%. Moreover, when the imbalance of the proportion becomes severe, it decreases to a rate of 69.0% in a 70/30 proportion and 54.4% in an 80/20 proportion. Finally, in the most extreme imbalanced scenarios (90/10 and 95/5 proportions), models achieve average rates of 38.1% and 15.4%, respectively. Thus, when a training set is slightly imbalanced (60/40), the prediction loss on failed firms is 5.4%. However, the loss continues to increase as the imbalanced proportion grows more severe, losing one-third of its prediction accuracy for failed firms (33.5%) and 82% in the most extreme case (95/5 proportion). Any degree of imbalance thus may be significantly detrimental to the prediction of firm bankruptcy. With regard to non-bankrupt firms, however, the specificity value is 83.6% on a balanced proportion, increases to 87.0% in a 60/40 proportion, 92.4% in a 70/30 proportion, and 94.0% in an 80/20 proportion. Finally, the rates are 95.2% and 96.7% in the two most imbalanced distributions (90/10 and 95/5). Therefore, specificity achieves an improvement that ranges from 6% in the least imbalanced training set to 11% in the most imbalanced scenario. However, bankruptcy prediction methods are not affected equally by imbalanced distributions, especially when used to predict firm failures. In this regard, though LDA achieves a sensitivity rate of 45.5% and LR and NN obtain sensitivity values of 46.6% and 47.8%, respectively, on average the SVM and RF methods significantly outperform the rest of the bankruptcy prediction methods, achieving

58.7% and 59.7% respectively. In contrast, there is a difference of only 3.3 percentage points among prediction methods with regard to specificity rates (89.5% with RF, 92.8% with LR). Accordingly, the SVM and RF methods emerge as more efficient than other methods in imbalanced datasets, because it provides more accurate predictions of firm failures. However, their predictions with regard to non-failed firms are slightly inferior.

To better explain and assess our previous results, in Table 5 we present the G-mean values that measure overall prediction in terms of a ratio of sensitivity and specificity. This table shows that losses in performance caused by less imbalanced proportions may be insignificant, because the average values are similar to those obtained with a balanced proportion (82.4% with 50/50, 81.9 with 60/40, and 79.7 with 70/30). Losses tend be more pronounced in the imbalanced distributions that follow, with an average loss of about 15% with an 80/20 proportion, about 30% with a 90/10 proportion, and greater than 50% with a 95/5 proportion. These results are not entirely unexpected; the G-mean takes into account the trade-off between sensitivity and specificity. All methods behave in similar ways until the distribution ranges around 80/20, because the increase in the specificity rate compensates for the slight decrease in the sensitivity rate. Accordingly, we observe eight imbalanced cases with 60/40 and 70/30 proportions, in which there are small performance gains compared with the balanced results. In imbalanced distributions at that point and above, losses in sensitivity appear, whereas specificity remains almost steady implying a significant reduction of the G-mean value. In general, all methods show losses in performance in each of the non-balanced distributions, though the SVM and RF are less affected. These results corroborate our initial finding that SVM and RF seem to be more suitable methods for imbalanced datasets.

Table 5

G-mean values achieved with prediction models on balanced and imbalanced training sets by sample

Training set:	50% healthy and 50% bankrupt				60% healthy and 40% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	81.7	82.1	80.9	82.4	79.9	80.4	80.3	81.7
LR	82.4	81.5	81.7	82.8	81.4	82.4	83.0	79.3
NN	82.3	82.2	80.8	84.9	82.2	81.0	82.3	82.1
SVM	83.2	83.7	81.4	84.4	84.0	83.3	81.8	85.1
RF	82.4	82.3	81.9	83.1	82.5	81.0	81.9	81.6
Training set	70% healthy and 30% bankrupt				80% healthy and 20% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	78.7	78.8	77.5	79.7	66.2	65.8	64.5	67.8
LR	78.9	79.0	77.6	81.0	66.9	66.8	64.4	69.4
NN	75.4	74.8	80.3	76.2	69.8	68.7	68.2	70.1
SVM	83.3	83.5	80.8	84.4	78.8	77.5	79.5	81.3
RF	81.7	78.9	81.9	81.1	76.4	72.1	76.4	72.4
Training set	90% healthy and 10% bankrupt				95% healthy and 5% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All

LDA	53.6	54.6	51.2	56.0	32.9	36.1	31.4	35.1
LR	56.6	59.2	58.2	55.3	33.0	27.9	34.9	32.4
NN	55.8	59.7	58.3	56.5	36.2	34.4	40.2	38.9
SVM	64.6	68.0	69.3	65.0	43.2	42.5	40.8	45.2
RF	68.1	62.0	66.0	59.5	45.2	45.8	43.3	42.7

Although sensitivity, specificity, and G-mean rates clearly reveal the effect of imbalanced distributions on prediction performance, we cannot make a general conclusion from these rates, because the firm size group is very uneven and evaluated on a given default threshold. Therefore, we computed the AUC, because it does not account for the two types of firm size, which makes it the most suitable measure to analyze the overall performance of methods for imbalanced datasets. To address the research questions -that is, which degree of imbalance significantly affects the performance of bankruptcy prediction methods and whether all prediction methods are equally sensitive to imbalanced distributions- we calculated the AUC values achieved by bankruptcy prediction methods with balanced and imbalanced distributions (Table 6). These results are complemented by results (Table 7) that indicate the degree of imbalance in which AUC values are statistically different at the 1% threshold, compared with the balanced proportion.

On the whole, AUC values are similar when the training set is moderately imbalanced. With a balanced distribution, AUC achieves an average value of 0.883, whereas for a 70/30 imbalanced distribution, the average value is 0.862, which is not significant (p = 0.134)[6]. The next level of imbalance (80/20 proportion) represents the barrier at which the method's performance is significantly disturbed, and the AUC decreases dramatically to an average value of 0.801. In the most extreme imbalanced scenarios, it achieves average values of 0.762 and 0.731on 90/10 and 95/5 proportions, respectively. Compared with the AUC of the balanced distribution, these values are significantly different at the 1% threshold. As a result, we establish that, starting with the 80/20 imbalanced distribution on the training set, the performance of bankruptcy prediction methods is significantly jeopardized.

Finally, the performance of all bankruptcy prediction methods are equally affected by imbalanced distribution, except for SVM and RF, which exhibit greater AUC values and seems insensitive to all but the most severely imbalanced distributions. Table 7 shows that SVM and RF are the least affected methods; the AUC differences are statistically significant at the 90/10 distribution in three of four samples and in two out of four respectively, and the other methods all show significant differences at the 80/20 distribution.

[6] We calculated the statistical differences between the two AUC values using Delong et al.'s (1988) non-parametric test.

The fact that RF and, especially, SVM can handle certain imbalanced datasets and lead to better performance is of a significant importance for business community because banks and financial institutions seem to be reluctant to these techniques; their bankruptcy prediction models mainly rely on parametric model (discriminant analysis and logistic regression). Thus, under certain imbalance property in real datasets for bankruptcy prediction, our results suggest that failure models currently used by banks lead to non-optimal performance and that non-parametric models would it make possible to better address for bankruptcy.

With these results, we make two general conclusions. First, any degree of imbalance somewhat damages a method's prediction capacity. Overall prediction performance is significantly affected starting with a 80/20 imbalanced distribution on the training set. Second, the SVM method is less sensitive to imbalanced distributions, which suggests its benefits as a way to deal with imbalanced datasets. Our finding that SVM leads to better results in imbalanced distributions is not entirely surprising; the data imbalance issue implies that classification rules are biased toward the majority class in the learning phase. The SVM follows a structural risk minimization strategy (Vapnik, 1998), compared to the conventional empirical risk minimization strategy followed by the other competing methods, relying not only on minimizing the number of classification errors in the learning process but on maximizing the margin between examples of both classes as well. This allows the SVM to avoid this majority class bias; and therefore, it is in general a more suitable and robust method for imbalanced datasets, as Wang and Japkowicz (2004) claim.

Table 6

AUC values achieved with prediction models on balanced and imbalanced training sets by sample

Training set:	50% healthy and 50% bankrupt				60% healthy and 40% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	0.873	0.881	0.865	0.883	0.857	0.873	0.849	0.866
LR	0.884	0.890	0.876	0.878	0.863	0.877	0.861	0.859
NN	0.887	0.889	0.868	0.898	0.873	0.871	0.854	0.882
SVM	0.890	0.892	0.870	0.895	0.879	0.877	0.856	0.883
RF	0.888	0.884	0.880	0.890	0.875	0.872	0.869	0.877
Training set	70% healthy and 30% bankrupt				80% healthy and 20% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	0.850	0.857	0.841	0.859	0.780	0.791	0.776	0.784
LR	0.858	0.870	0.853	0.857	0.778	0.779	0.785	0.771
NN	0.864	0.863	0.850	0.873	0.790	0.789	0.773	0.787
SVM	0.868	0.873	0.851	0.876	0.862	0.818	0.845	0.871
RF	0.870	0.859	0.864	0.862	0.843	0.804	0.830	0.807
Training set	90% healthy and 10% bankrupt				95% healthy and 5% bankrupt			
	Serv	Reta	Cons	All	Serv	Reta	Cons	All
LDA	0.758	0.736	0.755	0.734	0.718	0.701	0.697	0.719
LR	0.733	0.741	0.760	0.758	0.706	0.688	0.721	0.710

NN	0.766	0.757	0.743	0.764	0.717	0.715	0.738	0.724
SVM	0.789	0.784	0.793	0.790	0.747	0.752	0.769	0.764
RF	0.795	0.770	0.787	0.768	0.758	0.759	0.772	0.746

Table 7
Imbalanced proportion at AUC values compared to balanced distribution values at significantly different 1% threshold

	Serv.	Reta.	Cons.	All
LDA	80/20	80/20	80/20	80/20
LR	80/20	80/20	80/20	80/20
NN	80/20	80/20	80/20	80/20
SVM	90/10	80/20	90/10	90/10
RF	90/10	80/20	90/10	80/20

4.2. Results of sampling techniques in imbalanced datasets

After establishing that the bankruptcy prediction method's loss of performance is significant starting with an 80/20 imbalanced proportion, we explore the capacity of sampling methods to overcome this loss of extreme imbalanced proportions. The results were similar among the four samples used (service, retail, construction, and all sectors), so we averaged them to analyze the contributions of the sampling techniques[7]. We first computed the sensitivity, specificity, and G-mean values achieved when applying sampling methods to imbalanced training sets, to demonstrate the effect of these techniques in terms of predicting firm bankruptcy, non-bankruptcy, and overall outcomes. We present these values in Tables 8–10, in which Panel A of each table indicates the results achieved in the 80/20 imbalanced proportion with the four sampling techniques, Panel B indicates results of the 90/10 proportion, and Panel C indicates those of the 95/5 proportion.

Table 8

Sensitivity rates achieved with sampling methods on imbalanced training sets (in percentages)

Models	Panel A: 80NB/20B proportion				Panel B: 90NB/10B proportion				Panel C :95NB/5B proportion			
	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E
LDA	77.1	75.2	73.6	75.6	75.0	71.8	70.0	72.1	73.1	69.3	70.8	68.5
LR	77.7	75.9	72.4	76.5	75.6	72.5	73.9	74.6	73.0	70.3	71.5	71.1
NN	76.6	78.9	73.0	79.2	72.5	76.5	75.9	77.1	70.4	74.9	74.0	75.7
SVM	77.1	79.4	74.1	80.6	72.9	77.3	73.4	76.8	70.6	74.5	73.3	72.0
RF	79.8	76.5	81.1	77.8	76.7	73.2	78.0	75.2	74.8	71.5	76.2	72.1

Notes: R.O = random oversampling, R.U: = random undersampling, SMO = synthetic minority oversampling technique, E.E = EasyEnsemble. NB = non-bankrupt, B = bankrupt.

[7] To maintain conciseness and noting that 240 results (5 methods × 4 samples × 4 sampling techniques × 3 imbalanced proportions) are not readable, the following results present the average of each sample.

Table 9

Specificity rates achieved with sampling methods on imbalanced training sets (in percentages)

Models	Panel A: 80NB/20B proportion				Panel B: 90NB/10B proportion				Panel C :95NB/5B proportion			
	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E
LDA	80.5	79.3	81.2	77.5	80.0	76.7	79.5	78.2	79.3	75.5	78.6	73.6
LR	80.9	78.7	81.0	78.3	80.5	76.1	78.7	76.8	79.8	75.3	77.9	74.5
NN	79.1	80.7	84.7	80.0	77.3	80.6	82.9	78.9	77.6	79.1	81.7	76.0
SVM	78.5	81.5	83.9	79.3	78.6	80.9	81.3	77.8	79.4	79.8	80.1	76.3
RF	80.6	79.5	82.9	79.7	80.3	78.4	82.0	78.9	79.8	77.1	81.4	78.2

Table 10

G-mean values achieved with sampling methods on imbalanced training sets (in percentages)

Models	Panel A: 80NB/20B proportion				Panel B: 90NB/10B proportion				Panel C :95NB/5B proportion			
	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E
LDA	78.8	77.2	77.3	76.4	77.5	74.1	74.7	74.8	76.3	72.1	74.7	71.0
LR	79.4	77.0	76.7	77.2	78.2	74.0	76.3	75.4	76.4	72.7	74.3	72.6
NN	77.9	79.5	78.8	79.2	74.9	78.1	79.5	77.8	73.9	76.6	79.8	75.4
SVM	78.0	80.1	79.1	79.7	75.8	79.0	77.6	77.0	74.8	77.0	76.6	74.0
RF	80.1	77.9	81.9	78.7	78.4	75.7	79.9	77.0	77.3	74.2	78.7	75.1

Applying sampling techniques increases sensitivity rates (failed firms) and decreases specificity rates, compared with the results of the imbalanced proportion (Table 4). If we evaluate the results on failed firms (sensitivity) -the class most affected by the data imbalance issue- sensitivity rates increase significantly when we apply sampling methods to the imbalanced sets 76.9% of failed firms, on average, are now predicted correctly on the 80/20 imbalanced proportion, the 90/10 proportion achieves an average rate of 74.5%, and the 95/5 proportions achieve an average rate 72.4%. Thus, the effect of sampling techniques implies an average increment of 41.1 percentage points in the prediction of failed firms compared with results achieved in imbalanced distributions (Table 4). In contrast, with regard to non-failed firms (specificity), the models obtain an average rate of 79.2%, which is a 5.1 percentage point decrease relative to the original balanced distribution result. The effect of sampling techniques thus implies a trade-off between sensitivity and specificity, which is critical. The gain

obtained by correctly predicting failed firms is much greater than the loss of performance for non-failed firms. In reality, models that better predict fates of failed firms are preferred, due to the misclassification cost asymmetry between failed and non-failed firms. In turn, G-mean values increase significantly, approximating those obtained when the methods are designed with 70/30 proportions. Moreover, in almost all the cases, all methods behave similarly. That is, assuming that sampling techniques are used, one method does not outperform another. However, a method's performance depends on the sampling approach adopted. The oversampling approach (random oversampling and SMOTE) exhibits higher average values (77.5), relative to undersampling (76.1).

To provide conclusive results about the performance recovery capacity of sampling methods, we computed AUC values achieved in the various imbalanced proportions by sampling techniques (Table 11). These values complement those that indicate the performance recovery obtained by each model we designed (Table 12). Note that to analyze recovery capacity, it is necessary to account for its gain of performance by using a sampling technique with the loss of performance caused by imbalanced proportions. Therefore, we computed the recovery percentage as follows:

$$Rec = \frac{S-I}{B-I},$$

where S represents the AUC value achieved using a sampling method, I is the AUC obtained in a given imbalanced proportion, and B denotes performance in a balanced proportion.

These results, estimated with AUCs, show that even though sampling techniques obtain satisfactory recovery on method performance -that is, an average recovery of 43.9% the AUC values for the original balanced set are not reached (average AUC value of 0.815)[8]. However, the two most imbalanced samples (90/10 and 95/5 proportions) obtain the highest recoveries, 47.5% and 46.3%. Moreover, there are some discrepancies among both sampling techniques and prediction models. If we analyze recovery capacity by sampling techniques, the SMOTE technique achieves an average recovery of 54.7%, whereas the EasyEnsemble technique achieves a 40.9% average recovery. Random oversampling and random undersampling obtain average recoveries of 40.4% and 39.0%, respectively. These results highlight the superiority of the most sophisticated sampling techniques, SMOTE and EasyEnsemble, in dealing with data imbalances. These techniques represent further developments of random oversampling and random undersampling techniques, designed specifically to overcome the drawbacks and increase the performance of models in imbalanced datasets. However, when we evaluate recovery by type of model, we find that NN achieves the highest average performance recovery (57.0%), whereas LDA, LR, SVM and RF achieve performance recoveries of 49.8%, 50.5%, 19.7% and 41.8%, respectively. Therefore, the NN

[8] If we consider only the recoveries achieved by LDA, LR, NN and RF, we uncover an average recovery of 50.3%.

prediction model clearly outperforms the recovery obtained by the other prediction models. Yet the SVM prediction model achieves two negative and two small recoveries that condition its low-percentage recovery. These results can be explained by noting that SVM can learn from certain imbalanced datasets (Imam et al. 2006; Li et al. 2008). Therefore, given that SVM seems insensitive to an 80/20 imbalanced proportion, better performance can be achieved by using SVM directly in the 80/20 imbalanced proportion rather than in combination with a sampling technique. Moreover, some studies document that classifier performance in a certain imbalanced dataset may be superior or comparable to performance on a balanced dataset using a sampling technique (Batista et al., 2004; Japkowicz and Stephen, 2002). Our results thus suggest that performance recoveries differ slightly, depending on the sampling technique and the prediction model used; in concrete cases, the application of sampling techniques is not always beneficial.

Table 11

AUC values achieved with sampling techniques in imbalanced training sets

Model	Panel A: 80NB/20B proportion				Panel B: 90NB/10B proportion				Panel C :95NB/5B proportion			
	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E
LDA	0.844	0.831	0.840	0.824	0.823	0.798	0.804	0.813	0.800	0.771	0.789	0.763
LR	0.847	0.827	0.843	0.834	0.836	0.802	0.806	0.820	0.803	0.774	0.793	0.769
NN	0.818	0.840	0.856	0.853	0.806	0.834	0.843	0.837	0.793	0.826	0.832	0.809
SVM	0.822	0.842	0.851	0.850	0.808	0.831	0.839	0.826	0.797	0.829	0.836	0.786
RF	0.854	0.835	0.861	0.842	0.838	0.804	0.847	0.815	0.816	0.791	0.842	0.783

Table 12

AUC recovery obtained with sampling techniques (in percentages)

Model	Panel A: 80NB/20B proportion				Panel B: 90NB/10B proportion				Panel C :95NB/5B proportion			
	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E	R.O	R.U	SMO	E.E
LDA	66.6	52.6	62.3	45.1	59.8	40.3	44.9	51.9	55.0	37.7	48.5	32.9
LR	66.3	47.1	62.5	53.8	65.6	40.0	35.8	55.2	55.4	38.9	49.7	36.1
NN	33.0	55.0	73.0	68.0	38.3	60.1	67.1	62.3	43.2	63.5	67.3	53.0
SVM	-71.0	-18.4	5.2	2.6	19.4	42.8	51.0	37.7	30.2	55.0	60.4	21.7
RF	51.6	21.9	62.5	32.8	55.2	22.8	63.8	33.3	45.7	26.0	66.2	19.9

4.3. Model performance on different training set sizes

The use of sampling techniques requires modifying the original training set by duplicating the minority class or removing the majority class, so sample size is important. Different numbers of samples are processed in training sets according to size and sampling technique, to achieve balanced distributions. Thus, the performance of sampling techniques may depend on both training set size (Weiss and Provost, 2003) and prediction methods' performance (Back et al., 1997). To assess all models' performance in imbalanced datasets, sample size should be considered along with sampling techniques, and we perform such an assessment in this section to contextualize our previous results. We created a set of training sets of different sizes[9] (2,000, 3,000, and 4,000 firms) for all samples collected, using the same three imbalanced proportions. Although empirical sample size in bankruptcy prediction is diverse, most datasets contain fewer than 4,000 total samples (Kumar and Ravi, 2007). Thus, our samples offer a stratified representation. We randomly selected new samples from the database until we achieved the desired total of firms. Table 13 indicates their configuration.

Table 13: Training set configuration by size

	Training set proportions		
Nº firms in training set	80/20	90/10	95/5
Size of 1,500 firms			
B	300	150	75
NB	1200	1350	1425
Size of 2,000 firms			
B	400	200	100
NB	1600	1800	1900
Size of 3,000 firms			
B	600	300	150
NB	2400	2700	2850
Size of 4,000 firms			
B	800	400	200
NB	3200	3600	3800

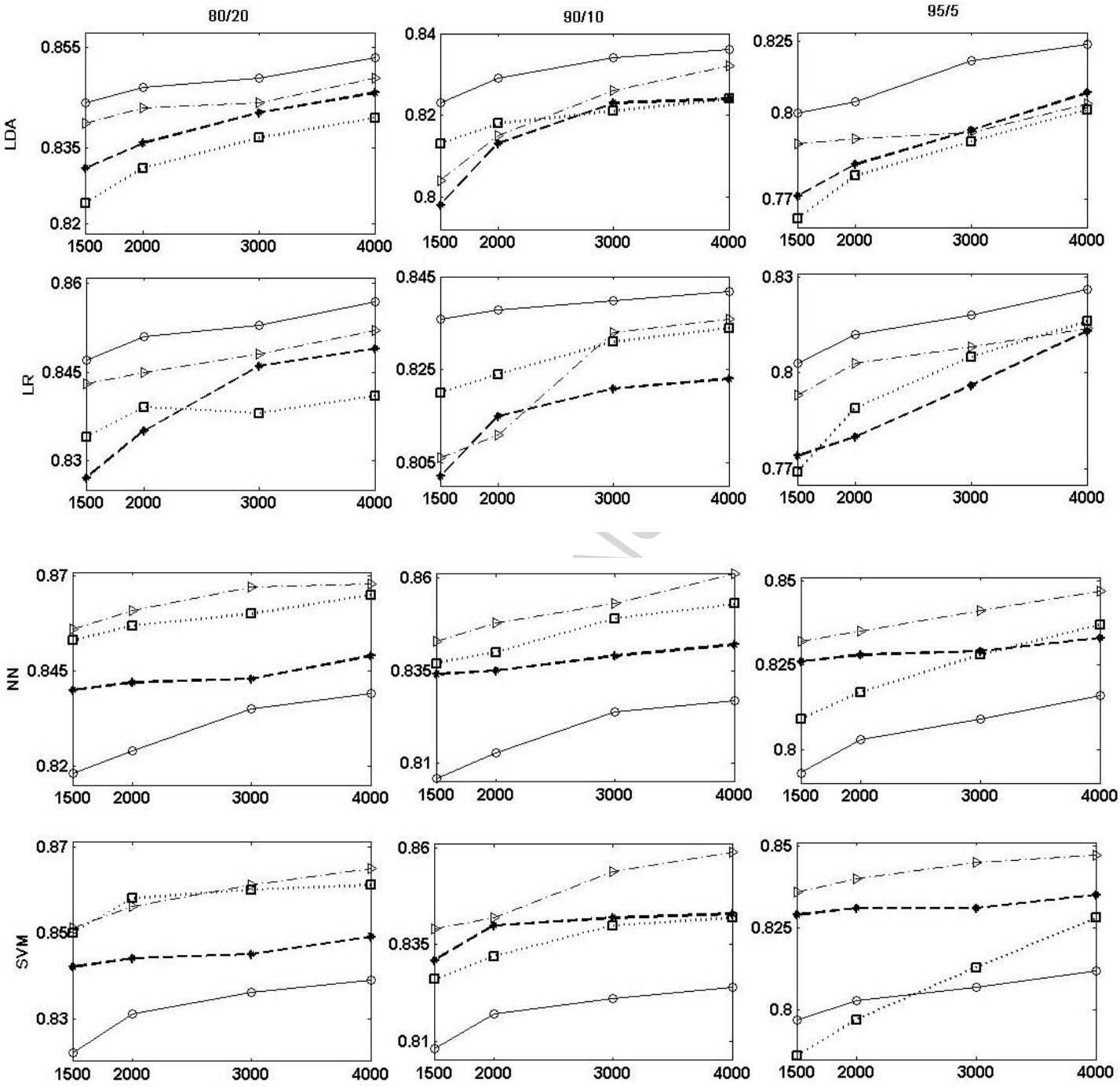
Notes: NB = non-bankrupt, B = bankrupt.

[9] We designed the training set size according to the firms available from our database, and 4,000 firms is the largest training set that we could create.

We then ran experiments to establish the relationships among model performance, sampling techniques, and training set size in imbalanced datasets. Figure 2 provides a graphical representation of the model's evolution of performance (measured by AUC values) by type of sampling technique and training set size. Its performance always improves when training set size increases, which confirms the impact of this factor on the prediction of bankruptcy in imbalanced distributions. Although a change in training set size from 1,500 to 3,000 firms shifts the performance curve upward (significant performance improvement), the curve begins to flatten in the next training size level (from 3,000 to 4,000 firms), such that performance remains almost steady. Thus, the relevance of training set size becomes less important as the set gets larger.

Figure 2 also shows that models do not behave in similar ways. That is, maximum performance does not always occur with the same sampling techniques, which indicates their sensitivity. When we analyze statistical methods, we note that random oversampling leads to better performance for LDA and LR in all imbalanced scenarios. In contrast, NN and SVM provide diverse results. Although they display the highest and roughly the same performance with the SMOTE and EasyEnsemble in the 80/20 imbalanced proportion, their performance with the latter sampling techniques falls considerably in the other imbalanced proportions; the SMOTE provides maximum performance. Therefore, oversampling is the optimal strategy, because it provides better and steadier performance in diverse scenarios. These results are not entirely odd and sound consistent with the literature that has analyzed the data imbalance issue. The Oversampling approach augments the sample space in a manner that generally improves learning; leading to models with an enhanced discriminatory power. Besides, the fact that SMOTE generates artificial data interpolated between existing minority samples seems to be a better choice, which might serve as an efficient solution for a real bankruptcy prediction problem. The creation of synthetic samples not only enhances the model performance but, it also provides new information that as the time goes by, it may acquire informative meaning to understand and assess business failure processes. In contrast, the Undersampling approach is sub-optimal in almost all scenarios, except those with less imbalanced proportions and larger training sets. This can be explained by the fact that the random elimination of majority class samples performed by this approach may lead to discard potentially useful information that could be important for the induction problem (Kotsiantis et al., 2006). Besides, even though this procedure can handle imbalance datasets, it might produce a small sample size, in which the amount of information possessed may be insufficient to determine an optimal model generalization. Thus, this approach presents two major drawbacks for its materialization in a real solution for bankruptcy prediction. On the one hand, its performance seems dependable to the sample size which is a major condition due to the lack of bankruptcy firms

samples. On the other hand, the elimination of data may be undesirable due to bankruptcy investigation is generally expressed as a function of the data.



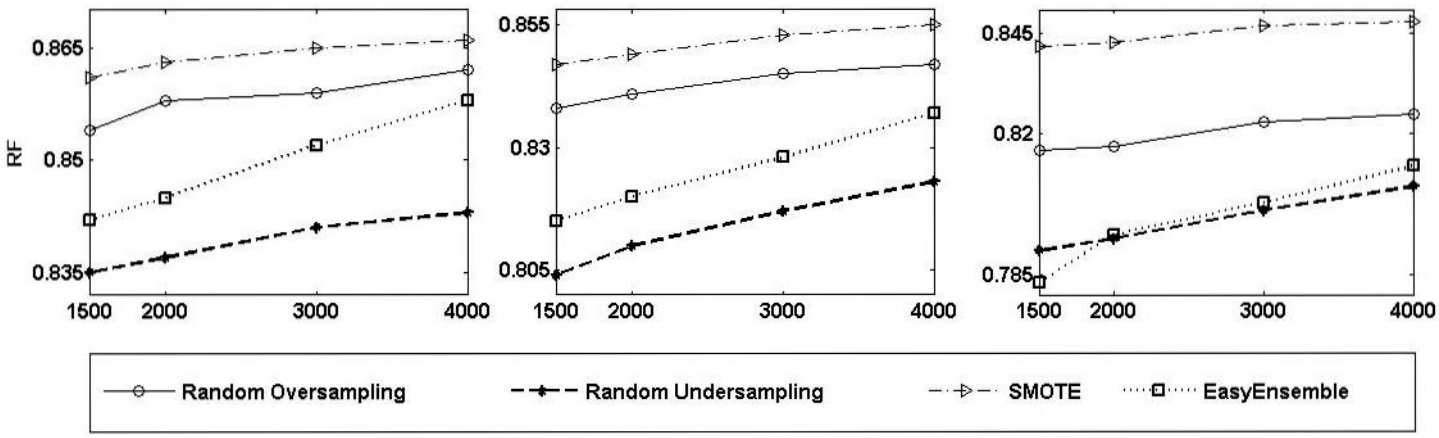


Fig. 2. Evolution of model performance by sampling technique and training set size (X-axis represents the sample size and Y-axis represents the AUC value).

5. Conclusion

We investigate the performance of bankruptcy prediction models in imbalanced datasets by analyzing three key notions: degree of imbalance, loss of performance, and sampling techniques. We establish which imbalanced distribution significantly damages prediction performance. Models built on training sets, in which bankrupt firms represent equal to or less than 20% of the total samples, suffer significantly diminished prediction performance. Although the performance of all classifiers is affected by imbalanced datasets, especially as that imbalance grows greater, the results that the SMV method is less sensitive. That is, it only suffers significant losses in performance in the most extreme scenarios (90/10 and 95/5 class proportions).

We also provide experimental results with regard to treatment methods and sampling techniques in imbalanced datasets. When we analyze the capacities of sampling techniques to recover prediction performance by balancing training sets, the results indicate an acceptable average recovery of 43.9%. Moreover, bankruptcy prediction models perform differently, depending on the sampling techniques used. In this regard, oversampling is a better choice, because it is most suitable for all type of prediction models and different training set sizes.

We also take a novel perspective that investigates the intercorrelations among the degree of data imbalance, the bankruptcy models' loss of performance, and sampling techniques. We thereby fill a significant knowledge gap and make two main contributions -one methodological and one empirical- to bankruptcy prediction literature.

The data imbalance issue is relevant in the bankruptcy prediction field, because the scarcity of bankrupt firms means that researchers must design models with imbalanced datasets. As a methodological contribution, we establish the limits of imbalanced distributions in datasets as they relate to bankruptcy prediction models. Researchers must be cautious when designing bankruptcy prediction models, because a moderately imbalanced training set (80/20 proportion) can be

detrimental to prediction performance. We also highlight the relevance of sample distributions for the design of bankruptcy prediction models. This finding is important for banks and other financial institutions; even though institutions have information on thousands of firms, their datasets often are limited by the number of bankrupt firms, and few institutions have more than 1,000 or so. In such scenarios, datasets exhibit imbalanced distributions, and the models likely cannot predict bankruptcy well, resulting in severe financial consequences.

With regard to our empirical contribution, our results confirm previous studies that show that applying sampling techniques to handle the data imbalance issue improves the performance of prediction methods. However, only about half of the performance loss generated by data imbalance can be recovered. Moreover, as the size of the training set increases (from 1,500 to 3,000 total samples), the performance curve of the sampling techniques moves upward and eventually begins to flatten. This empirical finding is crucial, because it shows that there is still much work to be done to overcome this significant concern.

Little research has been conducted on analyzing imbalanced datasets in the bankruptcy prediction field, even though in the real world, bankruptcy prediction datasets present vastly imbalanced distributions that hinder bankruptcy prediction performance. The implications of our findings thus are pertinent to both academics and financial institutions. Nonetheless, it is important to remark that our results should be interpreted cautiously. They reflect the information included in the data and the characteristics of the input data. Further research should therefore investigate data imbalance issue in bankruptcy prediction domain in order to empower our results. Moreover, most models are validated in experimental conditions that do not represent real-world scenarios, such as imbalanced datasets. We speculate that even though more sophisticated and complex models are being designed, they will fail to prevent real bankruptcies, because they are rendered suboptimal by data issues. Thus, this study can serve as a guide for researchers and academics to design bankruptcy prediction models in imbalanced datasets and develop new solutions to overcome the problem of data imbalance. It can also serve as an inspiration for developing a bankruptcy prediction model that delivers steady performance, regardless of data imbalance distributions. This effort will represent our next research direction.

Acknowledgement

We are very grateful to the two anonymous reviewers for their substantial contribution to the improvement of this article.

References

Altman, E. I. (1968). Financial Ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 889-609.

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation.* Oxford University Press.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions On Neural Networks, 12*(4), 929-935.

Back, B., Laitinen, T., Hekanaho, J., & Sere, K. (1997). The effect of sample size on different failure prediction methods. *Turku Centre for Computer Science Technical Report,* 155, 1-23.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *British Accounting Review, 38*(1), 63-93.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20-29.

Beaver, W. (1966). Financial ratios as predictor of failure. *Journal of Accounting Research, 4*, 71-111.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144-152.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446-3453.

Campa, D. and Camacho, M (2015). The impact of SME's pre-bankruptcy financial distress on earnings management tools. *International Review of Financial Analysis, 42*, 222-234.

Charitou, A., Lambertides, N., & Trigeorgis, L. (2007). Managerial discretion in distressed firms. *The British Accounting Review, 39*(4), 323-346.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1-6.

Chen, H.-J., Huang, S. Y., & Lin, C.-S. (2009). Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach. *Expert Systems with Applications, 36*(4), 7710-7720.

D'aveni, R. A. (1989). The aftermath of organizational decline: A longitudinal study of the strategic and managerial characteristics of declining firms. *Academy of Management journal*, *32*(3), 577-605.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837-845.

du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, *73*(10), 2047-2060.

du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, *242*(1), 286-303.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18-36.

Fernández, A., García, S., Luengo, J., Bernadó-Mansilla, E., & Herrera, F. (2010). Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study. *IEEE Transactions on Evolutionary Computation*, *14*(6), 913-941.

Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. *Expert Systems with Applications*, *41*(14), 6433-6445.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, 878-887.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, *37*(4), 543-558.

Imam, T., Ting, K. M., & Kamruzzaman, J. (2006). z-SVM: An SVM for improved classification of imbalanced data. In *Australasian Joint Conference on Artificial Intelligence*, 264-273.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429-449.

Kim, M. J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications*, *25*(4), 637-646.

Kim, M. J., Kang, D. K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, *42*(3), 1074-1082.

Kim, T., & Ahn, H. (2015). A hybrid under-sampling approach for better bankruptcy prediction. *Journal of Intelligence and Information Systems, 21*(2), 173-190.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25-36.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research, 180*(1), 1-28.

Lane, P. C., Clarke, D., & Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems, 53*(4), 712-718.

Leshno, M., & Spector, Y. (1996). Neural network prediction analysis: The bankruptcy case. *Neurocomputing, 10*(2), 125-147.

Li, X., Wang, L., & Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence, 21*(5), 785-795.

Lopez, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.

McKee, T. E., & Greenstein, M. (2000). Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. *Journal of Forecasting, 19*(3), 219-230.

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of Accounting Research, 22*(1), 380-395.

Messier, Jr., W. & Hansen, J., (1988). Inducing rules for expert system development: An example using default and bankruptcy data. *Management Science, 34*(12), 1403–1415.

Ohlson, J.A., (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464-473.

Ooghe, H., & Joos, P. (1990). Failure prediction, explanation of misclassifications and incorporation of other relevant variables: Result of empirical research in Belgium. Working paper, Department of Corporate Finance, Ghent University (Belgium).

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*. 106, 15-29.

Rosner, R. L. (2003). Earnings manipulation in failing firms. *Contemporary Accounting Research*, *20*(2), 361-408.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184-203.

Stein, R. M. (2007). Benchmarking default prediction models: Pitfalls and remedies in model validation. *Journal of Risk Model Validation*, *1*(1), 77-113.

Tang, T. C., Chi, L. C. (2005), Neural Networks Analysis in Business Failure Prediction of Chinese Importers : A Between-Countries Approach, *Expert Systems with Applications*, 29(2), 244-255.

Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*(4), 309-317.

Tian, S., Yu, Y., & Zhou, M. (2015). Data sample selection issues for bankruptcy prediction. Risk, *Hazards & Crisis in Public Policy*, *6*(1), 91-116.

Vapnik, V. (1998). Statistical learning theory. Wiley, New York.

Wald, A. (1944). On statistical problem arising in the classification of an individual into one of two groups, *Annals of Mathematical Statistics*, *15*(2), 145-162.

Wang, B. X., & Japkowicz, N. (2004). Imbalanced data set learning with synthetic samples. In *Proc. IRIS Machine Learning Workshop*.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315-354.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision support systems*, *11*(5), 545-557.

Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, *41*, 16-25.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82.

Biographical notes:

David Veganzones is currently pursuing the Ph.D. degree in management & finance at the Institute d'Administration des Enterprises (IAE), University of Lille, Lille, France. He is interested in various domains of bankruptcy prediction and the application of machine learning to corporate finance.

Eric Séverin is a Professor of finance at USTL (University of Lille) and he is a specialist in corporate finance. His research interests are twofold: bankruptcy prediction and the relationship between economics and finance.

Research highlights:

► An investigation of bankruptcy prediction in imbalanced datasets is proposed.

►The prediction losses increase as the imbalanced proportion grows more severe.

►Support Vector Machine method is less affected by imbalanced datasets than other prediction method.

►SMOTE outperforms other sampling techniques for all type of prediction models and different training set sizes.