revue *Recherche en Sciences de Gestion-Management Sciences-Ciencias de Gestión,* n°152 p. 127 à 153

Bankruptcy prediction modeling in real-world conditions: A contrast of boosting algorithm and logistic regression

Xavier Brédart
Professor
University of Mons
(Belgium)

Diego Andrés Correa-Mejía
Professor
University of Antioquia
(Columbia)

This article aims to build bankruptcy forecasting models using techniques that overcome the imbalanced data set issue. The imbalanced data set issue is overcome by applying logit, boosting, and over-sampling techniques to an imbalanced data set of 2,266 Belgian firms. The synthetic minority over-sampling technique (SMOTE) is used to test the accuracy of models on different proportions of imbalanced samples. The results demonstrate that using techniques that consider the imbalanced dataset issue provides better prediction accuracy, especially by reducing type I error rate, which is the costliest economic error. This survey offers important information for investors, suppliers, bankers, and governments.

Keywords: *Bankruptcy - Imbalanced dataset – Boosting – Resampling - Logistic regression – Belgium.*

Cet article vise à construire des modèles de prédiction de la faillite en utilisant des techniques qui prennent en considération les problèmes liés aux bases de données déséquilibrées, en appliquant des techniques de type logit, boosting et de suréchantillonnage à un ensemble de données déséquilibré de 2266 entreprises belges. La technique de suréchantillonnage des minorités synthétiques (SMOTE) est utilisée pour tester la précision des modèles sur différentes proportions d'échantillons déséquilibrés. Les résultats démontrent que l'utilisation de techniques prenant en compte le problème de déséquilibrage des données offre une meilleure précision de prédiction, notamment en réduisant le taux d'erreur de type I, qui constitue l'erreur économique la plus coûteuse. Cette étude offre des pistes intéressantes pour les investisseurs, les fournisseurs, les banquiers et les gouvernements.

Mots-clés: *Faillite - Base de données déséquilibrée – « Boosting » – Rééchantillonnage - Régression logistique – Belgique.*

Este artículo tiene como objetivo construir modelos de predicción de quiebra, utilizando técnicas que tengan en cuenta los problemas asociados con las bases de datos desbalanceadas, mediante la aplicación de técnicas logit, impulso y sobremuestreo a un conjunto de datos desequilibrado de 2266 empresas belgas. La técnica de sobremuestreo de minorías sintéticas (SMOTE) se utiliza para probar la precisión de los modelos en diferentes proporciones de muestras no balanceadas. Los resultados demuestran que el uso de técnicas que tienen en cuenta el problema del desbalanceo de datos ofrece una mejor precisión de predicción, en particular al reducir la tasa de error de tipo I, que es el error más costoso. Este estudio ofrece recomendaciones interesantes para inversores, proveedores, banqueros y gobiernos.

Palabras clave: *Quiebra - Base de datos desbalanceado – Impulso – Remuestreo - Regresión logística - Bélgica.*

Introduction

Because of its effect on shareholders, managers, and human resources, bankruptcy is a crucial topic in the field of corporate finance, and bankruptcy prediction has been well studied (e.g., Altman, 1968; Ohlson, 1980; Kim & Kang, 2010; Serrano-Cinca, Gutiérrez-Nieto, & Bernate-Valbuena, 2019). Bankruptcy prediction modeling is also important for the financial sector, especially for banks and rating agencies. Banks need effective models to manage the allocation of their resources to firms that will be able to reimburse them, and rating agencies must assess the health of firms to inform investors accurately.

Data and their characteristics are the most crucial elements of any prediction model (Anderson, 2007). Nevertheless, most bankruptcy prediction models use data sets that do not represent real-world conditions. Some models use paired samples of firms that contain the same number of failed and non-failed firms (Daily & Dalton, 1994; Ciampi, 2015), though bankruptcy is rarely observed in the real-world. Paired data optimize overall prediction accuracy if the model is run on an imbalanced data set, but these models do not take the disproportion between the number of failed and non-failed firms into account (Lopez et al., 2013), which results in a poor classification rate for the minority class (Wilson & Sharda, 1994). Specifically, a type I error (i.e., misclassifying a bankrupt firm as non-bankrupt) tends to be high in models using imbalanced data sets. This consideration is particularly important for creditors because it implies costs that represent total or partial losses of credits granted, solely due to a poor risk estimate. Several solutions, such as the sequential boosting technique and resampling, seek to resolve this issue. Although literature about imbalanced data sets is relatively abundant (e.g., Piri, Delen, & Liu, 2018; Lopez et al., 2013; Estabrooks, Jo, & Japkowicz, 2004; Saez et al., 2015; He & Garcia, 2009), few studies focus on imbalanced data sets in the bankruptcy prediction field (e.g., Kim, Kang, & Bae, 2015; Zhou, 2013; Veganzones & Séverin, 2018).

Using a data set of 2,266 Belgian firms, including 153 bankrupt firms and 2,113 non-failed firms, we compare the accuracy of different prediction models, using information from firms' balance sheets and income statements. First, we apply a logit modelization of the original data set. Second, we use boosting, which is an ensemble

technique that sequentially builds models that assign more weight to incorrectly classified observations. Third, because resampling methods provide better results than imbalanced distributions (Estabrooks, Jo, & Japkowicz, 2004), we resample the imbalanced data set by adding synthetic observations into the sample to create a balanced distribution.

The results show that the global accuracy rate is higher using the boosting algorithm than logistic regression. However, because these techniques still report high rates of type I errors, we test the models on different proportions. The results, presented using confusion matrixes with balanced and imbalanced data distribution and through received operation characteristic (ROC) curves, enable us to identify the precision of the prediction using the area under the curve (AUC).

We establish two main results. First, boosting provides better prediction results than logistic regressions, especially regarding type I error. Second, regarding the degree of the imbalance of the data sets, we achieve the best results with balanced (1:1) samples. A synthetic minority over-sampling technique (SMOTE) creates a balanced distribution that decreases type I errors in both logistic and boosting models.

The remainder of this paper is organized as follows: Section 2 provides a literature review of corporate failure models and imbalanced data set issues. Section 3 describes our research methodology. In Section 4 we present and discuss the results, and then Section 5 concludes.

1. – Literature review

According to Ben (2017) bankruptcy is presented when companies incur in non-payment debts. This situation affects both the companies and their different creditors. Companies can enter in a bankruptcy state for two reasons: default or if they do not have the resources to pay their obligations (Li & Faff, 2019).

Corporate failure modeling was pioneered by Beaver's (1966) discriminant analysis of a single financial ratio. Then Altman (1968), Olhson (1980), and Zmijewski (1984) developed statistical methods to discriminate between failed and non-failed firms. In the 1990s, some authors relied on artificial intelligence methods, such as

neural networks (Odom & Sharda, 1990), for corporate failure prediction modeling. Ensemble methods, such as boosting (du Jardin, Veganzones, & Séverin, 2017), also have been used for corporate failure prediction. Although some authors have extended detection models to include non-financial variables (e.g., Tobback et al., 2017; Ciampi, 2015), financial information represents the main input for bankruptcy prediction.

Most bankruptcy modelizations use balanced samples, including the same proportion of failed and non-failed firms. This "paired sample" (generally by size or industry) technique prevents the model from neglecting prediction accuracy rates for failed firms. Nevertheless, in this case sample selection bias might occur (Zmijewski, 1984). Among the few studies (Wilson & Sharda, 1994; McKee & Greenstein, 2000) that build prediction models using imbalanced data sets, closer to real-world conditions, the results indicate that models built using balanced samples outperform those built with imbalanced samples, especially for failed firms.

Researchers have tried to improve model accuracy, especially with regard to the classification rate of failed firms, for imbalanced data sets. According to Kang and Cho (2006), two methods traditionally have been used to resolve the issues of imbalanced data sets: resampling the data or assigning different weights (i.e., penalties) to observations, depending on their misclassification instances.

The first solution resamples the data set to make its distributions balanced. In the context of bankruptcy prediction, this technique manipulates the data from the original imbalanced data set to create a balanced data set that contains the same number of bankrupt and non-failed firms. This data manipulation uses standard classification techniques and ensures that the model accounts for prediction accuracy for the class of failed firms. To create these balanced data sets, two methods exist. Under-sampling removes observations from the majority class, whereas over-sampling duplicates or creates synthetic observations to increase the number of cases of the minority class. The under-sampling method reduces the time spent training the models but loses information, because observations get deleted (Seiffert et al., 2008). In the case of bankruptcy prediction, real-word conditions result in an enormous loss of healthy firms from the data set. In contrast, the over-sampling method does not imply any loss of information but requires more time

to train the models (Japkowicz & Stephen, 2002; Seiffert et al., 2008) and can lead to over-fitting.

Because bankruptcy is a rare event, bankruptcy prediction is often modeled using small databases. Moreover, the original sample is often divided into training and testing subsamples. Thus, the results can be very sensitive to the sample. Zhou (2013) and Kim and Ahn (2015) use sampling techniques on originally imbalanced data sets and report improved accuracy following the resampling. Various under- (e.g., random, easy ensemble) and over- (e.g., random, SMOTE) sampling techniques have been proposed. Veganzones and Sévérin (2018) report that a model using less than 20% of failed firms jeopardizes its ability to predict bankruptcy, and the SMOTE resampling technique that creates synthetic observations to increase the number of cases of the minority class outperforms other sampling techniques.

The second solution to the imbalanced data set issue, without proceeding to a resampling of the training data set, relies on cost-sensitive classification methods that assign penalties to misclassified instances. Unlike resampling techniques, these methods do not modify the data distribution, so they avoid the problems inherent to resampling techniques. Nevertheless, cost-sensitive classification methods might be highly sensitive to sample characteristics and potentially generate unstable classifiers (Kim, Kang & Bae, 2015). The boosting technique (Schapire, 1990) sequentially builds models in which a higher weight (i.e., penalty) is assigned to incorrectly classified observations. Because it provides more learning opportunities for minority class samples, which are more likely to be misclassified than majority class samples, boosting is an appropriate technique to solve data imbalance problems (Kim, Kang, & Bae, 2015) and to model bankruptcy prediction in real-world conditions, as shown in several studies (e.g., Kim, Kang, & Bae, 2015; du Jardin, Veganzones, & Séverin, 2017). According to du Jardin, Veganzones, and Séverin (2017), boosting leads to more accurate models than single models in the field of bankruptcy prediction.

For bankruptcy prediction, modeling real-world conditions implies the use of imbalanced data sets; using resampling and a cost-sensitive classification boosting technique might build more accurate prediction models. The purpose of this study is to determine the best solution to overcome the issue of imbalanced data sets in the context

of bankruptcy prediction, thus leading to better assessments of the default risk of firms.

2. – Methodology

2.1. Data

We gathered data from the Bureau Van Dijk Bel-First database, which provides financial information about Belgian firms. The initial database includes 7,814 firms. However, 5,548 firms were not considered in this study because they did not report the required financial information (i.e., the explanatory variables) to develop the prediction. After the elimination of the incomplete data, we identified 153 firms that went bankrupt in 2017 and 2,114 non-bankrupt firms in 2017, which form our final data set. As in most bankruptcy prediction modeling articles, we resort, in this paper, to the legal definition of bankruptcy (i.e. the recognition by the judge with regard to the criteria used in the law).

Because we aim to predict bankruptcy, we consider financial ratios of both types of firms (i.e., bankrupt and non-bankrupt) one year prior to bankruptcy (i.e., 2016). Table 1 reports the distribution of the sample between bankrupt and non-bankrupt firms by activity sector.

Table 1. Initial sample

Type	**Number of Firms**	**Proportion**
Non-bankrupt	2,113	93.2%
Bankrupt	153	6.8%
Total	2,266	100%
Activity Sector	**Number of Non-Bankrupt Firms**	**Number of Bankrupt Firms**
Agriculture, forestry, and fishing	36	1
Mining and quarrying	3	0
Manufacturing	136	12
Electricity, gas, steam, and air conditioning supply	2	0
Water supply: sewerage, waste	5	0

management, and remediation activities		
Construction	293	27
Wholesale and retail trade: repair of motor vehicles and motorcycles	464	32
Transportation and storage	66	10
Accommodation and food service activities	77	7
Information and communication	81	5
Financial and insurance activities	118	9
Real estate activities	157	10
Professional, scientific, and technical activities	359	20
Administrative and support service activities	84	8
Education	18	3
Human health and social work activities	148	7
Arts, entertainment, and recreation	22	0
Other service activities	42	2
Activities of households as employers	2	0
Total	2,113	153

The data distribution represents imbalanced information, because 93% of the companies are non-bankrupt, and only 7% are bankrupt. This data set mimics real-world conditions, where it is common to find more non-bankrupt firms than bankrupt firms. For this condition, bankruptcy prediction is known as a rare event (Calabrese & Osmetti, 2015).

2.2. Variables

The dichotomous, dependent variable is bankruptcy. It equals 1 when a company is bankrupt and 0 otherwise.

In accordance with Ben (2017) and Correa-Mejía et al. (2021), we consider liquidity, profitability, and debt ratios as the independent variables to predict bankruptcy. Free cash flow and current ratio can measure liquidity; according to Foerster et al. (2017) and Correa-García & Correa-Mejía (2021), they make it possible to evaluate the amount of cash that companies earn and their financial capacity to pay

short-term debts. In accordance with Nyitrai and Virág (2018), we use four measurements of profitability (i.e., earnings before interest, taxes, depreciation, and amortization [Ebitda], return on assets [ROA], return on equity [ROE], and net added value). According to Zhou and Lai (2017), these variables signal the efficiency of companies, so they can support predictions of future cash flows. Finally, debt concentration, debt level, and financial leverage can determine the portion of the resources committed to creditors (Zhou & Lai, 2017). In this context, these critical financial measurements might forecast bankruptcy. In addition, we use the minimum number of variables to build a functional predictive model. Table 2 shows the nine applied ratios in the forecast process.

Table 2. Financial measurements

Category	Variable	Calculation
Liquidity	Free cash flow	Net cash from operating activities + Capex
Liquidity	Current ratio	Current assets / Current liabilities
Profitability	Ebitda	Earnings before interest, taxes, depreciation, + amortization
Profitability	ROA	Net profit/Assets
Profitability	ROE	Net profit/Equity
Profitability	Net added value	Operating income – Purchases – Services and other goods
Debt	Debt concentration	Current liabilities/Total liabilities
Debt	Debt level	Total liabilities/Total assets
Debt	Financial leverage	Financial liabilities / Equity

One of the challenges in bankruptcy studies is variance stability; financial information presents different distributions, outliers, and asymmetry (Jones, Johnstone, & Wilson, 2017). These data characteristics affect bankruptcy predictions. Therefore, we

applied a data transformation method proposed by Yeo and Johnson (2000). The Yeo-Johnson transformation makes it possible to work with negative or zero values for the variables. This transformation can be represented as follows:

$$\psi(\lambda, y) = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda} if\ \lambda \neq 0, y \geq 0 \\ \log(y+1)\ \ if\ \lambda \neq 0, y \geq 0 \\ \dfrac{-[(-y+1)^{2-\lambda}-1)]}{2-\lambda}\ if\ \lambda \neq 2, y < 0 \\ -\log(-y+1)\ if\ \lambda = 2, y < 0 \end{cases},$$

where ψ (λ, y) is the Yeo-Johnson transformation, y is a list of numbers without restrictions, and λ is the parameter of the transformation.

2.3. Models

To address these predictions, we tested two different models: logistic regression and a boosting algorithm. Furthermore, we apply an over-sampling technique to increase the prediction accuracy of the models.

2.3.1. Logistic regression

This regression can be applied when the dependent variable only takes one of two values that are mutually exclusive (Pérez, Lopera, & Vásquez, 2017). Because the result of the logistic regression is either of two values, $Y \in \{-1,1\}$, we can calculate the probability that a certain event will occur according to the result of the independent variables, as follows:

$$P_i = \frac{e^z}{1+e^z}\ [1],$$

where P_i represents the likelihood that a specific firm enters bankruptcy, and z comprises the independent variables.

According to Calabrese and Osmetti (2013), there is a basic problem in the use of this regression to predict bankruptcy. Because bankruptcy is a rare event, with more non-failed than failed companies, the estimation reached through this model might

underestimate the probability of bankruptcy. Some authors, such Zhou and Lai (2017) and Le et al. (2018), have suggested resampling techniques to overcome this problem.

2.3.2. Boosting

A boosting algorithm combines different classifiers to produce a committee (Hastie, Tibshirani, & Friedman, 2008). According to Ridgeway (1999), this method is mainly used to solve classification problems, such as those that affect bankruptcy prediction. In this study, we use the algorithm AdaBoost.M1, proposed by Freund and Schapire (1997) specifically for binary problems, to predict bankruptcy. This model assigns the same weight, $1 / N$, to a data set, and the dependent variable can be one of two possible values $Y \in \{-1,1\}$. To conform the committee, the algorithm generates several iterations $m = 1,2,3, \dots, M$. In each iteration, the weights of all observations are modified according to their classification accuracy (Hastie, Tibshirani, & Friedman, 2008). In iteration m, the weights decrease for observations that are classified properly, but the weights increase for those that are misclassified. Table 3 depicts the steps proposed by Hastie, Tibshirani, and Friedman (2008) to develop the algorithm.

Table 3. Boosting algorithm

Initialize the observation weights $w_i = 1 / N$; $i = 1,2, \dots, N$.
For $m = 1$ to M:
Fit a classifier $Gm(x)$ to the training data using weights w_i.
Compute $err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$ [2]
Compute $\alpha_m = \frac{\log(1 - err_m)}{err_m}$
Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1,2, \dots, N.$
Output

$$G(x) = sign[\sum_{m=1}^{M} \alpha_m G_m(x)] \text{ [3]}$$

2.3.3. SMOTE algorithm

The SMOTE algorithm was proposed by Chawla et al. (2002) to create synthetic observations leading to a balanced data set (García, Marqués, & Sánchez, 2019). According to Kim, Kang, and Bae (2015), the algorithm generates a new sample by identifying specific observations with a K similar minority class. The new observations are calculated as follows:

$$X_{new} = X + rand\ (0,1) * (X_n - X) \text{ [4]},$$

where X_{new} is the new observation, X is the original data, and X_n is one of the K nearest neighbors to the original observation. Table 4 presents the steps to develop this algorithm, according to Chawla et al. (2002).

Table 4. SMOTE algorithm

Choose the K nearest neighbors to the original observations.

Measure the distance of the original observation and K samples as $(X_n - X)$

Multiply the distance $(X_n - X)$ by a random number between 0 and 1.

Add the multiplied distances to the original sample.

Output:

$$X_{new} = X + rand\ (0,1) * (X_n - X)$$

3. – Results

We consider the financial ratios primarily to identify whether the variables lead to different results for failed and non-failed companies. The tendencies in both groups of firms were analyzed according to differences in means. Table 5 contains the results of this analysis.

Table 5. Descriptive statistics

Variable	Mean Non-bankrupts	Sd Non-bank-rupts	Mean Bank-rupts	Sd Bank-rupts	Test of Equal Means
FCF	569.30	6291.09	87.4	404.08	3.2**
Current ratio	2.59	5.06	5.42	12.35	–3.48***
Ebitda	674.52	7611.96	78.64	252.39	3.51***
Net added value	1532.39	14390.42	252.99	506.57	3.12**
ROE	0.19	0.58	0.09	0.47	2.58*
ROA	0.05	0.11	0.02	0.17	1.66
Financial leverage	1.38	3.54	1.24	4.09	3.54***
Debt level	0.59	0.24	0.55	0.28	1.11
Debt concentration	0.51	0.36	0.7	0.32	–7.17***

Significance: *** < .001, ** p < .01, * p < .05, p < .1.

Regarding liquidity, non-bankrupt firms have higher free cash flows (t-statistic = 3.2, p < .01) than bankrupt firms, whereas bankrupt firms indicate a higher current ratio (t-statistic = –3.48, p < .001). In addition, non-failed companies achieve better profitability measurements than failed companies (Ebitda t-statistic = 3.51, p < .001; Net added value t-statistic = 3.12, p < .01; ROE t-statistic = 2.58, p < .05; ROA t-statistic = 1.66, p < .1). Finally, there is no clear separation between bankrupt and non-bankrupt companies according to the debt ratios; it is not possible to define the effect of the debt through differences in means. However, we still include this variable in the prediction process, because Son et al. (2019) demonstrate that companies with high debt levels are at a higher risk of bankruptcy.

We compare the boosting algorithm with logistic regression to establish which model more accurately predicts bankruptcy. Table 6 contains the confusion matrix, which demonstrates the accuracy of both models to classify non-failed and failed firms. During the prediction process, data were randomly divided into two groups to train and test the model. A random sample composed of 90% of the total data was used to train the model. Thus, 1,902 non-bankrupt and

138 bankrupt companies composed the sample that trained the model, and 211 non-bankrupt and 15 bankrupt firms were used to test the model (i.e., 10% of the total sample; Li and Faff, 2019; Le et al., 2019). We chose 90% of observations to train the model because there were fewer bankrupt firms, and it was necessary to have a large number of companies to train the model, to obtain more accurate results from the test sample (Li & Faff, 2019).

Table 6. Confusion matrix

	Boosting Algorithm		**Logistic Regression**	
Classification	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt
Bankrupt	73.3%	12.3%	0%	0%
Non-bankrupt	26.7%	87.7%	100%	100%
Total	**100%**	**100%**	**100%**	**100%**
Error type I	26.7%		100%	
Error type II	12.3%		0%	
Global accuracy rate	86.7%		93.2%	

The confusion matrix reveals that 73.3% of bankrupt and 87.7% of non-bankrupt companies are classified correctly using the boosting algorithm. In comparison, the logistic regression incorrectly predicts bankrupt companies, though all non-bankrupt firms are classified properly.

Type I error refers to the probability of classifying a bankrupt company incorrectly (Liang et al., 2016, Correa-Mejía & Lopera-Castaño, 2019). It directly affects the financial performance of companies, because it is not possible to recover the products or services sold or the cash flow from the sale. In contrast, a type II error classifies a non-bankrupt firm as bankrupt (Bauer & Agarwal, 2014). It also has a negative effect on the financial performance of companies, because their profits are reduced by an incorrect rejection of a customer that is not a credit risk. According to Liang et al. (2016) though, type I error is more critical, because it implies reductions in both cash flow and profit. The boosting algorithm thus achieves better prediction performance, in that it is less probable that companies incur

type I error using this method, compared with the logistic regression. The accuracy of both methods is demonstrated by the ROC curves in Figure 1.

Figure 1. ROC curves

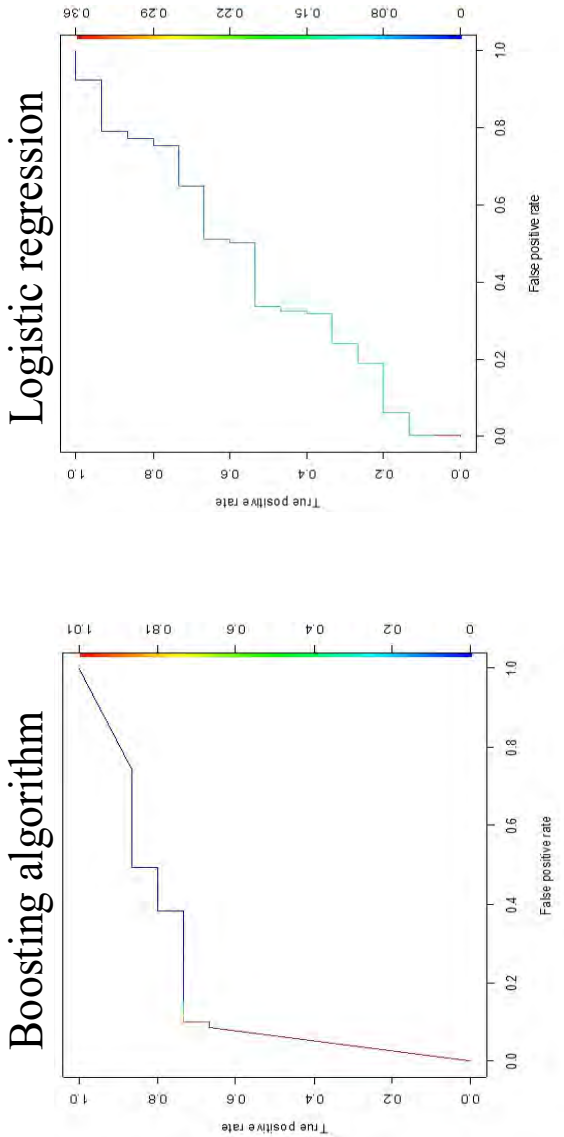


Figure 1. ROC curves

The ROC curve is a graphic technique that enables visual analysis of the accuracy of a test (Kovacova & Kliestik, 2017). The ROC curves in Figure 1 represent the prediction accuracy of both models. The boosting algorithm creates a greater AUC (= .79) than logistic regression (AUC = .57), indicating that it is possible to reach better prediction results for bankrupt and non-bankrupt firms using the boosting algorithm.

To verify the consistency in prediction accuracy of both models, we tested models with different proportions to reduce type 1 errors and evaluate its sensibility. That is, we recomposed the sample using the SMOTE algorithm with different proportions. As proposed by Kim, Kang, and Bae (2015), five groups were created, according to different balance rates (i.e., 1:1, 1:3, 1:5, 1:10, and 1:20), so that we could analyze the classification accuracy of both models at different imbalance levels. In this process, the same variables were used to predict bankruptcy. Table 7 shows the configuration of each group.

The SMOTE algorithm requires a given quantity of observations to estimate new observations for the minority class (i.e., over-sampling) and a given quantity of observations to estimate new observations for the majority class (i.e., under-sampling). Each group in Table 7 shows a different number of companies, because the quantities of the majority class (i.e., under-sampling) and the minority class (i.e., over-sampling) depend on the data set size and the class proportions (Chawla et al., 2002). In this context, we work with different numbers of firms to estimate both models for each proportion.

Table 7. Imbalanced data samples

Set	Training			Test		
	Bankrupt	Non-bankrupt	Total	Bankrupt	Non-bankrupt	Total
1:1	964	964	1,928	107	107	214
1:3	551	1,652	2,203	61	184	245
1:5	413	2,066	2,479	46	230	275
1:10	275	2,754	3,029	31	306	337
1:20	275	2,808	3,083	31	612	643

Both models were developed using the five new sample sets, and the confusion matrixes are in Table 8. The predictive power for bankrupt companies employing both models decreases as the imbalance increases. The results of the confusion matrixes are based on the test samples used in the prediction; we simulated each model 1,000 times using a random sample in each simulation. The results in Table 8 are the most accurate for each simulation. It was possible to create different imbalanced data sets because the SMOTE algorithm can generate synthetic observations using the distribution of initial data.

Table 8. Confusion matrixes using different imbalanced proportions

Boosting Algorithm												
	Initial Sample		1:1		1:3		1:5		1:10		1:20	
Classifica-tion	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt
Bankrupt	73.3	12.3	95.3	6.5	91.8	9.2	80.0	5.2	70	.7	48.4	1.8
Non-bankrupt	26.7	87.7	4.7	93.5	8.2	90.8	20.0	94.8	30	99.3	51.6	98.2
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Error Type I	26.7		4.7		8.2		20		30		51.6	
Error Type II	12.3		6.5		9.2		5.2		.7		1.8	
Global Accuracy	86.7		94.4		91		92.3		96.7		95.8	

Logistic Regression												
	Initial Sample		1:1		1:3		1:5		1:10		1:20	
Classification	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt
Bankrupt	0	0	85	25.2	26.2	6.5	17.8	.4	10	1.3	0	0
Non-bankrupt	100	100	15	74.8	73.8	93.5	82.2	99.6	90	98.7	100	100
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Error Type I	100		15		73.8		82.2		90		100	
Error Type II	0		25.2		6.5		.4		1.3		0	

Global Accuracy	93.2	79.9	76.7	85.9	90.6	95.2

Results shown in table 8 are percentages.

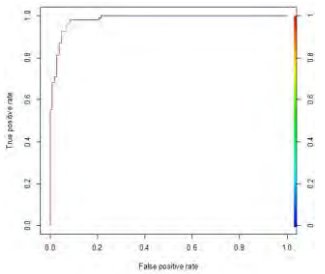
Although the prediction accuracy of both models decreases as asymmetry increases, the boosting algorithm offers better prediction results for bankrupt firms. The boosting algorithm reflects better global accuracy rates in all proportions of the sample. It is important to highlight that type I error increases when the imbalance in the data set is higher. In the case of the boosting algorithm, type I error is 4.7 when the data set is symmetric. However, when the data set is 1:20, the error rate is 51.6. In comparison, logistic regression has a type I error rate of 15 when the data set is symmetric and 100 when the data set is 1:20. Therefore, there is no way to classify a bankrupt company correctly using logistic regression when the data set is imbalanced at 1:20.

Type I error increases when the imbalance in the data set is greater. Therefore, it is more probable to incur this kind of error when the data set has fewer bankrupt firms. Nevertheless, through the boosting algorithm, it is possible to reduce the probability of both type I and type II errors, compared with logistic regression. The ROC curves, used to evaluate the accuracy of both models at the proposed imbalance rates, are in Figure 2.

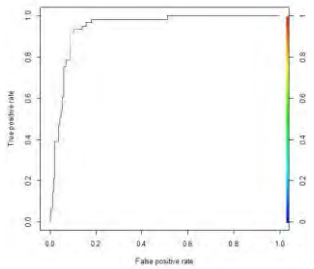
Figure 2. ROC curves

2.a. Boosting algorithm

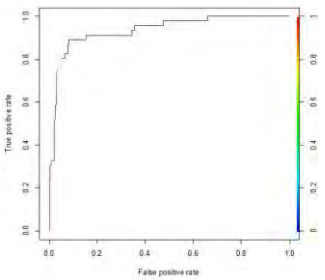
1:1



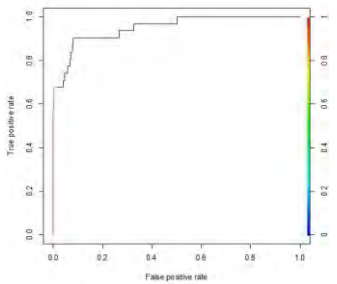
1:3



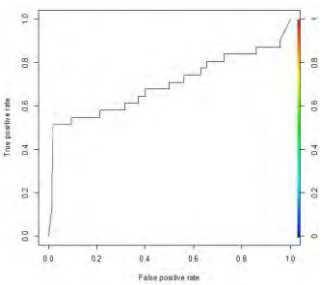
1:5



1:10

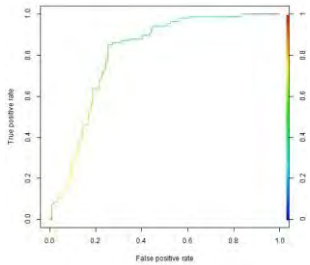


1:20

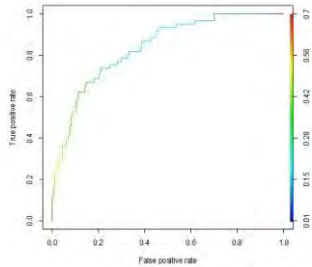


2.b. Logistic regression

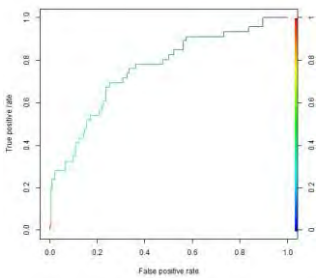
1:1



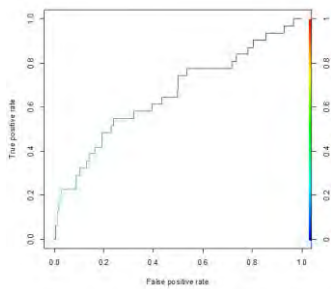
1:3



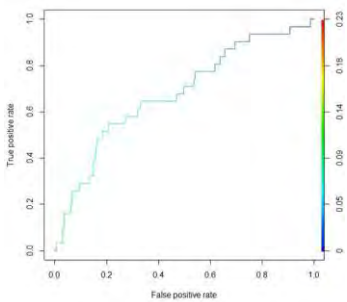
1:5



1:10



1:20



The AUC tends to decrease as the data imbalance increases. When the ROC curve has an inclination close to 45°, the model does not predict accurate bankruptcy, because the AUC is minimal (50). The ROC curve tends toward 45° when the data set is more imbalanced. Although AUC decreases when the imbalance data set increases, the boosting algorithm achieves a superior AUC than logistic regression. Thus, the average accuracy of the prediction increases when the boosting algorithm is employed, as the AUC results in Table 9 confirm.

Table 9. AUC results

Set	Boosting Algorithm	Logistic Regression
1:1	0.98	0.81
1:3	0.94	0.84
1:5	0.94	0.76
1:10	0.95	0.66
1:20	0.69	0.68

These AUC results confirm that the boosting algorithm has better accuracy than logistic regression in bankruptcy prediction. Despite the data imbalance, the boosting algorithm always has a superior AUC, which means that boosting classifies true positives and true negatives with a high accuracy rate. These results are consistent with Table 8, in which the global accuracy rates were lower using logistic regression.

When the boosting algorithm and logistic regression are compared using the original data set and different imbalance rates, the boosting algorithm is better at predicting performance. With the boosting algorithm, type I error is less frequent, and the ROC curve increases, which means that the AUC is higher and the accuracy of the prediction increases.

4. – Discussion and conclusion

Few studies (e.g., Wilson & Sharda, 1994; McKee & Greenstein, 2000) have built prediction models using imbalanced data sets (i.e., closer to real-world conditions). This research proposes a

bankruptcy prediction model that considers the issues created by imbalanced data sets. Although bankruptcy is rare in the real world, most bankruptcy prediction models use balanced samples of firms (i.e., 50 bankrupt firms and 50 non-bankrupt firms). Kang and Cho (2006) note that resampling the data or assigning different weights (i.e., penalties) to observations are two methods traditionally used to resolve the imbalanced data set issue, so we adopt boosting and over-sampling techniques to overcome the imbalanced data set issue. Our results show that the boosting algorithm is an appropriate model to forecast bankruptcy. Through the estimation of type I and type II errors, as well as an analysis of the ROC curves and AUC values, we find that there is a lower probability of type I errors and better performance in the boosting algorithm predictions compared with logistic regression. Moreover, balanced samples are preferable, because they prevent the model from specializing in the classification of the most represented group, leaving the minority category mainly misclassified. The use of the SMOTE technique creates balanced distributions and artificially increases the sample size, which decreases type I errors in both logistic and boosting models.

The best model uses the boosting algorithm on a balanced sample created with a SMOTE over-sampling technique. This model reports a global accuracy of 94.4 and a type I error rate of less than 5. Type I error is the probability of misclassifying a bankrupt firm as non-bankrupt, resulting in the total or partial loss of credits granted by creditors. This rate of 4.7 is much lower than the type I error rate reported for the logit model based on a balanced sample of 1:1 built using the SMOTE over-sampling technique (15) or that reported using the boosting model on an unbalanced sample of 1:20 (52).

Our results align with findings from Zhou (2013), Kim and Ahn (2015), and Veganzones and Sévérin (2018). The models achieve greater accuracy following resampling. The results also align with Kim, Kang, & Bae (2015) and du Jardin, Veganzones, and Séverin (2017), confirming that the boosting technique is suitable for bankruptcy prediction modeling on real-world conditions.

The experimental results also demonstrate that the boosting algorithm has an advantage over logistic regression for predicting bankruptcy in imbalanced and balanced data distributions. Similar to Kim, Kang, & Bae (2015), we find that the more imbalanced the data distribution, the less accurate the model will be. However, the results using the boosting technique indicate that it is an effective tool to

assess bankruptcy risk in real-world conditions (Kim, Kang, & Bae, 2015). Therefore, the decision-making process related to granting credits will be more precise using this algorithm.

This study offers important information for investors, suppliers, bankers, and governments. With the proposed model (i.e., boosting technique applied to a balanced database created through the SMOTE algorithm), organizations can reduce type I errors and avoid entering into corporate contracts with risky customers.

Nevertheless, certain limitations of this study should be mentioned. First, we rely on a database of firms from only one country, Belgium, which has specific characteristics that might influence the results. Second, the absence of comprehensive accounting information for all companies in the original database (Section 3.1) limited our ability to consider some firms' information and reduced the sample in the study from 7,814 to 2,266 companies. Third, though most bankruptcy prediction models are based solely on financial variables, the inclusion of non-financial variables into models can improve their accuracy (e.g., Ciampi, 2015; Tobback et al., 2017). Further research thus might consider CEO characteristics, corporate governance policies, and management styles, for example.

REFERENCES

ALTMAN, E. I., «Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy», *The Journal of Finance*, vol. 23, n° 4, 1968, p. 589-609.

ANDERSON, R., *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press, 2007.

BAUER, J., & AGARWAL, V., «Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test», *Journal of Banking and Finance,* vol. 40, n° 1, 2014, p. 432–442.

BEAVER, W. H., «Financial Ratios as Predictors of Failure», *Journal of Accounting Research*, vol. 4, n° 71, 1966, p. 71-111.

BEN, S., «Bankruptcy prediction using Partial Least Squares Logistic Regression», *Journal of Retailing and Consumer Services*, vol. 36, 2017, p. 197–202.

CALABRESE, R., & OSMETTI, S. A., «Improving forecast of binary rare events data: A gam-based approach», *Journal of Forecasting*, vol. 34, n° 3, 2015, p. 230–239.

CALABRESE, R., & OSMETTI, S. A., «Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model», *Journal of Applied Statistics,* vol. 40, n° 6, 2013, p. 1172–1188.

CHAWLA, N. V, BOWYER, K. W., HALL, L. O., & Kegelmeyer, W. P., «SMOTE: Synthetic Minority Over-sampling Technique», *Journal of Artificial Intelligence Research,* vol. 16, 2002, p. 321–357.

CIAMPI, F., «Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms», *Journal of Business Research*, vol. 68, n° 5, 2015, p. 1012–1025.

CORREA-GARCÍA, J. A., & CORREA-MEJÍA, D. A., «Importancia del estado de flujos de efectivo para la gestión financiera sostenible», *Cuadernos de Contabilidad*, 22, 2021, p. 1–32.

CORREA-MEJÍA, D. A., & LOPERA-CASTAÑO, M., «Pronóstico de la insolvencia empresarial en Colombia a través de indicadores financieros», *Panorama Económico*, vol. 27, n° 2, 2019, p. 510–526.

CORREA-MEJÍA, D. A., MURILLO-PALACIOS, M. C., & VÉLEZ CARDONA, N., «Los indicadores financieros: Herramienta para evaluar el principio de negocio en marcha», *Desarrollo Gerencial*, vol. 13, n° 2, 2021, p. 1–24.

DAILY, C. M., & DALTON, D. R., «Corporate governance and the bankrupt firm: An empirical assessment», *Strategic Management Journal*, vol. 15, n° 8, 1994, p. 643–654.

DU JARDIN, P., VEGANZONES, D., & SÉVERIN, E., «Forecasting Corporate Bankruptcy Using Accrual-Based Models», *Computational Economics*, vol. 54, n° 1*,* 2017, p. 7-43.

ESTABROOKS, A., JO, T., & JAPKOWICZ, N., «A Multiple Resampling Method for Learning from Imbalanced Data Sets», *Computational Intelligence*, vol. 20, n° 1, 2004, p. 18–36.

FOERSTER, S., TSAGARELIS, J., & WANG, G., «Are Cash Flows Better Stock Return Predictors Than Profits?» *Financial Analysts Journal*, vol. 73, n° 1, 2017, p. 73–99.

FREUND, Y., & SCHAPIRE, R. E., «A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting», *Journal of Computer and System Sciences,* vol. 55, n° 1, 1997, p. 119–139.

GARCÍA, V., MARQUÉS, A. I., & SÁNCHEZ, J. S., «Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction», *Information Fusion,* vol. 47, 2019, p. 88–101.

HE, H., & GARCIA, E. A., «Learning from Imbalanced Data», *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, n° 9, 2009, p. 1263–1284.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. California, Estados Unidos. Springer, 2008.

JAPKOWICZ, N., & STEPHEN, S., «The class imbalance problem: A systematic study», *Intelligent Data Analysis,* vol. 6, n° 5, 2002, p. 429–449.

JONES, S., JOHNSTONE, D., & WILSON, R., «Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks», *Journal of Business Finance and Accounting,* vol. 44, n° 1, 2017, p. 3–34.

KANG, P., & CHO, S. *EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems*. In I. King, J. Wang, L.-W. Chan, & D. Wang (Eds.), «Neural Information Processing» (Vol. 4232, pp. 837–846), 2006.

KIM, M., KANG, D., & BAE, H., «Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction», *Expert Systems with Applications*, vol. 42, n° 3, 2015, p. 1074–1082.

KIM, M.-J., & KANG, D.-K., «Ensemble with neural networks for bankruptcy prediction», *Expert Systems with Applications*, vol. 37, n° 4, 2010, p. 3373–3379.

KIM, T., & AHN, H., «A Hybrid Under-sampling Approach for Better Bankruptcy Prediction», *Journal of Intelligence and Information Systems*, vol. 21, n° 2, 2015, p. 173–190.

KOVACOVA, M., & KLIESTIK, T., «Logit and Probit application for the prediction of bankruptcy in Slovak companies», *Equilibrium-Quarterly Journal of Economics and Economic Policy*, vol. 12, n° 4, 2017, p. 775–791.

LE, T., SON, L. H., VO, M. T., LEE, M. Y., & BAIK, S. W., «A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset», *Symmetry,* vol. 10, n° 7, 2018, p. 1–12.

LE, T., VO, B., FUJITA, H., NGUYEN, N., & WOOK, S., «A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting», *Information Sciences*, vol. 494, 2019, p. 294–310.

LI, L., & FAFF, R., «Predicting corporate bankruptcy : What matters? » *International Review of Economics and Finance*, vol. 62, 2019, p. 1–19.

LIANG, D., LU, C. C., TSAI, C. F., & SHIH, G. A., «Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study», *European Journal of Operational Research,* vol. 252, n° 2, 2016, p. 561–572.

LÓPEZ, V., FERNÁNDEZ, A., GARCÍA, S., PALADE, V., & HERRERA, F., «An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics», *Information Sciences*, vol. 250, 2013, p. 113–141.

MCKEE, T. E., & GREENSTEIN, M., «Predicting bankruptcy using recursive partitioning and a realistically proportioned data set», *Journal of Forecasting*, vol. 19, n° 3, 2000, p. 219–230.

NYITRAI, T., & VIRÁG, M., «The effects of handling outliers on the performance of bankruptcy prediction models», *Socio-Economic Planning Sciences*, vol. 67, 2018, p. 1–9.

ODOM, M. D., & SHARDA, R. *A neural network model for bankruptcy prediction*. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 163–168 vol.2). San Diego, CA, USA: IEEE, 1990.

OHLSON, J. A., «Financial Ratios and the Probabilistic Prediction of Bankruptcy», *Journal of Accounting Research*, vol. 18, n° 1, 1980, p. 109-131.

PÉREZ, J., LOPERA, M., & VÁSQUEZ, F., «Estimación de la probabilidad de riesgo de quiebra en las empresas colombianas a partir de un modelo para eventos raros", *Cuadernos de Administración*, vol. 30, n° 54, 2017, p. 7–38.

PIRI, S., DELEN, D., & LIU, T., «A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets», *Decision Support Systems*, vol. 106, 2018, p. 15–29.

RIDGEWAY, G., «The state of boosting», *Computing Science and Statistics,* vol. 31, 1999, p. 172–181.

SÁEZ, J. A., LUENGO, J., STEFANOWSKI, J., & HERRERA, F., «SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering», *Information Sciences*, vol. 291, 2015, p. 184–203.

SCHAPIRE, R. E., «The strength of weak learnability», *Machine Learning*, vol. 5, n° 2, 1990, p. 197–227.

SEIFFERT, C., KHOSHGOFTAAR, T. M., Van HULSE, J., & NAPOLITANO, A., «RUSBoost: Improving classification performance when training data is skewed», 2008 19th International Conference on Pattern Recognition, 2008, p. 1–4.

SERRANO-CINCA, C., GUTIÉRREZ-NIETO, B., & BERNATE-VALBUENA, M., «The use of accounting anomalies indicators to predict business failure», *European Management Journal*, vol. 37, n° 3, 2019, p. 353-375.

SON, H., HYUN, C., PHAN, D., & HWANG, H. J., «Data analytic approach for bankruptcy prediction», *Expert Systems with Applications*, vol. 138, 2019, p. 790–784.

TOBBACK, E., Bellotti, T., MOEYERSOMS, J., STANKOVA, M., & MARTENS, D., «Bankruptcy prediction for SMEs using relational data», *Decision Support Systems*, vol. 102, 2017, p. 69–81.

VEGANZONES, D., & SÉVERIN, E., «An investigation of bankruptcy prediction in imbalanced datasets», *Decision Support Systems*, vol. 112, 2018, p. 111–124.

WILSON, R. L., & SHARDA, R., «Bankruptcy prediction using neural networks», *Decision Support Systems*, vol. 11, n° 5, 1994, p. 545–557.

YEO, I.-K., & JOHNSON, R. A., «A new family of power transformations to improve normality or symmetry", *Biometrika,* vol. 87, n° 4, 2000, p. 954–959.

ZHOU, L., «Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods», *Knowledge-Based Systems*, vol. 41, 2013, p. 16–25.

ZHOU, L., & LAI, K. K., «AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data», *Computational Economics,* vol. 50, n° 1, 2017, p. 69–94.

ZMIJEWSKI, M. E., «Methodological Issues Related to the Estimation of Financial Distress Prediction Models», *Journal of Accounting Research*, vol. 22, 1984, p. 59-82.