ANALES de Edificación

Received: 20-03-2021
Accepted: 01-04-2021

Anales de Edificación
Vol. 7, Nº1, 9-18 (2021)
ISSN: 2444-1309
Doi: 10.20868/ade.2021.4766

Algoritmos de Random Forest como alerta temprana para la predicción de insolvencias en empresas constructoras.
Random Forest algorithms as early warning tools for the prediction of insolvencies in construction companies.

José Ignacio Sordo Sierpe[a], Mercedes del Río Merino[b], Álvaro Pérez Raposo[c], Verónica Vitiello[d]
(ignacio.sordo@alumnos.upm.es, ignacio.sordo@urtinsa.com; mercedes.delrio@upm.es; alvaro.p.raposo@upm.es; veronica.vitiello@unina.it).

[a] PhD, Universidad Politécnica de Madrid. CFO Grupo Arpada. Madrid, Spain.
[b] Professor. Departamento de Construcciones Arquitectónicas y su Control. ETSEM- Universidad Politécnica de Madrid. Madrid, Spain.
[c] Full Professor. Departamento de matemática aplicada. ETSEM- Universidad Politécnica de Madrid. Madrid, Spain.
[d] Universidad Federico II. Naples, Italy

***Resumen*— La preocupación de la Unión Europea por evitar que las empresas lleguen a un procedimiento de insolvencia motivó la promulgación de la Directiva (UE) 2019/1023 del Parlamento Europeo y del Consejo, y su transposición obligatoria a las regulaciones de los Estados miembros antes del 17 de julio de 2021. Esta Directiva establece que los deudores deben tener acceso a herramientas de alerta temprana para detectar situaciones de insolvencia inminente. Esta investigación tiene como objetivo contribuir al desarrollo de este tipo de herramientas de alerta temprana para un sector muy específico: la construcción residencial y no residencial. La metodología se ha dividido en dos fases, cada una con su propio objetivo específico: (1) seleccionar las variables predictoras que mejor puedan explicar el modelo (para ello se han utilizado técnicas estadísticas tradicionales); y (2) seleccionar los algoritmos que proporcionen la mayor precisión para el modelo de herramienta de alerta temprana entre cinco algoritmos Random Forest. El objetivo principal de esto es obtener señales de alerta con la suficiente antelación para poder detectar situaciones de insolvencia. El objetivo fundamental es lograr un modelo sin utilizar las cuentas de pérdidas y ganancias de las constructoras investigadas. Esto es así para evitar la falta de objetividad que pueden tener los ingresos y, por tanto, los resultados contables en este sector. Se obtuvieron porcentajes de precisión superiores al 85% tres años antes de que ocurriera la insolvencia utilizando únicamente ratios de balance. El principal valor es poder aplicar la herramienta de alerta temprana de forma sencilla, utilizando pequeñas cantidades de datos, especialmente para el deudor, que puede reaccionar con la suficiente antelación para evitar una situación financiera potencialmente irreversible.**

Palabras Clave— Advertencia temprana, Random Forest, construcción.

***Abstract*— The European Union's concern with preventing companies from reaching insolvency proceedings motivated the enactment of Directive (EU) 2019/1023 of the European Parliament and of the Council, and its mandatory transposition into Member States' regulations by July 17, 2021. This Directive states that debtors must have access to early warning tools to detect situations of**

imminent insolvency. This research aims to contribute to the development of such early warning tools for a very specific sector: residential and non-residential construction. The methodology has been divided into two phases, each with its own specific objective: (1) to select the predictor variables that can best explain the model (traditional statistical techniques have been used for this purpose); and (2) to select the algorithms that provide the greatest precision for the early warning tool model from among five Random Forest algorithms. The main objective of this is to obtain warning signs sufficiently enough in advance that insolvency situations can be detected. The fundamental aim is to achieve a model without using the profit and loss accounts from the construction companies under investigation. This is so to avoid the lack of objectivity that income, and therefore accounting results, may have in this sector. Accuracy percentages of over 85% were obtained three years before insolvency occurred using only balance sheet ratios. The main value is to be able to apply the early warning tool in a simple way, using little amounts of data, especially for the debtor, who can react early enough to avoid a potentially irreversible financial situation.

Index Terms— Early warning, Random Forest, construction.

I. Introduction

The construction industry is the world's largest industry and one of the most dynamic in the global economy. Its importance is due to its extraordinary contribution to the distribution of wealth, to the well-being of society, and the large number of workers it employs.

The construction sector in Spain contributes 6.5% of the GDP (Spanish National Classification of Economic Activities CNAE 41) with the sector bringing in a total of 1,202 million euros and 1,277,900 directly related jobs, according to the latest publication of the National Statistics Institute.

Since 2008, 63.2% of companies in the sector have been created, with 50% of them having been created since 2012 (Fundación Laboral de la Construcción), which is evidence of the high mortality rate of construction companies.

From 2008 to the end of 2019 (provisional data), 25.18% of the companies that have filed for insolvency proceedings in Spain belonged to the construction sector. This is a statistic that is far above other sectors such as commerce (5.75 percentage points), industry and energy (8.23 percentage points), and hotel and catering (20.62 percentage points).

The European Parliament Directive adopted in June 2019 notes that Member States must ensure "that debtors have access to one or more clear and transparent early warning tools which can detect circumstances that could give rise to a likelihood of insolvency and can signal to them the need to act without delay".

The Directive also states that these early warning tools may include: alert mechanisms in case the debtor has failed to make certain types of payments; advisory services provided by public or private organizations; and incentives for third parties that have relevant information about the debtor (tax and social security administrations, etc.). These mechanisms are used to warn the debtor of any negative developments.

The Directive re-emphasizes that Member States shall ensure that such early warning tools are publicly available online, that they are easily accessible, and that they are presented in a user-friendly format.

This Directive, whose transposition was scheduled for 17 July, has not yet been transposed in its entirety in the different Member States, as many EU countries (including Spain) have made use of Article 34.2 of the Directive and have requested a one-year extension for its application.

A. State of the Art

Insolvency tests should be one of the first tools that any early warning system should incorporate since, once tax or social security defaults occur, it is more than likely that the desired objective has been reached too late.

Predictive mechanisms for corporate insolvency emerged in the 1930s.

A chronological analysis of the different predictive models shows that there was a first stage, known as the descriptive stage, in which there were very few studies in the period from 1930 to 1966.

There was also a second stage, known as the predictive stage, which covers the period from 1966 to the present day. Within this predictive stage, we find two different techniques; statistical prediction models and artificial intelligence techniques. The latter of the two began to be developed at the beginning of the 1990s, in parallel with the advance of computer systems (although both models coexist today and the appearance of the latter has not made the former disappear).

Focusing on the second stage and reviewing these models at the international level, we first have Beaver's univariate models, which look to explain independent variables and their influence on solvency/insolvency separately (Beaver, 1966).

Two years later, Altman's Z-score model (1968) appeared, introducing multivariate analyses with the multiple discriminant analysis technique. This model is undoubtedly the most well-known model to date. Over the years, this model has been reformulated and adapted to other types of companies. Other authors who have developed models based on multiple discriminant analysis include Meyer and Pifer (1970), Deakin (1972), Edmister (1972), Blum (1974), Dambolena et al. (1980), Taffler (1982), Micha (1984), Laffarga et al. (1985), Gombola et al. (1987), and Laitinen (1992).

James Ohlson (1980) developed a business failure prediction model which, unlike Altman's model, used logistic

regression analysis. Four years later, Zmijewski (1984) used a probit model.

The early 1990s saw the beginning of models being based on computational techniques, using all kinds of machine learning algorithms.

In 1990, Odom and Sharda (Alaka, 2018) applied a Neural Network (NN) model for the first time and since then new models and techniques have not stopped appearing. Some examples include the Support Vector Machine (SVM), used by Shin et al. (2005) and Ming and Lee (2005) to predict insolvency in Korean companies; the Decision Tree (DT), used by Cho et al. (2010); Case-Based Reasoning (CBR), used by Jegon et al. (2012); and Genetic Algorithms (GA) used by Divsalar et al. (2011). All these models demonstrate the continual interest in the use of computational tools for insolvency prediction.

In the last five years, the use of different Machine Learning algorithms has grown exponentially around the world. To cite just a few works from this year and the past by country: in Italy (Perboli & Arabnezhad, 2021) Random Forest, Gradient Boosting, Logistic Regression, and Neuronal Networks have been used; in Spain, Neural Network algorithms have been applied to the restaurant sector (Becerra-Vicario, Alaminos, Aranda, & M., 2020); in Turkey, models such as the Decision Tree, Random Forest, AdaBoost, and others have been used (Tabbakh, Kumar, & Janjhi, 2021); in India, Random Forest, Logistic Regression, and SVM algorithms have been used (Arora, 2020); and in Taiwan, SVM, Naive Bayes, K-NN, Random Forest, and other models have been implemented (Wang & Liu, 2021).

B. Revenue recognition in construction companies

The recent modification of the regulations developed by RD1/2021 of 12 January modifying the 'Plan General de Contabilidad' (Spanish General Accounting Plan), and specifically the resolution of 10 February 2021, issued by the 'Instituto de Contabilidad y Auditoría de Cuentas' (Spanish Accounting and Auditing Institute), laying down rules for the recording, valuation, and preparation of the annual accounts for the recognition of income from the delivery of goods and the provision of services, states the following in its article 11:

1.- The objective of measuring the degree of progress is to represent the activity of the company in transferring the control of goods or services committed to the customer.

2. The company shall apply a single method for measuring progress and the same method for similar obligations and in similar circumstances.

3. The procedures for measuring progress include two types of methods:

(a) Output methods

In this method, revenue is recognized by directly measuring the value of goods or services transferred to the customer to date (e.g. certifications of work already completed), and is relative to the remaining goods or services.

As García Castellví (2005) notes, to arrive at the results for the year, in the case of advance certifications issued or work completed pending execution, income shall be given by the equation:

Income for the year = Certified work + Work executed pending certification - Work certified pending execution.

(b) Input methods

Under this method, revenue is recognized on the basis of the costs of production employed by the entity in relation to the total costs that the entity expects to incur in satisfying the obligation, excluding any of the factors of production that do not represent the activity undertaken to be able to transfer to the customer.

This would be the method known as the percentage of realization method and its calculation formula would be:

Percentage = Costs incurred/ (Costs incurred + Costs pending)
Revenue for the year = Total revenue foreseen in the contract x percentage

C. Cost structure of a construction project

It is important to know the cost structure of a construction project in order to better understand how the two methods of revenue recognition affect practical application. It is also important to know the cost structure as using early warning tools that include this magnitude, as well as the other magnitudes that are directly influenced, may result in a subjective model that alters the value of the objective set.

Construction projects have four large groups of costs, each of which accounts for practically a quarter of the execution budget, and which we can simplify as follows:

1.- Earthworks, foundations, and structure

2.- Masonry, roofing, waterproofing, and insulation

3.- Carpentry, Flooring, Miscellaneous

4.- Installations

The execution period of all these units has an average duration of 18 months, and the distribution of accumulated costs over time of a residential building of average quality without special foundations incurs 60% of the costs in the first 12 months and the remaining 40% in the last 6 months.

Taking these aspects into account, in the output method, revenue recognition is closely linked to the execution phase of the work and is highly dependent on the margin with which each item that makes up the work has been contracted. As such, in each execution phase, the result can vary greatly.

In the input method, revenue recognition and the margin have a more linear distribution but are highly dependent on a

TABLE I
RATIOS

No.	PREDICTOR VARIABLES	Category
19	Short-term liabilities/Total liabilities	Indebtedness
10	Long-term liabilities/Current liabilities	Indebtedness
18	Short-term liabilities/Long-term liabilities	Indebtedness
20	Total Debt/ (Total Assets-Current Liabilities)	Indebtedness
21	Long-term debt/ (Total Assets-Current Liabilities)	Indebtedness
3	Current Assets/Total Assets	Structure
6	Fixed Assets/Total Assets	Structure
7	Working capital/ Current liabilities	Structure
8	Short-term debt/Total Assets	Structure
15	Non-current assets/Long-term debt	Structure
16	Working Capital/ (Total Assets-Current Liabilities)	Structure
17	(Current Assets-Current Liabilities)/Current Assets	Structure
22	Working Capital/Total Assets	Structure
4	Current Assets - Stock/ Current Liabilities	Liquidity
5	Liquid Assets/ Current Liabilities	Liquidity
1	Total Debt/Total Assets	Solvency
2	Current Assets/Current Liabilities	Solvency
9	Liquid Assets/Total Assets	Solvency
11	(Current Assets-Stock/Current Liabilities)/Total Assets	Solvency
12	(Current Assets-Stock/Current Liabilities)/Current Liabilities	Solvency
13	(Current Assets-Stock/Current Liabilities)/Net Worth	Solvency
14	Current Assets/Total Debt	Solvency

correct estimate of the costs to be incurred.

In the end, the subjectivity of these methods based on analytical accounting, their high variability, as well as the possibility that a company with financial problems may resort to "earnings management" practices is what causes us not to use ratios that take into consideration the amount of income. As a consequence, the results for the companies are also not considered.

II. Objectives and Methodology

The main objective of the research is to find models that show greater accuracy in the prediction of insolvency in construction companies that also do so sufficiently enough in advance to serve as an early warning sign of this circumstance.

In terms of specific objectives, in this work, we look to develop two aims:

1.-To find the predictor variables (ratios) using only items from the balance sheet.

2.-To compare the Random Forest algorithms that achieve the highest accuracy.

In this study, we used the annual accounts published in the Mercantile Register, which is accessible in the SABI database (Iberian Balance Sheet Analysis System, owned by INFORMA, S.A.).

We selected all companies in CNAE 41.2 (residential and non-residential construction) that filed for insolvency proceedings between 2010 and 2019, as well as all those that were active as of December 2019. Another of the requirements for selection was that the turnover of the companies exceeded 6 million euros per year in the most recent fiscal year in the available accounts.

With these conditions, we obtained a total of 127 companies that had filed for insolvency proceedings and 631 active companies.

The methodology for reaching each of the specific objectives was as follows.

A. Methodology for predictor variables

For the analysis of the explanatory variables of the model, we used the ratios that have been most frequently used in the different insolvency prediction studies.

The ratios were ordered numerically by the number of times they appeared in works related to the subject according to

TABLE II

	R1	R2	R3	R4	R5	R6
Valid	561	561	561	561	561	561
Mean	0.666	2.635	0.775	1.604	0.342	0.225
Median	0.706	1.389	0.833	1.120	0.162	0.167
Std. Deviation	0.222	9.015	0.202	3.921	0.736	0.202
IQR	0.312	0.864	0.259	0.652	0.335	0.259
Shapiro-Wilk	0.963	0.126	0.884	0.166	0.363	0.884
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001
Range	1.480	193.269	0.984	63.451	10.400	0.984
Minimum	0.019	0.284	0.016	0.024	0.000	0.000
Maximum	1.500	193.553	1.000	63.475	10.400	0.984

	R7	R8	R9	R10	R11	R12
Valid	561	561	561	561	561	561
Mean	0.730	0.521	0.649	0.211	3.702e -4	0.004
Median	0.297	0.539	0.681	0.121	1.714e -4	3.397e -4
Std. Deviation	2.247	0.232	0.218	0.232	6.181e -4	0.047
IQR	0.605	0.376	0.324	0.271	3.674e -4	7.956e -4
Shapiro-Wilk	0.255	0.978	0.961	0.816	0.552	0.046
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001
Range	39.873	1.099	0.992	0.997	0.006	0.925
Minimum	-0.560	3.008e -4	0.006	3.084e -5	1.237e -7	1.789e -7
Maximum	39.313	1.100	0.998	0.997	0.006	0.925

	R13	R14	R15	R16	R17	R18
Valid	561	561	561	561	561	561
Mean	0.002	1.519	41.772	0.491	0.303	181.105
Median	5.540e -4	1.132	2.296	0.567	0.280	7.289
Std. Deviation	0.007	2.250	453.084	0.475	0.328	1.739.745
IQR	0.001	0.511	4.501	0.573	0.376	25.959
Shapiro-Wilk	0.230	0.264	0.058	0.756	0.887	0.072
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001
Range	0.129	40.207	10.177.130	5.842	3.515	32.425.285
Minimum	-0.003	0.064	0.000	-4.835	-2.520	0.003
Maximum	0.126	40.271	10.177.130	1.006	0.995	32.425.288

	R19	R20	R21	R22
Valid	561	561	561	561
Mean	0.789	2.205	0.287	0.254
Median	0.879	1.483	0.207	0.204
Std. Deviation	0.232	2.390	0.303	0.236
IQR	0.271	1.892	0.358	0.306
Shapiro-Wilk	0.816	0.724	0.773	0.966
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001
Range	0.997	29.943	3.715	1.452
Minimum	0.003	-11.034	-0.003	-0.504
Maximum	1.000	18.909	3.712	0.948

Tascón and Castaño (2012). As a result, ratio 1 (Total Debt/Total Assets) appeared on at least 18 occasions and ratio 2 (Current Assets/Current Liabilities) on 14 occasions. This makes a total of 22 ratios (Table 1).

We took the entire population of both classes and did the relevant data cleaning for missing data.

The method used to analyze the values that were considered as outliers was the Tukey (1977) method, with values being called extreme outliers if they were outside 3 times the interquartile range.

As can be seen in Table 2, regarding the descriptive statistics, after applying the Shapiro-Wilk test, the p-value of all the variables was less than 0.001. As such, the null hypothesis of normality was not fulfilled for a confidence interval of 95%. For this reason, we have to conclude that the variables do not follow a normal distribution.

As we were interested in selecting those variables that presented statistically significant differences between the two study groups (solvent and insolvent) and as we knew that the variables did not follow a normal distribution, we needed to apply non-parametric techniques.

We used the Brown-Forsythe test to assess whether there was an equality of variances in the two groups (homoscedasticity) and the Mann-Whitney test to test the means. As there were not many observations, we applied a significance level of 1. In other words, we estimated that there were statistically significant differences for a 99% confidence interval.

Next, we checked the correlation (with Spearman's method) between these variables to eliminate all those with a correlation higher than (+-) 2/3.

B. Machine Learning

We applied different Random Forest algorithms to the ratios we selected from the statistical analysis explained in the previous section.

Random Forest is one of the most powerful machine learning algorithms. It is an ensemble method and this type of method combines the predictions of several machine learning algorithms together to make more accurate predictions than when using an individual model.

We divided the sample into a random partition of 80% for model training and 20% for validation. For the 80% sample, we made adjustments to balance the two classes (the 20% sample subset is left with the real data).

There are generally three types of adjustment: reducing the number of samples in the larger class (undersampling), artificially increasing the number of samples in the smaller classes (oversampling) and a mixed option of both, reducing and enlarging simultaneously, which is the option we used to balance the data.

The training sample consisted of 450 observations while the validation sample consisted of 111 observations.

The training was carried out only for year 3 prior to the declaration of insolvency, and then in order to test the models obtained, they were applied on the validation sample reserved for the fiscal year n-3, and for the entire sample for year n-2.

Once we trained each of the algorithms, we made a comparison between the results obtained in each of the models, using two important metrics: Accuracy and Cohen's Kappa index.

Accuracy represents the total number of hits obtained divided by the total number of observations, while the Kappa index measures the agreement observed in a data set with respect to what could occur simply due to chance. If the index were zero, it would mean that the observed agreement coincided with what would be expected due to chance. As such, the higher the index, the lower the probability that the accuracy obtained is due to chance.

We also paid special attention to the metrics sensitivity or true positive rate and specificity or true negative rate.

III. RESULTS

With regard to the first part of data cleaning and the observation of extreme outliers, only the ratio number R21 presented two values in companies classified as solvent that presented negative equity, and, as this circumstance is a cause of the dissolution of the company, they were eliminated. In fiscal year n-3, which is the fiscal year in which the predictive model was prepared, there were, therefore, 561 observations consisting of 482 solvent companies and 79 companies classified as insolvent.

A. Results of the selection of predictor variables for the model.

For the selection of the predictor variables of the model, once the Brown-Forsythe test had been applied, the ratios in which the null hypothesis of equality of variances was not satisfied were found to be the following: R1, R4, R5, R7, R9, R10, R11, R13, R14, R19, and R21 (Table 3), and; applying the Mann-Whitney test, the ratios in which the null hypothesis of equality of means was not satisfied were found to be the following: R1, R4, R5, R9, R10, R11, R12, R13, R14, R15, R18, R19, and R21 (Table 4).

Therefore, the variables that presented statistically significant differences within the two study groups, since they fulfilled the hypotheses of differences in variances and means, were the following: R1, R4, R5, R9, R10, R11, R13, R14, R19, and R21. After applying Spearman's method for correlation analysis and eliminating the strongly correlated variables, the final selection of variables was as follows: R1, R4, R5, R9, R9, R11, R13, R14, and R19, which have the correlation found in Figure 1.

TABLE III
BROWN-FORSYTHE TEST

Brown-Forsythe Test (alpha = 0.01)		
R1	**R8**	**R15**
statistic : 37.38432	statistic : 0.6911356	statistic : 0.1029797
num df : 1	num df : 1	num df : 1
denom df : 121.9915	denom df : 109.4543	denom df : 190.2615
p.value : 1.204508e-08	p.value : 0.4075883	p.value : 0.7486349
R2	**R9**	**R16**
statistic : 2.598114	statistic : 16.36592	statistic : 0.06321762
num df : 1	num df : 1	num df : 1
denom df : 543.7047	denom df : 119.213	denom df : 132.5215
p.value : 0.1075713	p.value : 9.307057e-05	p.value : 0.8018703
R3	**R10**	**R17**
statistic : 0.0186803	statistic : 8.984707	statistic : 0.5522183
num df : 1	num df : 1	num df : 1
denom df : 116.1315	denom df : 98.68959	denom df : 116.4076
p.value : 0.8915236	p.value : 0.00344418	p.value : 0.4589098
R4	**R11**	**R18**
statistic : 14.65658	statistic : 144.8069	statistic : 0.6931077
num df : 1	num df : 1	num df : 1
denom df : 541.1116	denom df : 521.4298	denom df : 455.6836
p.value : 0.0001440621	p.value : 1.32747e-29	p.value : 0.4055454
R5	**R12**	**R19**
statistic : 57.35325	statistic : 3.498725	statistic : 8.984707
num df : 1	num df : 1	num df : 1
denom df : 553.1508	denom df : 481.0594	denom df : 98.68959
p.value : 1.537337e-13	p.value : 0.06202306	p.value : 0.00344418
R6	**R13**	**R20**
statistic : 0.0186803	statistic : 35.05521	statistic : 0.247748
num df : 1	num df : 1	num df : 1
denom df : 116.1315	denom df : 491.8931	denom df : 110.0651
p.value : 0.8915236	p.value : 6.019901e-09	p.value : 0.6196583
R7	**R14**	**R21**
statistic : 14.59712	statistic : 20.82645	statistic : 18.47168
num df : 1	num df : 1	num df : 1
denom df : 558.352	denom df : 556.7314	denom df : 85.52091
p.value : 0.0001480611	p.value : 6.191544e-06	p.value : 4.546357e-05
		R22
		statistic : 0.4552885
		num df : 1
		denom df : 105.4976
		p.value : 0.5013102

B. Results from applying machine learning.

We selected the following algorithms as models for the insolvency test:

Normal Random Forest
Weighted Subspace Random Forest
Global Random Forest
Regularized Random Forest
Conditional Inference Random Forest

TABLE IV
MANN-WHITNEY TEST

R1	W = 25689, p-value = 6.379e-07
R2	W = 17341, p-value = 0.2037
R3	W = 18290, p-value = 0.5751
R4	W = 12355, p-value = 5.591e-07
R5	W = 10270, p-value = 5.163e-11
R6	W = 19788, p-value = 0.5751
R7	W = 15871, p-value = 0.0177
R8	W = 19947, p-value = 0.4968
R9	W = 23743, p-value = 0.0004281
R10	W = 23479, p-value = 0.0008859
R11	W = 6864, p-value < 2.2e-16
R12	W = 6832, p-value < 2.2e-16
R13	W = 9431, p-value = 6.273e-13
R14	W = 11741, p-value = 4.639e-08
R15	W = 14450, p-value = 0.0005903
R16	W = 17653, p-value = 0.2995
R17	W = 17341, p-value = 0.2037
R18	W = 14599, p-value = 0.0008859
R19	W = 14599, p-value = 0.0008859
R20	W = 21731, p-value = 0.04385
R21	W = 26409, p-value = 3.418e-08
R22	W = 17735, p-value = 0,329

The five trained Random Forest algorithms achieved an accuracy of over 83% in the validation sample in year n-3 and hit rates equal to or above 80% in the insolvent companies.

The true positive rate improved in two algorithms (RRF and WSRF) from year n-3 to year n-2 and the true negative rate improved in four algorithms.

The most accurate model in terms of sensitivity or true positive rate was the Conditional Inference Random Forest model, with a hit rate of 86.67% at three years before insolvency and 85.71% at two years before insolvency.

The most accurate model in terms of specificity or true negative rate (in our case the accuracy of detecting solvent companies), was the normal Random Forest model, with a hit rate of 88.54% at three years before insolvency and 89% at two years before insolvency.

The full metrics of the five models in the validation sample can be seen in Table 5.

IV. Conclusions

The proposed early warning model is very simple as only the five assets and two balance sheet items from the company's annual accounts need to be entered online:

- Non-current assets
- Current assets
- Stock
- Liquid assets
- Net worth
- Non-current liabilities
- Current liabilities

With the combination of these data, the six input variables in the model are formed, and the model returns the probability of insolvency under each of the three best Random Forest algorithms that have demonstrated the greatest precision in the samples studied.

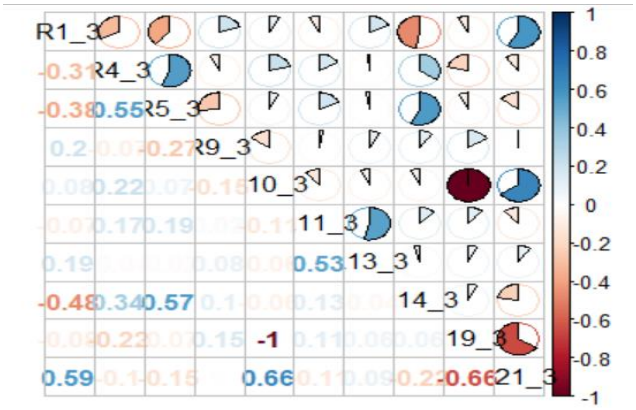

Fig. 1. Correlation matrix between explanatory variables.

Based on these three probabilities, both the debtor and all interested third parties (workers, financial institutions, suppliers, etc.) will have a first approximation of the company's solvency sufficiently in advance.

The proposed model responds to the needs set out in the Directive:

- Anticipation: The earlier a debtor can detect its financial difficulties and take appropriate measures, the greater the chance of avoiding imminent insolvency, as the Directive points out in recital 22.
- Accessibility: The tool is available online.
- Ease: With very little information being needed to enter into the tool, results can be obtained.
- Clarity: The response of the model in a way that shows the probability of insolvency under three different assumptions is clear and easy to understand.

This research has been carried out in a very little studied sector in the prediction of insolvency, and future lines of research should be aimed at seeking predictor variables that better explain in advance the phenomenon of insolvency, which is understood as a gradual process of deterioration. Therefore, relying on new machine learning algorithms, we must find patterns that, with enough time in advance, can better explain this deterioration.

TABLE V
FULL METRICS FOR YEARS n-3 AND n-2

RANDOM FOREST N-3	RANDOM FOREST N-2	WEIGHTED SUBSPACE RF N-3	WEIGHTED SUBSPACE RF N-2
Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 12 11 Solvent 3 85 Accuracy: 0.8739 95% CI: (0.7974, 0.9293) No Information Rate: 0.8649 P-Value [Acc > NIR]: 0.45813 Kappa: 0.5595 Mcnemar's Test P-Value: 0.06137 Sensitivity: 0.8000 Specificity: 0.8854 Pos Pred Value: 0.5217 Neg Pred Value: 0.9659 Prevalence: 0.1351 Detection Rate: 0.1081 Detection Prevalence: 0.2072 Balanced Accuracy: 0.8427 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 64 55 Solvent 13 445 Accuracy: 0.8821 95% CI: (0.853, 0.9073) No Information Rate: 0.8666 P-Value [Acc > NIR]: 0.1487 Kappa: 0.586 Mcnemar's Test P-Value: 6.627e-07 Sensitivity: 0.8312 Specificity: 0.8900 Pos Pred Value: 0.5378 Neg Pred Value: 0.9716 Prevalence: 0.1334 Detection Rate: 0.1109 Detection Prevalence: 0.2062 Balanced Accuracy: 0.8606 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 12 15 Solvent 3 81 Accuracy: 0.8378 95% CI: (0.7559, 0,901) No Information Rate: 0.8649 P-Value [Acc > NIR]: 0.835242 Kappa: 0.4813 Mcnemar's Test P-Value: 0.009522 Sensitivity: 0.8000 Specificity: 0.8438 Pos Pred Value: 0.4444 Neg Pred Value: 0.9643 Prevalence: 0.1351 Detection Rate: 0.1081 Detection Prevalence: 0.2432 Balanced Accuracy: 0.8219 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 63 68 Solvent 14 432 Accuracy: 0.8579 95% CI: (0.8267, 0.8854) No Information Rate: 0.8666 P-Value [Acc > NIR]: 0.7522 Kappa: 0.5261 Mcnemar's Test P-Value: 4.832e-09 Sensitivity: 0.8182 Specificity: 0.8640 Pos Pred Value: 0.4809 Neg Pred Value: 0.9686 Prevalence: 0.1334 Detection Rate: 0.1092 Detection Prevalence: 0.2270 Balanced Accuracy: 0.8411 'Positive' Class: Insolvent
RANDOM FOREST GLOBAL N-3	**RANDOM FOREST GLOBAL N-2**	**REGULARIZED RF N-3**	**REGULARIZED RF N-2**
Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 12 13 Solvent 3 83 Accuracy: 0.8559 95% CI: (0.7765, 0.9153) No Information Rate: 0.8649 P-Value [Acc > NIR]: 0.67152 Kappa: 0.5187 Mcnemar's Test P-Value: 0.02445 Sensitivity: 0.8000 Specificity: 0.8646 Pos Pred Value: 0.4800 Neg Pred Value: 0.9651 Prevalence: 0.1351 Detection Rate: 0.1081 Detection Prevalence: 0.2252 Balanced Accuracy: 0.8323 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 60 67 Solvent 17 433 Accuracy: 0.8544 95% CI: (0.823, 0.8822) No Information Rate: 0.8666 P-Value [Acc > NIR]: 0.8214 Kappa: 0.5062 Mcnemar's Test P-Value: 8.975e-08 Sensitivity: 0.7792 Specificity: 0.8660 Pos Pred Value: 0.4724 Neg Pred Value: 0.9622 Prevalence: 0.1334 Detection Rate: 0.1040 Detection Prevalence: 0.2201 Balanced Accuracy: 0.8226 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 12 15 Solvent 3 81 Accuracy: 0.8378 95% CI: (0.7559, 0,901) No Information Rate: 0.8649 P-Value [Acc > NIR]: 0.835242 Kappa: 0.4813 Mcnemar's Test P-Value: 0.009522 Sensitivity: 0.8000 Specificity: 0.8438 Pos Pred Value: 0.4444 Neg Pred Value: 0.9643 Prevalence: 0.1351 Detection Rate: 0.1081 Detection Prevalence: 0.2432 Balanced Accuracy: 0.8219 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 61 77 Solvent 16 423 Accuracy: 0.8388 95% CI: (0.8062, 0.8679) No Information Rate: 0.8666 P-Value [Acc > NIR]: 0.9761 Kappa: 0.478 Mcnemar's Test P-Value: 4.918e-10 Sensitivity: 0.7922 Specificity: 0.8460 Pos Pred Value: 0.4420 Neg Pred Value: 0.9636 Prevalence: 0.1334 Detection Rate: 0.1057 Detection Prevalence: 0.2392 Balanced Accuracy: 0.8191 'Positive' Class: Insolvent
CONDITIONAL INFERENCE RF N-3	**CONDITIONAL INFERENCE RF N-2**		
Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 13 16 Solvent 2 80 Accuracy: 0.8378 95% CI: (0.7559, 0,901) No Information Rate: 0.8649 P-Value [Acc > NIR]: 0.835242 Kappa: 0.5022 Mcnemar's Test P-Value: 0.002183 Sensitivity: 0.8667 Specificity: 0.8333 Pos Pred Value: 0.4483 Neg Pred Value: 0.9756 Prevalence: 0.1351 Detection Rate: 0.1171 Detection Prevalence: 0.2613 Balanced Accuracy: 0.8500 'Positive' Class: Insolvent	Confusion Matrix and Statistics Reference Prediction Insolvent Solvent Insolvent 66 91 Solvent 11 409 Accuracy: 0.8232 95% CI: (0.7896, 0.8535) No Information Rate: 0.8666 P-Value [Acc > NIR]: 0.9987 Kappa: 0.469 Mcnemar's Test P-Value: 5.192e-15 Sensitivity: 0.8571 Specificity: 0.8180 Pos Pred Value: 0.4204 Neg Pred Value: 0.9738 Prevalence: 0.1334 Detection Rate: 0.1144 Detection Prevalence: 0.2721 Balanced Accuracy: 0.8376 'Positive' Class: Insolvent		

REFERENCES

Alaka, H. A. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert System with Applications, 164-184.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 589-610.

Altman, E. (2005). An emerging market credit scoring system for corporate bonds. Emerging markets review, 311-323.

Altman, E., Iwanicz-Drozdowska, M., Laitinen, E., & Suvas, A. (2017). Financial distress prediction in an international context. A review and empirical analysis of Altman's Z-score model. Journal of International Financial Management & Accounting, 131-171.

Arora, I. (2020). Prediction of corporate bankruptcy using financial ratios and news. International Journal of Engineering and Management Research, 82-87.

Beaver, W. (1966). Financial ratios as predictors of failure: An empirical research in accounting selected studies. Journal of Accounting Research, 71-111.

Becerra-Vicario, R., Alaminos, D., Aranda, E., & Fernández-Gámez, M. A. (2020). Deep recurrent convolutional neuronal network for bankruptcy prediction: A case of the restaurant industry. Sustainability, 1-15.

Beneish, M. (1999). The D¡detection of earnings

manipulation. Financial Analysts Journal, 24-36.

Blum, M. (1974). Failing company discriminant analysis. Journal of accounting research, 1-25.

Deakin, E. (1972). A discriminant analysis of predictor of business failure. Journal of accounting research, 167-179.

Edmister, R. (1972). An empirical test of financial ratio analysis por small business failure prediction. Journal of Financial and Quantitative Analysis, 1477-1493.

Garcia Catellví, A. (2005). El cálculo del resultado de los proyectos desarrollados por constructoras. Partida Doble, 42-53.

Tukey, J. W. (1977). Exploratory data analysis. Massachusset: Audison-Wesley Publishing Company.

Lantz, B. (2013). Machine Learning with R. Birmingham: Packt Publishing, Ltd.

Meyer, P. A., & Pifer, H. W.(1970). Prediction of bank failures. The Journal of Finance, 853-868.

Ohlson, J. (1980). Financial ratios and the `probabilistic of bankruptcy. Journal of accounting research, 109-131.

Perboli, G., & Arabnezhad, E. (2021). A Machine Learning-based DSS for mid and long-term company crisis prediction. Expert Systems with Applications, 1-12.

Rosner, R. (2003). Earnings manipulations in failing firms. Contemporary accounting research, 361-408.

Sajjan, R. (2016). Predicting bankruptcy of selected firms by applying Altman's Z-score model. International Journal of Research-Granthaalayah, 152-158.

Tabbakh, A., Kumar, J. S., & Janjhi, N. (2021). Bankruptcy Prediction using Robust Machine Learning Model. Turkish Journal of Computer and Mathematics Education, 3060-3073.

Tascón, M., & Castaño, F. (2012). Variables y modelos para la identificación del fracaso empresarial: Revisión de la investigación empírica reciente. Revista de Contabilidad, 7-58.

Wang, H., & Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data. Plos One, 1-17.

Zmijewski, M. (1984). Methodological Issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 59-82.