Insolvency prediction for Portuguese agro-industrial SME: Tree Bagging Methodology

José Augusto Canto, Amélia Cristina Ferreira Silva, Gabriela Leite, Carlos Machado-Santos

Abstract
The aim of this study lies on the empirical application of the tree bagging methodology, in order to predict the insolvency of Portuguese Small and Medium-sized Enterprises (SME) in the agro-industrial sector, one year in advance. The database consists of financial indicators of 243 companies, available at SABI (Iberian Balance Analysis System), all from agro-industrial sector. The proposed model reveals a robust result when compared with traditional parametric models.
The results show that two indicators – "short-term liquidity" and "capacity to generate results appropriate to the size" – were the most statistically relevant, both in the Proposed Model and the Logistic Regression model.

Keywords: Insolvency; Bagging; Decision Tree; Overfitting; Agro-industrial; Financial indicators.

1. INTRODUCTION

Insolvency is a natural phenomenon for firms that operate in open market economies. However, the presence of potential insolvency undermines economic transactions, which are based on trust. In this context, it is of crucial importance for economic agents the use of models that may predict and anticipate insolvency situations, reducing financial risks of economic operations.

Throughout the years, various techniques have been used to develop insolvency forecast models, according to Breiman (2001) there are two cultures in the use of forecast mathematical models: The first one, traditional in the statistics community, named *data modelling culture*, assumes in a general fashion the $r(x) = \beta 0 + \beta ixi$ model. Its main objective is the interpretation of the βi parameters, subjected to the hypotheses of normality, linearity, and homoscedasticity to validate a theory. The second one, depends on the evolution of computers, named *algorithmic modelling culture* which dominates the *machine learning* community, the algorithms verify automatically the relations between variables, not subjected to the hypothesis of the traditional models.

In the 60s, with the publication of Altmam (1968), the insolvency forecast studies had an important boost by correlating the various financial indicators through linear analysis models of *data modelling culture.* In the 80s, the linear analysis models shared the prediction study space with the logistical models, which present their results in the form of accumulated probability, an improvement in the interpretative quality of prevision, by substituting the linear scores of parametric models.

The technological evolution of the 90s, brought to light alternatives on the study of insolvency forecast, by incorporating *machine learning* algorithms, accrued from the *algorithmic modelling culture*, of which are examples the Decision Trees, Neural Networks Theory, Genetic Algorithm Theory, and Fuzzy Algorithm Theory.

Amongst the options, Quinlan (1986) highlights: a) greater ease of comprehension, for being greatly intuitive; b) the ability to deal with absent and extreme values; c) besides dealing very well with normally distributed variables, the Decision Trees Algorithm automatically detects non-linear interactions and adjusts itself to them. The classical methods suffer greatly with these problems.

However, the *algorithmic* tends to generate *"overfitting"* models, a problem confirmed by (Kothari and Dong, 2001). This happens when the original set of items is well classified by the model, but it presents an important risk of lowering its performance with new data. For this reason, the *tree bagging* technic of Breiman (1996) associates the *bagging* process with the Decision Trees to reduce this model's instability.

In the *bagging* methodology, each tree replica works as a trained classifier, the set of replicas generates a committee of trees, which through voting forecast a new datum.

The goal of this study is to apply the *Tree-Bagging* in order to predict, one year in advance, the insolvency of Portuguese SME from the agro industrial sector, and provides 3 critical contributions: (1) it presents a technical alternative from the *algorithmic modelling culture* with potential for identifying the complex and non-linear relations which are present in SME data, for the prevision of insolvency one year in advance. (2) it shows that the alternative technique is as much or more capable of predicting the insolvency of these Portuguese SME as the traditional statistic methods, represented by the model of Logistic Regression (3) the empirical results, besides suggesting relevance of the predictive capacity of the alternative model, also reinforce the importance of short term liquidity and investment profitability indexes to anticipate the insolvency of the Portuguese SMEs of the agro industrial sector.

2. LITERATURE REVIEW

Beaver (1966), through univariate discrimination, presented the first paper with statistical techniques, by employing countable data to predict bankruptcy. From that moment on, the amount of research connecting financial indicators grew all over the planet, to address problems of insolvency forecast, bankruptcy, and financial hardship.

Altman (1968) gave momentum to the study of forecast models, in spite of the result of its discriminating function, known as Z Score, not being very intuitive. Perhaps for that reason, during the 80s, the models of discriminating analysis gradually came to share space with logistical analysis, models which don't need to assume the premise of discriminating analysis of multivariate normality assumption, embodying the effects of non-linearity. On that technique, logistically distributed financial indicators are used.

Ohlson's (1980) logistical analysis used eight financial indicators, and was able to identify, one year in advance, bankruptcy of companies with 89% precision rate. Platt and Platt (1991), whilst elaborating their models, advised the usage of financial indicators standardized by the sector, instead of absolute indexes from the companies. Huang, et al (2017) have developed some work with a sample containing financial indicators from 156 Chinese solvent companies and 156 insolvent ones, collected (2000 - 2011) in order to compare accuracy between discriminating analysis models and logistic analysis, the result was the same 74,2 %. Hensher, et al (2007) and Shumway (2001) also used financial indicators and the logistic technique to anticipate bankruptcy with good results, 92% and 88% respectively.

Although the use of the *algorithmic modeling culture* is still recent on the financial projects, there are several papers being published. For example, Auria *et al.* (2009), Brown (2012), Butaru *et al.* (2016) and Sealand (2018) have deeply studied the prediction of financial problems by analysing the credit risk with the use of algorithms. Other references in the field can be found in the studies of Dietterich (2000), Deng (2016), Addo, et al (2018) Tokpavi (2018). These authors compared the results obtained with the traditional statistical model of Logistic Regression. In this paper we follow this approach and look at the problem of bankruptcy prediction in terms of several financial ratios which are intrinsically linked to the financial strength of Portuguese SME of agro-industrial sector.

Liao, et al (2014) through a sample of financial indicators from 63 insolvent and 2680 solvent companies, verified an accuracy of 94,91% with the Bagging methodology, and of 92,44% with the Discriminating Analysis. Nagaraj and Sridhar (2015) with the same goal, and using a sample of financial indicators of 107 bankrupt and 143 non-bankrupt companies, found an accuracy of 97,4% with the Bagging methodology and 97,2% with the Logistic Analysis model.

It's thus verified that, in a general fashion, the financial forecast papers have in common the use of sets of financial indicators on the country of origin of the research as a data source; concern for defining the timeline of the dataset and comparative study of techniques, as for their performance in terms of prevision accuracy.

According García et. al. (2019:89), "unlike the statistical models, machine learning and computational intelligence methods do not assume any specific prior knowledge, but instead they automatically extract information from past observations. These are represented by a set of explanatory variables, which usually correspond to financial ratios, macroeconomic indicators and sociodemographic characteristics, either straightforwardly represented as continuous variables or discretized as qualitative information".

A brief search, in Web of Science Core Collection, for articles published in journals, in the last five years, with the TOPIC: "Bagging" AND "bankruptcy", result in twenty papers, of which 1 was duplicate. After an initial screening, we excluded 3 papers. The most relevant information extracted from each of the 16 remained papers is presented in the following table.

Tab. 1. *Literature review 2016-2020*

Article	Objectives	Empirical application	Conclusions

Pisula (2020)	To solve the bankruptcy prediction problem from the perspective of learning with label proportions, which can not only overcome the limitation that massive training data is hard to be labeled, and to improve a framework for the applications of machine learning models in bankruptcy prediction.	Australian; Japanese, German, Polish	The proposed methods can not only explicitly model the unknown instance-level labels and the known label proportions under a large-margin framework, but also improve the performance through introducing ensemble learning strategies. Extensive experiments on the benchmark datasets demonstrate their efficiency and superiority on solving the problem of bankruptcy prediction.
Chen, et. al. 2020	To develop a scoring model (with good classification properties) that can be applied in practice to assess the risk of bankruptcy of enterprises in various sectors.	Poland	The GBM-based ensemble classifier model present superior classification capabilities. The approach presented in the paper can be used not only to assess the risk of bankruptcy of enterprises by market analysts and regional analysts, but also in banking activities to assess credit risk for corporate loans.

Lahmiri, et. al. 2020	To assess the relative performance of existing state-of-the-art ensemble learning and classification systems with applications to corporate bankruptcy prediction and credit scoring. The considered ensemble systems include AdaBoost, LogitBoost, RUSBoost, subspace, and bagging ensemble system.	Polish	AdaBoost ensemble learning and classification system is effective as it yields to lowest misclassification rate with relatively less complexity represented by number of weak learners and processing time. Ensemble classification systems are useful intelligent tools for classification of financial data.
Shrivastava et. al. (2020)	To create an efficient and appropriate predictive model using a machine learning approach for an early warning system of bank failure.	India	Application to various stakeholders like shareholders, lenders and borrowers etc. to measure the financial stress of banks.
Guo et. al. (2019).	To present a novel multi-objective particle swarm optimization for credit scoring (MOPSO-CS), and MOPSO-CS focuses on enhancing credit scoring models based on LDA in three aspects: (i) to construct a higher accuracy credit scoring model which is easy to be interpreted; (ii) to find the most suitable cut-off for discriminating “good credit” customers and “bad credit” customers; and (iii) to improve the sensitivity of the classifier by using multi-objective particle swarm optimization.	UK German Taiwan	Compared with black box technologies such as ANN and SVM, the credit score function proposed is more comprehensible. The example and experimental studies based on benchmark data sets and real-world data sets confirm that the proposed method outperforms the counterparts in term of sensitivity while maintaining acceptable accuracy.

García et. al. (2019)	To gain some insight into the potential links between the performance of classifier ensembles (BAGGING, AdaBoost, random subspace, DECORATE, rotation forest, random forest, and stochastic gradient boosting) and the positive sample types.	Australian, Finland, Polish, Japanese, German, Taiwan, Iranian	The analysis on each category of databases has shown that the performance of any ensemble configuration depends on the types of samples available in the data set. This finding can be especially useful when one has decide which classifier to apply for a particular problem in hand, thus avoiding to choose by a trial-and-error approach the most appropriate prediction model.
Sánchez-Medina et. al. (2019).	To analyze the effect of the normative change that took place in Spain in December 2010, related to opinions modified for going-concern uncertainties. Until that date, the auditor's uncertainty about the company's going-concern status led to a qualified opinion. However, under the new regulation, it became an opinion that included an explanatory paragraph stating the reasons for concern, which was considered less serious.	Spain	A change in the norm that catalogs the going-concern issue as less serious made auditors more likely to report this situation, thus questioning the audit quality. The users of accounting information must pay special attention to auditors' behavior when regulatory changes occur in the auditing field. With the proposed classifiers, it would be possible to establish, with a high level of accuracy, whether the auditors' opinion was coherent with the financial situation of any SME before the regulatory change.

Xia et. al. (2018).	To propose a novel heterogeneous ensemble credit model that integrates the bagging algorithm with the stacking method. The proposed model differs from the extant ensemble credit models in three aspects, namely, pool generation, selection of base learners, and trainable fuser. This paper also considers the relationship between the number of iterations (i.e., T) and model performance.	Australian German	The proposed stacking model significantly outperforms the benchmark individual and homogeneous ensemble models. The empirical results reveal that 40–60 iterations are suitable for the proposed stacking model. Furthermore, interpretability should be highlighted to achieve a balance among accuracy, complexity and interpretability of a real-world credit scoring model.
Sun et. al. (2018).	To propose a new DT ensemble model for imbalanced enterprise credit evaluation based on the synthetic minority over-sampling technique (SMOTE) and the Bagging ensemble learning algorithm with differentiated sampling rates (DSR), which is named as DTE-SBD (Decision Tree Ensemble based on SMOTE, Bagging and DSR).	China	The comparation among the six models of pure DT, over-sampling DT, over-under-sampling DT, SMOTE DT, Bagging DT, and DTE-SBD indicate that DTE-SBD significantly outperforms the other five models and is effective for imbalanced enterprise credit evaluation.

Dahiya et. al. 2017	To present a feature selection-based hybrid-bagging algorithm (FS-HB) for improved credit risk evaluation.	German	The hybrid FS-HB algorithm performed best for qualitative dataset with less features and tree-based unstable base classifier. Its performance on numeric data was also better than other standalone classifiers, whereas comparable to bagging with only selected features.
Zhu et. al. 2017).	To apply an compare six methods, i.e., one individual machine learning (IML, i.e., decision tree) method, three ensemble machine learning methods [EML, i.e., bagging, boosting, and random subspace (RS)], and two integrated ensemble machine learning methods (IEML, i.e., RS–boosting and multiboosting),.	China	The IEML methods acquire better performance than IML and EML method. In particular, RS–boosting is the best method to predict SMEs credit risk among six methods.
Barboza (2017)	To test machine learning models (support vector machines, bagging, boosting, and random forest) to predict bankruptcy one year prior to the event, and compare their performance with results from discriminant analysis, logistic regression, and neural networks.	USA	The bagging, boosting, and random forest models outperform the others techniques, and all prediction accuracy in the testing sample improves when the additional variables are included.

Ekinci & Erdal (2017)	To compare three common machine learning models grouped in the following families of approaches: (i) conventional machine learning models, (ii) ensemble learning models and (iii) hybrid ensemble learning models.	Turkey	The hybrid ensemble machine learning models clearly outperforme over conventional base and ensemble models. These results indicate that hybrid ensemble learning models can be used as a reliable predicting model for bank failures.
du Jardin (2016)	To suggest a set of profiles that closely mirror the various situations firms may experience at a given moment of their existence, before going bankrupt, then to build as many models as there are profiles. These profiles are estimated using a vector quantization method (Kohonen map).	French	Ensemble models seem to capture some variations within the decision space that individual models do not, thanks to the diversity they generate randomly, while profile-based models designed with these same techniques are also able to capture such variations, but more accurately, and this time not by chance but through to the knowledge they convey about bankruptcy.
Yao & Lian (2016)	To propose a new Support Vector Machine (SVM) based ensemble model (SVM-BRS) to address the issue of credit analysis. The model combines random subspace strategy and boosting strategy, which encourages diversity.	German	The proposed model has the potential to generate more accuracy classification. The ensemble model performs better than a single model.

Chang et. al. (2016)	To propose a decision tree-based short-term default credit risk assessment model to assess the credit risk. This paper integrates bootstrap aggregating (Bagging) with a synthetic minority over-sampling technique (SMOTE) into the credit risk model to improve the decision tree stability and its performance on unbalanced data.	Taiwan	The classifying recall rate and precision rate of the proposed model was obviously superior to the logistic regression and Cox proportional hazards models

As we can see from table 1, the topic of bankruptcy is as important as the credit rating. Despite the recent contribution on the topic come from various parts of the world, there is an emphasis on Asia. In general, investigation demonstrate the superiority of computational methods over statistical techniques. However, machine learning models offer a black box from which we only get the result, but we do not know of explain them.

3. METHODOLOGY

The proposed methodology uses the *Tree Bagging* technique for supervised training of the constituted examples of financial indicators. The use of financial indicators for training, assumes the premise of information accumulation, consequential from a set of observed (like heightened demand) and non-observed (like managerial characteristics) factors on countable demonstrations. According to Beaver (1966), the same will happen with the financial indicators, which justifies its use as a predictors or estimators of the company's insolvency probability.

$$Prob\ (insolvency) = f(financial\ indicators)$$

The concept of insolvency is applied to the supervised training orientation and it is in accordance with the Article 3 (2) of the Insolvency and Corporate Recovery Code, described by Figueiredo (2018): "it is considered that in insolvency situations the debtor is unable to fulfil his overdue obligations, are also considered insolvent when its passive is superior to the active, evaluated according the applicable accounting standards".

The *tree bagging* technic is explained by He *et al.* (2005) and Guoh *et al.* (2004). It is a classifier generated by Decision Trees replicas, which are the algorithms built by a function known as impurity-function. The function seeks to minimize the margin of error thoroughly by recursive process. It is minimal when all the data belong to the same type and maximal when the data are distributed linearly through the various types.

According to Sutton (2005) the impurity-functions – Entropy Function and Gini Index – are listed as being more used in the classification tree.

$$Entropy(N) = \sum_{j=1}^{m} -p_j \, log_2 p_j$$

$$Gini(N) = \sum_{j \neq m}^{m} -p_j \, p_{m=}1 - \sum_{j=1}^{m} P_j{}^2$$

Where: N is the set of examples; m is the set of types: p_j is the proportion of N which belongs to type j, then we have: $p_j = \frac{|N_j|}{N}$

The growing tree procedure tries to find an optimal way by the attribute's selection. One of the known measures of the attribute's selection is the Information Gain.

$$\Delta Gain(N,t) = Entropy(N) - Entropy(N_l) - Entropy(N_r\ )$$

Where: t is the current attribute; $Entropy(N)$ is the impurity of the current node; $\Delta Gain(N,t)$ is the gain of the attribute t above the set N.

For each replica, a Decision Tree, which works as a trained classifier, is generated. The set of replicas generates a committee of trees, which predicts new data through vote. It is reasonable to suppose that this prediction is stronger than the prediction of only one tree.

To generate multiple Decision Tree versions, the *Bagging* method builds *bootstrap* samples from the set of original data. According to Breimam (1996) a training set $\mathcal{L}$ consists of $\{(x_i\ ,\ y_i)\ , i\ =\ 1; : : : ; N\}$ data, where N is the quantity of examples; x_i attributes or input variables; y_i variables' answers or types used for the training.

If the input is x we can estimate y by the predictor $\varphi(x_i\ , \mathcal{L}\ )$. Now, suppose a set of predictors $\{\mathcal{L}_k\}$, each one with N independent observations, originated by the same subjacent distribution $\mathcal{L}$, with the purpose of improving the learning of one single $\varphi(x_i\ , \mathcal{L}\ )$. The authorization for working with the sequence of the set predictors is restricted $\{\varphi(x_{i,}, \mathcal{L}_k)\}$.

If $\varphi(x_i\ , \mathcal{L}\ )$ predicts a type $j\ \ \in\ \{1, \ldots, j\ \}$ then one method to aggregate $\varphi(x_{i,}, \mathcal{L}_k)$ is by majority voting. To do $N_j\ =\ \#\{k; \varphi(x_{i,}, \mathcal{L}_k) = j\}$ in order to find $\varphi_A(x) = arqmax_j\, N_j$.

Usually there is only one training set $\mathcal{L}$ without replicas, which conducts to the process of finding φ_A. To that end, copies of the *bootstrap* samples $\{\mathcal{L}^{(B)}\}$ are made from $\mathcal{L}$ to $\{\varphi(x_i\ , \mathcal{L}^{(B)}\ )\}$.

If y is a type, as in work, we take $\{\varphi(x_i\ , \mathcal{L}^{(B)}\ )\}$ to do the voting in order to find $\varphi_B(x)$. We call this procedure *"bootstrap aggregation"* also known as bagging.

Each of the Decision Trees is only trained with 63 % of observations, because of the random choice of n between N observations with replacement. This portion of data is known as *"in-bag"* data, while 37% of hidden observations are the *"out-of-bag"* observations. The "out-of-bag" observations are not used to build nor prune any tree, but to provide better error estimation for each of the tree nodes, besides other generalization errors for the predictors originated from *"bagging"*.

The *"out-of-bag"* observations' calculated errors are used to estimate the force of prediction and the attributes' input variable importance. As the ability of prediction is more dependent on the important attributes and less dependent on the less important attributes, we can use this idea to measure the importance of each attribute. We can understand the importance of this attribute by exchanging randomly the data and investing in the increase of the error.

The technic that will serve as a traditional statistical reference to validate the proposed methodology uses logistically distributed accountable indicators, in a form of cumulative probability between the 0 and 1 values. It provides a better interpretative quality for the forecast to present the probability form. This is a significant attribute in the decision making. The logistical distribution described by Zavgren (1985) is a special function type identified as a cumulative logistical function.

$$Pi = E(Y = 1/ Xi) = (e^{B0+B1x}) / (1 + e^{B0+B1x})$$

One of the first relevant studies of logistical analyses Ohlson (1980) used eight financial indicators and was able to identify with a precision of 89% company bankruptcies a year in advance. Hensher *et al.* (2007) and Shumway (2001) also used financial indicators with 92% and logistical technic with 88% to anticipate bankruptcy.

The main purpose is to build an insolvency forecast model for the Portuguese agro-industrial SME using the methodology called *tree bagging*. The validation of the proposed model is followed by the methodology related to the use of the statistical traditional model as a performance parameter.

The experiments made in this study are divided into two groups: adjustments and tests with logistical modelling and the proposed model. The experiments are made separately having in common only the data definition phases.

The methodological description, summarized in Figure 1, includes the experimental methodology (it omits any research references). It is divided into five steps: (1) data description (indicators); (2) data cleansing; (3) variables selection; (4) adjustment (or training); and (5) tests.

In the data description we describe the indicators which constitute the potential input variables for the predicting model. At the data cleansing, variables selection and tests, the used strategies are explained.

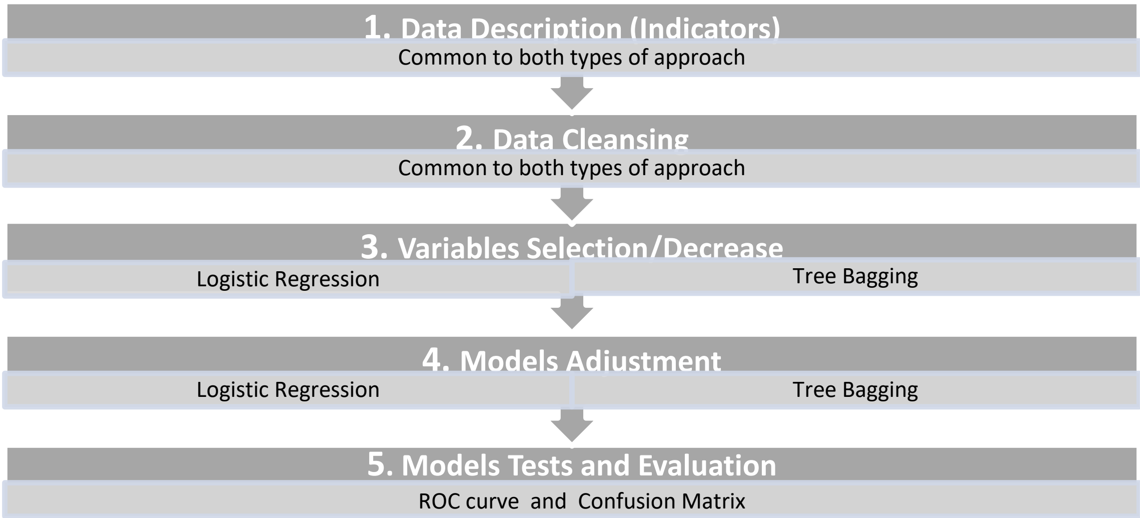

Fig. 1. *Experimental methodology*

The database contains European financial indicators covered by SABI (Insolvency and Corporate Recovery Code) research tool database. The initial database had 2,236 Portuguese SME of agro-industrial sector: agriculture, animal production, hunting and activities related to the forestry, forest exploitation, food industries, beverages, tobacco, leather and cork. Although the database includes SMEs from quite different sub-sectors, we are working with the "average risk". So, we decide to ignore potential heterogeneity across companies in the various sub-sectors.

The SME European concept was adopted as published by the Official Journal of the European Union (20.05.2003): "The category of micro, small and medium-sized enterprises (SME) consists of companies employing less than 250 people whose annual turnover does not exceed EUR 50 million or whose total annual balance sheet does not exceed EUR 43 million".

All SME organized on *"cross-section"* observe the 2007-2017 timeframe of the annual publication of the corporate financial indicators contained in database.

Criteria from the initial database were adopted to select the final sample. The first criterion was the extraction only of the SME base with complete financial indicators in the series. The companies were divided into two types: solvent companies and insolvent companies.

Adopted company selection criterion for the insolvent type: Company published one year before Equity became negative in a series of at least three consecutive negative years, and company published one year before leaving base by default. The criterion adopted to select solvent company, does not reflect negative equity in the period 2007-2017.

The adopted criterion to choose the indicators include the data integrity related with the implementation of Accounting Normalization System on 1st January 2010. All the solvent companies were collected in 2017. After 2010, insolvent companies were collected due to the criterion of three consecutive balance sheets with negative equity.

After the selection of the companies, 11 financial indicators were selected, as shown in Table 2. There is no theory for the choice of financial indicators, the adopted criterion encompassed the tradition of usage in similar papers, the integrity and availability of datum in the database, there was no selection of indicators pondered by quantity of workers, such as workplace productivity, as not to mix with other non-pondered financial indicators.

Tab. 2. *Used indicators*

Indicators	Formula
Current liquidity ratio	Current Assets / Net Liabilities
Liquidity ratio	(Current Assets - Inventories) / Net Liabilities
Shareholder liquidity ratio	Equity / Fixed Liabilities
Solvency ratio	(Equity / Total Assets) * 100
Leverage	((Fixed Liabilities + Financial Debts) / Equity) * 100
Profit margin	(Earnings Before Tax / Operating Income) * 100
Shareholder liquidity ratio	(Earnings Before Tax / Equity) * 100
Return on Capital Employed	(Earnings Before Tax + Financial Expenses And Similar Expenses) / (Equity + Fixed Liabilities)) * 100
Return on Total Assets	(Earnings Before Tax / Total Assets) * 100
Ability to cover interest	Earnings Expense / Financial Expenses and Similar Expenses
Stock Turnover	Operating income / inventory

Source: Self elaboration

Data cleansing is a treatment made for the selected data. It ensures the quality (completeness, veracity and integrity) of the presented facts. Common tasks of the data cleansing are: (i) fill in missing values, (ii) identify *outliers* and (iii) soften noises and correct erroneous or inconsistent information. Besides the identified missing *"outliers"* data which were inconsistent with the reality, this work required adjustments in the data for the first two tasks.

The predictor variables selection is going to be made separately with some common considerations. For example, existence of high correlation between predictor variables. To select modelling variables for Logistic Regression, a parametric Wald test is applied where the null hypothesis was verified at the 5 % level. For the *Tree-Bagging* modelling the importance of the attributes measured by the classification error of the "*out-of-bag"* observations are verified. The process comprises the successive removal of the predictor variables to verify the variation of the classification error with the lack. According to Arlot *et al.* (2010) the 10-fold cross-validation error is tested and the set of indicators with the smallest error is selected in order to find the best set of predictors. The samples are divided into ten *"folds"* parts during the process. Nine are used for the training and one for testing in a circular and successive manner.

Models are separately adjusted, and, on the Logistic regression, the coefficient values are generated to set the logit company insolvency predictor function. In the *bagging* methodology, 200 Decision Trees are generated, and classification error is verified. An error reduction in the number of *bootstrap's* copies is expected. The 200 trees together, form the vote committee, on which each *bootstrap* copy has a vote to forecast the SME insolvency. Thus, the methodology faces *overfitting* problem of the decision tree.

After the adjustment phase, the models are tested and evaluated through statistical tests. Models are evaluated by the amount of arrangements and error types. When solvent companies are differentiated from insolvent companies, two types of errors can occur: error type I, related to an insolvency result when the company is solvent and error type II, which represents the possibility to select the company as solvent when it is insolvent. To verify the correct answers and errors, the *Machine Learning* methodology uses a medical method used to evaluate the health exams quality. Method that uses the Confusion Matrix table to account the results and the ROC curve tool that allows exam evaluation at several cut points.

The Confusion Matrix and ROC (*Receiver Operating Characteristic*) Curve tools offer effective measures of performance by showing the correct and incorrect classification numbers versus foretold classifications for each type with a set of dichotomist examples.

The Confusion Matrix, shown on table 3, includes the necessary data for the calculations of metrics named by precision, specificity and sensitivity.

Tab. 3. Confusion Matrix Model applied for the insolvency forecast

Forecast Insolvent	TP	FP
Forecast Solvent	FN	TN
Types	Insolvent	Solvent

Legend: TP – True positive; TN - True negative; FP - False positive; FN – False negative.

The FP result is related to the Error Type I and FN is related to the Error Type II. Precision measures the probability that the test result is correctly classified, by the total examples: $(TP + TN)/T$. Sensitivity corresponds to the probability that the test correctly classifies a company as insolvent: $TP/(TP + FN)$. Specificity corresponds to the probability that the test correctly classifies a company as solvent: $TN/(FP + TN)$.

ROC curves, as it is shown in figure 2, represent the sensitivity and specificity for all the possible cut-off values under the curve. It will be used overall to evaluate the used methodologies in this work.

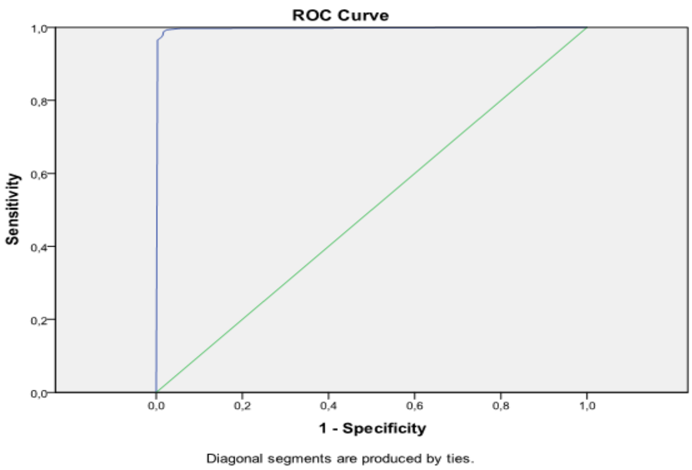

Fig. 2. *Example of a ROC curve graph extracted from SPSS platform*

3. RESULTS AND DISCUSSION

In the initial database, from the 2,236 Portuguese SME of agro-industrial sector 2,058 companies were identified as being solvent and 178 as being insolvent. As we can see, it was an unbalanced sample. It is explained by Drummond *et al.* (2003) that the precision and generalization capacity of models for the problem selection suffers from the influence of the sample size, the number of attributes and data balance which implies selection restrictions.

When the problem of data imbalance was prioritized, the adopted solution was to balance the sample for 356 companies. It was reduced to 243 companies, 122 solvent and 122 insolvent, after the data cleansing process, outliers and missing data.

In an effort to adjust the model's complexity to the size and quality of the available sample, the attributes selection process was separated by methodology. Thus, the initial attributes were restricted to the more significant and more important ones. All the experiments were made by using the computational platform Matlab® from Mathworks.

3.1 Selection of the input variables

To synthetize and simplify, the variables or attributes assumed input numbers.

Tab. 4. Numerical match of the attributes

1	Return on equity
2	Return on invested capital
3	Return on total assets
4	Profit margin
5	Ability to cover interest
6	Stock Turnover
7	Current liquidity ratio
8	Liquidity ratio
9	Shareholder liquidity ratio
10	Solvency ratio
11	Leverage

Source: Self elaboration

As shown in table 4 above, it was verified through the correlation matrix described in table 4 the explanatory variables with high correlation, before being applied in the specific methodologies to select the input variables. For the 0.5 threshold it is verified that attributes (1 and 2), (1 and 3), (1 and 4), (1 and 11) are related and it is not recommended for them to be together in the selection of variables. The same applies to the variables (2 and 3), (2 and 4), (3 and 4), (3 and 10), (7 and 8) and finally (8 and 10).

Tab. 5. Correlation Matrix

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1	1										
2	0.622	1.000									

3	0.720	0.692	1.000								
4	0.610	0.524	0.781	1.000							
5	0.215	0.182	0.225	0.127	1.000						
6	0.079	0.100	0.239	0.104	0.048	1.000					
7	0.133	0.117	0.182	0.146	0.028	-0.031	1.000				
8	0.150	0.136	0.301	0.196	0.122	0.103	0.778	1.000			
9	0.074	0.012	0.148	0.074	0.143	0.071	0.115	0.141	1.000		
10	0.386	0.240	0.504	0.419	0.062	0.142	0.425	0.518	0.287	1.000	
11	-0.562	-0.169	-0.340	0.343	-0.023	-0.082	-0.124	-0.155	-0.114	-0.471	1.000

Source: Impressed result from the Matlab

Besides the correlation level between the input variables being verified, the spurious possibility relation between the input and output variables was also seen. In the research, the output variable used for the supervised training process is dichotomous. This has a direct relation with the net equity of the SME, the value one (1) stands for solvent and zero (0) for insolvent.

To avoid artificial cause-effect relations between input and output variables, the input variables 1, 2, 9, 10 and 11 were not used in the supervised learning process because they contain the equity attribute in their formations.

In the Wald test for the logistic regression the p-value statistic is obtained through the comparison between maximum resemblance estimate of the $\widehat{\beta_J})$, and its pattern error estimate. The rate resulted from the $H_0: \beta_j = 0$ hypothesis and has the normal pattern distribution.

$$W_j = {}^{\widehat{\beta_J}}\!\Big/_{DP\widehat{\beta_J}}$$

The p-value is defined as $P(\lfloor Z \rfloor > W_j)$, where Z stands for the random variable of the normal pattern distribution.

The Wald test is used to select the set of the 6 most significant attributes. In table 5 we can verify that variables 3 and 8 reject the null hypothesis of 5 % significance level (Return on Total Assets and Liquidity Ratio). Description of the table: First column - estimated variables; β_j – constants correspond to each estimated variable; $DP\widehat{\beta_J}$ – coefficients pattern error; Wald – for each coefficient to test the null hypothesis that corresponds to coefficient zero against the alternative hypothesis different from zero; pValue – *p*-value for *F*-statistic of hypothesis test corresponds to coefficient equal or not to zero. If the value is higher than 0.05, the variable is not significant at the 5 % of significance level compared with other models' variables.

Tab. 6. Estimated coefficients Logistic Regression

Estimated Variables	$\boldsymbol{\beta_j}$	$DP\widehat{\beta_J}$	**Wald**	pValue

Intercept	0.6641	0.3828	1.7348	0.0827
3	-0.3693	0.0580	-6.3659	1.9417and -10
4	-0.0313	0.0438	-0.7151	0.4745
5	0.0013	0.0011	1.1633	0.2446
6	0.0022	0.0023	0.9437	0.3453
7	0.2214	0.3023	0.7323	0.4639
8	-1.2665	0.5527	-2.2902	0.0220

Source: Result from Matlab software

It is necessary to inspect how the set error varies with the accumulation tree in order to estimate the attributes' importance when the *"tree bagging"* methodology is used for the variable's selection. The estimators' importance can be seen through the random permutation of out-of-bag data, by removing the estimator and verifying the error increase because of its lack. The largest error increment means the estimator is more important.

Initially it is verified how the observation error varies with the increase of the set of trees. An error reduction with the increase in number of trees is expected. In Figure 3 the variation graph of the error with the number of trees is shown. 200 trees were generated and the graph clearly shows the decreased error, which means that *"tree bagging"* process seems appropriate for that purpose.

It is recommended for the classification problems, like it is shown in this study that the minimum size of the end nodes equals to one. In addition, the square-root of the total number of attributes is selected randomly for all division of node decisions.

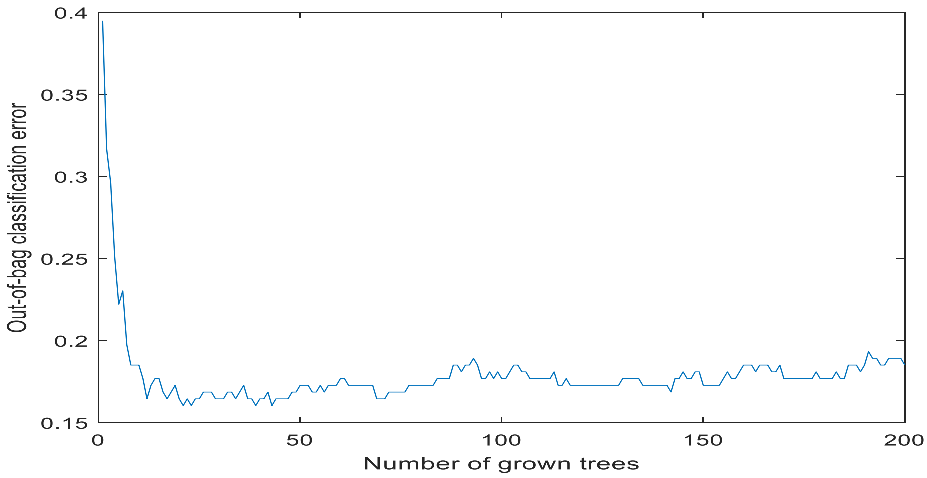

Fig. 3. *Variation of out-of-bag error with the number of created trees, Matlab*

Figure 4 shows the attribute's importance measured for the error classification of the *"out-of-bag"* observations. Because of the data permutation, the increase of the classification error shows the attribute's importance.

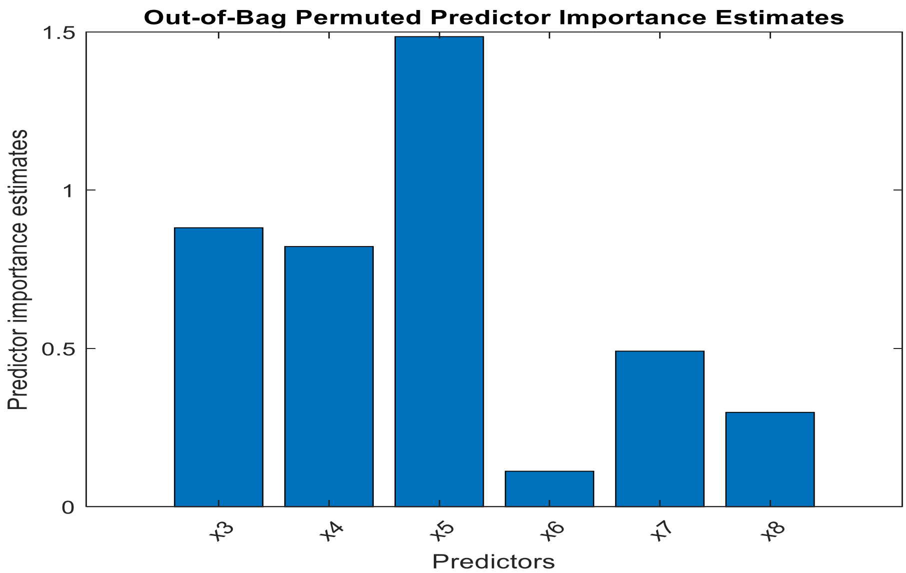

Fig. 4. *Importance of the attribute, measured as an out-of-bag classification error, Matlab*

In the order suggested by the *tree bagging* method, the importance of the six most important variables is 5, 3, 4, 7, 8 and 6. However, the attribute 7 has a strong correlation with attribute 8. Thus, the attribute 7 was excluded from the list.

The process was repeated when five attributes were selected. The result from Figure 5 has confirmed the previous selection.

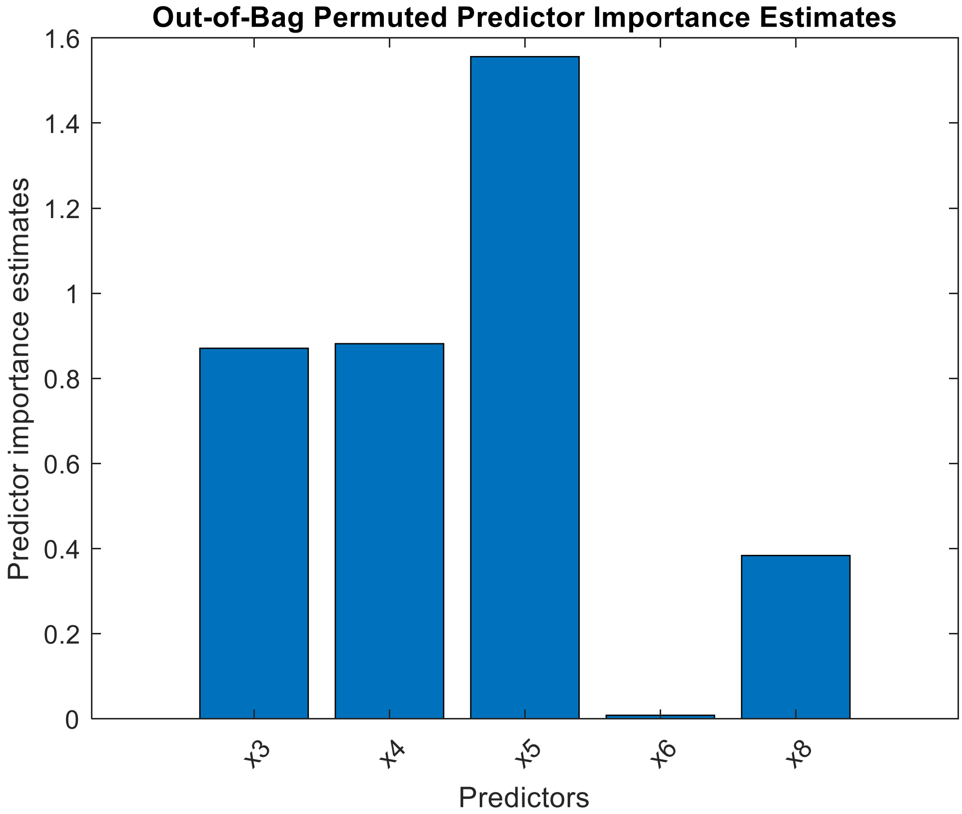

Fig. 5. *Importance of attributes among 5 attributes selected,Matlab*

From the set of five variables, another set of variables was selected. Variable 6 was discarded because it had a very distinct importance. The following step was the testing of four possible combinations with the remaining variables.

The combinations have generated models of three variables, as it is represented in Table 7. Its representation shows the 10-fold crossed validation error as a variable's selection criterion.

Tab. 7. Three tested attributes combination.

Combination attributes	**Crossed validation error**
{3,4,5}	0.1975
{3,4,8}	0.2016
{3,5,8}	0.1893
{8,5,4}	0.1934

From the obtained results, the selected variables are 3, 5, 8 (Return on total assets, Liquidity Ratio, Ability to cover interest) in order to present the smallest validation error.

3.2. Adjustments and Results Evaluation

As a result of the Logistic Regression Model adjustment, the predictor equation is described – insolvency probability for a SME a year in advance:

$$P(Y=1) = {}^{1}/_{(1 - e^{-g(x)})}$$

Where $g(x) = \beta_0 + \beta_1 x_1 + \cdots \beta_j x_j$;

Result from the Logistic Regression adjustment: $g(x) = \beta_0 + 0{,}664 - 0.3693.$ Return on total assets $-$ $0.2665.$ Liquidity ratio.

In the *bagging* methodology, the base of the proposed system is the Decision Tree. Where supervised learning used as input a set of three most important indicators: x_3 = Return on total assets; x_8 = Liquidity ratio; x_5 = Ability to cover interests. For the output for the training process output values 0 and 1 were adopted, which represent the insolvency and solvency type.

A set of 200 trees has been created for the adjustment, with a minimum size of end nodes equal to one. The square root of the attribute's total number for each division of decision nodes was randomly selected. The observation error varied with the increase of the set of trees and it is expected that the error reduces with the number of trees. In Figure 6 the variation graph of the error with the number of trees is shown and it clearly shows the decrease of the error. It means that the adjustment of the *bagging* model was appropriate.

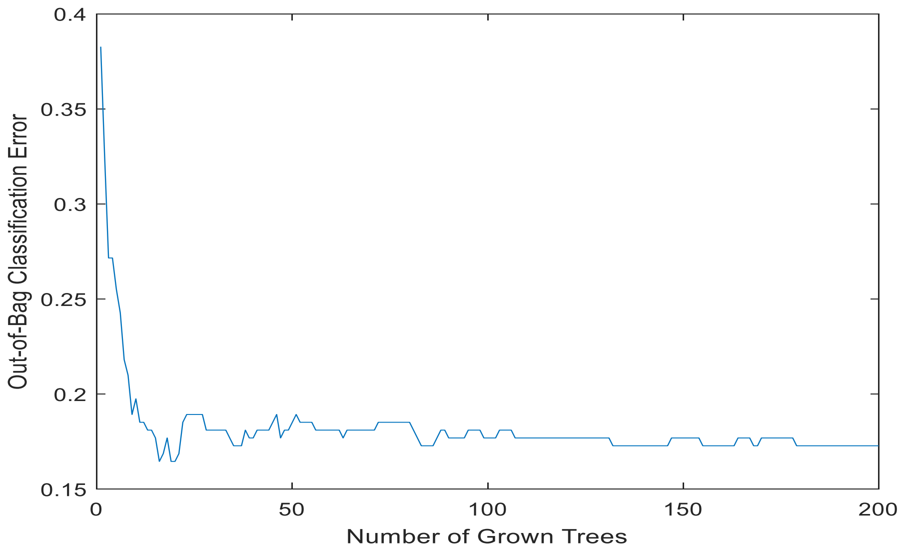

Fig. 6. *Number of bootstrap copies x classification error*

The results were evaluated through the metrics precision, sensitivity and specificity, with the calculation based on the data presented on Confusion Matrix and AUC metric of curve ROC. All the data were extracted from the adjusted model in Matlab platform.

Tables 8 and 9 present the results of the Logistic Regression model's adjustment.

Tab. 8. Confusion Matrix

Insolvent	104	20
Solvent	30	89
	0	1
	Type	Predict

Tab. 9. Metrics for Evaluation

Precision	$\frac{104+89}{243} = 79.4\ \%;$
Sensitivity	$\frac{104}{104+30} = 77.61\ \%$
Specificity	$\frac{89}{89+20} = 81.65\ \%$

Tables 10 and 11 present the results of the Tree-Bagging model's adjustment.
Tab. 10. Confusion Matrix

Insolvent	101	18
Solvent	27	97
	0	1
	Type	Predict

Tab. 11. Metrics for Evaluation

Precision	$\frac{101+97}{243} = 81{,}48\ \%$;
Sensitivity	$\frac{101}{101+27} = 78.91\ \%$
Specificity	$\frac{97}{97+18} = 84.35\ \%$

Tab. 12. Consolidated Results Confusion Matrix and AUC

	Precision	Sensitivity	Specificity	AUC
Logistic regression	79.4 %	77.61 %	81.65 %	0.89
Tree-Bagging	81.48 %	78.91 %	84.35 %	0.92

The metrics presented in table 12 suggest superiority of Tree-*Bagging* model in comparison with the traditional model of Logistic Regression selection. The Precision test presented 81.48% of probability to adjust the forecast state of Portuguese SME insolvency for agro-industrial sector a year in advance, while the traditional model presents 79.4% of probability. The Sensitivity test of the proposed model presented 78.91% of probability to foresee insolvency, given that the SME was insolvent and the traditional presented 77.61%. The specificity test of the proposed model presented the probability of 84.35% to foresee the solvency given that the SME was solvent, while the traditional model presented 81.65%.

The estimate 0.92 of the adjustment test of AUC measure of ROC curve in proposed methodology. The result 0.89 of traditional model points out the superior quality of the adjusted methodology, proposed to foresee insolvency of SME, when the cut point of the sensitivity and specificity measures are changed.

4. CONCLUSION

Estimates of the evaluation measures of the proposed model tests compared to the traditional Logistic Regression model, more specifically the Sensitivity measure, which has a 78.91% probability of predicting insolvent companies when they are insolvent, suggests the validation of the *Tree-Bagging* methodology for forecasting insolvency of Portuguese SME of agro-industrial sector, a year in advance.

When the analysis is improved, the estimates are in accordance with the study of Edmister (1972), which states that with the right financial reasons and by using the discriminant analysis technique one can predict, with anticipation and some reliability, the bankruptcy of a small company.

As a side observation, the selection of the most important model indicators, in order to anticipate the insolvency of SME in the studied sector, suggests the need for effective monitoring of short-time liquidity effects. Additionally, in the long term, it suggests the importance for an appropriate relation between the result generation capacity and the SME investments. The results also suggest the importance of developing studies based on Tree-Bagging methodology for a better understanding of the insolvency phenomenon.

Even though the paper has important practical contributions, we recognize some limitations regarding the methodology, namely the potential bias introduced in the model by ignoring the possible heterogeneity across companies in the various sub-sectors.

REFERENCES

Addo, P. M., Guegan, D. & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. SSRN. https://doi.org/10.2139/ssrn.3155047

Altmam, E.I. (1968). Financial ratios discriminant: analysis and the prediction of corporate bankruptcy. Journal of Finance, 23, 589-609.

Arlot, Sylvain; Celisse, Alain. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79.

Auria, L., Moro, R. A. (2009). Support Vector Machines (SVM) as a Technique for Solvency Analysis. SSRN. https://doi.org/10.2139/ssrn.1424949

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Beaver, W.H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research, 4 (supplement), 71-111.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140https://doi.org/10.1007/BF00058655

Brown, D. R. (2012). A Comparative Analysis of Machine Learning Techniques For Foreclosure Prediction. Nova Southeastern University. Retrieved fromhttps://nsuworks.nova.edu/gscis_etd/105/

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. Journal of Banking & Finance, 72, 218-239.

Chang, Y. C., Chang, K. H., Chu, H. H., & Tong, L. I. (2016). Establishing decision tree-based short-term default credit risk assessment models. Communications in Statistics-Theory and Methods, 45(23), 6803-6815.

Chen, Z., Chen, W., & Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. Expert Systems with Applications, 146, 113155.

Dahiya, S., Handa, S. S., & Singh, N. P. (2017). A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. Expert Systems, 34(6), e12217.

Deng, G. (2016). Analyzing the Risk of Mortgage Default. University of California. Retrieved from https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Grace_Deng_thesis.pdf

Dietterich, T. (2000). An empirical comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Machine Learning, 40(2): 139-157.

Drummond, C.; Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: why undersampling beats over-sampling. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets.

du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. European Journal of Operational Research, 254(1), 236-252.

Edmister, R.O. (1972). An empirical test of financial ratio: analysis for small business failure prediction. Journal of Financial and Quantitative Analysis, 7, 1477-1493.

Ekinci, A., & Erdal, H. İ. (2017). Forecasting bank failure: Base learners, ensembles and hybrid ensembles. Computational Economics, 49(4), 677-686.

Figueiredo, H.M. (2018). 'O problema da recuperação de empresas em Portugal : Analise Crítica', Dissertação Mestrado, Instituto Superior de Contabilidade e Administração de Coimbra https://comum.rcaap.pt/bitstream/10400.26/23121/1/Helena_Figueiredo.pdf

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. Information Fusion, 47, 88-101.

Guo, H., Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the data boost-IM approach. SIGKDD Explorations, 6(1).

Guo, Y., He, J., Xu, L., & Liu, W. (2019). A novel multi-objective particle swarm optimization for comprehensible credit scoring. Soft Computing, 23(18), 9009-9023.

He, Y., Kamath, R. (2005). Bankruptcy prediction of small firms: in individual industries with the help of mixed industry models. Asia-Pacific Journal of Accounting & Economics, 12 (1), 19-36.

Hensher, D.A., Stwart, J. (2007). Forecasting corporate bankruptcy: optimizing the performance of the mixed logit model. Abacus, 43 (3), 241-364.

Hsiao, S.,Whang, T. (2009). A study of financial insolvency prediction model for life insurers. Expert Systems with Applications, 36 (3), 6100-6107.

Huang, J., Wang, H., Kochenberger, G. (2017). Distressed Chinese firm prediction with discretized data. Manag. Decis. 55, 786–807.

Kothari, R., Dong, M. (2001). Decision trees for classification: a review and some new results. Lecture Notes in Pattern Recognition, World Scientific Publishing. p. 241-252.

Lahmiri, S., Bekiros, S., Giakoumelou, A., & Bezzina, F. (2020). Performance assessment of ensemble learning systems in financial data classification. Intelligent Systems in Accounting, Finance and Management, 27(1), 3-9.

Liao, J.J., Shih, C.H., Chen, T.F., Hsu, M.F. (2014). An ensemble-based model for two-class imbalanced financial problem. Econ. Model. 37, 175–183

Nagaraj, K., Sridhar, A.(2015). A predictive system for detection of bankruptcy using machine learning techniques. Int. J. Data Min. Knowl. Manag. Process 5, 29–40

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

Onan, A. (2019). Consensus clustering-based undersampling approach to imbalanced learning. Scientific Programming, 2019.

Pisula, T. (2020). An Ensemble Classifier-Based Scoring Model for Predicting Bankruptcy of Polish Companies in the Podkarpackie Voivodeship. Journal of Risk and Financial Management, 13(2), 37.

Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, p. 81-106.

Sánchez-Medina, A. J., Blázquez-Santana, F., & Alonso, J. B. (2019). Do Auditors Reflect the True Image of the Company Contrary to the Clients' Interests? An Artificial Intelligence Approach. Journal of Business Ethics, 155(2), 529-545.

Sealand, J. C. (2018). Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models. Slippery Rock University. Retrieved from

Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. Cogent Economics & Finance, 8(1), 1729569.

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. Journal of Business, 74, 101-124.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425, 76-91.

Sutton, C.D. (2005). Classification and Regression Trees,Bagging, and Boosting. Elsevier B.V. Handbook of statistics, 24, ISSN: 0169-7161

Tokpavi, H. S. H. C. S. (2018). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. https://www.researchgate.net/publication/318661593

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Systems with Applications, 93, 182-199.

Yao, J., & Lian, C. (2016). A new ensemble model based support vector machine for credit assessing. International Journal of Grid and Distributed Computing, 9(6), 159-168.

Zavgren, C.V. (1985). Assessing the vulnerability of failure of American industrial firms: a logistic analysis. Journal of Business.1, 19-45.

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. Neural Computing and Applications, 28(1), 41-50.