Firms Default Prediction with Machine Learning

Tesi Aliaj[1], Aris Anagnostopoulos[1], and Stefano Piersanti[1,2(✉)]

[1] Sapienza University of Rome, Rome, Italy
tess.aliaj@gmail.com, {aris,piersanti}@diag.uniroma1.it
[2] Bank of Italy - Statistical Data Collection and Processing Directorate, Rome, Italy

Abstract. Academics and practitioners have studied over the years models for predicting firms bankruptcy, using statistical and machine-learning approaches. An earlier sign that a company has financial difficulties and may eventually bankrupt is going in *default*, which, loosely speaking means that the company has been having difficulties in repaying its loans towards the banking system. Firms default status is not technically a failure but is very relevant for bank lending policies and often anticipates the failure of the company. Our study uses, for the first time according to our knowledge, a very large database of granular credit data from the Italian Central Credit Register of Bank of Italy that contain information on all Italian companies' past behavior towards the entire Italian banking system to predict their default using machine-learning techniques. Furthermore, we combine these data with other information regarding companies' public balance sheet data. We find that ensemble techniques and random forest provide the best results, corroborating the findings of Barboza et al. (Expert Syst. Appl., 2017).

1 Introduction

Bankruptcy prediction of a company is, not surprisingly, a topic that has attracted a lot of research in the past decades by multiple disciplines [2,4–6,8–11,13–15,19,20,22,23]. Probably the main importance of such research is in bank lending. Banks need to predict the possibility of default of a potential counterparty before they extend a loan. An effective predictive system can lead to a sounder and profitable lending decisions leading to significant savings for the banks and the companies and, most importantly, to a stable financial banking system. A stable and effective banking system is crucial for financial stability and economic recovery as well highlighted by the recent global financial crisis and European debt crisis. According to Fabio Panetta, general director of the Bank of Italy, referring to Italian loans, "The growth of the new deteriorated bank loans and the slowness of the judicial recovery procedures have determined

a rapid increase in the stock of these assets, which in 2015 reached a peak of 200 billion, equal to 11% of total loans."[1]

Of course, despite the plethora of studies, predicting the failure of a company is a hard task, as demonstrated by the large number of approaches used over time and the results of prediction still not certainly excellent.

Most related research has focused on *bankruptcy* prediction, which takes place when the company officially has the status of being unable to pay its debts (see Sect. 3). However, companies often signal much earlier their financial problems towards the banking system by going in *default.* Informally speaking, a company enters into a default state if it has failed to meet its requirement to repay its loans to the banks and it is very probable that it will not be able to meet his financial commitments in the future (again, see Sect. 3). Entering into a default state is a strong signal of a company's failure: typically banks do not finance a company into such a state and it is correlated with future bankruptcy.

Firms bankruptcy prediction and more generally creditworthiness assessment of the companies can be very important also in *policy decisions*, such as for example the policies of assignement of public guarantee programs [3].

In this paper we use historic data for predicting whether a company will enter in default. We base our analysis on two sets of data. First, we use historic information from *all the loans* obtained by *almost all the companies* based in Italy (totaling to around $800K$ companies). This information includes information on the companies credit dynamics in the past years, as well as past information on relations with banks and on values of protections associated to loans. Second, we combine these data with the balance sheets of $300K$ of these companies (the rest of them are not obliged to produce balance sheets). We apply multiple machine-learning techniques, showing that the future default status can be predicted with reasonable accuracy. Note that the dimensions and the information in our dataset exceeds significantly those of past work [4,5], allowing to obtain a very accurate picture of the possibility to predict over various economic sectors.

Contributions. To summarize the contributions of our paper are:

1. We analyze a very large dataset ($800K$ companies) with highly granular data on the performance of each company over a period of 10 year. To our knowledge, this is the most extensive dataset used in the literature.
2. We use these data to predict whether a company will default in the next year.
3. We combine our data with data available from company balance sheets, showing that we can improve further the accuracy of predictions.

Roadmap. In Sect. 2 we present some related work. In Sect. 3 we provide definitions and we describe the problem that we solve. In Sect. 4 we describe our datasets and the techniques that we use and in Sect. 5 we present our results. We conclude in Sect. 6.[2]

---

[1] Fabio Panetta, Chamber of Deputies, Rome, May 10, 2018.

[2] The views expressed in the article are those of the authors and do not involve the responsibility of the Bank of Italy.

2   Related Work

There has been an enormous amount of work on bankruptcy prediction. Here we present some of the most influential studies.

Initially, scholars focused on making a linear distinction among healthy companies and the ones that will eventually default. Among the most influencing pioneers in this field we can distinguish Altman [2] and Ohlson [20], both of whom made a traditional probabilistic econometric analysis. Altman, essentially defined a score, the Z discriminant score, which depends on several financial ratios (working capital/total assets, retained earnings/total assets, etc.) to asses the financial condition of a company. Ohlson on the other side, is using a linear regression (LR) logit model that estimates the probability of failure of a company. Some papers criticize these methods as unable to classify companies as viable or nonviable [6]. However, both approaches are used, in the majority of the literature, as a benchmark to evaluate more sophisticated methods.

Since these early works there has been a large number of works based on machine-learning techniques [16,18,21]. The most successful have been based on decision trees [12,15,17,24] and neural networks [4,7,11,19,23]. Typically, all these works use different datasets and different sets of features, depending on the dataset. Barboza et al. [5] compare such techniques with support vector machines and ensemble methods showing that ensemble methods and random forests perform the best.

These works mostly try to predict bankruptcy of a company. Our goal is to predict default (see Sect. 3). Furthermore, most of these papers use balance-sheet data (which are public).

Recently, Andini et al. [3] have used data from Italian Central Credit Register to assess the creditworthiness of companies in order to propose an improvement in the effectiveness of the assignment policies of the public guarantee programs.

Our dataset contains credit informations on the past behavior of loan repayment for each single firms of a very large set of companies. To our knowledge, this is the most extensive dataset used in the literature.

3   Firm-Default–Prediction Problem

There are many technical terms used to characterize debtors who are in financial problems: illiquidity, insolvency, default, bankruptcy, and so on. Most of the past research on prediction of failures addresses the concept of *firm bankruptcy*, which is the legal status of a company, in the public registers, that is unable to pay its debts. A firm is in *default* towards a bank, if it is unable to meet its legal obligations towards paying a loan. There are specific quantitative criteria that a bank may use to give a default status to a company.

3.1   Definition of *Adjusted Default Status*

The recent financial crisis has led to a revision and harmonization at international level of the concept of loan default. In general, default is the failure to pay

interest or principal on a loan or security when due. In this paper we consider the classification of *adjusted default status*, which is a classification that the Italian National Central Bank (Bank of Italy) gives to a company that has a problematic debt situation towards the entire banking system. It represents a supervisory concept, whose aim is to extend the default credit status to all the loans of a borrower towards the entire financial system (banks, financial institutions, etc.). The term refers to the concept of the Basel II international accord of *default* of customers. According to this definition, a borrower is defined in default if its credit exposure has became significantly negative. In detail, to asses the status of adjusted default, Bank of Italy considers three types of negative exposures. They are the following, in decreasing order of severity: (1) A *bad (performing) loan* is the most negative classification; (2) an *unlikely to pay* (UTP) loan is a loan for which the bank has high probability to loose money; (3) A loan is *past due* if it is not returned after a significant period past the deadline.

Bank of Italy classifies a company in *adjusted default*, or *adjusted non performing loan* if it has a total amount of loans belonging to the aforementioned three categories exceeding certain pre-established proportionality thresholds [1]. Therefore, a firm's adjusted default classification derives from quantitative criteria and takes into account the company's debt exposure to the entire banking system.

If a company enters into an adjusted-default status then it is typically unable to obtain new loans. Furthermore, such companies are multiple times more likely to bankrupt in the future. For instance, out of the $13K$ companies that were classified in a status of adjusted default in December of 2015, 2160 (16.5%) were no longer active in 2016, having gone bankrupt or being in another similar bad condition. On the other hand, only 2.4% of the companies that were not in adjusted default status became bankrupt.

In this paper we attempt to predict whether a company will obtain an adjusted default status, although for brevity we may call it just default.

4   Data and Methods

In this section we describe the data on which we based our analysis and the machine-learning techniques that we used.

4.1   Dataset Description

Our analysis is based on two datasets. The first and most important in our work is composed of information on loans and the credit of a large sample of Italian companies. The second reports balance sheet data of a large sub-sample of medium-large Italian companies.

Credit data. The first dataset consists of a very large and high granular dataset of credit information about Italian companies belonging to the Italian central credit register (CCR). It is an information system on the debt of the customers

of the banks and financial companies supervised by the Bank of Italy. Bank of Italy collects information on customers' borrowings from the intermediaries and notifies them of the risk position of each customer vis-à-vis the banking system. By means of the CCR the Bank of Italy provides intermediaries with a service intended to improve the quality of the lending of the credit system and ultimately to enhance its stability. The intermediaries report to the Bank of Italy on a monthly basis the total amount of credit due from their customers: data information about loans of 30,000 euro or more and non-performing loans of any amount. The Italian CCR has three main goals: (1) to improve the process of assessing customer creditworthiness, (2) to raise the quality of credit granted by intermediaries, and to (3) strengthen the financial stability of the credit system.

The crucial feature of this database is the high granularity of credit information. It contains information for about $800K$ companies for each quarter of the the period of 2009–2014. The main features are shown in Table 1.

Table 1. Main attributes for the loan (L) and the balance-sheet (B) datasets.

ID	Description	ID	Description
L1	Granted amount of loans	**B1**	Revenues
L2	Used amount of loans	**B2**	ROE
L3	Bank's classification of firm	**B3**	ROA
L4	Average amount of loan used	**B5**	Total turnover
L5	Overdraft	**B6**	Total assets
L6	Margins	**B7**	Financial charges/operating margin
L7	Past due (loans not returned after the deadline)	**B8**	EBITDA
L8	Amount of problematic loans		
L9	Amount of non-performing loans		
L10	Amount of loans protected by a collateral		
L11	Value of the protection		
L12	Amount of forborne credit		

The Balance-sheets dataset. Our second dataset consists of the balance-sheet data of about $300K$ Italian firms. They are generally medium and large companies and they form a subset of the $800K$ companies with loan data. It contains balance-sheet information for each year from 2006 to 2014. The main features include those that regard the profitability of a company, such as return of equity (ROE) and return of assets (ROA); see Table 1 for a more extended list. Typically balance sheet data are public data and have been used extensively for bankruptcy prediction (e.g., see Barboza et al. [5] and references therein).

4.2 Machine-Learning Approaches

As we explain in Sect. 2, the first approaches for assessing the likelihood of companies to fail were based on some fixed scores; see the work by Altman [2]. Current approaches are based on more advanced machine-learning techniques. In this paper we follow the literature [5] by considering a set of diverse machine-learning approaches for predicting loan defaults.

In the first test we used five well-known machine-learning approaches. We provide a brief description of each of them, as provided by Wikipedia.

Decision Tree (DT): One of the most popular tool in decision analysis and also in Machine Learning. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Random Forest (RF): Random forest are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. Random decision forests correct for decision trees' habit of overfitting to their training set.

Bagging (BAG): Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging was proposed by Leo Breiman in 1994 to improve classification by combining classifications of randomly generated training sets.

AdaBoost (ADA): AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, in 2003. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier.

Gradient boosting (GB): Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction.

Except for these standard techniques, we also combined the various classifiers in the following way. After learning two versions of each classifier, one with the default parameters of the `Python scikit` implementation and one with optimal parameters (to this end, we have used an exhaustive search over specified parameter values for each classifier, using the `sklearn.model.selection.GridSearchCV`), we execute all of them (10 in total) and if at least 3 classifiers predict that a firm will default then the classifier predicts default for that firm. The number 3 was chosen after experimentation. We call this ensemble approach **COMB**.

5   Experimental Results

The main goal of our study is to evaluate the extent to which we can predict whether a company will enter in a default state using data from past years. In particular, our goal is to predict whether a company that by December 2014 is not in default, will enter in default during one of the four trimesters of 2015. To do the prediction, we initially used data from the period of 2006–2014; however we noticed that using loan data running earlier than five trimesters before 2015 did not help. Therefore, for all the experimental results that we report here, we use the loan data from the last quarter of 2013 plus those from the four quarter of 2014 and the entire balance-sheet dataset from 2006 to 2014.

5.1   Evaluation Measures

We use a variety of evaluation measures to assess the effectiveness of our classifiers, which we briefly define. As usually, in a binary classification context, we use the standard concepts of true positive (TP), false positive (FP), true negative (TN), false negative (FN): For instance, FN is the number of firms that

	Predicted Default	Predicted Not Default
Default	**TP**	**FN**
Did not default	**FP**	**TN**

defaulted during 2015 but the classifier predicted that they will not default.

We now define the measures that we use:

- Precision: $\mathbf{Pr} = \dfrac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$
- Recall: $\mathbf{Re} = \dfrac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$
- F1-score: $\mathbf{F1} = 2 \cdot \dfrac{\mathbf{Pr} \cdot \mathbf{Re}}{\mathbf{Pr} + \mathbf{Re}}$
- Type-I Error: $\mathbf{Type\text{-}I} = \dfrac{\mathbf{FN}}{\mathbf{TP} + \mathbf{FN}}$
- Type-II Error: $\mathbf{Type\text{-}II} = \dfrac{\mathbf{FP}}{\mathbf{TN} + \mathbf{FP}}$
- Balanced Accuracy: $\mathbf{BACC} = 2 \cdot \frac{\mathbf{TP \cdot TN}}{\mathbf{TP + TN}}$

5.2   Datasets

In Sect. 4.1 we already described the datasets that we use. We perform two families of experiments. In the first one, we use only the loan data (as typically performed by Bank of Italy) to assess the probability of default. Then, we also combine this information with balance-sheet data. We have balance-sheet data

available for 300K (out of 800K) companies. We decided to limit our study to these 300 K companies, as this allows us to compare the results in the case that we use only loan data and in the case that we use both loan and balanced-sheet data.

5.3   Balanced Versus Imbalanced Classes

The classification problem that we deal is very imbalanced: around 4.3% of the firms were in a default state in 2015. Therefore, as performed in prior work [5] we consider two cases. First we use the entire dataset, second we also create a balanced version by selecting all the firms that defaulted ($13.2K$) and an equal number of random firms that did not default, creating in this way a down-sampled balanced dataset of $26.4K$ firms.

5.4   Baselines

We evaluated the techniques presented in Sect. 4.2. To assess their effectiveness, we compare them with three basic approaches. The first one is a simple multino-mial Naïve Bayes (**MNB**) classifier. The second is a logistic regression (**LOG**) classifier. Finally, we created the following simple approach. We first measured the correlation of each feature with the target variables (refer to Table 1). We found the most significant ones (i.e., the ones that are mostly correlated with the target variable) are L3 (a bank's classification of the firm) and L7 (amount of loans not repaid after the deadline) for the loan dataset, and B2 (ROE) and B3 (ROA) for the balance sheet dataset. Then we built the simple classifier that outputs *default* if at least one of L3 or L7 are nonzero and *not default* otherwise for the loan dataset. We call this baseline **NAIVE**.

   We gather the classification approaches that we use in Table 2.

Table 2. Baselines and classification algorithms.

ID method	Description
NAIVE	Naive classifier based on features correlation with target
MNB	Multinomial Bayesian classifier
LOG	Logistic Regression
GB	Gradient Boosting
RF	Random Forest
DT	Decision Tree
BAG	Bagging
ADA	AdaBoost
COMB	Combined method based on multiple classifiers

5.5 Prediction of Adjusted Default

We are now ready to predict whether companies will enter into an adjusted default state, as we explained in Sect. 3.1.

First we present the results for the original, imbalanced dataset. In Table 3 we present the results when we use only the loan dataset, whereas in Table 4 we present the results when we also use the balance-sheet data. The first finding is that the evaluation scores are rather low. This is in accordance to all prior work, indicating the difficulty of the problem. We observe that the machine-learning approaches perform better than the baselines, and the various algorithms trade off differently over the various evaluation measures. Random forests perform particularly well (in accordance with the findings of Barboza et al. [5]) and our combined approach (**COMB**) is able to trade off between precision and recall and give an overall good classification. Comparing Table 3 with Table 4 we see that the additional information provided by the balance-sheet data helps to improve the classification.

Table 3. Imbalanced training set; loan data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.25	0.11	0.16	0.89	0.04	0.54
MNB	0.95	0.05	0.09	0.95	0.02	0.52
LOG	0.44	0.01	0.02	0.99	0.01	0.50
GB	0.63	0.22	0.33	0.78	0.01	0.61
RF	0.61	0.21	0.31	0.79	0.01	0.60
DT	0.27	0.29	0.28	0.71	0.03	0.63
BAG	0.53	0.19	0.28	0.81	0.01	0.59
ADA	0.56	0.20	0.30	0.80	0.01	0.60
COMB	0.52	0.32	0.40	0.68	0.01	0.66

In Tables 5 and 6 we present the results for the balanced dataset. There are some interesting findings here as well. First, as expected the classification accuracy improves (similarly to [5]). Second, we notice that the **NAIVE** classifier performs well (expected, as feature L3 takes into account several factors of the company's behavior); however the type-II error is high. Overall, **COMB** approach remains the best performer.

Table 4. Imbalanced training set; loan and balance-sheet data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.29	0.14	0.20	0.89	0.06	0.55
MNB	0.95	0.06	0.09	0.95	0.03	0.52
LOG	0.46	0.02	0.03	0.99	0.02	0.50
GB	0.63	0.23	0.34	0.77	0.01	0.61
RF	0.68	0.25	0.37	0.75	0.01	0.62
DT	0.28	0.32	0.30	0.68	0.04	0.64
BAG	0.59	0.21	0.31	0.79	0.01	0.60
ADA	0.61	0.26	0.36	0.74	0.01	0.63
COMB	0.55	0.36	0.43	0.64	0.01	0.67

Table 5. Balanced training set; loan data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.24	0.78	0.37	0.28	0.50	0.62
MNB	0.43	0.08	0.14	0.88	0.03	0.51
LOG	0.36	0.21	0.26	0.79	0.03	0.59
GB	0.23	0.67	0.34	0.33	0.10	0.78
RF	0.16	0.73	0.26	0.27	0.17	0.78
DT	0.10	0.69	0.17	0.31	0.30	0.69
BAG	0.16	0.69	0.25	0.31	0.17	0.76
ADA	0.24	0.65	0.35	0.35	0.09	0.78
COMB	0.20	0.69	0.31	0.31	0.13	0.78

Table 6. Balanced training set; loan and balance-sheet data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.25	0.77	0.38	0.23	0.49	0.64
MNB	0.44	0.09	0.15	0.91	0.03	0.53
LOG	0.36	0.22	0.28	0.78	0.03	0.60
GB	0.19	0.78	0.30	0.22	0.15	0.81
RF	0.18	0.80	0.30	0.20	0.16	0.82
DT	0.10	0.71	0.18	0.29	0.26	0.72
BAG	0.17	0.75	0.27	0.25	0.17	0.79
ADA	0.18	0.76	0.29	0.24	0.16	0.80
COMB	0.19	0.84	0.31	0.16	0.16	0.84

5.6   A Practical Application: Probability of Default for Loan Subgroups

We now see an application of our classifier in an applied problem faced by Bank of Italy. We compare the best performing classifier (**COMB**) with a method commonly used to estimate the probability of one-year default by companies at aggregate level.

Consider a segmentation of all the companies (e.g., according to economic sector, geographical area, etc.). Often there is the need to estimate the *probability of default* (PD) of a loan in a given segment. A very simple approach, which is actually used in practice, is to simply take the ratio of the companies in the segment that went into default at year $T + 1$ over all the companies that were not in default in year T. We use this method as a baseline.

We now consider a second approach based on our classifier, which we call **COMB**. We estimate the PD by considering the amount of companies in the segment that are expected (using the **COMB** classifier) to go into default at year $T + 1$ compared to the total loans existing for the segment at the time T.

We use two different segmentations. A coarse one, in which the segments are defined by the economic sector (e.g. mineral extraction, manufacturing), and a finer one, which is defined by the combination of the economic sector and the geographic area, as defined by a value similar to the company's zip code.

In Table 7 we compare the two approaches for estimating the PD. As expected, in both segmentations the classifier-based approach is a winner, with the improvement being larger for the finer segmentation. In many cases the two approaches give the same result, typically because in these cases there are no companies that fail (PD equals 0).

Table 7. Comparison of the standard approach to estimate PD with the classifier-based one. "Mean error" is the average error between the predicted PD value and the real one. "Var error" is the variance of the error. "Superiority percentage" is the percentage of segments in which the predictor is better than the other; in the remaining ones we have the same performance.

Coarse segmentation			Fine segmentation		
	Baseline	COMB		Baseline	COMB
Mean error	0.11	0.048	Mean error	0.088	0.036
Var error	0.056	0.016	Var error	0.06	0.025
Superiority percentage	25.1%	45.6%	Superiority percentage	6.1%	19.5%

6   Conclusion

Business-failure prediction is a very important topic of study for economic analysis and the regular functioning of the financial system. Moreover the importance of this issue has greatly increased following the recent financial crisis. There have

been many recent studies that have tried to predict the failure of companies using various machine-learning techniques.

In our study, we used for the first time credit information from the Italian Central Credit Register to predict the banking default of Italian companies, using Machine Learning techniques. We analyzed a very large dataset containing information about *almost all the loans of all the Italian companies*. Our first findings is that, as in the case of bankruptcy prediction, machine-learning approaches are able to outperform significantly simpler statistical approaches. Moreover, combining classifiers of different type can lead to even better results. Finally, using information on past loan data is crucial, but the additional use of balance-sheet data can improve classification even further.

We show that the combined use of loan data with balanced-sheet data leads to improved performance for predicting default. We conjecture that using loan data in the prediction of bankruptcy (where, typically, only balance-sheet data are being used) can improve further the performance.

Nevertheless, prediction remains an extremely hard problem. Yet, even slight improvement in the performance, can lead to savings of multiple hundreds of euros for the banking system. Thus our goal is to improve classification even further by combining our approaches with further techniques, such as neural-network based ones. Some preliminary results in which we use only neural networks are encouraging, even though are worse than the results we report here.

References

1. Methods and Sources: Methodological notes (2018). available on the website of the Banca d'Italia. https://www.bancaditalia.it/pubblicazioni/condizioni-rischiosita/en_STACORIS_note-met.pdf?language_id=1
2. Altman, E.: Predicting financial distress of companies: revisiting the z-score and zeta. In: Handbook of Research Methods and Applications in Empirical Finance, vol. 5 (2000)
3. Andini, M., Boldrini, M., Ciani, E., de Blasio, G., D'Ignazio, A., Paladini, A.: Machine learning in the service of policy targeting: the case of public credit guarantees, vol. 1206 (2019). https://www.bancaditalia.it/pubblicazioni/temi-discussione/2019/2019-1206/en_tema_1206.pdf
4. Atiya, A.: Bankruptcy prediction for credit risk using neural networks: a survey and new results. IEEE Trans. Neural Netw. **12**, 929–935 (2001). https://doi.org/10.1109/72.935101
5. Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. Expert Syst. Appl. **83**(C), 405–417 (2017). https://doi.org/10.1016/j.eswa.2017.04.006
6. Begley, J., Ming, J., Watts, S.: Bankruptcy classification errors in the 1980s: an empirical analysis of Altman's and Ohlson's models. Rev. Acc. Stud. **1**, 267–284 (1996)
7. Boritz, J., Kennedy, D., Albuquerque, A.D.M.E.: Predicting corporate failure using a neural network approach. Intell. Syst. Account. Finance Manag. **4**(2), 95–111 (1995). https://doi.org/10.1002/j.1099-1174.1995.tb00083.x

8. Chen, M.Y.: Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Comput. Math. Appl. **62**(12), 4514–4524 (2011). https://doi.org/10.1016/j.camwa.2011.10.030
9. Cho, S., Hong, H., Ha, B.C.: A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. Expert Syst. Appl. **37**(4), 3482–3488 (2010). http://www.sciencedirect.com/science/article/pii/S0957417409009063
10. Erdogan, B.: Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. J. Stat. Comput. Simul. - J STAT COMPUT SIM **83**, 1–13 (2012)
11. Fernández, E., Olmeda, I.: Bankruptcy prediction with artificial neural networks. In: Mira, J., Sandoval, F. (eds.) IWANN 1995. LNCS, vol. 930, pp. 1142–1146. Springer, Heidelberg (1995). https://doi.org/10.1007/3-540-59497-3_296
12. Gepp, A., Kumar, K.: Predicting financial distress: a comparison of survival analysis and decision tree techniques. Procedia Comput. Sci. **54**, 396–404 (2015)
13. Kumar, P.R., Ravi, V.: Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. Eur. J. Oper. Res. **180**(1), 1–28 (2007). https://doi.org/10.1016/j.ejor.2006.08.043
14. Lee, S., Choi, W.S.: A multi-industry bankruptcy prediction model using backpropagation neural network and multivariate discriminant analysis. Expert Syst. Appl. **40**(8), 2941–2946 (2013). https://doi.org/10.1016/j.eswa.2012.12.009
15. Lee, W.C.: Genetic programming decision tree for bankruptcy prediction. In: 9th Joint International Conference on Information Sciences (JCIS 2006). Atlantis Press (2006)
16. Lin, W.Y., Hu, Y.H., Tsai, C.F.: Machine learning in financial crisis prediction: a survey. IEEE Trans. Syst. Man Cybern. - TSMC **42**, 421–436 (2012). https://doi.org/10.1109/TSMCC.2011.2170420
17. Martinelli, E., de Carvalho, A., Rezende, S., Matias, A.: Rules extractions from banks' bankrupt data using connectionist and symbolic learning algorithms. In: Proceedings of Computational Finance Conference (1999)
18. Nanni, L., Lumini, A.: An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Syst. Appl. **36**(2), 3028–3033 (2009). https://doi.org/10.1016/j.eswa.2008.01.018
19. Odom, M., Sharda, R.: A neural network model for bankruptcy prediction. In: Proceedings of the 1990 IJCNN International Joint Conference on Neural Networks, vol. 2, pp. 163–168 (1990)
20. Ohlson, J.A.: Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. **18**(1), 109–131 (1980)
21. Sarojini Devi, S., Radhika, Y.: A survey on machine learning and statistical techniques in bankruptcy prediction. Int. J. Mach. Learn. Comput. **8**, 133–139 (2018). https://doi.org/10.18178/ijmlc.2018.8.2.676
22. Wang, G., Ma, J., Yang, S.: An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Syst. Appl. **41**(5), 2353–2361 (2014). http://www.sciencedirect.com/science/article/pii/S0957417413007872
23. Wang, N.: Bankruptcy prediction using machine learning. J. Math. Finance **07**, 908–918 (2017). https://doi.org/10.4236/jmf.2017.74049
24. Zhou, L., Wang, H.: Loan default prediction on large imbalanced data using random forests. TELKOMNIKA Indones. J. Electr. Eng. **10**, 1519–1525 (2012)