POLITECNICO DI TORINO
Repository ISTITUZIONALE

A Machine Learning-based DSS for mid and long-term company crisis prediction

(Article begins on next page)

24 April 2024
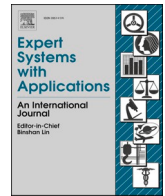
A Machine Learning-based DSS for mid and long-term company crisis prediction

Guido Perboli [b,a], Ehsan Arabnezhad [b,a]

[a] *ARISK s.r.l., Milan, Italy*
[b] *ICT for City Logistics and Enterprises Center, Politecnico di Torino, Turin, Italy*

A B S T R A C T

In the field of detection and prediction of company defaults and bankruptcy, significant effort has been devoted to evaluating financial ratios as predictors using statistical models and machine learning techniques. This problem becomes crucially important when financial decision-makers are provided with predictions on which to act, based on the output of prediction models. However, research has shown that such predictors are sufficiently accurate in the short-term, with the focus mainly directed towards large and medium-large companies. In contrast, in this paper, we focus on mid- and long-term bankruptcy prediction (up to 60 months) targeting small and/or medium enterprises. The key contribution of this study is a substantial improvement of the prediction accuracy in the short-term (12 months) using machine learning techniques, compared to the state-of-the-art, while also making accurate mid- and long-term predictions (measure of the area under the ROC curve of 0.88 with a 60 month prediction horizon). Extensive computational tests on the entire set of companies in Italy highlight the efficiency and accuracy of the developed method, as well as demonstrating the possibility of using it as a tool for the development of strategies and policies for entire economic systems. Considering the recent COVID-19 pandemic, we show how our method can be used as a viable tool for large-scale policy-making.

1. Introduction

Due to the huge impact of company crises on the economy and society (even global debt), they have long been an interesting topic of investigation–even more so at present–in order to accurately study the consequences of bankruptcy and to find ways to avoid it. To reduce the effects of such crises, companies may apply for economic help/funds from financial institutions, while decision-makers in the financial system try to identify those companies that are highly likely to declare bankruptcy in the future. For this reason, company crisis/bankruptcy prediction aims to assess the financial health and future performance of a company. The existing literature has mainly focused on financial aspects, obtaining results with a good prediction rate in the short-term; normally 12 months (Altman, 2014; Altman, Iwanicz-Drozdowska, Laitinen, & Suvas, 2016). However, due to different regulations, these methods tend to be more accurate for large and medium–large companies. This is true, in particular, in countries where the number of the Small and Medium Enterprises (SMEs) is high and those characterized by a plethora of small companies (Altman, Esentato, & Sabato, 2020; Altman, Danovi, & Falini, 2013).

This paper contributes to the literature along two lines. First, we

introduce a state-of-the-art insolvency prediction model as a decision support system (DSS) based on machine learning. One of the novelties of our work consists of a two-round tuning algorithm for the machine learning module when the data set is highly unbalanced. In the first round, outperforming companies are chosen (by threshold), in order to make the data set more robust and to distinguish between the companies that will face a crisis and those that will not, with an improvement of up to 11% in accuracy, with respect to previous works Son, Hyun, Phan, and Hwang (2019). The machine learning module was tuned using financial statement data of more than 160,000 Italian SMEs that were operational by the end of 2018, combined with the data of approximately 3,000 bankrupted companies, covering the period of 2001–2018. Extensive computational results demonstrate the accuracy of our method, not only in the short-term (12 months), but also in the medium (36 months) and long (up to 60 months)-term. Second, we illustrate how our system can be used by company owners and decision-makers as a viable strategic tool. We applied our DSS to two different settings: The Italian SME system before the COVID-19 pandemic and the post-COVID-19 economy, using the DSS to evaluate the financial policies of the Italian government and testing different variants of the policies on the total set of SMEs in the Piedmont region.

E-mail addresses: guido.perboli@polito.it (G. Perboli), ehsan.arabnezhad@polito.it (E. Arabnezhad).

This paper is organized as follows: In Section 2, we explore the literature in this field, highlighting the main gaps. Section 3 is devoted to presenting the whole DSS, while Section 4 describes the data and the machine learning module; whose performances are discussed in Section 5. Its application to the Italian SME system is discussed in Section 6, including the use of the DSS in the post-COVID-19 economic crisis. Finally, Section 7 summarizes the results and presents possible future research directions.

2. Literature review

Financial institutions, fund managers, lenders, governments, and financial market players started to develop models to efficiently assess the likelihood of company default almost a century ago when, in 1932, Patrick (1932) performed a multivariate analysis on 20 companies (Patrick, 1932). Meanwhile, researchers and practitioners have developed several quantitative approaches. In 1966, Beaver (1966) applied a *t*-test to obtain the significance of each ratio for each company (Beaver, 1966). Altman (1968) used multiple discriminant analysis (MDA); however, the false statistical assumptions underlying the MDA approach led researchers to concentrate their efforts on the development of conditional probability models (logistic regression) based on data sets (Ohlson, 1980). Herein, we analyze the literature along three axes: The prediction method type, the horizon of the prediction, and the data types that are incorporated into the prediction model.

2.1. Prediction methods

Traditional methods rely on statistical models. Altman (1968) generated a score, by which observations could be classified into good and bad payers (Altman, 1968). Following his work, other applications have been developed by specializing the model to specific sectors and segmentations (Altman et al., 2020; Altman et al., 2016; Altman, 2014). In contrast to Altman (1968), Ohlson (1980) was one of the first researchers to apply logistic regression analysis for default estimation (Ohlson, 1980). Ohlson (1980)'s model determines the default probability of a potential borrower. Several subsequent studies have sought to perform similar tests, thanks to the relative ease of carrying out discriminant analysis and logistic regression (Hillegeist, Keating, Cram, & Lundstedt, 2004; Upneja & Dalbor, 2001; Chen, Chollete, & Ray, 2010). The advantages of these models are twofold: first, their ability to derive an analysis of the certainty (probability) of the results and, second, to evaluate the effect of each feature individually. Despite their wide adoption in both research and industry, these classes of models have become inaccurate and the need for enhancements in the modeling of default risk has been suggested (Begley, Ming, & Watts, 1996). Additionally, in order to be accurate, they require tuning with respect to different markets (e.g., the parameters need tuning differently for industry and services). In addition, they have limited ability to enhance the predictive results, normally not over 12 months (Altman, 2014; Altman et al., 2016). Moreover, they cannot be automatically incorporated into large time-series data and rely on the standard mean-value theory; however, for the most part, extreme events are the key factors and, therefore, extreme-value theory might provide better insight (Baldi, Manerba, Perboli, & Tadei, 2019; Perboli, Tadei, & Gobbato, 2014).

To overcome the limitations of statistical models, studies that use pattern recognition methods have been actively developed in the field of machine learning (Linden, 2015; Barboza, Kimura, & Altman, 2017), demonstrating how machine learning models can outperform traditional classification methods. Some of these works have made use of artificial intelligence systems, such as neural networks and genetic algorithms (Odom & Sharda, 1990; Coats & Fant, 1993; Boritz, Kennedy, & Albuquerque, 1995). Several new works have also demonstrated the power of ensemble models to deal with imbalanced data sets (Brown & Mues, 2012; Kim, Kang, & Kim, 2015). The difference between parametric and

Table 1
Summary of machine learning models.

Algorithm family	Linear	Accurate	Easy to interpret	Scalable	Algorithms
Linear models	Yes	No	Yes	Yes	Linear regression and logistic regression
Basic models	Possibly	No	Yes	Possibly	Naïve Bayes, DT, and KNN
Ensemble models	No	Yes	Yes	Yes	RF, AdaBoost, and gradient boosting
SVM	No	Yes	No	No	SVM
Deep learning	No	Yes	No	Yes	MLP classifier and MLP regressor

non-parametric methods to analyze the credit risk of SMEs, in particular, have been discussed in depth by Figini, Bonelli, and Giovannini (2017), where multivariate outlier detection techniques have been used to enhance the results (Figini et al., 2017). In an interesting case study of neural networks, Brédart (2014) used a limited number of features/ratios on Belgian SMEs, improving upon the performance of previous works (Shah & Murtaza, 2000; Becerra, Galvão, & Abou-Seada, 2005).

As shown in the summary in Table 1, ensemble methods (i.e., bagging Breiman (1996), boosting Freund, Schapire, & Abe (1999), and stacking) generally outperform the other methods. Gradient boosting is a powerful ensemble method which has recently gained attention from researchers regarding company insolvency, for which it has been shown to be one of the best methods (Friedman, 2001). Gradient boosting is an additive model that operates on weak learners (e.g., decision trees), until the model no longer improves the results based on a loss function.

To the best of our knowledge, at present, the best results have been obtained in Son et al. (2019), where the authors applied XGBoost to a data set audited by a Korean credit rating agency. Despite the good accuracy, the main problems are that it is difficult to understand the prediction capacity in the mid-term (over 24 months)—which is a classic problem in all models considered as providing a financial rating—and that the accuracy is highly dependent on external factors, such as the presence of a regulation that obliges companies to undergo an external audit (which is compulsory only for a subset of SMEs and is dependent on the country's regulations).

2.2. Time horizon of the predictions

Traditional models are generally accurate up to 12 months, with some cases in which the prediction can maintain sufficient accuracy (approximately 70%) for up to 24 months (Altman, 2014; Altman et al., 2016; Altman et al., 2013; Hillegeist et al., 2004; Upneja & Dalbor, 2001; Chen et al., 2010; Altman et al., 2020). Even in the case of machine learning, good prediction results are typically only obtained when considering the short-term (Son et al., 2019; Barboza et al., 2017). An ideal prediction model should be able to make mid-term forecasts. In fact, many studies have shown how failure process symptoms can be traced back to 5–8 years before the failure actually occurred (Argenti, 1976; Hambrick & D'Aveni, 1988; Luoma & Laitinen, 1991; Ooghe & Prijcker, 2008). Thus, there exists a need for studies with a longer horizon than a few months.

2.3. Data dimensionality and predictors tuning

In the field of bankruptcy prediction, a great attention must be paid to the so called the curse of dimensionality. Normally a large number of indicators are usually involved and the training data are insufficient to cover the decision space. The feature selection (or feature reduction) problem addresses this issue by removing irrelevant, redundant and correlated features, improving the accuracy and the compactness of the
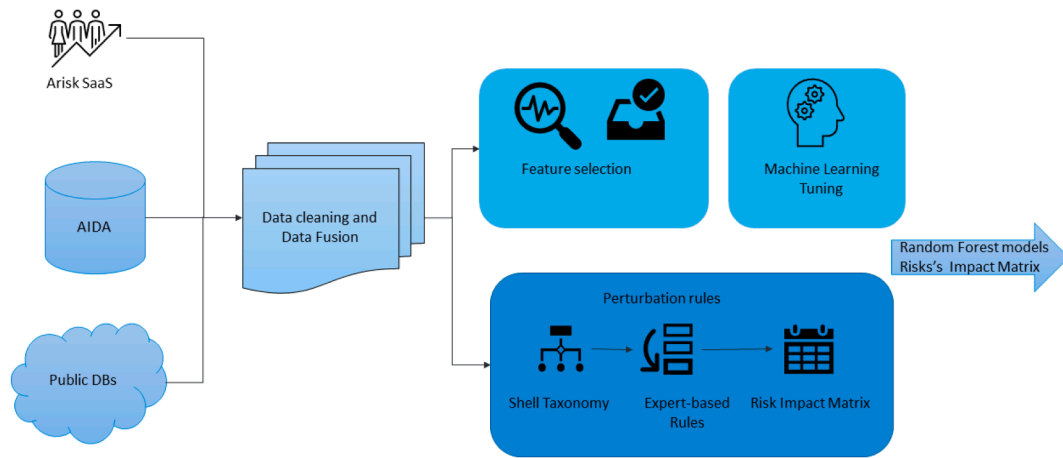
Fig. 1. Decision support system – system training and tuning.

classification models, decreasing their computational effort, and facilitating their adoption. Feature selection can be modeled as a combinatorial optimization problem. It can be solved by a greedy approach, as in our case. When the results are not satisfactory, more complex methods can be used. The most promising methods are based on evolutionary algorithms, due to their flexibility and capacity to properly span the potentially enormous solution space given by the combinations of the features (Chen, Ribeiro, Vieira, Duarte, & Neves, 2011; Chen, 2011; Pellegrino, Perboli, & Squillero, 2019). A similar optimization problem arising in machine learning and deep learning applications is the optimal setting of the predictor parameters. In fact, the final performances are strictly related to that aspect (the so-called hyper parameter optimization or hyper parameter tuning problem). These methods include grid search, random grid search, Bayesian model-based optimization, and evolutionary algorithms (Feurer & Hutter, 2019; Chou, Hsieh, & Qiu, 2017; Chen, 2011). Approaches as random grid search give good results when a high computational power is available (in our case, a High Performance Computing facility), while evolutionary

algorithms showed a better performances when it was difficult to reduce a priori the search space and a larger solution space needed to be explored. Unfortunately, the mostly part of these approaches have been tested on limited data sets or on short-term predictions (12–18 months).

2.4. Data incorporated in the prediction model

The general improvement over time in the accuracy of traditional models should be linked to the selection of ratios and indices to be included in the statistical model. However, as highlighted by Balcaen and Ooghe (2006), who reviewed business failure studies over the last 35 years, there has been little consensus on which variables are the best in discriminating between failed and non-failed firms. Moreover, most of the literature has focused on financial data, disregarding non-financial data. For a detailed discussion about this topic, the reader should refer to the recent paper by Altman et al. (2016), in which a deep review of the topic has been given (Altman et al., 2016). The literature has shown, in any case, that the introduction of non-financial data can improve the
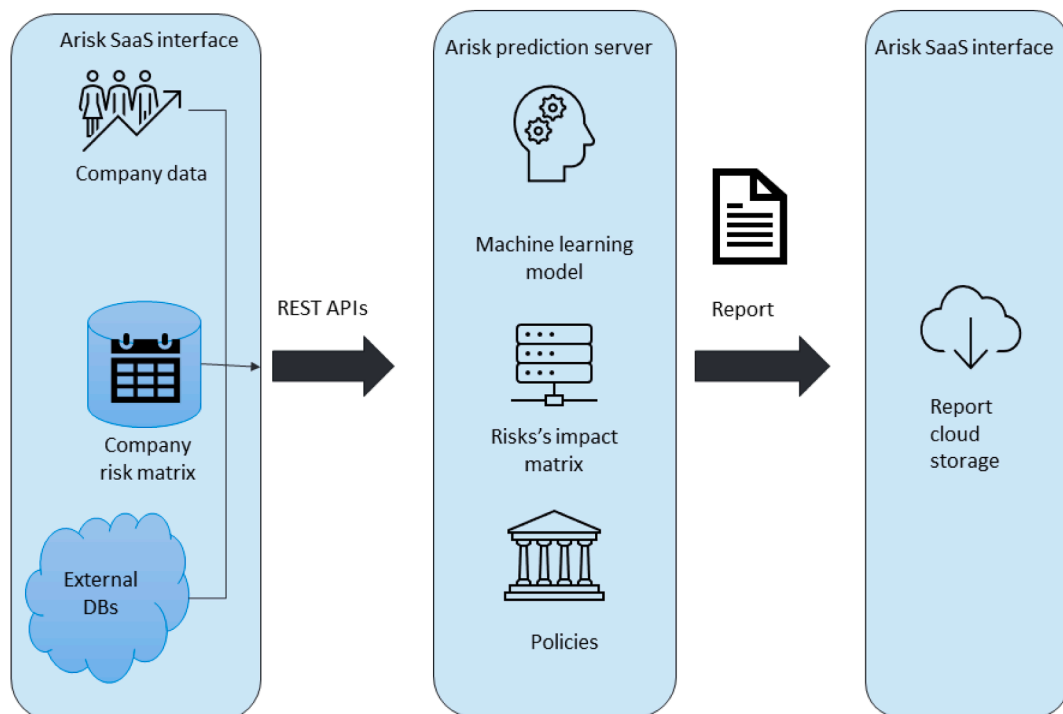

Fig. 2. Decision support system – machine learning prediction module.

performance and time horizon of both traditional and machine learning models (Altman et al., 2016; Son et al., 2019). Unfortunately, to date, relevant studies have tried either to determine whether there is a correlation between bankruptcy and non-financial variables, as done by Altman et al. (2016), or have just added one or two variables related to the organization of the company (normally the industry type and the presence of an external audit, such as in Son et al. (2019)).

2.5. Literature gaps and paper contribution

From our analysis of the relevant literature, a gap between the best practice available methods and the market needs became evident. In fact, there is no model in the literature or in the market that is accurate both in the short (one/two years) and in the mid-term (up to five years), which is adaptable to different markets with a standard (and possibly automated) tuning and able to incorporate and analyze the effects of both financial and non-financial variables. In this paper, we attempt to address these needs by introducing a machine learning-based DSS which is capable of providing accurate predictions both in the short- and mid-term; as well as a new method for the tuning of machine learning methods in the case of unbalanced data, which can improve their overall performance.

3. The decision support system

Our Decision support system (DSS) considers, but is not limited to, financial data. It can collect, catalog and incorporate several types of risks. The present version collects info related to budget and financial data, company organization data, family risk matrices related to cash flows, supply chain management and …The overall DSS structure is shown in Figs. 1 and 2. The DSS is developed by ARISK, a fin-tech spin-off of Politecnico di Torino providing business interruption prediction services to SMEs, It is split into two different sections: a training and tuning module and a prediction server.

The training and tuning module (see Fig. 1) collects data from public databases as public financial data (in Italy the Italian Camera di Commercio), a set of indexes and ratios from AIDA Bureau van Dick, as well as, whether available, data from the proprietary interface by ARISK to collect additional data. Then the data are cleaned, normalized and merged. Data are thus split between core and non-core sets. The core data represents the features of the machine learning module, while the non-core data are data that are not directly incorporated in the machine learning. An example can be qualitative data coming from specific industrial sectors.

Core data are then managed by the machine learning pipeline for reducing the features first (feature selection procedure) and then the Machine Learning algorithm is chosen and tuned. At the moment, our system considers a wide set of Machine Learning systems, including Random Forest, XGBoost, Logistic regression, and Neural Networks. The outputs of this pipeline are the binary files of the predictors then passed to the prediction module.

Non-core data are considered as secondary data that are not directly incorporated in the Machine Learning predictor, but whose effects are simulated as perturbations of the Machine Learning features. This is done by a specific pipeline. The non-core data are first classified by a tree-based taxonomy, based on the SHELL-based taxonomy by Cantamessa, Gatteschi, Perboli, and Rosano (2018). The methodology adopted for the analysis of startup failure is based on the SHELL model, originally implemented to classify aviation accidents and errors, and here adapted to the entrepreneurship sector. The SHELL model, whose name derives from the initial letters of its components, Software, Hardware, Environment, Liveware People and Liveware Environment, was developed by Hawkins in 1975 basing on the original work proposed by Edwards in 1972 under the name SHELL model (Hawkins, 1993). Specifically, the SHELL model requires analyzing how each person acted and interacted with the other four components. The

different interactions between the person and each of the other components are considered as the human possibility, while a mismatch between the central Liveware and any other four components leads to a source of human error. Moreover, the SHELL methodology adapted to the analysis of the startup failures presented excellent behaviors compared to other results in the literature (Cantamessa et al., 2018). For the aforementioned reasons, we decided to adopt the basic framework of the SHELL for Startups model and to incorporate it in our system. The output of the SHELL model is then joint with expert-based rules mapping the effect of the different components on the core features and creating a risk impact matrix to be used for perturbing the Machine Learning module in the prediction module.

The prediction of a company's business interruption and bankruptcy risk phase is performed by the system depicted in Fig. 2. Given the data of a single company and its risk Matrix obtained by applying the risk impact matrix, a request to the prediction server is sent by REST APIs. The server checks the data, gives the core data to the Machine Learning module, while the non-core data are processed by the Risk's Impact Matrix and then the corresponding oscillations of the core data are introduced in the Machine Learning obtaining the effect on the prediction given by the non-core data prediction. For each set of core and non-core data 5 predictions are created (12, 24, 36, 48, and 60 months), plus a series of performance indexes related to national and international regulations and are then merged in a report. The report gives to the user (entrepreneur, bank, assurance, policy-maker) a detailed description of the company situation, as well as the key aspects that should be considered for reducing the business interruption and bankruptcy risks within a continuous improvement process.

4. Machine learning prediction

This section describes our Machine Learning applied to financial data and used to predict the company bankruptcy, but the process is generic and can be repeated and applied to other data types too. The process works as follows:

- Data cleaning. Data of companies that went in bankruptcy and those which are still active are collected and cleaned;
- Data fusion and first balanced data set creation. Data from the two sets of companies are joint. Due to the heavy unbalancing between bankruptcy and active companies, a balanced data set is obtained by sampling the active data set;
- Data split. To evaluate the performance of a machine learning algorithm, we need to split the data. One part (train) is used for the algorithm to learn how to predict future instances and another part (test) is to examine how good is our algorithm about predicting future samples. This is done by using the Python's Scikit-learn library, setting the test datatset equal to the 20% of the total. Further validation approaches to spot over-fitting or under-fitting have been tested on the data (k-fold, k = 10) (Burnham & Anderson, 2004; Cai, 2014).
- Feature reduction. Being the initial set of financial features composed by about 170 indexes and ratios, this number is reduced with an iterative procedure.
- Hyper Parameter Tuning. The parameters of the Machine Learning method are tuned. To enhance the performance of the Machine Learning module, we need to tune these parameters to get the best results. These parameters may be different From one classification task to another. For this step, we used a tuning approach based on Grid-search (Bergstra & Bengio, 2012). Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it).
- Final data set creation. The data set used in the previous step was built to be representative of some attributes related to geographic dispersion and the industry type. However, as previously stated, one
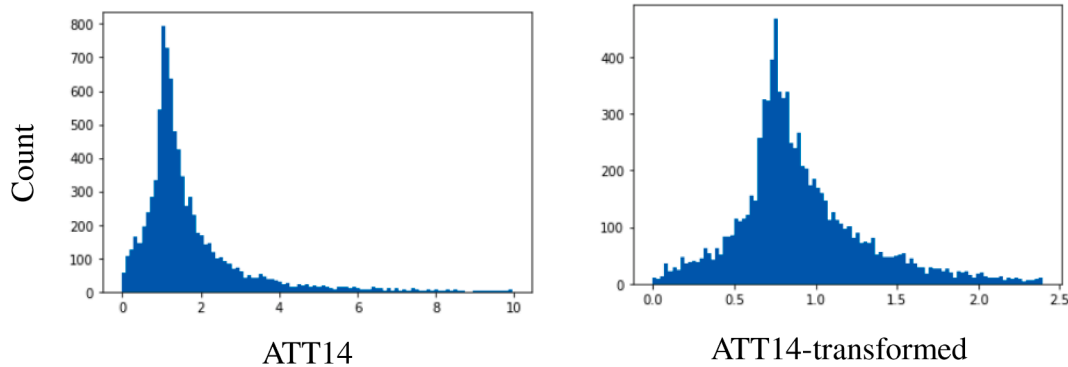
Fig. 3. Example of distribution after the standardization of a feature.

of the contributions of this paper is the introduction of a procedure to obtain a sample that increases the performances of the Machine Learning. In this step, the final data set is created.

In the following, we give more details concerning the main phases.

4.1. Data cleaning

The company data that we used in this work was made of financial information on Italian companies from 2001 to 2018. This data is extracted from the AIDA database, the largest financial and organizational database managed by Bureau van Dijk/Moody's (Bureau, 2020). All types of companies are either limited companies or joint-stock companies. Out of which we collected bankrupted companies such that, all of them had revenues between 1 million to 40 millions of euros in at least one of the last 5 years of life before they go bankrupt and a company lifetime of at least 10 years. For each company, we collect the last 5 years of financial data of such companies and save them into 5 different data sets which are roughly made up of 3000 companies. If a bankrupted company has less than 5 official financial reports, it is removed from the data set.

The active companies are composed of all active companies in 2018 and we collected, again, all financial data of those companies which their revenue is the same as bankrupted companies in the last 5 years of life (i.e., the last 5 years before 2018). For these companies, we kept the last year's information. The number of all active companies in 2018 is more than 160,000. We will refer in the following to the companies that went bankrupt as Class 1 and the active companies as Class 0. Preprocessing data is an important task and it is necessary for improving machine learning metrics. This is because data is mostly noisy, it sometimes has missing and also false values. We applied missing value imputation and standardizing to our train and test (Barnard & Meng, 1999; Hilbe, 2009). We replaced all missing values with zero and we applied standard scaling to both. As shown by Fig. 3, standardizing of data seems to cause the distribution of data to be close to a normal distribution and this will result in improving the prediction. Just for the oversampling part, it is done only on the training set by the Imblearn Python library.

We should notice that if we build our data set in this way the data set would become highly imbalanced and this will affect the result of our machine learning model regarding the recall of confusion matrix. So just to remedy the negative effect of this, first we sampled 6000 active companies out of 160,000 and then start to merge. By doing this we still preserve the imbalanced nature of the data set but in a controlled way meaning that by doing so we somehow sacrifice precision in favor of recall since finding all companies that most likely will declare bankruptcy status is more important. Now for each year of information of bankrupted companies, we add the same sampled of active companies and build our final data set which has 5 different parts (year1, year2, year3, year4, and year5) and is composed of 8959 companies.

Table 2
Data set features.

Feature	Feature value	Feature type
ATT10	Absolute value	Revenue/profit
ATT11	Index/percentage (%)	Revenue/profit
ATT12	Absolute value	Revenue/profit
ATT13	Absolute value	Revenue/profit
ATT14	Index/percentage (%)	Revenue/profit
ATT15	Index/percentage (%)	Cost/debt
ATT16	Absolute value	Cost/debt
ATT17	Index/percentage (%)	Cost/debt
ATT18	Absolute value	Cost/debt
ATT19	Index/percentage (%)	Cost/debt
ATT20	Absolute value	Production
ATT21	Absolute value	Production
ATT22	Index/percentage (%)	Revenue/profit
ATT23	Absolute value	Production
ATT24	Index/percentage (%)	Cost/debt

4.2. Feature reduction

At the beginning of the process when we collected financial information of companies, there were more than 170 different financial and operational features for each company. So we reduced the dimensions of our data set by an iterative feature removal process. More in details, at every step we remove one feature and we discard it if the precision score of a simple classification task didn't change more than 1 percent. By repeating this procedure we removed more than 150 features from our data set and we were left with 15 most important financial features. Table 2 reports a summary of the features. As in many other papers we cannot give the detailed list of the feature sets after the tuning, being under a Non-Disclosure Agreement. On the other side, and differently from the majority of the other works, we give to the reader an idea of the feature data types by providing some feature information. In more detail, data are split into category types (Profit, Cost, and Production) and feature value type (index or absolute value).

4.3. Evaluation

There are many metrics to evaluate a machine learning classification method (which is the effort to assign each future sample to its correct class). Depending on the nature of classification the trade-off between false-positives and false-negatives must be taken into account. In Powers (2011) some metrics for classification problems are introduced. In literature AUC is the most widely used metric for evaluation of a machine learning classifier, however, it is not a perfect metric when different classifiers are used (Hand, 2010; Hand, 2009). Moreover, the AUC curve is the metric used in previous works and so we adopt it to obtain comparable results (Barboza et al., 2017; Son et al., 2019) After training the data on the training set, we used AUC and confusion matrix

Table 3
Summary of the first and second rounds of prediction.

Algorithm	After	First Round						Second Round					
		AUC	F1	Pre	Rec	Matt	Log-loss	AUC	F1	Pre	Rec	Matt	Log-loss
Random Forest	1st year	0.78	0.74	0.74	0.74	0.41	0.53	0.88	0.83	0.83	0.83	0.65	0.40
	2nd year	0.79	0.74	0.74	0.74	0.42	0.51	0.89	0.84	0.84	0.84	0.66	0.37
	3rd year	0.82	0.78	0.78	0.78	0.49	0.48	0.90	0.86	0.86	0.87	0.69	0.35
	4th year	0.85	0.80	0.80	0.80	0.53	0.45	0.91	0.86	0.86	0.86	0.71	0.32
	5th year	0.87	0.81	0.81	0.81	0.58	0.41	0.93	0.88	0.88	0.88	0.74	0.29
XGBoost	1st year	0.77	0.75	0.75	0.75	0.41	0.54	0.88	0.84	0.84	0.84	0.66	0.37
	2nd year	0.79	0.75	0.75	0.75	0.41	0.51	0.90	0.85	0.85	0.85	0.67	0.35
	3rd year	0.82	0.77	0.77	0.77	0.48	0.48	0.92	0.87	0.87	0.87	0.71	0.30
	4th year	0.85	0.80	0.80	0.79	0.53	0.44	0.93	0.89	0.89	0.89	0.75	0.27
	5th year	0.88	0.82	0.82	0.82	0.58	0.40	0.94	0.89	0.89	0.89	0.79	0.25
Logistic Regression	1st year	0.71	0.67	0.70	0.66	0.30	0.61	0.78	0.73	0.75	0.73	0.38	0.53
	2nd year	0.74	0.68	0.71	0.67	0.31	0.60	0.81	0.75	0.76	0.74	0.42	0.50
	3rd year	0.76	0.69	0.72	0.69	0.37	0.57	0.83	0.76	0.78	0.75	0.48	0.48
	4th year	0.79	0.71	0.74	0.70	0.43	0.55	0.86	0.77	0.79	0.76	0.51	0.44
	5th year	0.81	0.74	0.76	0.74	0.46	0.52	0.88	0.80	0.81	0.79	0.57	0.41
Neural Network	1st year	0.72	0.70	0.71	0.69	0.32	0.61	0.85	0.80	0.80	0.80	0.59	0.44
	2nd year	0.73	0.71	0.72	0.71	0.34	0.62	0.88	0.81	0.81	0.81	0.63	0.40
	3rd year	0.79	0.75	0.75	0.74	0.45	0.54	0.90	0.84	0.84	0.84	0.67	0.35
	4th year	0.81	0.75	0.76	0.75	0.47	0.52	0.91	0.85	0.85	0.85	0.69	0.34
	5th year	0.84	0.77	0.78	0.77	0.50	0.50	0.93	0.86	0.86	0.86	0.76	0.30

in addition to Matthews score to evaluate the results on the test set. Since we used cross-validation, in each run, on the train and validation sets, we evaluated those metrics but rather than reporting all of them we averaged them to compare with the result of the test set to see if we are running into over-fitting or not.

4.4. Two-rounds data set creation

In this section, we are going to explain the difference between our approach and all other previous researches which are done in this field, where we add another round of prediction based on the output of the previous round and build our data set again to further differentiate between active and failed companies. Active companies (Class 0) are companies still active, but they contain companies that will go in bankruptcy in the next years. Thus, to consider this aspect, we use a two-step procedure.

In the first round, along with getting the results that inform us which companies will go bankrupt in the five consecutive years, we also get the probability of going bankrupt. A classifier will output a company as bankrupt if the probability of bankruptcy for that company is more than 50%. Note that this is not true for all classifiers where some of them do not operate on the probabilities and instead on discrete zero or one values.

So, for each company and each year, we have a probability for which that company will go bankrupt. At this point, we tried to find a threshold that tells us which companies are less likely to go bankrupt and sample those companies and build our active companies again as we did in Section 4.1. We sampled those companies as active if for all 5 consecutive years the probability of going bankrupt is less than the threshold. For example, if we set the threshold to 20%, we will have 80,000 companies for which the probability of going bankrupt in all 5 years is less than 20% (the outperforming companies). We again sample 6000 active companies out of this number (80,000) and we build our data set as described in Section 4.1.

We set the threshold to 20, 30, 40, 50, 60% of bankruptcy and for each, we extract the new active companies from the active companies in the first round and we test the Machine Learning method.

5. Computational results

In this section, we discuss the results of our model, which is described in Section 4. We also compare our results with the those of Son et al. (2019) and Figini et al. (2017). The performance is measured by using

the following metrics: AUC, F1, Matt Coefficient, Log-Loss, Precision and Recall. Concerning the performance measures, for sake of brevity, we do not include their definition. The interested reader can refer to Matthews (1975) and Japkowicz and Shah (2011).

Concerning the two-phase data set creation, we found a good threshold to be 60%. By setting this threshold, we were actually able to both maintain the companies doing very well as active for the next 1–5 years and, at the same time, to not bias our data set toward bankrupted companies. Empirically, the total number of companies going bankrupt each year should be considered as roughly 3%; the 60% threshold could confirm this fact. The results of the method, applied to different types of machine learning methods, are presented in Table 3, where "first round" presents the results of the standard sampling, while the "second round" presents the performance of the two-phase approach. For the sake of simplicity, we present just the results using the best threshold (60%). As witnessed by the results, our two-phase procedure obtained better performance than standard stratification.

In our work, the gradient boosting (GB) algorithm outperformed the logistic regression and neural network models, while achieving a similar (but slightly better) performance when compared to random forest. The GB algorithm achieved better a log loss (GB = 0.25) score than the other three models (log loss: RF = 0.29, NN = 0.3, and LR = 0.41). These results also confirm the results presented in Son et al. (2019) and Figini et al. (2017). From Fig. 4, it is clear that logistic regression was the worst classifier, as also shown in Table 3. The neural network suffered from over-fitting; its accuracy was not even as high as that of RF. Moreover, we can observe that, for this particular task, the gradient boosting model generalized the best, followed by random forest; demonstrating the generally better behavior of the ensemble methods. In our problem data are highly unbalanced. Moreover, it is quite difficult to implement solutions as the resampling of the minority class (bankruptcy companies). This is due to the nature of the features, some of them presenting a high level of sensitivity. Moreover, the majority class (the active companies) contains a subset of them that are presently active, but that will go in bankruptcy in the next 5 years (normally, about between the 5% and the 10% every 10 years). Thus, to overcome this difficulty, we implemented the two-rounds data set procedure presented in SubSection 4.4. As shown both in Table 3 and Fig. 5,6, carrying out the second round improved the AUC metric by at least 7%; for example, for the random forest model, the AUC of the first round was 78%, which was increased to 88% in the second round. In addition the same results are confirmed by the improvement of F1, recall and precision. In fact for all the algorithms and all the years we can see a boost in the performances. Among
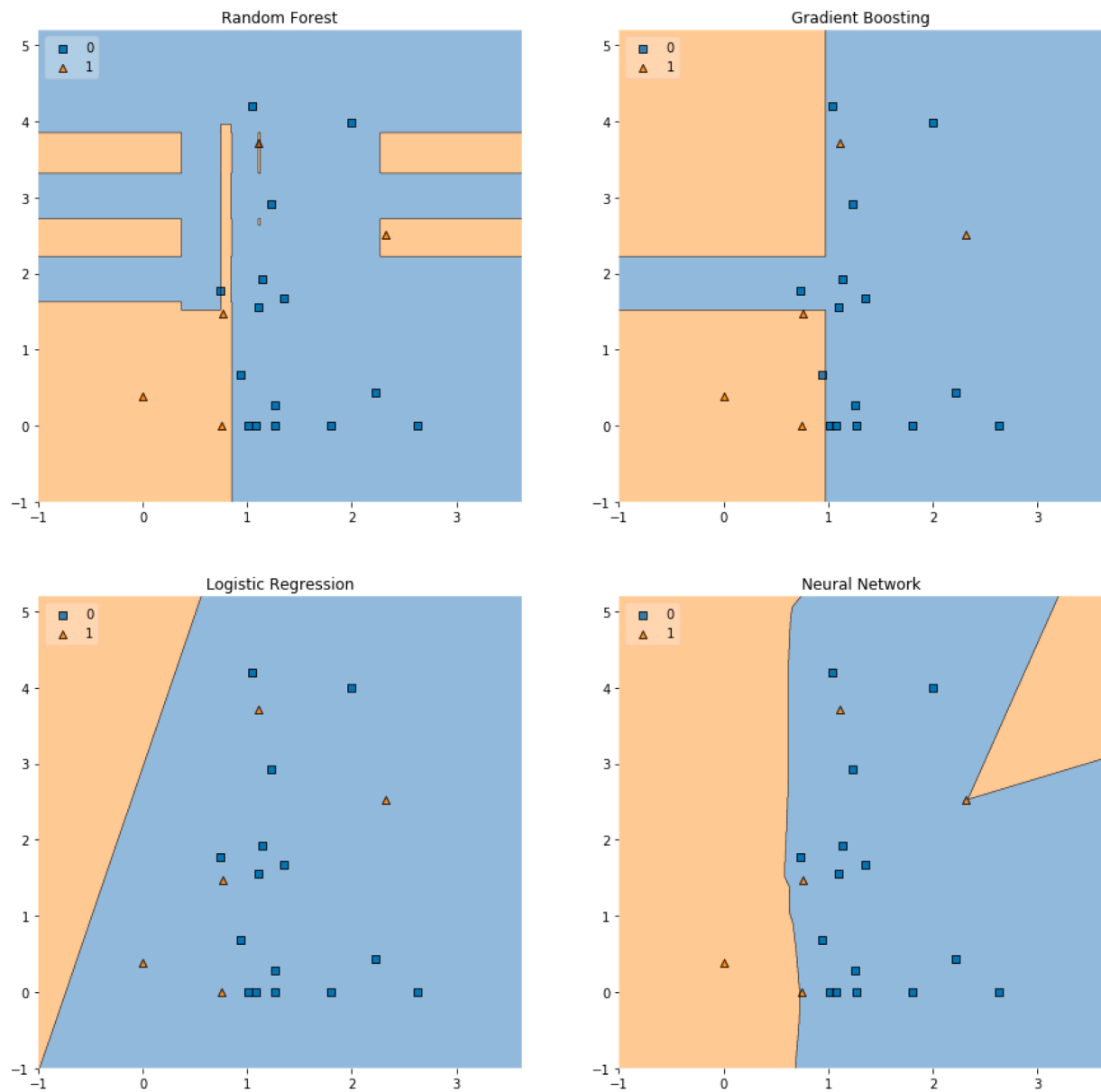
Fig. 4. Decision boundary of the different classifiers on our data set.
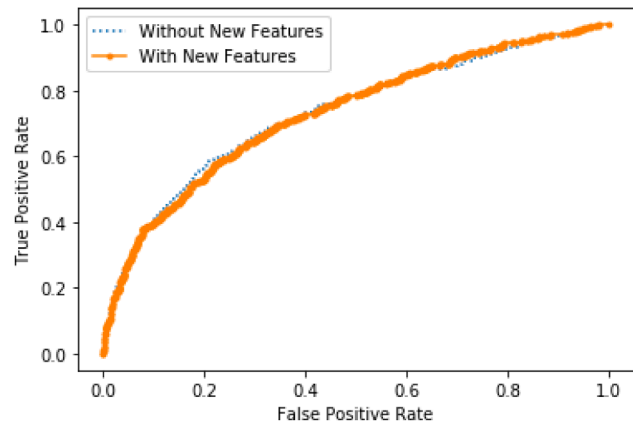

Fig. 5. XGBoost on the "1st Year" data set, with and without additional features of activity type and geolocation.
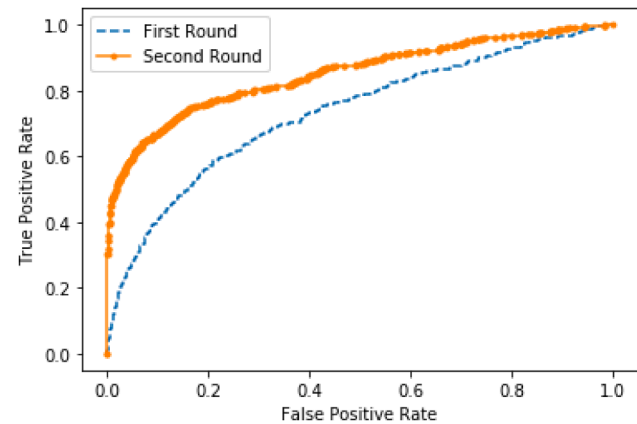
Fig. 6. The ROC curves of the random forest on the "1st Year" data set for the first vs. second rounds.
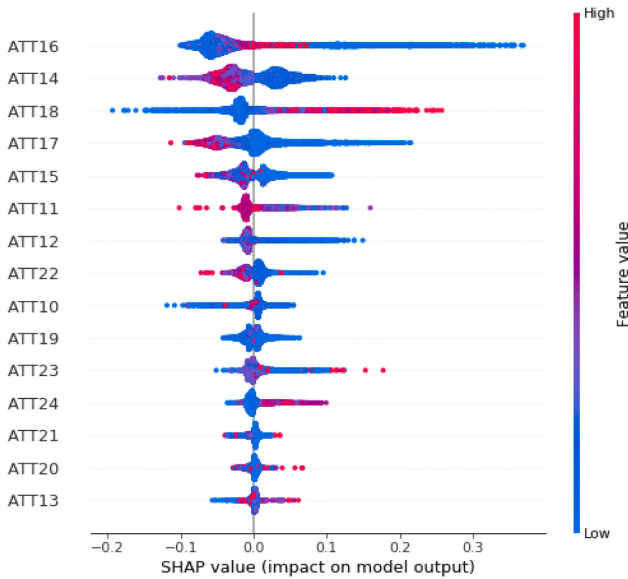
Fig. 7. Attribute importance.

Table 4
Italian companies demographic info.

Segmentation	Cluster	Number of companies
Region	Northeast	39,775
	Northwest	53,045
	Center	36,724
	South	32,058
Activity	Industry	69,351
	Commerce	46,524
	Public	3,630
	Service	42,097
Revenue	< 5 Mil	127,009
	5 ⩽ X < 10 Mil	17,965
	10 ⩽ X < 15 Mil	6,519
	⩾ 15 Mil	10,109

the ensemble models, GB performed better than random forest (see Table 3). If we consider the best results of our work (i.e., GB and random forest) to the best results of the other two papers, the AUC score of our solution (after the second round) achieved comparable or better results than those achieved in Son et al. (2019) and Figini et al. (2017). This is because we used only 15 financial and operational features as independent variables, while the other papers used more than 40 features. This demonstrates the effectiveness of our procedure for feature selection and the importance of financial data for prediction. Moreover, we present an almost constant prediction rate up to five years, in comparison to the other papers, where 18 months were considered, at most.

A study on the type of industrial activity and the geographical location of companies was conducted, in order to observe the effects of these features on company crises. In particular, these classifications normally being considered for arbitrary and historical reasons, we wanted to test whether our machine learning module was independent of them. This is particularly crucial for industrial activity, as it might not catch the real activity of the company. Imagine, for example, a company engaged in precision agriculture. It will be classified as agriculture (which is normally identified by a moderate level of automation, a low level of innovation, and a low or moderate level of digital revolution) while, in terms of the type of activity, it will probably be more similar to a company of the same size in Industry 4.0, which means a high level of innovation and digital penetration. Industrial activity was coded according to the ATECO code, which represents the type of activity of a company. It is a six-digit code, in which the first two digits reveal the greater area of activity. We divided this area into four sub-areas-namely, industry, commerce, public, and services- and then assigned a relevant code to each company. The mapping between ATECO codes and industry was done according to the classification of the Italian Ministry of Economy. The second feature was the geographical feature where the company was established, which was classified according to the province and region in which the company's headquarters were located.

In Fig. 5, we can see that, with these additional features, there was almost no improvement in the AUC of the GB model (as well as for the

other models). This makes sense when considering that most companies tend to slightly change their activity type during their lifetime, as well as their location; although it may be the case that they simply register the company in one place (mostly because of heavy bureaucracy procedures) and carry out work in another place (e.g., another city).

As discussed above, a model which can accurately interpret the results is highly preferable in this field. In this research, in order to explain the decision made by the machine learning model, the SHapley Additive exPlanations (SHAP) method was used, as described in Lundberg et al. (2020), Lundberg et al. (2018), and Lundberg and Lee (2017). The results are shown in Figs. 7 and 8, where the overall importance of the variables that affect the model are shown from top to bottom. The SHAP value may be the only method which can deliver a full explanation. In situations where the law requires explainability—such as the EU's "right to explanation"—SHAP may also be the only legally compliant method, as it is based on a solid theory and distributes the effects fairly. In Fig. 8, we evaluate the outcome of our analysis for a particular random company. We can see that the base value (average target probability without any prediction) was approximately 0.35. Moreover, we observed the features that most affected said company's target evaluation. As can be seen, these features pushed the outcome (0.28) from a base value toward non-bankruptcy. This means that, while the average probability that any company would become bankrupt by chance was approximately 35%, this particular company experienced a 28% probability of going bankrupt, based on our evaluation.

6. Application to a real economic system: the Italian case study

In this section, we show how our machine learning algorithm can be used as a predictive tool for an entire economic system, focusing on Italian SMEs. In more detail, we considered all 160,000 companies with revenues between 1 and 40 million euros in at least one of their fiscal years in the interval 2013–2018 (2018 was the last fiscal year for which the official annual balance sheets were available, due to the COVID-19 crisis) and that were still active at the end of 2019 (i.e., no bankruptcy, merger/acquisition, or displacement event was documented in the official Italian records). The results were validated by a group of experts, led by the former President of the Italian Companies and Exchange Commission (CONSOB); that is, the authority of vigilance on the stock market and the banking system in Italy. Some classifications of the companies, according to the geographic area of their headquarters, the
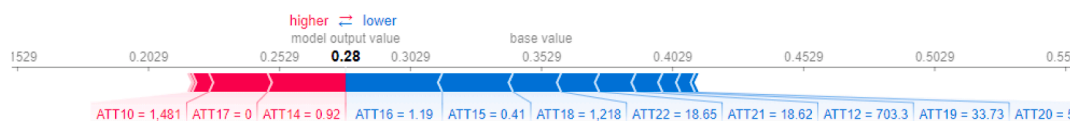

Fig. 8. Attribute importance affecting one company.

Table 5
Italian companies bankruptcy with regard to the activity.

Risk of bankruptcy	Activity	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	Industry	69,351	58%	18%	24%
	Commerce	46,524	56%	19%	25%
	Public	3,630	51%	23%	26%
	Service	42,097	47%	21%	32%
multirow420ptMiddle-term	Industry	69,351	53%	19%	28%
	Commerce	46,524	51%	21%	28%
	Public	3,630	48%	23%	29%
	Service	42,097	43%	22%	35%
Long-term	Industry	69,351	52%	19%	29%
	Commerce	46,524	50%	22%	29%
	Public	3,630	46%	23%	31%
	Service	42,097	42%	22%	36%

Table 6
Italian companies bankruptcy with regard to company location.

Risk of bankruptcy	Location	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	Northeast	39,775	60%	18%	22%
	Northwest	53,045	57%	18%	25%
	Center	36,724	49%	20%	31%
	South	32,058	48%	22%	30%
Middle-term	Northeast	39,775	56%	19%	25%
	Northwest	53,045	53%	20%	27%
	Center	36,724	44%	22%	34%
	South	32,058	44%	23%	34%
Long-term	Northeast	39,775	55%	19%	26%
	Northwest	53,045	52%	20%	28%
	Center	36,724	43%	22%	35%
	South	32,058	42%	23%	35%

Table 7
Italian companies bankruptcy with regard to company revenues (millions of euros).

Risk of bankruptcy	Revenue	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	< 5	127,009	51%	20%	29%
	5 ≤ X < 10	17,965	64%	18%	18%
	10 ≤ X < 15	6,519	65%	16%	19%
	≥ 15	10,109	65%	16%	19%
Middle-term	< 5	127,009	47%	21%	32%
	5 ≤ X < 10	17,965	60%	19%	21%
	10 ≤ X < 15	6,519	61%	17%	22%
	≥ 15	10,109	60%	18%	22%
Long-term	< 5	127,009	46%	22%	32%
	5 ≤ X < 10	17,965	60%	18%	22%
	10 ≤ X < 15	6,519	61%	17%	22%
	≥ 15	10,109	59%	18%	23%

type of activity according to their ATECO codes, and their revenues, are presented in Table 4. The largest proportion of companies was settled in the north of Italy, with revenues between 1 and 5 million euros. The activity type was more distributed, with a slight predominance of the "industry" category.

Table 5 displays the results of the prediction of bankruptcy related to the activity, as provided by the ATECO codes. The probability of a company crisis was computed for short (one year), medium (three years), and long (five years) periods, where the severity of the probability was considered low if the probability was under 50%, medium if it was between 50% and 70%, and high if it was greater than 70%. For each row, we report the total number of companies for each cluster (i.e., industry, commerce, public, and service). Then, for each probability range, the percentage related to the total number of companies in each cluster having the given probability of company crisis is given. The companies working in the "service" cluster were those with the largest percentage of medium and high risk, particularly considering the long-term prediction, while the other sectors presented better performance; especially in the short-term. Comparing short- and long-term predictions, there was an increment of approximately 4%, in terms of companies with a high probability of company crisis in the next 60 months. By considering their direct revenues and the indirect effects of an eventual crisis, we were able to estimate the impact of these companies as being 30 billion euros of direct revenue and 80 billion euros by also including the indirect effects, equal to approximately 3% of the Italian GDP. This proves the economic value of having mid- and long-term predictions, in terms of economic impact.

Table 6 reports the same data clustered by geographic location. The companies were grouped according to the four standard Italian clusters (i.e., northeast, northwest, central, and south). The largest number of companies with a high company crisis risk were those in central and southern Italy, while those with the lowest risk were located in the northeast. It is also worth mentioning that the companies in the north also presented the largest number of companies with low probability (up to 50%).

Regarding the size of the company, in terms of yearly revenue (see Table 7), the companies were grouped into four clusters (less than 5 million euros, up to 10 million euros, up to 15 million euros, and over 15 million euros). These results show how companies up to 5 million euros were those with the largest percentage of high risk and also those with the smallest amount of low risk. This aspect was particularly relevant when moving from the short-term to the middle- and long-term, with approximately 5% of the companies moving to moderate and high probability levels.

Given the recent COVID-19 crisis, we applied our machine learning DSS to the Italian case by simulating the effect of the lockdown and the effects of the Italian government's policy for financially supporting companies. In this case, we focused on the Piedmont area, due to the possibility of obtaining direct data and checking the results with the help of the group of experts led by the former President of CONSOB and some policy-makers of the Regional Council of Piedmont. Moreover, the

Table 8
The main characteristics of Piedmont companies with revenue between 1 and 40 million euros.

Companies	12,707
Revenues (K euros)	5005
EBITDA (K euros)	452
Employees	22.58
Shareholders	3.03

Table 9
The bankruptcy of Piedmont companies pre- and post-COVID-19, as well as after the financial support policy (at 10%, 20%, or 30% of a company's past-year revenues).

Risk of bankruptcy	Prob. <50%	50% ⩽Prob. < 70%	Prob. ⩾70%	Mean risk
Pre-COVID-19	70.7 %	28.7 %	0.6 %	29 %
Post-COVID-19	13.8 %	84.6 %	1.6%	39%
Loan 10% of revenues	15.8%	82.7%	1.5 %	40 %
Loan 20% of revenues	27.7 %	71.5%	0.7 %	33 %
Loan 30% of revenues	20.1%	79.0 %	0.9%	38%

sample was representative, in terms of company mix and revenues, and presented a very favorable pre-COVID-19 situation. Table 8 summarizes the characteristics of the sample. The first row reports the number of companies. As for the Italian case, the companies were SMEs with revenue between 1 and 40 million euros. The other rows report the mean revenues and EBITDA (in K euros), the mean number of employees, and the mean number of shareholders. Our sample was responsible for 49% of the GDP of Piedmont (65 over 132 billion euros) and included approximately 270,000 direct employees in total. To simulate the situation pre- and post-COVID-19, we decreased the revenues of each company by a percentage equal to 30% (estimation made by CONFINDUSTRIA, the main Company Association in Italy). As per the Piedmont Regional Council and the Italian government, no differentiation concerning the revenue per sector was applied. We then applied the policy of providing financial support to companies, in the form of a loan granted by the government equal to a given percentage of their previous year's revenue. We simulated percentages equal to 10%, 20%, and 30% of the previous year's revenue. The risk was computed for the middle-term (three years). The results of the simulation are reported in Table 9. In the table, the pre- and post-COVID-19 situations without any public policy and with the policy of the loans granted by the government (with the three different percentages), the number of companies (percentage of the total) with low (risk under 50%), medium (risk between

50% and 70%), and high (more than 70%) risks of bankruptcy, as well as the mean risk of all 12,707 companies are shown. It is worth noting that the initial situation was quite good, with just 0.6% of the companies at high risk pre-COVID-19. Post-COVID-19, the number of high-risk companies tripled, but the worst result was that the percentage of low-risk companies became just 13.8%, compared to the original 70.7%. This was due to two effects: The loss of revenue and already being on the border between low and medium risk (which was the case for almost half of the companies). The effect of the policy of giving loans of a certain percentage of the revenues of a company not working properly if the percentage was low (10%) was inconsistent while, with 20% and 30%, the effect was more consistent. The 30% policy provided a low deterioration of the general situation, due to the increase in mid-term debt, with the financial support given at a very low interest rate to be refunded in a fixed time (five years in Italy). Moreover, in terms of the mean risk over the full set of companies, the best policy was the 20% one, with mean risk almost returning to that of the pre-COVID-19 value.

We also performed an analysis of the effects on the economic system caused by the lockdown and the policy with a loan equal to 20% of a company's past-year revenues, which is summarized in Fig. 9. The index of unemployed workers was forecasted to increase by 12%, with the Italian temporary lay-off numbers increasing of 160% with respect to the previous year. Moreover, most of the companies would be out of the Basel III and other short-term financial stress tests, leaving them incapable of receiving a loan from banks and the bank system, blinded in terms of evaluation of the future performance of their customer portfolios (Georg, 2011; Altman, 2020). The policy of providing the grant for 20% of a company's revenues would bring the high-risk companies back to the pre-COVID-19 situation, but needing approximately five years to return to the same situation as in 2018, under the hypotheses of a reduction in GDP of 10% in 2020 with an increase of GDP of 6% in 2021, 5% in 2022, and a loan payment of 10 years. Pay-back of the loan in five years, as per the hypotheses of the Italian government, might abolish the effect of the financial support, bringing the point of return of the investment to 8.5 years and increasing the stress of companies so much that it increases their risk back to high or medium.

7. Conclusions and future developments

In this paper, we considered the challenge of forecasting company crises using machine learning techniques. The machine learning training step was enhanced by use of a two-phase training procedure, which was able to improve the performance of the considered machine learning methods. We demonstrated how we were able to accurately forecast the presence of a crisis up to 60 months in advance, starting from operational and financial data. Moreover, we introduced our machine learning module into a DSS and applied it to Italian SMEs, in order to
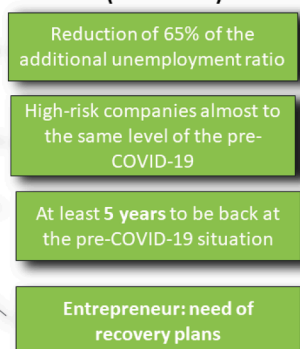

Fig. 9. Summary of the post-COVID-19 and the post-government policy (20% of the revenues) effects.

analyze the Italian economic system. Finally, we used the DSS as a support tool for validating public policies related to the economic shock resulting from the COVID-19 pandemic.

Future developments include the introduction of additional data from other risk sources, such as cybersecurity and seismic data, and to explicitly include the dynamic evolution of the system into the machine learning module, as well as including the presence of a certain level of uncertainty by incorporating extreme value theory (Perboli et al., 2014).

CRediT authorship contribution statement

Guido Perboli: Conceptualization, ML models implementation, ML models testing, Case study definition and supervision, Case study analysis. **Ehsan Arabnezhad:** ML models testing, Case study analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

While working on this paper, Guido Perboli was the head of the Urban Mobility and Logistics Systems (UMLS) initiative of the interdepartmental Center for Automotive Research and Sustainable mobility (CARS) at Politecnico di Torino, Italy, as well as the R&D Director of ARISK, a spin-off of Politecnico di Torino.

References

Altman, E., Esentato, M., & Sabato, G. (2020). Assessing the credit worthiness of italian smes and mini-bond issuers. *Global Finance Journal, 43*.

Altman, E., Iwanicz-Drozdowska, M., Laitinen, E., & Suvas, A. (2016). Financial and nonfinancial variables as long-horizon predictors of bankruptcy. *Journal of Credit Risk, 12*(4), 49–78.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.

Altman, E. I. (2014). Predicting financial distress of companies: Revisiting the Z-Score and ZETA models. In A. R. Bell, C. Brooks & M. Prokopczuk (Eds.), Handbook of research methods and applications in empirical finance (pp. 428–456). Edward Elgar Pub.

Altman, E. I. (2020). Covid-19 and the credit cycle. SSRN paper n. 3604361, DOI: 10.2139/ssrn.3604361. URL:https://ssrn.com/abstract=3604361.

Altman, E. I., Danovi, A., & Falini, A. (2013). Z-score models' application to italian companies subject to extraordinary administration. *BANCARIA, 04*, 24–37.

Argenti, J. (1976). *Corporate collapse; the causes and symptoms*. McGraw Hill.

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

Baldi, M. M., Manerba, D., Perboli, G., & Tadei, R. (2019). A Generalized Bin Packing Problem for parcel delivery in last-mile logistics. *European Journal of Operational Research, 274*(3), 990–999.

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

Barnard, J., & Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From aids to nhanes. *Statistical Methods in Medical Research, 8*(1), 17–36.

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

Becerra, V. M., Galvão, R. K., & Abou-Seada, M. (2005). Neural and wavelet network models for financial distress classification. *Data Mining and Knowledge Discovery, 11*(1), 35–55.

Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Review of Accounting Studies, 1*(4), 267–284.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

Boritz, J. E., Kennedy, D. B. & Albuquerque, A. d. M. e. (1995). Predicting corporate failure using a neural network approach. Intelligent Systems in Accounting, Finance and Management 4, 95–111.

Brédart, X. (2014). Bankruptcy prediction model using neural networks. *Accounting and Finance Research, 3*(2), 124–128.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

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

Bureau van Dijk (2020). Aida – italian companies database. URL:http://aida.bvdinfo.com/.

Burnham, K. P., & Anderson, D. R. (2004). *Model selection and multimodel inference*. New York: Springer. URL:http://link.springer.com/10.1007/b97636.

Cai, E. (2014). Machine learning lesson of the day – overfitting and underfitting.

Cantamessa, M., Gatteschi, V., Perboli, G., & Rosano, M. (2018). Startups' roads to failure. *Sustainability, 10*(7), 2346.

Chen, J., Chollete, L., & Ray, R. (2010). Financial distress and idiosyncratic volatility: An empirical investigation. *Journal of Financial Markets, 13*(2), 249–267.

Chen, M.-Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications, 62*(12), 4514–4524.

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, J. C. (2011). A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Systems with Applications, 38*(10), 12939–12945.

Chou, C.-H., Hsieh, S.-C., & Qiu, C.-J. (2017). Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. *Applied Soft Computing, 56*, 298–316.

Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. *Financial Management, 22*(3), 142–155.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning* (pp. 3–33). Springer International Publishing.

Figini, S., Bonelli, F., & Giovannini, E. (2017). Solvency prediction for small and medium enterprises in banking. *Decision Support Systems, 102*, 91–97.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence, 14*(771–780), 1612.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Georg, C. -P. (2011). Basel III and systemic risk regulation – What way forward? Working Papers on Global Financial Markets 17, Graduiertenkolleg 'Konstitutionelle Grundlagen globalisierter Finanzmärkte - Stabilität und Wandel'. URL:http://hdl.handle.net/10419/94477.

Hambrick, D. C., & D'Aveni, R. A. (1988). Large corporate failures as downward spirals. *Administrative Science Quarterly, 33*, 1–23.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning, 77*(1), 103–123.

Hand, D. J. (2010). Evaluating diagnostic tests: The area under the roc curve and the balance of errors. *Statistics in Medicine, 29*(14), 1502–1510.

Hawkins, F. H. (1993). *Human factors in flight*. Aldershot, England: Ashgate.

Hilbe, J. M. (2009). Data analysis using regression and multilevel/hierarchical models. *Journal of Statistical Software, 30*(3).

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of Bankruptcy. *Review of Accounting Studies, 9*(1), 5–34.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. USA: Cambridge University Press.

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

Linden, H. (2015). *Synthesis of research studies examining prediction of bankruptcy* (p. 88). Aalto University School of Business.

Lundberg, S. & Lee, S. -I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 2017-Decem, pp. 4766–4775). Curran Associates Inc.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence, 2*(1), 2522–5839.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749.

Luoma, M., & Laitinen, E. (1991). Survival analysis as a tool for company failure prediction. *Omega, 19*, 673–678.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure, 405*(2), 442–451.

Odom, M. D. & Sharda, R. (1990). A neural network model for bankruptcy prediction. In 1990 IJCNN international joint conference on neural networks (pp. 163–168). IEEE.

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

Ooghe, H., & Prijcker, S. D. (2008). Failure processes and causes of company bankruptcy: A typology. *Management Decision, 46*, 223–242.

Patrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant, 2*, 598–605.

Pellegrino, S., Perboli, G., & Squillero, G. (2019). Balancing the equity-efficiency trade-off in personal income taxation: An evolutionary approach. *Economia Politica, 36*(1), 37–64.

Perboli, G., Tadei, R., & Gobbato, L. (2014). The multi-handler Knapsack problem under uncertainty. *European Journal of Operational Research, 236*, 1000–1007.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technology*.

Shah, J. R., & Murtaza, M. B. (2000). A neural network based clustering procedure for bankruptcy prediction. *American Business Review, 18*(2), 80.

Son, H., Hyun, C., Phan, D., & Hwang, H. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications, 138*, Article 112816.

Upneja, A., & Dalbor, M. C. (2001). An examination of capital structure in the restaurant industry. *International Journal of Contemporary Hospitality Management, 13*(2), 54–59.