Università
Ca'Foscari
Venezia

Master's degree
in Economics and Finance

Final thesis

Robust logistic regression for SMEs' default prediction

Supervisor:
Ch. Prof. Lisa CROSATO

Co-supervisor:
Ch. Prof. Stefano CAMPOSTRINI

Graduand:
Kevin Dalla Mora
Matriculation number 868254

Academic year
2021-2022

Typeset by LaTeX2e

"Research is immersion in the unknown.
We just don't know what we're doing.
We can't be sure whether we're asking the right question
or doing the right experiment until we get the answer or the result".
Martin A. Schwartz

"A discordant small minority
should never be able to override
the evidence of the majority of the observations".
Peter J. Huber,
Elvezio M. Ronchetti

Abstract

Predicting SMEs' default and financing promising firms means protecting 99% of all enterprises in the EU, as well as the largest part of the European value added and jobs. Accordingly, there is a vast literature studying SMEs' default in European Countries, mainly based on accounting indicators. Logistic regression is the benchmark model for classification of default, due to remarkable performances comparable with those of machine learning methods, with an immediate interpretation.

The goal of the thesis is to search for alternative methods such as robust logistic regression to predict SMEs' default in Italy. Firstly a comprehensive bibliographic research on SMEs' default prediction is carried out, followed by the description of the collection and creation of a large dataset of balance sheets downloaded from Aida database. Thereafter the available libraries in R are used to apply robust logistic regression to classify defaulted firms within the collected data, rearranging the functions where needed. Lastly, a comparison of classification rates, the significance and relevance of the coefficients with the standard logistic regression outcome is performed to contextualize the results within the relevant literature.

The aim is to point up that although new methods should be taken into consideration, the logit model remains the cornerstone of credit risk evaluation, besides credit scoring.

Note to the reader

Typeset by LaTeX

The thesis has been written with LaTeX$_{2\epsilon}$ (LaTeX home page), which is a high-quality typesetting system; it includes features designed for the production of technical and scientific documentation. LaTeX is the *de facto* standard for the communication and publication of scientific documents. LaTeX is available as free software. At present time, TeX is a registered trademark of the *American Mathematical Society* (AMS). The program uses various extensions to increase potentialities which are identified with the symbols AMS-LaTeX, which states for "*LaTeX with AMS's extensions*".

The usage of LaTeXhas been integrated with extensions which let me add hyperlinks, table of contents, list of figures, list of tables, all correctly ordered. The work let me improve the knowledge of this typesetting system to show all the possibilities this kind of language is able to produce, which in this thesis are only partly examined in depth by the author.

Contents

List of Figures

List of Tables

Introduction

The category of micro, small and medium-sized enterprises (SMEs) is made up of enterprises which employ fewer than 250 persons and have an annual turnover not exceeding EUR 50 million, and/or an annual balance sheet total not exceeding EUR 43 million. Nonetheless, as stated by the European Commission, SMEs are the engine of the European economy. They drive job creation and economic growth and ensure social stability. In 2013, over 21 million SMEs provided 88.8 million jobs throughout the EU. Nine out of every 10 enterprises is a SME, and SMEs generate two out of every three jobs. SMEs also stimulate an entrepreneurial spirit and innovation throughout the EU and are thus crucial for fostering competitiveness and employment as stated by the European Union [56]. In this context the relative importance of SMEs is even higher in Italy (European Commission, 2019), where they generate 66.9% of the overall value added in the national "non-financial business economy", exceeding the EU average of 56.4%. The share of employment generated by SMEs is also greater, at 78.1%, compared to the EU average of 66.6%. Micro firms are particularly important, providing 44.9% of employment compared to the EU average of 29.7%.

Given that the 99% of Italian enterprises is composed of SMEs, the aim of this research is to provide the necessary tools and structure to examine in depth the workings of default prediction models for this kind of firms, first considering the logistic regression and then comparing the results with more robust approaches. In particular, a large amount of financial and non-financial information, grouped in different macro-categories and adequately cleaned and processed for missing values, will be used to build a database including all Italian SMEs registered in the Bureau van Dijk's database between 2018 and 2019. To analyse and predict the default probability one-step forward three key sectors of the Italian economy will be selected, namely the manufacturing sector, the construction one and the trade.

Starting from a stepwise logistic regression, which is a baseline modelling procedure in credit risk research, the focus will shift towards robust logit models to take outliers into account. Robust models will include the Bianco-Yohai estimator, a more robust version of Pregibons's estimator that works with a bounded function ρ to downweight influential observations in the design space. Another contribution to assess the impact of the outliers, the Forward Search by Atkinson and Riani [9], will be explored and revised to model validation tools necessary for classification and in particular to investigate the characteristics of firms identified as outliers. The whole analysis will be carried out through a standard training and test set protocol, where the best model estimated on the training set has been selected via cross out-of-sample validation.

Results show that robust models could be more parsimonious and slightly better than the standard logit model, particularly in correctly classifying defaulted firms, while the forward search singles out as outliers exactly the good companies with bad indicators and the bad firms with good indicators. This is particularly true in the modelling stage and helps in improving classification of failed firms in the test set, although to a different extent across sectors. Other inter-sectoral differences can be found in the selected indicators.

The research is divided into the following sections: the first section reviews the literature regarding SMEs' default prediction, showing that recent events, such as the new Basel II Accords which established that banks should have developed credit risk models specifically addressed to SME, laid the foundations of further research. The second section summarizes the relevant notions of credit scoring modelling, especially what logit and robust logit analysis mean. The third section regards the construction of the Dataset to perform the analysis. The fourth section dives into the empirical results coming out from the sectoral analysis of the Italian economy. Conclusions of the work and discussion are presented in the last section.

Chapter 1

Literature review

The prediction of default is a relative young field of research and dates back to the work of Altman [2] in 1968, followed by critics and improvements in the 1980s by Edmister [27], Ohlson [47] and Gentry, Newbold, and Whitford [29]. To start the chapter with, it would be useful to recall some important concepts regarding default prediction, which is more properly an art, rather than an exact science. The economic development of the 20th century, especially during the first half and in times of crisis, forced banks and agencies to think about the kind of information needed to assess the creditworthiness of merchants and enterprises.

The evaluation and analysis of financial ratios to predict the default has long been at the core of financial institutions to grant loans to those who deserved it on one side and refuse it to those who did not merit it on the other side. That was the primary role of banks, before the recent investment banking and trading activities jumped in to transform the sector radically. The sharpening of credit risk models required not only the work of practitioners, but also the studies of academicians, being of course the starting point referred to US publicly manufacturing corporations for which comprehensive financial data were obtainable.

However, in the last twenty years[1], as new credit score models came into view and got frequently updated in an international regulatory framework, more and more researchers started considering also Small and Medium Enterprises (SMEs) classification and forecasting. Using the words of Ciampi et al. [21, p. 2146]: *"The interest in the field started in the early 1970s. However, it was only after 2004, the year of the initial publication of the Basel II Accord, which linked the minimum required levels of the capital of financial institutions to the level of creditworthiness of their clients more strictly, where we recorded a gradual but growing interest by scholars"*.

[1] To be honest, Altman [2, p. 609] already in 1968 said in his most famous paper that *"an area for future research* [...] *would be to extend the analysis to relatively smaller asset-sized firms and unincorporated entities where the incidence of business failure is greater than with larger corporations"*.

SMEs play an important role in the economic system of many European countries, in particular in Italy, where they represent almost the entire social fabric. As Covid-19 hit, the economic impact on those firms got sharper because generally they are characterized by over-indebtedness, relational financing, lack of transparency and are not enough resilient when it comes to downturn situations. Therefore they need to be assessed with carefulness through an early warning system based on a set of relevant key risk indicators to prevent possible losses for banks and correctly reallocate scarce resources to the economy.

Before building a new brick in the research, let us briefly discuss the previous literature on SMEs failure prediction, based on recent papers, referring to [21] for an exhaustive and in-depth analysis[2]. The author of the "z-score" Altman [2] refers to studies which suggest the potential of ratios measuring profitability, liquidity and solvency as significant indicators capable of predicting default. Gentry, Newbold, and Whitford [29] focus on the classification of failed and non-failed companies based on cash-flow funds flow components, yet they do not find significant improvements, except for the dividend funds flow component. In 2003 Lehmann [39] analyses the impact of qualitative information to perform credit risk modelling and forecasting of SMEs' default. The following year Dietsch and Petey [26] employ the logit model to capture SMEs' portfolio default correlation. The authors look at a large sample of German and French SMEs, divided further by size, and show a low correlation between small and medium-sized enterprises and the 'state of the economy'. Furthermore, the results of portfolio simulations reveal a positive relationship between asset correlation and the probability of default - the concentration of default being higher in the riskiest classes, causing chain defaults. This relationship may not be so strong for small and medium enterprises because they give a diversified contribution to the economy and this could be a benefit when building an efficient portfolio[3]. Instead, this behaviour can be observed for credit portfolios consisting of large corporate exposures, so-called concentrated portfolios. Kolari, Ou, and Shin [37] consider the US banking industry and test the hypothesis whether the profitability of banks specializing in small business lending is better than diversification (the point of view is different in this case). The outcome is that indeed lending to SMEs can play a positive role due to their positive effect on ROA and ROE across different bank size groups. A further step is made by Sohn and Kim [54] whose paper suggests a random

[2] The bibliometric analysis of Ciampi et al. [21] gathers together several papers related to five different clusters in SMEs' default literature: the firm-bank relationship; the 'core' SME default prediction-modelling literature; the innovation-related variables in predicting SMEs' default; the critical variables for small company success; the prediction models based on longitudinal data.

[3] Dietsch and Petey [26, p. 786]: *"Even if smaller SMEs are on an individual basis riskier than the large SMEs, their very weak sensitivity to systematic risk and the positive effects of large portfolio diversification invite the exposures on these firms to be treated as retail exposures"*.

effect logistic regression model to explain the probability of default of SMEs considering either financial and non-financial characteristics. The results indicate a higher prediction accuracy, strengthened by including some macroeconomic variables, such as CPI or exchange rate.

An important piece of work relates to Altman, Sabato, and Wilson [6]. Their starting point regards the fact that SMEs do not have the same amount of information as listed firms, so that in order to elaborate a credit risk model it is practical to take also non-financial statements into account. Following the previous studies, the model covers a high number of firms before the Great Recession, proves to be significant in its predicting power, and improves its accuracy by adding non-financial characteristics of the companies such as age, financial reporting, compliance and trade credit relationships. Psillaki, Tsolas, and Margaritis [50] express themselves in the same way in their paper which does not produce a comprehensive model for bankruptcy prediction, rather focusing on non-financial performance indicators useful to predict business failure. In the end, it shows that more efficient firms, in terms of less distance from the industry's best practice frontier, are less prone to fail. Moro and Fink [45] concentrate on 'relationship lending', which is a key driver in the Italian bank-centred system. Results show that trust between SMEs and banks could be a win-win strategy for both actors: on one side firms would gain a more easy access to credit, on the other side the bank manager would reduce the costs of monitoring and control and improve the decision making process by having more soft information.

Recent studies include Mannarino and Succurro [41] who attempt to verify if and to what extent financial ratios affect the probability of default in different Western Europe convergence regions. Results indicate that the financial structure is a significant factor, with differences among financial ratios and countries. For instance in Italy, where internal resources are extremely important due to the higher difficulties to access external financial resources (bank-based economy), debt and cash flow ratios are relevant. Ciampi [18] focuses on the improvement of Small Enterprises' default prediction accuracy by combining economic-financial variables with corporate governance variables, which are CEO-duality, size and composition of the board of directors and ownership concentration. The conclusion points that banks need to carefully include more and more qualitative variables in their rating systems in order to significantly reduce the errors of not granting credit to the healthier small firms. Ju, Jeon, and Sohn [34] consider time-varying covariates using the Cox proportional hazard model to take into account the effects of changing economic indicators after a loan application is successfully approved. Stress test results show that firms with a high level of marketability factors (market potential, product competitiveness) are significantly affected by economic conditions in terms of technology credit risk, especially during a recession. Moreover, they exhibit higher loan default rates than those with high scores in management or profitability. Barreto Fernandes and Artes [12] investigate the neighbourhood relationship between SMEs, which is treated as a risk factor and included as an explanatory variable in the logistic model analysis. The empirical results show an improvement in the evaluation of the probability of default of SMEs. Gabbianelli [28] includes firm-territory relationship variables in a logistic regression model to improve the accuracy of small business default prediction models. The results are in line with the previous literature ([19], [20]).

Lastly, Ciampi et al. [21] review SMEs default models by constructing a database based on queries to summarize the state of the art. It is patent noticing that the interest of scholars has been increasing over the last twenty years due to the development of the Basel II Accord (and updates), in particular concerning the internal rating system to evaluate and differentiate between large corporate firms and SMEs. Moreover to improve accuracy rates, practitioners have started considering non-financial variables, such as corporate governance characteristics, history, size or audit qualifications, bank-firm relationship, which should be further investigated. The authors advise to take advantage of big data analytics and AI techniques to pave the way for future research.

Chapter 2

Research methodology

2.1 Theoretical framework

The aim of this section is to underline the salient aspects of credit scoring modelling for SMEs through an overview of the main statistical models. One could start by recalling that, according to the working paper of the Basel Committee on Banking Supervision [11] (BCBS), the statistical analysis of rating systems and score functions is based on the assumption that there are two categories of obligors of a bank: obligors that will be in default at some predefined time horizon and obligors that will not be in default at this time horizon. Usually, it is not known in advance whether an obligor belongs to the first or to the second category. Banks therefore face a dichotomous (or binary) classification problem as they have to assess an obligor's future status by using present available characteristics only.

The procedure of applying a classification tool to an obligor for an assessment of her or his future status is commonly called discrimination. The main construction principle of rating systems can be described as "the better a grade, the smaller the proportion of defaulters and the greater the proportion of non-defaulters that are assigned this grade". Consequently, the more the defaulters' distribution on the grades and the non-defaulters' distribution on the grades separate, the better the rating system will discriminate. The discriminatory power of a rating system thus denotes its ability to distinguish ex ante between defaulting and non-defaulting borrowers and it can be assessed using a number of statistical measures of discrimination. A credit default scoring model is therefore considered optimal when it is compared to others based on several validation techniques, such as the accuracy rate and AUROC. Using different datasets and different reference countries, it is preferable to consider a multiplicity of models rather than just one to optimize classification.

When default models are considered, one needs to face the problem of the relative small number of defaulted firms compared to that of active firms, which, although structural, poses a series of issues in modelling. Just to mention, the well-known E. Altman [3] started analysing the issue of 'zombie' companies, which actually continue to survive although they do not have healthy financial indicators. This impacts the models of credit scoring leading to misclassification of such firms.

Lastly, in a classification model one can incur in two type of errors: the type I error, when a good credit is judged as bad one; the type II error, when a bad credit is considered a good one. Indeed, the second mistake has more serious implications [1] for a bank profitability due to misallocation of scarce financial resources.

2.2 The models

2.2.1 Linear discriminant analysis

The first model developed by Altman [2] is the linear discriminant analysis. Traditional studies suggest the potential of ratios measuring profitability, liquidity and solvency as relevant indicators capable of predicting default. The multiple discriminant analysis model theorized by Altman allows to predict the corporate companies that will default or not by looking at balance sheet data (linearly combined inputs) with significant effectiveness in a limited time horizon (2-3 years). The inputs are linearly combined to determine the default of companies which is taken as the dependent variable in a qualitative form. Besides the Z-score, it has implications for the evaluation of loans to consumers and businesses, investments and internal control procedures.

Following this path, Edmister [27] chooses the most indicating financial ratios to predict small business failure. A number of hypothesis is assumed, mainly the fact that the ratios' average level and their trend are predictors of default, and the multiple discriminant analysis (MDA) is used as statistical method. The author concludes that *"since ratios tend to be very similar in their information content, great care has to be taken to select a group that is as diverse as possible"* [27, p. 1491], so that it is better to have a few independent variables than to have too many multicollinear variables. Further research confirms that the linear discriminant analysis is generally based on strong basic assumptions difficult to be fully justified, mainly the linear relationship between the variables and the normal distribution of the independent variables. Furthermore, the variance/covariance matrices for the two groups (failed/non failed firms) must be equal. As a consequence new studies switched to logistic regression.

2.2.2 Logistic regression analysis

Logistic regression is one of the most popular probabilistic model for classification. It predicts the probability of one event (out of two alternatives) taking place by modelling the log-odds for the event as a linear combination of one or more independent variables, which are called 'predictors' or explicative variables. Formally, there is a single binary dependent variable, coded by an indicator variable, where the two values are labelled "0" (solvent firm) and "1" (defaulted firm) or viceversa, while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

From a mathematical point of view[4] suppose that we have k independent observations $y_1, ..., y_k$, and that the $i-th$ observation can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i).$$

Further we suppose that the logit of the underlying probability π_i is a linear function of the predictors

$$logit(\pi_i) = \boldsymbol{x_i'\beta},$$

where x_i is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients. Exponentiating and solving for the probability π_i in the logit model we obtain the more complicated model

$$\pi_i = \frac{exp\{\boldsymbol{x_i'\beta}\}}{1 + exp\{\boldsymbol{x_i'\beta}\}}.$$

While the left-hand-side is in the probability scale, the right-hand-side is a non-linear function of the predictors, and there is no simple way to express the effect on the probability of increasing a predictor by one unit while holding the other variables constant. One of the first probabilistic model of bankruptcy used the maximum likelihood estimation (MLE), with the log-likelihood function given by:

$$logL(\beta) = \sum[y_i log(\pi_i) + (n_i - y_i) log(1 - \pi_i)],$$

where π_i depends on the covariates x_i and a vector of p parameters $\boldsymbol{\beta}$ through the logit transformation. At this point we could take first and expected second

[4] See the chapter 3 of Rodríguez [52] for further details.

derivatives to obtain the score and information matrix and develop a Fisher scoring procedure for maximizing the log-likelihood. Given a current estimate $\hat{\beta}$ of the parameters, we calculate the linear predictor $\hat{\eta} = x_i^{'}\hat{\beta}$ and the fitted values $\hat{\mu} = logit^{-1}(\eta)$. With these values we calculate the working dependent variable z, which has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} \cdot n_i$$

where n_i are the binomial denominators. We then regress z on the covariates calculating the weighted least squares estimate $\hat{\beta} = (X^{'}WX)^{-1}X^{i}Wz$, where W is a diagonal matrix of weights with entries $w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i$. The resulting estimate of β is used to obtain improved fitted values and the procedure is iterated to convergence.

One of the first application of this model is given by Ohlson [47] which assesses a conditional logit model for traded companies. The failure/non-failure dichotomy is a crude approximation of the space of results and possible decisions. His paper does not want to investigate bankruptcy theories or the usefulness of indicators, but is a statistical analysis. The assessment of the predictive capacity is based on the description of the classification and the verification of which of the proposed models minimizes the sum of the percentage errors. The conclusion emphasizes the point concerning the lag in the publication of financial data by unhealthy companies due to distress, which is an element that should be considered more properly in the forecasting model. Moreover, the author urges the need to use additional forecasters to improve the estimate. Altman and Sabato [5] indicate that refining the logistic regression model for the forecast of default, i.e. by creating a distinct rating system between SMEs and large corporates, banks could benefit both directly entrusting healthy SMEs and indirectly, by reducing the level of capital regulations to be set aside according to current regulation.

2.2.3 Robust regression

The concepts and methods of robust statistics originated in the 1950s (the technical term “robust” was coined in 1953 by George Box). However, the concepts of robust statistics had been used much earlier. Robust statistics assesses the changes in estimates due to small changes in the basic assumptions and creates new estimates that are insensitive to small changes in some of the assumptions. Robust statistics is also useful to separate the contribution of the tails from the contribution of the body of the data. A general theory of robustness was developed by the statistician Huber [32], who introduced modern methods such as the M-estimators which are a generalization of maximum likelihood estimators (MLEs). Afterwards, other authors tried to explore the robust logistic regression with different results, some of which will now be presented.

In 1996, moving from the assumption that the ML-estimate is extremely sensitive to the presence of anomalous data in the sample, Ana M. Bianco and Victor J. Yohai developed a new class of robust and Fisher-consistent M-estimates for the logistic regression models, whose estimates were consistent and asymptotically normal. They studied the behaviour of different datasets under outlier contamination to prove that a small fraction of arbitrary outliers in the sample has a small effect when computing the robust ML-estimate. From another point of view in 2001 Atkinson and Riani [10] suggested a simple robust method for the detection of atypical and influential observations in binomial data, based on a forward search procedure which orders the observations from those most in agreement with a specified generalized linear model to those least in agreement with it. The aim was again developing a robust estimator in detecting masked multiple outliers and showing the structure of the data.

These robust models have been tested in some recent papers, some of which will be below recalled, before diving into the econometric models and formulas behind. Hauser and Booth [31] use the Bianco-Yohai estimator to significantly improve the classification and prediction of default since it behaves better when dealing with outliers, which are supposed to be misclassified bankrupted firms. Although there are only two possible outcomes, it is always easier to predict the default of non-bankrupted firms since these are more numerous in the data set and in the real world. If there are outliers in the data sample, the BY robust logistic regression will result in significantly different estimated coefficients and better bankruptcy prediction. If there are no significant outliers in the data sample, the BY robust logistic regression will produce essentially the same results as ML logistic regression. Also Miyamoto [43] tests the BY estimator to investigate the independent variables needed for credit risk assessment of SMEs in Japan, showing that complete data and data treated with multiple imputed methods underperform based on the accuracy ratio and the area under the receiver operating curve with respect to BY.

Further developments imply more advanced statistical analysis. For example Komarek and Moore [38] evaluate the logistic regression against modern machine learning algorithms to prove that the predictive performance of the first is better when accelerated by a conjugate gradient approximate linear solver. Cheng [17] considers as a starting point the fact that a "complete case analysis" is often a waste of information, because the omitted units carry information with respect to the relation between the observed covariates and the outcome variable. Consequently, the study uses a EM algorithm in combination with the forward search algorithm to detect multiple outliers in glm with incomplete data.

Lastly, Atkinson and Riani [10] explore a forward search algorithm to detect the influential multiple outliers that have a clear effect on residuals and test statistics, while they do not properly fit the structure of general linear models. Therefore these observations should be taken into account when performing a prediction analysis.

After this parenthesis, the two robust regression models, namely the econometric model of Bianco and Yohai [25] and the forward search of Atkinson and Riani [9], will be defined mathematically. To start with, the estimator of the ML is defined as:

$$\hat{\gamma}_n^{ML} = \underset{\gamma}{argmax}\, logL(\gamma; X_n) = \underset{\gamma}{argmin} \sum_{i=1}^{n} d(z_i^t \gamma; y_i),$$

where $logL(\gamma; X_n)$ is the log-likelihood function calculated in γ and $d(z_i^t\gamma; y_i)$ is the deviance function given by

$$d(z_i^t\gamma; y_i) = -y_i\, logF(z_i^t\gamma) - (1 - y_i)\, log\{1 - F(z_i^t\gamma\}.$$

The generalization of the first equation consists of replacing the deviance function by another one to define the estimator $\hat{\gamma}_n$ which is $\hat{\gamma}_n = \underset{\gamma}{argmin} \sum_{i=1}^{n} \varphi(z_i^t\gamma; y_i)$ with the following properties: φ is a positive, continuous and differentiable function; $\varphi(s; 0) = \varphi(-s; 1)$ for any score s, where a score value $s_i = z_i^t\gamma$ is obtained as a linear combination of a given parameter vector γ; $lim_{s \to \inf} \phi(s) = 0$ implying that a large negative score is not contributing to the objective function. A more robust model was first introduced by Pregibon in 1982, followed by Bianco and Yohai in 1996, who elaborated a more consistent version by working with a bounded function ρ and defining

$$\hat{\gamma}_n = \underset{\gamma}{argmin} \sum_{i=1}^{n} \{\rho(d(z_i^t\gamma; y_i)) + C(z_i^t\gamma; y_i)\},$$

with $C(z_i^t\gamma; y_i)$ being a bias correction term. Bianco and Yohai suggested using the following function

$$\rho(t) = \begin{cases} t - \frac{t^2}{2c} & \text{if t} \leq \text{c} \\ \frac{c}{2} & \text{otherwise} \end{cases},$$

where c is a tuning parameter. To make a comparison, while the maximum likelihood estimator is $\theta_{ML}(s) = -ln(1 - F(s))$, the BY estimator is $\theta_{BY}(s) = \rho(-ln(1 - F(s))) + G(F(s)) + G(1 - F(s)) - G(1)$.

The other method to detect important observations which may strongly affect the generalized linear model fitted to data is the forward search. It is a general robust method extensively described by Atkinson and Riani [9] with the aim of identifying the outliers which strongly influence the estimates and may be masked using standard deletion diagnostic procedures or simple logistic regression. The details of the model will be used in the thesis to assess the effectiveness of the models previously developed. The procedure consists of three main steps (following Grossi and Bellini [30]):

1. The forward search algorithm starts with the selection of a basic subset free from outliers based on the p parameters of the model. In the case of generalized regression models, squared deviance residuals are used in order to select the units belonging to the basic subset which does not contain atypical observations. In formula the initial subset is such that:

   $$d^2_{[med],S^p_*} = min_j(d^2_{[med],S^p_j}),$$

   where $d^2_{[l],S^p_j}$ is the lth ordered squared residual among d^2_{i,S^p_j}, $i = 1, ..., n$

   $$med = p + \left[\frac{n-p}{2}\right].$$

2. Starting from the initial subset, the forward search selects the $m+1$ units with the smallest squared deviance residuals, the units being chosen by ordering the observations according to their degree of accordance to the underlying model using the squared deviance residuals d^2_{i,S^m_*}. At each step we collect the estimators $\hat{\beta}_{FS}$ resulting from the maximum likelihood estimations. In most moves from m to $m+1$ just one new units joins the subset, but it is important to underline, as shown later on, that *"sometimes two or more units join S^m_* as one or more leave. Such an event is unusual, only occurring when the search includes one unit which belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course several other units then must leave the subset"* Atkinson and Riani [10, p.64]. Graphically, the introduction of outliers is signalled by sharp changes in the curves which monitor parameter statistics at every step.

3. The last step regards the monitoring of statistics, such as parameter estimates, t-values, and so on along each step of the search to obtain an overview of the structure of the data. Among others, one can use the Cook's distance, which *"measure[s] the effect of deletion of a single observation and so may be liable to masking when several outliers are present. The forward search*

overcomes this masking, with abrupt changes in parameter estimates indicating influential observations, which can be detected through the monitoring of a "forward version" of the Cook statistic D_i" Atkinson and Riani [9, p.34].

2.2.4 Other models

In order to give a complete picture of the state of the art regarding credit default classification, other classifiers and their main features will be discussed briefly. The most important one is the Artificial Neural Networks (ANN), which is a model that simulates the way the human brain works, i.e. in a non-linear way and by examples. The main pro consist in having a high precision rate, but the inability to explain the process from input to output (black box) and time constraints (long process to build an optimal network) are cons that must be taken into account. For example Akkoç [1] develops a three-step hybrid model that combines the Artificial Neural Network with the Neuro Fuzzy applying it to the creditworthiness of Turkish consumers applying for a credit card and finds that some variables have an explanatory power regarding the reasons for which a credit request is refused (low level of education or work maturity). Ciampi and Gordini [19] test the default prediction accuracy of the ANN model against the multiple discriminant analysis and logistic regression using data from Italian SMEs divided into business sectors, geographic areas and size, with proper financial-economic ratios as independent variables. All the three models show satisfying results, whose accuracy improve by considering the three divisions either separately or in twos. Whatever the level of aggregation the analyses is made, ANNs are better. This research is conducted after Ciampi et al. [20] show that either LDA and LRA could provide significant information regarding future SE default, even though some caution should be exercised in applying statistical methods and interpreting results, due to the general opacity of small companies.

The Support Vector Machine (SVM) model produces a binary classifier through a non-linear mapping of the input vectors into the high dimensional feature space. There are some advantageous features, namely there are only two free parameters to be chosen (the upper bound and the kernel parameter), the solution is optimal and unique, this type of classifier minimizes the upper bound of the actual risk and not the empirical one. Kim and Sohn [36] applies it to predict the default rate of Korean SMEs using four categories of input variables: SMEs characteristics, financial ratios, technology evaluation and economic indicators. The results outperform existing methods, such as LR and ANN, so that it can be considered as a valid alternative method. Nehrebecka [46] makes a comparison between logistic regression and SVM through the transformation of raw data using weight of evidence to evaluate the best results in terms of credit scoring.

Since the number of default SMEs is relatively lower than the non-default ones, the problem of underestimation of PDs could be risky for banks. As a consequence Calabrese and Osmetti [14] propose the Generalized Extreme Value (GEV) distribution which is suitable to model extreme values and rare events data. The model is tested against the logistic regression and is considered to be a good regression model to identify defaults. [7] go further and analyse and compare UK and IT predictors of SMEs insolvencies applying the GEV, BGEVA and logistic additive models to data and two methods for missing values, i.e. the weight of evidence and imputation. Other less relevant statistical models include Multiple Criteria Decision Aid (MCDA) [8] and Grabit [53].

2.3 Evaluating a model

The goodness of fit statistic for which one speaks about in logistic regression is the deviance to refer to the residual sum of squares. As a general definition following Atkinson and Riani [10] the deviance is ϕ times the log-likelihood ratio test for comparing the model with parameters β in the linear predictor to the saturated model for which the parameter estimates β^{max} are such that the fitted means $\hat{\mu}_i$ equal the observations y_i; that is,

$$D(\beta) = 2\phi\{L(\beta_{max}) - L(\beta)\}$$

or explicitly

$$D = 2\sum\{y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i) log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\}.$$

In a perfect fit the ratio observed over expected is one and its logarithm is zero so the deviance is zero. Besides this measure one can also compute the residual deviance. This deviance is compared with that coming from the null model (the one in which the linear predictor contains only a constant). The difference between the null deviance and the residual deviance is called the explained deviance.

Another important statistic to evaluate the goodness of fit is the so-called pseudo-R^2 measure. Unlike ordinary least square R^2, log-likelihood-based pseudo-R^2 does not represent the proportion of explained variance but rather the improvement in the model likelihood over a null model. The multitude of available pseudo-R^2 measures and the absence of benchmarks often lead to confusing interpretations and unclear reporting, although almost all pseudo-R^2

are influenced to some extent by sample size, number of predictor variables, and number of categories of the dependent variable and its distribution asymmetry. In this thesis the McFadden's R^2 will be used, which is defined as $1 - \frac{deviance}{null\ deviance}$ where a small ratio (and thus a final value close to 1) indicates that the specified model is better than an intercept-only model.

Lastly, to measure the performance of a model one can add another criteria based on likelihood, to take the number of parameters required to fit the model into account. The ultimate objective is to include only the relevant and explicative variables, trying to have the shortest or more succinct computational description. These measures are, among others, the Akaike (1974) information criterion (AIC)

$$AIC = -2logL(\hat{\theta_k}) + 2k$$

and the Schwartz (1978) Bayesian information criterion (BIC)

$$AIC = -2logL(\hat{\theta_k}) + 2logN,$$

where k is the number of parameters, N the size of the sample, and $\hat{\theta}$ the k-dimensional vector of parameter estimates. $L(.)$ is the likelihood function. The best model is supposed to minimize the selected criterion. Both BIC and AIC penalize models with many parameters and thereby reduce overfitting.

2.4 Classification metrics

The usual starting point for measuring the effectiveness of the achieved classification is the confusion matrix[5]. Generating the predicted classes based on the typical 50% cutoff for the probabilities, it shows a cross-tabulation of the observed and predicted classes. The positive class represents the not-normal class or behaviour, so it is usually less represented than the other class. The negative class, on the other hand, represents normality or a normal behaviour. For the two classes (0 "Non-bankrupt", 1 "Bankrupt") important information can be obtained, which are now briefly described:

- The accuracy ratio tells us how many right classifications are made out of all the classifications. It tells the ratio of “trues" to the sum of “trues" and “falses", $\frac{TP+TN}{TP+FP+FN+TN}$. The confusion matrix considers also the balanced

[5] The information described below are provided by the Caret package which contains tools developed to create a unified interface for modelling and prediction, streamline model tuning using resampling, provide a variety of helper and increase computational efficiency, using parallel processing. Therefore it was possible to define measures for predicted classes which are mostly used in the classification process.

	Actual Class	
Predicted	Event	No Event
Event	TP	FP
No Event	FN	TN

accuracy which is an average between sensitivity and specificity. The overall accuracy rate is computed along with a 95 percent confidence interval for this rate (using binomial test) and a one-sided test to see if the accuracy is better than the "no information rate", which is taken to be the largest class percentage in the data.

- Recall or Sensitivity detects the proportion of actual defaults correctly classified $\frac{TP}{TP+FN}$.

- Specificity is the ability to correctly classify a firm as non-default and usually tends to be inversely proportional to sensitivity $\frac{TN}{TN+FP}$.

- Cohen's Kappa tells how much better the classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. This measure is more informative than overall accuracy when working with unbalanced data. For a good model Cohen's kappa is close to 1.

- The AUROC (Area Under the Receiver Operating Curve) indicates how well the probabilities from the positive classes are separated from the negative classes. The higher the AUROC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the higher the AUC, the better the model is at distinguishing between SMEs defaulted and survived. The ROC curve is plotted with Sensitivity against "1 - Specificity" where the former is on the y-axis and the latter is on the x-axis. An excellent model has AUC near to 1, meaning that it has a good measure of separability. A poor model has an AUC near 0 which means it inverts all the classifications and when AUC is 0.5, it means the model has no class separation capacity whatsoever (same as flipping a coin).

- McNemar's test indicates the change in the proportion of non-default firms following the intervention. The null hypothesis shows that the two probabilities for each outcome are the same.

- The H-Measure normalizes the classifiers' cost distribution based on the expected minimum misclassification loss since the AUC suffers from using different misclassification cost distributions for different classifiers.

Chapter 3

The construction of the Main Dataset

In this chapter the work done to collect the data for the computation of the results will be explained. The activity included the search for the financial and non-financial indicators, the construction of the dataset and data cleaning and manipulation before performing the empirical analysis. The sample was drawn from AIDA, a financial dataset powered by The Bureau van Dijk Database, which gives necessary access to the economic and financial information about Italian small- and medium-sized companies in great detail. AIDA helps in the research by selecting from a variety of criteria: one can go through by trends and over multiple years as well as combine many criteria and use full Boolean logic (and, or, not). When a group of companies has been identified they can be compared against each other. One can also create and customize indicators, calculate sector averages, modify report layouts, customize data layouts, comparisons and evaluations and create both tables and graphs. The data downloaded are presented in the appendix (table 21).

Before moving the attention to the selection of the independent variables, one needs to focus on the definition of the legal status regarding default, which is the categorical and dependent variable. Lin, Ansell, and Andreeva [40] address the problem of different definitions of default, from bankruptcy to financial distress, to develop a broader and more robust accounting-based model and show that the model's accuracy vary depending on the default definition, although profit and growth related variables constantly remain important in distinguishing between healthy and insolvent companies.

Another perspective regarding default concerns firms which recover from a bad stressed condition, touching the default by a hair's breadth. For example Wolter and Rösch [58] focus on cure event studies, in which resurrected firms are no longer treated as defaulted firms, instead they are included in a revised default model that distinguishes between definitely defaulted firms and healthy/recovered ones with a two-step procedure. The Cure After Default Model provides significant influence on the default risk, therefore could be considered as additional information into credit risk modelling.

In the Italian crisis and insolvency law the default is intended as the end of the firm's activity, i.e. when the company enters a bankruptcy procedure and is subject to liquidation and the remaining assets are used to pay creditors and shareholders, based on the priority of their claims. With the new Code of the Business Crisis and Insolvency ("CCII"), which fully replaced the Italian Bankruptcy Law on July 2022, the logic has changed from stressing 'failure' to a new focus which displays a preference for business continuity and favours composition with creditors over liquidation. The organization of early and timely intervention can significantly increase the chances of a successful business recovery or, at least, resolve the crisis in the least traumatic way.

From a financial risk point of view, as stated by the CRR [48, art. 189] *"a default shall be considered to have occurred with regard to a particular obligor when either or both of the following have taken place:*

(a) the institution considers that the obligor is unlikely to pay its credit obligations to the institution, the parent undertaking or any of its subsidiaries in full, without recourse by the institution to actions such as realising security;

(b) the obligor is past due more than 90 days on any material credit obligation to the institution, the parent undertaking or any of its subsidiaries".

In this work data were collected for the fiscal years 2018, 2019 and 2020 by focusing on two defined groups of firms, namely active firms which are healthy and currently operative, and defaulted firms, which do intentionally comprehend firms with one of the statutes "bankruptcy", "default of payment", "receivership", "in liquidation", "dissolved (bankruptcy)" to take into account as many enterprises as possible. Consistently with the literature, dissolved firms that no longer exist as a legal entity were excluded when the reason for dissolution is not specified or due to merger or de-merger. The dependent variable was therefore considered as default in t+1 when the firm is in a defaulted status and the last balance sheet is t and survived in the other cases.

The Dataset was downloaded during the week of 14-18th of March 2022 and the analysis was performed with the software R. A disclaimer should be inserted here regarding the information used (Ciampi and Gordini [19]):

- Small firms have fewer legal obligations regarding data disclosure than larger firms so less information is publicly available. Some ratios are ineffective below certain dimensional levels and manoeuvrable as audits are not compulsory;

- More often than not Italian SMEs do not have the owners separated from the managers so that could have a negative impact on accuracy;
- Relationship banking could have a much higher significance than accounting balance sheets.

The fundamental economic and financial and non-financial indicators included as independent variables are divided into 7 categories based on previous literature, (see table 22 for a general overview): Leverage, Liquidity, Profitability, Coverage, Activity, Non-financial information and Economic indicators. For all of these macro-aggregate a brief description will be now given.

Solvency ratios, also called leverage ratios, measure a company's ability to sustain operations indefinitely by comparing debt levels with equity, assets, and earnings. In other words, these ratios help investors assess a company's ability to meet its long-term obligations. They also explain how the company has been financed (debt or equity).

- The short-term-debt-to-equity ratio is a measure of how investors evaluate a company's short-term leverage. High values and an increasing trend are assessed negatively and interpreted as a deterioration of company's creditworthiness, because of the growing influence of short-term debt on the level of equity.
- Debt-to-equity, known as gearing ratio, constitutes a broad category of financial ratios and refers to total debt along with total equity, an expression of the percentage of company funding through short and long-term borrowing.
- The assets-to-equity ratio reveals the proportion of an entity's assets that has been funded by shareholders. A low ratio indicates that a business has been financed in a conservative manner, with a large proportion of investor funding and a small amount of debt. A low ratio should be the goal when cash flows are highly variable, since it is quite difficult to pay off debt in this situation. A higher ratio is tolerable when a business has a long history of consistent cash flows, and those cash flows are expected to continue into the future.
- The equity-to-debt measure strengthens how much the firm's assets can decline in value before the liabilities exceed the assets and the firm becomes insolvent.
- The short-term-debt-to-total-debt is an indicator which measures in percentage terms the relative weight of short-term financing capital sources

(current liabilities) on the total of short- and long-term third-party capital sources (current liabilities and long term liabilities). Companies which are reliant on short term funding are more vulnerable to liquidity shocks than those with longer-term debt finance as debt facilities can be withdrawn immediately. While companies with short term financing are likely to have a lower cost of debt than those with longer-term financing, should interest rates rise, those with short term financing will see rates rise faster.

- The total-debts-to-total-assets ratio measures a company's leverage with that of other companies in the same industry. This information can reflect how financially stable a company is. The higher the ratio, the higher the degree of leverage and, consequently, the higher the risk of investing in that company is expected.

Liquidity ratios are used to determine whether a company is able to pay off its short-term debt obligations.

- The cash-to-total-assets ratio measures the portion of a company's assets held in cash or marketable securities. Although a high ratio may indicate some degree of safety from a creditor's viewpoint, excess amounts of cash may be viewed as inefficient.

- The tangible-to-total-assets ratio measures the percentage of a company's physical assets whereas the intangible-to-total-assets ratio considers a company's brand value and other intangible aspects of its valuation. Collateral is generally one of the instruments used by banks to assess creditworthiness and it is expected to have a negative relationship with the probability of default as more solid firms should repay debts. Moreover goodwill could be considered as future growth opportunities and may add value to the firm.

- The working-capital-to-total-asset ratio is a measure of the net liquid assets of the firm relative to the total capitalization ([2]), so it determines the short-term company's solvency.

- The current ratio measures whether a company has sufficient short-term assets to cover its short-term liabilities.

- The quick ratio compares current liabilities only to those assets that can be readily turned into cash.

- The solvency ratio is calculated by dividing the net assets by total assets and represents how effectively a company funds its assets with shareholder equity, as opposed to debt.

The profitability of a company refers not only on the margins generated, but also on the assets that must be employed to generate those profits.

- ROE is a well-known measure of the profitability of the equity, while ROI measures the profitability of a company's investments without regard to the way the investment is financed.
- ROCE is a metric for analysing profitability and for comparing profitability levels across companies in terms of capital. Two components are required to calculate return on capital employed: earnings before interest and tax (EBIT) and capital employed (total assets - current liabilities).
- ROS indicates the profitability of the sales, while ROA is used to determine how efficiently a company uses its assets to generate a profit.
- EBIT-to-total-assets is a measure of the true productivity of the firm's asset, taking also tax and leverage factors into account.
- Retained-earnings-to-total-assets is a measure of cumulative profitability, as it considers the ability of the enterprise to accumulate profits over time. Younger firms have less reinvested earnings, so a low indicator potentially equates to a higher default rate.
- Productivity ratios measure in general the goods and services/added value produced (output) to the number of labour (input) required for the production process.
- The research-&-development-to-sales ratio is a measure to compare the effectiveness of R&D expenditures between companies in the same industry. It is calculated as R&D expenditure divided by total sales.

Coverage ratios are extremely important for banks as they are used to determine how easily a company can pay interest and capital on outstanding debt.

- The interest coverage ratio indicates the ability of the company to cover interest expenses through the economic margins (gross profit and EBIT) and through net profit.
- The debt-to-EBITDA ratio is a measure of a company's ability to pay off its incurred debt. A high ratio result could indicate a company has a too-heavy debt load.
- The cash-flow-to-debt is a type of coverage ratio and can be used to determine how long it would take a company to repay its debt if it devoted all

of its cash flow to debt repayment. Cash flow is used rather than earnings because cash flow provides a better estimate of a company's ability to pay its obligations. In the banking sector the Debt Service Coverage Ratio is a well-known and extensively adopted indicator.

Activity ratios are the keys to analyse how effectively and efficiently small and medium businesses are managing the assets to produce sales.

- The asset turnover ratio measures the value of a company's sales or revenues relative to the value of its assets. The higher the asset turnover ratio, the more efficient a company is at generating revenue from its assets. Conversely, if a company has a low asset turnover ratio, it indicates it is not efficiently using its assets to generate sales.
- Account payable/COGS is a short-term liquidity measure used to quantify the rate at which a company pays off its suppliers. Accounts payable turnover shows how many times a company pays off its accounts payable during a period. A decreasing turnover ratio indicates that a company is taking longer to pay off its suppliers than in previous periods.
- Account receivable/sales measures the rate at which accounts receivable are being collected on an annual basis. A low accounts receivable to sales ratio is almost always favourable, as it means that the company's cash collection cycle does not represent a great liquidity risk. The bulk of the company's sales goes into its cash account, which can then be used to finance the business.

Non-financial information is capable of yielding valuable information and improve credit rating systems which are based solely on quantitative information by considerable amounts ([39]).

- Size is found to have a relevant meaning in different studies (e.g Dietsch and Petey [26] Michala, Grammatikos, and Ferreira Filipe [42]). When dividing SMEs into groups based of the number of employees and assets, the impact of the macro-economy on small and medium companies is relatively softer than on micro companies, which seem to have be more vulnerable to economic fluctuations and have less healthy years on average.
- The categories of business sector or the geographic location captures possible effects of the typically diverse (economic and financial, structural and behavioural) profiles of firms operating in different industries and influence the likelihood of the firm's default/non default Ciampi [18].

- The management factor concerns education, professional and industry experience of the top and middle management, the quality of management information systems (controlling, accounting) which allow for timely and up-to-date information about financial and operational risks, and the existence of a plausible long-term business strategy for the company [39], therefore it should be strictly connected to the health of a firm.

Economic indicators (Kim and Sohn [36], Sohn and Kim [54], Michala, Grammatikos, and Ferreira Filipe [42], Ju, Jeon, and Sohn [34], Wolter and Rösch [58]) relate to systemic macro-economic variables (GDP, CPI, unemployment rate, Oil price) potentially associated with credit risk of firms. Recent studies in default prediction models these macroeconomic conditions with the core assumption that an increase in the economic sentiment indicator results in lower distress rates.

New challenges and techniques will be developed in the future to calibrate the internal ratings with relevant changes in the predictive variables using the variables from the last two categories, but this thesis will only refer to the solid traditional foundations for simplicity.

Chapter 4

Empirical results

4.1 A general overview

The main characteristics for all of the accounting ratios are summarized in table 22 in order to give a first concise picture of the Italian business context before the beginning of the pandemic. At a first glance, leverage ratios reveal the typical over-indebtedness of the Italian SMEs, see the debt/equity ratio values oscillating between 1.89 in 2018 and 1.66 in 2020 on average. Defaulted companies show even more higher level of both short and long liabilities compared to total assets and that is quite non-surprising. When considering liquidity ratios, attention is given to the value of net working capital since in the defaulted firms it is negative. It is to remind that negative working capital describes a situation where a company's current liabilities exceed its current assets as stated on the firm's balance sheet, so that the company is in financial distress. Profitability ratios and coverage ratios follow the same path with negative values on average associated with defaulted firms, while activity ratios do not show considerable large differences. Finally, the main indicators reveal "on average" that between 2018 and 2020 revenues and profits decreased while the total assets rose, being the number of employees almost the same.

Before starting with the modelling, a spatial comparison was done considering how the different regions (Figure 3) and sectors (Figure 4) behaved with respect to the default variable. Not surprisingly, half of the active Italian SMEs are based in Lombardia, Lazio, Campania and Veneto while one third of the bankrupted firms are located in Lombardia, Lazio and Veneto. The percentages of defaulted and not defaulted are quite different in either year for several regions, see for example Lombardia 15.47% bankrupted vs. 20.83% non-bankrupted in 2018 or vice-versa Puglia 5.34% vs. 8.09%, although the proportions remained quite stable for the same regions. The trigger event of the pandemic is shocking in the data in the sense that one can see some changes in numbers, especially taking into account the most affected regions, Lombardia and Veneto, which form one third of the total defaults between 2020 and 2021 (+5.84 basis point on aggregate). In Campania, Emilia-Romagna, Lazio, Liguria, Lombardia, Toscana, Valle D'Aosta the differences of the percentages between active and non-active firms turned out to be lower, meaning that the percentage of the defaulted firms increased. In other words the concentration of active firms (always in the sense of those firms which stayed alive) is quite similar to the one of bankrupted ones (Lombardia is emblematic in that sense, from 21.02/14.77% to 22.27/20.86%).

Moving to the analysis by industry, it can be observed that first of all the main sectors of the Italian economy relate to manufacturing, construction and trade, followed by a variety of services, from food and accommodation to technical and scientific ones. A second observation concerns the proportions of defaults versus non-defaults: the percentages of bankrupted is higher in wholesale trade and retail trade, in construction and in services, while lower in real estate and manufacturing. That is confirmed in 2019 but the situation changes a little bit in 2020 where one can see an expected increase in absolute values of the defaults, whose percentages relative to 2019 skyrocketed in the Construction sector (16% to 18.25%) and the Real Estate Activities (8.68% to 12.38%).

In general, when estimating a default model for small and medium enterprises, predictive models have better performances when trained for a specific sector as this avoids pooling heterogeneous firms. As highlighted by Rikkers and Thibeault [51, p. 208] *"economic intuition suggests that for three reasons industry effects should be an important component in bankruptcy prediction. First, financial ratios differ between industries, because industries differ with respect to factors of production, product life cycles, competitive structure and distribution modes, which cause industry differences in various measures of financial condition. Second, industries face different levels of competition and therefore, the likelihood of bankruptcy can differ for firms in different industries. And third, accounting standards might differ between industries"*. All of this relates to industry effects in banking prediction which could affect the variables included in the model, the coefficients and the slope. Disadvantages in dealing with different sectors exist, for example building industry specific models leads to a range of models and being the models based on small samples, they might develop less robust. Besides, the development and regular validation of the models according to the Basel II requirements is time consuming and costly. The option to incorporate industry effects in default prediction models is feasible and quite satisfying, being in line with new studies on non-financial information, but it is not so straightforward to implement dummy variables, therefore it was decided to conduct the analysis for the leading sectors separately, where a high number of defaults is observed. Moreover all the sectors involving public intervention and less competition, such as health care and social assistance and utilities, which are indirectly controlled by national governments, were excluded due to the complexity in distinguishing between defaulted, non-defaulted and zombie enterprises, which are kept alive due to public choices. Lastly, the finance and insurance sector were excluded due to the fact that the balance sheets contain less explanatory variables, e.g. turnover is not indicative of the profitability of the firms (most revenues come from provisions and other income results).

4.2 Manufacturing

The analysis was carried out firstly on the manufacturing sector, since Italy is the second largest manufacturing country in Europe and particularly strong in sectors such as machine tools, fashion, food products, automotive and pharmaceuticals. The sector was suitable for the analysis as it is characterized by the presence of small and medium-size firms, which are found mainly in north-eastern and north-central Italy. Successful Italian manufacturers tend to be export driven and invest more in advanced manufacturing technologies. In 2020, the year of the Covid-19 outbreak, Italian GDP decreased by 8.9% and Italian manufacturing suffered correspondingly with exports decreasing 11.5% and revenue by 8.9%, but bounced back the following year with rapid improvement in factory production and book orders.

To define the sample the database previously built for 2018[6] was filtered using ATECO 2007 codes (from 10 to 33), collecting all the variables and the dependent variable default in 2018, which is equal to 1 if the firm considered defaulted in 2019, otherwise 0 if it remained alive[7]. Moreover a strictest definition of SMEs was implemented, namely only firms with annual turnover of fewer than 50 million euros, the number of employees lower than 250 and a balance sheet of fewer than 43 million euros were retrieved. The resulting dataset contained 121.988 SMEs with a proportion of 2.9% defaulted firms.

The average values and standard deviations of the variables separately for survived and defaulted firms are reported in table 23. In particular, if one considers the five categories it can be seen that, as expected, active firms exhibit

[6] Since the financial statements of 2019 (published in 2020) and 2020 (published in 2021) could include references to Covid-19 effects, the research focused on 2018 data.

[7] An early warning: e.g. when pricing a loan, a bank usually takes into account both backward-looking information (balance sheets and non-financial information of previous years) and a minimum of one-year forward looking information (business plans, future cash flow) to predict the default of the firm. Therefore financial institutions actually attempt to know in advance the prospective economic status of the firm, well-before the release of the financial statements, and give it a scoring which is periodically updated during the lifetime of the banking product

better figures: indeed the gearing ratio is highest for troublesome firms; current, liquidity and solvency ratio shows the resilience of active firms; profitability variables (such as ROE, ROA, ROS) exhibit negative signs for defaulted firms; coverage reveals weaknesses in defaulted firms (EBIT/Interest expenses and Cash Flow/Debt are negative); activity ratios are quite similar.

Before handling the prediction models, the presence of NAs in the dataset was controlled and taken care for. The cleaning procedure consisted in two steps:

- Firstly the dataset was filtered to evaluate only variables which did not contain more than 15% of missing values. Exceeding that threshold presumably meant the variable would not be explicative enough for the model;
- Secondly the dataset was reduced to consider only firms which actually had information (the problem of missing information in SMEs is not new in literature and assessed in different ways), as a consequence firms with more than one missing value were excluded from the analysis. This manipulation actually could give more robustness to the results, although it could also result in a biased analysis, a drawback and a warning one needs to take into account.

All in all, the procedure scaled down the dataset to 102.571 SMEs. Thereafter the variables known to behave according to highly right-skewed distributions were log-transformed. To perform the analysis the dataset was divided into training set (70%) and testing set (30%) and the model's parameters were estimated and validated. More in detail, one hundred iterations were performed, with the creation of sub-training set and validation set via random sampling so that at each trial, the assessment was done on the validation set from a different training model. Finally the models' performance was evaluated on the hold-out sample.

4.2.1 Glm

Two issues were addressed regarding the computation of the logistic regression model. The first one was that of imbalanced classification: in general it refers to the fact that one class outnumbers other class by a large proportion. Class imbalance could lead to a reduction of accuracy for a number of reasons: ML algorithms struggle with accuracy because of the unequal distribution in dependent variable; the performance of existing classifiers gets biased towards majority class; the algorithms are accuracy driven i.e. they aim to minimize the overall error to which the minority class contributes very little; ML algorithms assume that the data set has balanced class distributions; they also assume that errors obtained from different classes have same cost. As a consequence the method of undersampling[8] was adopted to modify the imbalanced data into balanced distribution. The undersampling was applied to the training data to fit the performance metrics.

The second issue regarded the selection of variables to be included in the model. The R package Bloor[9] was used to build a regression model from a set of candidate predictor variables by entering and removing predictors based on Akaike information criterion, in a stepwise manner until there was no variable left to enter or remove.

Table 1: *Logistic regression for the best model on the test sample, Manufacturing*

	Dependent variable: Default.2018
	response
Total Debts/Total Assets	1.833***
	(0.415)
Turnover per employee	-0.457^{***}
	(0.102)
ROS	-0.043^{***}
	(0.012)
Tangible Assets/Total Assets	-1.930^{***}
	(0.433)
Net working capital	-0.0005^{***}
	(0.0002)
P&L	-0.002^{*}
	(0.001)
Constant	1.426***
	(0.532)
Observations	718
Log Likelihood	-422.658
Akaike Inf. Crit.	859.316
Note:	*p<0.1; **p<0.05; ***p<0.01

To perform the analysis on the test set, the best model for the in-sample simulations, based on the Area under the receiver operating curve (AUROC), was selected since this is an indicator of the accuracy of the model in case of imbalanced data. The resulting regression model is summarized in table 1 and the plot of the balanced sample is in figure 5. All the variables in the best model were significant. As expected, it was found that the gearing ratio was positively related to failure propensity, indicating that the higher the amount of debts the

[8] This method works with majority class. It reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles.

[9] Tools for building binary logistic regression models downloadable from https://blorr.rsquaredacademy.com/.

less likely a company is to survive. The profitability indicator of turnover per employee, which expresses the revenues from sales and services relative to the number of employees in the company, was associated with a lower probability of default, in line with Calabrese and Osmetti [14]. Profit margin revealed a sign which is consistent with the previous literature ([7] and [24]). The impact of tangible assets was negative. In particular collateral has a beneficial effects on borrowers' behaviour by increasing the probability of repayment, therefore reducing the probability of default of such companies (see Psillaki, Tsolas, and Margaritis [50]). The prediction model showed a good accuracy ratio (74.51%, with a standard range between 50% and 80%, see [43]). Sensitivity is 72.05% and specificity is 74.54% revealing an overall good classification.

The search for a better model led the way to different simulations, by excluding the variables which were highly correlated with the computation of the VIF[10], to make sure the baseline model offered a fair correct classification rate relative to other models. In particular, the baseline model referred to the general model as-is, coming from the stepwise process, while the other simulations were built upon three different variations: the first simulation did not include the variable 'sales', the second one the variable 'employees' and the third one did not include both the variable 'sales' and considered only firms with at least one employee. All in all, the baseline model seemed to be acceptable (Table 2), although the third simulation is more balanced in discriminating between bankrupted and non-bankrupted firms (see the higher H-Measure), therefore it was decided to move on to testing robust models.

Table 2: *Classification metrics on the test set, logistic regression via stepwise approach, Manufacturing*

Logistic regression	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7451	0.7205	0.7454	0.8178	0.3178	214	7262	21259	83
Simulation 1	0.6949	0.7244	0.6946	0.7942	0.2716	205	8065	18342	78
Simulation 2	0.7129	0.7032	0.7130	0.7947	0.2744	199	7580	18827	84
Simulation 3	0.7392	0.7348	0.7392	0.8119	0.3193	205	6872	19480	74

4.2.2 Robust methods

To perform the comparison, the glmrob package was used. The models were specified by giving a symbolic description of the linear predictor and a description of the error distribution. Two methods were selected, namely Mqle, which fits a generalized linear model using Mallows or Huber type robust estimators, as described in Cantoni and Ronchetti [15] and Cantoni and Ronchetti [16]. The other method BY, available for logistic regression (family = binomial) only, is the

[10] The Variance Inflation Factor quantifies the severity of the multicollinearity.

Bianco-Yohai estimator, where algorithmic parameters are const 0.5 and maxhalf 10 maxit 1000. The results are shown in table 3.

Table 3: *Classification metrics on the test set, robust methods, Manufacturing*

Model BY	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7428	0.7138	0.7431	0.8076	0.3081	212	7326	21195	85
Simulation 1	0.7457	0.6926	0.7463	0.8042	0.2959	196	6699	19708	87
Simulation 2	0.7475	0.6926	0.7481	0.8064	0.3069	196	6652	19755	87
Simulation 3	0.7595	0.7133	0.7600	0.8199	0.3423	199	6324	20028	80

Model Mqle	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7436	0.7104	0.7440	0.8080	0.3062	211	7302	21219	86
Simulation 1	0.7596	0.6855	0.7604	0.8013	0.3060	194	6326	20081	89
Simulation 2	0.7618	0.6890	0.7626	0.8026	0.3093	195	6270	20137	88
Simulation 3	0.7616	0.7133	0.7621	0.8210	0.3439	199	6270	20082	80

The accuracy ratio and the ROC curve were quite similar between glm and glmrob for the baseline model, however on the other three simulations the robust model performed better (figure 6) based on AUC, in particular in detecting non-defaulted firms, highlighted by a higher specificity.

Table 4: *Comparison between logistic and robust models, Manufacturing*

Dependent variable: default	*Logit*	*Mqle*	*BY*
Parameter Estimate			
Number of employees	-0.1559	-0.1652^{*}	-0.1464^{*}
	(0.098)	(0.0895)	(0.0857)
Total Debts/Total Assets	1.9801^{***}	2.1972^{***}	2.2994^{***}
	(0.4393)	(0.4167)	(0.4301)
Turnover per employee	-0.4261^{***}	-0.5202^{***}	-0.4941^{***}
	(0.1003)	(0.0972)	(0.1047)
ROS	-0.0297^{***}	-0.0527^{***}	-0.057^{***}
	(0.0107)	(0.0105)	(0.0106)
Tangible Assets/Total Assets	-1.9577^{***}	-1.6830^{***}	-1.7000^{***}
	(0.4719)	(0.4429)	(0.4554)
Net working capital	-0.0005^{**}	-0.0007^{***}	-0.0005^{***}
	(0.0002)	(0.0002)	(0.0001)
EBIT/Interest expenses	-0.0011	-0.0003	-0.002
	(0.0008)	(0.0008)	(0.0009)
Constant	1.4170^{*}	1.8405^{**}	1.6186^{**}
	(0.5998)	(0.5678)	(0.6197)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

To understand the extent to which outlying observations influenced the models, the value and significance of the estimated coefficients were compared. Looking at table 4, it can be observed that Mqle and BY delivered quite similar results in terms of classification, although they had different results for the variables included, see for example the change in the significance level for the Number of Employees (now significant at the 10% level). The findings for the manufacturing sector confirmed what pointed out by Hauser and Booth [31, p. 577]: *"at worst, the BY robust logistic regression makes no changes in the estimated regression coefficients and has the same classification and prediction results as ML logistic regression"*. On one side this is strong evidence that BY robust logistic regression should be used as a robustness check on ML logistic regression, as well as for prediction when outliers exist in the data set. On the other side one should consider that there could be no particularly outlying observations or that further robustness checks should be made. This introduced the further topic of discussion, the role of the outliers.

4.2.3 Identification of outliers

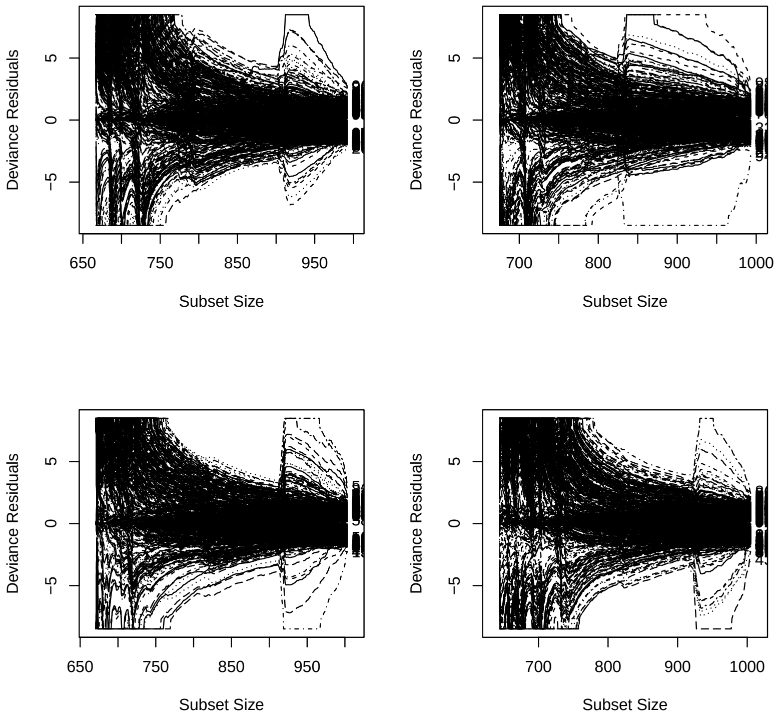
After having considered the Bianco-Yohai regression, the attention was focused on the detection of the outliers using the forward search algorithm to see whether an improvement of the estimation and the classification metrics could be achieved. Since most of the research in this field is actually pioneering in a way, Cheng [17] and Atkinson and Riani [9] were followed for the technical realization in Rstudio. But before diving into what it was actually done, the importance of this method should be reminded for two reasons: first of all, multiple outliers may strongly affect the generalized linear model fitted to data, as may unidentified distinct subset. The issue of "masking" could be found in the estimation outputs previously obtained, i.e. the coefficients, therefore this method in general intends to show the structure of the data step by step. Secondly, *"the forward analysis is not only an alternative way of looking at the data but also leads naturally to the definition and calculation of a robust and fully efficient forward search estimator"* [10], that is the purpose should be to obtain a model which should also have better performance measures, in terms for example of AIC (a more parsimonious model).

To start with, the variables of the best model presented above were selected and the cross-validation scheme was repeated, collecting the information of the training set and the validation set on one side and the hold-out sample on the other side. Next, the fwdglm function[11] was used to apply the forward search approach to robust analysis in generalized linear models. Plotting the results the first impression was that of many iterations in which there was a steadily upward trend in the deviance residuals meaning that outlying observations were present

[11] I thank Prof. Luca Scrucca for porting the model described in the book of Atkinson and Riani [9] to R.

and added (for reference, see the Forward Search graphical output in figure 1).

Figure 1: *Example of absolute values of deviance residuals as the subset size increases, Manufacturing*



Afterwards the goal was to design a way to reduce the effects of the outliers in the training set and the methodological procedure was the following:

1. By considering the last 30% of the forward search step-by-step process it was chosen to cut the process in the step where one could recognize influential observations, i.e. a cluster of firms entering the search and having absolute deviance residuals larger than 2.

2. The multiple masked outliers found were either deleted from the training set or their categorical class was switched from 0 to 1 or vice-versa. The second idea came from Atkinson and Riani [9, p.257] that showed in the vasoconstriction example that the deviance explained of the modified data improved by making this exchange[12].

The computation of the average values and the standard deviations of the variables (table 5) confirmed that the detected firms were indeed outliers. Of course

[12] *"It is to be expected that if these two observations are switched from one to zero the fit of the model will improve. We begin our numerical investigation of the effect of near perfect fit on t values and deviances by comparing analyses of the original data with data modified by making this exchange".* [...] *"modifying the data has caused an appreciable increase in the deviance explained by the model from 24.81 to 46.47. Since this is not a residual deviance but the difference between the residual deviance for the fitted model and the null model with just a constant, the values do have a meaning".*

many explanations for the misclassification of these firms could be found in the literature from an economic point of view, but here one may only recall all the studies related to ′zombiism′ ([3]) that focuses on insolvent firms, i.e. with bad indicators, which are indeed still functioning over a relatively long time period due to inefficient market competition and the relentless support of investors and government. The other side of the coin is represented by small and medium enterprises which could have had positive numbers in 2018, however for a numerous number of reasons faced a sudden turmoil, e.g. downward trend, strong competition, management swings. This kind of situations more related to macro-economic and culture variables rather than financial indicators makes the model imperfect and perfectible.

Table 5: *Outliers′ descriptive statistics characteristics, Manufacturing*

	Mean_0	Sd_0	p_0	p_{25}	p_{50}	p_{75}	p_{100}
Total Debts/Total Assets	0.84	0.18	0.09	0.77	0.86	0.93	4.28
ROS	-2.43	11.01	-50.00	-5.46	1.09	3.32	28.79
PL	-65.30	261.51	-4594.00	-29.00	0.00	4.00	485.00
Tangible Assets/Total Assets	0.15	0.17	0.00	0.02	0.08	0.21	0.96
Net working capital	-87.74	883.04	-22079.00	-50.00	8.00	64.00	8455.00
Turnover per employee	4.24	0.83	0.79	3.72	4.23	4.75	7.99
	Mean_1	Sd_1	p_0	p_{25}	p_{50}	p_{75}	p_{100}
	0.70	0.22	0.06	0.56	0.74	0.87	1.26
	2.69	10.26	-48.01	0.48	3.12	7.22	29.82
	5.70	154.40	-1520.00	-3.75	4.00	20.75	1078.00
	0.18	0.21	0.00	0.02	0.10	0.29	0.96
	117.25	950.29	-5591.00	-23.75	44.50	179.75	10409.00
	4.66	0.98	1.52	3.99	4.64	5.27	7.97

The results of the classification metrics for the in-sample (table 6) showed a slight increase in the default prediction of the survivors (better specificity) and a resulting fair correct classification rate in terms of AUC. The H-measure metric and the error rate revealed a better performance compared to the last step, i.e. the standard logistic regression. These results need to take into account the fact that the number of outliers of class 1 (defaulted) are on average twice the number of non-bankrupted firms discovered by the forward search, but overall the resulting validation models seem to be more predictive.

Table 6: *Classification metrics on the validation set, forward search, Manufacturing*

	Accuracy	Sensitivity	Specificity	AUC	H-Measure	Error rate
Last step	0.7048	0.7934	0.7039	0.8267	0.3188	0.2952
Deletion	0.7269	0.7934	0.7262	0.8278	0.3400	0.2731
Substitution	0.7292	0.7887	0.7286	0.8277	0.3404	0.2708

The kernel of the probability density function (figure 2) clarifies the nature of the outlying observations compared to the rest of the sample in the best training sample selected through the best AUC. Indeed one can see that the probability density functions are swapped for the misclassified firms, while the kernel for the modified training set (deletion or substitution) are squeezed to the left and right.

Figure 2: *Kernel probability density function, best training set, forward search, Manufacturing*

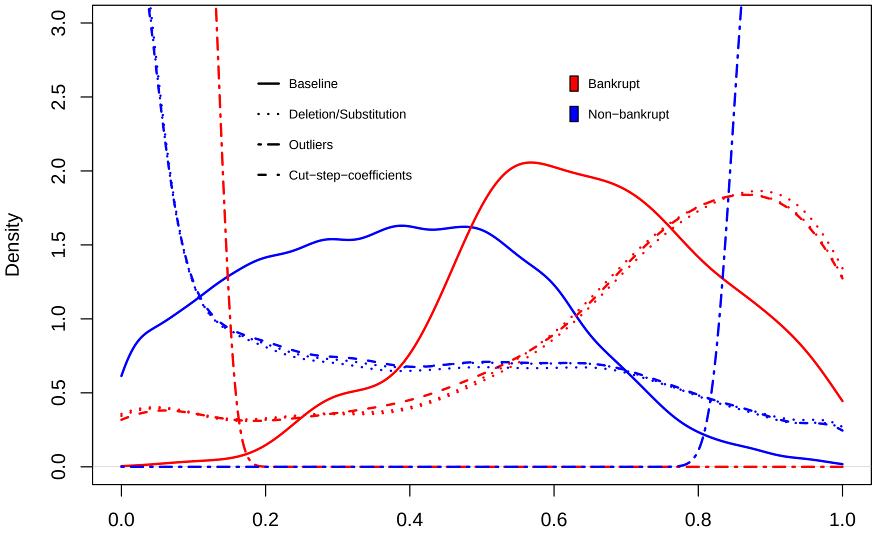


Figure 8 shows the trajectories of classification metrics for sensitivity and specificity during the forward search for the logit link regarding all the iterations ([30]). From the graph it is interesting to note that in almost all the simulations adding observations to the initial subset causes a decrease of error in classifying unhealthy firms (red lines). On the opposite the proportion of correctly classified solvent firms decreases very quickly (blue lines). This behaviour can be explained considering that observations which are included in the last steps are unhealthy firms which are financially similar to solvent firms and wrongly influence the classification rule (see also Grossi and Bellini [30]). Second, the spikes seen in the figure are indeed an indication of presence of outliers which distorts the metrics. The graph helps in explaining the higher sensitivity in the previous table 6. Remind that from a financial institution's point of view, it is more serious to misclassify an insolvent firm as healthy than the opposite, so that actually the forward search confirms the reliability of the logit classifier once again.

In the hold-out sample (table 7) the model which presented the best AUC for the training set was tested. The scores were globally quite similar to that of the logit model, which remained the best in terms of AUC and H-measure metrics, and that was not actually surprising since in general, it is difficult to classify correctly outlying observations in the testing set if these are eliminated or modified in the training set. Even if one picks the best model out of the validation list you still do not have the right balanced dataset and tools to reduce misclassification in the hold-out sample.

Table 7: *Classification metrics on the test set, forward search, Manufacturing*

	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Last step	0.7043	0.7542	0.7038	0.8091	0.3169	224	8449	20072	73
Deletion	0.7267	0.7340	0.7267	0.8031	0.3043	218	7796	20725	79
Substitution	0.7289	0.7340	0.7288	0.8029	0.3038	218	7735	20786	79

Table 8: *Estimates and statistics of the three models of the forward search, Manufacturing*

Dependent variable: default	*Last step*	*Deletion*	*Substitution*
Parameter Estimate			
Profit & Loss	-0.0009^{**}	-0.0126^{***}	-0.0133^{***}
	(0.0003)	(0.0016)	(0.0016)
Total Debts/Total Assets	2.0423^{***}	1.6286^{***}	1.6983^{***}
	(0.3525)	(0.4417)	(0.4392)
Turnover per employee	-0.4447^{***}	-0.6389^{***}	-0.6650^{***}
	(0.0848)	(0.1123)	(0.1120)
ROS	-0.0397^{***}	-0.0233^{*}	-0.0237^{*}
	(0.0081)	(0.0105)	(0.0105)
Tangible Assets/Total Assets	-1.6788^{***}	-4.1839^{***}	-4.3680^{***}
	(0.3763)	(0.5332)	(0.5306)
Net working capital	-0.0003^{**}	-0.0046^{***}	-0.0049^{***}
	(0.0001)	(0.0005)	(0.0005)
Constant	1.1214^{*}	2.9435^{***}	3.0520^{***}
	(0.4600)	(0.5800)	(0.5792)
AIC	1138.1	761.27	768.11
Pseudo R^2	0.1817	0.41	0.4476

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Once the performance of the three models was compared to see that the results were not so different each other, the analysis shed light on the significance of the models and their relative coefficients (table 8), from which one could say that the only difference was in the variable tangible assets/total assets: tangibles were associated with a reduction of 98% in the relative risk of default, compared to the 81% of the logit model. On the other side, leverage multiplied by 7.71 the probability of default compared to survived firms in the first model, whereas the odds ratio were lower for the forward search models (5.10 and 5.46 respectively). The deviance residuals looked good since they were close to being centred on 0 and were roughly symmetrical. However, taking a look at the Akaike Information Criterion and the Pseudo R^2 the simple logistic regression had less predictive power in the proportion of outcome compared to the robust models, indeed a point in favour of the forward search design.

4.3 Construction

In the last decade the building sector has experienced a declining path, as highlighted by the European Construction Sector Observatory [23]: *"total investment by the narrow construction sub-sector declined by 34.4%, from EUR 8.1 billion in 2010 to EUR 5.3 billion in 2020. At the same time, the gross operating rate of the broad construction sector, an indicator of the sector profitability, stood at 18.5% in 2018, being a 0.9 percentage point above the 2010 level (17.6%)"*. In 2018 the expectation of a continued modest rebound of Italian construction performance did not materialize, quite the opposite the sector continued experiencing subdued demand and further deteriorating profit margins. Investments in new residential buildings grew just 2%, and the backlog of unsold houses still amounted to 1.4 million units. From a financial point of view the sector is generally characterized by high indebtedness, poor financing conditions, lack of liquidity, bad payment experience. Lack of financial support also contributed to the pronounced increase in construction costs for residential buildings. After the pandemic, to sustain the recovery of the economy a series of tax rebate schemes for energy efficiency renovations have been implemented (Superbonus 110%, National Rental Fund, Earthquake Bonus) which followed other previous measures such as the Eco-Bonus approved with the 2016, 2018 and 2019 Stability Law. These tax deductions were supposed to be claimed for renovations carried out on the common parts of residential buildings and for interventions on the building envelope, aimed to improve energy performance, and at the same time relaunch a creaking sector, whose relationship with banks seems to be quite hard.

To define the sample the database previously built for 2018 was filtered using ATECO 2007 codes from 41 - *Construction of buildings* - to 43 - *Specialised construction activities*, collecting all the variables and the dependent variable default in 2018, which is equal to 1 if the firm considered defaulted in 2019, otherwise 0 if it remained alive. A strictest definition of SMEs was implemented, namely only firms with annual turnover of fewer than 50 million euros, the number of employees lower than 250 and a balance sheet of fewer than 43 million euros were retrieved. The resulting dataset contained 125.042 SMEs with a proportion of 3.4% defaulted firms.

The average values and standard deviations of the variables separately for survived and defaulted firms are reported in table 24. In particular, considering the five categories it can be observed that, as expected, active firms show better coefficients: indeed the gearing ratio is higher for troublesome firms (6.07 vs. 0.80); current, liquidity and solvency ratio show the resilience of active firms; profitability variables (such as ROE, ROA, ROS, EBITDA/Total Assets) exhibit negative signs for defaulted firms; coverage reveals weaknesses in defaulted firms (EBIT/Interest expenses and Debt/EBITDA ratio are negative); activity ratios are quite similar. The exception can be found in the leverage ratios since debt/equity ratio and leverage are higher for survived firms, so that one does not expect these ratios to be in the regression and to classify firms correctly.

4.3.1 Glm

Table 9: *Logistic regression estimates for the best model on the test sample, Construction*

	Dependent variable: Default.2018
	response
Total Debts/Total Assets	0.3464^{*}
	(0.17)
Sales	-0.1572^{***}
	(0.0224)
ROE	-0.0043^{***}
	(0.0013)
Tangible Assets/Total Assets	-2.056^{***}
	(0.2652)
Net working capital	-0.0002^{*}
	(0.0001)
Current Assets	-0.1296^{***}
	(0.0387)
Total shareholders funds	-0.0002
	(0.0001)
Constant	1.1956^{***}
	(0.1870)
Observations	2031
Log Likelihood	-1242.438
Akaike Inf. Crit.	2500.9
Note:	*p<0.1; **p<0.05; ***p<0.01

The analysis and the research design were devised as that of the manufacturing sector so one can refer to section 4.2 for all the detailed explanations. Of course it was necessary to take into account the fact that the number of missing values in certain ratios were different depending on the sector, for example here the variables ROS, Turnover per employee and added value per employee were excluded in the cleaning procedure because of missing values above 15%, while in the manufacturing sector they were included. Ultimately, the performance of the logistic regression via stepwise approach (both direction backward and forward) showed results in which one could see the unbalance between specificity and sensitivity, indeed the H-measures and AUC were lower compared to the manufacturing sector (see the first block of table 10). Among the four models, the simulation 3 was the best in terms of AUC and H-measure (with a slightly lower sensitivity than the baseline).

The variables included in the baseline model (table 9) were satisfying (in line with other papers, see [55]) with sales, ROE, tangibles and current assets significant at the 1% level. The signs of the indicators were correct, being only the gearing ratio positive and significant at the 10% level. Four out of seven indicators were noticed also in the model for manufacturing obtained from the stepwise logistic regression, with a clear difference: there the gearing ratio was significant with a large effect, here the coefficient is lower and not significant at the 5%.

4.3.2 Robust methods

The different simulations and methods showed that overall, as in the manufacturing industry, the accuracy ratio and the ROC curve were quite similar between `glm` and `glmrob`, where the only aspect which could be relevant is the improvement in the classification of non-defaulted firms using the Mqle method: specificity increased in the baseline model and in the three simulations, with sensitivity slightly lower than BY (table 10). It should be recalled that by adopting a stricter definition of Small and Medium Enterprise and eliminating observations with lacking information, outliers could have been already excluded, which in turn explains why the classifiers did not present such patent differences. One important thing to highlight is the equivalence between the baseline model and the second simulation, which is a consequence of the variable "Employees" not having effect at all in changing the best model and therefore the classification metrics.

Table 10: *Classification metrics on the test set, logistic regression via stepwise approach and robust methods, Construction*

Logistic regression	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7039	0.6003	0.7060	0.7179	0.1682	356	8945	21476	237
Simulation 1	0.7352	0.5900	0.7380	0.7246	0.1862	341	7921	22312	237
Simulation 2	0.7039	0.6003	0.7060	0.7179	0.1682	356	8945	21476	237
Simulation 3	0.7148	0.5870	0.7167	0.7307	0.2007	172	5781	14621	121
Model BY	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7068	0.5970	0.7090	0.7175	0.1694	354	8853	21568	239
Simulation 1	0.7366	0.5848	0.7395	0.7244	0.1852	338	7876	22357	240
Simulation 2	0.7068	0.5970	0.7090	0.7175	0.1694	354	8853	21568	239
Simulation 3	0.7443	0.5529	0.7470	0.7148	0.1975	162	5161	15241	131
Model Mqle	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7386	0.5750	0.7418	0.7179	0.1690	341	7854	22567	252
Simulation 1	0.7379	0.5865	0.7408	0.7254	0.1897	339	7837	22396	239
Simulation 2	0.7386	0.5750	0.7418	0.7179	0.1690	341	7854	22567	252
Simulation 3	0.7562	0.5358	0.7594	0.7303	0.1999	157	4909	15493	136

A review of the estimated coefficients, considering the first simulation where AUC was higher, confirmed that BY and Mqle logistic regression and ML logistic regression were not that different, although more similarities in terms of significance could be found between Mallows quasi-likelihood estimators and logit ones (e.g. total shareholders funds significant for the Logit and Mqle models but not for BY, see table 11). The results in terms of business sector is in line with the previous literature [20], [19].

Table 11: *Comparison between logistic and robust models, Construction*

Dependent variable: default	*Logit*	*Mqle*	*BY*
Parameter Estimate			
Total Assets Turnover	-0.1820^{***}	-0.2060^{***}	-0.1926^{***}
	(0.0634)	(0.0643)	(0.0720)
Total Debts/Total Assets	0.4167^{**}	0.3152^{*}	0.4610^{***}
	(0.1750)	(0.1783)	(0.1723)
Total shareholders funds	-0.0003^{**}	-0.0004^{**}	-0.0002
	(0.0001)	(0.0002)	(0.0001)
ROE	-0.0043^{***}	-0.0054^{***}	-0.0053^{***}
	(0.0013)	(0.0013)	(0.0014)
Tangible Assets/Total Assets	-2.175^{***}	-2.1361^{***}	-2.217^{***}
	(0.2607)	(0.2634)	(0.2925)
Current Assets	-0.2284^{***}	-0.2338^{***}	-0.2408^{***}
	(0.0387)	(0.0383)	(0.0426)
Number of Employees	-0.1845^{***}	-0.1941^{***}	-0.1903^{***}
	(0.0615)	(0.0631)	(0.0729)
Constant	1.3537^{***}	1.4876^{***}	1.399^{***}
	(0.2025)	(0.2068)	(0.2210)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

4.3.3 Identification of outliers

From then the forward search algorithm was applied to identify and remove the outliers to get robust estimates and attempt to have better classification results. The results in table 12 revealed that the sensitivity and the H-measure were higher in the case of deletion of outliers or substitution (1 to 0 or vice-versa) than in the logit output, while specificity gave an equal result only in the case of substitution. Overall, if one compares the construction sector with the previous one, the accuracy is better (0.7871/0.7629/0.7867 vs. 0.7048/0.7269/0.7292) because specificity is bigger, but sensitivity is quite lower.

Graphically the kernel of the probability density function (fig. 9) confirmed the output given with little difference between the deletion and the substitution

Table 12: *Classification metrics on the validation set, forward search, Construction*

	Accuracy	Sensitivity	Specificity	AUC	H-Measure	Error_rate
Last step	0.7871	0.5972	0.7908	0.7576	0.2085	0.2129
Deletion	0.7629	0.6422	0.7652	0.7572	0.2237	0.2371
Substitution	0.7867	0.6137	0.7901	0.7569	0.2234	0.2133

procedure, with the outliers correctly unrecognised by a standard model. The descriptive statistics of these outlying firms (table 13) were in line with what expected, expect for the gearing ratio, which is somehow counter-intuitive, with a value for the bankrupted firms higher than that of the defaulted one. In a way, this is reflected in the coefficients of the best training models reported below (table 14), with the gearing ratio being not significant for the "constructed" logit models on one side. What draws the attention here is the different effect of the ratio tangible assets/total assets compared to the "last step" model.

Table 13: *Outliers' descriptive statistics characteristics, Construction*

	Mean_0	Sd_0	p_0	p_{25}	p_{50}	p_{75}	p_{100}
Total Debts/Total Assets	0.78	1.37	0.00	0.60	0.86	0.96	112.93
ROE	-19.03	38.20	-150.00	-30.35	-6.66	0.07	114.59
Sales	1.36	1.91	0.00	0.00	0.00	2.77	10.17
Tangible Assets/Total Assets	0.03	0.10	0.00	0.00	0.00	0.01	1.00
Net working capital	90.05	673.06	-15699.00	3.00	22.00	136.00	13295.00
Total shareholders funds	56.51	390.71	-26569.00	9.00	18.00	58.00	11114.00
	Mean_1	Sd_1	p_0	p_{25}	p_{50}	p_{75}	p_{100}
	1.12	6.90	0.02	0.46	0.74	0.90	179.52
	20.24	42.22	-147.84	-1.25	12.42	47.51	146.05
	4.69	2.31	0.00	3.66	5.11	6.23	10.41
	0.12	0.22	0.00	0.00	0.02	0.13	1.00
	259.31	1153.18	-3347.00	7.00	45.00	200.50	24230.00
	174.50	542.13	-1729.00	15.00	52.00	161.50	8704.00

The plot of the classification metrics for the training set (fig. 10) in 100 simulations showed the hiatus between sensitivity and specificity along the steps of the forward search, as opposed to the manufacturing sector, in which a cross and an improvement in recognizing the defaulted firms was seen. One could conclude that the model is unable to progressively obtain a more balanced trade-off between correctly classification of defaulted firms and non-defaulted ones. Given that the forward search algorithm did not work well for the hold-in sample it was decided not to proceed with the test set, for which the average II error type is above 0.40. Results suggest that it is actually difficult to predict the default status in the building industry due to not homogeneous aggregate information.

Table 14: *Estimates and statistics of the three models of the forward search, Construction*

Dependent variable: default	*Last step*	*Deletion*	*Substitution*
Parameter Estimate			
Sales	-0.2126^{***}	-0.7488^{***}	-0.5668^{***}
	(0.0195)	(0.0520)	(0.0356)
Total Debts/Total Assets	0.2553^{*}	-0.1232	-0.1000
	(0.1364)	(0.1346)	(0.0759)
Total shareholders funds	-0.0002^{*}	-0.0040^{***}	-0.0033^{***}
	(0.0001)	(0.0006)	(0.0005)
ROE	-0.0051^{***}	-0.021^{***}	-0.0185^{***}
	(0.0012)	(0.0021)	(0.0018)
Tangible Assets/Total Assets	-1.9158^{***}	-14.4678^{***}	-11.6784^{***}
	(0.2555)	(1.1842)	(0.8980)
Net working capital	-0.0002^{**}	-0.0026^{***}	-0.0021^{***}
	(0.0001)	(0.0002)	(0.0002)
Constant	0.8560^{***}	4.8073^{***}	3.5270^{***}
	(0.1334)	(0.3276)	(0.2077)
AIC	2531.6	1028	1329.2
Pseudo R^2	0.1177	0.5470	0.5263

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

4.4 Trade

The last part of this chapter focused on the wholesale and retail trade sectors (Ateco code 45, 46, 47) since it represents the backbone of the Italian economy together with the manufacturing and the construction industries. To remind the classification, wholesalers act as marketing intermediaries that neither produce nor consume the finished product, but instead sell to retailers, other merchants, and/or to industrial, institutional, and commercial users. The type of goods traded can be agricultural raw materials and live animals, food, beverages and tobacco, machinery, equipment and supplies and so on. Retailers instead are organised to sell merchandise (generally without transformation) in smaller quantities to the general public for personal or household consumption, and to other business and institutional clients.

Taking a look at ISTAT statistics [33], in the retail sales one can see an up-and-down trend from 2018 (the Eu country with lowest growth) till the beginning of the 2020 and than a huge drop due to the Covid-19 restrictions. Analogously, the turnover for the wholesale and retail trade and repair of motor vehicles and motorcycles flexed between 2018 and 2020. The persistent tensions between the USA and China with protectionist barriers, the uncertainties related to Brexit negotiations and the slowdown of the automotive branch reduced confidence and investments towards the Italian economy well before the pandemic hit and were only partly offset by the developing e-commerce.

To define the sample the database previously built for 2018 was filtered using ATECO 2007 codes from 45 - *Wholesale and retail trade and repair of motor vehicles and motorcycles* - to 47 - *Retail trade, except of motor vehicles and motorcycles*, collecting all the variables and the dependent variable default in 2018, which is equal to 1 if the firm considered defaulted in 2019, otherwise 0 if it remained alive. As in the previous example, a strictest definition of SMEs was implemented, namely only firms with annual turnover of fewer than 50 million euros, the number of employees lower than 250 and a balance sheet of fewer than 43 million euros were retrieved. The resulting dataset contained 153.571 SMEs with a proportion of 1.83% defaulted firms.

The average values and standard deviations of the variables separately for survived and defaulted firms are reported in table 25. Even in this case, considering the five categories it can be seen that, as expected, active firms showed better coefficients for 2018: indeed the gearing ratio was higher for troublesome firms (7.42 vs. 0.76); current, liquidity and solvency ratio explained the resilience of active firms, even though the defaulted firms had also positive signs; profitability variables (such as ROE, ROA, ROS, EBITDA/Total Assets) exhibited negative signs for defaulted firms; coverage revealed weaknesses in defaulted firms (EBIT/Interest expenses and Debt/EBITDA ratio are negative); activity ratios were quite similar. All in all, most indicators confirmed what a bank expects to see if a counterpart downgrades or goes bankrupt on one side or stays alive and manages to pay the obligations on the other, even though non-defaulted SMEs' financial and non-financial information are not generally complete or sufficient for an exhaustive valuation and monitoring (see the high percentage of missing values for some ratios, a remarkable characteristic for the type of firms covered).

4.4.1 Glm

The analysis and the research design were devised as that of the building and manufacturing sector so the reader can refer to section 4.2 for all the detailed explanations. Of course one needs to take into account the fact that the number of missing values in certain ratios were different depending on the sector, for example here the variables ROI, Turnover per employee and added value per employee were excluded in the cleaning procedure due to missing values above 15%. It can be noticed that the performance of the logistic regression via stepwise approach (both direction backward and forward) was more in line with the construction sector, in terms of AUC and H-measure, with a lower sensitivity. Among the different simulations, the second one[13] presented the best classification metrics, with the variables sales, ROS[14], tangibles and gearing ratio significant at the 5% level (table 15).

Table 15: *Logistic regression estimates for the best model on the test sample, Trade*

	Dependent variable: Default.2018
	response
Total Debts/Total Assets	0.3634^{**}
	(0.1790)
Sales	-0.2277^{***}
	(0.031)
ROS	-0.0350^{***}
	(0.005)
Tangible Assets/Total Assets	-1.7345^{***}
	(0.3021)
Net working capital	-0.0002^{*}
	(0.0001)
Total shareholders funds	-0.0002
	(0.0002)
Constant	1.1374^{***}
	(0.2070)
Observations	2083
Log Likelihood	-1293.14
Akaike Inf. Crit.	2600.3
Note:	*p<0.1; **p<0.05; ***p<0.01

The sign of the indicators were plausible, such that a higher indebtedness leads to a higher probability of default, while a positive sign of equity, tangibles, turnover and short-term business solvency means a negative effect on the probability of default of the counterparty. Even in this case the effect of the total debts/total assets ratio is not comparable with that of the manufacturing sector (0.3634 vs. 1.833) but the other coefficients are indeed similar.

4.4.2 Robust methods

For the robust models BY and Mqle, table 16 contains all the information about the classification metrics for the hold-out sample. A comparison revealed that the robust checks led overall to a better classification for survived firms (average specificity 0.7648 and 0.7813 against 0.7303 of the logistic regression),

[13] That is, leaving the variable 'Employees' out of the sample.

[14] Although it may sound strange, the variance-inflation factor did not assess a strong collinearity between sales and ROS

reflected in a higher accuracy, nonetheless sensitivity remained as is (or got even worse, see the column corresponding to the simulations for Mqle). A positive aspect about robust methods was also the H-Measure, which was higher than the stepwise logistic regression for all the simulations presented.

Table 16: *Classification metrics on the test set, logistic regression via stepwise approach and robust methods, Trade*

Logistic regression	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7325	0.5737	0.7352	0.7250	0.1823	405	11012	30578	301
Simulation 1	0.7163	0.5775	0.7186	0.7087	0.1649	425	12095	30892	311
Simulation 2	0.7464	0.5897	0.7491	0.7293	0.1810	434	10787	32200	302
Simulation 3	0.7161	0.5723	0.7181	0.7151	0.1990	273	9555	24340	204
Model BY	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.7500	0.5878	0.7528	0.7325	0.1885	415	10283	31307	291
Simulation 1	0.7554	0.5503	0.7589	0.7087	0.1781	405	10363	32624	331
Simulation 2	0.7638	0.5788	0.7670	0.7276	0.1924	426	10017	32970	310
Simulation 3	0.7772	0.5388	0.7806	0.7193	0.2060	257	7437	26458	220
Model Mqle	Accuracy	Sensitivity	Specificity	AUC	H-Measure	TP	FP	TN	FN
Baseline model	0.8007	0.5255	0.8054	0.7314	0.1940	371	8093	33497	335
Simulation 1	0.7652	0.5462	0.7690	0.7141	0.1863	402	9931	33056	334
Simulation 2	0.7680	0.5802	0.7712	0.7292	0.1980	427	9834	33153	309
Simulation 3	0.7763	0.5430	0.7796	0.7214	0.2064	259	7472	26423	218

The output in terms of coefficients for the glm and glmrob is shown in table 17 and considers the baseline models. The AUC of the BY estimation was the best (0.7325) and the Mqle estimation had better performance than the logit regression in terms of specificity (0.8054 vs. 0.7352), even though the sensitivity was really low (0.5255 vs. 0.5737). What stands out from the table is the different magnitude from the logistic regression in the variables total debts/total assets, cash flow/EBITDA[15] (not significant for the robust models) and employees (not significant for the BY estimator). The signs were not conflicting each other.

4.4.3 Identification of outliers

From then the forward search algorithm was applied to identify and remove the outliers to get robust estimates and try to have better classification results. In table 18, which contains the model with the best AUC for the three simulations (last step equivalent to logistic regression, deletion procedure of outliers and substitution 1 to 0 and vice-versa), it was observed that in the case of deletion or substitution the AUC and the H-measure were higher than the logit output. A positive aspect was found in terms of accuracy, reflecting the fact that the misclassified firms recognized and thrown out of the sample, thanks to the forward algorithm, were above all defaulted firms. Indeed the robust estimation

[15] The sign here is reversed, since the ratio actually assesses a company's efficiency in converting its profits into cash and generally a higher cash conversion ratio is better than a lower.

Table 17: *Comparison between logistic and robust models, Trade*

Dependent variable: default	*Logit*	*Mqle*	*BY*
Parameter Estimate			
Sales	-0.2046^{***}	-0.2063^{***}	-0.2544^{***}
	(0.0411)	(0.0386)	(0.0405)
Total Debts/Total Assets	0.6485^{***}	1.3181^{***}	1.1201^{***}
	(0.1969)	(0.1918)	(0.1856)
Net working capital	-0.0002^{*}	-0.0002^{**}	-0.0001
	(0.0001)	(0.0001)	(0.0001)
ROS	-0.0347^{***}	-0.0415^{***}	-0.0356^{***}
	(0.0053)	(0.0047)	(0.0043)
Tangible Assets/Total Assets	-1.5771^{***}	-1.5291^{***}	-1.7072^{***}
	(0.3082)	(0.2954)	(0.3131)
Cash Flow/EBITDA	0.0601^{*}	0.0048	0.0147
	(0.0309)	(0.0127)	(0.0137)
Number of Employees	-0.2287^{***}	-0.1694^{**}	-0.0641
	(0.0724)	(0.0681)	(0.0692)
Total Assets Turnover	0.0886^{*}	0.0678	0.0890^{*}
	(0.0523)	(0.0480)	(0.0473)
Constant	0.7693^{***}	0.4387^{**}	0.6859^{**}
	(0.2316)	(0.2199)	(0.2265)

Note: ${}^{*}p < 0.1$; ${}^{**}p < 0.05$; ${}^{***}p < 0.01$

limited the influence of the outliers by lowering the error rate and increasing correct detection of non-defaulted firms from 0.7345 to 0.7630/0.7645 respectively. Since the AUC was well above 70% the results could not be considered of poor quality, but the classification metrics of the manufacturing sample remain stronger.

Table 18: *Classification metrics on the validation set, forward search, Trade*

	Accuracy	Sensitivity	Specificity	AUC	H-Measure	Error rate
Last step	0.7328	0.6364	0.7345	0.7503	0.2203	0.2672
Deletion	0.7606	0.6194	0.7630	0.7529	0.2431	0.2394
Substitution	0.7623	0.6272	0.7646	0.7554	0.2480	0.2377

Graphically the kernel of the probability density function once again (fig. 11) confirmed the output given with little difference between the deletion and the substitution procedure, with the outliers correctly misclassified by a standard model (see also the table 19 for the details of the descriptive statistics). What is relevant here is to acknowledge once again the fact that the outliers presented the expected mean for all the variables.

Table 19: *Outliers' descriptive statistics characteristics, Trade*

	Mean_0	Sd_0	p_{0}	p_{25}	p_{50}	p_{75}	p_{100}
Total Debts/Total Assets	0.86	0.32	0.02	0.75	0.90	0.98	12.96
ROS	-8.97	12.57	-49.98	-14.99	-5.13	0.46	29.01
Sales	4.69	1.69	0.00	3.78	4.79	5.71	10.81
Tangible Assets/Total Assets	0.08	0.15	0.00	0.00	0.02	0.08	0.99
Net working capital	-43.78	592.19	-20341.00	-9.00	5.00	34.00	10422.00
Number of employees	0.82	0.83	0.00	0.00	0.69	1.39	5.50
	Mean_1	Sd_1	p_{0}	p_{25}	p_{50}	p_{75}	p_{100}
	0.74	0.28	0.05	0.57	0.79	0.91	2.83
	4.82	7.91	-34.42	0.84	3.03	8.32	29.28
	5.99	1.88	0.00	4.84	6.10	7.24	10.45
	0.10	0.17	0.00	0.00	0.03	0.12	0.97
	136.87	669.44	-5937.00	2.00	31.50	144.00	11837.00
	1.11	0.96	0.00	0.00	1.10	1.79	4.53

Chapter 5

Conclusion

This thesis intended to investigate to what extent robustification of generalized linear model through different methods could improve the classification metrics, especially discovering defaulted enterprises, and the impact on the estimated coefficients. Using financial ratios and qualitative data, three important sectors of the Italian economy were selected, in which SMEs are a vast majority, to design and compare the well-known logistic regression with the three robust checks (Bianco-Yohai, Mallows quasi-likelihood estimator and Forward Search), to take possible outliers into account. From the analysis above one can confirm that there are indeed key informative indicators which are explicative for the classification across sectors, namely turnover (and the related profit margin), the gearing ratio (a measure of a firm's total financial health), tangibles (a collateral safeguard for banks), net working capital (a measure of a company's liquidity and short-term financial health) and the turnover per employee (activity ratio), while others are taken into consideration depending on the sector, see for example EBIT/interest expenses, total shareholders' fund and number of employees. This is an important aspect to consider when dealing with portfolio analysis in terms of credit risk in order *"to capture the possible effects of the typically diverse (economic and financial, structural and behavioural) profiles of firms operating in"* [...] *"different categories of business sectors"* Ciampi [18, p. 1017].

Taking a look at the different sectors, the manufacturing one gives the better classification metrics in terms of Area under the Curve and H-Measure, while the other two are less predictive. The different behaviour is not a novelty, rather it was observed also in other studies such as [20] and [19], hence the importance of using a sectoral approach to deal with classification. The aim of testing robust models was accomplished, although results did not show such significant improvements in the estimation. BY and Mqle models give results which are slightly better than the ML estimation, with some differences being observed in the coefficients. The lesson is that the use of BY robust logistic regression provides another tool to analyse the ML regression results, improving classification accuracy as far as outliers are detected in the sample. Along these lines, the forward search is an extremely powerful tool to recognize masking outliers in a step-by-step process and on the whole managed to get better training performances by little, so that it could be considered as a starting point to build more advanced model to assess misclassified firms. Beyond that, the conclusion wants to remind the importance of the broadly used logit model as a cornerstone, confirmed by the words of Hauser and Booth [31, p. 581]: *"robust logistic regression should be used as a robustness check on ML logistic regression"*.

Limitations This thesis has of course some limitations. Since bankruptcy is legislated differently depending on the country of interest, there are constraints when generalizing studies. A bankrupted company in Italy may not be considered failed in another country of the EU because of how the countries rule when firms become financially distressed. Moreover, in this study the definition of default was strict, but there is indeed literature showing that different definitions of financial distress prove different potential in bankruptcy prediction [40].

Several bankruptcy prediction models were proposed within this research area, with some features in common, for example cross-validation schemes. In this study the pair sampling was used in the training set, but no further rebalancing schemes were implemented [57]. In reality, one should always keep in mind the bankruptcy rate is not as high as healthy companies which makes the proportion off, hence the imbalanced dataset.

This study is concise in the way it selected the variables to be entered in the model based on previous consolidated literature to avoid complexity, storage troubles and bizarre results. However, an increasing shift can be seen from the accounting-based variables to non-financial information regarding, e.g. the geographical area (spatial dependence [12]), the innovation-related variables, or the management ability. Updating credit scoring models with machine learning which is able to handle more variables together linked with non-linear relationship, is a new focus, with the aim of trying to explicate and interpret such modelling.

This thesis did not want to be exhaustive on the topic of default classification, but rather it was a journey to explore a few statistical methods. Default prediction, as consistently proved, relies in two milestones: logistic regression, which seems to be an evergreen method to classify with overall accuracy around 75% in general, and firm's financial and economics ratios as ingredients, even though relationship lending and soft information as well as sustainability indicators and will play a major role in future studies. There is always room for improvements.

Appendix

Table 20: *Location of the ratios and variables considered in the literature*

Most relevant coefficients		Academic papers
Leverage	Short-term Debt/Equity	Dietsch and Petey [26]; Altman and Sabato [5]; Altman, Sabato, and Wilson [6]; Pederzoli and Torricelli [49]; Psillaki, Tsolas, and Margaritis [50]; Angilella and Mazzù [8]; Altman, Esentato, and Sabato [4]
	Debt/Equity	Ciampi et al. [20]; Sohn and Kim [54]; Kim and Sohn [36]; Psillaki, Tsolas, and Margaritis [50]; Lin, Ansell, and Andreeva [40]; Calabrese and Osmetti [14]; Ciampi and Gordini [19]; Andreeva, Calabrese, and Osmetti [7]
	Assets/Equity	Sohn and Kim [54]; Pederzoli and Torricelli [49]; Wolter and Rösch [58]; Altman, Esentato, and Sabato [4]; Altman, Esentato, and Sabato [4]
	Equity/Debt	Altman [2]; Dietsch and Petey [26]; Altman and Sabato [5]; Altman, Sabato, and Wilson [6]
	Short-term Debt/Total Debt	Altman, Esentato, and Sabato [4]
	Short-term Debt/Total Assets	Michala, Grammatikos, and Ferreira Filipe [42]; Wolter and Rösch [58]; Sigrist and Hirnschall [53]; Altman, Esentato, and Sabato [4]
	Total Debts/Total Assets	Ohlson [47]; Altman and Sabato [5]; Ciampi et al. [20]; Psillaki, Tsolas, and Margaritis [50]; Lin, Ansell, and Andreeva [40]; Mannarino and Succurro [41]; Ciampi and Gordini [19]; Ciampi [18]; Sigrist and Hirnschall [53]; Altman, Esentato, and Sabato [4]

Liquidity	Cash/Total Assets	Dietsch and Petey [26]; Altman and Sabato [5]; Altman, Sabato, and Wilson [6]; Pederzoli and Torricelli [49]; Mannarino and Succurro [41]; Wolter and Rösch [58]; Angilella and Mazzù [8]; Altman, Esentato, and Sabato [4]
	Intangible Assets/Total Assets	Altman and Sabato [5]; Psillaki, Tsolas, and Margaritis [50]; Angilella and Mazzù [8] Altman, Esentato, and Sabato [4]
	Tangible Assets/Total Assets	Psillaki, Tsolas, and Margaritis [50]; Wolter and Rösch [58]; Altman, Esentato, and Sabato [4]
	Net working capital/Total Assets	Altman [2]; Ohlson [47]; Dietsch and Petey [26]; Altman and Sabato [5]; Altman, Sabato, and Wilson [6]; Pederzoli and Torricelli [49]; Psillaki, Tsolas, and Margaritis [50]; Altman, Esentato, and Sabato [4]
	Quick Assets/Current Assets	Altman, Sabato, and Wilson [6]
	Current ratio	Ohlson [47]; Altman, Sabato, and Wilson [6]; Ciampi et al. [20]; Pederzoli and Torricelli [49]; Lin, Ansell, and Andreeva [40]; Calabrese and Osmetti [14]; Ciampi and Gordini [19]; Mannarino and Succurro [41]; Ciampi [18]; Gabbianelli [28]; Andreeva, Calabrese, and Osmetti [7]; Sigrist and Hirnschall [53]; Altman, Esentato, and Sabato [4]
	Quick ratio	Dietsch and Petey [26]; Ciampi et al. [20]; Calabrese and Osmetti [14]; Ciampi and Gordini [19]; Ciampi [18]; Andreeva, Calabrese, and Osmetti [7]; Altman, Esentato, and Sabato [4]
	Solvency ratio	Calabrese and Osmetti [14]; Andreeva, Calabrese, and Osmetti [7]
Profitability	ROE	Sohn and Kim [54]; Ciampi et al. [20]; Kim and Sohn [36]; Lin, Ansell, and Andreeva [40]; Calabrese and Osmetti [14]; Mannarino and Succurro [41]; Ciampi and Gordini [19]; Ciampi [18]; Andreeva, Calabrese, and Osmetti [7]; Gabbianelli [28]; Altman, Esentato, and Sabato [4]

	ROA	Altman and Sabato [5]; Sohn and Kim [54]; Pederzoli and Torricelli [49]; Kim and Sohn [36]; Nehrebecka [46]; Altman, Esentato, and Sabato [4]
	ROTA	Altman [2]; Ohlson [47]; Pederzoli and Torricelli [49]; Psillaki, Tsolas, and Margaritis [50]; **lin**; Michala, Grammatikos, and Ferreira Filipe [42]; Ciampi and Gordini [19]; Angilella and Mazzù [8];
	EBITDA/Total Assets	Altman and Sabato [5]; Psillaki, Tsolas, and Margaritis [50]; Wolter and Rösch [58]; Altman, Esentato, and Sabato [4]
	Retained earnings/Total assets	Altman [2]; Altman and Sabato [5]; Altman, Sabato, and Wilson [6]; Sigrist and Hirnschall [53]; Altman, Esentato, and Sabato [4]
	ROI	Ciampi et al. [20]; Kim and Sohn [36]; Ciampi and Gordini [19]; Calabrese and Osmetti [14];Ju and Sohn [35]; Ju, Jeon, and Sohn [34]; Ciampi [18]; Gabbianelli [28]
	ROCE	Lin, Ansell, and Andreeva [40]; Andreeva, Calabrese, and Osmetti [7]; Nehrebecka [46]
	ROS	Dietsch and Petey [26]; Altman and Sabato [5]; Sohn and Kim [54]; Ciampi et al. [20]; Kim and Sohn [36]; Pederzoli and Torricelli [49]; Psillaki, Tsolas, and Margaritis [50]; Ciampi and Gordini [19]; Ciampi [18]; Gabbianelli [28]; Altman, Esentato, and Sabato [4]
	Cash flow/EBITDA	Gentry, Newbold, and Whitford [29]; Ciampi et al. [20]; Ciampi and Gordini [19]; Ciampi [18]
	Turnover per employee, Added value per employee, Long term assets per employee	Sohn and Kim [54]; Ciampi et al. [20]; Lin, Ansell, and Andreeva [40]; Calabrese and Osmetti [14]; Ciampi and Gordini [19]; Ciampi [18]
	R&D/Sales	Angilella and Mazzù [8]
Coverage	EBITDA/Interest expenses	Altman and Sabato [5]; Altman, Sabato, and Wilson [6]; Ciampi et al. [20]; Michala, Grammatikos, and Ferreira Filipe [42]; Ciampi and Gordini [19]; Ciampi [18]; Altman, Esentato, and Sabato [4]

	EBIT/Interest expenses	Altman and Sabato [5]; Pederzoli and Torricelli [49]; Mannarino and Succurro [41]; Andreeva, Calabrese, and Osmetti [7]; Altman, Esentato, and Sabato [4]
	Interest expenses/turnover	Ciampi et al. [20]; Ciampi and Gordini [19]; Ciampi [18]; Altman, Esentato, and Sabato [4]
	Bank loans/turnover	Ciampi et al. [20]; Calabrese and Osmetti [14]; Ciampi and Gordini [19]; Ciampi [18]
	Net financial position/turnover	Ciampi et al. [20]; Ciampi and Gordini [19]; Ciampi [18]
	Cost of debit (%)	Ciampi et al. [20]; Ciampi and Gordini [19]; Ciampi [18]
	Debt/EBITDA	Ciampi et al. [20]; Ciampi and Gordini [19]; Lin, Ansell, and Andreeva [40]; Calabrese and Osmetti [14]; Altman, Esentato, and Sabato [4]
	Cash flow/Debt	Ohlson [47]; Dietsch and Petey [26]; Ciampi et al. [20]; Lin, Ansell, and Andreeva [40]; Michala, Grammatikos, and Ferreira Filipe [42]; Ciampi and Gordini [19]; Ciampi [18]; Gabbianelli [28]; Nehrebecka [46]
Activity	Sales/Total Assets	Altman [2]; Altman and Sabato [5]; Sohn and Kim [54]; Ciampi et al. [20]; Kim and Sohn [36]; Pederzoli and Torricelli [49]; Psillaki, Tsolas, and Margaritis [50]; Lin, Ansell, and Andreeva [40]; Mannarino and Succurro [41]; Ciampi and Gordini [19]; Gabbianelli [28]; Altman, Esentato, and Sabato [4]
	Account payable/Cogs	Altman and Sabato [5]; Altman, Esentato, and Sabato [4]
	Account receivable/Sales	Altman and Sabato [5]; Altman, Esentato, and Sabato [4]
Non-financial information	Size	Ohlson [47]; Dietsch and Petey [26]; Altman, Sabato, and Wilson [6]; Psillaki, Tsolas, and Margaritis [50]; Michala, Grammatikos, and Ferreira Filipe [42]; Mannarino and Succurro [41]; Ciampi and Gordini [19]; Ciampi [18]
	Geographical area	Michala, Grammatikos, and Ferreira Filipe [42]; Ciampi [18]; Barreto Fernandes and Artes [12]; Gabbianelli [28]

	Technology/Innovation	Sohn and Kim [54]; Kim and Sohn [36]; Ciampi and Gordini [19]; Mannarino and Succurro [41]; Ju and Sohn [35]; Ju, Jeon, and Sohn [34]; Angilella and Mazzù [8]; Gabbianelli [28]
	Sector/Industry/Market	Lehmann [39]; Dietsch and Petey [26]; Sohn and Kim [54]; Altman, Sabato, and Wilson [6]; Kim and Sohn [36]; Psillaki, Tsolas, and Margaritis [50]; Ju and Sohn [35]; Michala, Grammatikos, and Ferreira Filipe [42]; Wolter and Rösch [58]; Ju, Jeon, and Sohn [34]; Ciampi [18]; Angilella and Mazzù [8]; Gabbianelli [28]; Nehrebecka [46]; Sigrist and Hirnschall [53]
	Audit accounts	Altman, Sabato, and Wilson [6]; Kim and Sohn [36]; Ciampi [18]; Ju, Jeon, and Sohn [34]
	Management knowledge	Lehmann [39]; Sohn and Kim [54]; Kim and Sohn [36]; Psillaki, Tsolas, and Margaritis [50]; Ju and Sohn [35]; Ju, Jeon, and Sohn [34];Ciampi [18]
	Age	Sohn and Kim [54]; Altman, Sabato, and Wilson [6]; Kim and Sohn [36]; Mannarino and Succurro [41]; Michala, Grammatikos, and Ferreira Filipe [42]; Andreeva, Calabrese, and Osmetti [7]; Sigrist and Hirnschall [53]
	Relational financing	Lehmann [39]; Moro and Fink [45]; Nehrebecka [46]
Economic indicators	Oil price	Sohn and Kim [54]; Kim and Sohn [36]
	CPI	Kim and Sohn [36]; Ju, Jeon, and Sohn [34]
	GDP	Michala, Grammatikos, and Ferreira Filipe [42]; Ju, Jeon, and Sohn [34]; Wolter and Rösch [58]
	Unemployment rates	Michala, Grammatikos, and Ferreira Filipe [42]; Ju, Jeon, and Sohn [34]

Table 21: *Structure of the dataset*

1	VAT number	41	Debt/EBITDA ratio
2	Company name	42	Cash flow/Debt
3	Website	43	Total assets turnover (times)
4	Last accounting closing date	44	Accounts payable
5	Legal status	45	Accounts receivable
6	ATECO 2007 code	46	Number of employees
7	Commune	47	Revenues from sales and services th EUR
8	Region	48	Total assets th EUR
9	Short-term debt/equity	49	P&L th EUR
10	Debt/equity ratio	50	Innovative PMI
11	Leverage	51	Number of advisors
12	Sharefunds/Liabilities	52	Number of directors managers
13	Short-term debt/Total Debt	53	No of available years
14	Short-term debt/Total Assets	54	Total Current Assets th EUR
15	Total Debts/Total Assets	55	Net financial position th EUR
16	Cash/Total Assets	56	Total shareholders funds th EUR
17	Intangible Assets/Total Assets	57	Net working capital th EUR
18	Tangible Assets/Total Assets	58	EBITDA th EUR
19	Net working capital/Total Assets	59	Cash Flow th EUR
20	Quick Assets/Total Assets		
21	Current ratio		
22	Liquidity ratio		
23	Solvency ratio		
24	ROE		
25	ROA		
26	ROTA		
27	EBITDA/Total Assets		
28	ROI		
29	ROCE		
30	ROS		
31	Cash Flow/EBITDA		
32	Turnover per employee th EUR		
33	Added value per employee th EUR		
34	R&D/Sales		
35	Interest Operating profit		
36	EBIT/Interest expenses		
37	Interest/turnover		
38	Banks/turnover		
39	Net Financial Position/Turnover		
40	Cost of debit		

Table 22: *Descriptive statistics of the Main Dataset*

	av 18 sur	sd 18 sur	av 18 def	sd 18 def	av 19 sur	sd 19 sur	av 19 def	sd 19 def	av 20 sur	sd 20 sur	av 20 def	sd 20 def
Short-term debt/equity	5.88	38.69	2.80	46.50	5.71	36.99	2.28	42.25	4.47	36.35	1.22	38.90
Debt/equity ratio	1.92	21.41	0.57	22.76	1.84	19.77	1.01	20.37	1.72	21.22	0.20	22.17
Leverage	9.76	48.92	4.25	53.44	9.49	47.34	4.12	48.40	8.00	47.32	2.70	45.87
Sharefunds/Liabilities	2.31	10.69	1.15	8.23	2.39	10.84	1.32	9.23	2.56	10.98	2.26	12.28
Short-term debt/Total Debt	0.79	0.31	0.84	0.29	0.79	0.30	0.84	0.30	0.74	0.31	0.80	0.33
Short-term debt/Total Assets	0.57	4.75	5.36	36.06	0.57	4.88	6.00	40.96	0.53	4.79	6.75	43.69
Total Debts/Total Assets	0.73	5.22	6.29	39.02	0.73	5.36	6.99	43.57	0.73	5.36	8.23	47.16
Cash/Total Assets	0.16	0.21	0.20	0.28	0.16	0.21	0.19	0.28	0.20	0.22	0.20	0.28
Intangible Assets/Total Assets	0.05	0.13	0.06	0.15	0.05	0.13	0.05	0.15	0.05	0.13	0.04	0.13
Tangible Assets/Total Assets	0.22	0.28	0.11	0.21	0.21	0.28	0.11	0.22	0.22	0.28	0.12	0.25
Net working capital/Total Assets	0.10	4.75	-4.58	36.05	0.10	5.00	-5.23	40.95	0.14	4.79	-5.99	43.89
Quick Assets/Total Assets	0.52	0.32	0.67	0.34	0.82	0.28	0.87	0.27	0.83	0.27	0.86	0.30
Current ratio	1.80	1.62	1.30	1.57	1.83	1.64	1.31	1.63	2.02	1.74	1.37	1.78
Liquidity ratio	1.51	1.57	1.13	1.52	1.54	1.60	1.14	1.57	1.72	1.70	1.18	1.70
Solvency ratio	32.24	29.38	26.89	39.66	32.96	29.51	27.28	40.21	33.85	30.54	30.03	41.89
ROE	11.94	34.32	-2.51	44.87	11.82	34.30	-5.65	43.86	9.79	35.33	-6.92	38.60
ROA	0.62	30.07	-40.05	120.31	0.49	30.85	-39.51	118.56	-0.76	31.57	-30.32	106.51
ROTA	3.29	30.83	-37.15	119.77	3.01	31.57	-36.29	116.10	1.15	32.11	-27.98	104.15
EBITDA/Total Assets	6.23	30.71	-32.89	117.19	5.97	31.44	-32.05	113.71	3.82	31.75	-24.78	101.77
ROI	6.47	10.27	0.90	11.89	6.44	10.26	0.28	11.63	5.00	10.76	-0.30	10.78
ROCE	14.68	72.07	14.12	150.01	13.83	72.19	10.33	151.72	9.45	76.58	4.16	130.92
ROS	4.32	11.28	-5.09	16.97	4.19	11.31	-5.78	17.45	2.88	13.21	-6.04	18.66
Cash Flow/EBITDA	0.54	14.50	1.23	18.46	0.58	13.91	1.34	15.44	0.63	14.64	1.19	17.39
Turnover per employee	204.74	294.32	132.13	248.25	202.46	291.81	116.01	217.73	191.17	283.04	116.62	240.29
Added value per employee	45.04	46.63	19.85	39.54	45.18	46.53	18.52	39.74	41.45	48.07	16.29	40.83
R&D/Sales	0.01	1.86	0.03	1.82	0.02	2.98	0.11	7.99	0.02	1.88	0.05	3.79
Interest/EBITDA	45.82	75.50	46.77	79.01	46.59	75.78	49.21	80.11	51.11	78.64	54.91	85.58
EBIT/Interest expenses	37.17	159.47	-35.32	195.88	36.26	160.78	-45.07	199.34	31.41	165.83	-46.92	206.01
Interest/turnover	2.21	7.33	3.24	10.70	2.08	7.11	3.92	12.38	2.00	6.87	4.50	13.51
Banks/turnover	11.13	18.42	8.73	18.00	10.92	18.24	7.91	17.88	14.51	20.89	7.35	18.00
Net Financial Position/Turnover	0.32	13.65	0.54	18.70	0.23	14.16	1.09	26.45	0.20	14.11	1.04	29.19
Cost of debit	4.98	4.20	4.81	4.89	4.92	4.16	4.26	4.85	3.85	3.73	3.28	4.36
Debt/EBITDA ratio	1.76	35.00	-3.14	58.67	1.66	34.96	-5.55	64.06	1.69	39.20	-8.37	74.51
Cash flow/Debt	1.58	16.31	-2.83	42.26	1.61	17.94	-3.94	42.11	0.84	14.91	-3.32	35.39
Total Assets Turnover (times)	0.98	0.95	0.69	1.00	0.99	0.96	0.59	0.95	0.84	0.84	0.35	0.76
Accounts payable	112.41	97.52	116.45	123.82	109.99	96.50	118.47	127.55	113.10	96.68	115.48	129.55
Accounts receivable	108.68	129.88	113.60	170.86	104.91	127.59	112.24	176.14	107.61	128.80	113.84	186.98
Number of employees	6.73	16.91	3.61	13.16	6.66	16.60	2.84	10.77	6.77	16.94	1.58	7.85
Sales	1398.80	5230.38	423.73	2734.15	1288.04	4878.01	303.14	2978.24	1198.97	4533.82	168.20	1326.11
Total Assets	3050.91	197447.86	826.29	7962.68	2985.62	188582.75	735.66	9219.08	3294.01	147124.61	771.55	4731.11
P&L	65.52	2083.84	-117.31	1655.61	58.69	2692.36	-92.38	1149.87	51.15	2632.76	-64.17	1157.88
Innovative PMI	0.00	0.05	0.00	0.02	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01
Number of advisors	0.17	0.71	0.06	0.38	0.16	0.69	0.07	0.43	0.17	0.70	0.10	0.52
Number of directors managers	2.87	3.79	2.08	2.26	2.81	3.70	2.14	2.33	2.82	3.74	2.21	2.75
No of available years	7.83	2.78	6.81	3.35	7.39	3.10	7.20	3.25	7.14	3.34	8.21	2.78
Current Assets	1250.80	12279.73	545.04	5023.02	1237.44	15501.40	495.26	3438.64	1345.61	16119.28	550.94	3552.49
Net financial position	20.70	265.26	33.89	181.85	14.39	267.45	27.99	182.41	4.64	282.09	10.58	200.68
Total shareholders funds	1098.48	22229.90	-190.09	6710.54	1123.40	23785.12	-256.93	8463.36	1346.86	23400.05	-451.33	5742.59
Net working capital	379.68	8095.98	-167.16	4357.51	391.38	13872.71	-201.20	4250.95	488.76	9805.16	-256.44	4365.31
EBITDA	122.86	815.14	-64.03	1100.16	119.22	797.07	-59.88	676.37	111.76	947.56	-37.75	781.45
Cash Flow	50.32	138.55	-26.84	129.38	50.73	139.38	-28.46	125.01	45.49	147.32	-24.32	116.55

Table 23: *Descriptive statistics for the manufacturing sector in 2018*

	NA_0	Mean_0	Sd_0	NA_1	Mean_1	Sd_1
Short-term debt/equity	0.11	5.13	27.64	0.01	2.44	41.19
Debt/equity ratio	37.80	1.69	13.27	1.60	1.18	19.46
Leverage	0.13	7.93	32.91	0.01	3.39	47.31
Sharefunds/Liabilities	6.53	1.27	5.32	1.07	0.68	6.16
Short-term debt/Total Debt	0.43	0.83	0.22	0.06	0.86	0.26
Short-term debt/Total Assets	0.01	0.57	4.60	0.01	5.74	34.00
Total Debts/Total Assets	0.01	0.69	4.73	0.02	6.71	37.37
Cash/Total Assets	0.00	0.13	0.17	0.00	0.16	0.27
Intangible Assets/Total Assets	0.00	0.04	0.09	0.00	0.05	0.13
Tangible Assets/Total Assets	0.00	0.21	0.21	0.00	0.14	0.24
Net working capital/Total Assets	0.01	0.14	4.60	0.01	-4.98	33.99
Quick Assets/Total Assets	0.00	0.56	0.25	0.00	0.66	0.32
Current ratio	2.84	1.79	1.34	0.24	1.14	1.39
Liquidity ratio	2.48	1.46	1.27	0.23	0.98	1.34
Solvency ratio	1.02	30.66	25.02	0.93	21.80	38.55
ROE	5.56	14.09	29.73	1.49	-4.29	44.48
ROA	0.03	2.48	19.56	0.09	-39.10	118.03
ROTA	0.03	5.26	20.58	0.08	-36.72	118.47
EBITDA/Total Assets	0.03	8.54	20.58	0.08	-31.65	114.01
ROI	50.48	8.48	9.84	2.22	2.04	12.08
ROCE	0.22	15.80	54.73	0.09	14.76	147.91
ROS	9.27	5.06	9.02	1.48	-6.18	16.94
Cash Flow/EBITDA	0.48	0.73	8.19	0.08	1.20	14.95
Turnover per employee	13.41	210.10	239.39	1.41	128.48	211.68
Added value per employee	13.63	52.68	40.11	1.54	21.55	36.85
R&D/Sales	48.27	0.02	0.70	1.75	0.04	0.85
Interest/EBITDA	25.00	41.37	68.03	2.17	38.01	69.63
EBIT/Interest expenses	13.67	40.16	137.84	0.98	-32.93	177.71
Interest/turnover	2.88	1.21	3.83	0.59	3.18	9.68
Banks/turnover	40.31	16.22	19.63	1.92	14.97	22.62
Net Financial Position/Turnover	0.01	0.14	7.28	0.00	1.53	23.26
Cost of debit	57.30	4.67	3.93	2.23	4.90	4.65
Debt/EBITDA ratio	38.08	2.32	23.25	1.63	-1.88	55.11
Cash flow/Debt	53.15	1.47	10.90	2.13	-2.05	25.38
Total Assets Turnover (times)	0.34	1.09	0.70	0.08	0.70	0.93
Accounts payable	40.20	119.52	77.56	1.97	135.39	115.66
Accounts receivable	34.09	120.08	95.04	1.79	137.77	167.22
Number of employees	0.00	14.52	24.01	0.00	5.83	14.46
Sales	0.00	3227.00	7116.05	0.00	677.54	2457.88
Total Assets	0.00	3630.08	31592.06	0.00	1234.85	7444.21
P&L	0.00	126.96	1504.93	0.00	-163.22	1331.00
Innovative PMI	0.00	0.00	0.06	0.00	0.00	0.03
Number of advisors	0.00	0.31	0.94	0.00	0.11	0.53
Number of directors managers	0.00	3.42	4.32	0.00	2.23	2.60
No of available years	0.00	8.32	2.54	0.00	7.37	3.19
Current Assets	0.00	2329.32	8314.79	0.00	747.54	4612.29
Net financial position	50.91	51.45	331.20	1.78	78.06	225.65
Total shareholders funds	0.00	1366.11	12925.62	0.00	-325.68	2907.47
Net working capital	0.00	743.35	16519.40	0.00	-346.29	2379.80
EBITDA	0.00	302.81	1101.30	0.00	-84.31	889.61
Cash Flow	6.14	107.76	195.30	0.13	-45.05	173.85

Table 24: *Descriptive statistics for the construction sector in 2018*

	NA_0	Mean_0	Sd_0	NA_1	Mean_1	Sd_1
Short-term debt/equity	0.29	8.38	52.24	0.01	5.18	53.68
Debt/equity ratio	50.53	3.12	33.74	2.05	0.92	24.14
Leverage	0.54	13.68	67.91	0.02	6.33	63.36
Sharefunds/Liabilities	24.40	1.61	8.22	1.92	0.98	6.14
Short-term debt/Total Debt	0.74	0.77	0.32	0.06	0.81	0.32
Short-term debt/Total Assets	0.00	0.61	6.49	0.02	4.82	34.63
Total Debts/Total Assets	0.01	0.80	6.27	0.02	6.07	38.10
Cash/Total Assets	0.00	0.13	0.19	0.00	0.17	0.27
Intangible Assets/Total Assets	0.00	0.02	0.08	0.00	0.02	0.09
Tangible Assets/Total Assets	0.00	0.15	0.24	0.00	0.07	0.18
Net working capital/Total Assets	0.00	0.18	6.49	0.02	-3.97	34.62
Quick Assets/Total Assets	0.00	0.53	0.35	0.00	0.65	0.37
Current ratio	10.57	1.92	1.61	0.43	1.50	1.69
Liquidity ratio	5.91	1.40	1.46	0.34	1.17	1.57
Solvency ratio	1.36	28.62	27.35	0.88	23.88	37.62
ROE	7.93	12.68	35.33	1.50	-0.45	42.74
ROA	0.07	2.00	26.94	0.09	-27.05	99.13
ROTA	0.06	4.73	28.16	0.09	-24.15	97.90
EBITDA/Total Assets	0.06	6.56	28.28	0.09	-21.08	92.38
ROI	61.90	6.20	9.96	2.60	0.26	10.85
ROCE	0.37	17.59	67.37	0.09	13.76	129.39
ROS	27.43	5.68	11.06	1.99	-2.52	16.51
Cash Flow/EBITDA	1.25	0.74	10.78	0.12	1.39	20.72
Turnover per employee	40.16	158.61	221.92	2.42	135.30	251.71
Added value per employee	40.54	42.43	38.23	2.49	24.30	36.32
R&D/Sales	47.12	0.00	0.09	2.08	0.00	0.01
Interest/EBITDA	39.41	42.76	72.74	2.68	41.10	71.33
EBIT/Interest expenses	22.38	38.72	161.36	1.47	-28.09	190.52
Interest/turnover	12.42	3.05	9.83	1.02	3.96	12.06
Banks/turnover	59.71	11.15	19.08	2.57	8.57	18.41
Net Financial Position/Turnover	0.11	0.82	18.26	0.01	0.98	19.89
Cost of debit	72.81	5.20	4.29	2.90	4.10	4.58
Debt/EBITDA ratio	51.26	2.17	50.18	2.11	-4.62	84.23
Cash flow/Debt	68.44	1.38	12.89	2.81	0.24	34.57
Total Assets Turnover (times)	0.72	0.84	0.90	0.12	0.54	0.92
Accounts payable	57.91	132.24	107.56	2.60	129.17	134.83
Accounts receivable	52.63	136.24	152.23	2.45	129.38	185.34
Number of employees	0.00	4.31	10.38	0.00	2.38	10.33
Sales	0.00	710.61	2297.43	0.00	327.80	1989.70
Total Assets	0.00	1915.16	24652.60	0.00	1283.23	11230.40
P&L	0.00	12.54	666.64	0.00	-194.09	3316.36
Innovative PMI	0.00	0.00	0.02	0.00	0.00	0.00
Number of advisors	0.00	0.09	0.47	0.00	0.05	0.31
Number of directors managers	0.00	2.39	2.51	0.00	2.06	2.32
No of available years	0.00	7.98	2.75	0.00	7.83	3.03
Current Assets	0.00	1363.68	22618.35	0.00	1088.77	10620.35
Net financial position	55.32	40.27	251.50	2.21	41.34	206.34
Total shareholders funds	0.00	489.62	5159.61	0.00	-396.01	4239.69
Net working capital	0.00	498.11	3874.51	0.00	3.46	3722.93
EBITDA	0.00	62.67	604.70	0.00	-119.02	2049.39
Cash Flow	0.87	30.17	106.83	0.09	-25.69	118.29

Table 25: *Descriptive statistics for the wholesale and retail trade sector in 2018*

	NA_0	Mean_0	Sd_0	NA_1	Mean_1	Sd_1
Short-term debt/equity	0.11	6.23	34.60	0.01	2.02	43.11
Debt/equity ratio	44.85	1.38	11.79	2.17	0.72	19.94
Leverage	0.14	8.85	39.40	0.01	2.97	47.45
Sharefunds/Liabilities	13.44	1.19	5.62	1.54	0.94	7.82
Short-term debt/Total Debt	0.79	0.85	0.23	0.09	0.86	0.26
Short-term debt/Total Assets	0.00	0.64	3.47	0.02	6.52	40.79
Total Debts/Total Assets	0.00	0.76	3.74	0.02	7.42	44.04
Cash/Total Assets	0.00	0.17	0.20	0.00	0.19	0.27
Intangible Assets/Total Assets	0.00	0.04	0.10	0.00	0.05	0.14
Tangible Assets/Total Assets	0.00	0.13	0.19	0.00	0.08	0.17
Net working capital/Total Assets	0.00	0.16	3.47	0.02	-5.71	40.79
Quick Assets/Total Assets	0.00	0.54	0.29	0.00	0.65	0.33
Current ratio	4.10	1.76	1.38	0.33	1.23	1.44
Liquidity ratio	3.42	1.25	1.31	0.31	0.98	1.36
Solvency ratio	1.66	28.58	26.52	1.11	24.17	38.73
ROE	8.13	14.65	33.35	1.80	-2.39	46.15
ROA	0.05	0.84	28.75	0.12	-44.47	124.13
ROTA	0.05	3.68	29.83	0.12	-41.73	125.92
EBITDA/Total Assets	0.05	6.01	29.66	0.11	-38.06	123.63
ROI	58.13	8.16	10.12	2.90	1.14	12.23
ROCE	0.44	18.36	74.00	0.12	20.36	156.67
ROS	10.92	3.47	9.00	1.76	-6.18	16.57
Cash Flow/EBITDA	0.82	0.70	7.72	0.11	1.14	8.71
Turnover per employee	27.09	338.20	397.65	2.03	220.63	339.89
Added value per employee	26.64	46.49	47.30	2.21	19.69	49.19
R&D/Sales	48.41	0.00	0.17	2.15	0.00	0.02
Interest/EBITDA	32.39	40.89	71.16	2.79	43.65	78.03
EBIT/Interest expenses	19.06	39.26	150.08	1.38	-35.60	189.61
Interest/turnover	3.92	1.02	3.65	0.70	2.65	8.67
Banks/turnover	47.21	10.44	16.33	2.49	9.77	18.31
Net Financial Position/Turnover	0.01	0.02	5.67	0.00	0.32	13.39
Cost of debit	68.52	5.44	4.35	3.02	5.52	5.08
Debt/EBITDA ratio	45.32	2.24	24.24	2.22	-1.84	46.63
Cash flow/Debt	63.58	1.46	14.33	2.89	-2.23	31.66
Total Assets Turnover (times)	3.22	1.37	1.01	0.24	0.86	1.09
Accounts payable	47.21	106.67	87.54	2.53	115.87	119.50
Accounts receivable	42.99	86.90	103.13	2.36	98.90	158.39
Number of employees	0.00	5.45	12.11	0.00	2.34	8.24
Sales	0.00	2222.23	7209.72	0.00	681.76	4685.49
Total Assets	0.00	1513.86	4846.05	0.00	675.86	10582.19
P&L	0.00	43.10	492.67	0.00	-86.90	1009.11
Innovative PMI	0.00	0.00	0.03	0.00	0.00	0.00
Number of advisors	0.00	0.13	0.59	0.00	0.04	0.31
Number of directors managers	0.00	2.45	2.93	0.00	1.82	1.69
No of available years	0.00	7.60	2.83	0.00	6.50	3.36
Current Assets	0.00	1143.64	3144.03	0.00	408.58	1468.51
Net financial position	50.32	16.02	264.19	2.26	36.04	177.88
Total shareholders funds	0.00	458.58	4158.44	0.00	-128.42	12939.75
Net working capital	0.00	341.83	1559.90	0.00	-246.34	7758.72
EBITDA	0.00	105.76	488.13	0.00	-55.58	950.39
Cash Flow	1.51	49.46	128.07	0.08	-30.02	133.61

Figure 3: *List of survived and defaulted SMEs per region (period 2018-2020)*

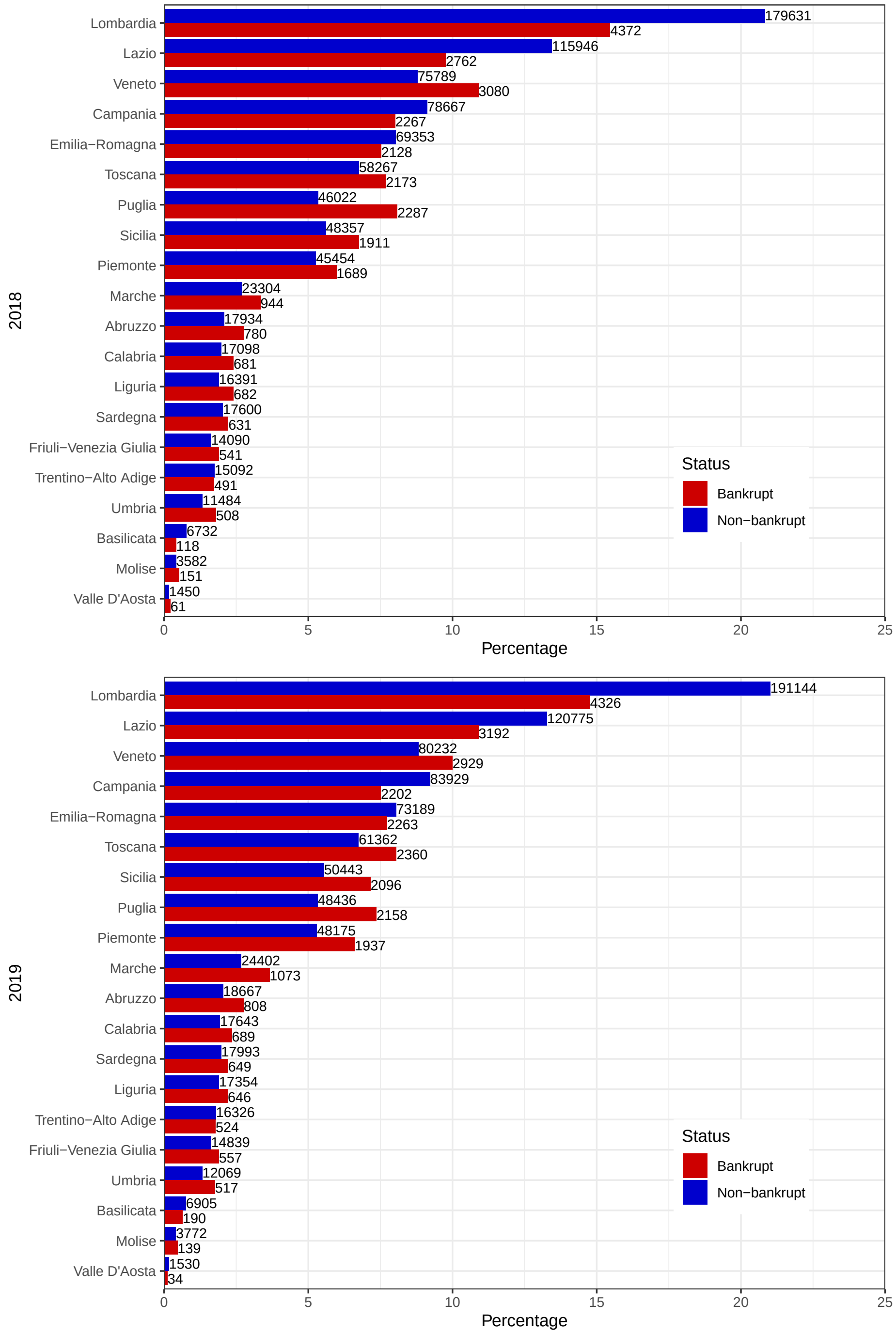


Figure 3: *Continued*

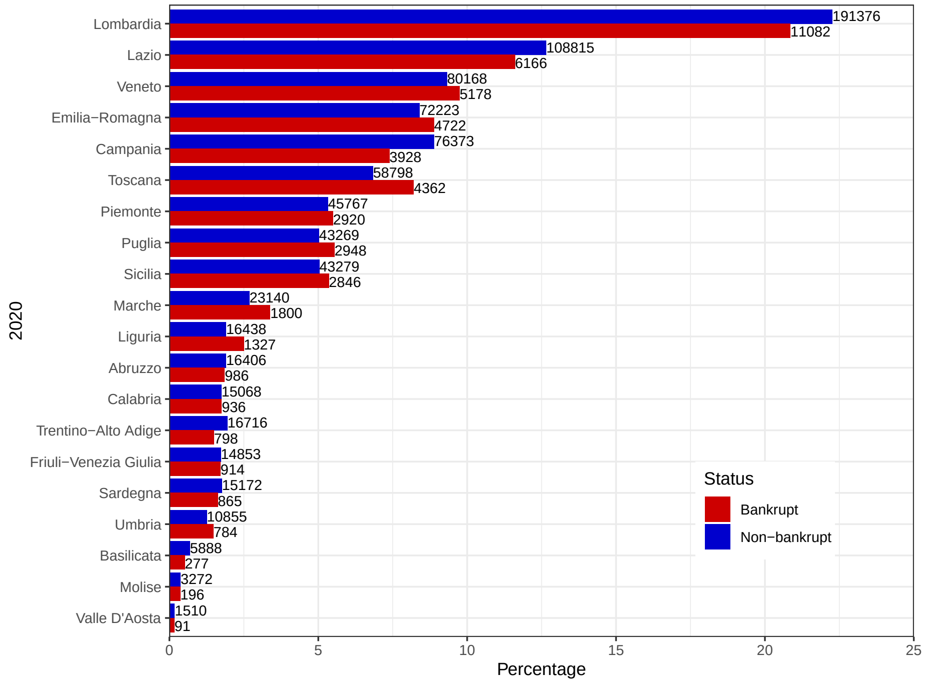


Figure 4: *List of survived and defaulted SMEs per sector (period 2018-2020)*

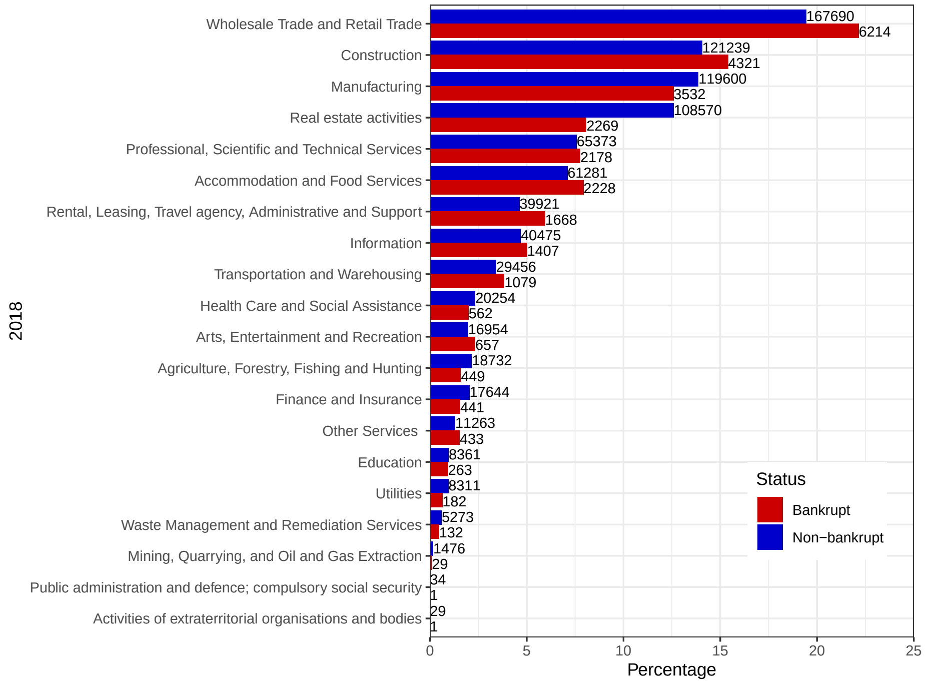


Figure 4: *Continued*

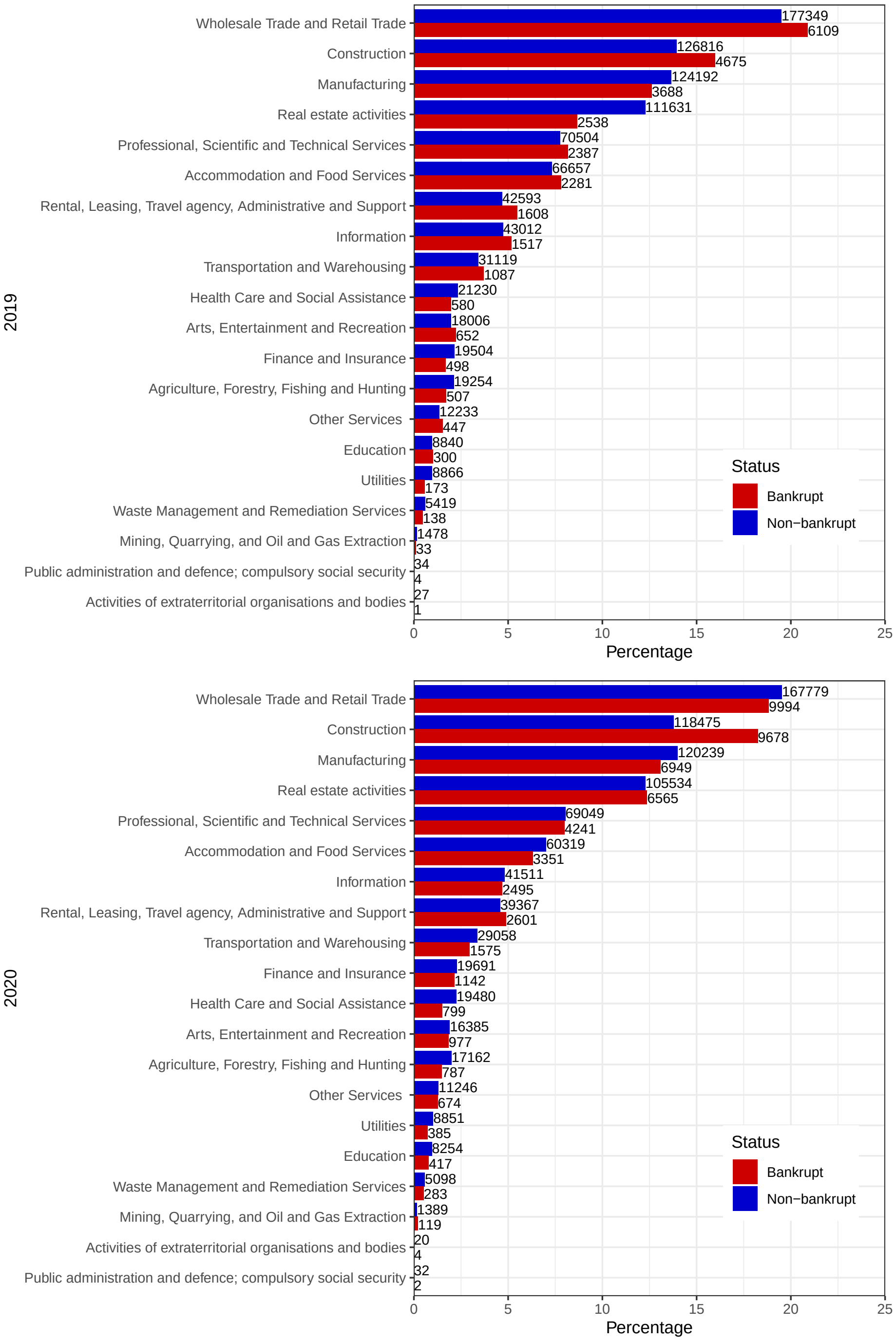


Figure 5: *Example of logistic regression's plots for a balanced training set, Manufacturing*

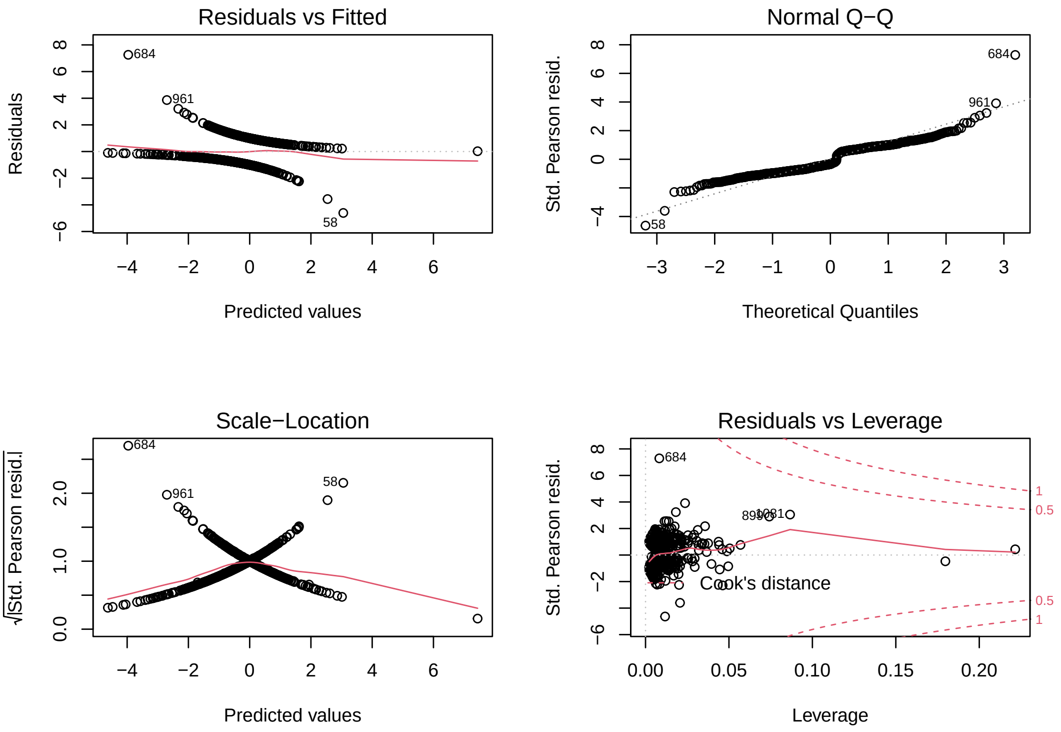


Figure 6: *AUROC plot, simulation 3, Manufacturing*

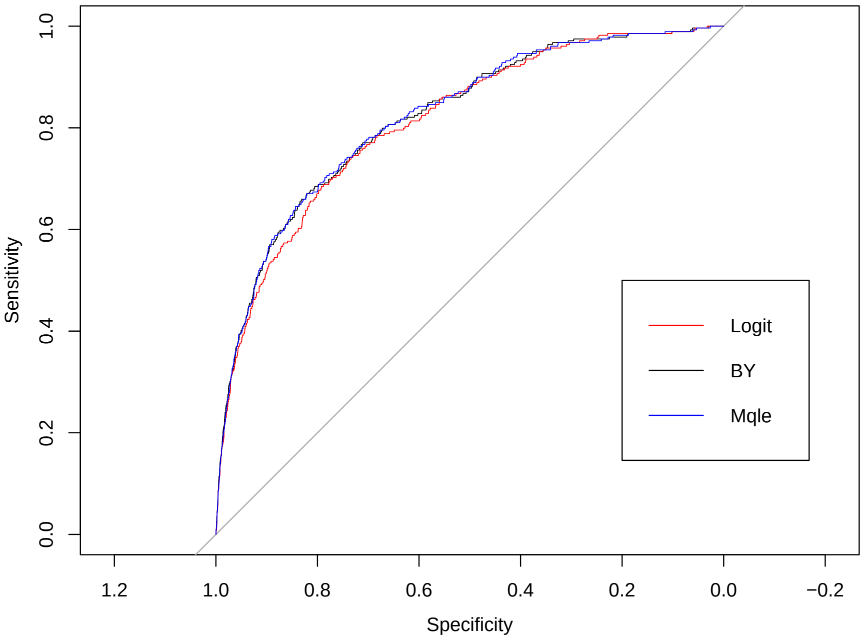


Figure 7: *Graphical example of outliers' detection, Manufacturing*

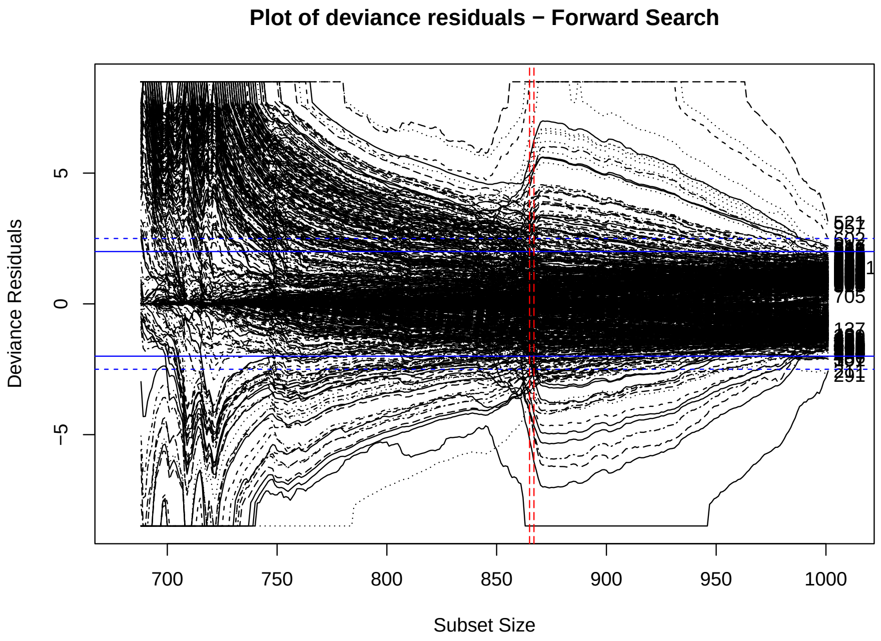


Figure 8: *Classification metrics during the forward search for all the iterations, training set, Manufacturing*

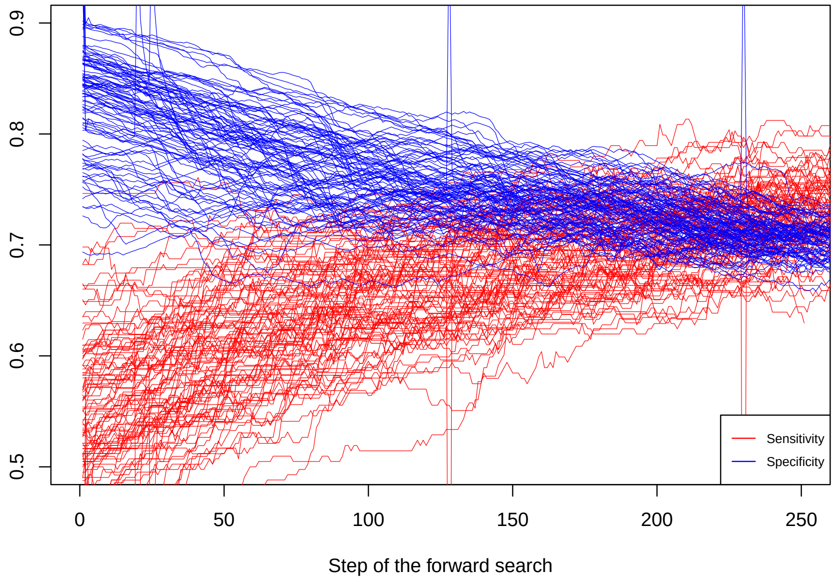


Figure 9: *Kernel probability density function training set, forward search, Construction*

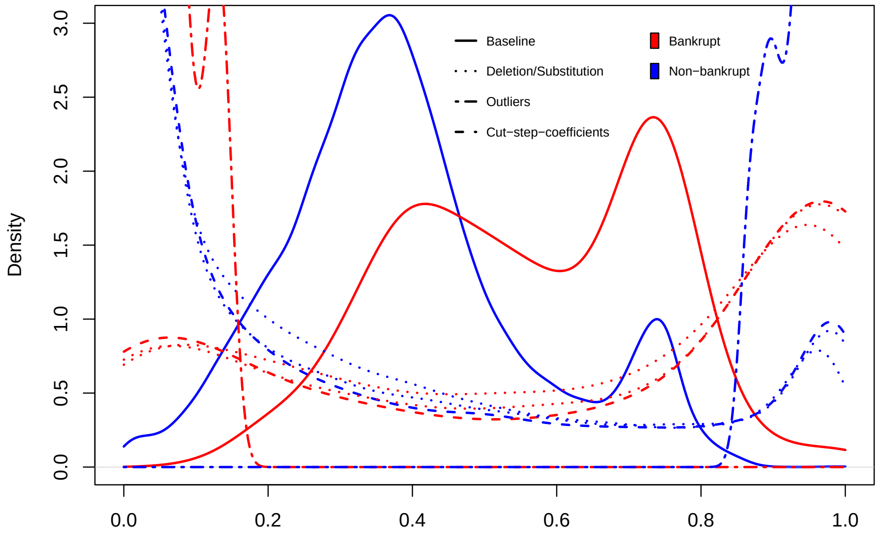


Figure 10: *Classification metrics during the forward search for all the iterations, training set, Construction*

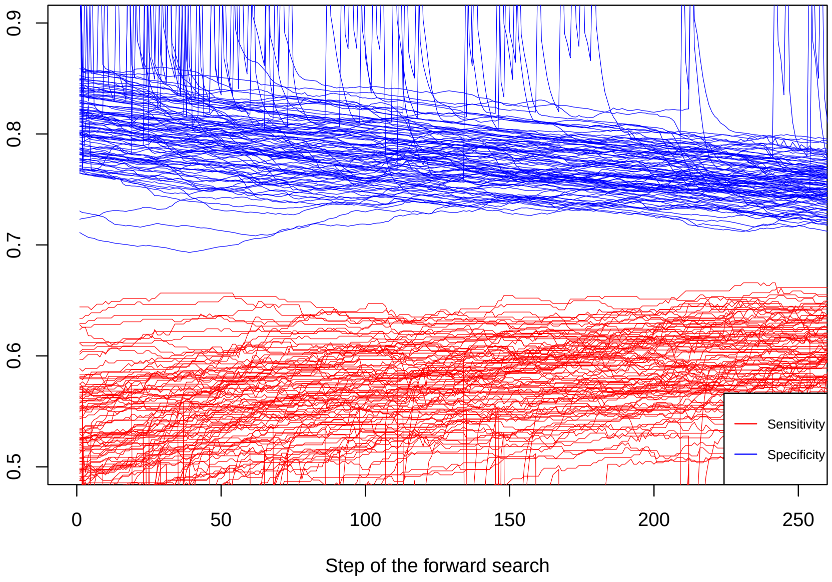


Figure 11: *Kernel probability density function training set, forward search, Trade*

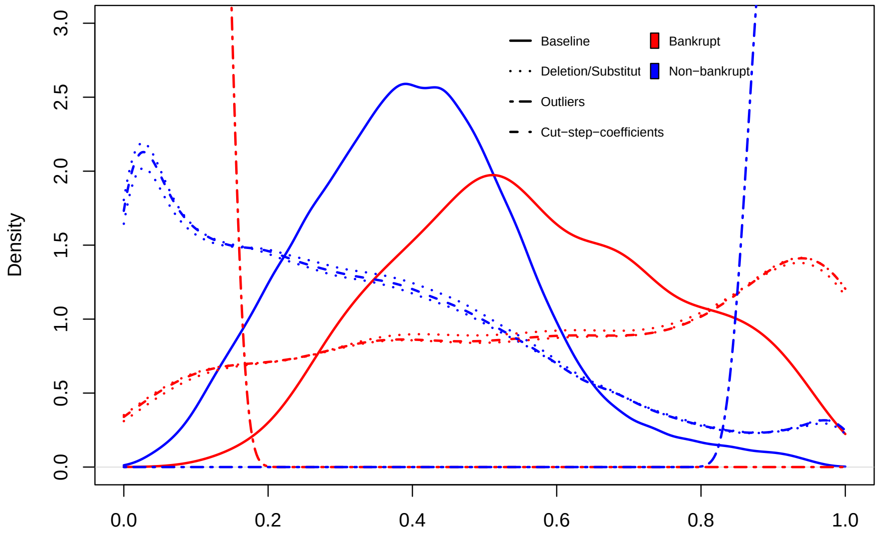
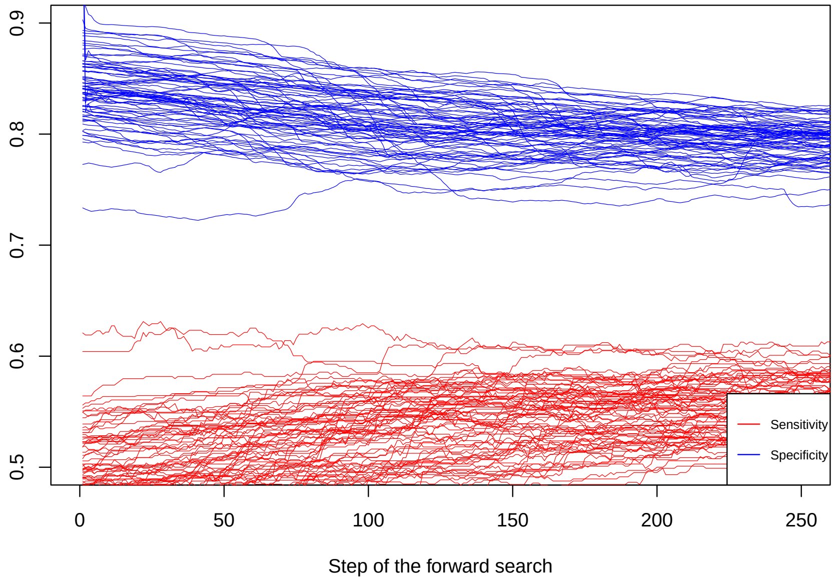


Figure 12: *Classification metrics during the forward search for almost all the iterations, training set, Trade*



Bibliography

[1] S. Akkoç. "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data". In: *European Journal of Operational Research* 222.1 (2012), pp. 168–178. URL: https://www.sciencedirect.com/science/article/pii/S0377221712002858.

[2] E. Altman. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". In: *The Journal of Finance* 23.4 (1968), pp. 589–609. URL: http://www.jstor.org/stable/2978933.

[3] E. Altman, Rui Dai, and Wei Wang. "Global Zombies". In: (2021). URL: http://dx.doi.org/10.2139/ssrn.3970332.

[4] E. Altman, M. Esentato, and G. Sabato. "Assessing the credit worthiness of Italian SMEs and mini-bond issuers". In: *Global Finance Journal* 43 (2020). URL: https://www.sciencedirect.com/science/article/pii/S1044028317304891.

[5] E. Altman and G. Sabato. "Modelling Credit Risk for SMEs: Evidence from the U.S. Market". In: *Abacus* 43.3 (2007), pp. 332–357. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6281.2007.00234.x.

[6] E. Altman, G. Sabato, and N. Wilson. "The Value of Non-Financial Information in SME Risk Management". In: *SSRN* (2008). URL: http://dx.doi.org/10.2139/ssrn.1320612.

[7] G. Andreeva, R. Calabrese, and S.A. Osmetti. "A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models". In: *European Journal of Operational Research* 249.2 (2016), pp. 506–516. URL: https://www.sciencedirect.com/science/article/pii/S0377221715007183.

[8] S. Angilella and S. Mazzù. "The financing of innovative SMEs: A multicriteria credit rating model". In: *European Journal of Operational Research* 244.2 (2015), pp. 540–554. DOI: https://doi.org/10.1016/j.ejor.2015.01.033.

[9] A. Atkinson and M. Riani. *Robust Diagnostic Regression Analysis*. New York: Springer Verlag, 2000.

[10] A. Atkinson and M. Riani. "Regression Diagnostics for Binomial Data from the forward search". In: *Journal of the Royal Statistical Society Series D (The Statistician)* 50 (Mar. 2001). DOI: 10.1111/1467-9884.00261.

[11] Basel Committee on Banking Supervision. "Studies on the Validation of Internal Rating Systems". In: *BIS Working Paper No. 14* (May 2005). URL: https://www.bis.org/publ/bcbs_wp14.htm.

[12] G. Barreto Fernandes and R. Artes. "Spatial dependence in credit risk and its improvement in credit scoring". In: *European Journal of Operational Research* 249.2 (2016), pp. 517–524. URL: https://EconPapers.repec.org/RePEc:eee:ejores:v:249:y:2016:i:2:p:517-524.

[13] A.N. Berger and G.F. Udell. "A more complete conceptual framework for SME finance". In: *Journal of Banking and Finance* 30.11 (2006), pp. 2945–2966. URL: https://www.sciencedirect.com/science/article/pii/S0378426606000938.

[14] R. Calabrese and S.A. Osmetti. "Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model". In: *Journal of Applied Statistics* 40.6 (2013), pp. 1172–1188. URL: https://doi.org/10.1080/02664763.2013.784894.

[15] E. Cantoni and E. Ronchetti. "Robust Inference for Generalized Linear Models". In: *Journal of the American Statistical Association* 96.455 (2001). URL: http://www.jstor.org/stable/2670248.

[16] E. Cantoni and E. Ronchetti. "A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures". In: *Journal of Health Economics* 25.2 (2006), pp. 198–213. URL: https://EconPapers.repec.org/RePEc:eee:jhecon:v:25:y:2006:i:2:p:198-213.

[17] T.C. Cheng. "Robust Diagnostic for the Logistic Regression Model". In: *Department of Statistics, National Chengchi University* (2002). URL: http://thesis.lib.nccu.edu.tw/record/#G0090354008.

[18] F. Ciampi. "Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms". In: *Journal of Business Research* 68.5 (2015), pp. 1012–1025. URL: https://www.sciencedirect.com/science/article/pii/S0148296314003221.

[19] F. Ciampi and N. Gordini. "Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises". In: *Journal of Small Business Management* 51.1 (2013), pp. 23–45. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-627X.2012.00376.x.

[20] F. Ciampi et al. "Are Credit Scoring Models Able to Predict Small Enterprise Default? Statistical Evidence from Italian Small Enterprises". In: *International Journal of Business and Economics* 8.1 (2009), pp. 3–18. URL: https://ssrn.com/abstract=2214957.

[21] F. Ciampi et al. "Rethinking SME default prediction: a systematic literature review and future perspectives". In: *Scientometrics* 126.3 (2021), pp. 2141–2188. URL: https://doi.org/10.1007/s11192-020-03856-0.

[22] European Commission. *Commission Recommendation of 6 May 2003 concerning the definition of micro,small and medium-sized enterprises (Text with EEA relevance) (notified under document number C(2003) 1422)*. 2003. URL: http://data.europa.eu/eli/reco/2003/361/oj.

[23] European Commission. *European Construction Sector Observatory Country profile Italy*. 2021. URL: https://single-market-economy.ec.europa.eu/sectors/construction/observatory/country-fact-sheets/italy_en.

[24] L. Crosato, C. Liberati, and M. Repetto. *Look Who's Talking: Interpretable Machine Learning for Assessing Italian SMEs Credit Default*. 2021. URL: https://arxiv.org/abs/2108.13914.

[25] C. Croux and G. Haesbroeck. "Implementing the Bianco and Yohai estimator for logistic regression". In: *Computational Statistics & Data Analysis* (2003). URL: https://EconPapers.repec.org/RePEc:eee:csdana:v:44:y:2003:i:1-2:p:273-295.

[26] M. Dietsch and J. Petey. "Should SME exposures be treated as retail or corporate exposures? A comparative analysis of default probabilities and asset correlations in French and German SMEs". In: *Journal of Banking and Finance* 28.4 (2004), pp. 773–788. URL: https://www.sciencedirect.com/science/article/pii/S0378426603001997.

[27] R. Edmister. "An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction". In: *The Journal of Financial and Quantitative Analysis* 7.2 (1972), pp. 1477–1493. URL: http://www.jstor.org/stable/2329929.

[28] L. Gabbianelli. “I modelli di previsione delle insolvenze e le piccole imprese: evidenze empiriche in una prospettiva territoriale”. In: *Sinergie Italian Journal of Management* (Jan. 2018), pp. 117–139. DOI: 10.7433/s101.2016.08.

[29] J. Gentry, P. Newbold, and D.T. Whitford. “Classifying Bankrupt Firms with Funds Flow Components”. In: *Journal of Accounting Research* 23.1 (1985), pp. 146–160. URL: http://www.jstor.org/stable/2490911.

[30] L. Grossi and T. Bellini. “Credit Risk Management through Robust Generalized Linear Models”. In: Data Analysis, Classification and the Forward Search (2006), pp. 377–386. DOI: 10.1007/3-540-35978-8_42.

[31] R. Hauser and D. Booth. “Predicting Bankruptcy with Robust Logistic Regression”. In: *Journal of data science: JDS* 9 (Jan. 2011), pp. 565–584. DOI: 10.6339/JDS.201110_09(4).0006.

[32] P.J. Huber and E.M. Ronchetti. *Robust Statistics*. Wiley, 1981.

[33] Istat. *22 COMMERCIO INTERNO E ALTRI SERVIZI*. 2020. URL: https://www.istat.it/it/files//2020/12/C22.pdf.

[34] Y.H. Ju, S.Y. Jeon, and S.Y. Sohn. “Behavioral technology credit scoring model with time-dependent covariates for stress test”. In: *European Journal of Operational Research* 242.3 (2015), pp. 910–919. URL: https://www.sciencedirect.com/science/article/pii/S0377221714008765.

[35] Y.H. Ju and S.Y. Sohn. “Updating a credit-scoring model based on new attributes without realization of actual data”. In: *European Journal of Operational Research* 234.1 (2014), pp. 119–126. URL: https://www.sciencedirect.com/science/article/pii/S037722171300163X.

[36] H.S. Kim and S.Y. Sohn. “Support vector machines for default prediction of SMEs based on technology credit”. In: *European Journal of Operational Research* 201.3 (2010), pp. 838–846. URL: https://EconPapers.repec.org/RePEc:eee:ejores:v:201:y:2010:i:3:p:838-846.

[37] J.W. Kolari, C.C. Ou, and G.H. Shin. “Assessing the Profitability and Riskiness of Small Business Lenders in the Banking Industry”. In: *Journal of Entrepreneurial Finance and Business Ventures* 7 (2006), pp. 1–26. URL: https://digitalcommons.pepperdine.edu/jef/vol11/iss2/2.

[38] P. Komarek and A.W. Moore. “Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs”. In: 2003, pp. 163–170. URL: https://proceedings.mlr.press/r4/komarek03a.html.

[39] B. Lehmann. “Is It Worth the While? The Relevance of Qualitative Information in Credit Rating”. In: *SSRN Electronic Journal* (Apr. 2003). DOI: 10.2139/ssrn.410186.

[40] S.M. Lin, J. Ansell, and G. Andreeva. “Predicting default of a small business using different definitions of financial distress”. In: *Journal of the Operational Research Society* 63.4 (2012), pp. 539–548. URL: https://doi.org/10.1057/jors.2011.65.

[41] L. Mannarino and M. Succurro. *The impact of financial structure on firms' probability of bankruptcy: a comparison across Western Europe convergence regions*. Working Papers 201305. Università della Calabria, Dipartimento di Economia, Statistica e Finanza "Giovanni Anania" - DESF, 2013. URL: https://EconPapers.repec.org/RePEc:clb:wpaper:201305.

[42] D. Michala, T. Grammatikos, and S. Ferreira Filipe. “Forecasting distress in European SME portfolios”. In: *EIF Working Paper* 17 (2013), pp. 117–139. URL: http://www.eif.org/news_centre/publications/EIF_Working_Paper_2013_17.html.

[43] M. Miyamoto. “Predicting Default for Japanese SMEs with Robust Logistic Regression”. In: *International Journal of Economics, Commerce and Research (IJECR)* 6 (June 2016). URL: https://ssrn.com/abstract=2838141.

[44] M. Mohammed. “Robust logistic regression in the presence of high leverage points”. In: *Journal of Al-Qadisiyah for computer science and mathematics* 11 (Sept. 2019), pp. 1–11. URL: https://qu.edu.iq/journalcm/index.php/journalcm/article/view/581.

[45] A. Moro and M. Fink. “Loan managers' trust and credit access for SMEs”. In: *Journal of Banking and Finance* 37.3 (2013), pp. 927–936. URL: https://www.sciencedirect.com/science/article/pii/S0378426612003342.

[46] N. Nehrebecka. “Predicting the Default Risk of Companies. Comparison of Credit Scoring Models: Logit Vs Support Vector Machines”. In: *Econometrics. Advances in Applied Data Analysis* 22.2 (2018), pp. 54–73. URL: https://EconPapers.repec.org/RePEc:vrs:eaiada:v:22:y:2018:i:2:p:54-73:n:5.

[47] J.A. Ohlson. “Financial Ratios and the Probabilistic Prediction of Bankruptcy”. In: *Journal of Accounting Research* 18.1 (1980), pp. 109–131. URL: http://www.jstor.org/stable/2490395.

[48] European Parliament. *Capital Requirements Regulation (CRR): REGULATION (EU) No 575/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012*. 2013. URL: https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/504.

[49] C. Pederzoli and C. Torricelli. "A parsimonious default prediction model for Italian SMEs". In: *Universita di Modena e Reggio Emilia, Facoltà di Economia "Marco Biagi", Centro Studi di Banca e Finanza (CEFIN)* 5 (Jan. 2010). URL: https://ideas.repec.org/p/mod/wcefin/0022.html.

[50] M. Psillaki, I.E. Tsolas, and D. Margaritis. "Evaluation of credit risk based on firm performance". In: *European Journal of Operational Research* 201.3 (2010), pp. 873–881. URL: https://EconPapers.repec.org/RePEc:eee:ejores:v:201:y:2010:i:3:p:873-881.

[51] F. Rikkers and A.E. Thibeault. "Default prediction of small and medium-sized enterprises with industry effects". In: *International Journal of Banking, Accounting and Finance* (2011). URL: https://ideas.repec.org/a/ids/injbaf/v3y2011i2-3p207-231.html.

[52] G. Rodríguez. *Lecture Notes on Generalized Linear Models*. 2007. URL: https://data.princeton.edu/wws509/notes/.

[53] F. Sigrist and C. Hirnschall. "Grabit: Gradient tree-boosted Tobit models for default prediction". In: *Journal of Banking and Finance* 102 (2019), pp. 177–192. URL: https://www.sciencedirect.com/science/article/pii/S0378426619300573.

[54] S.Y. Sohn and H.S. Kim. "Random effects logistic regression model for default prediction of technology credit guarantee fund". In: *European Journal of Operational Research* 183.1 (2007), pp. 472–478. URL: https://www.sciencedirect.com/science/article/pii/S0377221706010393.

[55] Tserng et al. "Prediction of default probability for construction firms using the logit model". In: *Journal of Civil Engineering and Management* 20 (2014). DOI: 10.3846/13923730.2013.801886.

[56] European Union. *User guide to the SME Definition*. 2015. URL: https://ec.europa.eu/regional_policy/sources/conferences/state-aid/sme/smedefinitionguide_en.pdf.

[57] D. Veganzones and E. Séverin. “An investigation of bankruptcy prediction in imbalanced datasets”. In: *Decision Support Systems* 112 (2018), pp. 111–124. URL: https://www.sciencedirect.com/science/article/pii/S0167923618301088.

[58] M. Wolter and D. Rösch. “Cure events in default prediction”. In: *European Journal of Operational Research* 238.3 (2014), pp. 846–857. URL: https://www.sciencedirect.com/science/article/pii/S0377221714003889.

Acknowledgements

At the end of the thesis I would like to profoundly thank my supervisor Professor Lisa Crosato who gave me the opportunity to get inside a challenging argument and to show myself outside the comfort zone, debating statistical methods which are out of the scope of several courses and keeping the motivation alive. This study let me dive into the world of credit risk from an econometric prospective and get an insight of what machine learning actually can do.

I would like to thank the Ca' Foscari University of Venice for organizing the course in Economics and Finance with plenty of skilful professors, especially my fixed-term employer Professor Simone Righi, with whom I could work for tutorials dedicated to bachelor's students via both on-line and in-person meetings. This activity and the Erasmus Programme during my bachelor's degree gave me a solid and multifaceted background very useful in all the job interviews I had.

Finally I would like to thank my family for their financial support, my colleagues, with which I shared these five years of university in San Giobbe, and my friends who always supported me in good and hard times with words and actions. Thank you for your love and support.