POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Default prediction of SMEs in the Italian Minibond market: an analysis under a new human perspective

TESI MAGISTRALE IN MANAGEMENT ENGINEERING – INGEGNERIA GESTIONALE

AUTHOR: MATTEO SPEDICATI, TOMMASO TREGROSSO

ADVISOR: GIANCARLO GIUDICI

ACADEMIC YEAR: 2022-2023

---

1. Literature Analysis

As a starting point of the work, an overview of main themes in discussion is given to introduce the specific analysis that is the core of the dissertation. With this, the related literature review is presented for each topic.

Firstly, it's described the important role played by SMEs at European and Italian level, in terms of percentage of SMEs out of all firms, value added and employment, highlighting their importance in the whole economic system. Then, a critical review of the literature related to SMEs' access to capital markets and SMEs' financing structure is done, to better understand which are the main determinants. In the second part, an analysis of the current financing situation in Europe and in Italy is made up, emerging that Italian SMEs are more dependent on bank loans. Finally, following the research results summarized in *Rapporto Cerved 2022*, the focus shifts to Italian SMEs' performances, to understand which are the conditions after Covid crisis and current energy shock.

Then, it's analysed in deep the mini-bond world, addressing the reasons that led to its birth, the characteristics of the instrument (advantages and disadvantages) and the regulatory framework (which facilitates its spread into SMEs market and providing data and statistics on Italian and foreign mini-bond market). In addition, some numbers for the Italian mini-bond market are given, resulting that the 2022 was a year of further diffusion of awareness among companies with the record of 190 new issuers. Looking also at the issuances' perspective, it is confirmed that minibonds hold their own "stable" market since the reforms, initiated by the 2012 Development Decree, came into force. Starting from 2013, we have a total sample of 1461 placements (268 in 2022).

Lastly, it is presented a deep overview of literature situation of the topic "*business failure prediction*", from traditional studies to new approaches for SMEs. The recent global financial crisis has resulted in numerous company failures in many countries, renewing the literature's interest in default risk forecasting models. Even though these models have been developed since the 1960s, a growing number of studies have been published in recent years, either proposing new approaches or contrasting various models that already exist to determine which has the highest predictive power. Nonetheless, the significance of implementing early warning systems has been highlighted by the current financial crisis. While warning system implementation and business failure forecasting are conceptually different, there is a risk of overlapping these concepts. Lots of model have been reviewed and most of them use financial ratios to predict bankruptcy, rather than corporate governance indicators. On the other hand, it is underlined how SMEs have specific characteristics that led to construct innovative model for their default prediction (financial ratios are no longer accurate enough). So, many models have been reviewed here too and lots of non-traditional quantitative methodologies and non-financial predictive variables that improve prediction accuracy have been briefly described.

2. Objectives and Data Collection

The objective of this dissertation is finding out if there is a relation between the default probability in the companies that issued minibonds focusing, on one hand, on the coexistence in the model of the classical predictors like the Assets or the Annual Interest Rate and ,on the other, new type of variable (the registration of intellectual properties and the composition of the Board of Directors).

For the analysis it has been taken in consideration the emissions from the 8th of April 2013 to the 31st of December 2020, the second type of filter is based on the elimination from the sample of the companies belonging to the financial sector (ATECO code K) and the emissions that exceed the amount of fifty millions of Euros.

The research questions that guided the research are:

RQ1: *Is it possible to observe any pattern between the different variables and the default of a firm?*
RQ2: *Are there any new types of variables that can perform a default probability prediction that are different from the more known indicators?*

The process of data collection consists in five distinct phases that brought to the classification of all the companies in the sample as defaulted or not. The phases are the following:

1. Query in the AIDA platform of all the companies and recognize those without any type of problem as "non default".
2. Research on Telemaco database for those firms that are not present in AIDA, the objectives are the same of the previous phase.
3. Check for all the companies that show problems and classification of all the observations as Default or No Default.
4. Discovery new variables for each company regarding three types of IPs: trademarks, designs, and patents. The research has been conducted through the database of the EUIPO (for trademarks and designs that are collected as a single variable) and EPO (for patents applications).
5. Research on the new variables in the field of the composition of the Board of Directors from the databases of AIDA and Telemaco, the new variables found regard the age of the members their gender distribution and the total number of members.

3. Data Analysis and Results

The core of this research is founding a predictive model that can represent the default probability of the firms using different sets of variables. Before the building of the model, univariate statistics on all the variables that will be necessary later, considering that there are both numeric, dichotomic and multilevel variables, have been conducted.
After the completion of this preliminary phase in which it is possible to observe differences between the distribution of the different variable, the next step is the multivariate analysis in order to create a logistic regression classifier. A stepwise procedure has been conducted that, starting from three models composed by all the variables in the sets, deletes the less significant variable until each model reach a stable solution, a solution has been considered as stable when all the variables that compose the set are significative in:

- Absence of randomness.
- Coefficient statistically different from zero.
- Collinearity that is non-significant from a statistical point of view.

The three models are composed from different sets of variables. for the detail of the variables that compose each set please refer to the Table 3.4 in the

main body of the dissertation, here will be named as conventional variables and non-conventional variables. The first model is composed only by conventional variables (e.g. the annual interest rate, the revenues, and the assets), the second one only by non-conventional variables (e.g. Number of patents applications and dimension of the Board of Directors) and the last one that include both of conventional and non-conventional variables.
At the end of this first phase the statistics of the model are the following (Table 1: Statistics of the models.):

Table 1: Statistics of the models.

Model 1	
N.Obs	969
AIC	483.89
BIC	503.40
Fisher Scoring	8
Model 2	
N.Obs	969
AIC	404.17
BIC	428.56
Fisher Scoring	6
Model 3	
N.Obs	969
AIC	460.12
BIC	489.38
Fisher Scoring	10

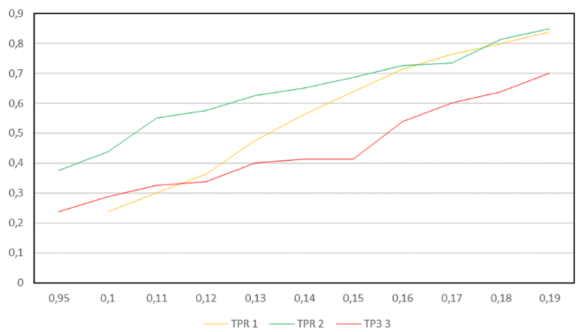
In all the statistics of the model it is possible to see how the second model results as the most performing, having the lowest AIC, BIC and reach the best likelihood in the lowest number of iterations.
The Table 2 represents the coefficient of the three logistic regressions.

Table 2: Betas of the Logistic Regressions.

Statistic	Beta	Beta	Beta
Intercept	-3.533	-3.603	-5.914
Age_BoD		0.041	0.0495
Listed	-2.345		-
Patents		-0.149	-0.335
Rate	34.725		37.63
Revenues	-2.01e-8		-1.46e-8
TM_D		-0.085	-
Tot_BoD		-0.321	-0.264

The following step is the analysis of the global performance of the three different models in terms of True Positive Rate and True Negative Rate. The three models perform similarly for the negatives rate, while the difference is quite prominent for the true positive rates. In this statistic we can see how the second model clearly represent the best fitting model, in fact, looking at the graph below it is possible to see how the second model reach strongly greater performance for low level of threshold while for the higher performs similarly to the first model.
Here we have another check of the correct choice that would be the second model.

Figure 1: performance of True Positive Rate to vary of threshold.



The last step in order to create an effective prediction model is determining the best threshold to be used.
From now on the lender point of view, called generally the bank, will be considered; the assumption, quite strong, needed only for the purpose of this paper is the following: in the market there is only one financial institution (the bank) that can grant all the debit required by the borrower. The objective of the bank is the maximization of the profits intended as the difference between the revenues and costs. Therefore, the bank is interested in the maximization of the following equation:

$$\pi = TP \times (r_{TP} - c_{TP}) + FP \times (r_{FP} - c_{FP}) + TN \times (r_{TN} - c_{TN}) + FN \times (r_{NF} - c_{FN})$$

The total number of True Positives, False Positives, True Negatives, and False Negatives is a function of the threshold exclusively (for the legend of the symbols, is recommended the redirection to the section 3.6.2. of the main body).
In this paper proxies have been created using the means of the interest rates and of the total amount lent, these type of indicators poorly fit the model and, therefore, the precise maximum results, has

not been chosen, but rather the point of elbow in which the profits start to become more stable and the advantages of the bank would be no more significant if compared to the strict policy that would be applied.
The thresholds that best fit the model and that could be the most useful in terms of prediction are:

- For the first model: 0.19.
- For the second model: 0.18.
- For the third model: 0.24.

In order to assess the goodness of the model obtained from the logistic regression a robustness test has been conducted using the classification trees. Trees are definitely clearer but poorer in performances if compared to the Logistic Regression.
The analysis conducted using classification trees shows comparable results in the types of correlations between the variables and the default probability.
It is possible to answer positively to both the research questions of this dissertation:

1. The correlation between the variables exists and it is shown in Table 2, in fact the presence of statistically significant coefficient means that the correlation exists, and the positive or negative sign give information about the type of correlation.
2. There is effectively a set of variable that performs better than the model built only on the more economic variables.

4. Comments on the results

The variables in this research have been grouped in two subset: the conventional variables and the non-conventional. From the first group the revenues and the interest rate resulted as the more significant variables. These two variables represent respectively the lender and the borrower perspective, higher interest rates lead to higher default probabilities, higher revenues strongly reduce this probability.

The second set of variables regards the innovation propension of the companies and the structure of their board of directors. The first aspect shows a negative correlation between the number of Intellectual Properties and the default probability, the second aspect shows positive correlation between the age of a board and the default, however, it is more interesting the strong positively correlation given by the total number of members in the board of directors: wider is the board clearer the firm success. The second important aspect on these variables is that the gender does not results significative in the model, this statement could be misleading because it might seem that there is no difference from a gender perspective, however, going in deep, it is possible to note that the differences in the gender gap are still incredibly strong, more that 80% of the members in the boards of directors are male, at the same point this high disparity makes more difficult to detect statistically the patterns.

5. Conclusions

The final output of this dissertation is the reasoning behind all the statistics developed previously. The most valuable resource on which a company can relies on is again the human brain, its inventing abilities, and the aptitude to cooperate brings better results than all the economic numbers that can be reached.
This paper has not at all the claim of being a whole comprehensive work, it is rather a starting point for research that put the human as the main resource also nowadays.

References

Angelino, A., Balda, F., Oliverio, F., Sampoli, L., & Schivardi, F. (n.d.). *Rapporto Cerved PMI 2022*.

Acknowledgements

We would like to express our gratitude, first of all, to our Professor Giancarlo Giudici, for the support that gives us during all the journey of this thesis.

Secondly, we would like to thank all the community of the Politecnico di Milano for all the competences and the development of our mind in these years.

Lastly, we would like to thank our beloved families, friends, and all people supporting us from the first day.

POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Default prediction of SMEs in the Italian Minibond market: an analysis under a new human perspective

TESI DI LAUREA MAGISTRALE IN
MANAGEMENT ENGINEERING
INGEGNERIA GETIONALE

Authors: **Matteo Spedicati**
Tommaso Tregrosso

Student ID: 10606292; 10613683
Advisor: Giancarlo Giudici
Academic Year: 2022-23

*"Learn from yesterday, live for today, hope for tomorrow.
The important thing is not to stop questioning."*

A. Einstein

*"It takes a great deal of bravery to stand up to our enemies,
but just as much to stand up to our friends."*

J.K. Rowling

Abstract

The default of companies has already been discussed for years in the global literature, in different form and treating companies of different typologies. The prediction analysis in the credit market is a key point in the assessment of debt from all the active agents, both from the lender and the borrower side. Banks and financial institutions for years exploit models and theories in the credit scoring algorithms and procedures, on the other hand the companies are interested in maintaining some indicators high enough to be considered safe and have loans granted.

In this dissertation the focus is on the firms that issued Minibonds in the Italian market, the company analysed are SMEs that historically are the core of the country and for which the access to credit could result more difficult than could be for bigger entities.

Nowadays the ever-increasing use and development of information technologies, like the machine learning algorithms or the rapid ascent of the artificial intelligence seemed to have left behind the attention on the human capital in its disparate forms and its abilities, for this reason will be analysed the performances of the firms that issued minibonds up to the end of 2020 looking at their capacity of developing new intellectual properties and the composition of the management.

The objective of the research is finding a prediction model that could classify the default of the companies through statistics that do not consider only the more economic variables, but rather new types of variables focused on the human capital trying to give the starting point on new typologies of decision for the credit market.

Key-words: default; minibond; SMEs; human capital; prediction

Abstract in Italiano

Il default delle aziende è già stato discusso per anni nella letteratura mondiale, in forme diverse e trattando compagnie di diverso tipo. L'analisi predittiva nel mercato del credito è un punto chiave nella valutazione del debito da parte di tutti gli agenti attivi, sia dal lato del finanziatore che del richiedente. Le banche e le istituzioni finanziarie sfruttano da anni modelli e teorie negli algoritmi e nelle procedure di credit scoring, mentre le aziende sono interessate a mantenere alcuni indicatori sufficientemente alti per essere considerate sicure e ottenere prestiti.

In questa tesi l'attenzione si concentra sulle imprese che hanno emesso Minibond nel mercato italiano; le aziende analizzate sono PMI che storicamente sono la spina dorsale del Paese e per le quali l'accesso al credito potrebbe risultare più difficile di quanto non lo sia per le imprese più grandi.

Al giorno d'oggi l'uso e lo sviluppo sempre maggiore delle tecnologie informatiche, come gli algoritmi di machine learning o la rapida ascesa dell'intelligenza artificiale, sembra aver lasciato indietro l'attenzione sul capitale umano nelle sue forme più varie e sulle sue capacità, per questo motivo verranno analizzate le performance delle imprese che hanno emesso minibond fino alla fine del 2020 guardando alla loro capacità di sviluppare nuove proprietà intellettuali e alla composizione del management.

L'obiettivo della ricerca è quello di trovare un modello di previsione che possa classificare il default delle aziende attraverso statistiche che non considerino solo le variabili più economiche ma piuttosto nuovi tipi di variabili incentrate sul capitale umano cercando di dare il punto di partenza su nuove tipologie di decisione per il mercato del credito.

Parole chiave: default; minibond; pmi; capitale umano; predizione.

Contents

Introduction

For many years, the issue of corporate default prediction has been covered in literature from all over the world, encompassing a variety of forms and company kinds. When evaluating debt, credit market prediction is an essential analysis for all the parts involved: lenders and borrowers. Models and other type of predictions have long been used by banks and other financial institutions in their credit scoring algorithms and processes. On the other side companies work hard to keep their indicators high enough to be considered safe and receive loans in the interim.

The companies that have issued Minibonds on the Italian market are the main subject of this dissertation. Small and medium-sized businesses (SMEs), which are historically the backbone of the nation, may have more difficulty obtaining finance than larger organizations.

The fast development and application of information technologies in the modern day, such artificial intelligence and machine learning algorithms, seems to have overcome the value of human capital in all of its forms and capacities. Consequently, the performance of the companies that issued minibonds will be investigated with an emphasis on the management team and the companies' capacity to create new intellectual property.

The goal of this dissertation is to create a predictive model that focus not only on economic predictor to categorize corporate defaults, but rather, on new factors focused on human capital.

In chapter 1 we will have a brief description of the main themes of this thesis and their already existing literature, then, in chapter 2, we will explain the procedure that have been used for the development of the model. In chapter 3 are reported the analysis and the results obtained, after this, in chapter 4, are presents some comments on the variables used and lastly, in chapter 5, we give a conclusive sum up and some hints for possible developments.

1 Literature Review

1.1. SMEs

1.1.1. Importance of SMEs in Europe

The definition of **Small-Medium Enterprise** (SME) is given by the European Commission through the Commission Recommendation of 6 May 2003 (EUROPEAN COMMISION, 2003).

"The category of micro, small and medium-sized enterprises (SMEs) is made up of enterprises which employ fewer than 250 persons and which have an annual turnover not exceeding EUR 50 million, and/or an annual balance sheet total not exceeding EUR 43 million."

Two more subgroups can be found within the general category:

- an organisation is considered small if it employs less than 50 people and has annual revenue and/or balance sheet totals under EUR 10 million;
- a microenterprise is defined as an enterprise which employs fewer than 10 persons and whose annual turnover and/or annual balance sheet total does not exceed EUR 2 million.

Class size	Number of enterprises		Number of persons employed		Value added	
	Number	Share	Number	Share	**Billion €**	Share
Micro	22 744 173	93,5%	38 790 351	29,4%	1419,4	18,6%
Small	1 332 200	5,5%	25 602 334	19,4%	1259,8	16,5%
Medium-sized	204 786	0,8%	20 493 722	15,5%	1266,5	16,6%
SMEs	24 281 159	99,8%	84 886 407	64,4%	3945,8	51,8%
Large	43 112	0,2%	46 918 978	35,6%	3673,8	48,2%
Total	24 324 271	100,0%	131 805 385	100,0%	7619,6	100,0%

Figure 1.1.: Summary of SMEs numbers in 2022 in Europe

In 2022, about 24.3 million SMEs were active in the EU-27 and these SMEs accounted for 99.8% of all enterprises in the non-financial business sector (NFBS). These SMEs employed 84.9 million people in the EU-27 in 2022. However, while almost all enterprises in the EU-27 NFBS were SMEs in 2022, the latter accounted for just under two-thirds of EU-27 NFBS employment and slightly more than half of EU-27 NFBS value added.

Figure 1.2: Share of different EU-27 SME size classes, employment and value added in NFBS.

Source: Calculations by the JRC based on Eurostat's Structural Business Statistics, Short*-Term Business Statistics and National Accounts Database*

The vast majority of SMEs in 2022 were micro-SMEs. However, these very small SMEs accounted for 36% of SME value added, and 46% of SME employment in the NFBS in 2022.

In terms of employment, micro enterprises account for a greater share of total SME employment than small SMEs (30%), and small SMEs account for more employment than medium-sized SMEs (24%).

The three SME size classes generated about the same proportion of SME value added in the EU-27 NFBS in 2022, with the share of value added generated by micro-SMEs (36%) being only slightly larger than the share generated by small and medium- sized SMEs (32% each).

These data are very important to figure out the crucial importance of SMEs in Europe: the European economy is totally dependent on SMEs and their growth becomes a fundamental topic for all the countries ((European Commission, Annual Report on European SMEs 2022/2023)

1.1.2. SMEs Financing Literature

As has been demonstrated previously, SMEs are vital to the European economy, so it is important to comprehend how they operate and, in particular, how they get their funding. The most significant research on SMEs' financing methods and financial structures will be provided in this section.

In their analysis, Bongini Paola et al., (2017) examine the factors that may affect SMEs' access to capital markets, specifically with regard to debt and equity. They also develop an indicator known as market suitability, which helps identify SMEs that may be a good fit for market-based financing.

The results demonstrate that a significant portion of firms suitable for market-based financing have unused potential, and that only a small number of Eurozone countries appear to have fully utilised their "potential" for capital market financing.

According to this paper, size, listed status, and growth potential are the three most crucial criteria taken into account when obtaining capital markets. It's interesting to note that there are significant differences between the nations as well. For example, states like Spain, Portugal, and Italy use fewer market-based instruments than their northern counterparts, most likely as a result of economic hardship. However, the situation changes significantly if the analysis is limited to medium-sized SMEs, with Southern European SMEs having a better position in terms of suitability for the capital market.

Turning our attention to bank lending, Demirguc-Kunt et al. (2008) noted that banks view the SME segment as highly profitable but also highly competitive in developed countries and unstable in developing ones. Because of this, they demonstrate that the lender environment—in this case, the SME segment—is the primary characteristic that banks look for when making financing decisions.

Ozturk & Mrkaic (2014) paper focuses on understanding other factors that affect SMEs' ability to obtain bank loans. The findings indicated that: a rise in bank funding costs is inversely correlated with firms' access to credit, albeit only in financially distressed economies; firms reporting a rise in their debt-to-asset ratios are substantially more likely to report a decline in their access to credit; the use of subsidies is inversely and positively correlated with all firms' access to credit, with the exception of large ones; and firm size and age are inversely and positively correlated with enhanced firms' access to credit.

Regarding the second section of this chapter, which deals with the financial structure, it is important to summarise the findings of N. Berger & F. Udell (1998). They explain that there are various optimal financial structures for different stages of a company's life cycle, ranging from equity capital—provided by venture capitalists and business angels—to short-term loans and, only subsequently, bonds, long-term loans, and other forms of financing. Therefore, knowing the age and stage of a firm's life cycle is essential to comprehending its financial structure.

In line with this theory, Lucey & Mac an Bhaird (2006) point out that as businesses get older, they rely more on retained earnings. However, they also noted that high growth businesses have more external equity financing and that there is a negative correlation between the fixed assets of their businesses and the personal collateral that entrepreneurs are required to provide. Therefore, age alone cannot account for the rationale behind a financial structure.

The leverage is another crucial factor that needs to be taken into account.Daskalakis & Psillaki, (2008) use a sample of French and Greek businesses to examine the factors that influence the capital structure of SMEs, or small and medium-sized businesses. They came to the conclusion that while firm size has a positive relationship with leverage, asset structure and profitability have a negative one.

Regarding growth, there is evidence to support the findings of Lucey, Mac, and Baird, showing a positive correlation between growth and debt levels and, consequently, leverage.

Finally, it goes without saying that one must take into account the fact that a nation's industrial and economic conditions can have an impact on the particular leverage that SMEs have within that nation.

Mantovani (2015) shows that companies with longer maturities within their balance sheet tend to be more transparent and perform better when examining the composition of financial liabilities. Benoit Coeurè first brought up the subject of transparency in 2013. He clarified that small and medium-sized enterprises (SMEs) typically have shorter credit histories, less informative financial statements, and higher fixed costs associated with external monitoring (EUROPEAN CENTRAL BANK, 2013). Because of all these factors, SMEs are less confident, which makes it harder for them to get bank financing. Another intriguing finding from Mantovani's paper is that, while high profitability companies rely more on shorter maturities, high growth companies rely more on longer maturities. This suggests that banks place more weight on a company's potential for growth than its existing state.

It is evident that credibility is essential for obtaining funding, but how can one obtain it? These days, banks consider a variety of factors in addition to economic drivers, such as managers' integrity, ethics, and dependability (Howorth & Moro, 2012). They show how extended ties between banks and businesses foster trust, which is essential for lowering agency issues and, in turn, debt costs. After examining the financial statements of Belgian SMEs from 1997 to 2010, Vander Bauwhede et al. (2015) discovered that the quality of financial reporting was a significant factor influencing debt accessibility. They demonstrated that the accuracy of a company's financial report and the cost of debt have a negative relationship.

Following the 2008 financial crisis and the Credit Crunch, the financing landscape underwent a significant upheaval as a number of new financing options proliferated, drastically altering the organisational structure of businesses. Taking these novel features into account, some scholars began examining the applications and effects of these new financing formats for SMEs.

First, Chludek (2011) provides an explanation of the significance of trade credit as a substitute source of funding. Contrary to popular belief, which views trade credit as highly costly, he provided evidence to the contrary, highlighting its significance as a significant component of SMEs' ideal financing mix. As previously stated, other research disputes this theory; however, one conclusion that is universally acknowledged is that there is a discernible difference between firms that are constrained and unconstrained in how they finance investments, with the former concentrating more on bank loans and the latter on alternatives like trade credit.

Numerous scholars find the emergence of the mini-bond market in several European nations to be highly intriguing. Schweizer et al. (2015), for instance, conducted research on the German situation in order to comprehend the high percentage (20%) of default that German mini-bonds encountered. By contrasting the default probability indicated by credit risk models, such as the Altman Z score, with the default probability based on a mini-bond's initial rating, they demonstrate how rating agencies can inflate ratings by giving too high of a rating. Due to the "window of opportunity" this produces for lower-quality companies to compete for funding, bondholders bear more of the risk rather than shareholders. High-quality companies are motivated to signal their quality through mini-bond underpricing in this context.

1.1.3. Performance analysis of Italian SMEs

1.1.3.1. Reaction after Covid crisis

Italian SMEs reacted promptly to the shock of the pandemic, demonstrating remarkable resilience in limiting the damage caused by the crisis and strong dynamism in rapidly recovering lost ground and trying to recovery. The system's positive response is linked to both endogenous and exogenous factors. Indeed, in recent years, SMEs have become stronger in terms of capital and are more solid and less financially exposed; at the same time, the prompt reaction of the government and European institutions has supported companies' liquidity needs and generated positive expectations about the recovery. Following analysis and numbers are taken from “Rapporto Cerved PMI 2022” (Angelino et al., n.d.) .

After the decline in 2020 (-1.8%), estimates on the number of active SMEs in 2021 show a significant increase (+4.2%), returning to values above the pre-Covid level (163,349 vs. 159,925 in 2019). The balance sheet figures for 2021 confirm the positive trend, with a strong rebound in all income statement indicators. After a decline in 2020 (-8.2%), turnover grew by 14.5% to +5.1% compared to 2019. The rebound was led by medium-sized enterprises (+15.3%), which were relatively less affected in 2020 (-5.8%). The gross operating margin of SMEs recorded record numbers: +23.7% over 2020 and +13% compared to pre-Covid, driven by growth in value added and corporate cost containment policies after the recessionary phase. The growth dynamic is driven by the excellent performance of industry (+16.4% over 2020) and especially construction

(+17.9% over 2020), both stimulated by the numerous incentives made available by the government. The return on equity (ROE) of SMEs has also almost completely recovered its pre-Covid levels, rising from 8% in 2020 to 11.8% in 2021 (it was 11.9% in 2019).

Figure 1.3: Changes % in Turnover for different sizes.

Figure 1.4: Changes % in Gross Margin for different sizes.

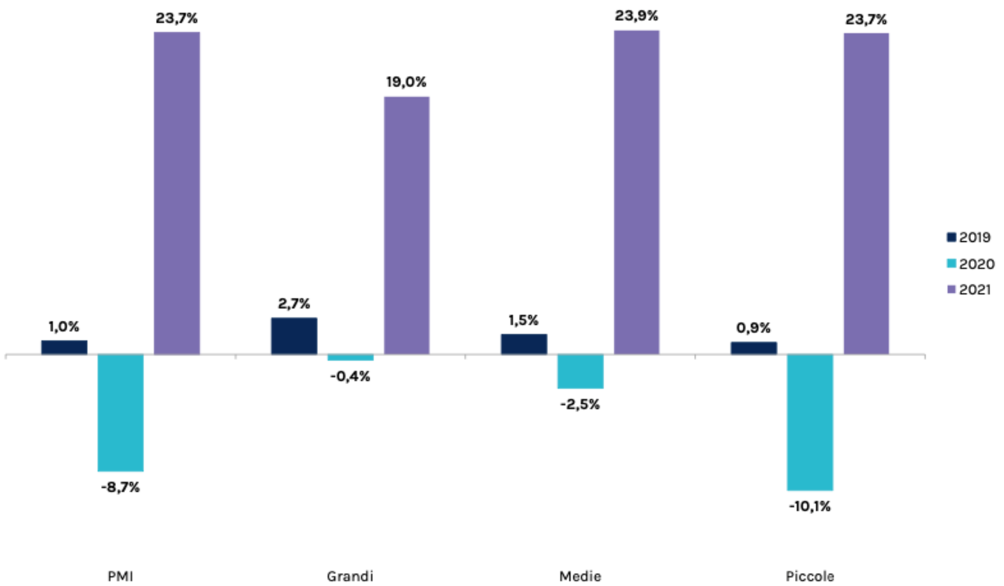
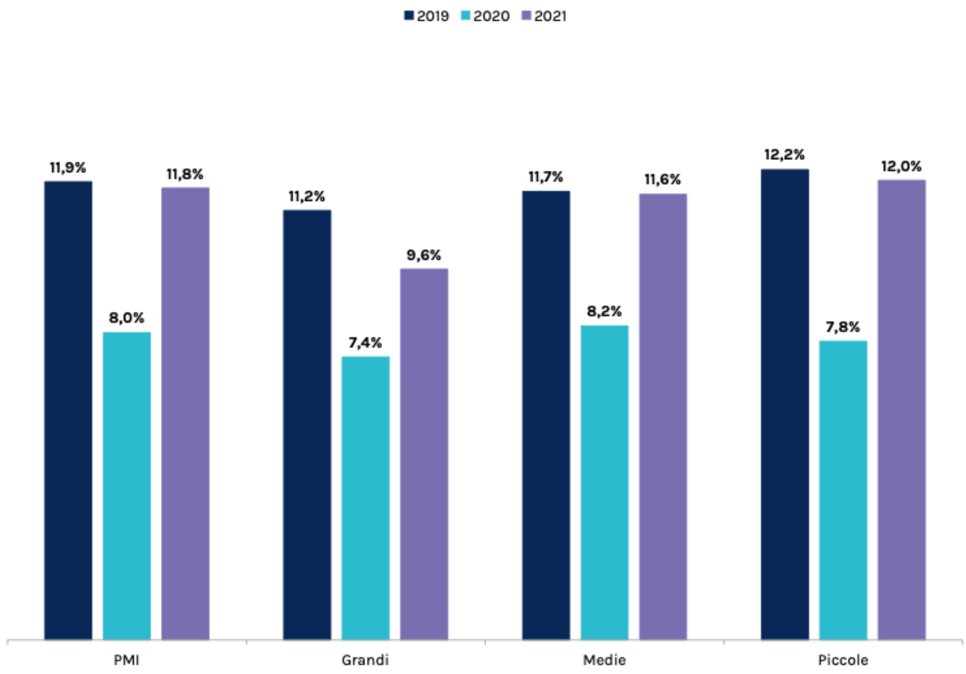


Figure 1.5: ROE for different sizes.



On the debt front, SMEs continued to increase their financial debts in 2021 (+7% for medium-sized companies, +6% for small ones), stimulated by low interest rates. Despite the sharp increase in debt over the past two years, SME leverage continued to decrease by 1.2 p.p. in 2021, from 61.1% to 59.9%, in contrast to large enterprises where it increased from 91.5% to 97.9%. The upward trend of SMEs is also certified by their payment habits and risk scores. Despite tighter deadlines, there is a reduction in the number of days SMEs are late in making payments to suppliers. Non-payments, after soaring in the most acute months of the pandemic, also show a clear decline. The economic-financial scores show a polarising trend in SME riskiness: solvent SMEs are increasing compared to before the pandemic, but risky companies are also growing.

Despite higher levels of riskiness compared to the pre-pandemic period, business closures continue to decline in 2021. Government guarantees have enabled the survival of SMEs with a very fragile financial structure, which, with the phasing out of liquidity support measures, could get into serious trouble, causing critical issues in the credit system.

1.1.3.2. Current situation with energy shock

Then, the prospects for post-pandemic recovery deteriorated in early 2022, with the outbreak of the Russian-Ukrainian conflict and the intensification of the energy shock. These two events triggered a series of geopolitical and macroeconomic changes that led to a rapid change of scenario.

Future scenarios are related to the escalation of the conflict, the evolution of the energy crisis, the ability to effectively implement the PNRR and the degree of tightening of ECB monetary policy. Now, it's difficult to find up-to-date data but we can find forecasts that distinguish two different scenarios: in the base scenario, real sales growth is expected to continue in 2023, while the most pessimistic scenario sees a reversal with declining revenues and gross margins. Actually, the forecasts reveal very heterogeneous trends that are mainly affected by the different exposure to the ongoing energy shock.

1.2. Bond and Minibond

A mini-bond is a debt security, either a bond (of any maturity) or a commercial paper (with maturity up to 36 months), issued by Italian companies – especially by small and medium enterprises, i.e. SMEs, and subscribed by professional and qualified investors (definition of mini-bond adopted by the Osservatorio Mini-bond of Politecnico di Milano's school of management). These instruments, in exchange for raising capital – that is reimbursed through a pre-defined schema, i.e. bullet or amortizing, - offer a remuneration, established in a reciprocal binding agreement, through the coupon payment. The issuance regulation that establishes the legal perimeter in which mini-bonds are treated is the art. 2410 – 2420 c.c. (i.e., Italian Civil Code), for which concern joint-stock companies (S.p.A. in Italian) and the art. 2483 c.c., for limited liability companies (S.r.l. in Italian), where the subscription of those instruments is limited only to supervised investors. In light of historical tendency of Italian firms to finance their need mainly through banks channels, especially among SMEs, until 2013 it was not common among unlisted companies to use mini-bonds to raise capital, even though these instruments can be perfectly compared with a well-known traditional fixed income security. Starting with the Development Decree, the government has introduced a series of tax incentives and regulations to encourage the development of this new instrument. The government's objective is to respond to the post-crisis credit crunch of 2008, the effects of which in the Italian market were amplified by the historical dependence of companies on bank credit as a source of financing.

The objective of this chapter is to provide a general overview of the mini-bond instrument by analysing the most important aspects. Since mini-bonds are a relatively new financial instrument, literature is still limited. The following chapter will be mainly based on the research activity conducted by the Osservatorio Mini-bond of Politecnico di Milano, extending it whenever possible (Osservatori Entrepreneurship Finance & Innovation & - Politecnico di Milano, 2023).

1.2.1. The reasons behind the creation of the mini-bond market

To understand the reasons that led the legislature to create the mini-bond market in Italy, it is necessary to better understand the economic and industrial context in which these instruments were created.

The effects of the 2007 financial crisis profoundly affected the Italian production system: the crisis was transmitted to the real economy and then created a vicious circle in which the sovereign debt crisis, fiscal policies, banks' capital tensions and regulatory constraints at European level played a significant role. In Italy, the effects of the crisis were amplified by the bank-centric nature of the production system, characterised by a weak capital structure and poor relations with foreign investors. In addition, the lack of Private Equity development in Italy, the administrative and fiscal complexity, the rigidity of the stock market and the fact that the bond market has historically been reserved only for large companies have all contributed to worsening the situation for Italian SMEs. Looking at Bank of Italy data, it is clear that SMEs, which make up an important part of Italian industrial world, have much more difficulty in accessing bank credit than large companies.

In addition, in Italy, the possibility of diversifying the sources of financing for Italian SMEs, through channels other than banks, is very limited due also to the difficulty in determining the creditworthiness of these companies. In other countries, this problem is solved thanks to the presence of specialised intermediaries that grant money to companies, mainly unlisted, through special credit funds.

Due to this situation, there was a sharp contraction in the supply of capital (credit crunch), coupled with an increase in the tax burden and difficulties in collecting debts from public institutions and private customers. The consequence has been that the

Italian production system, affected by these problems, has found increasingly tough to operate in a global competitive environment. From this, the need to structurally change the way the Italian production system is financed.

1.2.2. Differences between Mini-bond and traditional Corporate bond

From a technical point of view, mini-bonds are quite similar to traditional corporate bonds, whereby corporate bonds we mean medium- and long-term debt securities that can be issued by both a listed and an unlisted company in order to obtain financial resources to finance development projects and/or extraordinary investments. So, they represent an alternative channel to the banking system and thus allow for a diversification of financing sources

The duration must be longer than 36 months for the instrument to be considered a mini-bond and in most cases is around 4-5 years (other more short-term alternative debt instruments to the banking channel are promissory notes). The redemption modalities of the bond can be bullet or amortising and, in some cases, there can be an option to convert the bond into shares of the issuing company.

To resume, we can highlight the differences between mini-bonds and corporate bonds in the following points:

- Type of issuers: mini-bonds are almost always issued by SMEs and unlisted companies while corporate bonds are issued by large companies
- Availability of issuer information: in the case of traditional corporate bonds this information is normally widely available while it is not always available for SMEs issuing mini-bonds (e.g. business plans)
- The role of the Investor Relator: in the case of corporate bonds, issuers generally have an Investor Relator figure who is accustomed to dialoguing with the financial markets, while in the case of mini-bond issuers this figure is very often not present

- Liquidity of financial instruments: traditional corporate bonds generally have good liquidity in the secondary market, whereas in the case of mini-bonds, despite the introduction of a multilateral trading platform by Borsa Italiana, they remain very illiquid instruments.
- Market information: in the case of traditional corporate bonds, there is standardised and detailed information on the market, whereas the mini-bond market is characterised by greater information asymmetries and scarcity of information
- Availability of historical data: for corporate bond issuers, a track record will generally be available and historical series are also available in the market. These data are, however, not available for the mini-bond market.

1.2.3. Mini-bonds advantages

We have already mentioned that mini-bonds are financial instruments created to ensure easier access to credit for Italian SMEs. In reality, debt securities of this type were already available in the past for unlisted companies and therefore also for SMEs, with the exception of micro-enterprises; the great development witnessed in recent years, however, is due to changes in regulations that have strongly encouraged the use of these instruments.

The advantages and potential associated with these instruments for Italian SMEs are several. First, they provide the possibility for unlisted SMEs to access international capital markets, thus providing an important complementary channel to bank debt. Mini-bonds also allow a diversification of financing sources leading to an improvement in financial management and a mitigation of the risks associated with excessive dependence on the banking channel. In addition, the fact that these instruments are characterised by a medium-long maturity guarantees greater credit stability for a medium-long period, avoiding the risk of early requests for repayment of credit lines as can occur with bank debt.

Another interesting aspect is the effects on the balance sheet related to the extension of the average duration of the company's balance sheet liabilities downstream of the use of mini-bonds. This can lead to a general improvement in balance sheet ratios measuring the consistency between the liquidity of investments and the degree of collectability of financing sources. The improvement in these ratios and in the company's balance sheet and financial equilibrium also leads to an improvement in the company's creditworthiness. The consequence is that the potential for access to the banking channel also increases and it is easier to obtain low interest rates due to the fact that the company is considered more reliable.

We must also consider the size of the loan, as an SME can more easily convince investors in the capital market than a bank to grant a large loan relative to the company's turnover on the basis of the company's growth prospects. In this case, the mini-bond instrument is not an alternative to the banking channel but rather the only possibility to obtain capital for the realisation of certain investments. An example of this type could be the first mini-bond placed on the Extra-Mot pro market and issued by CAAR S.p.A., a company with a turnover of around EUR 5 million that was able to obtain a loan of EUR 3 million in this way, which it would have had difficulty obtaining from banks.

A further advantage, introduced by the new legislation, concerns tax benefits:

- The deductibility of the costs for issuing the mini-bond
- The exemption of withholding tax on income paid on securities traded on regulated markets or multilateral trading systems of EU Member States or other countries on the White List. The aim is to avoid double taxation on the company

- The deductibility of interest expenses, according to the same rules provided for listed companies, with reference to securities listed on regulated markets or multilateral trading systems and subscribed by qualified investors holding no more than 2% of the issuer's capital.

Finally, we have the qualitative advantages of mini-bonds especially related to the visibility that the issuance process provides to companies. In fact, often the reasons that drive a company to issue mini-bonds are related to marketing objectives.

1.2.4. Mini-bonds disadvantages

There are, of course, critical issues and potential disadvantages for enterprises associated with the issuance of mini-bonds. First, the issuance of mini-bonds requires an increase in transparency towards investors, not only in terms of the balance sheet, but also in terms of the goals, investments, and projects the company wants to realise. This increase in transparency, although in many cases having positive effects at the level of management control, comes at a cost and provides important information regarding the company's strategy for competitors.

In fact, the high illiquidity of securities requires in-depth due diligence on the issuing company to go into the sector in which it operates, its competitive positioning, corporate governance, strategy, financial structure, and historical and prospective financial-economic balances.

Then there are a series of costs related to the feasibility assessment of the transaction and the issuance process in general. These costs concern all the charges relating to the various parties involved (in-house consultants, advisors, auditing firms, arrangers, rating agencies, etc.) and obviously also the costs of listing on regulated markets. On this last point, it is important to emphasise that the ExtraMOT Pro segment of Borsa Italiana, the trading platform where mini-bonds are listed, actually provides for particularly cheap listing fees for SMEs.

Regarding the timing of issuance, in general raising capital through the financial markets takes longer than financing through the banking channel and this could be a problem for some companies.

Finally, it is important to emphasise that mini-bonds, unlike bank credit, are not a suitable solution for all types of companies. In fact, according to observations on the Italian market, mini-bonds must have a size limit of several million euros in order to be attractive on the financial market, especially to foreign investors. With the aim of overcoming this limit, it was decided to introduce the so-called *"basket bonds"*, i.e. innovative instruments that aggregate the needs of several companies in a single issue.

1.2.5. Regulatory framework

As anticipated before, since 2012 numerous legislative decrees were introduced in Italy, with the objective of making the bond market equally accessible to non-listed SMEs like in the other European countries. This section is based on the comprehensive summary proposed by Osservatorio Mini-bond of Politecnico di Milano .

1.2.5.1. "Development" and "Development-Bis" Decrees

The Development Decree Law (D.L.) No. 83 of June 22, 2012, which was later converted into Law No. 134 of August 7, 2012, brought the mini-bond tax system in line with that of listed companies, which had historically benefited from a more favourable scheme, and unlisted companies. More precisely, unlisted businesses can deduct interest costs for IRES up to 30% of gross profit when they issue debt instruments that are traded on regulated markets. In contrast, the interest expense is only deductible in the following situations: 1) the securities are held by qualified investors (banks, insurance companies, investment funds) that do not hold more than 2% of the issuer's capital; and 2) the income beneficiary is a resident in Italy or another nation on the White List (a list of nations that permit an adequate exchange of information). This is the case when the mini-bonds are not traded on regulated markets or are exchanged in multilateral trading facilities. The "Development" Decree also included additional tax benefits. First, it extended the D.L. 239/1996 tax system on interest received by investors to joint stock companies traded on the MTS and to unlisted companies. In addition, all issuing fees (advisers, arrangers, rating agencies, etc.) are deducted in the year that the cash principle-based expenses are incurred. Lastly, it eliminated the cap on the amount of debt instruments that unlisted companies could issue for more than twice the equity of their shareholders, given that the securities would be listed on multilateral trading facilities or regulated markets.

1.2.5.2. "Destination Italy" Decree

Within the framework of D.L. No. 145/2013, the "Destination Italy" Decree introduced certain measures to support medium- to long-term alternatives to bank financing. For instance, it expanded the scope of Law 130/1999, which deals with the securitization process, to include bonds (excluding hybrid and convertible instruments) in order to encourage the establishment of new funds that will purchase mini-bonds, thereby reinforcing the demand of the instrument. Additionally, it removed the 20% withholding tax from interest and revenue payments made on mini-bonds subscribed by funds whose capital is exclusively used to purchase mini-bonds and whose shares are held by professional investors.

Furthermore, this law made it possible for SMEs, who typically don't have properties with high values, to get financing using their assets without being prohibited from using them. This guarantee lowers the capital cost while offering investors more protection.

Additionally, the Central Guarantee Fund for SMEs was extended by the "Destination Italy" Decree to include mini-bonds subscribed by qualified investors. Specifically, single issues account for 40% of the fund, while mini-bond portfolios account for 60%. The requirements are:

- the issue must be used to finance the business and not to substitute other debts;
- the securities must be subscribed after the management board decision to approve the guarantee;
- the maturity must be between 36 and 120 months;
- the amount covered by the Fund must not be supported by other guarantees.

The maximum guarantee in the event of a single issue is € 1.5 million, and 50% of the subscription for amortised bonds and 30% for bullet bonds. Guarantees for mini-bond portfolios consisting of individual subscriptions up to 3% in amount can range from € 50 million to € 300 million.

1.2.5.3. "Competitiveness" Decree

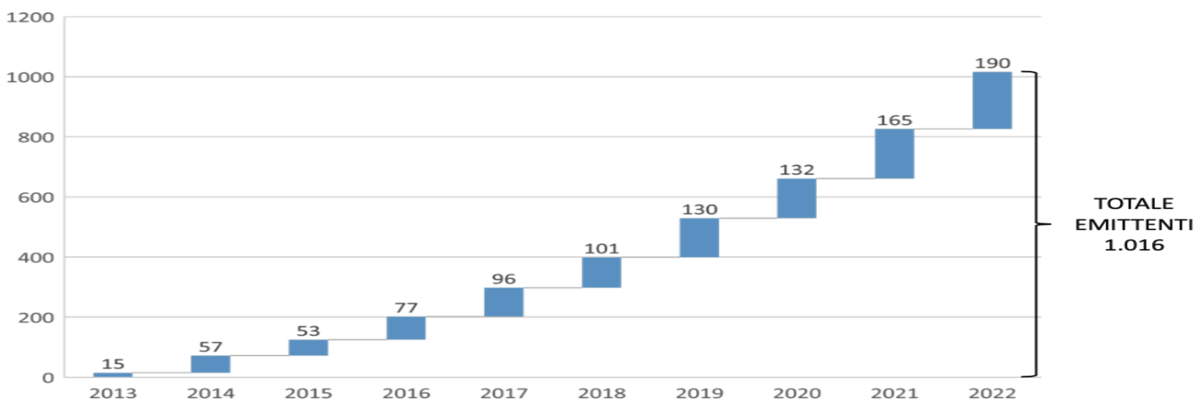
A number of initiatives were introduced by the D.L. 91/2014 ("Competitiveness" Decree) to increase the competitiveness of Italian businesses and attract foreign investors. In particular, it is established that businesses may receive direct financing from investment funds, insurance companies, and securitization companies. In order to attract foreign capital, it eliminated the withholding tax on medium-long term loans made by international investment funds, insurance companies, and securitization firms. Additionally, it eliminated the 26% withholding tax that applied to interest and proceeds from unlisted bonds that were sold to institutional investors. Lastly, it extended the substitute tax to sales of guaranteed receivables.

1.2.6. Italian Minibond Market: issuers and issuances' analysis

Following the results of *"9° report Italiano sui Minibond"* (Osservatori Entrepreneurship Finance & Innovation & - Politecnico di Milano, 2023), the number of companies that placed minibonds for the first time in 2022 (added to those that had also placed securities in the past) was 190 compared to 165 in 2021; this is a new record in the

Italian industry, as Figure 1.6 shows. In total, there were 254 issuers in 2022, up from 200 in 2021.

Figure 1.6: New Italian mini-bonds issuers year over year.



The 2022 was thus a year of further diffusion of awareness of minibonds among companies; to note that basket bond transactions played a key role.

The majority of 2022 issuers are SMEs (according to the definition adopted at European level). In particular, 178 out of 254 comply with the size requirements; the percentage is 70.1%, higher than in 2021 (it was 66.5%).

The issuing companies analysed, which placed minibonds under €50 million from 2013 to 2022, form a sample of 1,016 companies. The total number of SMEs is 663, so 65.3% of the total.

Figure 1.7: Segmentation of issuers for turnover size since 2013

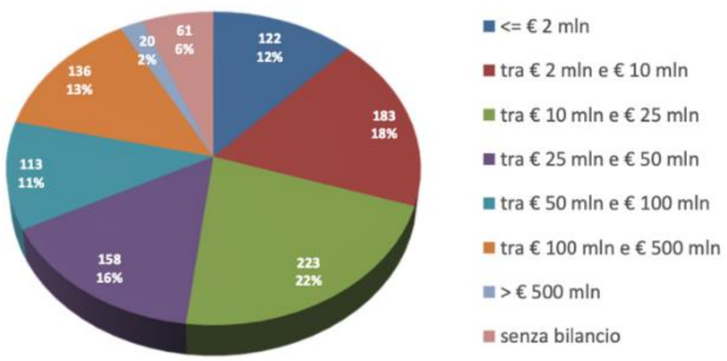


Figure 1.8: Segmentation of issuers for turnover size in 2020,2021 and 2022

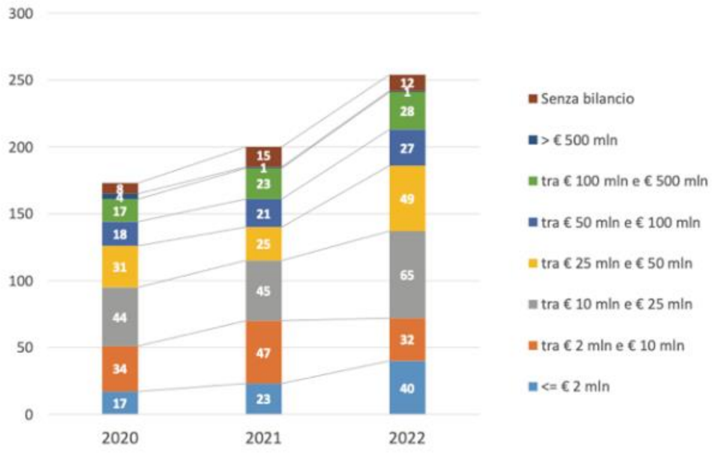
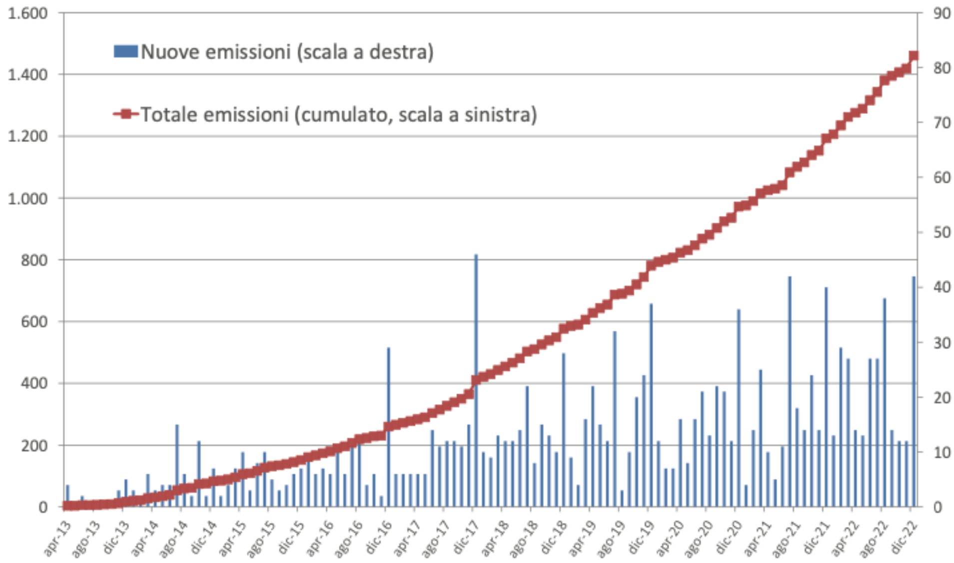


Figure 1.7 segments the total sample according to the consolidated turnover in the last available financial statements before issue. It should be noted that the largest class is the one between € 10 and € 25 million (223 observations, or 21.9% of the total) followed by the group between € 2 and € 10 million (183 cases, or 18.0%). For 61 companies, no financial statements had yet been filed at the time of issue; these were almost always companies that were formally newly established but resulted from the demerger or restructuring of existing businesses. As Figure 1.8 shows, 2022 saw a good increase in the number of issuers with revenues between €10 million and €50 million, while those with revenues between €2 million and €10 million returned to 2020 levels. Issuers with annual turnover from € 50 to € 500 million are also growing. It therefore seems that in 2022 the market preferred more established and mature companies than in the previous year.

Figure 1.9: Italian mini-bond issuances since 2013.



The number of minibond issues surveyed by the Observatory during the year 2022 of an amount below € 50 million is 268. In contrast, 221 issues were recorded in 2021, an increase of +21.3%, which is very good in view of the worsening overall situation, with the war in Ukraine, inflation and rising interest rates. Issues were distributed fairly between the two half-year periods; as usual, there was a 'peak' at the end of the year (42 issues in December). Starting from 2013, we have a total sample of 1,461 placements.

Lastly, Figure 1.9 updates the cumulative year-end trend of minibond issues since the reforms initiated by the 2012 Development Decree came into force. It is confirmed that minibonds hold their own "stable" market.

1.3. Default prediction

1.3.1. General overview

In the fields of finance and economics, the applicability of business failure prediction is seen as a major issue. Forecasting bankruptcy on time and preventing default could allow to adopt some actions to restore firms' financial situations. There are several models to predict bankruptcy, such as discriminant analysis (Beaver, 1966;Altman, 1968), logit and probit models (Charitou et al., 2004; Jones & Hensher, 2007), artificial neural networks (Wilson & Sharda, 1994), genetic algorithms (Kingdon & Feldman *, 1995), survival analysis (Gepp & Kumar, 2008; Shumway, 2001) and recursive partitioning algorithm (RPA) (Frydman et al., 1985).

These models choose the most pertinent financial ratios that impact the chance of default to effectively identify enterprises based on their financial health and estimate the risk of business failure. Although these models differ, they all have one crucial feature: the distressed status of a company, represented by a binary variable, must be known beforehand (that is, when the event of bankruptcy has already occurred: Amendola et al., 2017).

However, managers, financial institutions, practitioners, and other interested stakeholders need to prevent the distress or, alternatively, accelerate the liquidation; avoiding letting assets of the distressed firms lose their value so as to keep the indirect costs of bankruptcy from increasing (Bisogno & De Luca, 2015). As such, it may not be sufficient to simply develop a model to predict the probability of bankruptcy. The actual economic situation, which has been significantly impacted by the global financial crisis, has highlighted how crucial it is to implement early warning systems in order to avoid bankruptcy.

It is important to underline the difference between forecasting models and early warning systems. The former are primarily concerned with determining which variables could more accurately predict the probability of default and calculating the failure risk. They are powerless to stop a potential problem for a company. Conversely, early warning systems consist of tools for monitoring and detecting warning events to predict and ward off an upcoming business crisis and implement a timely intervention. Consequently, putting such a system into place entails realizing that company failure is a process that progresses through a number of troubling situations, bankruptcy being only the last stage.

Despite this distinction, the terms "early warning systems" and "forecasting models" are commonly used interchangeably, which poses the risk of conceptually overlapping them and hiding the complexity and dynamic nature of business failure processes.

Since the 1930s, a significant amount of research has been done on the forecast of corporate financial difficulties. The earliest research focused on using ratio analysis to forecast bankruptcy; it progressed from univariate models (Beaver, 1966) to multivariate investigations, the most well-known of which was published by Altman (1968). Then, several strategies have been developed with the goal of offering more reliable and accurate models for bankruptcy prediction. In an attempt to identify the most effective failure indicators, researchers have largely implemented models that compare distressed and healthy organizations (Amendola et al., 2011; du Jardin, 2010).

The number of businesses in danger of going bankrupt increased significantly as a result of the Great Depression of 1929 and the most recent financial crisis, which started in 2008, which stimulated research on bankruptcy prediction. Now, the literature classifies forecasting models into two categories (Alaka et al., 2018). The first one consists of statistical models, which analyses two samples of healthy and distressed firms, and where the selection of financial ratios having a predictive ability is based on empirical studies. The model's parameters and default probability are then estimated using the chosen ratios.

The second group of approaches used to forecast the risk of default is based on soft computing techniques, whose primary assumption is that data can be incomplete and environmental conditions can change over time. Because of this, these approaches are designed to consider the possibility that certain parameters could change in response to shifting environmental factors; as a result, these models are dynamic and frequently known as learning systems.

Models adopted in previous studies can be classified in accordance with the selection of factors, that is, financial ratios used to predict bankruptcy. These factors are usually classified into several groups, expressing the performance of a firm (Maksimovic & Phillips, 2001), its liquidity and solvency (Rege, 1984), the leverage (Heiss & Köke, 2004), the size (Heiss & Köke, 2004) and so on. It is also worth noting that many studies have classified ratios into several categories. The most common are:

- Profitability ratios, such as Return on Equity (ROE), Return on Assets (ROA), Return on Sales (ROS), Net Income to Total Assets, Net Income to Total Debts, Sales to Fixed Assets, Finance Charges to Net Sales and so on.
- Liquidity and solvency ratios, such as Current Ratio, Quick Ratio, Current Assets to Fixed Assets, Inventory to Current Assets, Working Capital to Total Assets, Liquid Assets to Total Assets, Cash Flow and so on.
- Size and capitalisation, such as Total Assets, Long- term Assets to Total Assets, Net Worth to Fixed Assets, Net Worth to Total Debts, Total Debts to Total Assets and so on.
- Turnover ratios, such as Inventory to Sales, Accounts Receivable to Sales, Accounts Payable to Sales, Total Debts to Sales.
- Operating structure ratios, such as Labour Cost to Production Cost, Labour Cost to Net Sales, Finance Charges to Debt, Finance Charges to Financial Debt.

The main goal of this classification is to make ratio interpretation easier, supporting the selection of those having a good predictive ability. Several studies have also found that corporate governance indicators influence the prediction of bankruptcy (Brédart, 2014; Chen, 2014; Liang et al., 2016). These indicators can be classified into five groups, including board structure, ownership structure, cash flow rights, key person retained and others. While some authors (Chen, 2014; Lin et al., 2010) have demonstrated that the performance of models can be enhanced by combining financial ratios and corporate governance indicators, there is disagreement regarding the optimal variable set to be employed in the model estimation process to forecast business failure.

The models aiming to forecast bankruptcy, described so far, adopt different approaches (statistical vs. soft computing techniques) and select ratios to be used as predictors of financial distress. Despite the differences, these models are substantially based on an ex-post perspective, comparing samples of distressed and healthy firms, namely firms whose status is already known.

As a result, these models may not always be able to assist practitioners in making decisions when they are (or will be) involved in transactions with a declining firm. As mentioned earlier, financial distress is a dynamic process that changes over time, and bankruptcy is only the last step when no other options remain (Volkov et al., 2017). This would imply that in order to determine whether a struggling company has a future chance of surviving, other exit strategies should be taken into account. Legislation in many nations is steadily moving toward prompt intervention, sometimes without the courts' direct involvement, with corporate rescue as primary goal.

Focusing on the European context, some legislative innovations in insolvency laws have taken place in countries like Germany (1999 and 2012), England (Enterprise Act 2002), Poland and Romania (2003 and 2006), Spain (2004 and 2013), France (2006 and 2014), Finland (2007), Greece (2007 and 2012); Italy (2017) and so on. Although several differences still exist between these legislations, a common orientation towards corporate rescue can be observed, as an alternative to liquidation procedures (Nicolaes Tollenaar, 2017).

The EU itself has recently proposed a new directive, aiming to implement preventive restructuring frameworks, in order to give firms a second chance and increase the efficiency of restructuring, insolvency and discharge procedures. Even if this proposal has been criticised from a juridical perspective (Nicolaes Tollenaar, 2017), from an economic point of view, it intends to facilitate timely interventions.

From a theoretical viewpoint, this would mean investigating and implementing models to prevent financial distress. Indeed, in the light of a timely intervention, the main problem is not to forecast default risk but mainly to prevent it.

Finally, it should be noted that comparing models can be challenging due to the various definitions of business failure that have been accepted. It is crucial to remember that, as Schary (1991) has noted, bankruptcy is just one option for leaving the market. In light of this, researchers (Balcaen et al., 2012; Chancharat et al., 2010; Jones & Hensher, 2007) suggest focusing on the different forms of exiting the market, considering not only bankruptcy but also voluntary liquidation as well as merger and acquisition (M&A), which constitute out-of-court exit procedures. It should be further observed that the vast majority of previous studies, with few exceptions, have adopted an ex-post perspective, testing the predictive accuracy of the model they propose on a sample of failed and non-failed firms, that is, firms whose status is already known. However, in several circumstances (e.g., when a bank is going to decide whether or not to lend money to a firm), an ex-ante perspective is required. Therefore, there is a need to investigate from an ex-ante viewpoint the probability of moving from healthy status to another (for example, liquidation or bankruptcy).

More precisely, in light of the aforementioned and the fact that a company's failure is a process that changes over time, we would suggest that the following statuses should be taken into account when forecasting the risk of default:

- Default of payment, which occurs when a firm (more generally, a debtor) starts not to pay its debts regularly
- Insolvency proceedings, which occur when a firm is unable to pay its debts. In this case, even if the insolvency is declared, the firm remains active, though it is in administration or receivership or under a scheme of arrangement. During this period, the firm is usually placed under the protection of the law, continues operating and repaying creditors, while trying to reorganise its operating activities. At the end of the procedure, the firm will alternatively return to normal operating (the default of payment was thus temporary) or will be reorganised (parts of its activity can be restructured or sold) or be liquidated
- Bankruptcy, which occurs when a firm is formally declared distressed since it is not able to pay its creditors. The court will appoint an insolvency expert, whose main aim is to sell the assets and repay the debts. At the end of the procedure, the firm will be dissolved.
- In liquidation, which occurs when all the assets of the company are being sold, and the firm will be dissolved.

1.3.2. Focus on SMEs

Over the last dozen years, the topic of small and medium enterprise (SME) default prediction has developed into a relevant research domain (with the need to develop new SME default predictors), motivated by the enormous toll on SMEs caused by the 2007–2009 global financial crisis as well as the recent COVID-19 crisis.

Effective default risk prediction of small and medium enterprises has always been a concern of financial institutions and bank managers. In addition, with the adoption of Basel II in 2004, banks were forced to compute their capital requirements based on the ratings assigned to their borrowers by their internal rating systems, including SMEs.

For them, the construction of quantitative models for default prediction based only on financial ratios, seen in the previous section, is particularly complicated and the results that can be obtained are usually far less accurate than in the case of larger firms. There are many reasons for this, some of which are (Ciampi & Gordini, 2013):

- the fewer legal obligations of SMEs regarding accounting data disclosure than larger firms: less information is available, and what can be obtained is less reliable and less accurate.
- the greater impact of external events that change company structure or behaviour, and that may, for instance, modify a state of crisis, by allowing a firm's weak points to be strengthened (that is, financial intervention by the owners, the appointment of new managers or changes in strategy). When a SME is predicted as being likely to default (or not to default), there is a greater probability that the prediction will be inaccurate because external events might intervene and either save the firm or, alternatively, lead to an unexpected, and sudden, collapse.
- the fact that smaller firms have automatically smaller figures (including accounting figures), and even small changes have a more marked effect on ratios and percentages. To the extent that, in terms of what they can reveal about a firm, some ratios are completely ineffective below certain dimensional levels.

The regulatory changes in the banking industry, the abovementioned specific characteristics of SMEs, the fact that they play an important role in the world's economy and the lower prediction accuracy rates obtainable from SME accounting data, are all factors which make it essential for both banks and academics to construct and test default prediction models that are specific for SMEs and include other variables, in addition to economic–financial ones, related to other company features

This combination of concurrent use of non-traditional quantitative methodologies and non-financial predictive variables has allowed improvements in the prediction accuracy of SME models over the years (Ciampi and Gordini, 2013).

So far, several researchers have explored SME default prediction potential of innovation-related variables, also using big data and non-traditional methodologies based on machine learning/non-linear programming.

Several authors (Moon et al., 2011; Moon & Sohn, 2010; Sohn et al., 2012; Sohn & Kim, 2013) investigate the relationship between technological innovation and SME default risk. They find that the level of profitable technology owned by a SME together with the legal representatives' technological knowledge reduces the probability of default, especially by companies in high-tech industries.

One of the most relevant findings in this cluster is to demonstrate the effectiveness of non-financial factors related to product innovation (patents and brand products) as predictive variables together with financial ratios, macroeconomic indicators, and some characteristics of legal representatives such as age, gender, and the value of their real estate properties (Chi & Meng, 2019; Yu et al., 2019). Few recent studies in this cluster try to solve the methodological issues generated by the need to combine different sources of information to assess SME default risk better, suggesting the use of machine learning/non-linear programming tools such as MCDA, fuzzy clustering, and cognitive mapping.

In combining different sources of information, (Gonçalves et al., 2015) develop a multiple criteria decision analysis (MCDA) based system for SMEs credit risk assessment by including, in addition to the traditional financial variables, innovation-related variables (R&D capability and reinvention capability), as well as to human characteristics (honesty, friendliness and ethical posture) of the SME managers, commercial aspects, management experience and skills, and external factors (legislation, financial background and political instability). Using non-linear programming tools, Pan et al. (2017) starts a promising stream of research by analysing the SME default predictive power of big data related to real transaction-based trade areas. They find that adding a variety of Big Data types, such as those related to credit card sales, improves default prediction accuracy rates, especially for low-risk SMEs.

This chapter ends with the latest theories on new cluster of variables. In the following parts of the dissertation the objective will be to produce an analysis with different sets of predictors in order to understand better the possible magnitude of these new frontiers.

2 Objectives and Data Collection

In this chapter we will examine the process that guides this paper, analysing the data collected, how we have collected them, and the models used to reach the results.

We will focus on two types of variables:

- The "Conventional Variables" are all those variable that concern debt structure, its issue, or the structure of the company itself in economic terms.
- The "Non-Conventional Variables" are those variables that concern the innovation and intellectual properties inside the company, and the structure of the Board of Directors.

2.1. Objectives

The primary objective of this dissertation is to understand the connections between the default of the Italian companies that issued minibonds in comparison with some of their structural variables.

The default probability connected with the size and other conventional variables has been highly debated in the literature; therefore, we will focus on a different approach that consists in the use of non-conventional variables in order to understand if there is a correlation between the human capital, the ability to conduct research, to exploit designs, and the default probability.

We also analyse the different prediction ability of the models that will be used in three scenarios, first of all using the conventional variables, then a second using only the non-conventional ones, and the third one using both of them in a mixed model.

The final goal is to have a better prediction, or at least a comparable one, using the new types of variables rather than the prediction obtainable with classical predictors.

2.2. Research question

The main purpose of this dissertation is to understand, considering the sample already present in the Department of Management Engineering of the Politecnico di Milano, enriched by our research, which are the most significant variables in predicting the default probability of the firm that issued minibonds.

We will start with the univariate statistic to recognize some evidence. The first question of this dissertation, therefore, is:

Is it possible to observe any pattern between the different variables and the default of a firm?

The research to answer this question leads to a more in-depth discussion for which are the sets that better clarify the default prediction. In particular, if there are new types of variables that can perform better than the already well-known in predicting the default probability.

This leads to the second question:

Are there any new types of variables that can perform a default probability prediction that are different from the more known indicators?

The instruments that will be used to establish if there is a connection are two different econometric models.

We will carry out a logistic regression on three different sets of variables and then a classification tree will be used as robustness test.

We will conduct this part of the analysis dividing the variables in different sets, once created the models with the relative variables, we will analyse the performance of each model.

These research will give the possibility of understanding if the variables that regard the human capital and the innovation capacity can lead to new choices: from the lender perspective, looking at these new variables rather than the use of the classical predictors in the credit scoring, and from the borrower, enhancing the exploitation of the human capital in order to reach the goals needed to survive in a evermore dynamic environment that needs more proactive choices.

2.3. Data Collection

Our research started with data collection from the available sources that we had in possession. We used different databases from which we got the information we needed. Our starting point and the core for the analysis is the database developed by the team of the Department of Management Engineering of the Politecnico di Milano in which are present all the data about the minibond issuance from the 7th of November 2012 to the 17th of February 2023, and groups about two thousand issues.

For our purposes we did not focus on all of them, but only on all the emissions up to 31st of December 2020, the reason behind this choice is derived from the availability to check the status of the companies, in fact, for the emissions in 2021 we would have data only about the last two years, for issues of the 2022 we would have only the last Balance Sheet and, in some cases, the ones referred to the 2022 are not yet available; while for the emissions of 2023 we have no information at all on the Financial Statements.

Starting from the documentation of the companies in the database we took the Financial Statements for each year from the year of issuance of the bond to its expiration year. In order to get these documents, we used two different platforms:

- AIDA: "the online database created and distributed by Bureau van Dijk S.p.A., containing the financial statements, personal and merchandise data of all active and failed Italian capital companies (excluding Banks, Insurance companies, and Public Administration)" [1]. The advantage in this database is the opportunity of downloading the file needed in an Excel format or consult them in an online spreadsheet. The platform is not open source, and it is needed a log in.
- Telemaco: is the database released by the Italian Chambers of Commerce in which are presents all the information about the companies and entities

registered in each single Chamber of Commerce across the Country. This platform has the advantage of having records of all the firms in it, however, the great disadvantage of this tool is the need to download the entire Financial Statement year by year. Also, this platform is not open source and needs a log.

The phases descripted below are conducted manually since the data are not standardized enough to use an algorithm or an automation tool.

2.3.1.1. First analysis – Preliminary check

From the first analysis we get, from the platform AIDA, the majority of the documents of the companies that did not have any type of structural problem; the first difficult is on the typologies of firms available: the access in fact is limited to the capital companies that are still active excluding therefore all the commercial partnership, cooperatives, and failed companies. The second difficulty in AIDA is the format of the data presented. In Italy we have a several number of companies that deposit their Financial Statements in shortened form; in this type of form, we do not have the distinction between the different voices of the liabilities, that are presented in a unique voice or, in other cases, the firms accounted the minibonds as another type of debt different from bonds, still other times, in the shortened form the liabilities are split only in debt and non-financial liabilities, making in all these cases impossible understand the bond real properties.

In all the plain cases we researched the fiscal code of the legal entity, and we inspected the Balance Sheet voice “Obbligazioni” in the liabilities part for each year of the bond life. In the cases in which we could exclude any problem we checked the firm with a flag “OK”; this first label has been given only to those companies that respect the structure of the debt given in the database of the Osservatorio Minibonds of the Politecnico: it was needed a total equivalence of amount, maturity, and repayment scheme. In all the other cases we conducted more in-depth research. The first analysis ended with the identification of 543 companies that were in a completely clear situation.

2.3.1.2. Second analysis – Completion on Telemaco

The second analysis has been focused on the companies not found in AIDA, going through the second platform (Telemaco). We had ninety-five entities not found in the first database, and, in this case, we had to check their status one by one through their fiscal code. During this type of analysis, we have found different scenarios, the two principal ones were a failed company or an error in the scan of the fiscal code. At the end of this second part of analysis every firm in the database could be in an "OK" status or in a "NOT OK" one.

2.3.1.3. Third analysis – Detailed research for "NOT OK" companies

The third part analysis focused only on the "NOT OK" companies. We returned on the analysis on AIDA and checked all the companies that redeemed the bond in anticipation. In these cases, we flagged the entities with an "OK" status, since it was not a default situation, however, we labelled them with an "EARLY REDEMPTION" flag in a second space. After this, we analysed all the remaining firms downloading all the Financial Statements through Telemaco and scan them for each bond issue in the notes to the Financial Statements. This is a profoundly time-consuming activity since for each company and for each year we checked manually the value of the debt, considering a low standardization of the data it is possible to understand the complexity of this operation. As said before some firm released their financial statements in a shorted-form, other ones did not categorize the information in the bond part or even in the financial debt one. In other cases, the research of the correct voice in which were accounted the minibonds has been difficult because all the Financial Statement was simply a black and white scan from paper documents. With the completion of this third phase, we had a complete overview on the status of each company:

The main status (the use of this type of variable is fundamental for the analysis part) has been named "DEFAULT":

- "NO"; the company with this flag is a firm that during the analysis would be flagged without a default problem, it is important not to confuse the "NO" status with the perfect match between the information on the database and the ones found during this research phase, by way of illustration, some companies have been involved in a process of Merge & Acquisition and the new resultant firms take over the remaining debt repurchasing it regularly, in these occasions the firm is labelled as "NO".
- YES: the company with this flag is one that would figure out as a firm with default troubles.

Then we created a list of more detailed information, this variable will be named "Status":

- EARLY REDEMPTION: the company in this status reached an agreement with the bond holders and redeemed the bond before the maturity date; it relates to a main status NO.
- REDEFINED MATURITY: in this case there are the firms that have delayed the bond reimburse for any reason; it can be connected to either a NO or a YES status.
- FAILED; we are in the case of a failure proceeding in place or already terminated. It is linked with a YES status.
- TERMINATED; the company in this status is no longer active and could be associated with a NO or YES status.

- MERGED; the company with this flag has been included in an M&A agreement, according to the situation it could be either NO or YES.
- RESTRUCTURING AGREEMENTS: in this section we found companies that started this type of proceeding; we described this as YES cases since the firms involved in this type of proceeding are not repaying the bond issued.
- IN LIQUIDATION, the firms labelled in this way are in a liquidation phase. Even if it could be in both states in our database, they are classified as in a YES condition; the liquidation procedure has been started from an external agent and the debt has not been repaid following the predicted scheme.
- INSOLVENCY PROCEEDING: all the firms in this section are represented as YES ones.

For every firm there is also a more detailed description on happenings that makes clearer the reasons behind their allocation.

2.3.1.4. Fourth analysis – Collection on IPs information

The fourth analysis has been conducted in order to collect data about the presence of trademarks and registered designs at the EUIPO office on the online website of the above-mentioned office. The Office provides a free access repository in which all the trademarks, the designs, the specifics about the intellectual property, and their ownership are present. For each firm we have researched its company name because they are not registered with their fiscal code. This gave us additional difficulty in finding a unique code that identifies the firm, since the one used by this office is not the fiscal code. Once we have found the name used by the company in order to register the property (in some cases the firms have used different names or there are firms that have similar name), we collected the unique code used by the office, the so-called "Owner-ID" and through this ID we researched all the trademarks and designs registered in the Office.

To collect the number of patents retrieved worldwide we used a tool called ESPACENET. With the IDs already researched we collect the number of patents and applications registered around the world.

- EUIPO: is the European Union for the Intellectual Property Office, in this office are registered all the types of intellectual property like patents, trademarks and designs; it facilitates the diffusion and the standardization of these properties across the European Union.
- ESPACENET: is an open-source search engine developed by the EPO, the European Patents Office, thanks to which it is possible to explore all the information about an application of a patent, also if the invention would not be patented. The EPO is different from the EUIPO, since the first office includes countries not belonging to the European Union and regards only the patents procedures.

2.3.1.5. Fifth analysis – Collection of information on the composition of the Boards of Directors

The last part of the data collection has been characterized by a new, different approach to the two databases used in in the first three phases of the analysis. We collected information from Telemaco and AIDA about the number, gender distribution, and age of the representatives in the Board of Directors of each firm. For every company we took the following information:

- Age of the principal representative expressed in years (integer number). Named "Age_Ch"
- Gender of the principal representative indicated as a dichotomic variable. Named "Gender_Ch"
- Mean of age of the totality of the Board of Directors. Named "Age_BoD"
- Total number of members in the Board of Directors. Named "Tot_BoD"

- Gender percentage in the Board of Directors. Named "Gender_BoD"
- Age difference between the principal representative and the mean of the Board of Directors; this element has been standardized dividing the difference by the age of the principal representative. Named "STD_Diff_Age"

It is important to note that we identify the principal representative as the Chairman of the Board of Directors or, in the cases of a unique representative, the Sole Director.

With the completion of the fifth part, we completed the macro phase of the data collection.

3 Data Analysis and Results

3.1. Introduction

This section is the core of the paper and started with the initial database enriched by the findings of the previous part of our work, the objective is build a model that can predict the level of insolvency of a firm that issued minibond, the model will be based on the two types of variables (conventional and non-conventional) in order to understand better which are the correlations between them and the performances associated.

A crucial part of the work is how clean all the data found; at the beginning of this phase we have a dataset formed by 95 attributes of which the majority of them are not useful for the purpose of our analysis, for example we have the moment of issue and the moment of expiration of the bond, these two attributes could seems quite significative at a glance, however they are very significative from a descriptive point of view, the market conditions vary year by year and issuing a bond in a given moment rather than another could be critical for the survival of a company, but from a predictive point of view they are completely useless, since it is impossible going back in the past.

After this screening procedure we can reduce the field of attributes to twenty-eight different ones:

- ID: is the unique, progressive number that identifies each bond issue, it gives us a chronological order in the minibonds issued over the years.

- ISIN: is the unique number that identifies each bond issue from a legal point of view, in the financial market every instrument has an alphanumeric code that identifies it.
- CF: is the fiscal code of the entity that issued the minibond.
- Name: is the complete name of the company.
- Motivation: is the reason behind the bond issue, it is a factorial variable that can assume four distinct levels or a combination of them: external growth (M&A processes), refund or restructuring a debt, internal growth, Cashflows, ESG .
- Amount: is the total amount of the debt required, this value could be different from the real amount of money collected, its value is a number expressed in millions of Euros.
- Price: is the par value of the bond, it is a numeric variable.
- Rate: is the annual interest rate adjusted as a fixed annual rate, the values are numeric and expressed in percentual points.
- Covenants: this is a dichotomic variable that flags the presence of covenants.
- Warrants: in the original database the data are expressed as a percentage of the total amount, in order to increase the explainability of this variable we modified it in a dichotomic variable that signals if it is present or not any type of warrant.
- Status: look at the previous section.
- Default: look at the previous section.
- Patents: is the total number of Patents registered in any worldwide (it includes both the pending applications and the approved ones also if expired. It could be possible filter only the approved patents looking for the kind code "B1"). The variable is an integer.

- TM_D: is the total amount of trademarks and designs registered at the EUIPO Office. The values of this type of attribute are treated as an integer.
- ATECO: is the code that represents the belonging industry of the firm, all the firms registered at the Chamber of Commerce belong to a specific industry. The variable in this case is a factorial that includes only the letter of the specific ATECO code of the firm.
- Region: is the belonging region of the firm, the variable is considered as a factor, different for each region.
- Geo: is the distinction in North, Centre, South of Italy, or if it comes from a foreign country, also in this case the variable is represented as a factorial attribute.
- Listed: is the dichotomic variable that identifies if a company is listed or not, it values as one if the company is listed in the market.
- Assets: is the total amount of the assets of a firm, the value is the last amount of assets in the Balance Sheet of the year of issuing, it is an integer variable expressed in Euros.
- Revenues: is the total amount of the revenues of a firm, the value is the last amount of revenues in the Income Statement of the year of issuing, it is an integer variable expressed in Euros.
- Equity: it represents the total equity, the value is the last amount of the total equity in the Balance Sheet of the year of issuing, it is an integer variable expressed in Euros.
- Employees: is the number of employees registered in the business profile of the Chamber of Commerce, the value refers to the last update available at the date of issue of the minibond and is registered as an integer.

- Age_Ch: it is the age of the chairman or any other type of principal representative, it is expressed as an integer.
- Gender_Ch: It is the gender registered in the Chamber of Commerce; it is a dichotomic variable.
- Age_BoD: It is the mean of the ages of all the members of the Board of Directors, included the principal representative, it is expressed as a numeric variable.
- Tot_BoD: It is the total number of members of the Board of Directors, including the principal representative, it is expressed as an integer.
- Gender_BoD: It is the percentage of males and females in the composition of the Board of Directors, included the principal representative. In order to make calculations we chose the count of males as a +1 for every male member, then the sum is the total count divided the number of members inside the board; the variable is a numeric one defined between 0 and 1 both comprised.
- STD_Diff_Age This variable has been built starting from the previous ones , in particular it represents the difference of the ages between the chairman and the rest of the members of the Board of Directors, this difference has been standardized with the age of the principal representative. We built up this variable because in the vast majority of the firms analysed the principal representative is quite older than the other members. This variable is numeric, and it is defined in minus one plus one both excluded.

Once that the number of attributes is reduced, we proceeded with the conditional reduction of the number of observations.

Applying the filters described in the previous chapter, we have a new database of 969 observations defined by twenty-eight attributes that represent the core of our analysis.

For the analysis we used the RStudio tool, a computing environment that has already implemented inside different statistics tools, in fact the models used are already available in RStudio. It could be possible to use more classical and famous instruments like MATLAB, in that language the code would be heavier even if its programming language is more spread. Being a less known and more sectorial, we insert the scripts of codes in Appendix A.

3.2. Univariate analysis

The Table 3.1 shows the univariate analysis on all the variables that are a fundamental for the development of the model. The sample size is fixed, in detail, the total number of observations is 969 of which 889 that resulted in No Default, the remaining 80 observations are Default observations. In the Table 3.1 we represented the principal statistics, in detail we have reported: the mean, the standard deviation, the first quartile, the median, the third quartile, and the total number of observations. Each statistic has been calculated for the total amount of observations and then the same statistics have been calculated for the defaulted observation (labelled as "YES") and the non-defaulted ones (labelled as "NO").

For the dichotomic and multilevel variables, there is no sense in the calculation of the indicators in Table 3.1 . We report here the total count of the observation defaulted and non-defaulted for each non-numerical variable in Table 3.2 and Table 3.3.

It is important to note that the statistics result quite similar in the cases of the total sample and the non-defaulted companies. The reason behind it is that only the 8.256% of the observation are defaulted companies.

Table 3.1: Statistics on the numeric variables.

Variable	Subset	Mean	Std. Dev.	1st Quart	Median	3rd Quart	N.Obs
Age_BoD	TOTAL	57,040	8,471	52,000	56,333	61,333	969
	YES	60,780	9,679	54,950	60,330	67,250	80
	NO	56,700	8,278	51,800	56,333	61,000	889
Age_Ch	TOTAL	61,100	11,328	54,000	60,000	69,000	969
	YES	62,500	10,390	57,000	61,000	69,000	80
	NO	60,980	11,405	54,000	60,000	69,000	889
Amount	TOTAL	6,049	7,723	1,100	3,000	7,500	969
	YES	6,038	8,103	1,200	2,150	7,000	80
	NO	6,050	7,693	1,100	3,200	7,500	889
Assets	TOTAL	74,190	123,702	12,670	30,600	83,220	969
	YES	357,912	54,425	8,903	18,584	46,326	80
	NO	77,650	127,564	13,740	31,550	89,200	889
Employees	TOTAL	189,000	399,024	21,000	69,000	172,000	969
	YES	113,000	512,060	3,000	27,000	73,000	80
	NO	195,900	386,872	22,000	76,000	179,000	889
Equity	TOTAL	22,421	51,450	2,568	7,162	21,536	969
	YES	10,316	19,189	1,528	3,216	11,388	80
	NO	23,511	53,277	2,704	7,778	23,589	889
Gender_BoD	TOTAL	0,844	0,214	0,667	1,000	1,000	969
	YES	0,879	0,243	0,950	1,000	1,000	80
	NO	0,841	0,211	0,667	1,000	1,000	889
Patents	TOTAL	16,060	134,285	0,000	0,000	1,000	969
	YES	1,225	3,901	0,000	0,000	0,000	80
	NO	17,390	140,121	0,000	0,000	1,000	889
Price	TOTAL	98,080	9,822	100,000	100,000	100,000	969
	YES	99,370	0,971	98,000	100,000	100,000	80
	NO	97,960	10,243	100,000	100,000	100,000	889
Rate	TOTAL	0,0451	0,0182	0,0350	0,0450	0,0550	969
	YES	0,0578	0,0167	0,0500	0,0575	0,0646	80
	NO	0,0439	0,0179	0,0323	0,4300	0,0550	889
Revenues	TOTAL	56,510	127,552	6,619	19,630	51,920	969
	YES	15,495	35,824	1,356	6,991	15,622	80
	NO	60,200	132,120	7,620	22,710	56,560	889
STD_Diff_Age	TOTAL	0,050	0,143	0,000	0,019	0,141	969
	YES	0,021	0,104	0,000	0,000	0,058	80
	NO	0,052	0,146	0,000	0,025	0,145	889
TM_D	TOTAL	9,934	100,883	0,000	0,000	2,000	969
	YES	0,950	2,500	0,000	0,000	0,000	80
	NO	10,740	105,289	0,000	0,000	2,000	889
Tot_BoD	TOTAL	3,963	2,734	2,000	3,000	5,000	969
	YES	2,438	1,590	1,000	2,000	3,000	80
	NO	4,100	2,774	2,000	3,000	5,000	889

Table 3.2: First part of the statistics on non-numeric variables.

Variable	Total	YES	NO	Defaulted on Total	Difference with Total
Motivation 0	46	5	41	10,870%	2,614%
Motivation 01	1	0	1	0,000%	-8,256%
Motivation 02	26	1	25	3,846%	-4,410%
Motivation 023	1	0	1	0,000%	-8,256%
Motivation 1	95	5	90	5,263%	-2,993%
Motivation 12	29	3	26	10,345%	2,089%
Motivation 123	1	0	1	0,000%	-8,256%
Motivation 13	1	1	0	100,000%	91,744%
Motivation 2	579	55	524	9,499%	1,243%
Motivation 23	18	0	18	0,000%	-8,256%
Motivation 3	73	2	71	2,740%	-5,516%
Motivation 4	9	1	8	11,111%	2,855%
Motivation n.a.	90	7	83	7,778%	-0,478%
Region Abr	15	1	14	6,667%	-1,589%
Region Bas	5	1	4	20,000%	11,744%
Region Cal	4	0	4	0,000%	-8,256%
Region Cam	97	7	90	7,216%	-1,039%
Region EmR	84	3	81	3,571%	-4,685%
Region Foreign	1	0	1	0,000%	-8,256%
Region FVG	20	1	19	5,000%	-3,256%
Region Laz	45	5	40	11,111%	2,855%
Region Lig	28	3	25	10,714%	2,458%
Region Lom	234	27	207	11,538%	3,283%
Region Mar	33	1	32	3,030%	-5,226%
Region Mol	1	1	0	100,000%	91,744%
Region Pie	67	8	59	11,940%	3,684%
Region Pug	25	2	23	8,000%	-0,256%
Region Sar	6	0	6	0,000%	-8,256%
Region Sic	17	2	15	11,765%	3,509%
Region Tos	33	1	32	3,030%	-5,226%
Region TAA	79	7	72	8,861%	0,605%
Region Umb	17	1	16	5,882%	-2,374%
Region VdA	2	0	2	0,000%	-8,256%
Region Ven	156	9	147	5,769%	-2,487%
Total N.Obs	969	80	889	8,256%	0,000%

Table 3.3: Second part of the statistics on non-numeric variables.

Variable	Total	YES	NO	Defaulted on Total	Difference with Total
ATECO A	9	0	9	0,000%	-8,256%
ATECO B	6	0	6	0,000%	-8,256%
ATECO C	419	29	390	6,921%	-1,335%
ATECO D	42	4	38	9,524%	1,268%
ATECO E	37	3	34	8,108%	-0,148%
ATECO F	93	11	82	11,828%	3,572%
ATECO G	78	5	73	6,410%	-1,846%
ATECO H	13	1	12	7,692%	-0,564%
ATECO I	39	3	36	7,692%	-0,564%
ATECO J	100	5	95	5,000%	-3,256%
ATECO L	22	1	21	4,545%	-3,710%
ATECO M	88	11	77	12,500%	4,244%
ATECO N	28	5	23	17,857%	9,601%
ATECO Q	16	2	14	12,500%	4,244%
ATECO R	9	0	9	0,000%	-8,256%
ATECO S	5	0	5	0,000%	-8,256%
Covenants 1	515	39	476	7,573%	-0,683%
Covenants 0	454	41	413	9,031%	0,775%
Gender_Ch M	884	73	811	8,258%	0,002%
Gender_Ch F	85	7	78	8,235%	-0,021%
Geo North	670	58	612	8,657%	0,401%
Geo Centre	128	8	120	6,250%	-2,006%
Geo South	170	14	156	8,235%	-0,021%
Geo Foreing	1	0	1	0,000%	-8,256%
Listed 1	81	1	80	1,235%	-7,021%
Listed 0	888	79	809	8,896%	0,640%
Warrants 1	482	35	447	7,261%	-0,995%
Warrants 0	487	45	442	9,240%	0,984%
Total N.Obs	969	80	889	8,256%	0,000%

3.3. Description of the models elaborated

The multivariate analysis has been conducted with two different typologies models:

- Classification Tree: models based on trees enable the segmentation of the space of the of attributes in a number of regions in which the prediction is constant. The algorithms based on trees work dividing the number of observations in two different groups, the first one is the so-called training set, its main role is the build of the tree also knowing the objective variable; the second group of observation is the test set, in this case the objective is the validation of the tree generated previously. Obviously, the results on the test set are worse if compared with the ones obtained from the training set, however, only the results from the test group are useful in assessing the model quality. The algorithm can recognize autonomously the variable needed, eliminating the non-significant ones. Trees are useful for understanding and easily interpreting the results and possible patterns in the dataset, but their performances are typically worse if compared to different classification or regression methods like the Logistic Regression. In order to overcome the performance issue, the model can be empowered with the use of random forests. This development of the model consists in the creation of different trees randomizing the variables and the observations used. However, this new implementation of the model becomes difficult in understanding, losing its main advantage. For this reason, we did not use the random forest model in our analysis and the classification tree model is used to give a brighter explanation of the results; this model has also been used as a robustness test. The "height" of each couple of branches in which the tree is split, gave us known about the importance of the node that generated them. The model would predict the dichotomic variable Default.

Here there is an example of a classification tree in Figure , represented only for illustrative purposes.

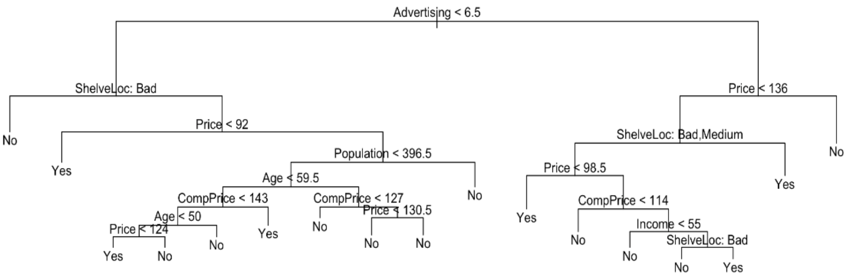


Figure 3.1: An Example of a Classification Tree

- Logistics Regression; this model consists in the attribution of a probability between 0 and 1 to each observation following the coefficients of the variables associated, the variable considered are all those available in the database, however, in the building phase of the model it is necessary paying close attention to all the variable that we are going to use, using an exaggerated number of variables the algorithm usually does not converge, in these cases is necessary to eliminate manually not significant variables. After this procedure the model can be refined and made more understandable through a stepwise process that consist in eliminating in each iteration the less significant variable; a variable is considered not significant when the associated coefficient is enough close to zero to make the variable incommensurable compared with the others. Then it is set a threshold between 0 and 1, the threshold probability, and for the observations that have a probability that the model estimated higher than the arbitrarily chosen threshold the dichotomic variable is set to 1 otherwise its value is zero. Choosing the threshold completely arbitrary is a not so rigorous method, in order to overcome this, it is possible to use two different systems: the ROC curve if we have well balanced dataset and the cost of false negative and false positive is quite the same. The second way is the plotting of a graph in which we can see the behaviour of the two index in correspondence of the increase of the probability, and choosing a probability that would prefer the true positive in spite of the true negatives or vice versa. This method is more rigorous than the classification tree but has a worse explainability if compared to the first one. The dichotomic variable that will be predicted is the Default variable.

In the following images it is possible to see: an example of ROC curve in **Error! Reference source not found.**, a graph with the ongoing of the index in **Error! Reference source not found.** and **Error! Reference source not found.** and the generic graph of a logistic regression with a single predictor in Figure .

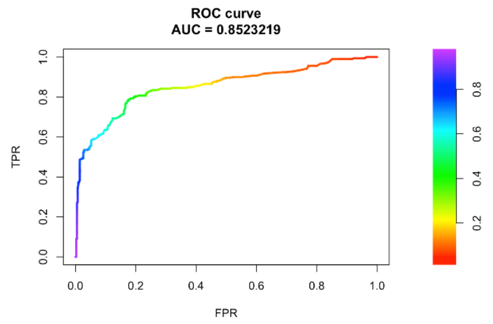


Figure 3.2: An Example of a ROC Curve

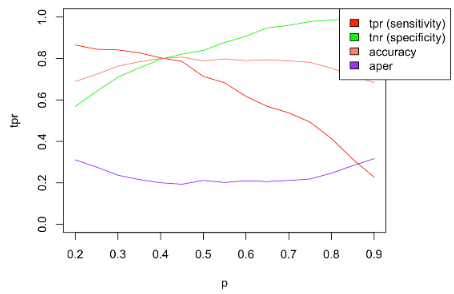


Figure 3.3: an Example of a TPR-TNR graph

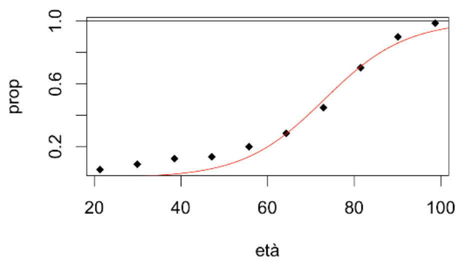


Figure 3.4 Example of a Generic Logistic Regression

The equation (3.1) is the generic form of a Logistic Regression:

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \cdots + \beta_n \times X_{nk} \qquad (3.1)$$

- Y is the probability associated with the default; it is defined in the space [0;1]. The higher values are connected to the positivity (positive index means default).
- k is the index of observations; it is defined in the space [1;969].
- β_0 is the intercept of the model it is defined in the space (-∞;+∞)
- β_n is the coefficient of the n^{th} variable, it is defined in the space (-∞;+∞)
- X_n is the value of the n^{th} variable for k^{th} observation, the definition space is variable for the type of variable.
- n is the index of variables [1;22]

We split for each model the database in three different subsets of variable: in the first case we have analysed only the conventional variable, in the second subset we have used only the non-conventional ones, while in the third we built-up the models using all the variables.

The division of the variables is the following:

Conventional Variables:

- Motivation.
- Amount.
- Price.
- Rate.
- Covenants.
- Warrants.
- ATECO.

- Region.
- Geo.
- Listed.
- Assets.
- Revenues.
- Equity.
- Employees.

Non-Conventional Variables:

- Age_Ch.
- Gender_Ch.
- Age_BoD.
- Tot_BoD.
- Gender_BoD.
- Patents.
- TM_D.
- STD_Diff_Age.

In Table 3.4 it is possible to find the all the variables that will be used in the analysis, split following the different models in which they appear

Table 3.4: Representation of all the variables used

Variable	Model 1	Model 2	Model 3
Age_BoD		Present	Present
Age_Ch		Present	Present
Amount	Present		Present
Assets	Present		Present
ATECO	Present		Present
Covenants	Present		Present
Employees	Present		Present
Equity	Present		Present
Gender_Ch		Present	Present
Gender_BoD		Present	Present
Geo	Present		Present
Listed	Present		Present
Motivation	Present		Present
Patents		Present	Present
Price	Present		Present
Rate	Present		Present
Region	Present		Present
Revenues	Present		Present
STD_Diff_Age		Present	Present
TM_D		Present	Present
Tot_BoD		Present	Present
Warrants	Present		Present

The composition of the database is the following:

- Number of observations, sample dimension: 969

Number of initial attributes: 24, the composition is 22 predictor X_i the status variable and the objective variable Y that is the dichotomic variable Default

3.4. Analysis before the building of definitive models

The analysis will be conducted through four different test:

- The first test is the Chi Square test, the objective is to understand if there is correlation between the X_i variable and the objective variable Y. The null hypothesis H_0 is that there is no correlation between the two variables. In the table are indicated the p-values.
- The second test is the Z-Test, the objective of this test is understanding the coefficient of the model β_i associated to the X_i variable. The null hypothesis H_0 the coefficient is equal to 0. In the table are indicated the p-values.
- The third test is the variance inflation factor also called VIF, the objective is to understand if there is a form of collinearity.
- The fourth test is the correlation rate matrix in order to understand which variable are correlated between them and how.

Beside the tests on the single variables, we have produced tests on the entire regressions, in detail we computed the following tests:

- AIC: The Akaike's information criterion gives us information about the relative quality of a model compared to the other two, we will prefer the model with the lowest value.
- BIC: The Bayesian information criterion gives us similar information of the AIC but with a higher penalisation on the complexity of the model in particular on the number of variables.
- Fisher scoring iteration: this number represents the number of iterations that the classification algorithm performs in order to reach the maximum likelihood, also in this case the lower the better.

We would like to remember that these tests will give an initial comparison between the models built, but obviously we will produce different test in order to assess the global performances.

For sake of completeness, we include in every table of the iterations the total number of observations.

In the iterations when the model reaches the definitive result the values are highlighted in green.

All the tests on the variables and the relevant statistics on the models are represented in a single table, the only test that will remain separated is the Correlation Rate Matrix that will be conducted for the first attempts in which will be important understanding which variables are correlated with each other.

In Table 3.5 it is possible to see a legend for the different levels of significance that will be used in the analysis.

Table 3.5: Legend of the levels of significance.

Symbol	Level of significance
***	99.9%
**	99.0%
*	95.0%
.	90.0%
	< 90.0%

3.4.1.1. First iteration

The first step to be conduct in order to reach our definitive model is a first attempt to design model without any type of check. The first attempt given in Table 3.6.

Table 3.6: Chi Square test for the first run and statistics on the models.

Models	Model 1	Model 2	Model 3
Test	Chi Squared	Chi Squared	Chi Squared
Statistics	p-value	p-value	p-value
Age_BoD		7.9e-6(***)	2.3e-6(***)
Age_Ch		0.145	0.203
Amount	0.679		0.679
Assets	3e-4(***)		4.1e-3(**)
ATECO	0.113		0.275
Covenants	0.323		0.323
Employees	0.409		0.497
Equity	0.962		0.871
Gender_Ch		0.859	0.733
Gender_BoD		0.869	0.796
Geo	0.623		0.649
Listed	6.4e-3(**)		7.2e-3(**)
Motivation	0.175		0.175
Patents		5.9e-3(**)	1.3e-3(**)
Price	0.125		0.125
Rate	6.8e-10(***)		6.8e-10(***)
Region	0.554		0.648
Revenues	1.4e-3(**)		3.3e-3(**)
STD_Diff_Age		0.178	0.424
TM_D		0.859	0.014(*)
Tot_BoD		4.5e-6	3.8e-3(**)
Warrants	0.560		0.560
N.Obs	969	969	969
AIC	542.11	509.16	520.97
BIC	839.56	503.55	857.43
Fisher Scoring	17	10	17

In this initial phase we conducted the second test only for the coefficient of the variables that are identified as multilevel factors. The reason behind this choice is that these types of variables are tricky to be used in R for their multiple levels in the analysis of the collinearity.

The multilevel factorial variables are the following: ATECO, Geo, Motivation, and Region. In order to understand if the following variables are significant, we conduct a Z test on the coefficients. The results are showed in Table 3.7.

Table 3.7: Z test for the first run.

Coefficient	Model 1	Model 2	Model 3
ATECOB	1		1
ATECOC	0.994		0.994
ATECOD	0.994		0.994
ATECOE	0.994		0.994
ATECOF	0.994		0.994
ATECOG	0.994		0.994
ATECOH	0.994		0.994
ATECOI	0.994		0.994
ATECOJ	0.994		0.994
ATECOL	0.994		0.994
ATECOL	0.994		0.994
ATECOM	0.994		0.994
ATECON	0.994		0.994
ATECOQ	0.994		0.994
ATECOR	1		1
ATECOS	1		1
GeoCentro	0.998		0.998
GeoNord	0.997		0.997
GeoSud	1		1
Mot01	0.991		0.990
Mot02	1		1
Mot023	0.651		0.519
Mot1	0.799		0.999

Mot12	0.206	0.454
Mot123	0.998	0.998
Mot2	0.998	0.249
Mot23	0.125	0.991
Mot3	0.992	0.714
Mot4	0.810	0.812
Motn.a.	0.701	0.755
RegionBas	0.277	0.857
RegionCal	0.996	0.996
RegionCam	0.412	0.336
RegionEmR	0.906	0.402
Regionforeign	0.998	0.998
RegionFVZ	0.988	0.550
RegionLaz	0.998	0.998
RegionLig	0.216	0.103
RegionLom	0.362	0.173
RegionMar	0.998	0.998
RegionMol	0.997	0.997
RegionPie	0.245	0.187
RegionPug	0.436	0.312
RegionSar	0.995	0.996
RegionSic	0.290	0.187
RegionTos	0.998	0.999
RegionTAA	0.286	0.174
RegionUmb	0.998	0.999
RegionVAo	0.999	0.999
RegionVen	0.406	0.199

After this second type of test, we can notice that in any case these variables result significant for our research. Before the second run of the algorithm, we can delete these variables from the analysis.

From this moment on we will conduct also the third and the fourth test, that will give us answers on the collinearity.

3.4.1.2. Second run

The results of the three models are in Table 3.8 with the relevant statistics for the variables and the entire logistic regressions, 18 predictors, objective variable: Default.

Table 3.8: Statistics on the second run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.14	0.111	-	-	-3.43	4.4e-4(***)	-	-	-7.99	0.036(*)	-	-
Age_BoD					0.164	0.071	7.9e-6(***)	52.365	0.177	0.071(.)	1.8e-5(***)	52.672
Age_Ch					-0.122	0.171	0.145	63.186	-0.128	0.442	0.081(.)	64.954
Amount	0.014	0.496	0.989	1.461					7.2e-3	0.733	0.989	1.438
Assets	8.2e-11	0.984	2.2e-4(***)	4.067					2e-9	0.641	4.7e-3(**)	4.174
ATECO	-	-	-	-					-	-	-	-
Covenants	0.060	0.816	0.687	1.083					0.206	0.437	0.687	1.686
Employees	3.4e-4	0.503	0.531	1.300					2.8e-4	0.519	0.674	1.746
Equity	-1.9e-9	0.858	0.795	3.603					-3.2e-9	0.780	0.933	3.983
Gender_Ch					-0.350	0.504	0.859	1.455	-0.424	0.442	0.747	1.347
Gender_BoD					-0.271	0.883	0.869	1.538	-0.019	0.978	0.989	2.189
Geo	-	-	-	-					-	-	-	-
Listed	-2.45	0.018(*)	1.2e-3(**)	1.036					-1.82	0.088(.)	7.3e-4(***)	1.089
Motivation	-	-	-	-					-	-	-	-
Patents					-0.150	0.047(*)	5.9e-3(**)	1.029	-0.210	0.469	1,5e-3(**)	1.101
Price	0.030	0.443	0.118	1,006					0.028	0.440	0.118	1.008
Rate	32.9	5.9e-7(***)	2.4e-10(***)	1.131					36.02	1e-6(***)	2.4e-10(***)	1.195
Region	-	-	-	-					-	-	-	-
Revenues	-2.3e-8	7.8e-4(***)	8.7e-5(***)	1.591					-1.8e-8	9.3e-3(**)	9.5e-4(***)	1.752
STD_Diff_Age					6.57	0.230	0.178	21.505	6.87	0.252	0.208	23.196
TM_D					-0.087	0.073(.)	0.859	1.040	-0.059	0.015(*)	0.015(*)	1.091
Tot_BoD					-0.271	5e-4(***)	4.5e-6(***)	1.240	-0.198	0.018(*)	3.6e-3(**)	1.320
Warrants	-0.601	0.023(*)	0.384	1.112					-0.632	0.022(*)	0.384	1.154
N.Obs	969				969				969			
AIC	490.84				409.16				468.71			
BIC	544.48				453.05				561.36			
Fisher Scoring	8				6				10			

In the Table 3.9 and Table 3.10 are represented the correlation rate matrix for the first two models (the third model is of difficult representation due to the high number of variables).

Table 3.9: Correlation rate matrix for the first model, second run.

	Amount	Price	Rate	Assets	Revenues	Equity	Employees
Amount	1						
Price	0,055	1					
Rate	0,144	0,020	1				
Assets	0,493	0,006	-0,080	1			
Revenues	0,228	-0,019	-0,070	0,643	1		
Equity	0,431	0,014	-0,112	0,868	0,392	1	
Employees	0,245	-0,027	-0,039	0,389	0,360	0,219	1

Table 3.10: Correlation rate matrix for the second model, second run.

	Patents	TM_D	Age_Ch	Age_BoD	Tot_BoD	Gender_BoD	STD_Diff_Age
Patents	1						
TM_D	0,107	1					
Age_Ch	0,005	0,031	1				
Age_BoD	0,031	-0,028	0,624	1			
Tot_BoD	0,100	0,023	0,097	-0,095	1		
Gender_BoD	-0,006	-0,002	-0,146	0,065	-0,213	1	
STD_Diff_Age	-0,039	0,061	0,639	-0,175	0,172	-0,219	1

Looking at the results obtained we can delete from the model the variables regarding the Equity and the STD_Diff_Age, we chose these two variables because they have high level of collinearity that can negatively influence the interpretability of the models.

3.4.1.3. Third iteration

The results of the three models are in Table 3.11 with the relevant statistics for the variables and the entire logistic regressions, 16 predictors, objective variable: Default.

Table 3.11: Statistics on the third run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.15	0.109	-	-	-3.41	5.1e-4(***)	-	-	-7.98	0.036(*)	-	-
Age_BoD					7.3e-3(**)	7.3e-3(**)	7.9e-6(***)	2.758	0.174	3.1e-3(**)	2.1e-5(***)	2.813
Age_Ch					-0.016	0.363	0.145	2.641	-0.126	0.314	0.161	2.700
Amount	0.013	0.509	0.989	1.444					6.6e-3	0.704	0.989	1.432
Assets	5.3e-10	0.833	2.24e-4(***)	1.486					9.9e-10	0.744	0.016(*)	1.431
ATECO		-	-	-					-	-	-	-
Covenants	0.061	0.812	0.687	1.082					0.209	0.415	0.601	1.081
Employees	3.5e-4	0.484	0.513	1.286					2.9e-4	0.479	0.575	1.346
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-0.311	0.552	0.869	1.455	-0.413	0.519	0.859	1.461
Gender_BoD					0.108	0.869	0.859	1.548	-0.021	0.976	0.869	1.535
Geo	-	-	-	-					-	-	-	-
Listed	-2.478	0.016(*)	1.2e-3(**)	1.024					-1.87	0.079(.)	1.1e-3(**)	1.074
Motivation	-	-	-	-					-	-	-	-
Patents					-0.150	0.047(*)	5.9e-3(**)	1.031	-0.021	0.47	1.7e-3(**)	1.112
Price	0.029	0.443	0.118	1.005					0.028	0.438	0.118	1.007
Rate	33.15	3.2e-7(***)	2.4e-10(***)	1.096					36.21	7.7e-7(***)	2.4e-10(***)	1.189
Region	-	-	-	-					-	-	-	-
Revenues	-2.3e-8	7.5e-4(***)	8.7e-5(***)	1.591					-1.7e-8	9.4e-3(**)	0,003(**)	1.686
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.077(.)	3.5e-3(**)	1.039	-0.056	0.218	0.016(*)	1.099
Tot_BoD					-0.301	9.9e-4(***)	4.5e-6	1.166	-0.201	0.016(*)	3.7e-3(**)	1.265
Warrants	-0.597	0.024(*)	0.384	1.133					-0.629	0.022(*)	0.384	1.150
N.Obs	969				969				969			
AIC	488.88				408.98				466.32			
BIC	537.64				447.991				549.22			
Fisher Scoring	8				6				10			

In the following two tables Table 3.12 and Table 3.13 are represented the correlation rate matrix for the first two models (the third model is of difficult representation due to the high number of variables).

Table 3.12: Correlation rate matrix for the first model, third run

	Amount	Price	Rate	Assets	Revenues	Employees
Amount	1					
Price	0,055	1				
Rate	0,144	0,020	1			
Assets	0,493	0,006	-0,080	1		
Revenues	0,228	-0,019	-0,070	0,643	1	
Employees	0,245	-0,027	-0,039	0,389	0,360	1

Table 3.13: Correlation rate matrix for the second model, third run.

	Patents	TM_D	Age_Ch	Age_BoD	Tot_BoD	Gender_BoD
Patents	1					
TM_D	0,107	1				
Age_Ch	0,005	0,031	1			
Age_BoD	0,031	-0,028	0,624	1		
Tot_BoD	0,100	0,023	0,097	-0,095	1	
Gender_BoD	-0,006	-0,002	-0,146	0,065	-0,213	1

Looking at the results obtained we can delete from the model the variables regarding the Amount and the Gender_Ch, the first variable has been deleted for its low level of significance, the second one has been removed in order to reduce the collinearity of the variables in the model.

3.4.1.4. Fourth iteration

The results of the three models are in Table 3.14 with the relevant statistics for the variables and the entire logistic regressions, 14 predictors, objective variable: Default.

Table 3.14: Statistics on the fourth run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.22	0.109	-	-	-3.46	4.4e-4(***)	-	-	-8.03	0.035(*)	-	-
Age_BoD					0.056	7.8e-3(**)	9.5e-6(***)	2.767	0.176	0.071(.)	1.9e-5(***)	2.813
Age_Ch					-0.016	0.360	0.144	2.655	-0.127	0.186	0.074(.)	2.700
Amount	-	-	-	-					-	-	-	-
Assets	1.1e-10	0.961	2.6e-4(***)	1.230					1.3e-9	0.582	6.8e-3(**)	1.431
ATECO	-	-	-	-					-	-	-	-
Covenants	0.069	0.788	0.580	1.079					0.213	0.421	0.580	1.081
Employees	3.9e-4	0.423	0.458	1.256					3e-4	0.486	0.617	1.346
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					0.1	0.855	0.856	1.081	-0.035	0.959	0.968	1.535
Geo	-	-	-	-					-	-	-	-
Listed	-2.45	0.017(*)	7.8e-4(***)	1.019					-1.85	0.081(.)	5.9e-4(***)	1.074
Motivation	-	-	-	-					-	-	-	-
Patents					-0.151	0.048(*)	5.9e-3(**)	1.026	-0.021	0.476	9.4e-4(***)	1.112
Price	0.030	0.439	0.119	1.005					0.029	0.435	0.119	1.007
Rate	34.30	4.6e-8(***)	6.9e-10(***)	1.018					36.92	1.2e-7(***)	6.9e-10(***)	1.189
Region	-	-	-	-					-	-	-	-
Revenues	-2.3e-8	5.8e-4(***)	6.9e-5(***)	1.532					-1.8e-8	8.9e-3(**)	8.3e-4(***)	1.686
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.036	-0.057	0.211	0.014(*)	1.099
Tot_BoD					-0.309	6.2e-5(***)	4.4e-6	1.138	-0.202	0.016(*)	3.1e-3(**)	1.265
Warrants	-0.554	0.029(*)	0.280	1.057					-0.610	0.023(*)	0.280	1.150
N.Obs	969				969				969			
AIC	487.29				407.32				462.91			
BIC	531.18				441.46				536.06			
Fisher Scoring	8				6				10			

In the following two tables Table 3.15 and Table 3.16 are represented the correlation rate matrix for the first two models (the third model is of difficult representation due to the high number of variables).

Table 3.15: Correlation rate matrix for the first model, fourth run.

	Price	Rate	Assets	Revenues	Employees
Price	1				
Rate	0,020	1			
Assets	0,006	-0,080	1		
Revenues	-0,019	-0,070	0,643	1	
Employees	-0,027	-0,039	0,389	0,360	1

Table 3.16: Correlation rate matrix for the second model, fourth run.

	Patents	TM_D	Age_Ch	Age_BoD	Tot_BoD	Gender_BoD
Patents	1					
TM_D	0,107	1				
Age_Ch	0,005	0,031	1			
Age_BoD	0,031	-0,028	0,624	1		
Tot_BoD	0,100	0,023	0,097	-0,095	1	
Gender_BoD	-0,006	-0,002	-0,146	0,065	-0,213	1

Looking at the results obtained we can delete from the model the variables regarding the Employees and the Age_Ch, the first variable has been deleted for its low level of significance, the second one has been removed in order to reduce the collinearity of the variables in the model.

3.4.1.5. Fifth iteration

The results of the three models are in Table 3.17 with the relevant statistics for the variables and the entire logistic regressions, 12 predictors, objective variable: Default.

Table 3.17: Statistics on the fifth run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.23	0.107	-	-	-3.58	2.7e-4(***)	-	-	-8.19	0.031(*)	-	-
Age_BoD					0.041	1.6e-3(**)	2e-5(***)	1.057	0.049	5.1e-4(***)	2.1e-5(***)	1.081
Age_Ch					-	-	-	-	-	-	-	-
Amount	-	-	-	-					-	-	-	-
Assets	1.8e-10	0.936	2.6e-4(***)	1.234					1.3e-9	0.563	6.8e-3(**)	1.242
ATECO	-	-	-	-					-	-	-	-
Covenants	0.075	0.769	0.580	1.082					0.218	0.409	0.580	1.082
Employees	-	-	-	-					-	-	-	-
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					-0.030	0.957	0.957	1.068	-0.200	0.724	0.726	1.074
Geo	-	-	-	-					-	-	-	-
Listed	-2.44	0.017(*)	7.8e-4(***)	1.019					-1.72	0.100	5.9e-4(***)	1.051
Motivation	-	-	-	-					-	-	-	-
Patents					-0.150	0.048(*)	5.9e-3(**)	1.024	-0.028	0.434	9.4e-4(***)	1.114
Price	0.030	0.440	0.119	1.005					0.028	0.448	0.119	1.006
Rate	34.61	3.1e-8(***)	6.9e-10(***)	1.014					36.99	1.1e-7(***)	6.4e-10(***)	1.071
Region	-	-	-	-					-	-	-	-
Revenues	-2.2e-8	5.4e-4(***)	6.9e-5(***)	1.289					-1.6e-8	6.2e-3(**)	8.3e-4(***)	1.327
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.034	-0.061	0.189	0.014(*)	1.085
Tot_BoD					-0.322	2.7e-5(***)	1e-6(***)	1.116	-0.25	0.003(**)	1,2e-3(**)	1.201
Warrants	-0.559	0.028(*)	0.280	1.057					-0.518	0.028(*)	0.280	1.076
N.Obs	969				969				969			
AIC	485				406.16				460.41			
BIC	524.85				435.42				523.80			
Fisher Scoring	8				6				10			

In the following three tables Table 3.18, Table 3.19, and Table 3.20 are represented the correlation matrix for the first two models (the third model is of difficult representation due to the high number of variables)

Table 3.18: : Correlation rate matrix for the first model, fifth run.

	Price	Rate	Assets	Revenues
Price	1			
Rate	0,020	1		
Assets	0,006	-0,080	1	
Revenues	-0,019	-0,070	0,643	1

Table 3.19: Correlation rate matrix for the second model, fifth run.

	Patents	TM_D	Age_BoD	Tot_BoD	Gender_BoD
Patents	1				
TM_D	0,107	1			
Age_BoD	0,031	-0,028	1		
Tot_BoD	0,100	0,023	-0,095	1	
Gender_BoD	-0,006	-0,002	0,065	-0,213	1

Table 3.20: Correlation rate matrix for the third model, fifth run.

	Price	Rate	Patents	TM_D	Assets	Revenues	Age_BoD	Tot_BoD	Gender_BoD
Price	1								
Rate	0,020	1							
Patents	0,012	-0,048	1						
TM_D	0,007	-0,026	0,107	1					
Assets	0,006	-0,080	0,232	0,072	1				
Revenues	-0,019	-0,070	0,142	0,050	0,643	1			
Age_BoD	0,010	0,026	0,031	-0,028	0,063	0,064	1		
Tot_BoD	0,042	-0,036	0,100	0,023	0,377	0,328	-0,095	1	
Gender_BoD	-0,027	0,057	-0,006	-0,002	-0,107	-0,022	0,065	-0,213	1

Looking at the results obtained we can delete from the model the variables regarding the Assets and the Gender_BoD, the first variable has been deleted in order to eliminate a part of the collinearity in the models, the second one has been deleted due to its low significative both for the Chi Square and the Z test.

3.4.1.6. Sixth iteration

The results of the three models are in Table 3.21 with the relevant statistics for the variables and the entire logistic regressions, 10 predictors, objective variable: Default.

Table 3.21: Statistics on the sixth run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.23	0.107	-	-	-3.603	1.9e-5(***)	-	-	-8.42	0.025(*)	-	-
Age_BoD					0.041	1.6e-3(**)	2.1e-5(***)	1.057	0.049	5.1e-4(***)	2.1e-5(***)	1.082
Age_Ch					-	-	-	-	-	-	-	-
Amount	-	-	-	-					-	-	-	-
Assets	-	-	-	-					-	-	-	-
ATECO	-	-	-	-					-	-	-	-
Covenants	0.075	0.769	0.580	1.082					0.259	0.409	0.580	1.078
Employees	-	-	-	-					-	-	-	-
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					-	-	-	-	-	-	-	-
Geo	-	-	-	-					-	-	-	-
Listed	-2.45	0.011(*)	7.8e-4(***)	1.005					-1.69	0.100	5.9e-4(***)	1.040
Motivation	-	-	-	-					-	-	-	-
Patents					-0.149	0.047(**)	5.9e-3(**)	1.024	-0.024	0.425	9.4e-4(***)	1.108
Price	0.030	0.439	0.119	1.005					0.028	0.445	0.119	1.006
Rate	34.66	2,7e-8(***)	6.9e-10(***)	1.007					37.18	7.3e-8(***)	6.9e-10(***)	1.057
Region	-	-	-	-					-	-	-	-
Revenues	-2.2e-8	1.7e-4(***)	6.5e-8(***)	1.007					-1.5e-8	6.1e-3(**)	1.7e-4(***)	1.134
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.034	-0.061	0.188	0.014(*)	1.082
Tot_BoD					-0.321	1.6e-5(***)	1e-6(***)	1.116	-0.236	3.4e-3(**)	1.5(**)	1.120
Warrants	-0.578	0.028(*)	0.280	1.052					-0.563	0.032(*)	0.280	1.062
N.Obs	969				969				969			
AIC	483.85				404.17				456.83			
BIC	517.98				428.56				510.47			
Fisher Scoring	8				6				10			

From now on we do not include the fourth type of research because the VIF test already let us confident about the absence of collinearity in the models, in order to be sure about the absence of collinearity we will continue in the assessment of the VIF test.

Looking at the results obtained we can delete from the model the variables regarding the Warrants and the Covenants, we chose to delete these two variable because are endogenous, the explanation will be discussed later.

3.4.1.7. Seventh iteration

The results of the three models are in Table 3.22 with the relevant statistics for the variables and the entire logistic regressions, 8 predictors, objective variable: Default.

Table 3.22: Statistics on the seventh run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-6.43	0.107	-	-	-3.603	1.9e-5(***)	-	-	-8.72	0.025(*)	-	-
Age_BoD					0.041	1.6e-3(**)	2.1e-5(***)	1.057	0.049	4.4e-4(***)	1.7e-5(***)	1.088
Age_Ch					-	-	-	-	-	-	-	-
Amount	-	-	-	-					-	-	-	-
Assets	-	-	-	-					-	-	-	-
ATECO	-	-	-	-					-	-	-	-
Covenants	-	-	-	-					-	-	-	-
Employees	-	-	-	-					-	-	-	-
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					-	-	-	-	-	-	-	-
Geo	-	-	-	-					-	-	-	-
Listed	-2.34	0.011(*)	1.3e-3(**)	1.002					-1.60	0.123	9.9e-4(***)	1.039
Motivation	-	-	-	-					-	-	-	-
Patents					-0.149	0.047(**)	5.9e-3(**)	1.024	-0.250	0.043(*)	8.8e-4(***)	1.122
Price	0.030	0.439	0.119	1.004					0.029	0.428	0.119	1.004
Rate	34.56	2,7e-8(***)	6.9e-10(***)	1.003					37.71	2.8e-8(***)	6.9e-10(***)	1.060
Region	-	-	-	-					-	-	-	-
Revenues	-2e-8	1.7e-4(***)	2,3e-8(***)	1.004					-1.3e-8	0.012(*)	7e-5(***)	1.096
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.034	-0.058	0.211	0.013(*)	1.091
Tot_BoD					-0.321	1.6e-5(***)	1e-6(***)	1.116	-0.227	0.004(**)	1,8e-3(**)	1.115
Warrants	-	-	-	-					-	-	-	-
N.Obs	969				969				969			
AIC	484.76				404.17				457.7			
BIC	509.14				428.56				501.58			
Fisher Scoring	8				6				10			

Looking at the results obtained we can delete from the model the variables regarding the Price and the TM_D from the third model.

3.4.1.8. Eighth iteration

The results of the three models are in Table 3.23 with the relevant statistics for the variables and the entire logistic regressions, 8 predictors, objective variable: Default.

Table 3.23: Statistics on the eighth run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-3.533	0(***)	-	-	-3.603	1.9e-5(***)	-	-	-5.91	0.025(*)	-	-
Age_BoD					0.041	1.6e-3(**)	2.1e-5(***)	1.057	0.048	4.4e-4(***)	1.8e-5(***)	1.084
Age_Ch					-	-	-	-	-	-	-	-
Amount	-	-	-	-					-	-	-	-
Assets	-	-	-	-					-	-	-	-
ATECO	-	-	-	-					-	-	-	-
Covenants	-	-	-	-					-	-	-	-
Employees	-	-	-	-					-	-	-	-
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					-	-	-	-	-	-	-	-
Geo	-	-	-	-					-	-	-	-
Listed	-2.345	0.022(*)	1.3e-3(**)	1.002					-1.67	0.123	7.8e-4(***)	1.032
Motivation	-	-	-	-					-	-	-	-
Patents					-0.149	0.047(**)	5.9e-3(**)	1.024	-0.384	0.043(*)	9.3e-4(***)	1.084
Price	-	-	-	-					-	-	-	-
Rate	34.725	1.2e-8(***)	5.5e-10(***)	1.003					38.24	2.8e-8(***)	5.5e-10(***)	1.067
Region	-	-	-	-					-	-	-	-
Revenues	-2.01e-8	2.2e-4(***)	1.6e-7(***)	1.002					-1.5e-8	0.012(*)	9e-6(***)	1.078
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.034	-	-	-	-
Tot_BoD					-0.321	1.6e-5(***)	1e-6(***)	1.116	-0.241	0.004(**)	1.9e-3(**)	1.117
Warrants	-	-	-	-					-	-	-	-
N.Obs	969				969				969			
AIC	483.89				404.17				457.84			
BIC	503.40				428.56				492.97			
Fisher Scoring	8				6				10			

Looking at the results obtained we can delete from the model the variable regarding the Listed from the third model, this variable does not result significant in this model looking at the tests conducted.

3.4.1.9. Nineth iteration

The results of the three models are in Table 3.24Table 3.23 with the relevant statistics for the variables and the entire logistic regressions, 8 predictors, objective variable: Default.

Table 3.24: Statistics on the nineth run.

Models	Model 1				Model 2				Model 3			
Test	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF	Z test		Chi Sq	VIF
Statistic	Beta	p-value	p-value		Beta	p-value	p-value		Beta	p-value	p-value	
Intercept	-3.533	0(***)	-	-	-3.603	1.9e-5(***)	-	-	-5.914	0(***)	-	-
Age_BoD					0.041	1.6e-3(**)	2.1e-5(***)	1.057	0.0495	3.9e-4(***)	3.8e-6(***)	1.089
Age_Ch					-	-	-	-	-	-	-	-
Amount	-	-	-	-					-	-	-	-
Assets	-	-	-	-					-	-	-	-
ATECO	-	-	-	-					-	-	-	-
Covenants	-	-	-	-					-	-	-	-
Employees	-	-	-	-					-	-	-	-
Equity	-	-	-	-					-	-	-	-
Gender_Ch					-	-	-	-	-	-	-	-
Gender_BoD					-	-	-	-	-	-	-	-
Geo	-	-	-	-					-	-	-	-
Listed	-2.345	0.022(*)	1.3e-3(**)	1.002					-	-	-	-
Motivation	-	-	-	-					-	-	-	-
Patents					-0.149	0.047(**)	5.9e-3(**)	1.024	-0.335	0.049(*)	9.3e-4(***)	1.075
Price	-	-	-	-					-	-	-	-
Rate	34.725	1.2e-8(***)	5.5e-10(***)	1.003					37.63	2.2e-8(***)	5.5e-10(***)	1.061
Region	-	-	-	-					-	-	-	-
Revenues	-2.01e-8	2.2e-4(***)	1.6e-7(***)	1.002					-1.46e-8	6.6e-3(**)	1.4e-6(***)	1.077
STD_Diff_Age					-	-	-	-	-	-	-	-
TM_D					-0.085	0.078(.)	3.5e-3(**)	1.034	-	-	-	-
Tot_BoD					-0.321	1.6e-5(***)	1e-6(***)	1.116	-0.264	5.4e-4(***)	1.4e-3(**)	1.091
Warrants	-	-	-	-					-	-	-	-
N.Obs	969				969				969			
AIC	483.89				404.17				460.12			
BIC	503.40				428.56				489.38			
Fisher Scoring	8				6				10			

With this last iteration we conclude the definition of the models. The expressions obtained are in the equation ( ), ( ), ( ). We can temporarily affirm that the second model is the best in terms of prediction from all the tests used, has the lowest AIC, BIC and reach in the best likelihood in the lowest number of iterations.

3.5. Logistic Regression Models

Logistic regression with the first set of predictors (3.2):

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} \tag{3.2}$$

- Y is the objective variable, in particular the variable is the Default probability. It will be classified depending on the threshold that will be defined in the following section.
- β_0 is the intercept of the equation, its value is -3.533.
- β_1 is the coefficient associated with the Listed variable, its value is -2.345.
- β_2 is the coefficient associated with the Rate variable, its value is 34.725.
- β_3 is the coefficient associated with the Revenues variable, its value is -0.0201. The value of the Revenues must be expressed in millions of Euros
- X_1 is the value of the variable Listed of the observation k.
- X_2 is the value of the variable Rate of the observation k.
- X_3 is the value of the variable Revenues of the observation k.

Logistic regression with the second set of predictors ( ):

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} + \beta_4 \times X_{4k} \tag{3.3}$$

- Y is the objective variable, in particular the variable is the Default probability. It will be classified depending on the threshold that will be defined in the following section.

- β_0 is the intercept of the equation, its value is -3.603.
- β_1 is the coefficient associated with the Age_BoD variable, its value is 0.0408.
- β_2 is the coefficient associated with the Patents variable, its value is -0.149.
- β_3 is the coefficient associated with the TM_D variable, its value is -0.0847.
- β_4 is the coefficient associated with the Tot_BoD variable, its value is -0.321.
- X_1 is the value of the variable Age_BoD of the observation k.
- X_2 is the value of the variable Patents of the observation k.
- X_3 is the value of the variable TM_D of the observation k.
- X_4 is the value of the variable Tot_BoD of the observation k.

Logistic regression with the third set of predictors ( )

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} + \beta_4 \times X_{4k} + \beta_5 \times X_{5k} \tag{3.4}$$

- Y is the objective variable, in particular the variable is the Default probability. It will be classified depending on the threshold that will be defined in the following section.
- β_0 is the intercept of the equation, its value is -5.914.
- β_1 is the coefficient associated with the Age_BoD variable, its value is 0.0495.
- β_2 is the coefficient associated with the Patents variable, its value is -0.335.
- β_3 is the coefficient associated with the Rate variable, its value is 37.63.
- β_4 is the coefficient associated with the Revenues variable, its value is -0.0146. The value of the Revenues must be expressed in millions of Euros.
- β_5 is the coefficient associated with the Tot_BoD variable, its value is -0.264.
- X_1 is the value of the variable Age_BoD of the observation k.
- X_2 is the value of the variable Patents of the observation k.

- X_3 is the value of the variable Rate of the observation k.
- X_4 is the value of the variable Revenues of the observation k.
- X_5 is the value of the variable Tot_BoD of the observation k.

3.6. Analysis after the building of the models

3.6.1. Analysis on the performance of the models

The next step in the analysis is the assessment of the best probability that will classify each observation in a dichotomic attribute with only two possible values: "NO" for the negative observation "YES" for the positive ones. With the term "Negative" we refer to a firm that will not be predicted as a defaulted one; the term "Positive" refers to a firm that will be predicted as a company that will be predicted as a defaulted firm. All the observations that result in a probability lower than a determined threshold (that could be arbitrary set) will be classified as positives the other as negative firms; we expect that an increase in the threshold level will lead to an increase in the observation classified as default. In order to understand the performance of each set of variable in the classification we built three graphs in Figure , Figure , and Figure with the ongoing of two indicators described in (3.5) and (3.6). It is important to note that the graphs in Figure and Figure are defined in thresholds p in [0;0.6] while the graphs in Figure is defined in thresholds p in [0;0.4], the reason behind is that over the above mentioned limits the True Positive Rate became equal to 1 while the True Negative Rate became equal to 0 and both constant.

$$TPR = \frac{TP}{TP + FN} \tag{3.5}$$

$$TNR = \frac{TN}{TN + FP} \tag{3.6}$$

- TPR is the True Positive Rate or sensitivity.
- TNR is the True Negative Rate or specificity.
- TP is the total amount of True Positives: with True Positive we refer to an observation that is positive and that is correctly predicted as positive.
- FP is the total amount of False Positives: with False Positive we refer to an observation that is negative but is wrongly predicted as positive.

- TN is the total amount of True Negatives: with True Negative we refer to an observation that is negative and that is correctly predicted as negative.
- FN is the total amount of False Negatives: with False Negative we refer to an observation that is positive but is wrongly predicted as negative.

The graphs in Figure , Figure , and Figure are the representations of the performances in the three different sets of variables.

From the graph in Figure we can see that True Positive Rate increase rapidly reaching an elbow before 0.2 as threshold, the True Negative Rate sharply decrease; the two lines meet in the point (0.11;0.24).

In the second graph in Figure we can see a less sharp ongoing both for the true negative rate and the true positive rate if compared with the first one; the two lines meet in the point (0.095;0.28).

The last graph in Figure represents the performance of the model built on the last set of variables, we can see a sharp decrease in the true negative rate, the increase of the true negative rate Is more linear; the two lines meet in the point (0.105;0.23).

Figure 3.5: Graph of the TPR-TNR for the first set of variables.

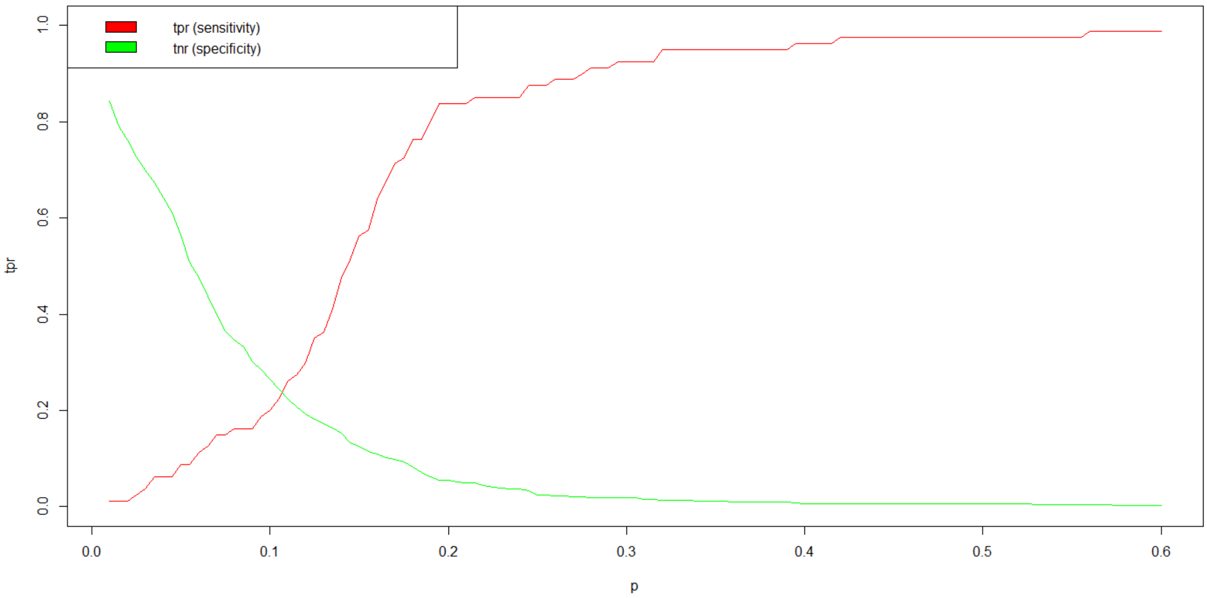


Figure 3.6: Graph of the TPR-TNR for the second set of variables.

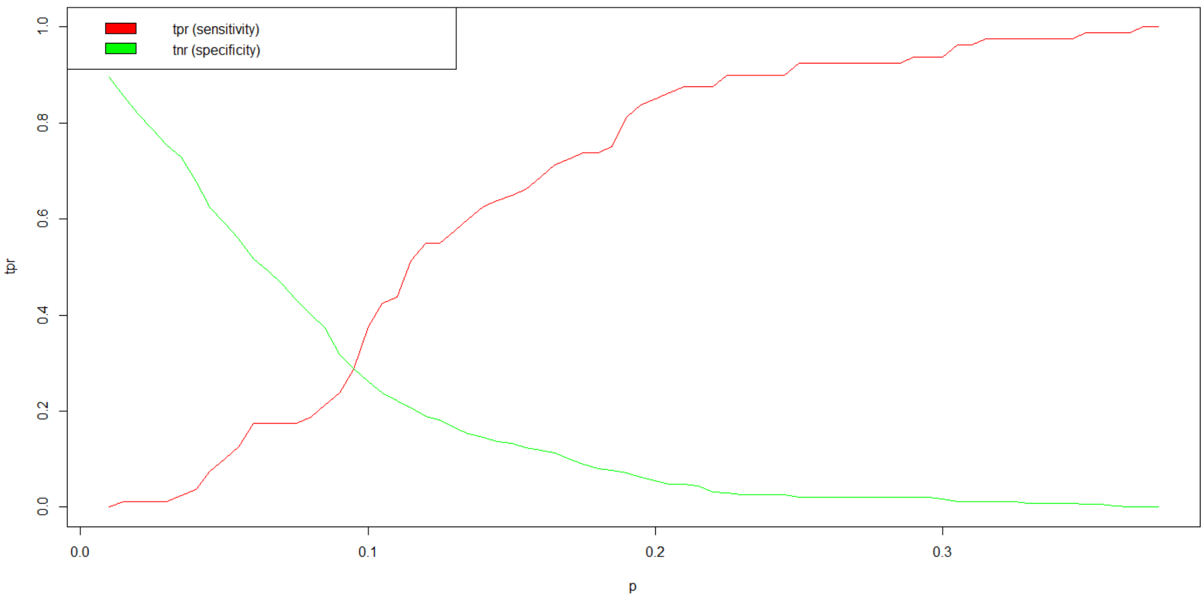
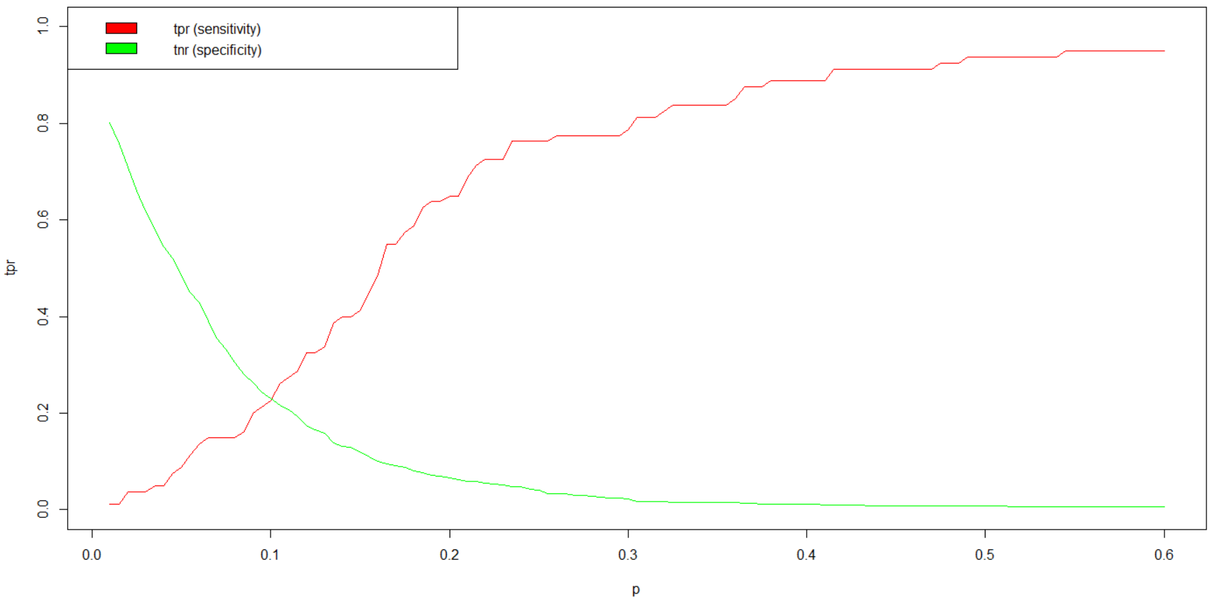
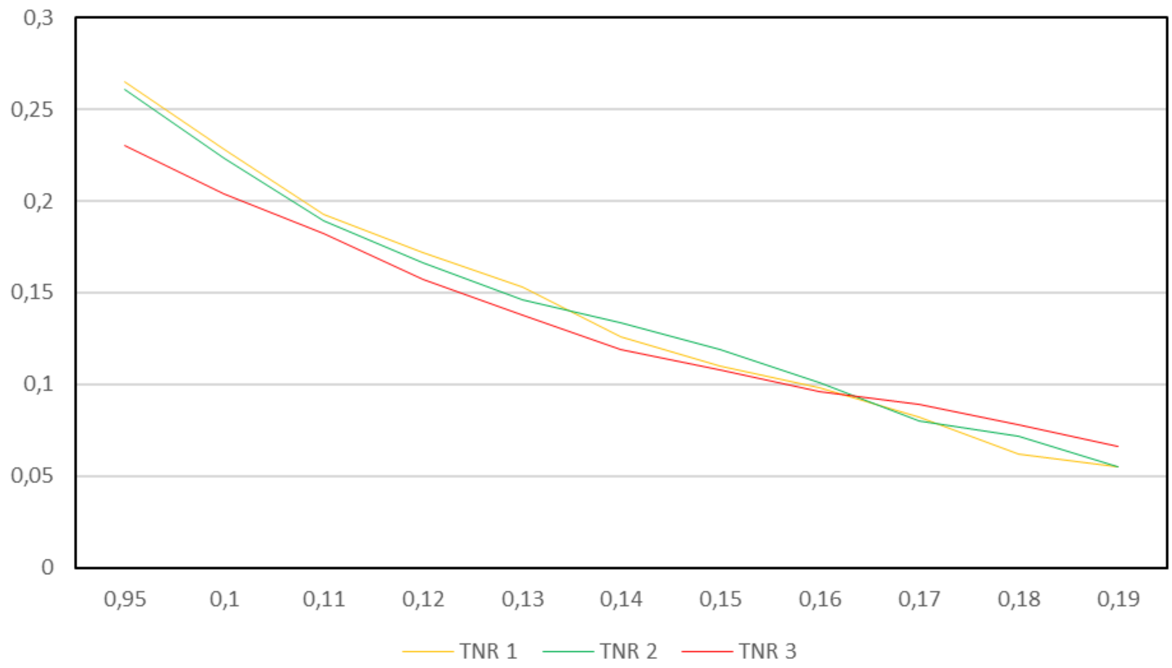


Figure 3.7: Graph of the TPR-TNR for the third set of variables.



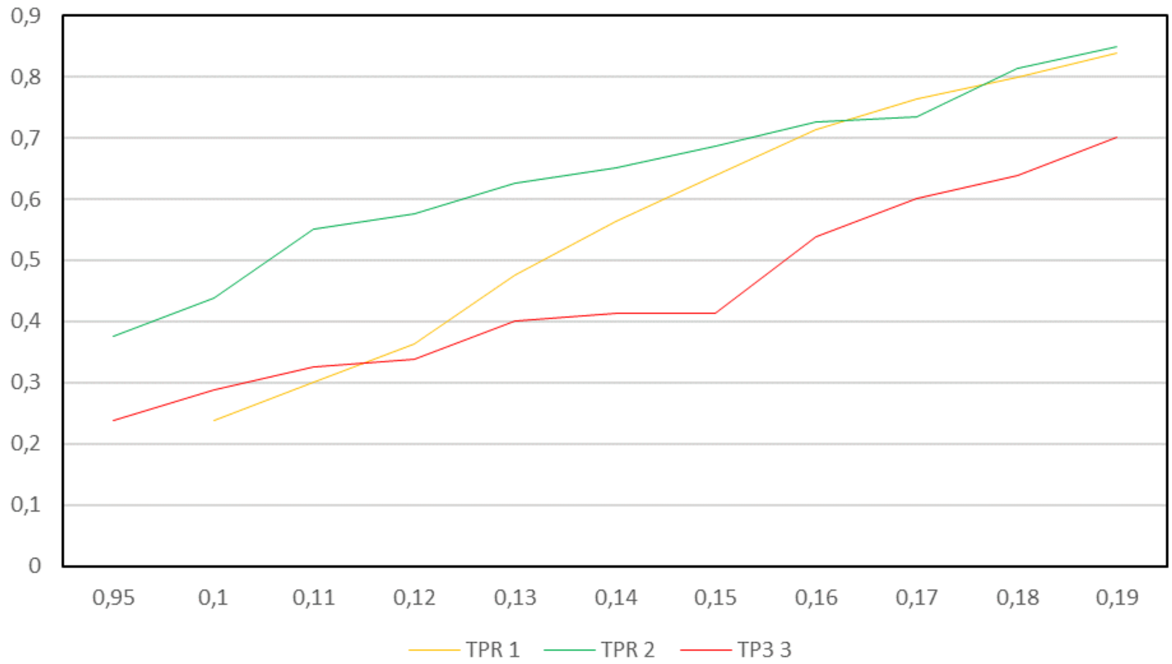
Now the objective is a first comparison of the two model. Our idea in order to solve this issue is the graphical comparison of the prediction ability combined with the numeric results on the two rates, we create two graphs in the space of threshold between 0.095 and 0.20 that will have on the x-axis the threshold as in the previous three graphs and on the y-axis in Figure the True Negatives Rate and in Figure the True Positive Rate comparing directly the three models.

Figure 3.8: Graph that compares all the tree models, TNR.



The first graph in Figure compares the three different models detailing the True Negatives Rates. Looking at the graph, there is not a clear preference between a model compared to the others.

Figure 3.9: Graph that compares all the tree models, TPR.



The graph in Figure compares the three True Positive Rate. In this case we can see a completely different situation: looking at the initial part of the graph (for lower thresholds) the second model performs much better than the model in which are used the conventional variables. Increasing the threshold, the first model reaches performance quite similar to that of the one built with the second set of variables.

This last graph shows that the best fitting model is the one composed by the non-conventional variables, confirming the first analysis results.

In order to be clearer we built three tables (Table 3.25, Table 3.26, and Table 3.27) that represent numerically the graph in Figure and Figure .

The tables below are composed as follow: the first column is the threshold used for the analysis, then we have the True Positives Rate, the True Negatives Rate, and their percentage variation from a threshold to the successive.

Table 3.25: Evolution of TPR-TNR of the first set of variables to vary of threshold.

Threshold	TPR	TNR	ΔTPR%	ΔTNR%
0.095	0.188	0.286	-	-
0.10	0.205	0.265	9.043	-7.343
0.11	0.238	0.228	16.098	-13.962
0.12	0.300	0.193	26.050	-15.351
0.13	0.363	0.172	21.000	-10.881
0.14	0.475	0.153	30.854	-11.047
0.15	0.563	0.126	18.526	-17.647
0.16	0.638	0.110	13.321	-12.698
0.17	0.713	0.098	11.755	-10.909
0.18	0.763	0.082	7.013	-16.327
0.19	0.800	0.062	4.849	-24.390
0.20	0.812	0.055	4.750	-11.290

Table 3.26: Evolution of TPR-TNR of the second set of variables to vary of threshold.

Threshold	TPR	TNR	ΔTPR%	ΔTNR%
0.095	0.300	0.287	-	-
0.10	0.375	0.261	25.000	-9.059
0.11	0.438	0.223	16.800	-14.559
0.12	0.550	0.189	25.571	-15.247
0.13	0.575	0.166	4.545	-12.169
0.14	0.625	0.146	8.696	-12.048
0.15	0.650	0.134	4.000	-8.288
0.16	0.687	0.119	5.692	-11.128
0.17	0.725	0.101	5.531	-15.126
0.18	0.735	0.080	1.379	-20.792
0.19	0.813	0.072	10.612	-10.000
0.20	0.850	0.055	4.551	-23.611

Table 3.27: Evolution of TPR-TNR of the third set of variables to vary of threshold.

Threshold	TPR	TNR	ΔTPR%	ΔTNR%
0.095	0.188	0.232	-	-
0.10	0.238	0.230	26.596	0.862
0.11	0.288	0.204	21.008	-11.304
0.12	0.325	0.182	12.847	-10.784
0.13	0.338	0.157	4.000	-13.768
0.14	0.400	0.138	18.343	-12.102
0.15	0.413	0.119	3.250	-13.768
0.16	0.413	0.108	0	-9.244
0.17	0.538	0.096	30.266	-11.111
0.18	0.600	0.089	11.524	-7.292
0.19	0.638	0.078	6.333	-12.360
0.20	0.700	0.066	9.718	-15.385

The final stage of our analysis will answer the question on how to choose this threshold and therefore which is the best threshold between the True Positive Rate and the True Negative Rate.

3.6.2. Analysis for the choice of the threshold

This phase of analysis focuses on the assessment of the best probability that will be used as threshold that classify the observations as positives or negatives.

The threshold is usually arbitrary, and when there is not a clear preference for the rate of true negatives or true positives, the best instrument is the ROC Curve. In this type of analysis, the threshold is chosen looking at the elbow of the curve, it will optimize the threshold considering of the same importance the increasing of the True Positives Rate and the True Negatives Rate. This is not the case that we have analysed since the cost associated to an error for a false positive is quite different if compared with a false negative. The objective now is to develop a function that will allows us to connect the cost of a misclassified element and the total misclassified by the model.

In order to solve this issue, it is necessary understanding the cost of a false positive or a false negative in order to choose a threshold and subsequently analysing the performances of the models. We will discuss the costs from the lender perspective, that from now on it is identified as a single bank that can grant all the bonds presents in the market; this assumption is necessary in order to compute the model and give us a simplification in not considering all the competitive mechanisms.

The cost of a true positive will be zero since the bank will not lend the money to a firm that then effectively results in a default status.

$$c_{TP} = 0$$

The cost of each false positive is the cost of debt sustained by the firm that issue a minibond, namely, the interest rate k_d multiplied for the amount that would be lent C.

$$c_{FP} = k_d \times C$$

The cost of each true negative is 0 since there are no additional costs.

$$c_{TN} = 0$$

The cost of a false negative is much higher than the false positive. In fact, if the bank grants a bond to a risky company and this last fails, the bank will lose all the amount lent.

$$c_{FN} = C$$

The revenues associated to a true positive can be considered as a saving for not landing money that a defaulted firm would not repay.

$$r_{TP} = C$$

The revenues associated to a false positive are zero since the bank will not receive any amount of money from the company because the bank choose not to grant the debt to a firm that could have repaid the debt instead.

$$r_{FP} = 0$$

The revenues associated to a true negative are the interest rate applied to the bank k_d multiplied for the amount C.

$$r_{TN} = k_d \times C$$

The revenues associated to a false negative will be zero since the firm will not be able to repay the debt.

$$r_{FN} = 0$$

Defined these coefficients, we can build a model of linear optimization, with the objective function as the bank profits, pointing at the maximization of function (3.7). Profits are defined as the difference between revenues and costs, as we can see in equation (3.7).

$$max. \pi = TP \times \ (r_{TP} - \ c_{TP}) + FP \times (r_{FP} - c_{FP}) + TN \times (r_{TN} - c_{TN}) + FN \times (r_{NF} - c_{FN}) \quad (3.7)$$

This equation is subject to the following bounds: (3.8) and (3.9).

$$TP + FP + TN + FN = tot\ observaations \quad (3.8)$$

$$TP, FP, TN, TP = f(p) \quad (3.9)$$

The equation (3.9) is the result of the part of the algorithm in the misclassification table, the misclassification table is the 2X2 matrix that combines the predictions made by the classification model with the real classification, this table is based on the Logistic Regression models built during the analysis.

We now compute a list as a legend of this optimization:

- c_{TP}: is the unitary cost of a True Positive.
- c_{FP}: is the unitary cost of a False Positive.
- c_{TP}: is the unitary cost of a True Negative.
- c_{FN}: is the unitary cost of a False Negative.
- r_{TP}: is the unitary revenues from a True Positive.
- r_{FP}: is the unitary revenues from a False Positive.
- r_{TN}: is the unitary revenues from a True Negative.

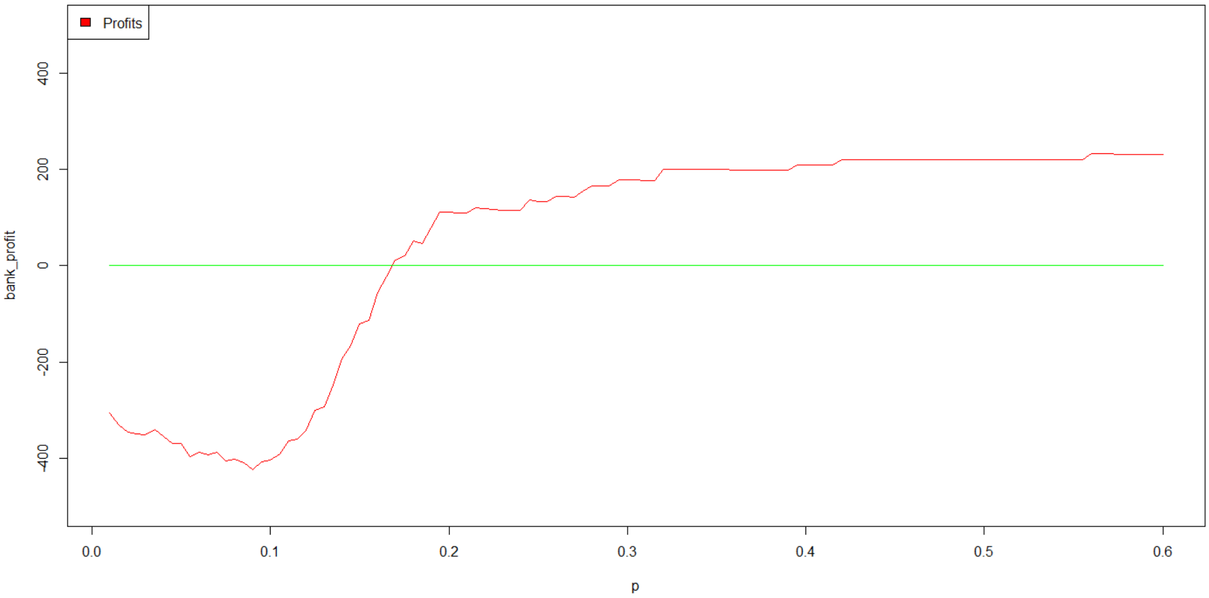
- r_{FN}: is the unitary revenues from a False Negative.
- k_d: is the interest rate applied to the debt.
- C: is the total amount of a debt.
- TP: is the total amount of True Positives.
- FP: is the total amount of False Positives.
- TN: is the total amount of True Negatives.
- FN: is the total amount of False Negatives.
- π: are the profits of the banks.

We can now define which is the best threshold to be used for each set of variable. It is important to clarify that we have already assessed the performance of the three model, but we will explain the procedure for all the three of them.

We analysed the ongoing of the profit function to vary of the threshold. In order to reach this objective, we build a graph that have on the x-axis the threshold and on the y-axis the profit level. Before describing the profits in each model, it is important to explain the value used for the estimation of k_d and C. k_d has been defined for the model as the mean of the annual interest rate of all the observations. The same has been done for C that is the average value of the total amount required.

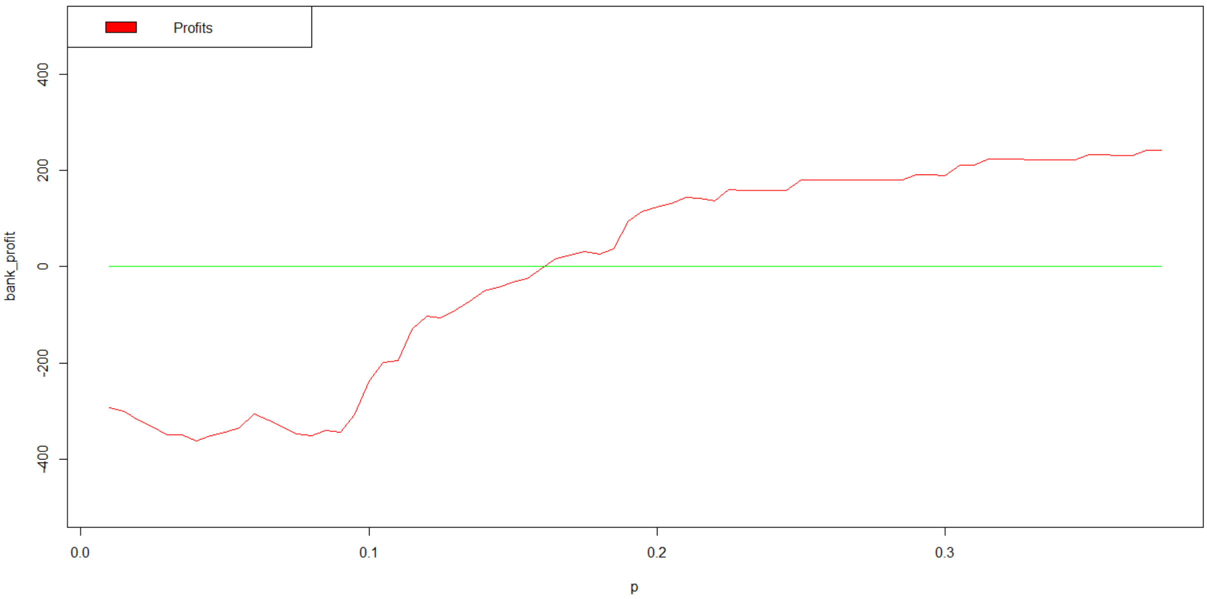
We will choose as optimum threshold a point of elbow since we are uncertain about the accuracy of the optimization function created in the analysis due to its roughness (in paragraph 5.2 we will discuss about possible developments and improvements of this function).

Figure 3.10: Profit function for the model obtained through the first set of variables.



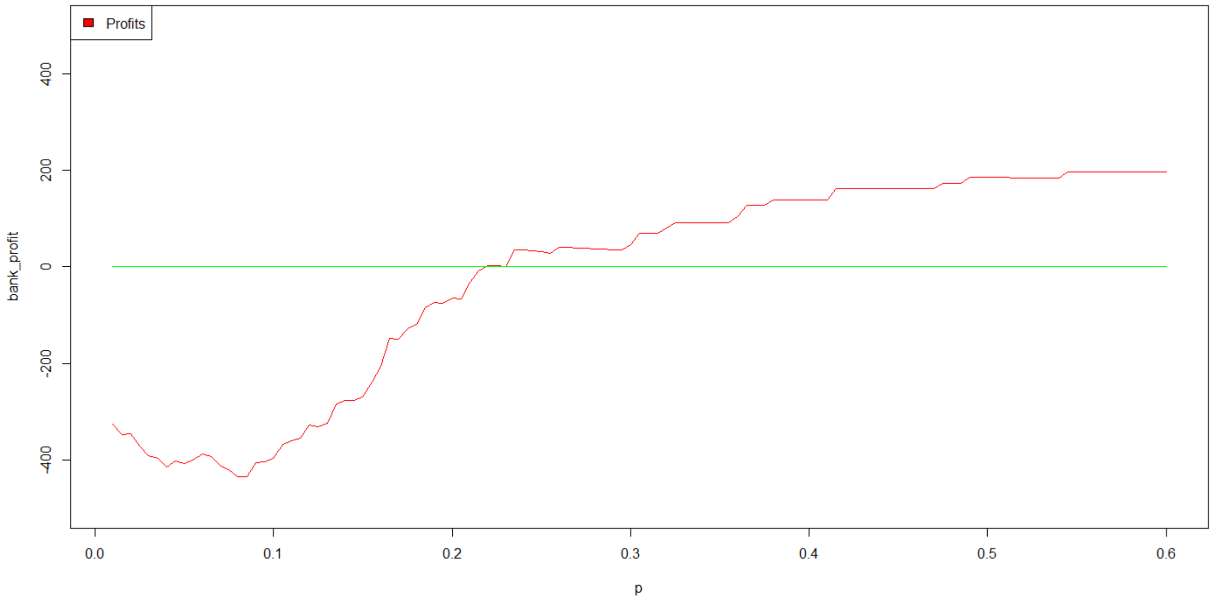
The graph in Figure shows the profit function of the model obtained through the first set of variables. The output corresponds to an optimal threshold of 0.19.

Figure 3.11: Profit function for the model obtained through the second set of variables.



For the graph in Figure that shows the profit function realized with the second set of variables, the distinction is not clear as in the first set of variables, however we can see small step in the positive region in 0.18

Figure 3.12: Profit function for the model obtained through the third set of variables.



The last graph in Figure shows the profit function realized with the third set of variables. This last graph highlights again the low performances of this third model, in fact the profits reach a step in the positive region of space only in 0.24.

Once the optimal threshold has been identified it could be possible to predict the output for new observations. In particular, having a new firm in analysis, it would be sufficient to solve the equation with the coefficients already found. If the value of Y (legend at page 83) is higher than the optimal threshold, it is more convenient for the bank not to grant the loan, otherwise to prefer lending money to the firm.

3.7. Sum up of the results

In this section we show a sum up of the results of all the analysis conducted. First of all, we report here the three regression models.

Logistic regression with the first set of predictors (3.2)

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} \tag{3.2}$$

- Y is the objective variable, in particular the variable is the Default probability. It will be classified depending on the threshold that will be defined in the following section.
- β_0 is the intercept of the equation, its value is -3.533.
- β_1 is the coefficient associated with the Listed variable, its value is -2.345.
- β_2 is the coefficient associated with the Rate variable, its value is 34.725.
- β_3 is the coefficient associated with the Revenues variable, its value is -0.0201. The value of the Revenues must be expressed in millions of Euros.
- X_1 is the value of the variable Listed of the observation k.
- X_2 is the value of the variable Rate of the observation k.
- X_3 is the value of the variable Revenues of the observation k.

Logistic regression with the second set of predictors ( ):

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} + \beta_4 \times X_{4k} \tag{3.3}$$

- Y is the objective variable, namely the Default probability. It will be classified depending on the threshold that will be defined in the following section.
- β_0 is the intercept of the equation, its value is -3.603.
- β_1 is the coefficient associated with the Age_BoD variable, its value is 0.0408.
- β_2 is the coefficient associated with the Patents variable, its value is -0.149.
- β_3 is the coefficient associated with the TM_D variable, its value is -0.0847.
- β_4 is the coefficient associated with the Tot_BoD variable, its value is -0.321.
- X_1 is the value of the variable Age_BoD of the observation k.
- X_2 is the value of the variable Patents of the observation k.
- X_3 is the value of the variable TM_D of the observation k.
- X_4 is the value of the variable Tot_BoD of the observation k.

Logistic regression with the third set of predictors ( )

$$Log\left[\frac{Y_k}{(1-Y_k)}\right] = \beta_0 + \beta_1 \times X_{1k} + \beta_2 \times X_{2k} + \beta_3 \times X_{3k} + \beta_4 \times X_{4k} + \beta_5 \times X_{5k} \tag{3.4}$$

- Y is the objective variable, namely the Default probability. It will be classified depending on the threshold that will be defined in the following section.
- β_0 is the intercept of the equation, its value is -5.914.
- β_1 is the coefficient associated with the Age_BoD variable, its value is 0.0495.
- β_2 is the coefficient associated with the Patents variable, its value is -0.335.
- β_3 is the coefficient associated with the Rate variable, its value is 37.63.
- β_4 is the coefficient associated with the Revenues variable, its value is -0.0146. The value of the Revenues must be expressed in millions of Euros.
- β_5 is the coefficient associated with the Tot_BoD variable, its value is -0.264.
- X_1 is the value of the variable Age_BoD of the observation k.

- X_2 is the value of the variable Patents of the observation k.
- X_3 is the value of the variable Rate of the observation k.
- X_4 is the value of the variable Revenues of the observation k.
- X_5 is the value of the variable Tot_BoD of the observation k.

The models perform in different ways:

- For what regards the True Negatives Rate all the three models perform similarly as the graph in Figure shows.
- For what regards the True Positives Rate, the third model is the worst in classifying the observations. The first model with thresholds lower than 0.12 performs badly, and could be compared with the third model, for the higher thresholds (more than 0.16) it reaches the performance of the second model. The second model results strongly the best with lower thresholds, for higher ones it performs in a similar way of the first model.

The last step is the assessment of the best threshold, in details:

- The first model with only conventional variables, best threshold is 0.19.
- The second model with only non-conventional variables, best threshold is 0.18.
- The third model with both conventional and non-conventional variables, best threshold is 0.24.

3.8. Robustness test

In order to have confirmation on the results obtained we conducted another test: we built a classification tree for each model. The objective is to understand if the variables used are the same of the regression models or are different and if their sense is the same. Initially we will check the variable used by the unpruned tree with the sensitivity related. Then the tree has been pruned through the respective function and are performed the same tests as before. The pruned trees are shown in Figure , Figure , and Figure .

3.8.1.1. Model with the first set of variables

In this model not all the variables are used. It counts only the following variables (automatically chosen by the algorithm as the more significative):

- Revenues (in Euros).
- Interest Rate.
- Par Value.
- Number of Employees.
- Listed Company.
- Assets (in Euros).
- Warrants.

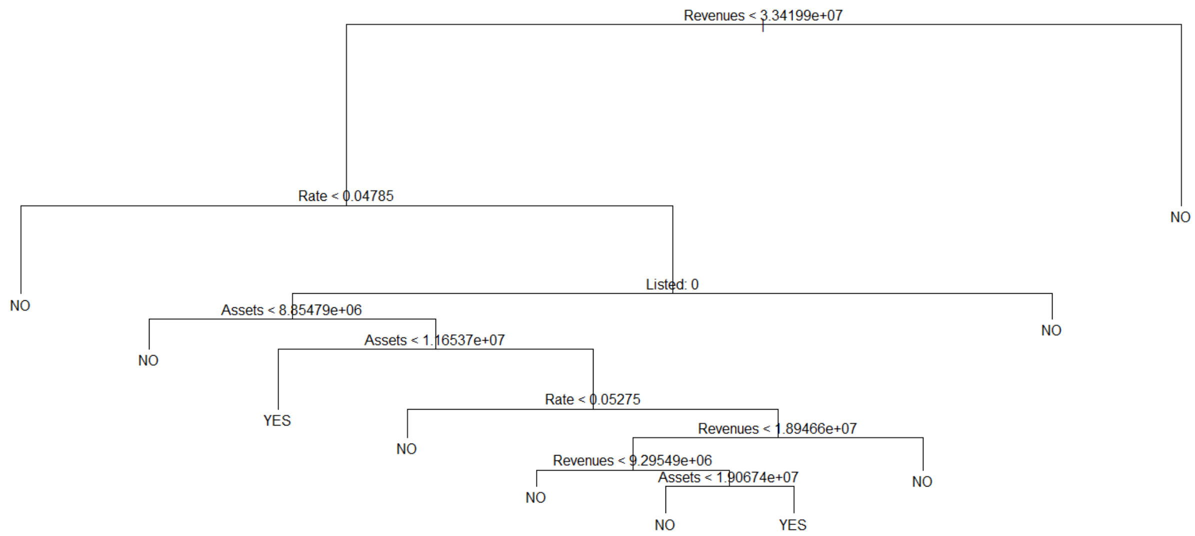
The True Positives Rate of the tree is 0.219, the pruned tree, in Figure , is composed by the following variables:

- Revenues (in Euros).
- Interest Rate.
- Listed Company.

- Assets (in Euros).

The True Positives Rate decreases to 0.125.

Figure 3.13: Pruned tree obtained from the first set of variables.



This first tree is quite difficult to be understood even if it has already been pruned. We see that the first node is represented by the revenues, companies with high revenues are predicted as no defaulted ones. The second important node is composed of the annual interest rate, companies with low interest rates are less likely to default. The third node is the dichotomic variable whether the listed company are flagged as no default; it could be possible to go in deep with the analysis of the nodes, but the remaining ones are less significant.

The comparison between the two models highlights an alignment for the variables and their interpretation. The first one is represented by the revenues and confirm the evidence from the logistic regression: in the classification tree, all the firms with high level of revenues are classified as no default ones, the same is for the annual interest rate, a high cost of the debt tends to increase the default probability for the firms, the third confirm derives from the listing of a company, a listed company is less likely to defaults and in this second model all the listed companies are classified as no default even in presence of low revenues. The tree model comprises also another variable: the value of the assets.

We can sustain that firms with high assets are considered safer, they can use their long-term assets as collateral in issuing new debts or it is possible to rely on short-term assets to absorb insolvency and liquidity risks that could happen in hard times, in the logistic regression model built we deleted this variable for the high collinearity with the other variables like the revenues. The classification tree analyses each variable individually and not as a combination of more than one, therefore is completely reasonable that the level of assets is a significative variable, even if less than the revenues.

3.8.1.2. Model with the second set of variables

In this model not all the variables are used. It counts only the following variables:

- Number of the members of the BoD.
- Average Age of the BoD.
- Difference of age between the Chairman and the BoD.
- Patents.
- Age of the Chairman.
- Gender of the members of the BoD.
- Trademarks and Designs.

The True Positives Rate of the tree is 0.219, the pruned tree, in Figure , is composed by the following variables:

- Number of the members of the BoD.
- Average age of the BoD.
- Gender of the members of the BoD.

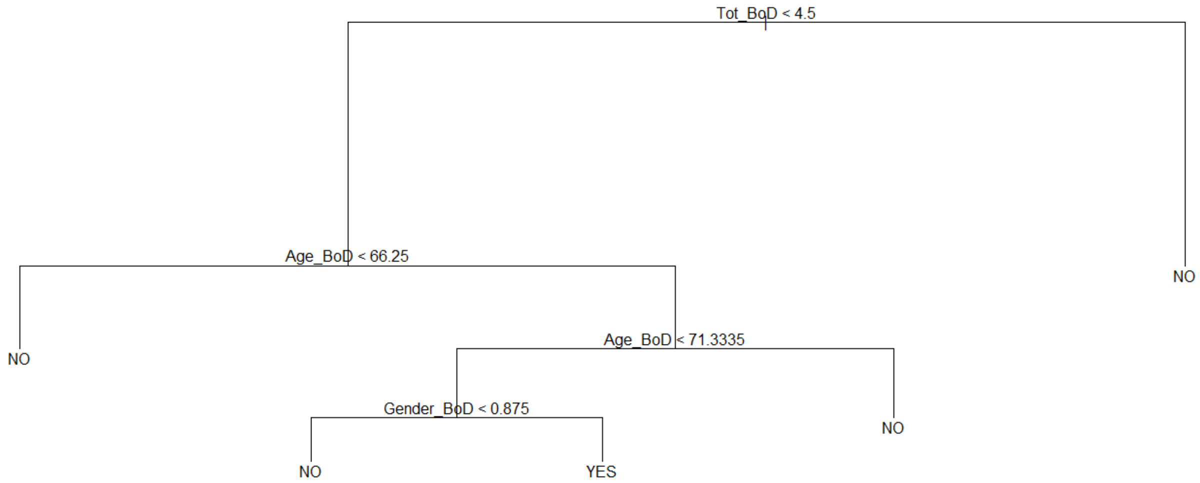
The True Positives Rate decreases to 0.188.

This second tree is clearer than the previous one. We can note as the first node is the total number of the members in the various boards, in particular larger boards (five people or more) are less vulnerable to default. On the second node we have the average age of the Board of Directors, in this case an older Board of Directors is more likely to default than a younger one. The third node instead rewards the older Board of Directors of the remaining ones, this node does not give great information, it is due to the composition of the data and the not great balance between the firm in default and the ones that are not. The last node splits the observations according to the percentage of male members inside the board, this could be a tricky point because, according to the model, the firms with a high percentage of male tent to default, however, more than the 80% of the total people present in the various boards are male, therefore this last point could be biased.

The model does not analyse the intellectual properties like the regression, for this aspect we do not have a clear double check, the reason behind the poor force of these variable can be understood thanks to Table 3.1. In fact, more than the 50% of the observations does not have any intellectual property, which makes difficult to understand if the absence of intellectual property could be a predictor of a defaulted firm.

We focus on the composition of the board of directors: a high number of members means a low default probability for both the models, the same for the boards age.

Figure 3.14: Pruned tree obtained from the second set of variables.



3.8.1.3. Model with the third set of variables

In this model not all the variables are used. It counts only the following variables:

- Revenues (in Euros)
- Number of the members of the BoD
- Interest Rate
- Gender of the members of the BoD
- Patents
- Assets (in Euros)
- Average age of the members of the BoD
- Age of the Chairman
- Difference of age between the Chairman and the BoD
- Trademarks and Designs
- Covenants
- Employees

The True Positives Rate of the tree is 0.313, the pruned tree in Figure , is composed by the following variables:

- Revenues (in Euros)
- Number of the members of the BoD
- Interest Rate
- Average age of the members of the BoD
- Employees
- Trademarks and Designs
- Age of the Chairman

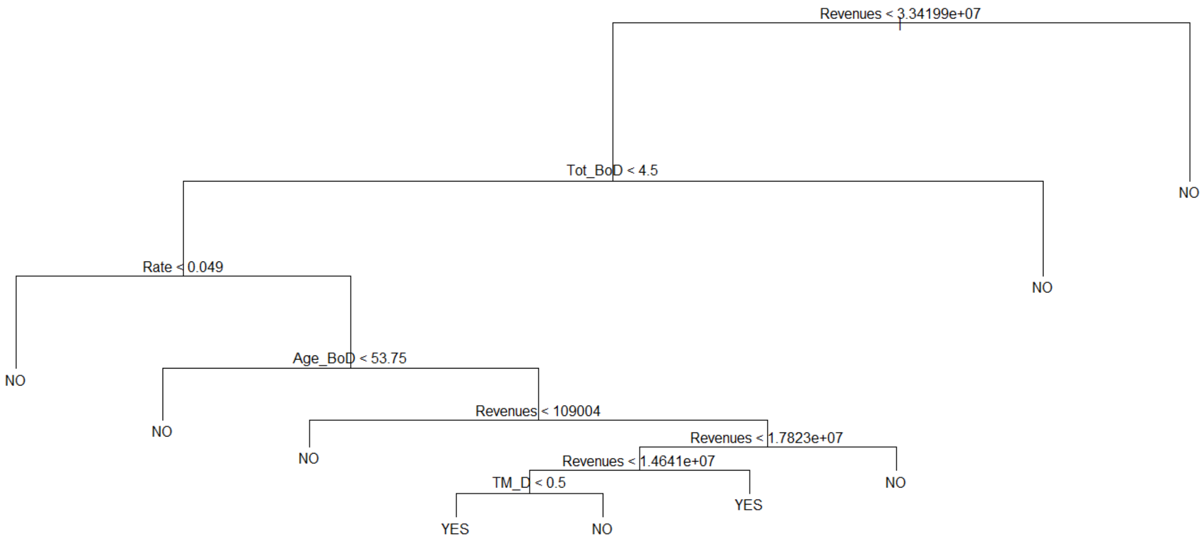
The True Positives Rate decreases to 0.281.

This last model is, like the first one, of hard readability for the number of nodes that composes it, however, focusing on the first nodes we understand how the two previous model coexist; the most important node is the amount of revenues and, like in the first model, we have negative correlation between revenues and default. As in the previous model we have the second node that connects the default and the dimension of the Board of Directors, the more are the component of the board, less is the default probability; the other nodes became less important.

We compare the structure of the tree and the logistic regression: the tree helps us in being more confident in all the aspects. The regression model and the tree collimate in the influence of the amount of revenues, in the linear regression we have a negative correlation and this is corroborated by the first node for which high level of revenues means no default; the second parameter regards the total number of directors and, as the coefficient is negative, the classification tree exclude from the defaulted firms those that have more than four people in the board; the third aspect agreeing with the logistic regression model is the interest rate, low interest rate results in low default probability.

The last aspect is the age of the members of the board: young boards reflect a low probability of default.

Figure 3.15: Pruned tree obtained from the third set of variables.



4 Comments on the results

This section of our dissertation focuses on the variables on which our model has been built on. We will comment all the variables that have some peculiarities.

4.1. Conventional Variables

The conventional variables cover all those aspects about the dimension of the firms, and all those types of information that could be detected from the Financial Statements. They have proved to be modest predictors for the assessment of the survival of the firms.

4.1.1. Amount

The amount variable is not considered in the model, this means that issuing a bond of few moneys or a big one it is not a determinant for the default of a company. First of all, it is important to remember that the firms took in consideration are those that issued less than 50 million of Euros in less than a month, therefore the amount itself has an upper bound. Reasoning about theory we can understand why the amount is not considered as a significative variable, we can assume that the firms in the database are of different dimensions, even remaining in the field of the SMEs. We also assume that for a firm will be convenient issuing new debt until the point of distress as already discussed in chapter 1.1, from theory about the capital structure and the leverage we know that this point is reached when the value of debt is close to the company value, it is reasonably that firms are not like to reach similar points. Stating this, if we consider that the dimensions of the firms are different, the model could not detect if the absolute value of the amount issued could be determinant in the default of a firm. A hint for the future developments could be the assessment of a new variable that will standardize the amount issued with the value of the firm in order to understand if there is a connection between the amount issued and the global level of debt in relation with the value of the firm.

4.1.2. Assets

In the model built on the classification trees the level of assets also appears in the significant variables. The correlation in this scenario is negative and the companies with an elevated level of assets are perceived as less risky. The reason behind that could be observed in two fields of the assets: high level of cash and liquid assets can help the firm in reaction to liquidity shocks in adverse moments and could avoid insolvency risks. On the other hand, looking at the long-term assets, the high availability of illiquid assets allows firms to exploit them as collateral in the debt issuance, giving the opportunity to get higher amounts of credit, requiring lower interest rates or more in general reach more favourable conditions in the credit market.

The level of assets would have resulted significative also in the regression models, however this variable has been deleted from this model due to its high level of collinearity that would have given distorted results in the algorithm.

4.1.3. Covenants

A useful comment for this variable is the particularity of not being so significant. Bonds with covenants are by definition riskier if compared to the ones without, the reason behind is that covenants are used as a safety net for the lender in order to reduce the risk of insolvency, so the expected output should be a positive correlation between the presence of covenants and a default, this event does not happen.

Even if there was this type of correlation we excluded it because we consider covenants as an endogenous variable, in fact, in a model with covenants, we could have translated the model in the following statement: “the company X has been perceived riskier because it has issued minibonds with some type of covenants” in reality the correlation cause-effect is exactly the opposite, the company is not risky because the debt granted has covenants; instead the true sentence should be the following “The company X issued a bond with some type of covenants because has been perceived riskier.”.

4.1.4. Employees

Employees have been for centuries the spine of a company; the labour force was spread in the majority of the of the operations in the productive cycle of a firm. This is no more the truth. The more mechanical operations that were been handmade, starting from the Sixties of the last century, has been substituted by machines, the theories in Operation Management, and the advent of the Lean Manufacturing moved the labour force to enhance new competences.

Another factor that needs to be considered is the predominance in the last years of the service sector; a low number of employees are able to offer services for which some decades ago were needed much higher numbers. The last aspect to be analysed is the advent of start-ups and hi-tech company, even if in Italy the phenomenon is less widespread then other countries like the United States, these new types of companies are becoming more frequent, and one of their peculiarity is the possibility of managing high cash flows with low human effort if compared with the industrial sector. All these arguments make us confident about the absence of correlation between the number of employees and the health of a firm.

4.1.5. Equity

Equity has been found as one of the worst predictors on which rely on, we can study this curious aspect looking for the reason behind it. The first reasoning that could be made is that it is not so important on how much is the capital involved in a company, but rather how the management use the money available for the firm, a firm with a low availability of money could exploit it much better than a big company that carries out poor managerial decisions.

The second reasoning is about the composition of the Liabilities part in the Balance Sheet, we saw that Assets are much more relevant than equity in the model, we did not delete them for their low level of significance , rather than for its high level of collinearity with the other variables that could give us a distorted model. Following the accounting principles, Assets are always equal to Liabilities and these lasts are divided in two sections: Equity and Debt; stating that, we can reasonably assume that all the significative part in the Liabilities is in the Debt part.

4.1.6. Geographical Area and Region

The geographical area and the belonging region of the firm can be analysed together since the first is a generalization of the second. We took them separated in the model because could be interesting if inside a determined geographical area of Italy, we could find some excellences or some deficiencies.

If we had found something in this sense we could rely on two aspect, the propension to entrepreneurship of some places in Italy is the first, difficult to be catch; the second aspect would be the policies of some Regional Governments, for example financial assistance expressed like low-interest loan or some type of soft loans; some other aids could also come from the Central Government like a low taxation.

It is interesting seeing how the intervention of the State could be an help in no showing this misalignment, in this paper the role of the politics has non be highlighted but could be possible that this type of aid helped in reaching equality in the different areas of our Country; examples of this aids are the low n.208 of the 2015 [4] that gives a tax offset to all the productive companies that buy and use instrumental assets in the South of Italy, or the decree law n.124 of the 2023 [5] that establish the Special Economic

Zone for the entire South of Italy, also in this case, the aid is in form of a tax offset for the companies belonging to this part of the Country.

It is possible a numerical interpretation looking at the number of companies that issues minibond: on 969 issues 670 (69.14%) are from the North of Italy, 170 (17.54%) from the South and 129 (13,32%) from the Centre. This such great disparity could be connected to two aspects: the number of companies active in the country and the financial education of the management. We will rely on an ISTAT research [6]; this research is a bit outdated (2012) but could give us an order of magnitude in the weights of the various areas of the Country: the Table 4.1 shows the number of firms operating in Italy and their dispersion across the Country.

Table 4.1: Distribution of the capital companies in Italy.

Geographical Area	Total number of firms	Percentage on the total
North	117,409	60.88%
Centre	44,719	23.19%
South	30,731	15.93%
Italy	192,859	100%

The numbers in the table make us confident about the first aspect already mentioned; we can see a high, even if less substantial, disparity between the three different areas of Italy.

Moving to the second aspect, we noticed that there are differences between the percentage in the companies that issued minibonds if compared with those that exist in the different parts of the Country, could be interesting in future developments to find a proxy and try to assess a possible connection between the financial education and the default probability, we expect an incredibly high correlation between the two variables.

4.1.7. Listed Companies

A listed firm is perceived as less risky in the models. When a company goes public needs to disclose information in order to reduce the asymmetries in the market and give more details to potential investors, banks and in general to all the market.

Listed companies usually have to reach significative importance in the belonging market in order to attract investors, moreover, going public is a process costly and complex that is not affordable for smaller companies. We can see that in this case the bigger companies tend to be safer than small businesses, looking at the Table 3.3, there is only one listed company that defaulted.

The second important finding in this table is that just eighty-one companies in the sample are effectively listed, this corroborates the difficulties in the listing process and, therefore, the companies that reached this goal are less likely to default.

Another aspect that is connected with the listed company is the possibility to attract foreign investors, the internationalization of a firm has not been analysed in this paper, however we will briefly discuss about it in the conclusion part when we will talk about potential starting point for the research.

4.1.8. Annual Interest Rate

The interest rate is present with high levels of certainty in all the models used and with the same meaning: the high cost of the debt leads to an increase in the default probability.

This is in line with theories from different aspects: first, the firms prefer lower interest rates in order to access the credit with advantageous conditions, like low-cost debt that it is easier to be repaid compared to an expensive one, on the other side, a lender is willing to lend money to a borrower at low interest rate only if it is perceived as a safe investment, the riskier the investment the higher the spread in the yield for the investment. This last point is also agreed with the trust concept that exist between the lender and the borrower, in fact in case of trust, the lender accept, ceteris paribus, a lower return for the investment because the investments itself results less risky. It is important to notice that trust does not have to be confused with the absence of market asymmetries.

4.1.9. Revenues

The revenues represent the most important variable in assessing the status of the firm. In all the models the revenues are correlated negatively with the default probability, this means that bigger firms tend to default less. The reasons behind this are clear from the investigation, the companies with elevated levels of revenue are the ones with a high market share and the highest volumes of sales and could exploit all the several advantages connected to the dimension of the firm such as reaching the critical mass or exploiting network and scale effects.

It is important to assess other types of analysis because the revenues are usually not a good predictor, before being available for the firm as profits, is needed to consider the cost structure and possible inefficiencies in the management. To complete the reason behind the revenues we should also transfer this economic view into a financial one. Obtaining reliable data about the cash logic it is almost impossible externally because free cash flows are not easy to be calculated and the process in order to obtain them is long and complex.

4.2. Non-conventional variables

The so called non-conventional variables that we analysed in this dissertation cover two field of the company. The first one regard the propensity of the firm in creating intellectual properties, for this reason we looked at the total amount of patents and applications registered, at the trademarks, and designs developed.

The second aspect that we have investigated is the composition of the Boards of Directors, in particular we looked at two different points: the dimensions in terms of people, the composition looking both at the age both at the rates in gender balance.

4.2.1. Age of the Board of Directors

The correlation between the age of the members of boards and the probability of default is positively correlated, this could be related to the increase in the average level of education. In particular, in the tertiary sector, education is becoming increasingly specialized and sectorial more than previous decades among young people, therefore this type of correlation could be seen more as a premium of the younger board rather than lack for the older ones, in other words we can imagine that for younger boards is more difficult failing, however, we are not saying that older boards tend to default more easily, also in this variable could be possible finding a correlation with the financial education and the new management styles. A second correlation that could be interesting for a possible presence is the correlation between the average age of the boards and the propension to internationalization.

4.2.2. Gender of the Board of Directors

This variable is less significant if compared to others, however, it is negative correlated with the probability of default. In this case understanding this variable could be a bit more complex: the variable is defined in [0;1], 0 is a board composed by only female members while 1 is the opposite. It is necessary to be aware about the nature of the data, in the sample less than 20% of all the people in a board of directors are women.

4.2.3. Patents

Patents results quite significant in the models analysed and always in the same manner: the more the patents correspond to the better conditions of the company. In this view we can affirm that the innovation capacity and propensity of firms in innovation is a core competence in a dynamic environment in which adaptability is crucial to exist and be competitive.

Moreover, the companies that patent intellectual properties must invest in human capital and resources to develop them, according with these peculiarities of patents, we see a high similarity between the intellectual properties and a more classical long-term investment. In the case of a patent, the long process of Research & Development could be compared with an investment in a productive plant or a speculative investment that requires several years in order to reach its payback point.

4.2.4. Trademarks and Designs

This two type of intellectual properties have been analysed along the research as a single variable due to their comparability; in fact, the both of them arise from artistic competences and creative abilities. The correlation between them and the default probability is negative, the answer to this result could be explained in terms of human capital: outlining a design or a trademark is a complex creative process that require high skilled employees, strong marketing research, and time to reach the right formula that is valuable for the company.

The comparison between these resources and other type of material or financial investments follows the same reasoning made before for patents. Trademarks and designs have a lower level of significance if compared to patents; this could be explained in the different process of acquiring this type of intellectual properties compared to the application process of the patents. Trademarks and designs have a much less strict procedure in order to be registered, for example designs became active already since the moment in which they are created like the procedures for copyrights, both of them are considered valid even before the application starts.

4.2.5. Members of the Board of Directors

The most important variable in the composition of the Board of Directors is represented by the total number of members that compose it. This variable is negatively correlated with the default probability; in detail, more are the members of the board the better is for the company.

It is interesting to analyse this aspect, having more members of the board enable the sharing of different backgrounds of competences and know how, reaching better the needs of the company, and creating sustainable competitive advantages. It is possible to imagine that more people that interact in the top-level management of a company could have a broader vision and, reasonably, they can have a more sectorial and deeper expertise in their field but sharing the competences in all the various sectors enable them to reach a more complete view of the company and this led to better management choices.

5 Conclusion and future developments

This last chapter is a summary of our results, in particular their meaning and the possibilities for future developments and limitations in the procedure of this dissertation.

5.1. Sum up of the main results

The research conducted in this dissertation lead us an important result. The most important set of variables considered is the second: the model that use predictors based on the human capital.

In fact, the variables considered in the second logistic regression come from two different fields of application that are the intellectual property development abilities and the composition of the Top Management level.

Talking about intellectual properties, we found a negative correlation between the number of IPs registered and the default probability of the companies analysed, considering Patents, Trademarks, and Design. This means that, empirically, there are evidence in affirming that for the companies investing in the research can become a driver for competitive advantages and a key resource for growth.

The results on the composition of the management lead evidence about two different aspects: first of all, the age of the member of the Board of Directors results positively correlated with the firms default probability and at the same time wider Boards of Directors lead to lower probabilities of default. It is important to note that we have also analysed the composition of the boards from the point of view of the genders, in this case we have not found any statistical evidence, however we should consider that more than 80% of the directors in the sample analysed are males and that more than 90% of the principal representatives of the firms are males too.

We conclude our dissertation with more global consideration. Companies across the world calculate their success in terms of economic indicator, focusing on numbers and productivity. Our research instead highlighted how, from different pathways, we reach the same results: could be for the invention ability or for the collaborative aptitude, the strongest resource on which the firm can rely is still the human brain.

In this dissertation we analysed different possibilities for future developments in the predictive research for default and bankruptcy using sets of predictors different from the classical financial ratios. Our research matches for some aspects and diverges for others principally with the research of paragraph 1.3.2 (here we do not report again all the studies and findings).

Our research confirmed that the non-financial indicators can be use as predictor for bankruptcy in the SMEs sector, in fact we used similar type of variables that regard the innovation capacity of a firm and the top-level governance.

There is a difference quite interesting that differing our research from the other already done. We used a model that includes both the financial predictors both the non-conventional ones; we expected a confirm of the previous studies in which a combination of the two sets of variables would have been resulted in the best set of predictors. The empirical evidence does not confirm these theories, on the contrary the model that combines the two typologies of variables is the one with the worst performance if compared to the other two.

For what regard the conventional variables we have a result that confirm the research already conducted Bongini et al. (2017). We saw that the listing process and the size of the company are strongly relevant in assessing the wealth of the companies as said in already existing papers in chapter 1.

In addition to these variables discussed up to this point, we tried to analyse another aspect that could be crucial from a social point of view: the composition of the board in terms of genders. We cannot distinguish a correlation from a statistical point of view between board composed by only men and boards with a mixed composition. However, we do not feel confident in affirm that this difference exists or not because the women in the boards are too few to make it understandable. We do not have the pretence to remark this aspect since out of our sphere of competence, we entrust that the experts of this sector will conduct the necessary research.

5.2. Limitations and future developments

This paper has absolutely not the intent to be a conclusion in this research field but rather we consider it as a starting point for the studies in the socio-economic fabric. In this last section we would like to suggest some new possibilities for future developments based on our limitations and difficulties.

First of all, our research is based on the minibond issuances up to 31st of December 2020 with a sample dimension of 969, however, just adding the data of 2021 and 2022 the sample would reach a dimension of 1459 observations. It is clear that a wider sample dimension would lead to more reliable results. In order to permit a speeded-up process of data collection we suggest the developing of a tool that automize the phases descripted in section 2.3..

Looking at the profits equation that we developed in paragraph 3.6.2, it is important to notice that could be interesting going deeper in the analysis of this optimization; for example could be useful expand the coefficients for the unitary costs and revenues, we used simply the sample mean in order to understand the magnitude of the profits, however, we profoundly know how using the average sometimes could be senseless, therefore could be interesting creating subfunction that will give more reliable, even if more complex, predictors.

The third aspect is for the choice of the variables: adding more variables could be useful in giving more precise predictions. Some examples could be the following: the addition of types of variables that regards the education of the management of the companies are probably relevant in this type of analysis and could became a great proxy for the financial education, that, as everyone, we expect to be incredibly correlated with the default propension of the firms. Another variable that could be chosen is the retribution of the members of the various boards, maybe in relation to the average compensation of a top-level manager in Italy, this could be done in order to understand if higher wages bring in higher motivation and then could be led to health signs. It could be interesting also going in deep with the internationalization, this might be done through the percentage of revenues coming from abroad countries or moreover the percentage of foreign debt lender or shareholders. The last tip on this point is the possibility of adding the Research and Development expenses in the various company in order to have a second indicator on the research ability.

Looking at the variable a great question that we have not found an answer for is the following: The third model is built on the mix of the two previous models, therefore the intuition should be that the third model has the better results. We demonstrated the opposite empirically and with statistical evidence, but which is the reason of this result?

The last point of our suggestion is the modification of some variables, an example might be standardizing them somehow in particular for that variables which resulted non significative like the amount (the meaning of its non-significance has been already discussed) or analysing the endogenous variables like covenants and warrants.

Bibliography

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Amendola, A., Bisogno, M., Restaino, M., & Sensini, L. (2011). Forecasting corporate bankruptcy: empirical evidence on Italian data. *EuroMed Journal of Business*, *6*(3), 294–312. https://doi.org/10.1108/14502191111170132

Amendola, A., Giordano, F., Parrella, M. L., & Restaino, M. (2017). Variable selection in high-dimensional regression: a nonparametric procedure for business failure prediction. *Applied Stochastic Models in Business and Industry*, *33*(4), 355–368. https://doi.org/10.1002/asmb.2240

Angelino, A., Balda, F., Oliverio, F., Sampoli, L., & Schivardi, F. (n.d.). *Rapporto Cerved PMI 2022*.

Balcaen, S., Manigart, S., Buyze, J., & Ooghe, H. (2012). Firm exit after distress: differentiating between bankruptcy, voluntary liquidation, and M&A. *Small Business Economics*, *39*(4), 949–975. https://doi.org/10.1007/s11187-011-9342-7

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, *4*, 71. https://doi.org/10.2307/2490171

Bisogno, M., & De Luca, R. (2015). Bankruptcy Efficiency and Indirect Costs in Italian SMEs: a Temporal Approach. *International Journal of Business Research and Development*, *3*(2). https://doi.org/10.24102/ijbrd.v3i2.407

Bongini Paola, Rossi Emanuele, & Rossolini Monica. (2017). Suitable or Non-Suitable? An Investigation of Eurozone SME Access to Market-Based Finance. In *CEPR Discussion Paper No. DP12006*.

Brédart, X. (2014). Financial Distress and Corporate Governance: The Impact of Board Configuration. *International Business Research*, *7*(3). https://doi.org/10.5539/ibr.v7n3p72

Chancharat, N., Tian, G., Davy, P., Mccrae, M., & Lodh, S. (2010). *Multiple States of Financially Distressed Companies : Tests using a Competing-Risks Model*.

Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting corporate failure: empirical evidence for the UK. *European Accounting Review*, *13*(3), 465–497. https://doi.org/10.1080/0963818042000216811

Chen, I.-J. (2014). Financial crisis and the dynamics of corporate governance: Evidence from Taiwan's listed firms. *International Review of Economics & Finance*, *32*, 3–28. https://doi.org/10.1016/j.iref.2014.01.004

Chi, G., & Meng, B. (2019). Debt rating model based on default identification. *Management Decision*, *57*(9), 2239–2260. https://doi.org/10.1108/MD-11-2017-1109

Chludek, A. K. (2011). A note on the price of trade credit. *Managerial Finance*, *37*(6), 565–574. https://doi.org/10.1108/03074351111134754

Ciampi, F., & Gordini, N. (2013). Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises. *Journal of Small Business Management*, *51*(1), 23–45. https://doi.org/10.1111/j.1540-627X.2012.00376.x

Daskalakis, N., & Psillaki, M. (2008). Do country or firm factors explain capital structure? Evidence from SMEs in France and Greece. *Applied Financial Economics*, *18*(2), 87–97. https://doi.org/10.1080/09603100601018864

Demirguc-Kunt, A., Bank, W., Martinez, M. S., & World Bank, P. (2008). *Bank Financing for SMEs Around the World: Drivers, Obstacles, Business Models, and Lending Practices*. https://www.researchgate.net/publication/23970207

Decreto-Legge 19 settembre 2023, n. 124: "Disposizioni urgenti in materia di politiche di coesione, per il rilancio dell'economia nelle aree del Mezzogiorno del Paese, nonché in materia di immigrazione.", 2023.

du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, *73*(10–12), 2047–2060. https://doi.org/10.1016/j.neucom.2009.11.034

EUROPEAN CENTRAL BANK. (2013). *Speech by Benoît Cœuré, Member of the Executive Board of the ECB, Eurofi High Level Seminar organised in association with the Irish Presidency of the Council of the EU, Contribution to plenary Session 11: Challenges and feasibility of diversifying the financing of EU corporates and SMEs, Dublin, 11 April 2013". URL: https://www.ecb.europa.eu/press/key/date/2013/html/sp130411.en.html*.

European Commision. (n.d.). *Annual Report on Europen SMEs 2022/2023*. https://doi.org/10.2760/028705

EUROPEAN COMMISION. (2003). *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*.

Frydman, H., ALTMAN, E. I., & KAO, D. (1985). Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. *The Journal of Finance*, *40*(1), 269–291. https://doi.org/10.1111/j.1540-6261.1985.tb04949.x

Gepp, A. ;, & Kumar, K. (2008). *The role of survival analysis in financial distress prediction*. http://www.it.bond.edu.au/publications/Theses.htm]

Goncalves, T. S. H., FERREIRA, F. A. F., JALALI, M. S., & MEIDUTĖ-KAVALIAUSKIENĖ, I. (2015). AN IDIOSYNCRATIC DECISION SUPPORT SYSTEM FOR CREDIT RISK ANALYSIS OF SMALL AND MEDIUM-SIZED ENTERPRISES. *Technological and Economic Development of Economy*, *22*(4), 598–616. https://doi.org/10.3846/20294913.2015.1074125

Heiss, F., & Köke, J. (2004). Dynamics in Ownership and Firm Survival: Evidence from Corporate Germany. *European Financial Management*, *10*(1), 167–195. https://doi.org/10.1111/j.1468-036X.2004.00244.x

Howorth, C., & Moro, A. (2012). Trustworthiness and interest rates: an empirical study of Italian SMEs. *Small Business Economics*, *39*(1), 161–177. https://doi.org/10.1007/s11187-010-9285-4

ISTAT, «I GRUPPI DI IMPRESA IN ITALIA,» 2012.

Jones, S., & Hensher, D. A. (2007). Modelling corporate failure: A multinomial nested logit analysis for unordered outcomes. *The British Accounting Review*, *39*(1), 89–107. https://doi.org/10.1016/j.bar.2006.12.003

Kingdon, J., & Feldman *, K. (1995). Genetic algorithms and applications to finance. *Applied Mathematical Finance*, *2*(2), 89–116. https://doi.org/10.1080/13504869500000006

Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, *252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

Legge 28 dicembre 2015, n. 208: "Disposizioni per la formazione del bilancio annuale e pluriennale dello Stato (legge di stabilità 2016)", 2015.

Lin, F., Liang, D., & Chu, W.-S. (2010). THE ROLE OF NON-FINANCIAL FEATURES RELATED TO CORPORATE GOVERNANCE IN BUSINESS CRISIS PREDICTION. *Journal of Marine Science and Technology*, *18*(4). https://doi.org/10.51400/2709-6998.1901

Lucey, B. M., & Mac an Bhaird, C. (2006). Capital Structure and the Financing of Smes: Empirical Evidence From an Irish Survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.905845

Maksimovic, V., & Phillips, G. (2001). The Market for Corporate Assets: Who Engages in Mergers and Asset Sales and Are There Efficiency Gains? *The Journal of Finance*, *56*(6), 2019–2065. https://doi.org/10.1111/0022-1082.00398

Mantovani, G. M. (2015). *The Maturity Drivers of Corporate Capital Structure of Private/Unlisted Companies*.

Moon, T. H., Kim, Y., & Sohn, S. Y. (2011). Technology credit rating system for funding SMEs. *Journal of the Operational Research Society*, *62*(4), 608–615. https://doi.org/10.1057/jors.2010.15

Moon, T. H., & Sohn, S. Y. (2010). Technology credit scoring model considering both SME characteristics and economic conditions: The Korean case. *Journal of the Operational Research Society*, *61*(4), 666–675. https://doi.org/10.1057/jors.2009.7

N. Berger, A., & F. Udell, G. (1998). The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle. *Journal of Banking & Finance*, *22*(6–8), 613–673. https://doi.org/10.1016/S0378-4266(98)00038-7

Nicolaes Tollenaar. (2017). *The European Commission's Proposal for a Directive on Preventive Restructuring Proceedings*.

Osservatori Entrepreunership Finance & Innovation, & - Politecnico di Milano. (2023). *9° Report italiano sui Minibond*.

Ozturk, B., & Mrkaic, M. (2014). SMEs' Access to Finance in the Euro Area: What Helps or Hampers? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2445463

Pan, H., Kang, M.-S., & Ha, H.-Y. (2017). Do trade area grades really affect credit ratings of small businesses? An application of big data. *Management Decision*, *55*(9), 2038–2052. https://doi.org/10.1108/MD-11-2016-0834

Rege, U. P. (1984). ACCOUNTING RATIOS TO LOCATE TAKE-OVER TARGETS. *Journal of Business Finance & Accounting*, *11*(3), 301–311. https://doi.org/10.1111/j.1468-5957.1984.tb00751.x

Schweizer, D., Proelss, J., & Mietzner, M. (2015). *Hidden Champions or Black Sheep? Evidence from German Mini-Bonds*. https://doi.org/10.13140/RG.2.1.2731.6323

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

Sohn, S. Y., Doo, M. K., & Ju, Y. H. (2012). Pattern recognition for evaluator errors in a credit scoring model for technology-based SMEs. *Journal of the Operational Research Society*, *63*(8), 1051–1064. https://doi.org/10.1057/jors.2011.105

Sohn, S. Y., & Kim, Y. S. (2013). Behavioral credit scoring model for technology-based firms that considers uncertain financial ratios obtained from relationship banking. *Small Business Economics*, *41*(4), 931–943. https://doi.org/10.1007/s11187-012-9457-5

Università di Pisa, «Descrizione data base Aida,» N.A..

Vander Bauwhede, H., De Meyere, M., & Van Cauwenberge, P. (2015). Financial reporting quality and the cost of debt of SMEs. *Small Business Economics*, *45*(1), 149–164. https://doi.org/10.1007/s11187-015-9645-1

Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision Support Systems*, *98*, 59–68. https://doi.org/10.1016/j.dss.2017.04.008

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, *11*(5), 545–557. https://doi.org/10.1016/0167-9236(94)90024-8

Yu, S., Chi, G., & Jiang, X. (2019). Credit rating system for small businesses using the K-S test to select an indicator system. *Management Decision*, *57*(1), 229–247. https://doi.org/10.1108/MD-06-2017-0553

A Appendix A – Script of the algorithms

In this appendix we attach the R scripts used for the built of the model. Important: these algorithms are pseudo-algorithms and for this reason should not be considered as runnable.

A.1. Algorithm of the logistic regression

Algorithm 1 Logistic Regression

```
avg_amount=mean(Database$Amount)
avg_rate=mean(Database$Rate)
tax=0.5
mod = glm(Database$Default ~ .-Status, family=binomial(link=logit), data = Database)
summary(mod)
BIC(mod)
coefficients(mod)
anova(mod, test="Chisq")
vif(mod)
tab<-cov2cor(cov(Numer))
tb<-as.data.frame(tab)
pred = mod$fitted.values
pred=as.numeric(pred)
pred_class=ifelse(pred>0.5,"NO","YES")
tab = table(pred=pred_class,true=Database$Default)
tab
misclass_table=as.matrix(table(pred_class, Database$Default))
tpr = misclass_table[2,2]/(misclass_table[2,2] + misclass_table[1,2])
tnr = misclass_table[1,1]/(misclass_table[1,1] + misclass_table[2,1])
aper  = (misclass_table[1,2] + misclass_table[2,1])/dim(Database)[1]
acc = 1 - aper
tpr
tnr
```

```
fn=misclass_table[1,2]
fp=misclass_table[2,1]
bank_profit=avg_rate*avg_amount*misclass_table[1,1]+(0-avg_amount)*misclass_table[1,2]+(0-avg_rate)*avg_amount*misclass_table[2,1]
x11()
p = seq(0.01,0.6, by = 0.005)
tpr = tnr = bank_profit = aper = rep(0,length(p))
for ( i in length(p):1)
{
  pred_class=ifelse(pred>p[i],"NO","YES")
 tab=as.matrix(table(pred_class, Database$Default))
  tpr[i] = tab[2,2]/(tab[2,2] + tab[1,2])
  tnr[i] = tab[1,1]/(tab[1,1] + tab[2,1])
  bank_profit[i] = avg_rate*avg_amount*tab[1,1]+(0-avg_amount)*tab[1,2]+avg_amount*tab[2,2]+(-avg_amount*avg_rate)*tab[2,1]
}
par(mar=c(5.1, 4.1, 4.1, 8.1), xpd=TRUE)
plot(p, tpr, col = "red", type = "l", ylim = c(0,1))
lines(p, tnr, col = "green", type = "l")
legend("topleft", inset=c(0,0), fill = c("red","green"), legend = c("tpr (sensitivity)","tnr (specificity)"))
p = seq(0.01,0.6, by = 0.005)
tpr = tnr = bank_profit = aper = rep(0,length(p))
for ( i in length(p):1)
{
  pred_class=ifelse(pred>p[i],"NO","YES")
  tab=(table(pred_class, Database$Default))
  bank_profit[i] = avg_rate*avg_amount*tab[1,1]+(0-avg_amount)*tab[1,2]+avg_amount*tab[2,2]+(-avg_amount*avg_rate)*tab[2,1]
}
par(mar=c(5.1, 4.1, 4.1, 8.1), xpd=TRUE)
plot(p, bank_profit, col = "red", type = "l", ylim = c(-500,500))
lines(p, tnr, col = "green", type = "l")
legend("topleft", inset=c(-0,0), fill = c("red"), legend = c("Profits"))
pred = mod$fitted.values
pred=as.numeric(pred)
pred_class=ifelse(pred>0.5,"NO","YES")
```

A.2. Classification Tree Algorithm

Algorithm 1 Classification Tree

```
site=(968/2)
site=as.integer(site)
set.seed(02091991)
train=sample(1:nrow(Database), 484)
a_train=Database[train,]
a_test=Database[-train,]
tree.a=tree( Default~.-Status ,a_train)
summary(tree.a)
plot(tree.a)
text(tree.a,pretty = 0 )
tree.pred_test = predict(tree.a,a_test,type="class")
tree.pred_training = predict(tree.a,a_train,type="class")
tab = confusionMatrix(tree.pred_test,a_test$Default)
tab
cv.a = cv.tree(tree.a,FUN=prune.misclass)
plot(cv.a$size,cv.a$dev,type="b")
prune.a = prune.misclass(tree.a,best=10)
plot(prune.a)
text(prune.a,pretty=0)
tree.pred = predict(prune.a,a_test,type="class")
tree.pred[1:6]
tab = table(tree.pred,a_test$Default)
tab = table(a_test$Default,tree.pred)
tab
(tab[1,2]+tab[2,1])/site
(tab[1,1]+tab[2,2])/site
tab = confusionMatrix(tree.pred,a_test$Default)
tab
```

List of Figures

List of Tables

Acknowledgments

We would like to express our gratitude, first of all, to our Professor Giancarlo Giudici, for the support that gives us during all the journey of this thesis.

Secondly, we would like to thank all the community of the Politecnico di Milano for all the competences and the development of our mind in these years.

Lastly, we would like to thank our beloved families, friends, and all people supporting us from the first day.