Universitat Politècnica de Catalunya (UPC) - BarcelonaTech
Facultat d'Informàtica de Barcelona (FIB)

Master in Innovation and Research in Informatics
Data Mining and Business Intelligence
Final Master Thesis Report

Defining Corporate Credit ratings for Spanish Companies using Machine Learning techniques

Submitted by
Francesc Busquet

Advised by
Lluís A. Belanche
Computer Science Department

June 11, 2018

Abstract

The financial Crisis of 2008 exposed the dubious practices followed by the major Credit Rating agencies, which because of conflicts of interests set the highest ratings available to products that presented major risks, being one of the key culprits of this crisis. For this reason, the creation of alternative measures of a firm's creditworthiness not being subject to conflicts of interest has become essential. Therefore, in this study we address this problem by implementing several techniques to assess a firm's creditworthiness in the area of Spain. To do so, a representative dataset of Spanish firms was used, testing different statistical and machine learning techniques. Results indicate that the best predictive accuracy is achieved using a Random Forest approach, achieving an accuracy near 90%, outperforming the vast majority of default prediction models (which usually achieve no more than 85% of accuracy) and achieving a result that is more than enough to produce a reliable metric to assess a firm's creditworthiness.

Acknowledgements

Firstly, I would like to thank my thesis Supervisor, PhD. Lluís A. Belanche, who has provided me invaluable guidance throughout the whole master thesis.

Furthermore, I would like to thank PhD. Argimiro Arratia, who has generously helped me in some of the doubts that I have had during the elaboration of this thesis.

Contents

List of Figures

List of Tables

Introduction

Corporate credit ratings have become one of the most common indicators of a firm's creditworthiness, reducing lender's uncertainty about the borrower's repayment capacity and, therefore, easing the credit flow. Corporate credit ratings must show a reliable (through-the-cycle) and trustworthy vision of a firm's creditworthiness. Additionally, those ratings must be based on objective criteria, in order to be comparable.

Credit ratings must be based on a huge set of dimensions that may affect a firm's creditworthiness, such as the risk involved with the country in which they operate, the risks involved in their industry or the risks involved to their own financial and business conditions. Therefore, setting a firm's credit risk is a tedious and costly process.

All those reasons have contributed to the establishment of a reduced number of independent agencies with vast knowledge in credit risk (Moody's, Standard and Poor's and Fitch) to provide the majority of corporate credit ratings. Even though, the 2008 financial crisis showed how those companies were also subject to conflicts of interest, due to the huge compensations received from their clients. High risk products were rated with the most positive available rating, being one of the major drivers of the financial crisis.

Thus, there is a high necessity for further independent and transparent corporate credit ratings. For this reason, this thesis is aimed at developing a corporate credit rating for Spanish companies using Machine Learning techniques, based on a firm's default prediction. Thus meeting all the previously mentioned characteristics of a firm's corporate credit rating (reliability, objectivity, trustworthiness), and decreasing the amount of time and resources needed by conventional systems to generate credit ratings, which usually takes some months and implies a high monetary cost. Moreover, the approach followed in this study was intended to apply to all kinds of businesses, independently of their size, therefore facilitating rating universality.

Main Objectives

The main objectives of this thesis are shown in the following list:

- Identify the main determinants of a firm's creditworthiness, by analyzing the main frameworks used on corporate credit risk analysis and the previous studies done in default or bankruptcy prediction.

- Construct a representative dataset for Spanish companies comprising all the determinants previously identified. Taking into account that all the determinants should be suitable either for big and small companies.
- Implement different statistical and Machine learning techniques, determining the best subset of predictors using a wrapper approach and the best combination of hyperparameters for each of them.
- Compare and evaluate the performance of the different models, selecting the best one, achieving an error rate lower than 15%[1].
- Generate a qualitative metric for predicted default probabilities, in order to mimic current corporate credit ratings.

Principal Contributions

Previous research has centered on bankruptcy prediction, while default and bankruptcy conditionants are highly similar, default prediction has been set a side due to the difficulty low number of available information about corporate default. Additionally, several of the previous studies that tried to predict corporate default missinterpreted default as bankruptcy.

For this reason, this is the first study predicting default in the Spanish area. Moreover, this study overcomes several of the problems spotted in previous literature, by using a cross-sectional sample, using balanced classes, and normalizing financial ratios using log transformation.

Additionally, in order to achieve reliable indications, not highly volatile with the economic cycle, a though-the-cycle approach was implemented, using information about the three previous years before defaulting.

Furthermore, this is the first study to use a random forest model for corporate default predicting. Achieving results superior to those achieved by methods used in previous literature (such as CART, C5.0 or Multilayer Perceptron).

Thesis Structure

This Master's Thesis is divided into six chapters starting from a theoretical viewpoint of corporate credit ratings to a practical view and implementation of a corporate credit rating system based on machine learning techniques for Spanish companies.

In chapter 1, we begin with a theoretical background of Corporate Credit Ratings, using economics and finance theory to describe the need for corporate credit ratings, the concept of corporate credit ratings and the factors that influence a firm's creditworthiness.

In chapter 2, we present a view of the progress done in this subject, presenting a historical

[1] 15% is the error rate achieved by top performing default prediction models, such as Thomson Reuters Starmine structural credit risk model("Starmine Quantitative Models", 2013)

review of corporate credit ratings literature, as well as providing a view of the different types of commonly used models for corporate credit assessment. Furthermore, we provide a critical view to some of the last studies in this topic, assessing the systematic deficiencies found on the reviewed literature.

In chapter 3, we describe the data sources used in order to define Corporate Credit Ratings in further chapters. Furthermore, using the theoretical background provided in chapter 1 and chapter 2, we proceed to define the drivers of a firm's creditworthiness (explanatory variables), in order to proceed to predict a firm's probability of default. Besides, we illustrate the methodology followed in the present study.

In chapter 4, we apply different models to predict a firm's default probability, starting from typical statistical models, such as logistic regression, decision trees to machine learning models, such as random forest, support vector machines and neural networks. Finally, we present the main results obtained using the models elaborated.

In chapter 5, we transform those probabilities to letter ratings, in order to mimic conventional credit ratings, thus providing rating comparison and making those ratings more easily understandable.

And finally, in chapter 6, we summarize the main contributions and results from this master's thesis. Additionally, we provide suggestions for further research in this topic.

1 | Corporate Credit Rating: Background

In today's world, lending money to a business or acquiring corporate bonds is a common practise, since they provide control and certainty of the investment made.

In corporate bonds and business lending, cash flows on investments are promised when the investment is made. Even though, in those transactions the borrower can be exposed to nonpayment risk. For this reason, the firms capacity and willingness to pay needs to be analyzed before investing in it.

Hence, in order to lend money to a company, the lender must have conducted a full due diligence analysis, examining all the assumptions and scenarios, so as to evaluate the risks involved.

Generally speaking, this risk should determine the interest rates on the borrowing: borrowers with lower default risk should pay lower interest rates than those with higher default risk. Therefore, the expected return on a corporate borrowing is likely to be related to the borrower's default risk.

Corporate credit ratings are one of the best-known forms of qualitative measurement to asses default risk. Those ratings describe the firm's creditworthiness and are usually assigned by an independent ratings agency, such as: Standard & Poor's, Moody's and Fitch (Big Three Credit Agencies).

This chapter discusses the need for corporate credit ratings by explaining some of the fundamental problems in the credit market. It also presents a further explanation of the corporate credit rating concept, as well as the general components that are considered in credit analysis to define the creditworthiness of a firm. Likewise, it presents the main objectives of corporate credit rating.

1.1 The need for Corporate Credit Ratings

Every business, independently of its size, is funded with a mix of debt (borrowed money) and equity (owner's funds).

According to Damodaran (Damodaran, 2010) each type of funding has its advantages and disadvantages. Debt has two main advantages:

- **Tax Benefit**: In many countries the interests paid on debt are tax-deductible, while equity cash flows are not (they are paid out of after-tax cash flows).
- **Debt payments impose an added discipline to the management**: Borrowing may increase the risk of default on projects with second-rate returns, which increases managerial discipline and impose a stricter project choice criteria and management (Jensen & Chew, 1995).

Whereas, debt also has several disadvantages:

- **Increased exposure to default**: Debt involves interest and principal payments, thus reducing the firms' future cash flows.
- **Agency Costs**: There is a conflict of interests between equity holders and debt holders, since equity investors tend to favor those actions that increase the value of their holdings, even at the expenses of increasing the risks faced by bondholders.
- **Reduced future flexibility**: Taking on debt implies introducing covenants that restrict the firms' flexibility. Furthermore, firms may prefer to preserve debt capacity for difficult or unforeseen situations (Modigliani & Miller, 1963).

The trade-off between the advantages and disadvantages of debt will define the optimal capital mix of a business, i.e. defining the amount to borrow.

As we have said one of the disadvantages of funding a firm with debt are the agency costs involved in issuing debt. This point needs to be emphasized since there exists an asymmetric relationship between investors and entrepreneurs. In other words, entrepreneurs have a privileged knowledge about their businesses and projects.

The relative lack of information of investors over entrepreneurs may cause adverse selection and moral hazard (De Servigny, Renault, & de Servigny, 2004). Generating one of the major inefficiencies in credit markets, thus constraining the availability of credit to businesses.

Adverse selection implies that creditors will not be able to distinguish good companies from bad companies, because of the creditors' lack of information. Hence, they will charge to all the companies the same interest rate, discouraging the good companies from its intentions to borrow. Therefore, bad companies will drive out good companies from the credit market (Akerlof, 1978).

Besides, moral hazard implies that the debtor will take actions that may negatively affect the creditor, because of the creditor's impossibility to track the debtor. Some examples of those activities are over-investment, or investment on high-risk projects (since part of the risk falls to the creditor, who has limited upside but potentially high downside risk (Damodaran, 2010)).

Additionally, more specifically, some authors point that some level of debt may act as a positive signal to creditors (Brealey, Leland, & Pyle, 1977; Ross, 1977), because the penalty involved by the debtor in case of bankruptcy and the rationale that the debtor (entrepreneur) must incur debt to finance projects that will increase its firm value. Furthermore, firms are found in a dynamic scenario, in which they have the need to sustain a good reputation as a debtor, in order to be able to raise credit in the future (Dewatripont, Tirole, et al., 1994). Although, problems aroused by the presence of asymmetric information are a reality in credit markets (Tang, 2009).

Economic theory offers several methods to overcome the problems derived from asymmetric information (moral hazard and adverse selection) between two parties (creditor and debtor):

- **Conditional allocation of control rights**: allowing the distribution of control rights between the creditor and the debtor (firm) in front of certain conditions, such as giving power to the debtors in case of distress (De Servigny et al., 2004). This method fulfills a twofold function: it mitigates the creditor's risks while incentiving the debtor to act appropriately, diminishing moral hazard.
- **Signaling**: the debtor (firm), in order to avoid adverse selection, can transmit some credible and verifiable information to the creditor showing its true creditworthiness. A classic example of signaling is provided in the context of the job market, where the education level is used as a signal of the quality of an employee to a prospective employer (Spence, 1978).
- **Screening**: similar to signaling, but undertaken by the uninformed agent, i.e. the creditor. One example, again in the labour market, could be an employer offering a menu of contracts, with different proportions of base salary and commissions. Potential employees with a higher level of productivity will choose contracts with a higher proportion of production related commissions (Laffont & Martimort, 2009).

Corporate credit ratings act as a signal to overcome adverse selection in the credit market, showing the debtor's (firm) creditworthiness and, thus, providing to the creditor information about its quality. As corporate bonds' interest rate is related to its default risk, which is measured by the corporate credit rating, high-rate corporate bonds will be priced to yield lower interest rate than low-rate corporate bonds (Damodaran, 2012).

Additionally, corporate credit ratings are usually elaborated by independent and prestigious agencies, which also monitor a firm's evolution along time, reflecting significant changes by updating a firm's rating. As confirmed by some studies, strong rated firms are less likely to engage in earnings management than other non or low rated firms (Li, 2017), thusly reducing moral hazard.

1.2 The concept of Corporate Credit Rating

Corporate Credit Ratings ultimately reflect a corporation's credit strength, being the most widely used measure of a firm's default risk.

In order to define a firm's corporate credit rating, a corporate credit analysis is needed, which will determine the firm's capacity and eagerness to pay its financial obligations in a timely manner, helping to distinguish good companies from bad companies.

There are several dimensions that can affect a firms creditworthiness and must be considered, in order to determine the firms' credit quality. Those dimensions can be classified in three blocks (Bilardello & Ganguin, 2005):

- **Country and Sovereign risks**: risks related to the country in which the firm operates and to the general macroeconomic conditions, such as changes in interest rates, economic cycle, etc.
- **Industry-specific risks**: risks related to the sector wither the firm operates, such as the sector's competitiveness, the nature and the life-cycle of the sector, etc.
- **Company-specific risks**: risks related to the firm, such as bad management of the firm, low sales consistency, poor financial flexibility, etc.

Hence, a business default will be ultimately conditioned by factors pertaining to those dimensions. For example a default could be due to industry specific risks, as a change in consumer's preferences, combined with company-specific risks, as a high level of debt.

The analysis of the risks associated to each of those dimensions is explained in a more detailed manner in the respective subsections that follows, pointing to the key concepts that must considered on each of those dimensions, in spite of defining the firm's creditworthiness.

1.2.1 Country and Sovereign risks

Country and Sovereign risks refers to those risks that are imposed by the country in which a firm operates, as well as the facilities provided by governments to economic growth and the correct performance of the private sector. Some of those risks may refer to a country regulatory framework, politico-economic instability, etc.

In economics, the Gross Domestic Product per capita (GDP per capita) is a commonly used measure of the performance of a country. Furthermore, by adjusting this measure by the relative cost of living and inflation rates of the countries (in other words, adjusting by purchasing power parity), this measure becomes a good indicator of the relative performance of a country.

In Table 1.1 we can observe the GDP per capita adjusted by purchasing power parity (PPP) for different countries. As we can see, there are notorious differences among those countries, this is due to divergences in the long-run economic growth of the different countries.

Growth in an economy is ultimately conducted by businesses, hence by analyzing the factors that accompany growth, we are analyzing the facilities/constraints imposed by a country's government and characteristics to the firms operating on it.

Table 1.1: 2016 GDP per capita (PPP) comparison

Spain	France	United Kingdom	Germany	Nigeria	Angola	Argentina
36,304.9	41,343.3	42,608.7	48,860.5	5,861.1	6,454.1	19,939.9

Data extracted from the World Bank

In 1956, two economists, Robert Solow and Trevor Swan, independently developed a quantitative model, called the Solow-Swan model, which attempts to explain the long-run economic growth. To do so, it determines the long-run economic growth as a function of the technology, the capital accumulation, the amount of labour and the saving rate (Barro & Sala-i Martin, 2007). Nevertheless, this model is not able to completely explain the economic growth, being the difference between the growth factor in the economy and the expected growth by the model, the Solow residual, i.e. that part of growth not measurable due to changes in the amount of capital, the amount of labour or the saving rate.

Initially, it was believed that this residual was explained by the growth attributable to the capital and the labour, even though those factors explain a low proportion of this residual. For this reason, several authors have tried to explain this residual; some authors (Hall & Jones, 1999; Acemoglu, Johnson, & Robinson, 2001) have found that this residual is strongly related to the quality of institutions. Moreover, they also point that growth is achieved, in part, thanks to innovators. And, since talent is randomly distributed, governments should favour the equality of opportunities among citizens. In a similar direction, other authors (Mankiw, Romer, & Weil, 1992) have found that this residual is also related to the human capital.

Institutional quality can be defined by four key points (Venard, 2013):

- Public sector is an efficient administrator, i.e. low corruption (economical and judicial), low bureaucracy and low tax evasion.
- Institutions do not distort the private sector performance, i.e. property rights are properly defined and there is no over-regulation.
- Efficient supply of public goods, such as public healthcare, education, etc.
- Allowance for political freedom.

Moreover, the concept of human capital refers to the level of education and the scientific talent of the population. Some studies have found a strong relation between the level of literacy (a measure of a population educational level) of a population and the subsequent rate of investment, and hence the subsequent rate of income growth (Romer, 1990).

Furthermore, other authors also point that a countries geography (Diamond & Ordunio, 2011) and culture are also key determinants on its income level. Since geography, conditions the weather and the availability of natural resources, as well as the number of pathogens that can cause human diseases (Dunn, Davies, Harris, & Gavin, 2010). While culture stimulates growth by two main channels: cultural attributes that encourage individual motivation, and attributes that promote social capital (Maridal, 2013).

Additionally, each country has specific fiscal (government revenue collection, such as taxes, and expenditure) and monetary policies (definition of the money supply size and growth rate, which in turn affects the interest rates; setting banks' reserve requirement, etc.). Those policies may also affect businesses' creditworthiness, for example a huge increase on income taxes may reduce severely the consumption level of individuals, negatively affecting businesses' earnings.

Thus, the stability or volatility of a country can be associated to many of the already mentioned factors. And this stability or volatility can be observed from numerous macroeconomic factors, such as the consumer spending, the inflation rate, the interest rates, etc. Likewise, the business cycle of an economy should be also considered, since economic cycles may severely affect the financial health of a business.

1.2.2 Industry-specific risks

Industry-specific risks refer to those risks imposed by the industry in which a firm operates. Thus, examples of those risks are changes in consumer preferences, technology and competitiveness of an industry.

Industry risks may impose a ceiling on businesses' profitability, since superior returns will attract competition as a magnet. Thereby, an industry competitiveness defines the attractiveness of an industry, shaping the level of returns obtained by the businesses in the sector.

One of the greatest exponents of industry and competition analysis is Michael Porter, who established a framework to analyze the attractiveness of an industry. He stated that the attractiveness of an industry is determined by five forces (Porter, 2008):

- **Threat of new entrants**: this force refers to the potential that new competitors come into the industry. This factor is determined by industry barriers of entry, i.e. barriers that protect an industry from newcomers. Some examples of entry barriers are economies of scale, switching costs, superior access to information, network economies, regulatory restrictions, high initial investment, etc.
- **Threat of substitutes**: this force refers to the existence (or potential existence) of products or service that meet the same basic need as the industry's product.
- **Bargaining power of customers**: this force refers to a threat imposed by customers; powerful customers will be able to force down the product's price or demand more value for the same price. Some factors that contribute to the power of bargaining of customers are their level of concentration, the existence of substitutes, the volume purchased, etc.
- **Bargaining power of suppliers**: this force refers to the threat imposed by suppliers; powerful suppliers will use their bargaining power to capture more value from their customers, by charging higher prices or supplying less value for the same price. Some factors that contribute to the bargaining power of suppliers are the customer's switching costs, the existence of substitutes, the supplier's level of concentration, etc.
- **Rivalry among existing competitors**: this force refers to the rivalry intensity among an industry's existing competitors, and it is conditioned by the four previous forces. Some factors that tend to increase an industries rivalry are slow industry growth, high exit barriers, and irrationally high commitment to the business from rivals.

Although, other authors(Greenwald & Kahn, 2005) simplify this approach by pointing that barriers of entry are the main force implied in determining the competitiveness of an industry. This is because barriers of entry (or to expand) limit the potential entrants to an industry, being the existing firms protected by those barriers and being able to achieve superior returns. While in an unprotected industry, other companies will flood in, driving down returns; being operational effectiveness the most important factor in the success of businesses in an unprotected industry.

There are two main indicators of entry barriers in an industry:

- **Market share stability among incumbent firms**: Market share stability reflects low competitiveness, indicating low or no expansion of incumbent firms in relative terms and low or null presence of new entrants.
- **Exceptional profitability among incumbent firms over a meaningful period**: Returns significantly above other industries are an indication of barriers of entry. Since industries with high returns act as a magnet to attract new competitors, anxious to capture part of those superior returns. Therefore driving down the industry's profitability. this is why returns on equity are highly similar among firms pertaining to different industries (Higgins, 2012).

Besides, there are other risks associated to a firm's industry, apart from its competitiveness. Those are the industry's nature and the current industry's life cycle phase. Regarding the industry's nature it is important to differentiate cyclical industries, those that are affected by the business cycle, from secular industries, those that are not affected from the business, such as the food industry.

Whereas, the industry life cycle, refers to the different stages that an industry experiments along time. To simplify, industries may experiment five stages during their life:

- **Start-up**: This phase represents the initial stage of an industry, where a new product is developed. A phase characterized by uncertainty, where information about the potential market is still limited. Consumers also need to obtain more information about the product of the industry, and get familiar with the offering. Hence, this phase tends to be featured by high business fragmentation in the industry and high levels of losses, due to high development and marketing expenses with low levels of sales.
- **Expansion**: This phase occurs after consumers have already acquired a clear knowledge and understanding about the industry's offering. Thus, business on the industry experiment a notorious sales growth. Even though, this phase is still not yet characterized by high levels of profitability, since business still incur in high expenditures in order to build a competitive advantage, i.e. acquiring superior performance and maintaining a solid position in the future.
- **High growth**: This phase is characterized by a demonstrated viability of the industry's product, where the potential market is well defined and still growing. In this phase, there will be high pressure of new entrants who will want to profit from the industry's expansion. For this reason, in this point is where entry barriers will be the only factor that will keep the industry's returns from degrading in relation to other industries. Moreover, it should be pointed that barriers of entry aroused only by economies of scale may be deteriorated, since industry growth may allow new entrants to capture a high enough market share as to also benefit from economies of scale(Greenwald & Kahn,

2005).

- **Mature growth**: This phase is characterized by a growth slowdown, where the product market has been consolidated. Because this growth slowdown, businesses on the industry set market share and cash flows as their main goals. This may lead to an increase of what Porter calls rivalry among existing competitors, which at the same time may deter the entrance of new entrants. Furthermore, this increase in competition may lead to a process of product homogenization, where firms on the industry end up competing on a pure operating efficiency basis.
- **Decline**: In this phase the industry's market starts to decrease, which may be due to product obsolescence or changes in the consumer's preferences. Hence, this phase usually indicates the end of viability of the industry's incumbent businesses. Therefore, this phase will be characterized by liquidations and/or redesign of incumbent firms.

1.2.3 Company-specific risks

In the two previous subsections we have analyzed the risks that may constrain a company's performance, hence we have related a company's performance to its environment. Although, each company operates in a different manner, thus each company needs to be analyzed in order to spot the risks that arise from their own financial and operating structure, as well as, due to its management.

The financial situation of a business can be analyzed through their accounting and financial statements. There are three main statements that should be analyzed to determine the financial situation, the position and the creditworthiness of a business:

- **Income Statement**: reflects the performance of a business during a certain period (usually of one year). This statement shows the revenues and expenses of a business, properly classified, allowing to distinguish operating and non-operating revenues and expenses. Hence, this statement is used to assess the profitability of a business, since it reports the net profit or loss incurred over the specified period. Even though, this statement is under an accrual basis of accounting, which means that it is not presented in a cash basis, i.e. revenues are reported when they are earned and that often occurs before they are paid by customers. The same happens with expenses, which are reported when they occur (or expire), which often differs from the moment in which the payment is made.
- **Balance Sheet**: it displays a company's assets, liabilities and shareholder's equity at a specific point in time. Thus, showing the financial position of a business, and allowing to spot the evolution of a business by comparing different periods' balance sheets.
- **Cash Flow Statement**: this statement reflects the cash inflows and outflows of a business, by adjusting net income for any non-cash expense. Hence, this statement shows the net change in cash and cash equivalents from start to end of a period. Thereby, providing a measure of liquidity.

Moreover, by observing the evolution of those statements we can assess a business consistency and stability, which is a key component of credit analysis. Hence, strong continued profitability will denote a strong competitive position, able to cope shifts in the business or the economy. While, strong continued liquidity will reflect the firm's ability to generate cash flows, being able to repay its financial obligations.

Finally, the balance sheet will provide measures of the company's financial structure, showing how debt grows in time and the amount of borrowed money in relation to a company's assets (high relative amounts of debt will increase a firm's probability of default). Hence, by analyzing the three statements and their evolution along time, we will obtain a general picture of the firm's profitability and creditworthiness. Furthermore, by observing the balance sheet, we will be able to determine a firm's financial flexibility, i.e. determining whether a firm in times of need, will have options for obtaining cash or not.

Nevertheless, corporate data by itself is not a good indicator for business comparison, since this information will be size dependent. Hence standardization of corporate information will be needed. The most common form of standardization comes from the use of ratios, which make comparison among firms and industries feasible. There are several kind of financial ratios, but the vast majority of them utilize information coming from the financial statements previously mentioned, being a good measure (in relative and absolute terms) of a companies' liquidity, profitability, operational efficiency, etc.

Another important factor to take into account is the firm's management, since management skill and ability plays a role at providing adequate liquidity (Bilardello & Ganguin, 2005). Furthermore, the relative position of a business among its competitors should be defined in order to spot future risks on its profitability.

1.3 Summary

In this chapter we have analyzed the benefits that debt offers to businesses. Similarly, we have also analyzed debt contracts through microeconomic theory, seeing that in those contracts the parties involved have different levels of information. For this reason, moral hazard and adverse selection is inevitable.

Although, microeconomic theory also offers solutions to those problems by stating a conditional allocation of control rights, by signaling or by screening. Also we argued that corporate credit ratings, which we defined as a qualitative measurement to assess a firm's creditworthiness, act as a signal to creditors. Thus avoiding adverse selection, and avoiding moral hazard since corporate credit ratings are usually set by independent credit agencies. Thereby, implying continued monitoring of a firms' financial health.

Furthermore, we defined the main determinants of a firm's creditworthiness, and thus of a firm's corporate credit rating. To define those determinants we followed the building block methodology, which relates the creditworthiness of a business to its environment, by analyzing the risks from the country in which it operates, as well as from the firm's industry. And, finally, analyzing the specific risks from a firm's financial and operational structure.

In order to analyze the country and sovereign risks we used concepts from macroeconomic theory, concretely by analyzing the drivers of economic growth, which is ultimately driven by businesses. Thus, we determined that growth was defined by factors such as a country's infrastructure, its institutional quality (referring to regulatory framework, public sector efficiency, the level of corruption, etc.) and its human capital.

Moreover, to analyze the risks associated to a firm's industry we used concepts from industrial organization economics, by stating that the competitiveness of an industry could be assessed by analyzing Porter's five forces, and paying special attention to an industry's entry barriers. Additionally, to fully assess the impact of the industry to a firm's creditworthiness we also pointed that the firms life-cycle should be analyzed.

Finally, in order to assess the firm's specific risks, we argued that a firm's operational and financial structure should be analyzed by analyzing its profitability, creditworthiness and financial flexibility using the three financial statements. Furthermore, we argued that other factors such as management could have a significant impact to a firm's creditworthiness.

2 | Present and Progress in Corporate Credit Rating

In the previous chapter we presented the concept of corporate credit ratings, as well as the dimensions involved in determining a corporate credit rating. However, it is not plausible to manually analyze every single aspect affecting a firm's creditworthiness and generate ratings that are consistent among them.

For this reason, one needs statistical approaches to determine a firm's creditworthiness and segregate those companies with low default risk from those with a high risk of default in an automated way.

Traditionally, it has been believed that corporate credit rating has relied on the use of simple models based on the analysis of a small set of key financial variables. Although, reality is completely different. Moreover, in the last years, due to the proliferation of machine learning and data mining, new opportunities to develop more complex methodologies in this area have appeared, bringing new and more sophisticated corporate credit rating models.

As we stated in the previous chapter credit ratings are associated to a firm's probability of default. Therefore, in this chapter we will analyze the most important contributions in assessing a firm's probability of default. Furthermore, studies on bankruptcy prediction have been crucial in the development and sophistication of default prediction, coming prior evidence on the role of the different default factors primarily from empirical bankruptcy research. Additionally, progress in individual credit scoring has also contributed to corporate creditworthiness assessment in several ways, such as the extrapolation of models applied to individual credit assessment to corporate credit assessment, thus some contributions on individual credit scoring will be also considered.

Therefore, in this chapter we will proceed to analyze the progress already done in the area of corporate credit rating, by reviewing the literature found in this area. Seeing how the techniques implemented in this area of research have grown in complexity during the recent years, adapting techniques from areas such as Data Mining and Machine Learning. Moreover, after reviewing the principal literature on default and bankruptcy prediction we will expose the main limitations incurred by the current models for default and bankruptcy prediction.

2.1 Literature Review

Formal analysis of a firm's creditworthiness started in 1932, when Paul Joseph FitzPatrick compared 13 financial ratios of 20 pairs of failed and successful firms, finding significant differences among them, specially on their liquidity and debt ratios (FitzPatrick, 1932). Hence, FitzPatrick was the precursor establishing a dependence between a firm's individual characteristics and its default probability.

In 1935, Raymond Smith and Arthur Winakor analyzed 183 bankrupt firms by analyzing the evolution of 21 financial ratios before they gone bankrupt. They found, that the net working capital to total assets ratio was among the most accurate indicators of a firm's failure (Smith, 1935).

In 1936, Ronald Fisher formulated the concept of Discriminant Analysis, using a linear combination of continuous variables in order to predict a categorical outcome (Fisher, 1936). Afterwards, in 1941, David Duran realized one of the first individual credit scoring models using discriminant analysis to segregate good from bad applicants using data from banks and other financial companies (Durand et al., 1941).

In 1942, Charles Merwin analyzed 939 small manufacturing firms, divided between successful and failing firms, during 1926 - 1936. Merwin found that there were three significant indicators in determining a firm's successfulness, those were the current ratio, the net working capital to total assets and the net worth to total debt (Merwin et al., 1942).

In 1945, Walter Chudson studied the patterns of corporate financial structures, finding that there was no monotonic financial structure among the firms analyzed. Even though he found that a firm's cash balance was related to a firm's industry and that profitable firms are likely to hold higher cash balances in proportion to their assets (Chudson et al., 1945).

In 1962, Nathaniel Jackendorff, compared the financial ratios of profitable and unprofitable firms, noticing that the current ratio and the net working capital to total assets ratio were consistently higher for profitable firms than for unprofitable firms, while debt to equity ratio was consistently lower for profitable firms (Jackendorff, 1962).

Additionally, the 1960's brought severe enhancements in the use of quantitative techniques to assess a firm's creditworthiness. This was due to several factors like the popularization of credit cards which severely increased the size of the population using credit, thus increasing the need for effective automated individual credit assessment models, which at the same time were extrapolated or gave insight to develop new methods to assess a firm's creditworthiness, specially in the ambit of SMEs[1].

In 1963, James Myers and Edward Forgy compared regression and discriminant analysis in individual credit scoring applications, finding that results were quite similar among both models, albeit discriminant analysis generated a better separation of groups at lower score levels, significantly diminishing potential losses. Furthermore, they affirmed that credit scoring acts as superior predictor than any qualitative expert's judgment (Myers & Forgy, 1963).

In 1966, William Beaver performed the first univariate analysis of corporate default risk,

[1] Small and Medium-sized enterprises

comparing the mean values of 30 financial ratios of 79 failed and 79 non-failed firms from different industries. He tested individually the ratios predictive power, finding that net income to total debt, net income to sales, net income to net worth, cash flow to total debt and cash flow to total assets were the variables with highest predictive power. Additionally, he suggested that the simultaneous use of multiple of those ratios could yield higher predictive power than by using them individually (Beaver, 1966).

Subsequently, in 1968, Edward Altman proceeded to perform the first multivariate analysis of corporate creditworthiness. He used multivariate discriminant analysis in order to predict manufacturing firms' bankruptcy (he used data for 33 pairs of bankrupt and non-bankrupt firms during 18 years), this model was called the Altman's Z-Score, since it classified the quality of any firm by assigning what he called a Z-score to it (Altman, 1968).

In 1971, Robert Edmister analyzed the use of financial ratios as discriminant predictors of small business failures, concluding that no single variable is able to predict failure as well as a small group of variables. Furthermore, he states that those additional variables must explain characteristics previously ignored by the current variable(s) (Edmister, 1971).

In 1977, Daniel Martin, first applied the logit regression in order to predict bank failure using past financial data. He analyzed 5,598 banks, 23 of which failed (Martin, 1977). Additionally, in 1980, James Olson applied logit regression to predict business failure, in order to avoid some of the constraints of the multiple discriminant analysis discriminant analysis, such as the requirement of identical variance-covariance for bankrupt and non-bankrupt firms, the requirement of normally distributed predictors and the lack of predicted default probabilities (Ohlson, 1980).

In 1975, Maurice Joy and John Tollefson criticized the methodology followed by the previous studies using Discriminant Analysis. They suggested that a model's predictive power should be evaluated using rigorous validation techniques and using data from a future period in the case of time series data (if the model was trained with data from moment t, the data used to evaluate the model should be from t + 1). Since previous studies were merely using a hold-out sample from the original sampled period, or even reclassified the sample used to estimate the parameters, in order to evaluate a model's predictive ability, thus optimistically biasing the estimated model's predictive power. Additionally, they also pointed that many of the actual studies introduced financial rations that had common numerators or denominators, hence introducing multicollinearity (Joy & Tollefson, 1975). Even though, in 1978, Edward Altman and Robert Eisenbeis clarified that multicollinearity is not a major problem (except in some particular cases) for Discriminant Analysis (Altman & Eisenbeis, 1978).

In 1984, Mark Zmijewski applied the first probit model to estimate financial distress prediction, he selected 3976 firms during 1972 - 1978, 96 of which failed. Additionally, Zmijewski pointed to another problem incurred by the majority of problem research on the area or financial distress prediction models: the estimation of models in nonrandom samples involving choice-based sample biases and sample selection biases (Zmijewski, 1984).

In 1988, due to the popularity acquired by expert systems, William Messier and James Hansen applied a decision tree learning algorithm (ID3) to predict loan defaults and bankruptcies, over-performing the results obtained using discriminant analysis (Messier Jr & Hansen, 1988).

Furthermore, in the 1990s the areas of machine learning and artificial intelligence gained vast popularity, this brought the popularization of sophisticated methodologies such as artificial neural networks. Moreover, by 1990 the competitive benefits of automated credit scoring and credit ratings were clear, they reduced the cost and limited the time of credit assessment. Additionally, in 1992 a new way to create nonlinear classifiers using Support Vector Machines was proposed by applying the kernel trick to maximum-margin hyperplanes (Boser, Guyon, & Vapnik, 1992) and in 1995 Corinna Cortes and Vladimir Vapnik proposed the soft margin implementation (Cortes & Vapnik, 1995).

In 1992 Margaret Dwyer compared classical statistical techniques (logistic regression and non-parametric discriminant analysis) and artificial neural network models (back-propagation and counter-propagation) in corporate bankruptcy prediction. Finding that the logistic regression and the back-propagation artificial neural network achieved the best results (Dwyer, 1992).

In 1995, Christopher Lacher et al. implemented an Artificial neural Network model for classifying the financial health of a firm, finding a neural network model to be more accurate than multiple discriminant analysis and, thus, also surpassing the limitations of multiple discriminant analysis (linear separability, multivariate normality and independence of predictive variables) (Lacher, Coats, Sharma, & Fant, 1995).

In 2003, Raphael Amit and Stewart Thornhill analyzed 339 Canadian corporate bankruptcies determining that there are different kinds of business' bankruptcies depending on the business' age: younger firms are more likely to become insolvent if they are not able to establish viable competitive positions before exhausting their initial asset endowments. While older firms are more likely to become insolvent due to obsolescence, losing relevance in a changing competitive environment (Thornhill & Amit, 2003).

In 2004 the Basel Committee developed new regulations and recommendations to the banking system under the Basel II framework, in order to deal with the complexity of the new risk-based rules. This framework brought two alternative ways of computing risk-weighted assets, basis for banks' capital to hold: the standardized approach and the internal ratings-based approach (Haselmann & Wahrenburg, 2016). This brought a new incentive for large banks to develop credit rating systems.

In 2005, Young Ryu and Wei Yue predicted firm bankruptcy using isotonic separation. This method was compared to other methods such as artificial neural networks or decision trees, obtaining better results for short-term bankruptcy prediction (Ryu & Yue, 2005). Furthermore, the same year, Jae Min and Young-Chan Lee implemented one of the first Support Vector Machine models for bankruptcy prediction, finding that Support Vector Machines outperforms Multiple Discriminant Analysis, logistic regression and three-layer back-propagation neural networks (J. H. Min & Lee, 2005).

In 2006, Sung-Hwan Min, Jumin Lee and Ingoo Han implemented an hybrid approach to predict firm's bankruptcy by using 32 financial ratios. This hybrid approach used genetic algorithms and support vector machines to predict a firm's bankruptcy, genetic algorithms were used in order to perform feature selection and to optimize the Support Vector Machine parameters (S.-H. Min, Lee, & Han, 2006). In 2007, Lili Sun and Prakash Shenoy used naïve Bayes Bayesian network models for bankruptcy prediction achieving results comparable to those achieved by Ohlson in 1980 (Sun & Shenoy, 2007).

In 2011, Mu-Yen Chen showed that traditional statistical methods are better handling large datasets without sacrificing prediction performance than intelligent techniques. Additionally, he also showed that by using principal component analysis the number of financial ratios could be reduced substantially in comparison to previous studies still obtaining highly-accurate forecasts. Furthermore, he found that the models that provided the best prediction performance for imminent corporate bankruptcy were decision tree models (C5.0 and CART) (Chen, 2011).

Table 2.1 shows an overview of the input variables considered by some of the studies considered in this section in a chronological manner. As can be seen on this table, the number of variables considered by more recent studies has increased notoriously. Nevertheless, many of the variables considered by early studies on default and bankruptcy prediction are still considered in today's most recent studies in this area.

Table 2.1: Overview of the input variables used by prior research

Study	**Input Variables**
Altman (Altman, 1968)	Working Capital/Total assets, Retained Earnings/Total assets, Earnings before interest and taxes/Total Assets, Sales/Total Assets, Market Value Equity/Total Debt
Martin (Martin, 1977)	Expenses/Operating Revenues, Net Liquid Assets/Total Assets, Loans/Total Assets, Gross Charge-offs/Net Operating Income, Net income/Total Assets, Commercial Loans/Total Loans, Loss Provision/(Loans + Securities), Gross Capital/Risk Assets
Ohlson (Ohlson, 1980)	Total liabilities/Total Assets, Current liabilities/Current Assets, Funds provided by operations/Total Liabilities, Net Income/Total Assets, log(total assets/GNP price-level index), Working Capital/Total Assets, binary variable indicating whether net income was negative for the last two years, binary variable indicating whether total liabilities exceeds total assets, change in net income
Ryu and Yue (Ryu & Yue, 2005)	Cash Flow/Total Assets, Cash/Sales, Cash Flow/Total Debt, Current Assets/current liabilities, current assets/total assets, Current assets/Sales, Earnings before Tax and Interests/Total Assets, Net Income/Total assets, Total Debt/Total Assets, Sales/Total Assets, Working Capital/Total Assets, Working Capital/Sales, Quick Assets/Total Assets, Quick Assets/Current liabilities, Quick Assets/Sales, Market Value of Equity/Total Capitalization, Cash/Current Liabilities, Current Liabilities/Equity, Inventory/Sales, Equity/Sales, Market Value of Equity/Total Debt, Net Income/Total Capitalization
Mind and Lee (S.-H. Min et al., 2006)	Growth rate of tangible assets, Ordinary income/total assets, Net income/total assets, Ordinary income/stockholders' equity, Net income/stockholders' equity, Ordinary income/sales, Net income/sales, Variable costs/sales, EBITDA/sales, Depreciation ratio, Interest expenses/total borrowings and bonds payable, Interest expenses/total expenses, Interest expenses/sales, Net interest expenses/sales, Interest coverage ratio, Break-even point ratio, Stockholders' equity/total assets, Cash flow/previous year's short term loan, Cash flow/short term loan, Cash flow/total loan, Cash flow/total debt, Cash flow/interest expenses, Fixed ratio, Fixed assets/stockholders' equity and long-term liabilities, Total borrowings and bonds payable/sales, Total assets turnover, Stockholders' equity turnover, Capital stock turnover, Operating assets turnover, Fixed assets turnover, Tangible assets turnover, Inventories turnover, Payables turnover, Gross value added/total assets and productivity of capital, Gross value added/property plant and equipment, Gross value added/sales, Solvency ratio, Ordinary income/ordinary expenses

Due to the increasing amount of variables associated to business' creditworthiness, in 2013, Petr Hajek and Krzysztof Michalak performed a Feature selection analysis in corporate credit rating prediction, comparing a selection approach based on wrappers and another based on filters. Finding that the use of wrappers outperformed the use of filters in this problem (in terms of model accuracy). To do so they used two datasets one for US (containing 852 companies and 81 variables) and one for Europe (containing 244 companies and 43 variables). Furthermore, for each variable they used the mean values calculated over the 2006-2008 period, in order to follow a though-the-cycle approach (Hajek & Michalak, 2013). As a side note, this study predicted directly corporate ratings defined by credit agencies, studies considered previously predicted business default and bankruptcy. Although conditionants to both concepts are strictly similar (if not the same depending on the definition of the rating used).

2.2 Limitations of Present Studies

Methods applied to corporate bankruptcy and default prediction can be criticized by several aspects. The nature of this kind of problems, where only a minority of businesses default or goes bankrupt, generates classes imbalances in the data, which have not been typically considered leading to classifiers biased towards the majority class. Furthermore, the recurrent use of accuracy as a metric of predictive power leads to over-optimistically results, falling into the accuracy paradox, due to the nature of this problem.

Additionally, current studies have overly-focused on the use of financial ratios as the sole predictor of financial default. While, several studies have also proved the importance of factors not directly related to a businesses finance, such as the business' age (Thornhill & Amit, 2003). Moreover, many of the studies developed in this area only do consider company-specific risks, not including factors associated to their industry or the country to which they operate (in cases where businesses from multiple countries are considered).

As we said, current studies rely on the use of financial ratios in order to predict corporate default, research has found that several financial ratios have leptokurtic and asymmetric distributions (Martikainen, Perttunen, Yli-Olli, & Gunasekaran, 1995; Ezzamel, Mar-Molinero, & Beech, 1987; Mcleay & Omar, 2000).Hence, strong evidence again using linear classifiers under those circumstances is provided, although there is still a heavy use of linear classifiers on default prediction in the presence of those conditions. Furthermore, several studies tend to use financial ratios that have the same denominator or numerator, introducing multicollinearity. Therefore, multicollinearity needs to be controlled in order to not alter the results.

Besides, many of the current studies perform variable selection using data from the same period under study, thusly optimistically biasing the results obtained. Likely, they also perform validation picking the same period, again achieving over-optimistically results, this point was already criticized in 1975 by Joy and Tollefson (Joy & Tollefson, 1975) as we previously reviewed.

Another important problem in default prediction (this problem is not present in bankruptcy prediction) is the existence of default under-reporting, in order to preserve a client's reputation since the borrower knows that soon he will be able to pay his obligations. Nevertheless, this problem is easily to solve by acquiring a sufficiently large dataset. Moreover, previous studies on default prediction tend to misinterpret the concept of default, confusing default as bankruptcy. One recent example of this misunderstanding can be seen in a study done by a team from the Valencia Polytechnic University (Sanfeliu, García, Martínez, & Clemente, 2013).

2.3 Summary

In this chapter we performed a review of all the literature that has contributed on the elaboration of methods to assess the creditworthiness of a firm. We have seen how the methods applied to assess a firm's creditworthiness have evolved along time driven by several factors such as the boom of credit cards, the boom of intelligent systems or the establishment of new banking regulations as Basel II.

As we saw formal analysis on bankruptcy analysis began in 1932, from this year to the end of the 1960s the analysis of a firm's creditworthiness was focused on the determinants that helped to explain the successfulness of a firm. But at the end of 1960s an important amount of literature on this topic appeared, which was focused on using statistical techniques, especially discriminant analysis and logistic analysis, to predict a firm's default or bankruptcy. Finally, at the end of 1980s the upswing achieved by machine learning and artificial intelligence, brought the implementation of techniques developed by those areas in several fields, such as default and bankruptcy prediction. This brought the implementation of decision tree models, artificial neural networks and support vector machines to predict corporate default and bankruptcy.

Additionally, it is worth mentioning that the current studies on default and bankruptcy prediction recurrently incur in some deficiencies in their implementation, like the under-representation of defaults or bankruptcies, the lack of consideration of country (in the case of considering several countries) and industry dimensions, etc. Therefore, those deficiencies have been taken into account on the present study, taking steps to avoid incurring on those mistakes, as will be explained in the following chapters.

3 Data

In Chapter 1 we analyzed the theoretical background that must be considered in order to evaluate a firm's creditworthiness. Subsequently, in Chapter 2, we analyzed the main contributions done in the elaboration of statistical and machine learning models for corporate creditworthiness assessment, also pointing their main deficiencies. In this chapter, by considering the theory exposed in chapter 1 and the considerations and limitations reported by previous research, we define the data sources and the exploratory variables considered in order to evaluate a company's creditworthiness.

This chapter provides an explanation of the criteria used in selecting the observations considered to elaborate this study. Additionally, it describes the initial set of variables considered in order to further develop models to predict the risk that a company will run out of money not being able to repay its financial obligations (i.e. default), considering some variables not contemplated other studies, which could also act as a good measure to predict default. Furthermore, previous studies were centered on predicting default by only considering business-specific risks, even though we will also consider other types of risk in order to generate more complete models.

Finally, this chapter also exposes the use of pre-processing techniques, in order to control the quality of the gathered data, and also exposes some of the main characteristics of the gathered data.

3.1 Data Sources

The data used in this study was retrieved from two databases: (1) Sistema de Análisis de Balances Ibéricos (SABI) and (2) Organisation for Economic Co-operation and Development (OECD) data. From SABI we extracted all the data related to a company's information (i.e. all variables from Table 3.1, except V29). While, from OECD data we extracted information related to the country in which the business is located.

In order to select the companies to analyze in this study, we selected all the Spanish companies that had formally defaulted in the last 18 years from the SABI database. Thus obtaining a cross-sectional dataset as suggested by Altman (Altman & Eisenbeis, 1978), in order to overcome the problems introduced by considering only a single moment of time as Joy and Tollefson showed (Joy & Tollefson, 1975). We considered that a company committed default either if it started suspension of payments proceedings or insolvency proceedings.

Additionally, for each financial variable we calculated a three year average mean, using data from the three previous years to the submission of its suspension of payments proceedings or insolvency proceedings. The reason to consider this approach is to mitigate the effect of the cycle, using an approach called "through the cycle" (Hajek & Michalak, 2013), obtaining a prediction less volatile and less dependent to changes in the business cycle (De Servigny et al., 2004).

After selecting the number of companies that had formally defaulted, we also selected companies that had not defaulted on the previous periods. We also selected a cross-sectional sample of those companies, selecting only those companies from similar sectors and regions to companies that had formally defaulted. Finally, we deleted all the companies which did not provide enough financial information to compute some of the independent variables considered in Table 3.1. All this process is explained in a more detailed way in Appendix B (concretely, sections 1 and 2).

After doing so, we obtained 4,371 companies that had formally defaulted in the years considered, which were paired with companies which had not defaulted. Therefore the final dataset included 8,742 companies.

3.2 Explanatory Variables

As we have seen in the previous chapters, corporate creditworthiness is affected by many determinants. Furthermore, previous studies have used an extensive number of different variables, in order to assess a firm's creditworthiness. For this reason, in contemplation of achieving the best description of a firm's default status, we use as many financial and economic measures as possible. Principally, we include parameters that have already been used and assessed by other authors in previous studies, as well as parameters that have not been considered by other authors but should be considered consistent with the theoretical background on credit rating exposed in Chapter 1.

It is worth mentioning that the majority of the variables used in prior research, as seen in Table 2.1, only point to what we called Business-specific risks in Chapter 1. Hence, those variables do not include country and industry risks, which as we pointed in Chapter 1 they are crucial on determining a firm's creditworthiness. Although, since this study only considers Spanish companies, country risk could be omitted. Nevertheless, since multiple periods are considered (cross-sectional data) we will include the GDP compound annual growth over the time period considered (three previous years) as a proxy for country and macroeconomic risks, thus considering variations in the risks associated to the Spanish economy.

Furthermore, due to the number of businesses contemplated in this study, we directly classified each firm to its main sector by considering the first two digits of its CNAE code, obtaining a good implicit representation of industry-specific risks.

Table 3.1 summarizes the variables initially considered, in order to predict a business' default probability. As can be seen, the variables collected present general information about the businesses, as well as information about the businesses' structure, operations, profitability, liquidity, efficiency and evolution. Furthermore, information about the business and the country and sovereign risks is included by the CNAE code and the Spanish risk premium, respectively, as previously mentioned. Another aspect to comment is that we added some financial ratios that have not been considered by previous literature such as the number of previous administrative claims. Additional information about the initial set of variables is provided in Appendix A, where Table A.1 describes the formulas used to calculate the ratios considered and Table A.2 shows a brief description of every variable considered.

Table 3.1: Initial set of variables considered

Variable	**Code**	**Variable**	Code
Number of employees	V1	Solvency Ratio (%)	V16
Number of directors & Managers	V2	Gearing (%)	V17
Import/Export activity	V3	Staff Costs-to-Operating Revenue	V18
Legal Form	V4	Total Assets per employee	V19
First two digits of CNAE code	V5	Operating Revenue per employee	V20
Return On Capital Employed (%)	V6	Sales Compound annual Growth (%)	V21
Return On Total Assets (%)	V7	Net Income Compound annual Growth(%)	V22
Profit Margin (%)	V8	Long-term Debt Compound annual Growth (%)	V23
Net Assets Turnover (%)	V9	Years after establishment age	V24
Interest Cover (%)	V10	Indebtedness	V25
Collection period (days)	V11	Cash-to-Current Liabilities (%)	V26
Credit period (days)	V12	Return On Shareholders' Funds	V27
Current Ratio (%)	V13	Shareholders' Liquidity Ratio	V28
Liquidity Ratio (%)	V14	Spanish GDP Compound annual Growth (%)	V29
Stock Turnover (%)	V15	Number of previous administrative claims	V30

After further analyzing the data retrieved, variables V22 ad V23 were removed from the initial set of variables considered, due to a significantly high presence of missing values as can be seen in Figure B.4.

Additionally, as observed in Figure B.8, financial ratios (from V6 to V28) do not follow a normal distribution as already pointed by some studies (Martikainen et al., 1995) (Ezzamel et al., 1987) (Mcleay & Omar, 2000), having leptokurtic and asymmetric distributions. For this reason, we applied a logarithmic transformation to those variables (since some of those variables took negative values, we applied this transformation: $sign(x) \times log(|x| + 1)$, where x represents each data point), as suggested by Martikainen, et al. (Martikainen et al., 1995).

3.3 Exploratory Analysis

Before proceeding to create models to predict a firm's probability of default, a general understanding of the data should be provided. Previous sections intended to so by providing general information about the dataset and its composition. Furthermore the current section builds up to this information by providing a brief but detailed view of the main data set characteristics.

Figure 3.1 shows the distribution of the pre-processed variables (i.e. after dealing with missing values and applying a logarithmic transformation to the financial ratios variables) for the data used in this study. As it can be seen, the dataset is composed of a majority of small and medium businesses, which is logical since they represent approximately the 99% of the Spanish businesses.

Additionally, as seen in V3 the mass of businesses considered in this dataset do not import or export. Moreover, as shown in V4 the most common legal form is limited liability company. Besides, it is important to mention that we did not plot the response variable (default) since the companies considered on this study are evenly spitted among defaulting and non-defaulting companies, as we have previously mentioned.

Additionally, comparing Figure 3.1 to Figure B.8, i.e. comparing transformed financial ratios to raw financial ratios' distributions, it is evident that by doing so normality is significantly improved on all the considered financial ratios.

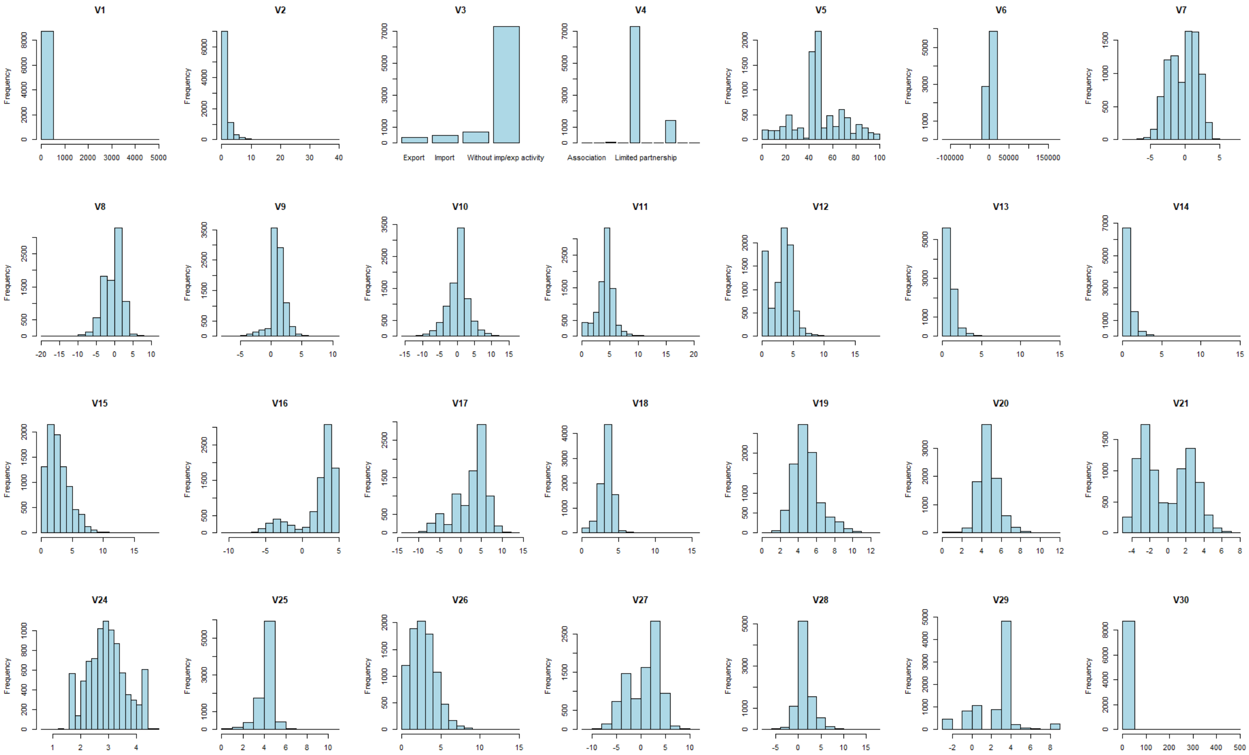


Figure 3.1: Pre-processed Variables Distribution

Likewise, Figure 3.2 shows the density plot of the different variables (except the variables V3, V4 and V5, which are plotted using a bar plot) segregating by their status (defaulting or non-defaulting). As it can be seen by observing the different density plots, variables such as V7, V10, V17, V21, V24, V26 and V29 present some notorious differences depending on the business status. Additionally, variables V1, V3, V4 and V5 corroborate the similar composition in terms of business characteristics (number of employees, import and export activity, legal form and sector) of the different defaulting and non-defaulting businesses.

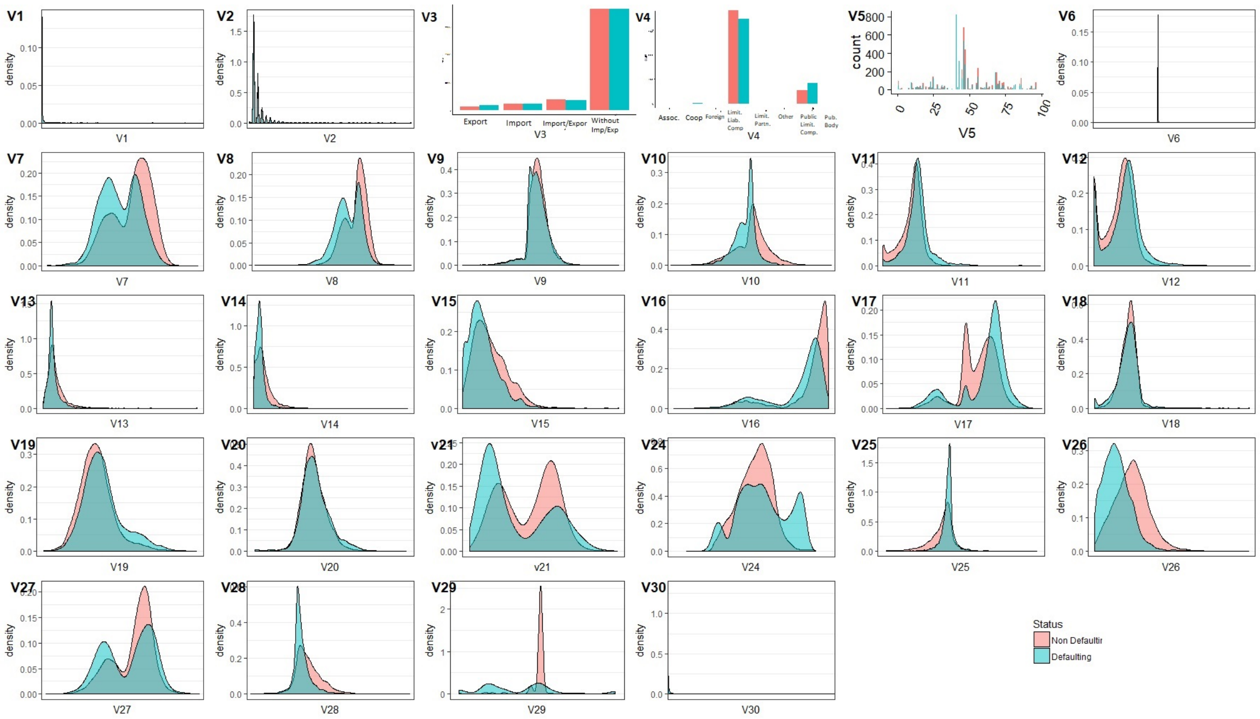


Figure 3.2: Class density plot for the different Variables (Non-defaulting businesses in red and defaulting businesses, overlapping regions mix both colors)

Figure 3.3 shows the Pearson (left) and the Spearman (right) correlation Matrix plot, in order to capture linear and monotonic relationships among variables, respectively. As this Figure indicates, there is a strong monotonic relationship among the different variables considered. Furthermore, almost all the variables are correlated in some way (by observing the Spearman correlation) with the response variable (Default). Additionally, as indicated previously those financial ratios that share a common numerator or denominator show a strong correlation among them. Hence subsequent use of feature selection may be beneficial to posterior fitted model's prediction power.

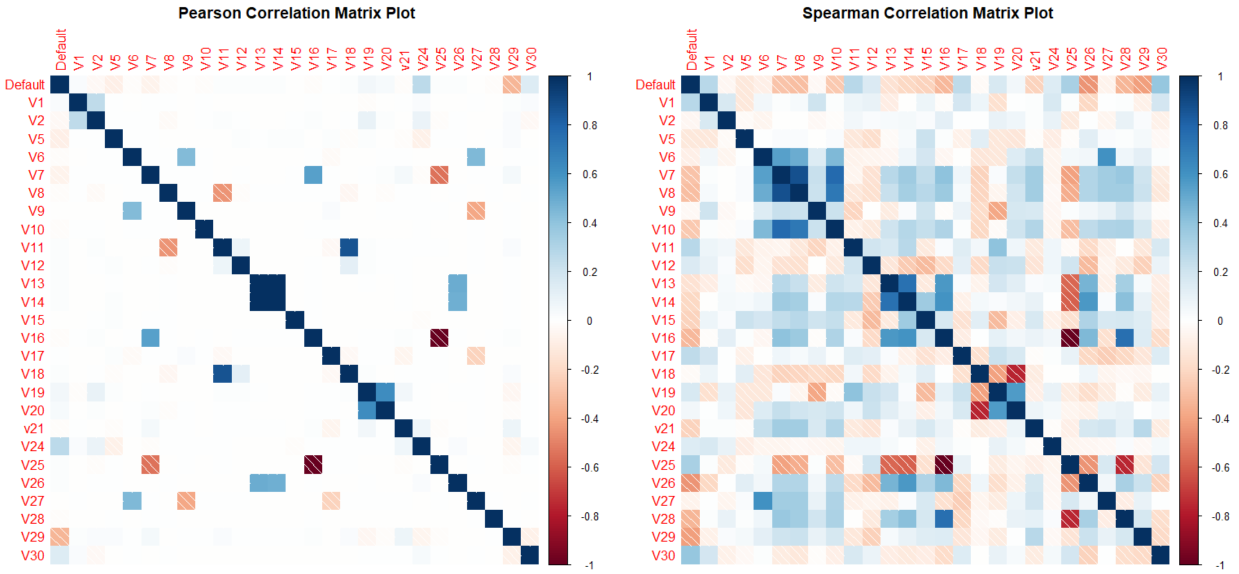


Figure 3.3: Correlation among different variables (including response variable)

Besides, we performed a Student's t-test for defaulting and non-defaulting businesses, in order to test whether the means of each numerical variable are different in the two groups. Hence, Figure 3.4 presents a graphical representation of the data, in which boxplots for defaulting and non-defaulting companies for each numerical variable are shown. Furthermore, the global mean for the variable is indicated with a dashed line, and the p-value resulting from the Student's t-test is shown in the upper left of each graphic.

By looking the boxplots we can identify differences between defaulting and non-defaulting companies. Nevertheless, for many of the variables there is some overlap between the two classes. Although, by examining the p-values associated to the Student's t-test, we can observe that almost all the numerical variables (except V6 and V9) present significant (95% confidence) differences among defaulting and non-defaulting companies, confirming what we initially intuited by examining the boxplots and the global mean.

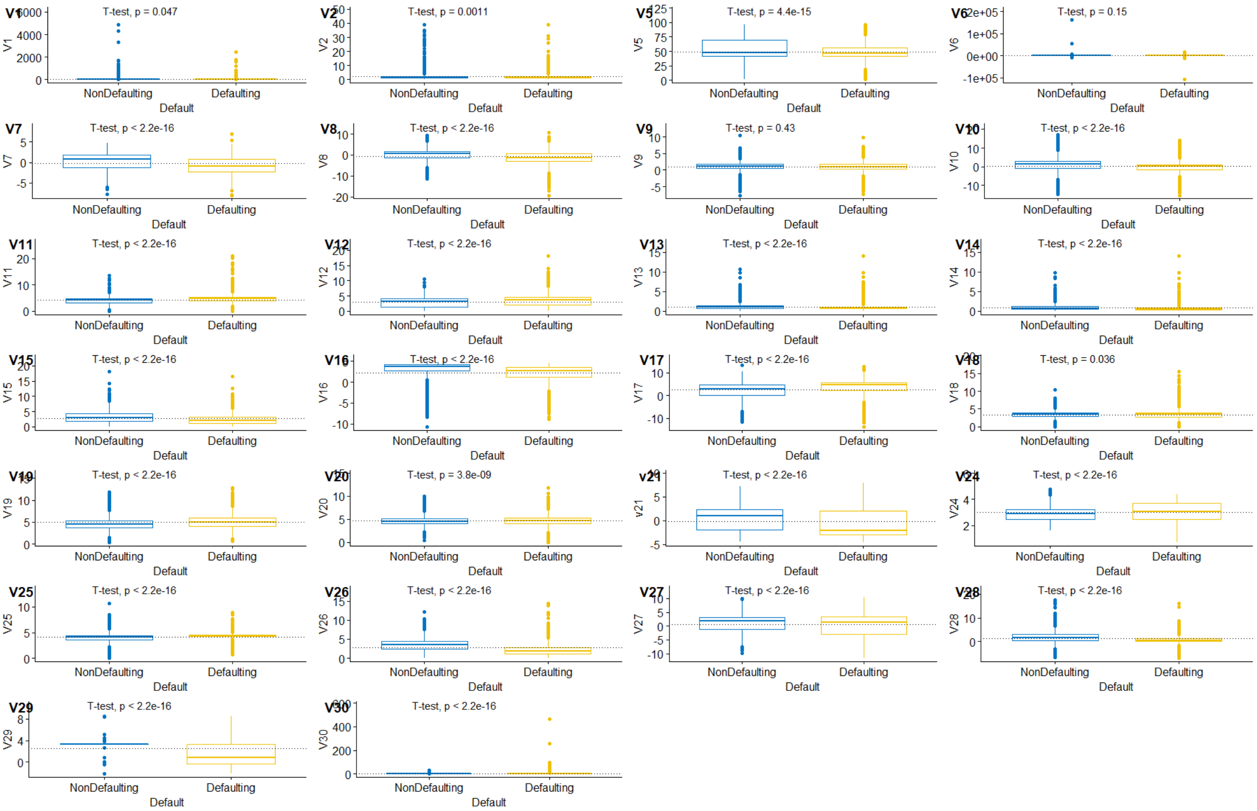


Figure 3.4: Boxplot for defaulting and non-defaulting companies and Student's t-test for each numeric variable

Finally, Figure 3.5 shows the firms representation on the space defined by the first three principal components, which explain approximately the 40% of the total inertia of the data, where defaulting firms are colored in cyan and non-defaulting firms are colored in red. As it can be seen, both classes seem to be quite differentiated, even though in some regions there is an overlap of classes, which may suggest that non-linear methods would achieve superior results than linear methods on this case. Although, due to the lower inertia explained by the first three principal components, this assumption could not be extrapolated in cases that higher dimensionality is added. Therefore linear methods should be tested to safely assess their prediction power.

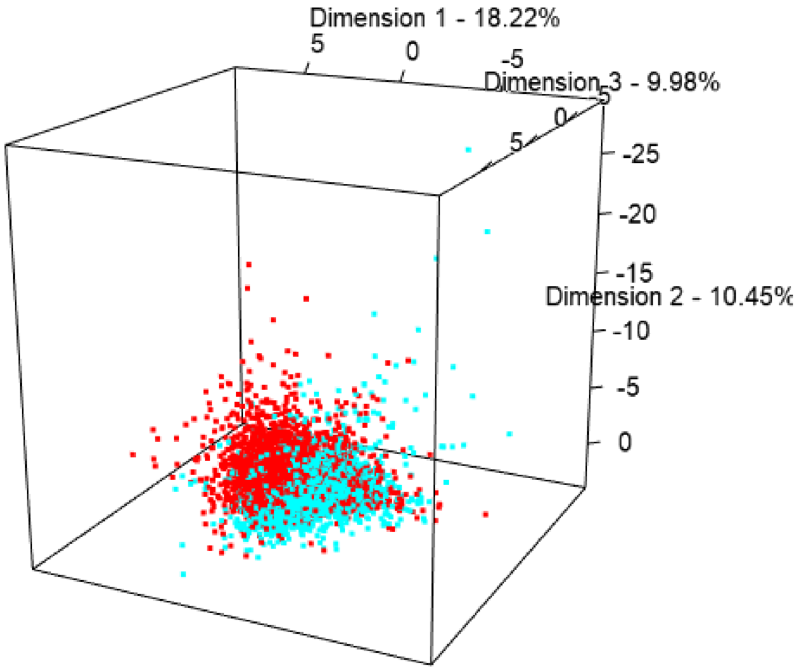


Figure 3.5: Defaulting (cyan) and non-defaulting (red) companies representation on the first three principal components

4 | Modelling

In the previous Chapter we provided a detailed view of the composition of the data used to perform this study, as well to its main characteristics. We exposed various problems regarding the data quality of the retrieved dataset, also explaining the procedures followed to overcome those problems, such as removal of variables and observations with a majority presence of missing values and the presence of non-normal distributions for financial ratios, which may have presented a potential distorting effect on the subsequent models.

Additionally, we explored the variables considered, finding that there exists a strong monotonic correlation between them, likewise among them and the reasons variable. Along with we also showed that the variables considered also present strong association to a firm's default.

In this chapter, we will implement different techniques, in order to find the function that connects the the response variable and the explanatory variables based on the retrieved data. First, we will implement commonly used techniques by previous literature on this kind of problems, like logistic regression and linear discriminant analysis, and then we will implement techniques more typical from machine learning, such as Support Vector Machines and Artificial Neural Networks.

Furthermore, it should be pointed that there exists a trade-off between prediction accuracy and model interpretability (James, Witten, Hastie, & Tibshirani, 2013). Even though, since in this study, our interest is in prediction and not in inference, more restrictive methods will be only preferred in case that they lead to superior or equal prediction power than more flexible techniques.

Moreover, due to the high correlation among explanatory variables, shown in the previous chapter, we apply a wrapper approach (considering that previous studies on this kind of problems have found that wrapper approaches overperform other feature selection approaches, such as filter-based feature selection (Hajek & Michalak, 2013)), in order to perform feature selection. Since, as we mentioned in the previous chapter, we believe that subsets of those variables can yield to higher predictive power.

4.1 Logistic Regression

Logistic regression, also called logit regression or logit model is one of the most commonly used methods by previous studies on default and bankruptcy prediction. It is a regression model that predicts the probability of a dichotomous outcome (defaulting company or not defaulting company, in our case) based one one or more explanatory variables.

Logistic regression is a linear model, which uses the logistic distribution, an S-shaped curve which asymptotically approaches to 0 and 1, in order to transform the predictions. This transformation leads to a loss of understanding of the predictions as a linear combination of the inputs as can be done in the linear regression.

In order to proceed to train a logistic regression in our data, first we applied Recursive Feature Elimination (RFE) with 10 fold Cross Validation (10 - CV), in order to select the best subset of variables. As it can be seen in Figure 4.1, the best accuracy[1] for the logistic regression is achieved by using 21 predictors (V1, V5, V7, V8, V10, V11, V12, V14, V15, V16, V17, V18, V19, V20, V21, V24, V25, V26, V28, V29 and V30).

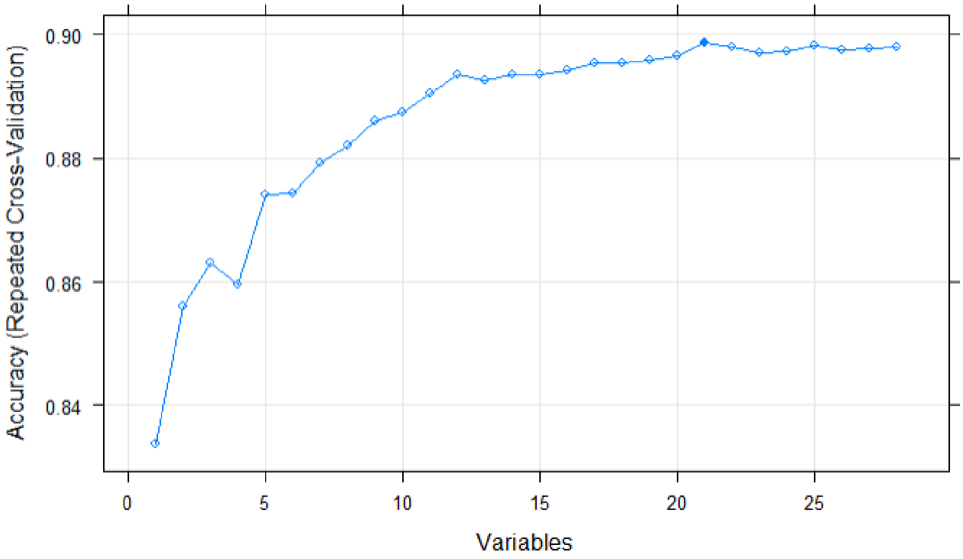


Figure 4.1: Recursive Feature Elimination Results for Logistic Regression (y-axis representing model accuracy and x-axis representing the number of features included)

4.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is another method commonly used in previous bankruptcy and defaulting prediction literature. This method is a generalization of the Fisher's Linear Discriminant analysis, in which a linear combination of features is searched, in order to separate two or more classes. To do so, the distribution of the predictors is modeled separately for each of the response's classes. Then, the Bayes' theorem those is used to generate conditional probabilities according to the different predictors. This approach counts with some advantages over logistic regression (James et al., 2013) such as higher stability in cases in which the predictors are approximately normal or in which the classes are well-separated. Additionally,

[1] this result cannot be taken as an overall performance metric of the model in question, otherwise it should be taken as a relative performance metric, since due to the nature of the method it will be upwards biased. Hence, it is a good measure to compare which predictors are the best, but not to assess overall model's prediction power

this method is more sutiable when there are more than two response classes.

As in the previous case, before proceeding to train a LDA model to our data, we applied RFE using 10 - CV, in order to select the best subset of variables. The RFE results obtained using LDA are shown in Figure 4.2, where we can see that the best accuracy is achieved using 23 predictors (V1, V5, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19, V20, V21, V24, V25, V26, V28, V29 and V30).

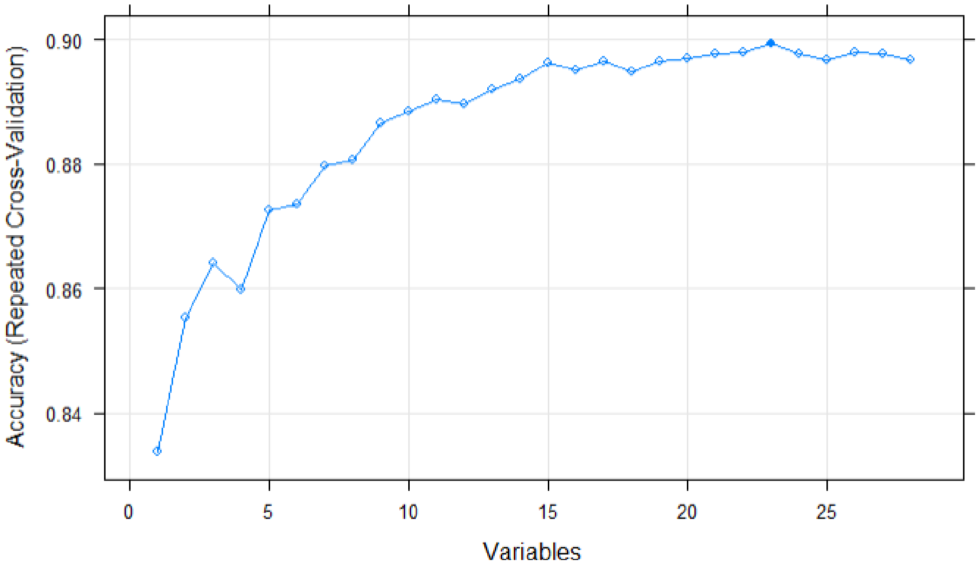


Figure 4.2: Recursive Feature Elimination Results for Linear Discriminant Analysis

4.3 Tree-based Methods

4.3.1 Decision Trees

Decision Trees are a predictive modelling technique used either for regression or classification. Decision trees are built by segmentation the different data points into homogeneous groups, based on their features. Hence, the whole mass of data points considered in the dataset, initially, generate a whole group of data which is subsequently divided by subgroups that meet some specified criteria. Thus obtaining a tree-like shape formed by nodes (groups of data) and branches (links).

In order to divide the different data points on each group a top-down greedy approach is considered (since it would be computationally infeasible to compute all the possible partitions), i.e. an approach which starts from the initial group (the whole dataset), successively splitting the predictor space and that does not generate the best possible result, since it searches for the best split at each particular step, without considering what splits could lead to a better tree in future steps. And this process continues until a stop criterion is reached, such as no region contains more than ten observations.

In order to define the best split, concepts such as the Residual Sum of Squares (RSS) for regression problems and the Classification Error Rate for classification problems are considered.

The splitting process just mentioned is likely to overfit the data, since the final model generated will be too complex and will mimic too strictly the training data. For this reason, a common strategy to solve this problem is applied by building a really large tree, which then is pruned to obtain a subtree. Since considering the prediction power of each possible subtree is infeasible, greedy approaches, such as cost complexity pruning are used.

Decision trees present several advantages (James et al., 2013) like their ease of understanding and communication, due to their intuitive graphical representation. Even though trees do not have the same level of predictive power as other classification approaches. Although, by aggregating several decision trees their prediction power can be enhanced.

In this case, as done previously, to start training a decision tree model (using CART algorithm), first we applied RFE using 10 - CV a subset of 27 predictors (only V3 was excluded from the optimal subset of predictors) were selected as the best one, as shown in Figure 4.3.

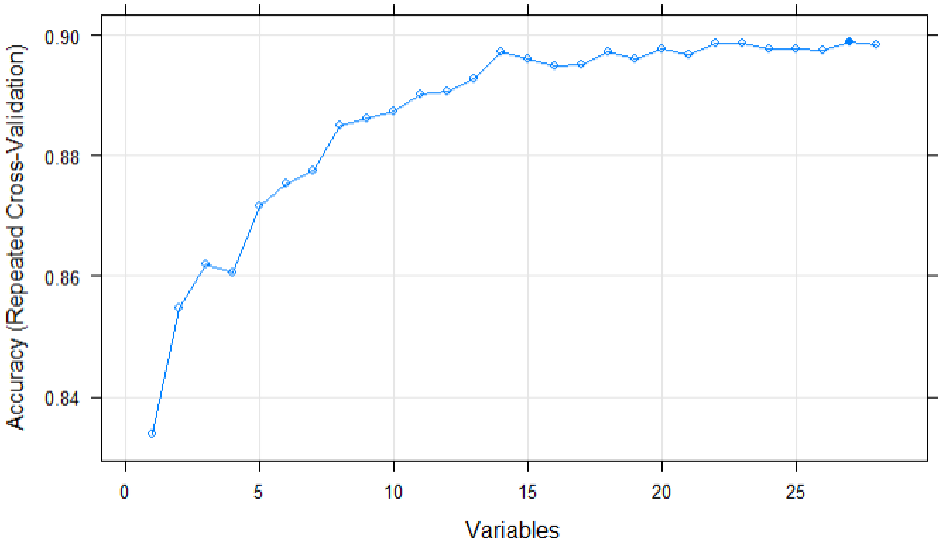


Figure 4.3: Recursive Feature Elimination Results for Decision Tree

Secondly, in order to train a Decision tree model, unlike in the case of the previous models, a complexity parameter must be defined. The complexity parameter penalizes larger trees, by preventing overfitting. Setting the complexity parameter to 0, would imply the creation of the largest decision tree. For this reason, we proceeded to perform a tuning process, in which several complexity parameter values were tested using 10 - CV. The results are shown in Figure 4.4, where accuracy is plotted against the different values tested for the complexity parameter. s shown in this Figure the optimal complexity parameter is around 0.001 (concretely, 0.001143903).

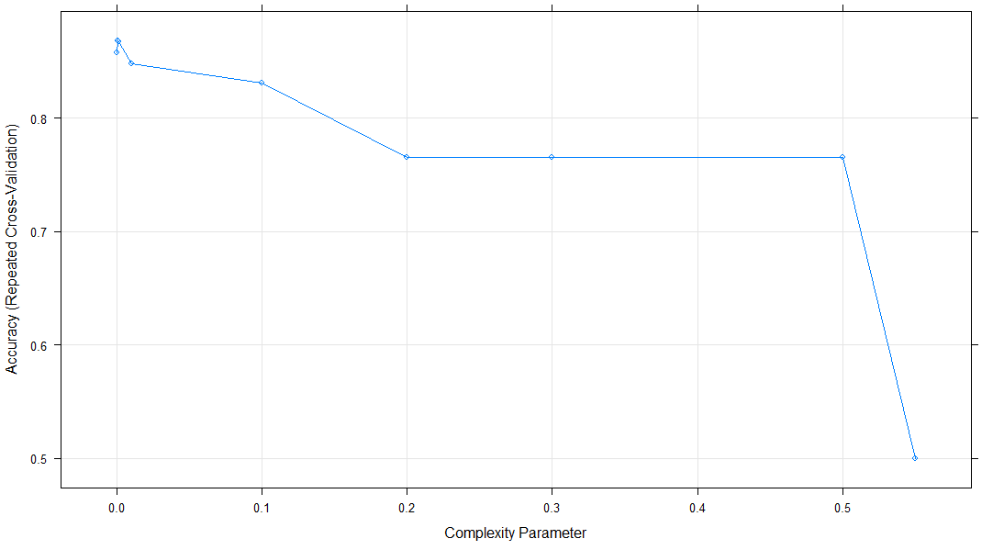


Figure 4.4: Parameter Tuning for Decision Tree

Figure 4.5 shows the representation of the optimal decision tree, nodes painted dark green contain a majority of non-defaulting businesses, while nodes painted with dark blue contain a majority of defaulting businesses (the proportions of defaulting and non defaulting businesses are shown in the middle of each node, being the proportion on the right for non-defaulting businesses and the proportion on the left for defaulting businesses). Additionally, the proportion in the bottom of each node, represents the percentage of businesses respect to the total found on that node. Finally under each node we find a split, which is determined by some condition (subsequent nodes on the left side fulfill that condition, while on the right side do not).

Consequently, by examining the different nodes and splits from the top to the bottom, we can have a clear vision of the different characteristics in which defaulting and non-defaulting firms. Broadly, we can observe that non-defaulting firms benefit from more positive values on their financial indicators.

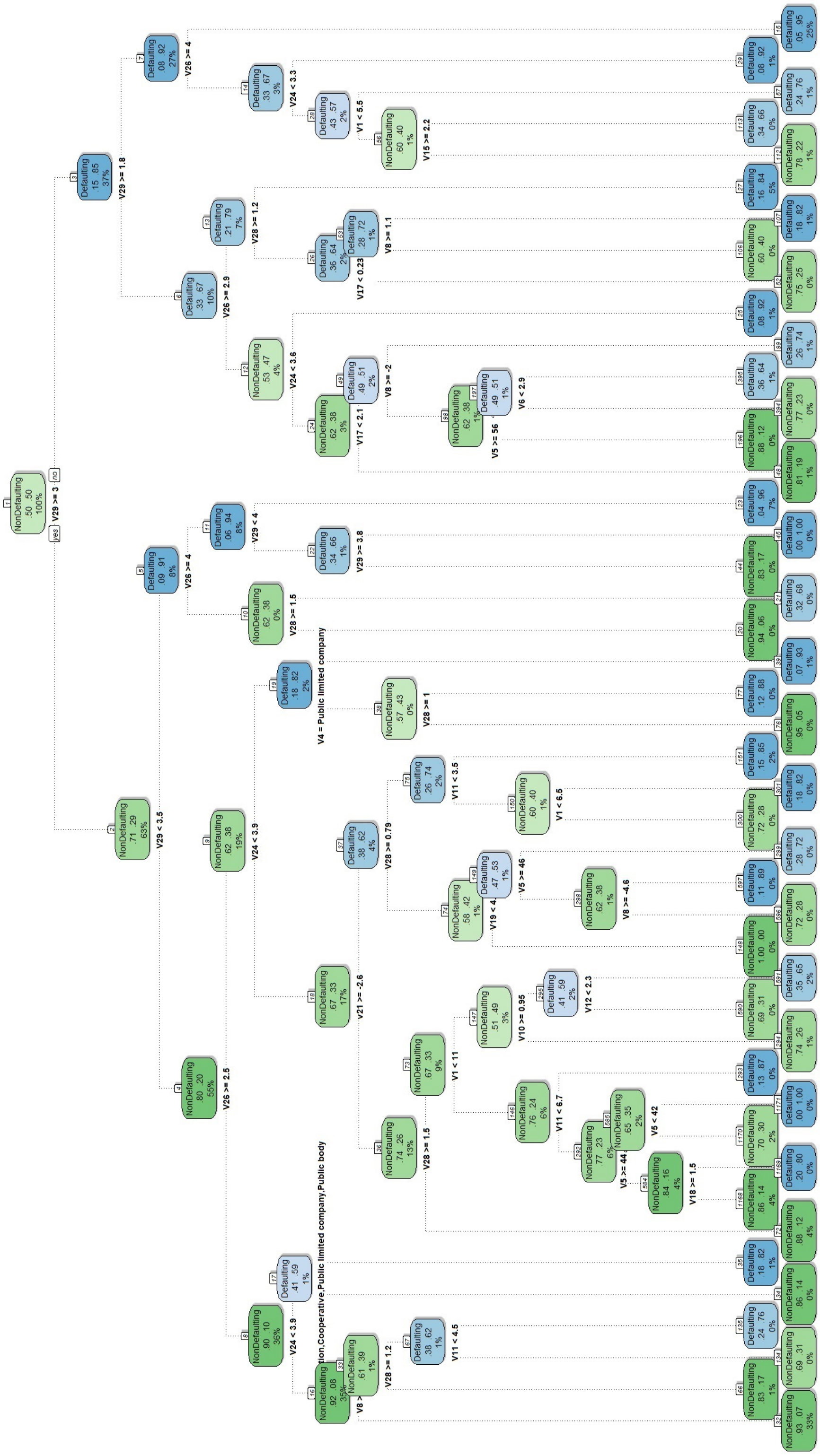


Figure 4.5: Decision Tree Plot

4.3.2 Random Forest

As mentioned earlier, decision trees' prediction power can be enhanced by aggregating several decision trees. One way to do so is by using Random Forests.

Hence the idea behind Random Forests is to build a number of decision trees on bootstrapped training samples. Additionally, in order to prevent correlation among the different built trees, for each time a split is considered we select a random subsample of predictors. This process, which can be thought as decorrelation of the trees(James et al., 2013), makes the average resulting tree less variable and more reliable.

This process of decorrelation is done because in case that there was one relatively strong predictor, all the aggregated trees would use that predictor as top split, being all of them quite similar and that introduces notorious correlation.

As in the previous cases, in order to proceed to train a Decision Tree model, first we applied RFE using 10 - CV, finding that the best subset of predictors is composed of 26 predictors (maintaining the same predictors as in the Decision Tree case, except V2. Therefore, V2 and V3 are excluded in this case), as shown in Figure 4.6.

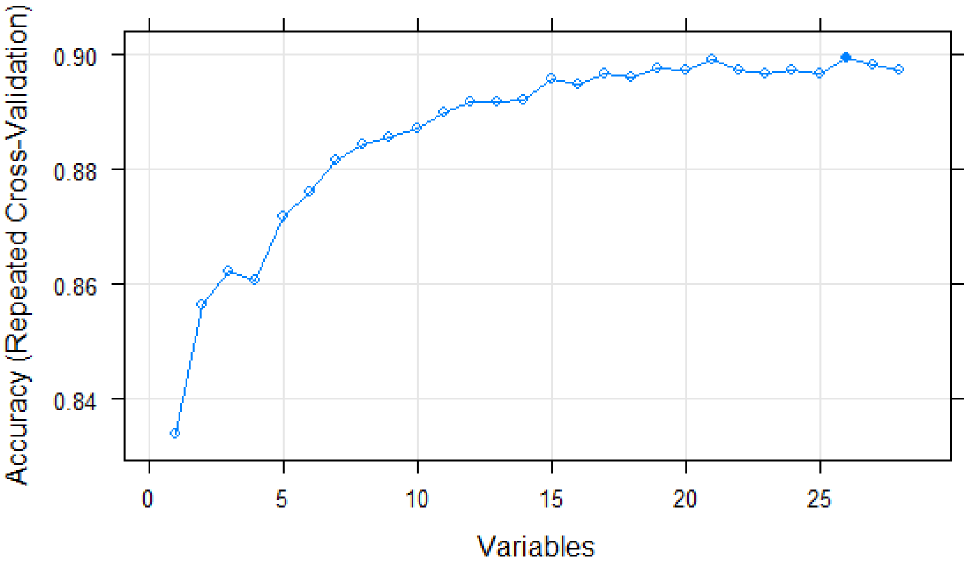


Figure 4.6: Recursive Feature Elimination Results for Random Forest

Afterwards, we proceeded to select the best number of trees to grow and the number of variables randomly sampled as candidates at each split by testing several combinations of those hyperparameters using 10 - CV. Figure 4.7 shows the results obtained using the different combinations of hyperparameters tested, and as shown in this Figure the best accuracy is achieved by using 2000 trees and randomly sampling 3 variables as candidates at each split.

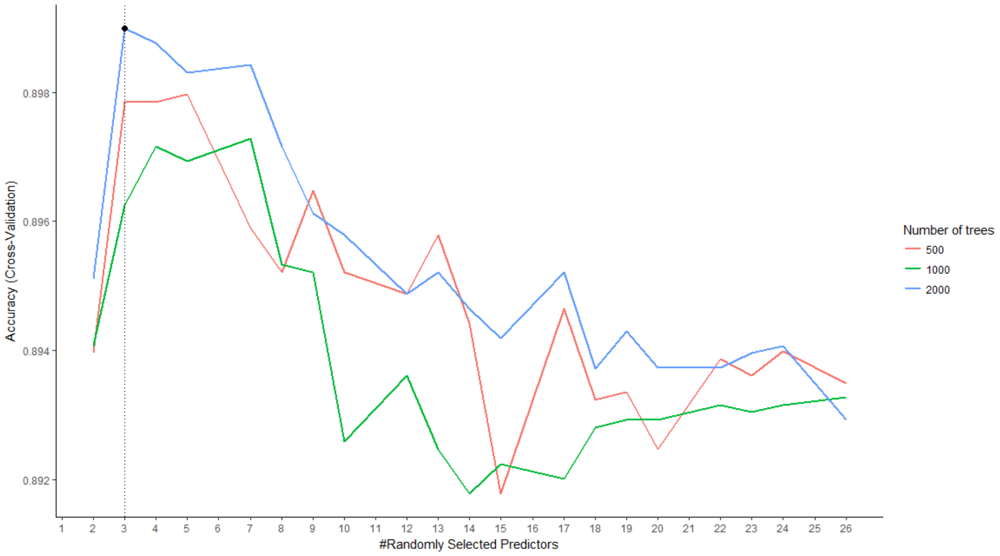


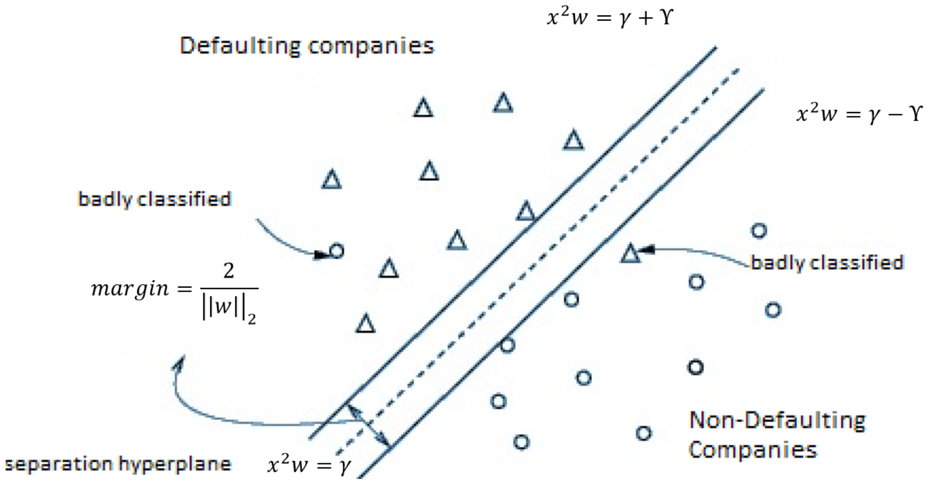
Figure 4.7: Parameter Tuning for Random Forest

4.4 Support Vector Machine

The Support Vector Machine (SVM) classifier is a generalization of a simple classifier, the maximal margin classifier. The maximal margin classifier is based on the idea that if our data can be perfectly separated using a hyperplane, then there will be an infinite number of such hyperplanes. Although, this method is based in the idea that the optimal separating hyperplane will be that one that is farthest from the training observations, this hyperplane is called the maximal margin hyperplane.

As we said the maximal margin classifier only works in the case in which the data is separable by a hyperplane, in this case the addition of a single observation can lead to dramatic change in the maximal margin hyperplane. For this reason, a generalization of this method was developed, which considered a classifier based on a hyperplane that does not perfectly separate the two classes, this is the case of the SVM classifier, and the fact that his classifier does not perfectly separate the two classes provides higher robustness to individual observations and a better classification of the test data on most of the cases.

The main goal of an SVM is is to maximize the distance between the hyperplane which separates the different classes and the training observations, while adding some flexibility to missclassify some points. This flexibility is regulated by the C parameter, which controls the bias-variance trade-off. When C is large the margin will be wide and many observations will violate the margin, i.e., there will be many support vectors. In this case the variance will be low, since changing one observation will not cause greater change on the model, even though the bias will be high. Figure 4.8 depicts the geometric interpretation behind the Support Vector Machine Classifier. Additionally, by using kernel functions nonlinear spaces can be transformed into linear spaces, thus also being able to appropriately classify classes that cannot be linearly separable a priori.



Source: Adapted image from Castro, J. (2017). Optimization Techniques for Data Mining notes.

Figure 4.8: Support Vector Machine Classifier.

Due to the importance of the kernels used in Support Vector Machine methods, we tested two different kernels: Linear and Radial basis function kernels. The following subsections show the optimal features and hyperparameters selected for each of those kernels.

4.4.1 Linear Kernel

The linear kernel represents the simplest kernel function and it is given by the inner product <x, y>, also an arbitrary constant can be added to this inner product.

Following the same approach as in the previous cases, first we performed selected the best subset of variables using RFE, as shown in Figure 4.9, where we can see that a subset of 27 variables (only V9 is discarded) are selected as optimal.

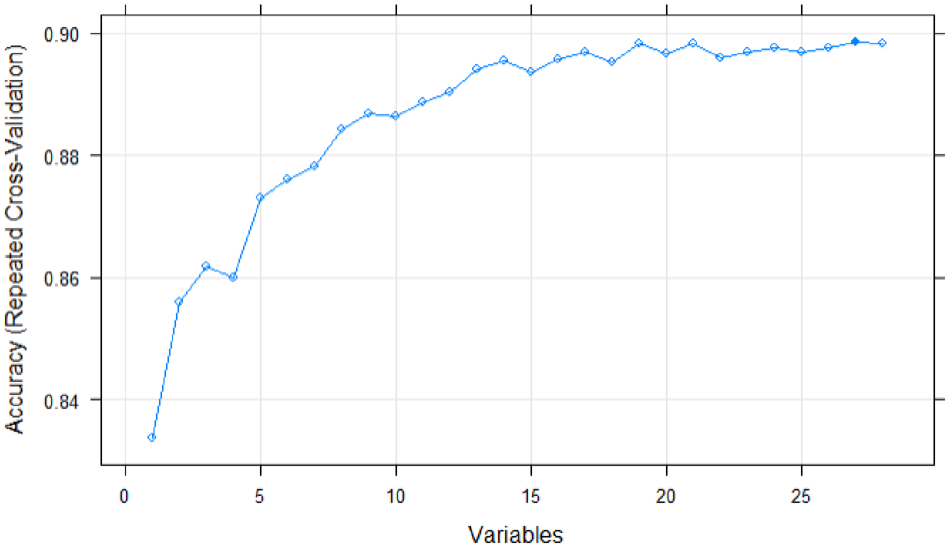


Figure 4.9: Recursive Feature Elimination for SVM - Linear kernel

Furthermore, as previously said, SVM allow some flexibility for misclassification by defining the C parameter. Additionally, the definition of this parameter will affect a model's prediction ability. For this reason, we proceeded to tune this parameter, by testing several values of C using 10 - CV, as shown in Figure 4.10, finding that a parameter C equal to 2 yielded the best results.

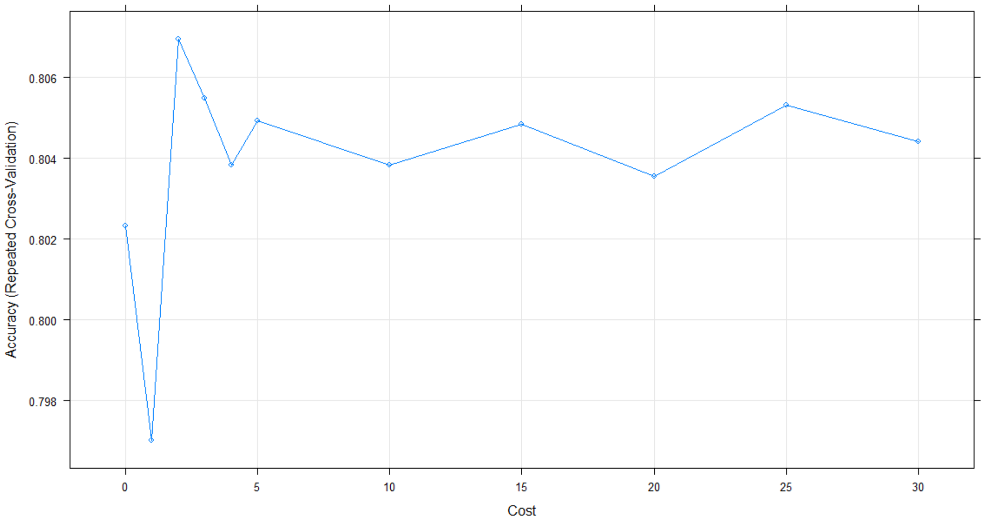


Figure 4.10: Parameter Tuning for SVM - Linear kernel

4.4.2 Radial Kernel

Radial Basis Function (RBF) kernel is one of the most commonly used kernels in machine learning, which is calculated as follows:

$$k(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

where sigma is an adaptable parameter, that plays a crucial role in the performance of the RBF kernel. Since the sigma determination plays a major role in the bias-variance trade-off. Therefore, subsequent tuning to decide its optimal value will be needed.

As in the previous cases, in order to train a SVM with RBF kernel, first we selected the best subset of variables using RFE, which in this case consisted of 24 variables (V1, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19, V20, V21, V24, V25, V26, V28, V29 and V30), as it can be seen in Figure 4.11.

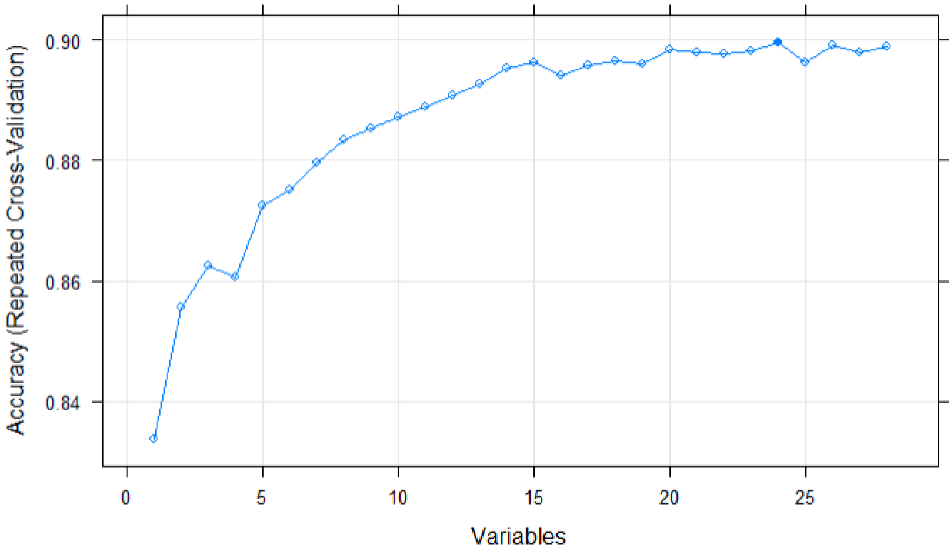


Figure 4.11: Recursive Feature Elimination for SVM - RBF kernel

Moreover, in this case two hyperparameters must be defined: as in the previous case the parameter C must be defined, and because of the use of the RBF kernel, sigma must also be defined. Therefore, Figure 4.12 shows the different accuracy achieved using 10-CV for different combinations of those parameters, seeing that the best accuracy is achieved using a parameter C equal to 4 and a sigma equal to 0.01.

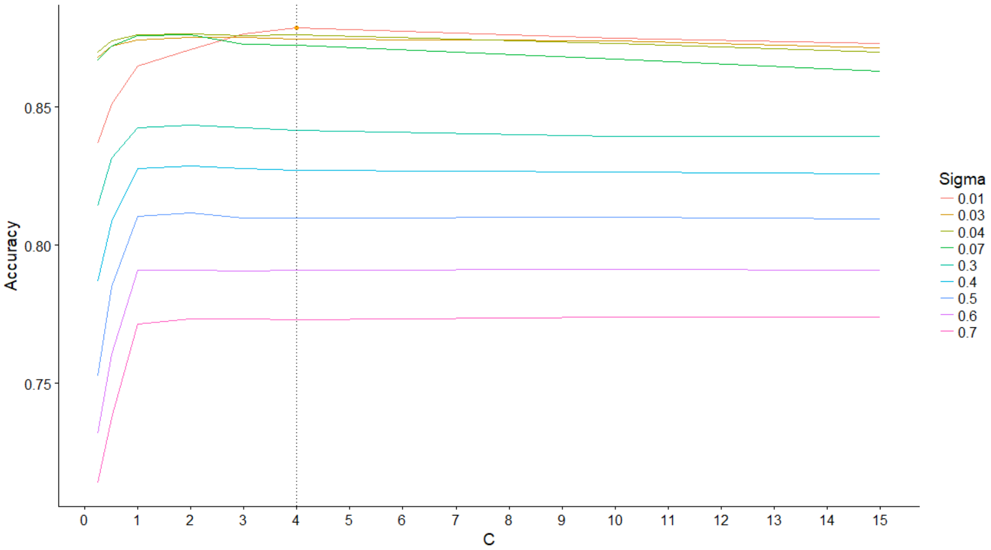


Figure 4.12: Parameter Tuning for SVM - RBF kernel

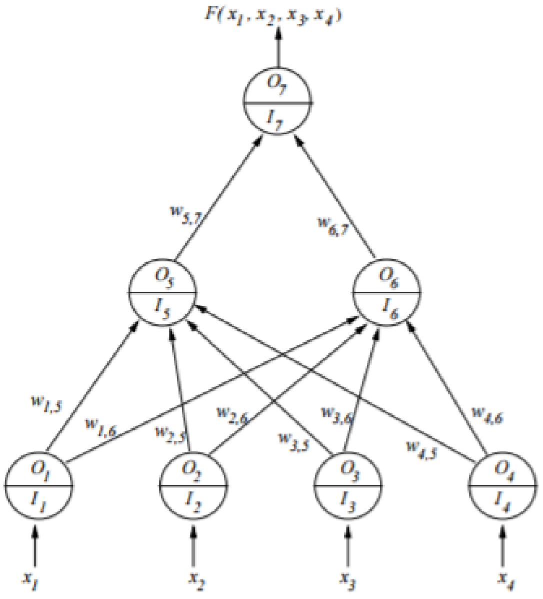
4.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are biologically inspired computing systems, which are based in the human brain structure. As the previous techniques, ANNs "learn" by considering examples.

ANNs are formed from different connected units or nodes called artificial neurons, which are organized in layers. Thus each artificial neuron is connected to some other neuron(s), to which information from one to another is transferred and further processed. Hence, each neuron has an input weight, a transfer function and an output.

Figure 4.13 shows a simple representation of a neural network, which is trained using four predictors. In this representation each circular node represents an artificial neuron, while each arrow represents a connection between two neurons. Furthermore, each circular node representation is divided into two subsections, characterized by a letter I and a letter O, which refer to Input and Output, referring to the fact that each artificial neuron receives and input "signal" which is transformed to an "output" signal, by using a transfer function. Then this output signal is weighted by the corresponding value w and sent to the different connected neurons. Subsequently, the weighted sum of a neuron's inputs constitutes its activation.

Thusly, during ANNs training, inter-unit connections are optimized until the prediction error is minimized.



Source: Castro, J. (2017) Optimization Techniques for Data Mining notes.

Figure 4.13: Artificial Neural Network

In order to train a single layer Artificial Neural Network (we train a single layer ANN, since adding additional layers would exponentially increase computational time, due to the high amount of hyperparameters and decisions to be optimized, without presumptively providing significant increases of performance), as previously done, first we selected the best subset of variables using RFE with 10 - CV. The results obtained for the different subsets of variables are shown in Figure 4.14, in which it can be seen that the best results were achieved using the totality of the predictors.

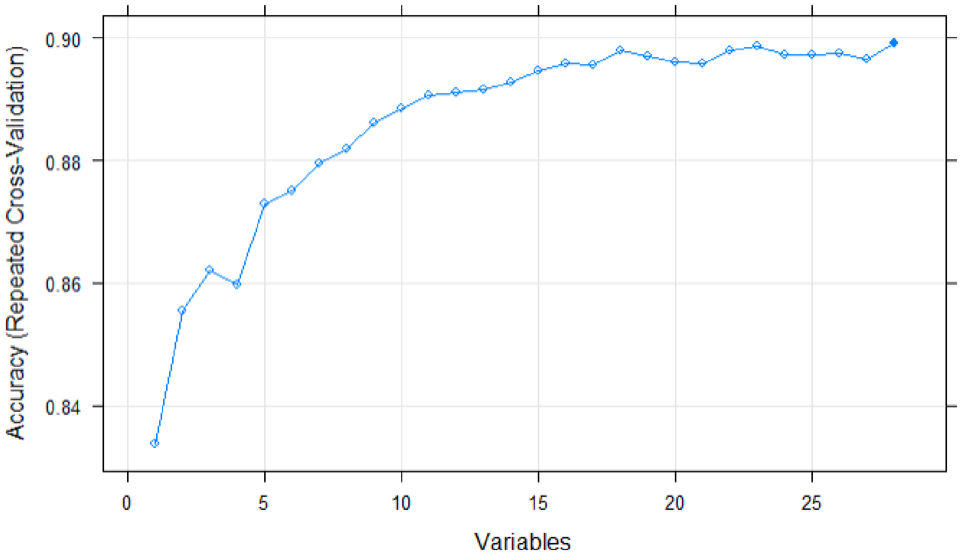


Figure 4.14: Recursive Feature Elimination for ANN

Additionally, in order to model a Neural Network two hypermarameters must be defined. Those are the number of units in hidden layer (since we only fit a single layer neural network) and the weight decay, which acts as a regularization parameter, preventing over-fitting. Figure 4.15 shows the accuracy obtained using 10-CV by using different combinations of those hyperparameters, seeing that the best result is achieved using a weight decay equal to 1 and 10 artifial neurons in the hidden layer.

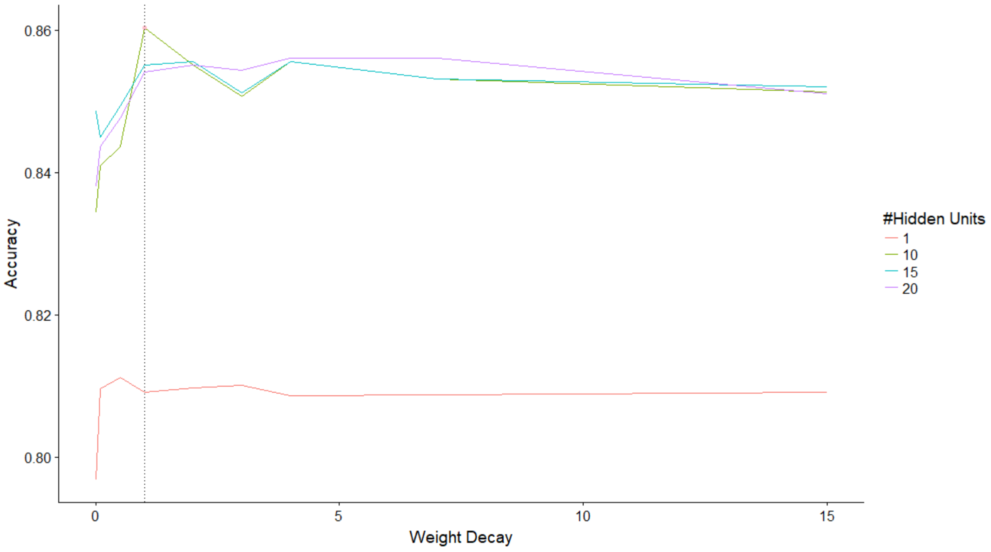


Figure 4.15: Parameter Tuning for Neural Network

4.5.1 Results

In this chapter we briefly analyzed a set of different techniques. Additionally, we selected the best subset of predictors for each of those techniques using a wrapper approach. And, finally, we selected the optimal hyperparameters for each of those models by using a hyperparameter tuning process.

Table 4.1 summarizes the optimal settings for each of the techniques applied on the previous chapter. Furthermore, it shows the main results obtained using those techniques using their optimal settings. As shown in this Table, linear models achieved the highest error rate, being highest for Linear Discriminant Analysis with an error rate of 20.81%. On the other hand, the best results were achieved by Random Forest, achieving an error rate equal to 10.1%, less than half of the error achieved using Linear Discriminant Analysis and almost 25% less than the second best performing method (decision tree). Moreover, as it can be seen in this table some of the previous assumptions, such as likely overperformance of non-linear methods over linear methods and the necessity to select a subset of features, have shown to improve the results obtained.

Additionally, Random Forest also overperforms all the other methods in terms of sensitivity and specificity, identifying almost the 90% (89.60%) of the defaulting companies, being the proportion for non-defaulting slightly better (90.02%). Hence, a significant level of detection of potential defaulting companies has been achieved using a Random Forest model. Those results are in accordance to some of the latest studies done in the field of bankruptcy detection, in which C5.0 and CART algorithms presented the best performance (Chen, 2011). Although, by using Random, results obtained by previous studies using decision trees are easily improved, specially in the case of default prediction.

For all those reasons, from all of the trained models, the Random Forest model is chosen as the optimal model to predict default among Spanish companies.

Table 4.1: Error rate, Sensitivity and Specificity for the different trained models

Model	**Hyperparameters selected**	**Number of Predictors**	**Error rate**	**Sensitivity**	**Specificity**
Logistic Regression	-	21	19.61%	80.80%	80%
Linear Discriminant Analysis	-	23	20.81%	79.65%	78.90%
Decision Tree	complexity parameter = 0.001143903	27	13.11%	87.60%	84.80%
Random Forest	#Randomly Selected Predictors = 3 Number of trees = 2,000	26	10.1%	89.60%	90.02%
SVM - Linear Kernel	C = 2	27	19.31%	81.26%	80.15%
SVM - Radial Kernel	C = 4 and sigma = 0.01	24	12.13%	89.49%	86.37%
Neural Network (1 layer)	size = 35 and decay = 3	28	13.97%	85.83%	86.24%

Therefore, the final set of predictors consists of 26 predictors, which are shown in Table 4.2, where specific sector information and the number of directors and managers are discarded (net income and long-term debt compound annual growth were previously discarded due to severe presence of missing values, as discussed in Chapter 3 and Appendix B).

Table 4.2: Final set of variables

Variable	**Code**	**Variable**	Code
Number of employees	V1	Solvency Ratio (%)	V16
Legal Form	V4	Gearing (%)	V17
First two digits of CNAE code	V5	Staff Costs-to-Operating Revenue	V18
Return On Capital Employed (%)	V6	Total Assets per employee	V19
Return On Total Assets (%)	V7	Operating Revenue per employee	V20
Profit Margin (%)	V8	Sales Compound annual Growth (%)	V21
Net Assets Turnover (%)	V9	Years after establishment age	V24
Interest Cover (%)	V10	Indebtedness	V25
Collection period (days)	V11	Cash-to-Current Liabilities (%)	V26
Credit period (days)	V12	Return On Shareholders' Funds	V27
Current Ratio (%)	V13	Shareholders' Liquidity Ratio	V28
Liquidity Ratio (%)	V14	Spanish GDP Compound annual Growth (%)	V29
Stock Turnover (%)	V15	Number of previous administrative claims	V30

5 Converting probabilities to Letter Ratings

In the previous chapter we analyzed different statistical and machine learning techniques, selecting the best subset of parameters using a wrapper approach for each of those techniques, as well as performing parameter tuning, in order to select the best combination of parameters for each. Finally, we determined that Random Forests using a set of 26 predictors outperformed all the other models, achieving an accuracy of approximately 90%, and we selected it as our final model.

Although, the results obtained from this model are represented as probabilities, while traditional corporate credit rating systems are usually represented by letter ratings. For this reason, in order to make our credit ratings comparable to conventional credit rating systems, and also easily understandable, we extrapolated the default probabilities obtained to letter ratings. Table 5.1 proposes a template to do, based on a three year average default rates of a Standard and Poor's static pool of rated agencies[1]. As it can be seen, a rating of AAA represents the highest rating granted, while a rating of D the lowest.

Table 5.1: Corporate Credit Rating Matrix

Rating	**Default probability (%)**
AAA	[0, 1)
AA	[1, 3)
A	[3, 5)
BBB	[5, 8)
BB	[8, 25)
B	[25, 45)
CCC	[45, 60)
CC	[60, 75)
C	[75, 90)
D	≥ 90

Figure 5.1 shows an example of the corporate credit ratings distribution of the cross-validated predictions using the Random Forest model. Because our dataset balances default-

[1] using Standard & Poor's Risk Solutions CreditPro

ing and non-defaulting classes, Figure 5.1 is negatively skewed. Even though, if we took a representative (in terms of default status) sample of Spanish firms, this distribution would be closer to a normal distribution, in which lower ratings would have a lower proportion of cases, while intermediate ratings would have a higher proportion of cases. Besides, the line shown in this Figure depicts the minimum default probability of each rating (Min DP).

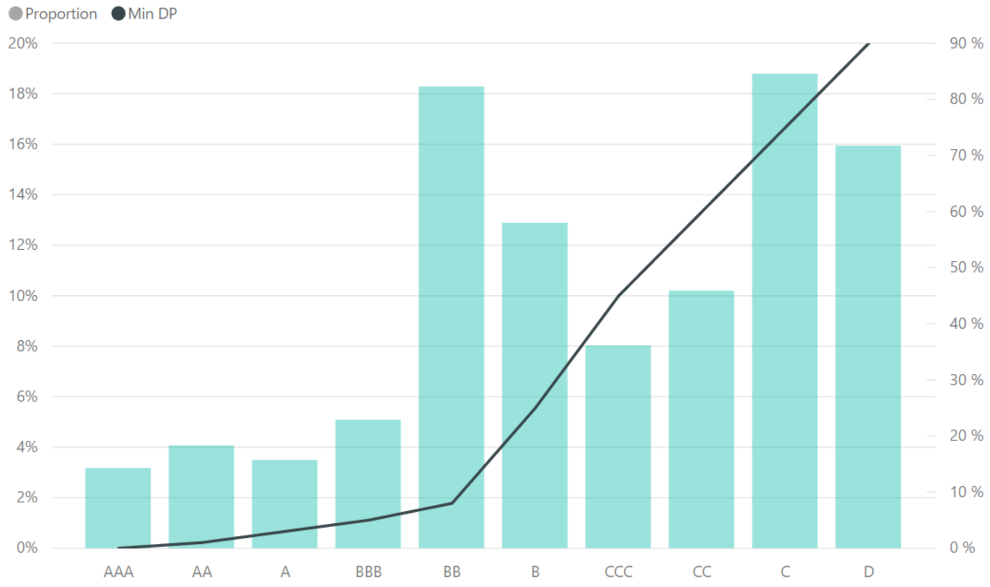


Figure 5.1: Distribution rate and minimum probability for each class

6 | Conclusions

In this Master Thesis we developed an automated corporate credit rating assessment method for Spanish companies, which was proven to be trustworthy, reliable, objective, comparable and free from conflicts of interest. Additionally, the developed method was suitable for all sizes of businesses from SMEs to big enterprises at no cost neither in time nor money.

In order to develop this corporate credit rating assessment method, firstly, we analyzed the theory behind corporate credit ratings, understanding their function and their utility, as well as all the dimensions involved in their determination. Subsequently, we reviewed the current research on bankruptcy and default prediction, listing all the concepts that affected a firm's creditworthiness, available for the majority of businesses (for example we did not consider variables related to a firm's capitalization). Doing so, we selected an initial set of 30 variables, two of which were initially discarded, due to high presence of missing values.

Consequently, taking into account some of the limitations of previous studies on this and related fields, we built a representative cross-sectional balanced data set of Spanish companies over which we implemented several statistical and machine learning techniques, setting the best subset of predictors using a wrapper approach and the best set of hyperparameters for each of them. Afterwards, we compared the prediction ability of each of the models elaborated (taking the optimal set of hyperparameters and subset of features), seeing how decision tree models over-performed the rest, specially the Random Forest model, which obtained an accuracy of approximately 90%. For this reason, we selected this model as the core of this corporate credit rating assessment method, using a subset of 26 predictors.

Finally, as a means to facilitate comparability to standard corporate credit ratings and make their understanding easier due to current standards, we converted the probabilities obtained from the random forest predictions to letter ratings, based on results obtained by analyzing conventional corporate credit ratings.

Additionally, it should be mentioned that due to the difficulty to obtain data about businesses default, there have not been many contributions on default prediction per se. Meanwhile, the majority of contributions to his field have come from bankruptcy prediction. For this reason, this study represents one of the first studies to predict Spanish businesses default probability.

6.1 Main Findings and implications

This thesis showed how machine learning methods are able to outperform conventional statistical methods in default prediction (logistic regression and linear discriminant analysis), showing how tree methods achieved the best prediction ability. Those results, are in the same line to Mu-Yen Chen (Chen, 2011), who showed that C5.0 and CART algorithms achieve the best prediction ability. Even though, we showed that Random Forest, in default prediction is superior to C5.0 and CART, achieving an error rate of only 10.1%, almost 25% lower than CART. Achieving an accuracy higher than the majority of previous studies (Mirzaei, Ramakrishnan, & Bekri, 2016; Kim & Sohn, 2010; Ramakrishnan, Mirzaei, & Bekri, 2015; Yeh, Wang, & Tsai, 2014)

Additionally, we showed that by excluding variables that are not common to the majority of businesses, significant prediction ability can be achieved. Thus in comparison to other studies we excluded common metrics such as market capitalization and any financial ratio related to market capitalization. Thus being this assessment method "universally" applicable and not only to listed companies as many of the corporate credit assessment methods developed by previous literature.

Moreover, again we corroborated the importance of financial ratios on a firm's creditworthiness assessment. Although, we found that measures that are not typically used in previous research on default and bankruptcy prediction also have shown substantial importance on predicting default, such as the previous number of administrative claims, the number of employees and the GDP compound annual growth over the period.

Hence, this automated credit rating assessment reduces the information asymmetry between corporate borrowers and lenders, by allowing any individual, firm or bank to check a firm's creditworthiness in a pretty simple manner. Additionally, internally, firms will be able to check how changes on their accounts or in the environment can affect their creditworthiness.

Besides, the rating system developed in this study provides an easy, free, quick and accessible solution to assess a firm's creditworthiness, reducing information asymmetry between debtors and creditors, without the existence of conflicts of interest as happens with rating agencies. Hence, providing an easy way to signal a business' financial position without the need to pay high amounts of money to credit rating agencies. Furthermore, it also provides insight in corporate bond pricing, since corporate bonds are highly correlated to their corporate rating (Damodaran, 2012). Also, it supplies to businesses a tool to check how different operations and hypothetical situations may affect their mid-term/long-term creditworthiness.

6.2 Research Limitations and directions for further research

The present study focused on Spanish companies default prediction, but this geographic scope could be extended using samples from other regions, such as France, Belgium, Italy... Although it should be noted, that some countries have significant differences between their accounting practices, which should be taken into account, in order to generate adequate predictions. Additionally, further variables referring to country risks should be added.

Additionally, survival analysis could be applied, in order to compare the approach followed in this study. Moreover, survival analysis could be used to assess recovery prospects, thus generating a corporate credit rating based in both default probability and recovery prospects.

Besides, further research on default prediction should ensure to avoid mistakes committed by previous literature, which have been repeatedly criticized, but are still common in today's studies on default prediction, such as the misinterpretation of the concept of default, maximization of accuracy in heavily unbalanced datasets, etc.

Bibliography

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, *91*(5), 1369–1401.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier.

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

Altman, E. I., & Eisenbeis, R. A. (1978). Financial applications of discriminant analysis: a clarification. *Journal of Financial and Quantitative Analysis*, *13*(1), 185–195.

Argenti, J. (1983). *Predicting corporate failure.* Technical Directorate of the Institute of Chartered Accountants in England and Wales.

Barro, R., & Sala-i Martin, X. (2007). *Economic growth.* Prentice Hall of India Private Limited.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71–111.

Bilardello, J., & Ganguin, B. (2005). *Fundamentals of corporate credit analysis.* McGraw-Hill.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Brealey, R., Leland, H. E., & Pyle, D. H. (1977). Informational asymmetries, financial structure, and financial intermediation. *The journal of Finance*, *32*(2), 371–387.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Acm sigmod record* (Vol. 29, pp. 93–104).

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

Chudson, W. A., et al. (1945). The pattern of corporate financial structure: a cross-section view of manufacturing, mining, trade, and construction, 1937. *NBER Books*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Damodaran, A. (2010). *Applied corporate finance.* John Wiley & Sons.

Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset* (Vol. 666). John Wiley & Sons.

De Servigny, A., Renault, O., & de Servigny, A. (2004). *Measuring and managing credit risk.* McGraw-Hill New York.

Dewatripont, M., Tirole, J., et al. (1994). *The prudential regulation of banks* (Tech. Rep.). ULB–Universite Libre de Bruxelles.

Diamond, J., & Ordunio, D. (2011). *Guns, germs, and steel.* Books on Tape.

Dunn, R. R., Davies, T. J., Harris, N. C., & Gavin, M. C. (2010). Global drivers of human pathogen richness and prevalence. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20100340.

Durand, D., et al. (1941). Risk elements in consumer instalment financing. *NBER Books*.

Dwyer, M. M. D. (1992). A comparison of statistical techniques and artificial neural network models in corporate bankruptcy prediction.

Edmister, R. O. (1971). *Financial ratios as discriminant predictors of small business failure* (Unpublished doctoral dissertation). The Ohio State University.

Ezzamel, M., Mar-Molinero, C., & Beech, A. (1987). On the distributional properties of financial ratios. *Journal of Business Finance & Accounting*, *14*(4), 463–481.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of human genetics*, *7*(2), 179–188.

FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies.*

Greenwald, B. C., & Kahn, J. (2005). *Competition demystified: a radically simplified approach to business strategy.* Penguin.

Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, *51*, 72–84.

Hall, R. E., & Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The quarterly journal of economics*, *114*(1), 83–116.

Haselmann, R., & Wahrenburg, M. (2016). *Banks' internal rating models-time for a change? the" system of floors" as proposed by the basel committee* (Tech. Rep.). White Paper Series.

Higgins, R. C. (2012). *Analysis for financial management.* McGraw-Hill/Irwin.

Jackendorff, N. (1962). *A study of published industry financial and operating ratios* (Vol. 52). Small Business Administration.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jensen, M., & Chew, D. (1995). Us corporate governance: Lessons from the 1980's.

Joy, O. M., & Tollefson, J. O. (1975). On the financial applications of discriminant analysis. *Journal of Financial and Quantitative Analysis*, *10*(5), 723–739.

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of smes based on technology credit. *European Journal of Operational Research*, *201*(3), 838–846.

Lacher, R. C., Coats, P. K., Sharma, S. C., & Fant, L. F. (1995). A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, *85*(1), 53–65.

Laffont, J.-J., & Martimort, D. (2009). *The theory of incentives: the principal-agent model.* Princeton university press.

Li, K. (2017). Monitoring role of credit rating agencies and corporate earnings management. In *Fma european conference.*

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The quarterly journal of economics*, *107*(2), 407–437.

Maridal, J. H. (2013). Cultural impact on national economic growth. *The Journal of Socio-Economics*, *47*, 136–146.

Martikainen, T., Perttunen, J., Yli-Olli, P., & Gunasekaran, A. (1995). Financial ratio distribution irregularities: implications for ratio classification. *European Journal of Operational Research*, *80*(1), 34–44.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of banking & finance*, *1*(3), 249–276.

Mcleay, S., & Omar, A. (2000). The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios. *The British Accounting Review*, *32*(2), 213–230.

Merwin, C. L., et al. (1942). Financing small corporations in five manufacturing industries, 1926-36. *NBER Books*.

Messier Jr, W. F., & Hansen, J. V. (1988). Inducing rules for expert system development: an example using default and bankruptcy data. *Management Science*, *34*(12), 1403–1415.

Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, *28*(4), 603–614.

Min, S.-H., Lee, J., & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert systems with applications*, *31*(3), 652–660.

Mirzaei, M., Ramakrishnan, S., & Bekri, M. (2016). Corporate default prediction with industry effects: evidence from emerging markets. *International Journal of Economics and Financial Issues*, *6*(3S).

Modigliani, F., & Miller, M. H. (1963). Corporate income taxes and the cost of capital: a correction. *The American economic review*, *53*(3), 433–443.

Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. *Journal of the American Statistical association*, *58*(303), 799–806.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Porter, M. E. (2008). *Competitive strategy: Techniques for analyzing industries and competitors.* Simon and Schuster.

Ramakrishnan, S., Mirzaei, M., & Bekri, M. (2015). Adaboost ensemble classifiers for corporate default prediction. *Research Journal of Applied Sciences, Engineering and Technology*, *9*(3), 224–230.

Romer, P. M. (1990). Human capital and growth: theory and evidence. In *Carnegie-rochester conference series on public policy* (Vol. 32, pp. 251–286).

Ross, S. A. (1977). The determination of financial structure: the incentive-signalling approach. *The bell journal of economics*, 23–40.

Ryu, Y. U., & Yue, W. T. (2005). Firm bankruptcy prediction: Experimental comparison of isotonic separation and other classification approaches. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *35*(5), 727–737.

Sanfeliu, C. B., García, F. G., Martínez, F. G., & Clemente, I. M. (2013). Default prediction of spanish companies. a logistic analysis. In *Intellectual economics* (Vol. 7, pp. 333–343).

Smith, R. F. (1935). *Changes in the financial structure of unsuccessful industrial corporations, by raymond f. smith... and arthur h. winakor...* University of Illinois.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281–306). Elsevier.

Starmine quantitative models. (2013). *Thomson Reuters*. Retrieved from https://www.thomsonreuters.com/content/dam/openweb/documents/pdf/financial/starmine-quantitative-models.pdf

Sun, L., & Shenoy, P. P. (2007). Using bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, *180*(2), 738–753.

Tang, T. T. (2009). Information asymmetry and firms' credit market access: Evidence from moody's credit rating format refinement. *Journal of Financial Economics*, *93*(2), 325–351.

Thornhill, S., & Amit, R. (2003). Learning about failure: Bankruptcy, firm age, and the resource-based view. *Organization science*, *14*(5), 497–509.

Venard, B. (2013). Institutions, corruption and sustainable development. *Economics Bulletin*, *33*(4), 2545–2562.

Yeh, S.-H., Wang, C.-J., & Tsai, M.-F. (2014). Corporate default prediction via deep learning. In *The 34th international symposium on forecasting (isf 2014)*.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59–82.

A Input Variables

Table A.1: Initial set of Input variables formulas

Variable	Formula
Number of employees	-
Number of directors & Managers	-
Import/Export activity	-
Legal Form	-
First two digits of CNAE code	-
Return On Capital Employed (%)	$\frac{Pre-Tax\ Profit\ +\ Interest\ Paid}{Shareholders'\ Funds\ +\ Non\ Current\ Liabilities} \times 100$
Return On Total Assets (%)	$\frac{Pre-Tax\ Profit\ (Loss)}{Total\ Assets} \times 100$
Profit Margin (%)	$\frac{Pre-Tax\ Profit(Loss)}{Operating\ Revenue} \times 100$
Net Assets Turnover (%)	$\frac{Operating\ Revenue}{Shareholders'\ Funds+Non\ Current\ Liabilities} \times 100$
Interest Cover (%)	$\frac{Operating\ Profit\ (Loss)}{Interest\ Paid} \times 100$
Collection period (days)	$\frac{Debtors}{Operating\ Revenue} \times 360$
Credit period (days)	$\frac{Creditors}{Operating\ Revenue} \times 360$
Current Ratio (%)	$\frac{Current\ Assets}{Current\ Liabilities} \times 100$
Liquidity Ratio (%)	$\frac{CurrentAssets-Stocks}{Current\ Liabilities} \times 100$
Stock Turnover (%)	$\frac{Operating\ Revenue}{Stocks} \times 100$
Solvency Ratio (%)	$\frac{Sharegolder's\ Funds}{Total\ Assets} \times 100$
Gearing (%)	$\frac{Non\ Current\ Liabilities+Loans}{Shareholders'\ Funds} \times 100$
Staff Costs-to-Operating Revenue	$\frac{Cost\ of\ Employees}{Operating\ Revenue}$
Total Assets per employee	$\frac{Total\ Assets}{Number\ of\ Employees}$
Operating Revenue per employee	$\frac{Operating\ Revenue}{Number\ of\ Employees}$
Sales Compound annual Growth (%)	$((\frac{Sales_{time2}}{Sales_{time1}})^{\frac{1}{number\ of\ years}}) - 1) \times 100$
Net Income Compound annual Growth (%)	$((\frac{Net\ Income_{time2}}{Net\ Income_{time1}})^{\frac{1}{number\ of\ years}}) - 1) \times 100$
Long-term Debt Compound annual Growth (%)	$((\frac{Long-term\ Debt_{time2}}{Long-term\ Debt_{time1}})^{\frac{1}{number\ of\ years}}) - 1) \times 100$
Years after Establishment date	$Information\ date - Establishment\ date$
Indebtedness	$\frac{Total\ Liabilities\ -\ shareholders'\ equity}{Total\ Liabilities}$
Cash-to-Current Liabilities	$\frac{Cash}{Current\ Liabilities}$
Return On Shareholders' Funds	$\frac{Pre-Tax\ Profit(Loss)}{Shareholders'\ Funds}$
Shareholders' Liquidity Ratio	$\frac{Shareholders'\ Funds}{Non-Current\ Liabilities}$
Spanish GDP Compound annual Growth (%)	$((\frac{Ending\ GDP}{Beginning\ GDP})^{\frac{1}{number\ of\ years}}) - 1) \times 100$
Number of previous administrative claims	-

Table A.2: Initial set of input variables description

Variable	Description
Number of employees	indicates a company's the number of employees
Number of directors & Managers	indicates a company's number of directors and managers
Import/Export activity	indicates whether a business exports its products services or not, and whether a company imports products/services or not
Legal Form	indicates the businesses' legal form (limited liability company, public limited company, association, etc.)
First two digits of CNAE code	CNAE codes are alphanumeric codes, which indicate a business' activity, the first two digits of a CNAE code indicate the division of a firm's activity
Return On Capital Employed (%)	profitability and efficiency measure, indicating the efficiency of a firm's employed capital.
Return On Total Assets (%)	indicates the effectiveness of a company's assets utilization to generate earnings
Profit Margin (%)	profitability measurement, which measures the amount of pre-tax profit generated with each euro of sales
Net Assets Turnover (%)	efficiency measurement, indicating the efficiency with which a company deploys its assets in generating revenue
Interest Cover (%)	debt and profitability ratio, indicating the effor needed by a company in order to pay interest on its outstanding debt
Collection period (days)	indicates the approximate amount of time (in days) needed to collect invoiced amounts from customers
Credit period (days)	indicates the approximate amount of time (in days) taken
Current Ratio (%)	liquidity measurement, which indicates a firm's ability to pay long and short-term obligations.
Liquidity Ratio (%)	liquidity measurement, whoch indicates a firm's immediate ability to pay its short-term obligations. It is also known as Quick ratio or Acid Test
Stock Turnover (%)	business performance measurement, indicating how fast a business sells its inventories.
Solvency Ratio (%)	financial leverage measurement, which determines the percentage of assets owned by shareholders.
Gearing (%)	financial leverage measurement, which indicates the relation between owner's equity and borrowed funds by the company
Staff Costs-to-Operating Revenue	profitability measurement, indicating the staff expenses in proportion to the operating revenue. It is also used as an efficiency measure, indicating how efficient is the staff allocation.
Total Assets per employee	average number of assets per employee
Operating Revenue per employee	efficiency measure, indicating the average revenue generated by each company's employee.
Sales Growth Rate (%)	average net sales growth rate during the period considered
Net Income Growth Rate(%)	average Net Income growth rate during the period
Long-term Debt Growth Rate (%)	average Long-term debt growth rate during the period
Years after Establishment	Number of years after business' establishment
Indebtedness	leverage measure, which indicates a company's proportion of borrowed funds.
Cash-to-Current Liabilities	liquidity measure, which measures a company's ability to meet short-term obligations.
Return On Shareholders' Funds	profitability ratio, indicating how efficiently a company is managing its shareholders' funds.
Shareholders' Liquidity ratio	leverage ratio indicating the amount of long term debt relative to shareholder's funds.
Spanish GDP Compound annual Growth	Mean annual growth of Spanish GDP over the period considered.
Number of previous administrative claims	Number of administrative claims presented to the company by the tax authorities or social security in a period of time..

B | Data pre-processing

As mentioned in Chapter 3 from this thesis we did not select the companies to be analyzed applying a size criterion, which could lead to a potential data quality problem. Due to, lower requirements of accounting information submission from those smaller businesses. Hence, not finding information for many of the variables considered for those businesses. Additionally, accounting information is also subject to veracity problems, hence multivariate outlier techniques must be applied, in order to detect potential anomalies.

Initially, we selected all the companies fulfilling the criteria specified in Chapter 3 (6,090 defaulting businesses and 11,764 non-defaulting businesses). After selecting this data we proceeded to check its quality, removing those companies that had a high proportion of missing variables. Thus obtaining 4,371 defaulting businesses which were paired with non-defaulting businesses of similar characteristics. Additionally, we also deleted two of the initial variables due to a high accumulation of missing values in them, those were V22 and V23. Hence the final dataset constituted of 4,371 defaulting firms and 4,371 non-defaulting and 29 explanatory variables.

B.1 Missing Data

Firstly, we analyze the proportion of missing values per variable and the companies' missingness pattern, discriminating by defaulting and not defaulting companies, since their status could lead to potential differences among the missingness patterns. Thereby, leading to different strategies to lead with missing data.

Figure B.1 shows the proportion of missing values per variable and the missingness pattern for defaulting companies. In this Figure, we can see that several variables accumulate a significant proportion of missing values. Additionally, we can also see that an important number of companies holds various variables with missing values.

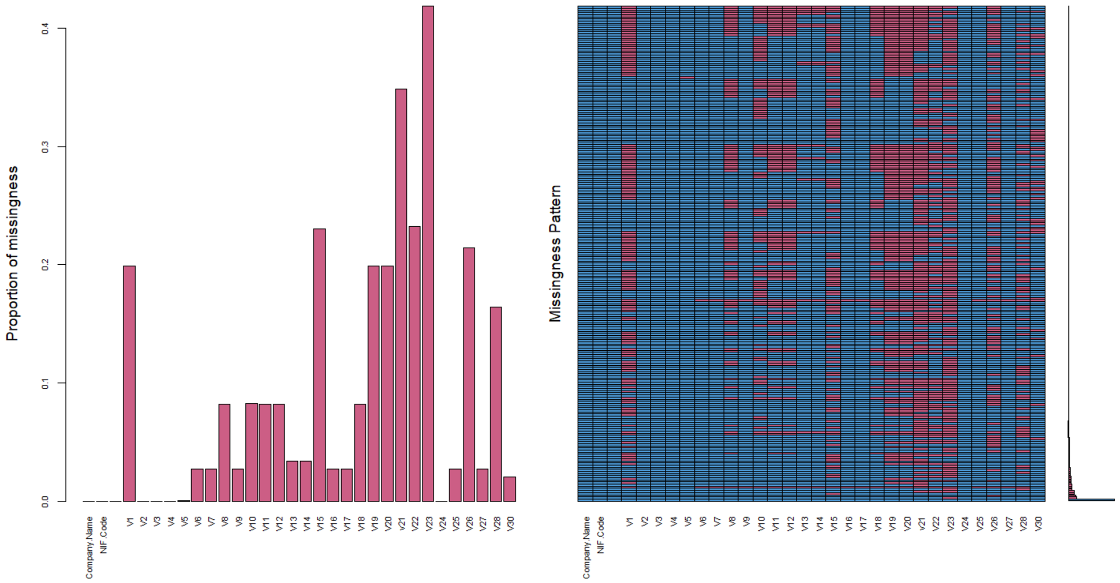


Figure B.1: Proportion of Missingness and Missingness Pattern for Defaulting Companies

Hence we removed all those companies having more than 3 variables with missing values (28.22% of the defaulting companies data). By doing so, the missingness pattern and the proportion of missing values per variable changed significantly, as we can see in Figure B.2.

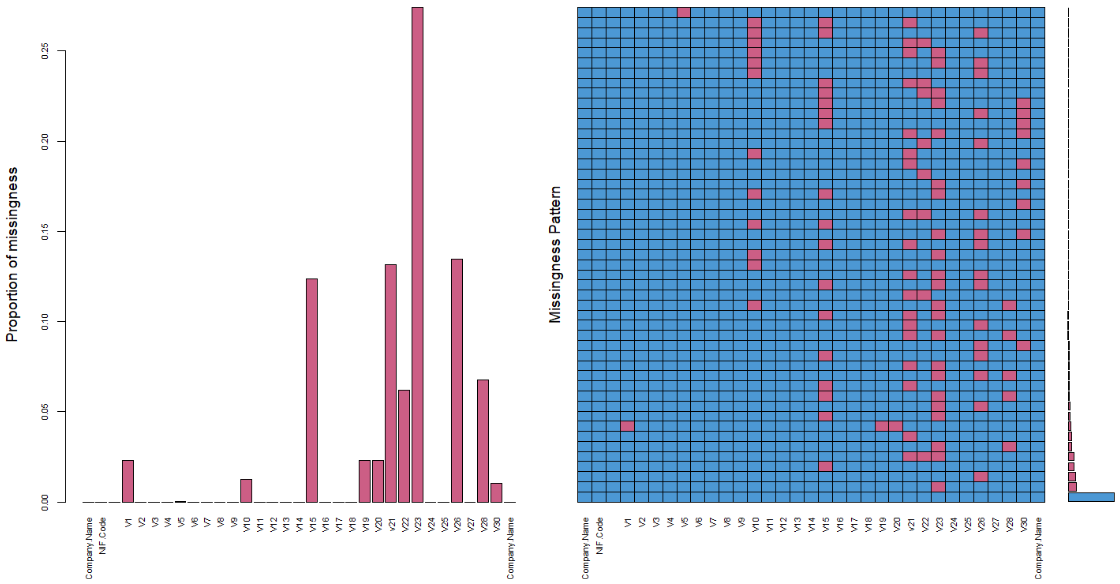


Figure B.2: Proportion of Missingness and Missingness Pattern for Defaulting Companies after removing companies with more than 3 variables with missing values

As seen in the previous Figure, by removing those companies with 3 or more missing variables, the proportion of missing values has been reduced substantially for all variables. Additionally, there's no systematic missingness pattern among companies. Furthermore, an important proportion of the defaulting companies (75.27%) have no missing values. For this reason, we proceed to use missing data imputation techniques, by using k-nearest-neighbors (KNN) algorithm, in order to match a missing value with its k nearest neighbors, taking k equal to 6.

After dealing with missing data on the defaulting companies, we proceeded to do the same for non-defaulting companies. Figure B.3 shows the proportion of missing values per variable, as well as the missingness pattern for Non-Defaulting Companies. As can be seen on this Figure, missingness represents an important problem in this case. Hence, in this case, we proceed to remove all those companies with more than 4 missing variables with the goal to improve as before the quality of our data. But as can be seen on Figure B.4, after doing so variables V22 and V23 still present notorious proportions of missing values, becoming uninterpretable. Therefore, we proceed to remove those two variables from our dataset.

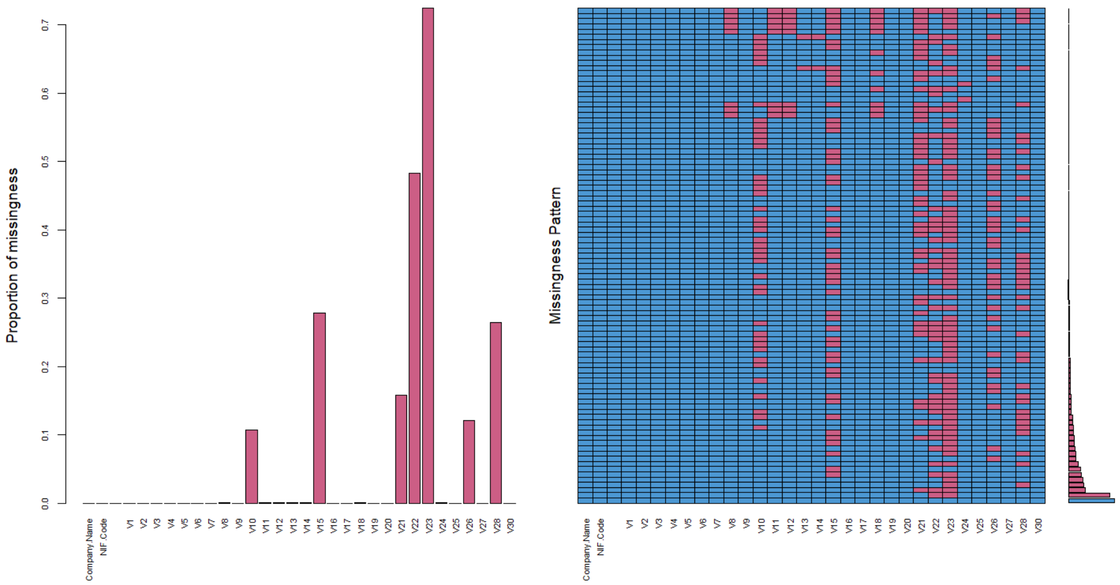


Figure B.3: Proportion of Missingness and Missingness Pattern for Non-Defaulting Companies

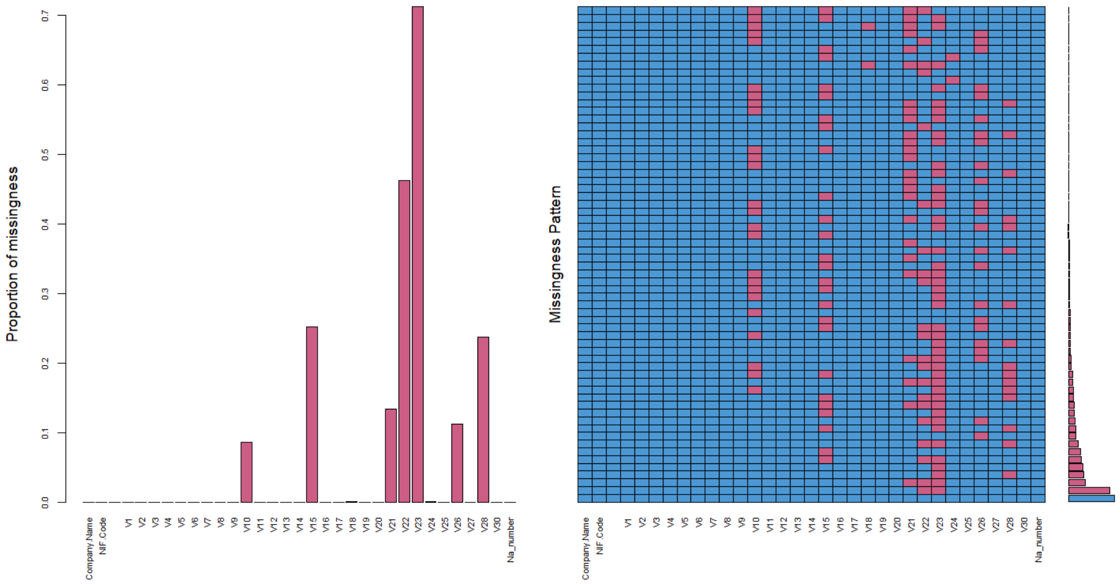


Figure B.4: Proportion of Missingness and Missingness Pattern for Non-Defaulting Companies after removing those companies with more than 4 variables with missing values

After removing V22 and V23 from the dataset, the missingness pattern and the proportion of missing values per variable become pretty similar to those seen for defaulting companies, as is evident from Figure B.5. For this reason, again, we proceeded to impute the missing values using KNN algorithm with k equal to 6.

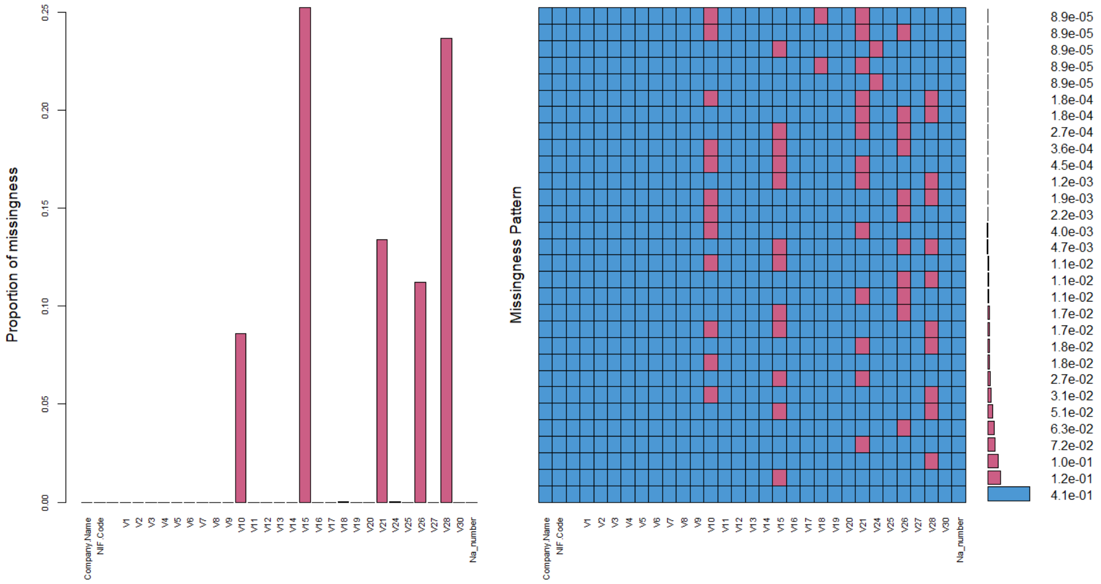


Figure B.5: Proportion of Missingness and Missingness Pattern for Non-Defaulting Companies after removing variables V22 and V23

B.2 Outliers

In order to detect outliers we followed an approach similar to the one followed for missing data. We treated defaulting and non-defaulting companies independently, in order to correctly spot outliers, due to their particularities because of their status. To detect outliers we used an algorithm called Local Outlier Factor (LOF) which detects outliers by measuring the local deviation of a particular data point in relation to its neighbours (Breunig, Kriegel, Ng, & Sander, 2000).

Figure B.6 plots the score obtained using LOF for each defaulting company. We can see, that there are three companies which have significant higher values than their neighbors, but by examining them we consider that those companies do not represent outliers, since all of their values are coherent (all of them have a relatively high number of employees, even though the number of employees for those companies is feasible and reasonable).

Additionally, Figure B.7 plots the score obtained using LOF for each non-defaulting company. In this case, there's only one company with a significant higher score than the others, but again by examining this company we did not consider it as an outlier, since this company accumulated significant losses among the years considered but by looking its financial information those loses were feasible according to its size.

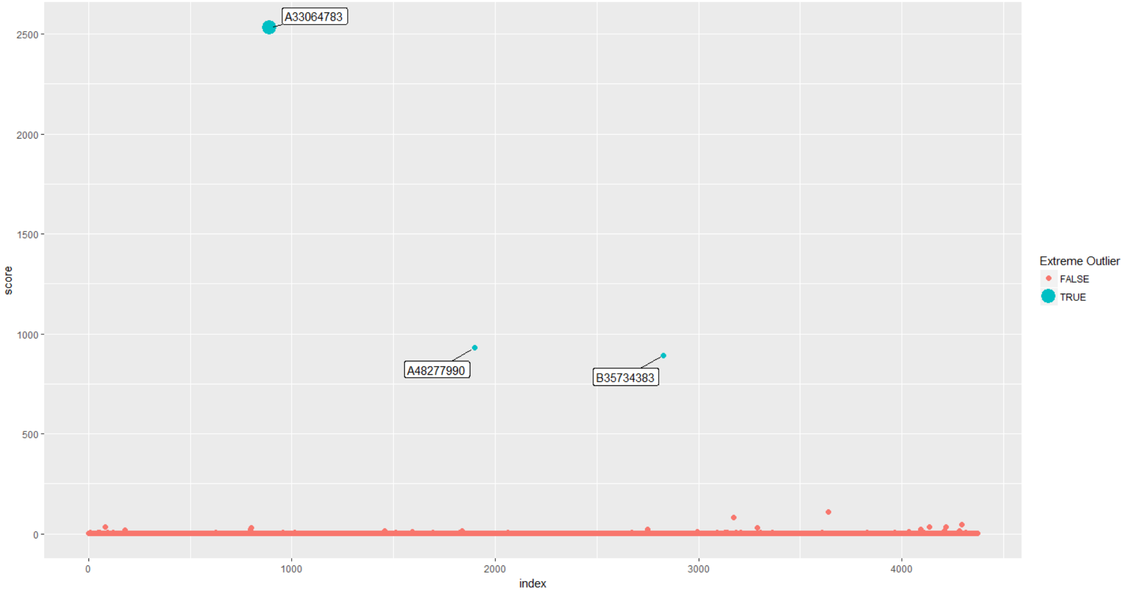


Figure B.6: Local Outlier Factor Scores for Defaulting Companies

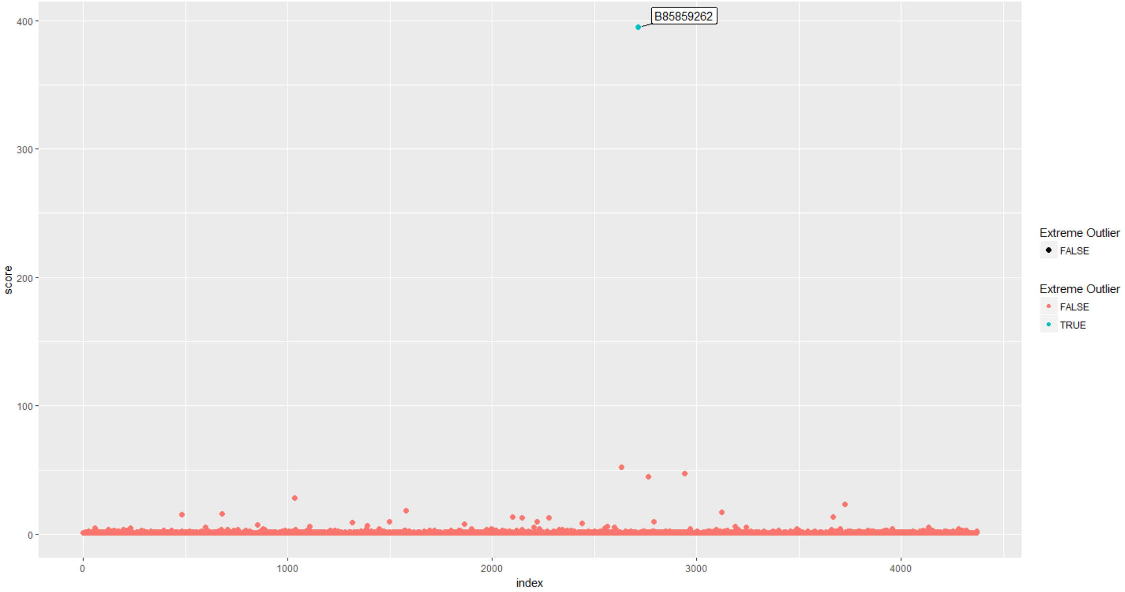


Figure B.7: Local Outlier Factor Scores for Non-Defaulting Companies

B.3 Variables

Figure B.8 shows the distribution of the different explanatory variables, for the totality of the data considered in this study (defaulting and non-defaulting companies). As can be seen from this Figure the majority of businesses Small and medium-sized enterprises (SMEs), additionally as depicted in this figure financial ratios do not follow a normal distribution. For this reason, we applied a logarithmic transformation to Variables V6 to V28 (due to the presence of negative values, for each point we converted its absolute value plus one, and preserving the sign after the logarithmic transformation, i.e. $sign(x) \times log(abs(x) + 1)$).

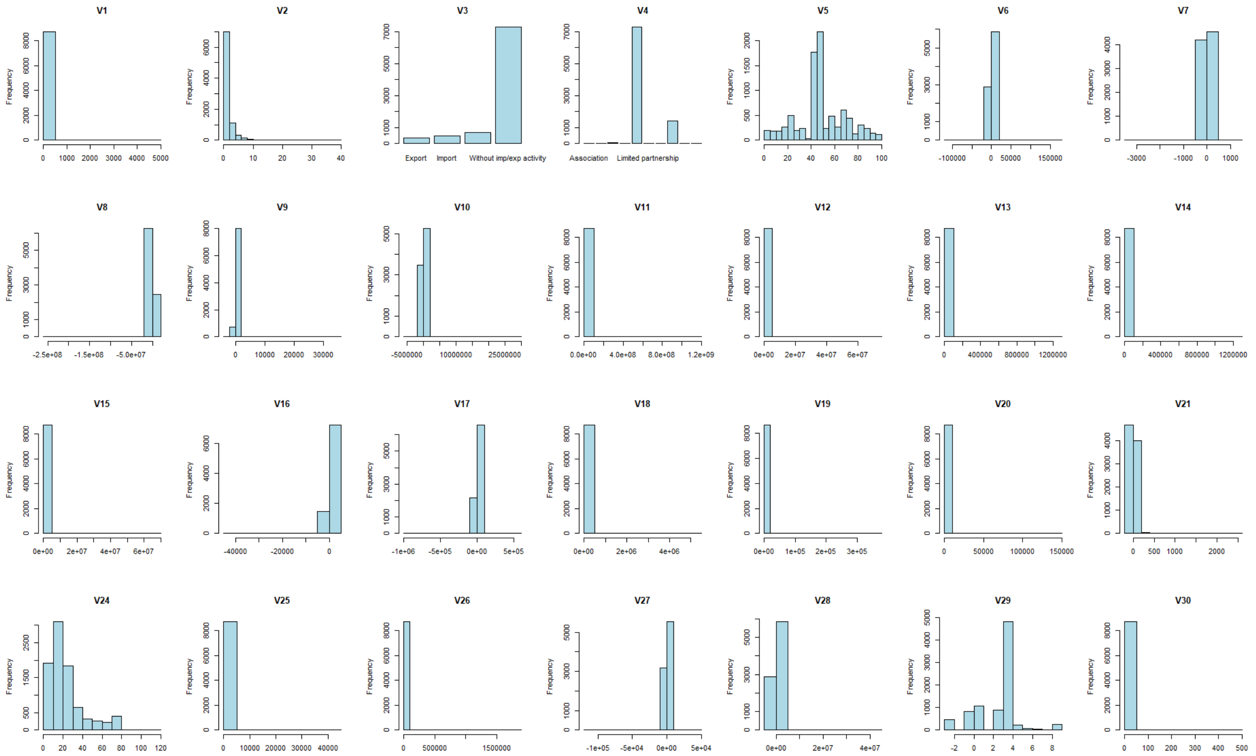


Figure B.8: Variables distribution

Glossary

adverse selection situation that happens in contexts affected by asymmetric information, in which those who have a higher amount of information benefit from it, at the expense of others. Thus leading to market inefficiencies. A common example is found in the secondhand market, where the seller may know about a car's defect, while the buyer doesn't. Thus, the seller may sell that car without disclosing the issue, charging the same price as if the car had not that issue (Akerlof, 1978). 2, 3, 9

Agency Costs Economic concept, which refers to conflicts of interest between individuals with different objectives, such as employer/employee, debtholders/shareholders, etc. 2

assets Usually firm's investment, i.e. resources which are expected to provide future earnings. 8

business cycle Rise or fall of a country Gross Domestic Product. 6

competitive advantage Attribute that permits an organization to sustain higher relative prices and or lower relative costs than its rivals in an industry. Thereby achieving relative superior performance. 7

corporate bonds Securities issued by a corporation, in order to raise financing, and sold to investors. 1

due diligence Process of investigation of a potential investment/product. 1

earnings management Usage of accounting techniques in order to produce financial reports that reflect an inordinately positive view of the business. 3

economies of scale Refers to reduced costs per unit as production increases. 6

filters Selection procedure based on general attributes like correlation with the variable to predict. 15

financial statements Formal record of the firm's financial activities and position. 8

Gross Domestic Product per capita Monetary measure of the market value of the final goods and services produced within a country in a determined time period (usually of one year), divided by the average country population in that period. 4

inflation rate Rate at which the general price level is increasing. 6

liabilities Firm's financial debt or obligations. 8

moral hazard change in one's behaviour due to a reduction of the risks assumed. A common example is in the insurance industry, in which an individual may be less inclined to take care of a belonging that has been insured against damage. 2, 3, 9

network economies Refers to increases of value to customers as the number of customers increases. A common example of economies of scale are social network or messaging apps. 6

operating efficiency Metric that measures the efficiency of the profit obtained in relation to operational costs. 8

principal components low dimensional representation of the data that contains the utmost amount of variability of the data. Each component is calculated as a linear combination of the different features of the data(James et al., 2013). 23

returns on equity Profitability measure that reflects the profit generated by a company in relation to the money that shareholders have invested. 7

RFE (Recursive Feature Elimination) greedy algorithm to find the best performing subset of features. This algorithm repeatedly constructs a model, by removing a portion of the worst features. Concretely, it starts building a model with all the features and ranking those features, then it removes the worst feature(s) building a new model with those features. This process is repeated, until all features have been removed, then the subset of features that presented a better performance over the others is selected as the best subset.. 26–28, 31, 33, 34, 36

shareholder's equity Equals to the difference between assets and liabilities, it represents the funds that would be returned to shareholders in case that all company's assets were liquidated and debts repaid. This is mainly composed by the company's share capitals and the retained earnings. 8

wrapper Iterative selection procedure, which selects features based on the performance evaluation of the classifier using those features. 15, 25, 37