Credit Risk Modelling

Diogo Rafael Dias Coreixas

Thesis to obtain the Master of Science Degree in

Industrial Engineering and Management

Supervisor: Prof. António Manuel da Nave Quintino

Examination Committee

Chairperson: Prof. Carlos António Bana e Costa

Supervisor: Prof. António Manuel da Nave Quintino

Member of the Committee: Prof. Maria Margarida Martelo Catalão Lopes de Oliveira Pires Pina

January of 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Abstract

Setting a credit limit to companies is a major source of concern in financial risk management. The methods that calculate the credit limit have been developed as a response to the market requirements and changes. A good and appropriate practice of setting credit limits has a huge impact on financial institutions savings, once this results in a reduction of the potential companies which can default in their financial obligations.

The present research aims to build reliable models, which automatically define the credit limit of the companies, according to their financial data. The financial dataset was provided from one of the biggest financial, containing information of thousands of corporations, since financial ratios to financial results. Before developing the models, it was necessary to process the data, in order to minimize the negative effects of its inherent aspects, which may impair the research, such as multicollinearity.

To build these models various technologies can be used, and in this study, methods such as statistical and AI methods will be tested. Considering these methods, the main objective was to find out which one can develop the most accurate model. Taking into account the final results, the model that revealed the best predictive performance was obtained by an Artificial Intelligence method, as it would be expected according to the academic literature. Despite the Multiplicative Model not being reliable at this stage of the research, it left strong evidences of being the analytical method used to credit limit calculation, requiring a deeper analysis, where the data are possibly divided into classes, and each class is defined by a different regression.

Keywords: *Credit limit, Multiple Linear Regression, Multiplicative Models, Multilayer Perceptron, Radial Basis Function.*

Resumo

Definir o limite de crédito de empresas é uma das maiores fontes de preocupação na gestão de risco financeiro. Os métodos que calculam os limites de crédito têm vindo a ser desenvolvidos, de forma a cumprir as exigências e mudanças que ocorrem no mercado. Uma boa e adequada prática de definição de limite de crédito tem um impacto enorme nas poupanças das instituições financeiras, dado que resulta numa redução do potencial de empresas que podem falhar as suas obrigações financeiras.

A presente pesquisa tem o objetivo de construir modelos fiáveis, que definem automaticamente o limite de crédito das empresas, de acordo com os seus dados financeiros. Os dados financeiros foram fornecidos por uma das maiores bases de dados financeira, contendo informações financeiras de milhares de empresas, desde rácios financeiros a resultados financeiros. Antes de desenvolver os modelos foi necessário processar os dados de forma a minimizar os efeitos negativos provenientes de aspetos inerentes dos dados, tal como a multicolinearidade, que pode condicionar a pesquisa.

Para construir estes modelos podem ser usadas várias tecnologias, sendo que neste estudo serão considerados métodos estatísticos e de inteligência artificial. Tendo em conta estes métodos, o principal objetivo é descobrir qual destes consegue desenvolver o modelo mais preciso. Considerando os resultados finais, o modelo que revelou o melhor desempenho preditivo foi obtido por um método de Inteligência Artificial, como seria expectável de acordo com a literatura académica. Apesar do modelo desenvolvido pelos Modelos Multiplicativos não ser fiável nesta fase do estudo, deixou fortes evidências de ser o método analítico usado para o cálculo do limite de crédito, exigindo uma análise profunda, sendo que existe a possibilidade de os dados serem divididos por classes, em que cada uma é definida por uma regressão diferente.

Palavras-chave: *Limite de crédito, Regressão Linear Múltipla, Modelos Multiplicativos, Multilayer Perceptron, Radial Basis Function.*

Acknowledgements

I would like to thank to my family for all support they gave me throughout my academic life that culminates in this thesis. I would also like to show my gratitude to my friends, who have always been willing to help me.

Finally, I would like to thank to professor António Quintino for his influence on this dissertation. As we know, the thesis is a work that takes many months to be done, and during this period the professor has always been committed to this research, being available whenever I needed, guiding me in the best way to develop this study, motivating me to always achieve the best possible results. Thus, professor António Quintino was an essential element in the development of this thesis.

Table of Contents

List of Figures

List of Tables

Glossary of Acronyms

AI- Artificial Intelligence

ANN- Artificial Neural Networks

CV- Coefficient of Variation

CRA- Credit Rating Agency

EBIT- Earnings before Interests and Taxes

EBITDA- Earnings before Interests, Taxes, Depreciation and Amortization

GDP- Gross Domestic Product

LE- Large Enterprises

LR- Linear Regression

MD- Mahalanobis Distance

MLP- Multilayer Perceptron

MLR- Multiple Linear Regression

MM- Multiplicative Models

RBF- Radial Basis Function

RNG- Random Number Generator

ROA- Return on Assets

ROCE- Return on Capital Employed

ROE- Return on Equity

SME- Small Medium Enterprise

VIF- Variance Inflation Factor

VIX- Volatility Index

1. Introduction

1.1. Contextualization of the problem

The credit concept has several meanings in the financial world. Despite being generally associated to banks or credit institutions by lending money, this concept may also be an exchange of goods and/or services in exchange for a deferred payment. This happens, for example, when a supplier sells goods to other company, and the last doesn't have the purchase money available, so the supplier by analyzing the customer financial data, defines the maximum amount of goods to give considering what is the maximum credit which the customer is able to repay plus interests. In this way, credit can be described as a contractual agreement in which a borrower receives something of value, and agrees to repay it at a later date to the lender (Kenton, Credit & Debt, 2019).

The big problem for lenders of granting credits, is the possibility of borrowers not meeting their obligations, and default in the repayment of the credit, and this is one of the main concerns for financial institutions (lenders) related with financial risk management. In order to prevent the occurrence of this phenomenon, it's necessary to define the maximum economic value that a credit provider should provide to a borrower, considering its capacity to repay the loan, by analyzing its financial data, and defining the credit exposure which the lender is willing to take from each client, called credit limit.

Although defining excessively low credit limit will result in a safer way to conduct the business, this option will have a negative impact on the performance of the business, once setting a trade credit can be seen as an advantage for sellers. By setting credits, sellers provide to buyers the opportunity to buy goods or services and delay its payment, attracting costumers that are not able to pay at the moment of exchange of goods or services, by taking costumer's default risk, which represents the risk of the costumer not fulfilling the payment agreement (Lou & Wang, 2017).

When a credit is granted the lender faces two broad types of risks: credit risk and market risk. Credit risk represents the likelihood of borrowers default on their obligations, resulting in losses for lender due to non-repayment from the borrower. The market risk reflects the variability in the value of their financial position due to changes in interest rates, exchange rates, taxes, and others. Basically, the market risk represents the conditions of the environment where the borrower is involved. In fact, market risk is very volatile, once market rates are constantly changing, and can be extremely difficult to predict them, whereas credit risk is usually constant considering the financial the financial status of the company because credit events are rare (Shilpi, 2014).

In this way, the credit limit of a company must be seriously studied, once if a credit is denied, a potential profitable customer may end up in a competitor company. Considering these factors, it's necessary to evaluate the scenarios of a client default in its obligations, and other of losing a profitable client by denning a credit, when making decisions about granting a credit.

When granting a credit, the future of enterprises is at stake, so it's mandatory to take some measures to manage the risk, in order to reduce the probability of lenders' business falling down, given that lenders depend on the repayment of loans from borrowers, and if they don't repay it, the business can go through difficulties.

Over the years, given the importance of setting credit limits, there are several corporations that have been studying the best model to adopt, aiming to find the most appropriate credit limit, in order to ensure that borrowers have conditions to repay the credit (plus taxes or interests, depending on the lender), minimizing the losses of the lender. To build these models various techniques are used, among them there are statistical tools, feeding a predictive model with companies past financial data, seeking to define the effect or relationship between each variable (which consists in a piece of financial information) and the credit limit, choosing which one fits better to include in the model. To perform these models, it's required to apply credit limits imposed by some financial entity, more precisely Credit Rating Agencies (CRA), and then, understand the influence of each variable (Bazzi & Hasna, 2015).The present research will test two statistical methods, Multiple Linear Regression (MLR) and Multiplicative Models (MM), since through the last it's only necessary to make a logarithmic transformation of all variables, and run a MLR analysis.

On the other hand, Artificial Intelligence (AI) methods can also be applied for the purpose of setting the most suitable credit limit, by using past financial data of companies, and through this, the computer can define the credit limit. Artificial Neural Networks (ANN) techniques (which is a branch of AI) allow the creation of models with high accuracy rates. This research uses two types of neural networks techniques: Multilayer Perceptron (MLP) and Radial Basis Function (RBF).

Both statistical and AI methods aim to facilitate the decision-making process about which is the most appropriate credit limit for each company, defining which is the maximum credit that can be granted to an enterprise.

1.2. Thesis goal

The present research aims to build models that are accurate in the credit limit calculation, facilitating the decision-making process. There are several statistical and AI techniques that can be applied to setting the credit limit, and this thesis focuses on the following methods: Multiple Linear Regression, Multiplicative Models, Multilayer Perceptron Neural Network and Radial Basis Function Neural Network. All these techniques are described and explained in further sections of the research, containing all details and considerations that must be done.

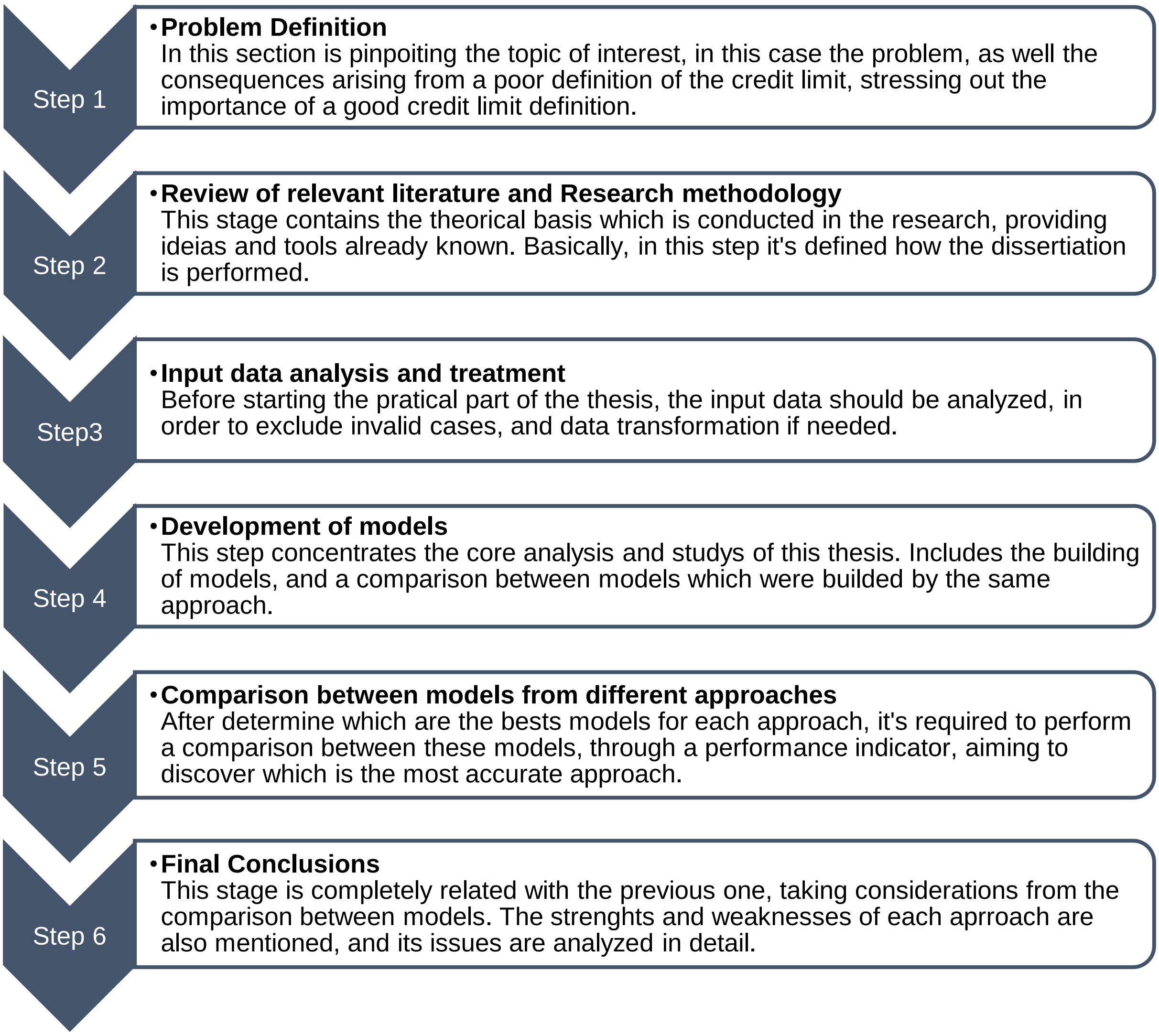
It should be noticed the fact that this subject is poor in terms of related literature, once it belongs to a sector where CRA face big competition between them, which means that their models are commonly unknown. However, there are some agencies which their models are relatively simple to discover when compared with others. The credit limit considered in this dissertation is relative to one of the most reliable credit limit (which cannot be mentioned for reasons of confidentiality), widely used in Europe, due to its accuracy in European companies, since geographic markets don't work in the same way. So, generally each model should be adapted considering various factors, such as geographic localization, once accurate models in Europe, may not be elsewhere, for example.

The models created in this thesis are intended to be implemented in response to real life problems, specifically those ones that large enterprises (LE) face, given that large part of its customers have to take a trade credit, adding the default risk to LE's business. In this way, this dissertation seeks to create reliable models of credit limit calculation, in order to establish alternatives to agencies, and so, LE become more independent, by not having to resort to external entities. As it known, the less a company depends on others, safer the business is, beyond the fact that in this case, the involvement of CRA represents a major financial expense.

To summarize, the results of this research can have a great impact on the way that the business is conducted, not just because it makes a company (lender) more autonomous, but also for the cost reduction, by not paying a license to CRAs.

1.3. Thesis Overview

To perform the present dissertation there are some steps that must be taken into account to understand the complexity of the problem in hands, in order to reach reliable models that set the credit limit for companies and to define which methods/methodologies can be applied in the problem solution.



Step 1

- **Problem Definition**
  In this section is pinpoiting the topic of interest, in this case the problem, as well the consequences arising from a poor definition of the credit limit, stressing out the importance of a good credit limit definition.

Step 2

- **Review of relevant literature and Research methodology**
  This stage contains the theorical basis which is conducted in the research, providing ideias and tools already known. Basically, in this step it's defined how the dissertiation is performed.

Step3

- **Input data analysis and treatment**
  Before starting the pratical part of the thesis, the input data should be analyzed, in order to exclude invalid cases, and data transformation if needed.

Step 4

- **Development of models**
  This step concentrates the core analysis and studys of this thesis. Includes the building of models, and a comparison between models which were builded by the same approach.

Step 5

- **Comparison between models from different approaches**
  After determine which are the bests models for each approach, it's required to perform a comparison between these models, through a performance indicator, aiming to discover which is the most accurate approach.

Step 6

- **Final Conclusions**
  This stage is completely related with the previous one, taking considerations from the comparison between models. The strenghts and weaknesses of each aprroach are also mentioned, and its issues are analyzed in detail.

Figure 1- List of main steps that must be taken to perform the research.

In the figure 1 are described the main steps that must be taken to perform this dissertation. A good definition of these steps is fundamental to generate a reliable research, once these steps coordinate the direction the study takes to solve the main problem.

It should be noted that, given that this thesis contains various approaches, the steps 2, 3 and 4 are repeated for each approach, once each one is conducted for different methods, that can only be done separately from others. So, for each approach it's required to review the relevant literature and define the research methodology, to make an input data treatment (if needed), and then to develop models for a certain approach.

1.4. Thesis Structure

This thesis is divided in multiple chapters, where each one has a different role. In the table below are disposed all chapters present in this document, as well as a description of the context of each one.

Chapter	Description
1. Introduction	In this section it is introduced the credit concept, giving a proper contextualization of the problem. It's also indicated the main goal of this thesis, and the impact in companies' real problems of achieving good results in this research. Lastly, are mentioned the main steps that must be taken, in order to perform a rigorous study.
2. Problem Definition	In this chapter it's described the credit importance, as well the benefits that can be taken from granting a credit. After analyzing the financial behavior of some European countries over the years, and the influence that the credit might have had on these countries. It's also provided a brief explanation about the CRAs, and their role in the in the market. At last, in this chapter is briefly presented the techniques that will be used in the present research.
3. Theoretical Background	The third section contains the theoretical basis of the research, containing all theories which support the methods and techniques that will be considered to perform the studies, building models that predict the credit limit of a corporation.
4. Input Data Collection, Analysis and Treatment	The fourth chapter focuses on the data that is used to perform the present research, including how the data was collected, a detailed description of the variables, and the pre-processing of the data through various analysis.
5. Model Building	This chapter describes the approach adopted for the development of the models, referring all the considerations taken, as well as the analysis of the created models.
6. Comparison between models	This is the last practical phase of the research, which aims to search for evidences to find out which developed model is the most accurate.
7. Conclusions	In the last section of the present dissertation the final results are analyzed and are mentioned the limitations experienced in the course of the study as well the further work that can be done in the future.

Table 1- Thesis structure and respective description.

2. Problem Definition

2.1. Credit Importance

In the past, credit was seen as only for individual consumers, not considering companies or institutions as it happens now. The life-cycle hypothesis was created by Modigliani and Brumber, presumes that consumers attempt to maintain their lifestyle over their lifetime considering their future income. So, taking debt in a early phase of their life is an expected behavior in order to maintain a desired lifestyle, supposing that their future income will be higher than their current income, and in the future their earnings will completely support their lifestyle (Soman & Cheema, 2002).

Over time the credit was applied to companies, having several benefits for both customers and suppliers:

- Suppliers are able to meet competition, once if competitors are making credits to customers, it's mandatory to follow this trend;
- Suppliers will increase sales, since that it's given to the customers the opportunity to buy with no prompt payment;
- When a customer has a good credit score (assess the past behavior on credit payments) allows him to obtain better interest rates and credit terms from lender (Woodruff, 2019).

The main disadvantage for suppliers in providing credits is that each client can be a potential bad debtor, and not fulfill its financial obligations, resulting in losses for lender.

A great example that shows the importance of the assessment of credits is the 2008 financial crisis, which was a dramatic moment for multiple major financial institutions, that during that time went bankrupt. This crisis was the worst economic and financial disaster since the Stock Market Crash of 1929. This crisis, in a brief explanation, can be explained by the facility of any entity getting credits, with low interests and taxes. The financial institutions that gave credits before this period did not make a deep research about its customers, analyzing if these had conditions to support the expenses related to the repayment of the credit. Through the facility to obtain a credit, the credit market became saturated, triggering several disastrous financial events around the world (BBC History Magazine, 2019).

Although the hardest time have already passed after this financial crisis, some consequences remain affecting population's lifestyle in many ways:

- Economic: investment, productivity, trend growth, jobs, real incomes, among others;
- Social: inequality, social tensions, political instability, etc;
- Long terms: pressure for economic reform, risk of hysteresis, protectionism (Riley, 2017).

As it known, the impact of the 2008 financial crisis was notorious, and because of that, this event was a turning point regarding the credit approach. Since this crisis the credit is seen differently, managers recognize the responsibility and the risks that are involved when setting a credit. The main lenders' concern has been to create a stricter approach that ensure a good accuracy, given that small improvements in accuracy can lead to decreased losses. With the development of credit modelling techniques, three main concepts were introduced in these:

- Risk management must account for unexpected losses, as well as accurately measure expected losses;

- Risk management must view risk from a portfolio perspective, taking into account correlations among assets, implying concerns for concentrated exposures to common risk factors;
- Risk management must develop measures of tail risk for assessing capital needs (Lang & Jagtiani, 2010).

There were several inklings that something bad could happen, and an example of this was the VIX (Volatility Index), which is a measure of a sentiment of a market. Basically, the higher the reading the more likely it is, investors believe that some event will occur, good or bad, in a near future.

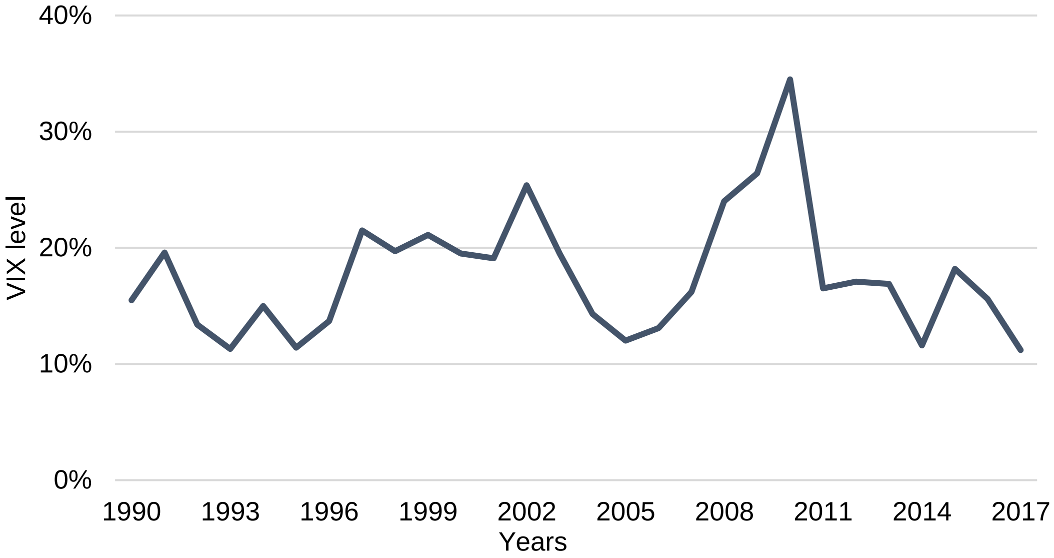


Figure 2- VIX level between 1990 and 2017 (Source: Schroders, Thomson Reuters Datastream).

Figure 2 reveals that since the beginning of 2005 the VIX level has begun to rise, and year after year the growth was more and more accentuated. In 2010 was recorded the higher VIX level ever seen, once this were unstable times, where was difficult to predict which scenario could happen in the future. Due to governments and banks interventions to stem the impact of the crises, the VIX decreased to historically low values, so confidence among investors grew once they don't feel that something will occur and cause a big change in the market (Brett, 2017).

2.2. Financial Overview of Europe

Given that the main goal of this research is to develop a model that can comprehends the credit limit informed by a world top credit rating agency, it's important to analyze the financial behavior and situation of some countries in Europe over the last years. The 2008 financial crisis had a huge impact on the European economy and lasted many years.

One of the best indicators of the financial status of a country is the deficit, which represents the amount by which its total expenses exceeds its tax revenues. The main objective of any government is to spend less than they earn, and when this happens, they reach a positive surplus. However, governments often fail to achieve this goal, being that total expenses are higher than governments revenues. In this case the government incurs a deficit, once they need to borrow the difference between these two financial data, increasing their debt (BBC News Business, 2013). The graph below shows the surplus or deficit (depending on whether it is greater than zero or not) as a proportion of GDP, which is the total value of goods and services produced by the economy each year.

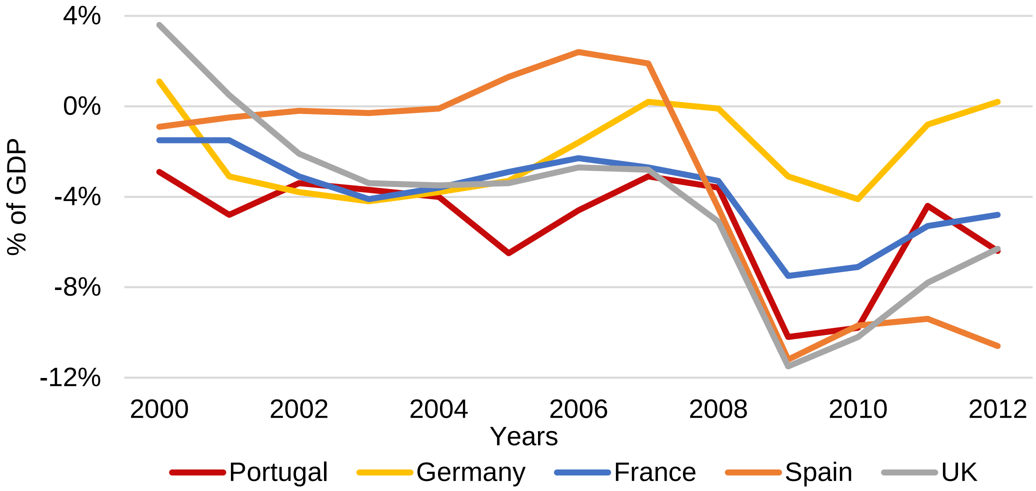


Figure 3- Deficit or Surplus in some countries of the Europe (Source: BBC News Business).

As it can be seen in figure 3, since 2007 the economy of various countries in Europe have begun to fall due to impact of the 2008 global financial crisis. Looking at the chart, it shows that Portugal, Spain and United Kingdom were the countries that suffered most from the crisis, recording a wide variation between 2007 and 2009. On the other hand, the country that best endured the crisis was Germany, although there was a small increase of deficit, this was not significant when compared with the others countries already mentioned.

As it is normal, the amount of credit that the lender is available to grant is highly dependent on the financial situation of the client. Therefore, it's important to assess the financial situation of the companies when a credit request is made. The financial autonomy appears as a measure which can be considered when assessing a company. This indicator shows the degree of an enterprise financial stability and leverage, evaluating the long-term solvency of the company. The smaller the ratio, more unstable the company is, which means that is more dependent from creditors, and it's calculated by dividing the total equity by the total assets of the company.

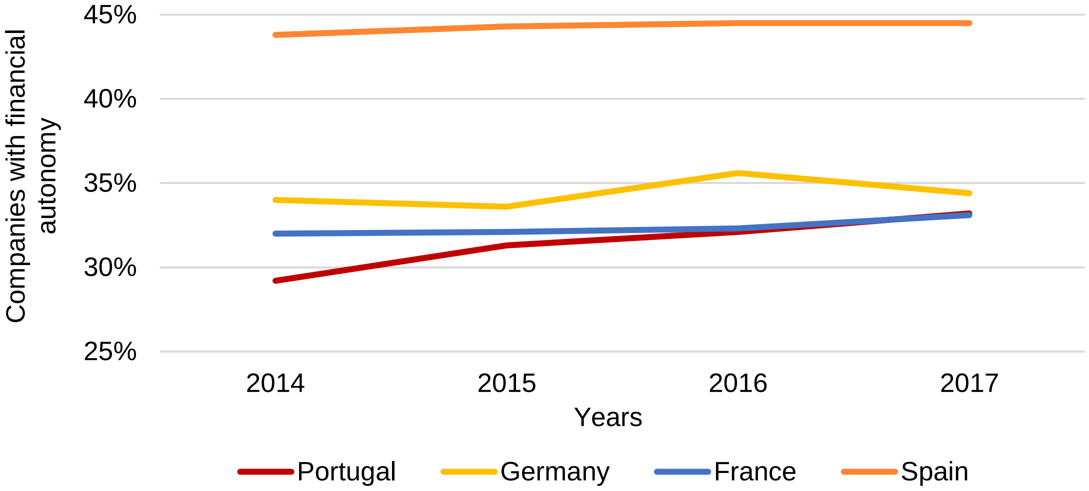


Figure 4- Percentage of companies with financial autonomy in some European countries (Source: Banco de Portugal).

In figure 4 is represented the percentage of companies with financial autonomy in different European countries between 2014 and 2017, and this indicator measures the corporation's ability to meet its financial commitments through its own capital, with no external funding from financial institutions or suppliers. Generally, for a company to be considered autonomous, this value must not exceed 0.5, but it depends on the market where each operates. As it can be seen in the chart, there are more companies in Spain with financial autonomy than in other countries, being around 45% of the companies. Although Portugal in 2014 had the lowest percentage of the companies with financial autonomy, about 29%, raised until 2017 to even slightly than France, that seems to had a constant behavior between this time period, regarding this indicator. Lastly, Germany did not have big changes in this period, but it needs to be highlighted the fact that german companies with financial autonomy never increased two consecutive years, remaining situated between 34% and 36%.

In order to assess the financial state of each country, it's also advisable to analyze the number of active businesses over the years. Generally, when the economy of a country is getting better or stable, it's expectable that the number of active companies in that country will not show major decreases. In this way, the business birth rate usually is higher than the business mortality rate during economic prosperity times. On the contrary, if time is of financial instability and crisis it's expectable that the business birth rate is lower than the business mortality rate.

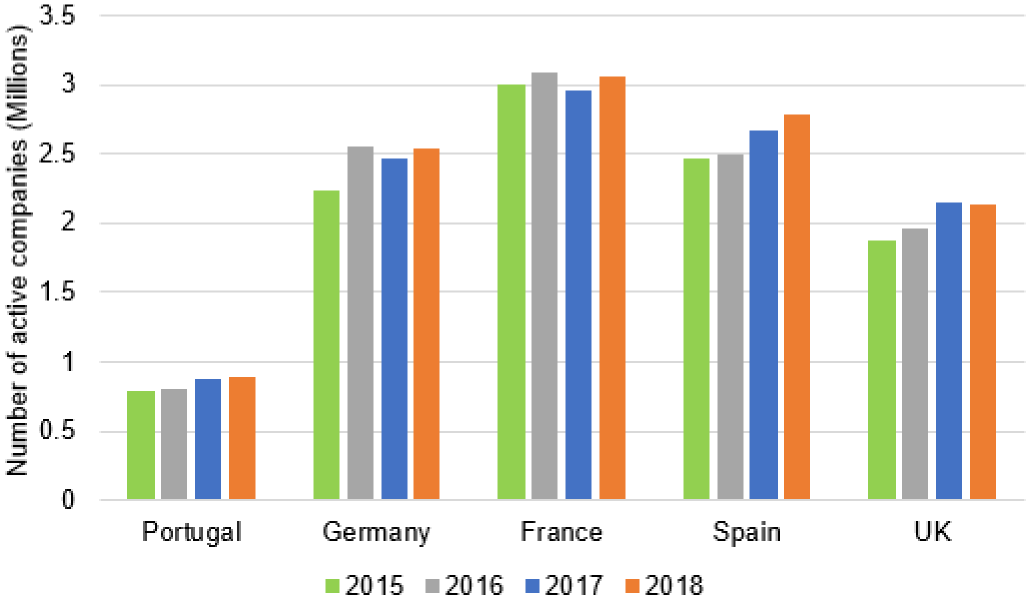


Figure 5- Number of active companies in some European countries over the years (Source: Structural Business Statistics Database- Eurostat).

In figure 5 are described the estimates of the number of active companies in various European countries between 2015 and 2018, covering the 'non-financial business economy', which includes industry, construction, trade, and services, but not enterprises in agriculture, forestry and the largely non-market service sectors such as education and health. As it can be seen, only Portugal and Spain increased the number of active companies in each year, in this time period. In the case of Germany and France the number of active companies didn't increase for two consecutive years, which may indicate that these countries have already reached a financial stable state. Lastly, United Kingdom increased the number of active companies until 2017, and in 2018 has suffered a non-significant decrease. Analyzing the countries set it's perceptible that in this time period Europe has lived with financial health, once there are no large negative variations in the number of active enterprises.

This economic stability was also achieved by a paradigm shift in the credit concept, if the credit was practiced as it was a few years ago, there would probably be more insolvencies in these years, which could result in a decrease of active companies, slowing the economic growth.

2.3. Credit Rating Agencies

CRAs play an important role in the world of credit, and the main objective of these enterprises is to assess the financial strength of borrowed entities, and their ability to meet their financial obligations. The ratings developed by these agencies provide a measurement of companies' solvency and the likelihood that they will default in their commitments. These ratings aim to be a support for investors or business partners, allowing them to assess the risk associated with their financial decisions, such as the amount of payment that they are comfortable to grant to their customers. The main agencies, that control about 95% of the rating business, are Standard and Poor's, Moody's Investor Services, and Fitch Group, and all three are North Americans (Corporate Finance Institute, 2019).

The ratings are scored through alphabetic codes, and they are very similar across agencies. The rating AAA for Fitch and Standard and Poor's and Aaa for Moody's is the highest credit rating which represents a negligible risk of default, and companies that belong to this rating are financially healthy. On the opposite side, the rating C for these three agencies is the worst possible, where corporations that included in this rating represent a high risk of default.

	Standard and Poor's	Moody's	Fitch
Investment Grade	AAA AA A BBB	Aaa Aa A Baa	AAA AA A BBB
Speculative Grade	BB B CCC CC C	Ba B Caa Ca C	BB B CCC CC C

Table 2- Rating Codes according each agency (Source: BBVA).

As it can be seen in the table 2, there are two major groups of ratings: Investment Grade (corporations with financial stability and low risk of default) and Speculative Grade (corporations with high risk of default). In this sector it's mandatory for agencies to make a detailed assessment, once agencies' decisions have a huge impact on the financial environment. For instance, losing a rating status may result in financing more expensive or it may be harder to obtain financing (Parient, 2017).

Although the CRAs mentioned above are market leaders, the reputations of these were damaged following the 2008 financial crisis, a large part of the risk managers took risk based in the public credit rating. Given that the 2008 financial crisis wasn't predicted by CRAs, so the role of these agencies was changed, and was created a regulation to control them. Another action that was taken, was the creation of new CRAs, with the purpose of not depending solely on the Big Three (Standard and Poor's, Moody's, Fitch). In this way, new agencies began to appear, and with new methodologies. Whereas the Big Three are reliable when assessing big organizations, new agencies (private agencies) have focused on the accurate assessment of Small and Medium Enterprises (SMEs), once it's required different approaches between these two types of organizations (Zanders, 2016).

To build their ratings, these agencies have to consider several factors that can affect the credit risk of certain company, such as the type of service that the companies perform, or the country where the company is based. These elements must be considered, since they have an impact on the companies' behavior, so companies should be evaluated in different ways when are in different environments. Each agency has its own methodology to assign companies to ratings, taking into account their financial indicators. The worse the credit rating, the higher the loan risk, which means that it's harder to the borrower receive a loan, and in cases that the company can receive it results in an increase of interest related with the loan. On the other hand, a good credit rating means that it's easier to take a loan to other organizations (Parient, 2017).

Credit rating of SMEs it's more complex when compared to big companies, once the management tools that this type of companies possess are insufficient, there are lack of management knowledge, among other issues. The great advantage of having recourse to private agencies is that there is a deeper research in the company considering the environment where it's involved, so, the assessment is made according the entity, and not only considering the debt issuances (Vairava Subramanian & Nehru, 2012).

2.4. Credit Limit Models

Being the credit so important for entrepreneurships, it's crucial to define well the credit limit for companies, in order to ensure that the supplier is not harmed. To prevent this from happening it's necessary to develop tools that facilitate customers financial analysis, and as mentioned before models appear as the best solution.

Searching for models to analyze the credit limit, there are two statistical techniques that can be used: Multiple Linear Regression or Multiplicative Models. MLR is a statistical method that uses explanatory variables as independent variables to predict the outcome of the dependent variable, calculating the linear relationship (positive or negative) between these variables. This technique, when using the right data, can make predictions about the desired variable based on the information given for another variables (Kenton, Multiple Linear Regression – MLR Definition, 2019). Although it is a method that is easy to understand and use, it has some limitations which can affect the validity of the model. In addition to all explanatory variables having to be independent of each other, the observations must be multivariate normal distributed, and being that the non-normality of the data could be as a result of its nature, it cannot be changed. However, there are various methods of data treatment such as logarithmic transformation, square root transformation, among others, that suggest normality but optimality is not commonly met (Paul E., Dan Dan, & I. Sidney, 2015).

In this way, to solve the problems with non-normality of the data it can be adopt MM. This method also predicts outcomes and linear relationship between independent variables and dependent variable, using a logarithmic transformation of the data. Applying this technique, the problem of non-normality of the data which is present in MLR models, is solved once it modifies the data distribution

(Osborne, 2002). So, the objective of an application of the logarithmic transformation is the reduction of the variability of data, and to make the data closer to the normal distribution (Wang, et al., 2014).

Technologic advances have allowed credit suppliers to acquire, manage and analyze financial data of borrowers aiming to build more robust and strong financial systems. In this way, Artificial Neural Networks techniques are emerging, having several advantages compared to statistical methods, which includes not having to assume certain data distributions. In order to implement this type of technology it's only required "training" samples to automatically extract the knowledge of how to set the credit limit, while taking into account the financial data present on these samples. These samples must be diversified terms of context, so that it will be possible to explore multiple companies' financial scenarios (Ghodselahi & Amirmadhi, 2011).

However, there are some disadvantages that should be referred, once it can be decisive when deciding which type of method or technology to adopt. The main drawback of AI methods is the complexity and vagueness of the models, given that these models work as a "black box" where it's not possible to assess the process that generate results, and to understand which variables are important for the model. Basically, in this case, where the goal is to define the credit limit for companies considering their financial data, the process doesn't tell how the credit was set, only calculates it taking into account the data that was submitted in the training sample, and "copy" the procedure to calculate the credit limit for companies which their credit limit is unknown. Apart from this disadvantage, there are others issues that can be found in AI methods such as: results can suffer of overfitting or overtraining to the training sample and the selection of the model architecture can be very complex (Danenas & Garsva, 2018).

Considering all the methods already mentioned, it's necessary to analyze each one, their strengths and weaknesses, and conclude which one will be the most accurate.

3. Theoretical Background

3.1. Multiple Linear Regression

Linear Regression is one of the most commonly used statistical methods to analyze the influence of variables in the outcome variable. This method can be simple or multiple, and the only feature that distinguishes them is the number of variables that predict the outcome. Multiple Linear Regression seeks to model the linear relationship between the dependent variable (the one which is intended to predict) and the independent variables, and these last ones explain the variation of the dependent variable. This technique has been widely applied in credit limit studies, in order to search which financial indicators or data are relevant when setting a credit (Abu Bakar & Mohd Tahir, 2009).

The general MLR model describes the relationship between k independent variables, X_j, and dependent variable, $\widehat{Y_\iota}$, as in the following equation:

$$\widehat{Y_\iota} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mathcal{E}_i \qquad (1)$$

There are $p=(k+1)$ parameters β_j, $j=0,1,\ldots,$ k, these values are regression coefficients, representing the effect on the dependent variable $\widehat{Y_\iota}$, of changing one unit in each explanatory variable in the model. $\mathcal{E}_i$ is the random error associated to each group of data, representing the possible difference between the observed and the predicted values. This model describes one hyperplane in k-dimensional space of regressors $\{X_j\}$.

When using this approach, there are several key assumptions that have to be made:

- The residuals of the regression (difference between observed and predict values) should be normally distributed;
- MLR assumes that there is no multicollinearity in the data. Multicollinearity indicates when independent variables are highly correlated with other;
- MLR assumes homoscedasticity, which indicate that residuals are the same among all values of the independent variables (Statistics Solution, 2019).

3.1.1. Stepwise Selection

In order to find the best model to calculate the credit limit, will be used for variables selection the Stepwise method. This technique is a combination of the forward and backward selection techniques, once in each step in which a variable is added to the models, all variables in the model are checked to analyze if each one is significant, and if there is a non-significant variable in the model, it's removed from there. It should be noted that the significance of a variable in the model can change when adding news ones to the model. So, one variable that was added to the model in the previous iteration, could be insignificant in the model in the next iteration. Using this selection method, it's required to define two significance levels: one for adding variables and other for removing them, through statistical F-test. Applying this test, it's necessary to define two cutoff values for F-test: F_{in} and F_{out}. If the F value of the variable is less than F_{in} the variable is added to the model, and if the F value of the variable which is already in the model is less than F_{out}, the variable is removed from the model. It's needed to give special attention when defining F values, once if F_{in} > F_{out} there is the possibility to the procedure get into an infinite loop, meaning that the process has no end (NCSS Statistical Software, 2016).

The major advantages of Stepwise method are:

- The ability to manage large amounts of potential predictor variables, choosing the best independent variables from the available ones;
- It's the fastest automatic model-selection methods;
- Analyzing the order in which variables are added or removed can provide valuable information about the effect of each one on the model (Stephanie, Stepwise Regression, 2015).

3.1.2. Least-Square Method

The Least-Square Technique aims to define the hyperplane in k-dimensional space that fits better according the data. As already mentioned above, residuals are the difference between observed and predicted value. This approach defines the plane in k-dimensional space in order to minimize the sum of the residuals square (A. Marill, 2004).

Considering that $\widehat{Y_i}$ is the predicted value through the model, and Y_i is the observed value, the residual of each observation (e_i) is calculated:

$$e_i = \widehat{Y_i} - Y_i \qquad (2)$$

Being n the number of observations, the Sum of Squares Error/Residuals (SS_E), which is the squared difference between the observed and the predicted value, is calculated through:

$$SS_E = \sum_{i=1}^{n} e_i^2 \qquad (3)$$

So, the main objective when using the Least-Square Method is to minimize SS_E. The smaller the SS_E the better the regression adjustment is, once the difference between the predicted and observed values is smaller.

3.1.3. Model's Significance (F-test)

The significance test seeks to check if any independent variable, X_j , in the model contributes significantly with information to explain linearly the variation of the dependent variable. To perform this test (F-test), the error term ($\mathcal{E}_i$) is required to be normal distributed and independently distributed with mean equal to 0. The hypothesis test to perform is:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ (null hypothesis)} \qquad (4)$$

$$H_1: \beta_j \neq 0 \text{ (alternative hypothesis)} \qquad (5)$$

The null hypotheses (4) means that there are no variables in the model that explain linearly the variation of the dependent variable, in other words, the model has no predictive capability. If the null hypothesis is rejected, then there is at least one independent variable in the model that explains linearly the variation of the dependent variable, and the alternative hypotheses must be accepted.

This test is based on the Sums of Squares, that are defined for equation (3), and the following ones, considering that $\overline{Y_i}$ is the mean of the observed observations (365 DataScience, 2020):

$$SS_T = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2 = SS_R + SS_E \tag{6}$$

$$SS_R = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2 \tag{7}$$

The main driver of this test is the F_0, once this value defines whether the null hypothesis can be rejected. The F_0 is calculated by:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} \tag{8}$$

The rejection criterion is given by:

$$F_0 > f_t = f_\alpha[k, n-p] \tag{9}$$

So, the null hypothesis is rejected with a significance degree α, if F_0 is greater than f_t (tabulated value for distribution F) with regression's degrees of freedom equal to k, and residuals' degrees of freedom equal to n-k. In the case which H_0 is rejected, it can be concluded with $(1-\alpha) \times 100\%$ of confidence, that the model is significant, and at least one predictor in the model is relevant for the response.

The analysis of the model's significance can be made through ANOVA table, where the p-value is given, and if this value is less than α, it is concluded that H_0 is rejected, so the model in consideration is significant. It must be noted that the fact the model is significant doesn't mean that this is the most appropriate model to predict the outcome variable (Bremer, 2012).

3.1.4. Coefficient of Determination (R^2)

There are several methods to assess the suitability of the model. In the measure of fitting the model, R^2 appears as one of the most widely used statistical tool, assessing the goodness of fit of the model. This coefficient estimates the proportion of the dependent variable that's explained through the regression of all predictors in the model (Renaud & Victoria-Feser, 2010). R^2 is given by:

$$R^2 = \frac{SS_R}{SS_T} \tag{10}$$

The coefficient ranges between 0 and 1, and the closer R^2 is to one, the greater is the proportion of the total variation of the dependent variable that is explained by the independent variables included in the regression model. However, a high R^2 doesn't necessarily mean that the regression model is a good adjustment, once when a variable is added to the model, the value of this coefficient always increases (when adding a variable to the model the Sum of Squares of the Regression rises). For this reason, including more independent variables in the model cannot be efficient, knowing that the more independent variables the model includes, the higher R^2. Thus, models with a high coefficient of determination can yield predictions unreliable, not being the best indicator about de degree of model adjustment (Schneider, Hommel, & Blettner, 2010).

Aiming to find a solution for this issue, some researchers prefer to use the adjusted coefficient of determination (R^2_{Adj}). R^2_{Adj} is calculated using the following equation:

$$R^2_{Adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right) \qquad (11)$$

This coefficient considers the number of independent variables that are included in the model, so if a non-explanatory variable is added to the model, the value of this coefficient generally decreases. In this way, when there is a big difference between R^2 and R^2_{Adj}, it means that there is the chance that non-significant variables are included in the model. Therefore, the adjusted coefficient of determination penalizes the insertion of unnecessary variables when calculating the dependent variable (Bremer, 2012).

3.1.5. Variable´s Significance

In MLR it's important to analyze each regressor that's included in the model, in order to find what will be the potential of each one. If the model is significant, by doing the F-test, there are evidences that at least one variable in the model is useful to predict the dependent variable. Making an individual analysis to the variables is essential, once when it's included a non-explanatory variable to the model, the degree of adjustment to data is weakened (Siegel, 2016).

To perform this analysis computationally, it's only necessary to observe the p-value related with each variable, and if this value is smaller than the significance degree α, it's concluded that the considered variable is significant in the model. On the other hand, if the p-value is greater than α, the variable in question is not relevant in the model, so it can be excluded from this one (Bremer, 2012).

The p-value for each independent variable tests the null hypothesis that the variable isn't correlated with dependent variable. If there is no correlation between them, the p-value will be higher than α and the null hypothesis is accepted. As mentioned above, if the p-value is smaller than the significance degree, the null hypothesis is rejected and it's assumed that the variable has effect on the dependent variable (Frost, How to interpret P-values and Coefficients in Regression Analysis, 2018).

3.1.6. Multivariate outliers' detection

As it's known the presence of outliers can have misleading effects on the results of statistical techniques. For this reason, it's mandatory to search for outliers that can affect the research and once the credit limit may depend on more than one variable, the outlier analysis must be multivariate (Riani & Zani, 1995).

One of the most used approaches for detecting outlier is by computing the Mahalanobis Distance (MD). MD measures the distance between two points in multivariate space. The Mahalanobis Distance calculates the distance from the centroid of the total of the observations to the observation i, through the following expression:

$$MD = \sqrt{(x_i - \bar{x})S^{-1}(x_i - \bar{x})^T} \qquad (12)$$

Where the parameters are:

- x_i- Vector for a given data observation;
- $\bar{x}$ – Arithmetic mean of the data set or estimated centroid of the data set;
- S – Sample covariance matrix.

MD for a data set with n independent variables has a chi-square distribution with n degrees of freedom. The classical approach of outlier detection uses the estimates of the Mahalanobis distance, by plugging in multivariate sample mean, $\bar{x}$, and covariance matrix, S, and tags as outlier any observation which has a MD^2 lying above a predefined quantile of the chi-squared distribution with n degrees of freedom (Ghorbani, 2019).

However, this technique involves serious drawbacks, especially when there is more than one outlier. The computation of the MD relies on the sample mean, which is not a robust indicator in the presence of several extreme values. Outliers don't necessarily have a high value of MD. In fact, a cluster of outliers could attract the centroid of the total of the observations and consequently increase the value of the variance. Consequently, there may be a possibility of outliers being considered as legitime observations, and this phenomenon is called as Masking effect. On the other hand, good observations may be incorrectly identified as outliers due to the influence of outliers on the mean, which is the Swamping effect (Chiang, 2007). In this way, this approach is sensitive to the presence of outliers, once its parameters are influenced in the same way by each observation.

On the other hand, in the present research it's mandatory to analyze each observation and understand the pattern where it's inserted. Being that, the credit limit is a variable that considers a wide range of values, there is the possibility of the credit limit being calculated considering different classes. In other words, different classes may have different regressions to calculate the credit limit, such as different weights for each explanatory variable included in the model. For example, credit limit below 30 thousand euros can be defined by regression P, while between 30 and 100 thousand euros is defined for regression L. For this reason, what may seem like a set of outliers, can in fact be evidences about a possible cluster of observation that are not described as the majority of observations, and assumes a different pattern.

It can be concluded that when studying the credit limit of a corporation, which has a wide range of values, can be complex to guarantee the non-legitimacy of an observation or a set of them. Therefore, great caution is necessary when assessing each observation, given that as opposed to being a set of outliers, it may be signs of the real behavior of the variable.

3.2. Multiplicative Models with logarithmic transformation

Logarithmic transformations are often recommended for skewed data, such as monetary measures. This technique generally has the effect of spreading out clumps of data and bringing together spread-out data.

Aiming to improve the normality of the variables, the logarithmic transformation of the variables is a commonly used method, when dealing with quantitative analysis of data. Many statistical procedures assume that the variables are normally distributed, which corresponds a significant violation. In this way, logarithmic transformation appears as one of the best solutions to counteract this trend. Before defining which transformation is more suitable in the data, it's crucial to investigate if the non-normality of the data is due to a valid reason, and not due to mistakes in data entry or non-declared missing values, for example. Among some valid reasons to accept the non-normality are the presence of outliers and/or the nature of the variable itself (Osborne, 2002).

Before applying a logarithmic transformation, the problem must be analyzed, in order to understand whether the transformed model doesn't compromise the original problem, given that the multiplicative model cannot be in accordance with original ideas (Teekens & Koerts, 1972).

In the following graphic is represented the distributions of the credit limit of the companies, that will be considered in the present research.

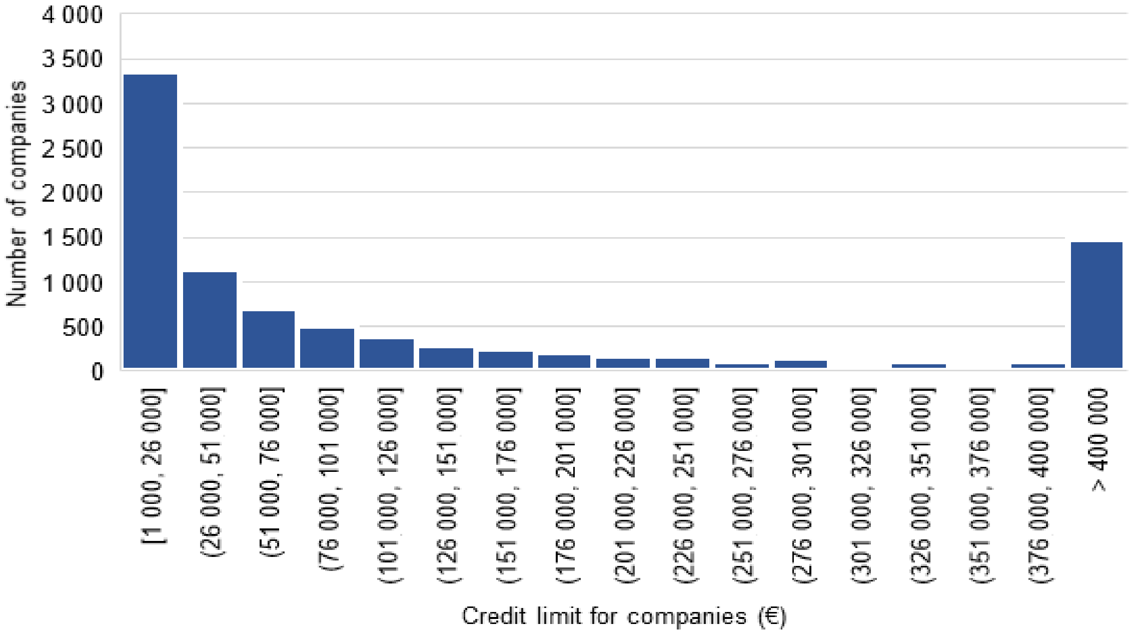


Figure 6- Histogram of distribution of credit limit for companies.

As it can be seen in figure 6 there isn't a normal pattern related with the distribution of the credit limit among companies, something that can be changed applying a logarithmic transformation, as it can be verified in the following graphic.

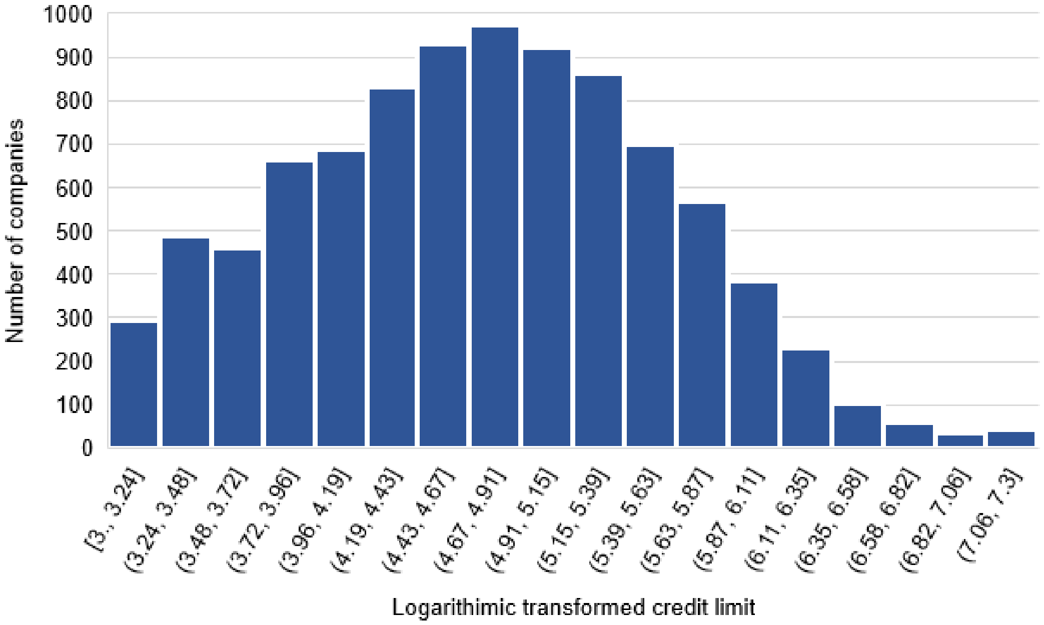


Figure 7- Histogram of distribution of credit limit for companies with logarithmic transformation.

When applying the logarithmic transformation, the distribution of the credit limit of the companies that will be used in this study will take a form similar to the normal one, as shown in figure 7, in accordance with the normality assumption which is made in several statistical procedures.

Comparing the last two graphics is clear the influence of the logarithmic transformation of the data. This transformation increases the normality of the data through a reduction of the relative spacing of the credit limit (in this case) on the right side of the distribution more than the credit limit on the left side. In this way, logarithmic transformation improves the normality of the data by changing the relative distances between data points.

Regarding logarithms, there are several bases, and in the present research is considered the decimal logarithm of base 10, once there is a big range in term of credit limit, or financial data. If the range were smaller it can be used a lower logarithm base, in order to avoid a loss of resolution, which is what happens when are considered higher bases than what is desirable. For example, the credit limit ranges between 0 and millions of euros, which is extremely wide, so a higher logarithm base should be applied (Osborne, 2002).

In order to handle situations where a non-linear relationship occurs between explanatory variables and dependent variable in a regression model, one of the most common ways to do it, is to logarithmically transform variables. Using this transformation allows to preserve the linear model, even if the relationship between variables is non-linear (Benoit, 2011).

Knowing that the general Linear Regression model is described by equation (1), applying the logarithmic transformation, results in:

$$log_{10}\widehat{Y_\iota} = log_{10}\beta_0 + \beta_1 log_{10}X_1 + \beta_2 log_{10}X_2 + \cdots + \beta_k log_{10}X_k \qquad (13)$$

Considering the following properties of logarithms:

$$Alog_{10}(x) = log_{10}(x^A) \qquad (14)$$

$$log_{10}(x) + log_{10}(z) = log_{10}(x\,z) \qquad (15)$$

The model can be described as:

$$\widehat{Y_\iota} = \beta_0 \bullet {X_1}^{\beta_1} \bullet {X_2}^{\beta_2} \bullet \ldots \bullet {X_k}^{\beta_k} \qquad (16)$$

In this way, using the MM approach, independent variables have a multiplicative relationship with the dependent variable instead of the usual additive relationship of the MLR.

One of the most important elements when making a data transformation is the interpretation of the results. In this case, performing a linear analysis with logarithmic transformation, the regression coefficients β_j, can be analyzed in a different way than in the traditional method. Applying this methodology, the coefficients are considered as elasticities of each independent variable included in the model. The elasticity gives the expected percentage of change in the dependent variable $\widehat{Y_\iota}$, when changing a certain dependent variable X_k. So, when increasing Z in X_k, $\widehat{Y_\iota}$ will suffer an increase of $\beta_k Z$ (Benoit, 2011).

Another very important concept about this approach is the residual interpretation. The residual in this case specifies the magnitude of the difference between the observed and the predicted value. Considering the following logarithmic property and the residual's calculation formula in MM:

$$log_{10}(x) - log_{10}(z) = log_{10}\left(\frac{x}{z}\right) \tag{17}$$

$$e_i = log_{10}(Y_i) - log_{10}(\widehat{Y}_\iota) \tag{18}$$

It results in:

$$e_i = log_{10}\left(\frac{Y_i}{\widehat{Y}_\iota}\right) \tag{19}$$

Analyzing the previous equation, the smaller the difference between observed and predicted value, the closer the ratio between them will be to 1. Applying the logarithm, the closer the residual is to 0, the better the accuracy of the model.

3.3. Artificial Neural Networks

Artificial Neural Networks are inserted in the Deep Learning models and have the objective to study regressions and classifications. Deep Learning is where a computer model learns to perform tasks directly from training data. Although the concept of ANN was introduced by Warren McCullock and Walter Pitts in the middle of the 19th century, only in the last few years has there been access to the required processing power for thus type of problems, and the data to train them was also scarce, something that has been changing due to the ease of obtaining data from Internet (Dotai, 2015).

ANN is an information processing technology that is inspired by the way that a human brain works. Basically, this technology is an attempt to simulate the network of neurons that make up a human brain, in order to learn and make decisions in a human way. This technology uses different layers of mathematical processing to analyze the information which is fed in the model.

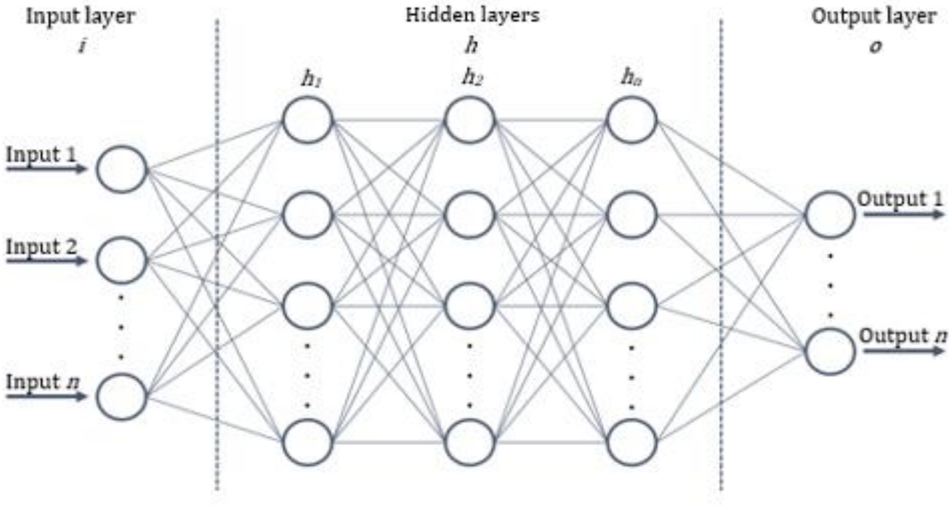


Figure 8- Artificial Neural Network Architecture.

As it can be seen in figure 8, the layers are composed by artificial neurons (also called as units), which are the cells responsible for receiving the input, process it and transform it in an output. The units responsible to transform the input into output are the hidden units, which corresponds to the main process, which is unknown. The connection between units represented in the previous figure are synapses which are really just weighted values, being that each connection has an assigned weight, and the higher the number, the greater influence one unit has on another (Marr, 2018).

When the model's information flows from the input units, through the intermediate calculations present in the hidden layers, and ends in the output it's called feedforward networks. No values are fed back to earlier layers (there are no loops), which doesn't happen in the feedback networks, where the connections between units form a cycle (Gutpa, 2017).

Another important concept is the transfer/activation function, which converts the input signals to output signals through mathematical equations in each neuron. In other words, this function decides whether a neuron should be activated or not by calculating the weighted sum. Without transfer function the model would not be able to recognize patterns, and learn from the training sample (Singh Chauhan, 2019).

There are various types of transfer functions, among them are:

- Threshold or Step function: this is a binary function, basically if the input value is above or below a certain threshold, the neuron is activated and sends the same signal to the next layer;
- Sigmoid function: this function has a "S" shaped curve which ranges between 0 and 1, and it's used mainly for models that predict a probability as an output;
- Hyperbolic Tangent function: this function is similar to Sigmoid function, but has a better performance, and ranges between -1 and 1;
- Linear/Identity function: this is the most used function, and creates an output signal proportional to the input multiplied by the weights for each neuron (Sharma, 2017).

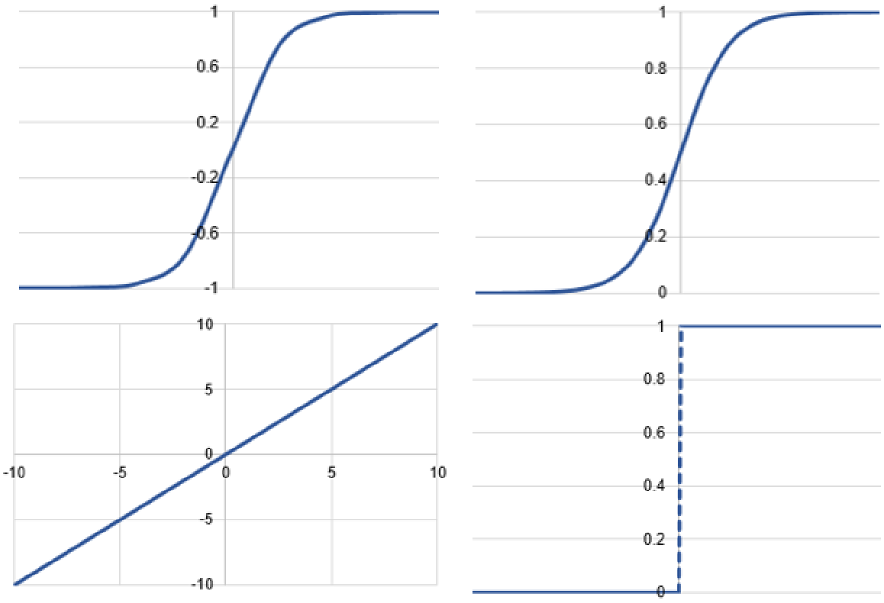


Figure 9- Plots of transfer functions: Hyperbolic Tangent (Top left corner), Sigmoid (Top right corner), Linear/Identity (Lower left corner) and Threshold (Lower right corner).

In figure 9 is described the effect of each transfer function on the input, and during the model's design it's indispensable to analyze which function is the most appropriate, considering the distribution of the variables which are being studied, as well consider between which values these variables can range. Given that, for example, applying a threshold function the variable being studied must be binary.

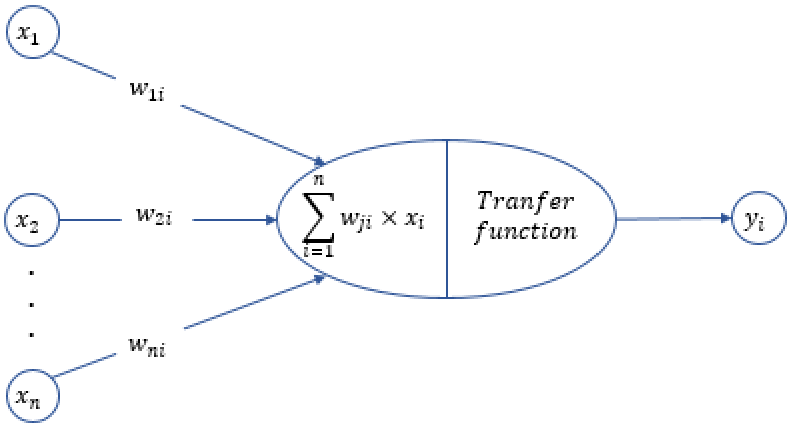


Figure 10- Outline of a neuron and the output process.

Considering the effect of the transfer function, the process to generate the output can be simplified as represented in figure 10, and in this case, it's only considered a single output. In this way, synapses and transfer functions are responsible to convert inputs in outputs and are adjusted to improve the model's accuracy.

3.3.1. Learning Mechanisms of ANN

One very important aspect, when dealing with ANN, is the information that is submitted to train the model, called as training set (Marr, 2018). In ANN there are different types of learning mechanisms, that are divided in three main categories:

- Supervised Learning: This method is defined by the input and output which are both provided in the training dataset. When applying this technique, the training set is composed by correct examples, and in this way the model analyzes the inputs and outputs, learning how these two elements are related, creating an algorithm to predict the outputs. The adjustment of the weights is made directly through the error, seeking to minimize them;
- Unsupervised Learning: Contrary to what happens in the previous method, in this one the training set only provides the input data, without the corresponding output. This technique is popular in applications to uncovering groups within data, or clustering. Should be noted that in this method there are no correct outputs in the training set, once the model itself analyzes the training data and recognizes patterns in data to create outputs (Rouse, 2019);
- Reinforcement Learning: Like in the Supervised Learning method, this method has a specific objective in the network. In this technique, there is no answer, although the model gives a sequence of actions to perform a given task, by considering a system of benefits and penalties which assesses the output and defines the values attributed to the synapses' weights. So, the output depends from this sequence of actions, and the objective is to reach the higher system result, reducing penalties and increasing benefits. The main advantage of this method is the stabilization control of uncertain non-linear systems in the presence of input constraints (Jiang, et al., 2017).

Considering the three types of learning mechanisms mentioned above, the one that is implemented in the present research is the Supervised Learning, once the available data contains inputs and the corresponding outputs, aiming to search for relation between these two elements.

The training set must be very diverse in term of cases, having several situations, in order to the model be able to distinguish each one, and in this way the accuracy of the model will be increased. So, the larger the amount of information given in the training set, the more reliable the model is.

3.3.2. Types of data set to be defined

Before creating the model, it's essential to define the three types of dataset that are required when building an ANN model:

- Training dataset: as mentioned previously, this set is used to train the model, fitting the parameters of the model;
- Validation dataset: this data is used to provide an unbiased assessment of a model fit on the training data while tuning model's hyperparameters, such as choose the number of hidden layers, initial weights and the type of activation function. In this way, the model doesn't learn from this data, it works as a configuration tuner, adjusting the model with the requirements of the study;
- Test dataset: this set aims to evaluate impartially the model fit on the training dataset, and it's only used when the model is completely trained. It must contain various cases that the model could face, once the model can be accurate in certain classes, and in others it can be inappropriate (Brownlee, 2017).

Considering the total amount available for the present research, there are some recommendations that have to be considered about how to split the total dataset into three categories referred above. This decision mainly depends on two factors: the total number of cases in the dataset and on the model, which is being trained. There are some models that need more data to train than others, and in these cases, it's required larger training set. On the other hand, models with few hyperparameters, usually are easier to validate, which means that the validation set can be reduced (Shah, 2017).

One of the major concerns when dealing with ANN models is a phenomenon called overfitting, and describes the cases when the training algorithm runs too long and the model generates values for weights that almost perfectly match the training data, but when the model is used to predict the outcome, its accuracy is very low. Typically, the longer the training stage, lower the error on the training set is, and when the error of this set is almost zero, it's very likely that the models suffers from overfitting.

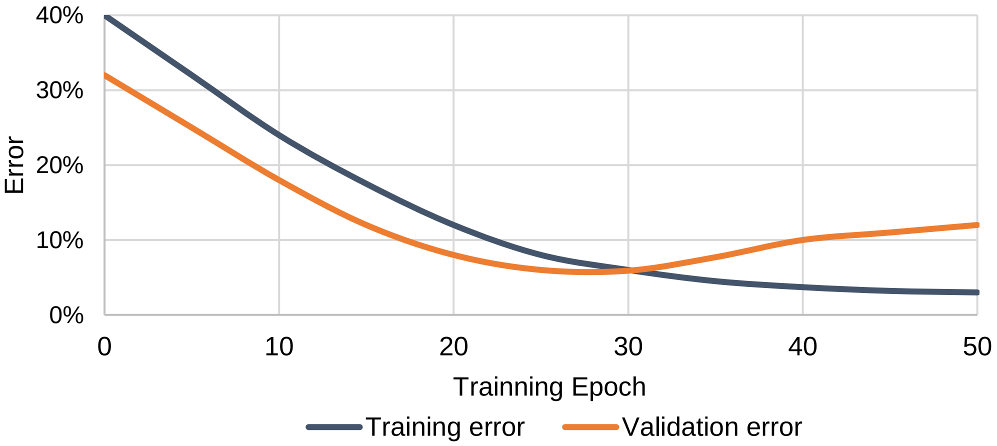


Figure 11- Training and Validation error (Source: Visual Study Magazine).

The train-validate test process identifies when the model overfitting starts to occur, in order to stop the train stage. With the aim of avoiding the occurrence of this phenomenon, the weights values are only updated if the validation set presents a better accuracy rate than the training set, and when this doesn't happen the likelihood of overfitting starts to appear. In figure 11, the training process should stop at the interception the training and validation error lines, which is in the 30^{th} training epoch/iteration, as it is increasingly likely the occurrence of overfitting (McCaffrey, 2015).

3.3.3. Multilayer Perceptron Neural Networks

The Multilayer Perceptron is one of the most ANN used techniques in the credit limit study, and over the years has been tested in several studies. MLP is a system where its network usually contains several layers, being that the first one is the input layer (having no computational role in the model), and the last is the output layer, as it happens in an ANN model. Between these two layers there are an arbitrary number of layers of perceptron (called neurons or units in ANN models) which are able to solve complex problems. The higher the number of layers, the more complex the system will be (Kang, 2017).

When dealing with MLP are often applied supervised learning models, once the input and output are submitted, and these models analyze them, searching for relationships between those two. The model training involves adjusting parameters, such as weights, in order to minimize the error. In the initialization stage, it's recommended that the weights are small enough around the origin, so that the transfer function operates in its linear regime (Kotu & Deshpande, 2019).

For the adjustment of the parameters, it's used the backpropagation algorithm, which basically, initializes the model with random weights values, and through an error measurement these values are refined, once the values change towards a reduction of the of the overall error of the network. With backpropagation, the weights adjustment of the inputs is simplified, given that this technique allows the performance of backward passes which attempt to minimize the difference between the real and the predicted values. The conservation of the transfer function variance it's a desirable property, once this allows information to flow well upward and downward in the network (Nicholson, 2017).

To apply this method is also mandatory to define a learning rate, determining the amount in which the algorithm is moving in every iteration. In most cases, is defined a constant rate, but in some cases, decreasing the learning rate over the time is efficient, and in this way narrow the grid search to the area where it's obtained the lowest validation error (Kumar Kain, 2018).

3.3.4. Radial Basis Function Neural Networks

Radial Basis Function is another commonly used type of ANN in credit limit study. RBF distinguishes from others methods due to its faster learning speed, robustness, accuracy and parallelism (is defined as the case where the same or different inputs are applied to different computational units and the outputs are used independently or are collected in some way to form the total output). This technique is a type of feedforward ANN, which contains exactly three layers, concretely one for inputs, one hidden layer, which is composed by neurons with radial basis activation functions, and at last the output layer, which performs a weighted linear combination of the results from the previous layer (Faris & Mirjalili, 2017).

This method can be described as: *“Each hidden input represents a particular point in input space, and its output, or activation, for a given instance depends on the distance between its point and the instance- which is just another point. Intuitively, the closer these two points, the stronger the activation”* (Witten & Pal, 2017).

Several RBF have been studied, but the most widely used is the Gaussian type, once it’s not only suitable in generalizing a global mapping, but also in refining local features without changing the already learned mapping (Sadeghkhani, Ketabi, & Feuillet, 2012).

The speed of this method is explained by the way it works, apart from the fact that it’s composed by only one hidden layer, which means less complexity of the system. Before starting the study, it’s required to define the minimum desired error (when the model reaches this value, the study stops) or the number of training iterations that must be completed with no improvement (Ahmadian, 2016).

The output of RBF is dependent on three parameters: the input vector, the center and the width of the respective neuron. The input vector is defined for having one dimension for each explanatory variable, and the center of the neuron corresponds to a point with as many dimensions as the input vector. In this way, the similarity of these two vectors must be analyzed through the distance between them. Finally, the spread controls the smoothness of the drop seen in the function, which evaluates the distance between input and center vectors, for greater distances (Militký, 2011).

3.3.5. Comparison between RBF and MLP models

It's useful to make a comparison between RBF and MLP models, in order to understand the differences between these two methods:

- Training stage in RBF models is faster than MLP, given that RBF models only have one hidden layer, and MLP models may have more, which results in more time to train the model;
- The interpretation of the RBF hidden layer it’s easier than MLP models. Once again due to the number of hidden layers;
- The classification of RBF models takes more time than MLP models;
- MLP is applied to a wide variety of problems, more than RBF, given that the first deals better with complex systems;
- Usually the hidden and output layers of MLP models use a non-linear classifier, but when this method is used to solve non-linear regression problems the output layer is linear. On the other hand, in RBF models, the hidden layer is always non-linear, whereas the output layer is linear (Chandradevan, 2017).

Despite the several advantages of the RBF models, there is a major limitation related with this method, which is the possibility of the training process to be trapped in a minimum local, without evaluating the surrounding area. MLP models ensures a better accuracy when dealing with complex problems, where contains a lot of data to be analyzed (Faris & Mirjalili, 2017). For this reason, is expected that the MLP model will have better results than the RBF model, once this study incudes a large amount of data, which increases the complexity of the problem. Therefore, the latter only get better results if the problem at hands is simple to solve and it doesn’t require a sophisticated system.

4. Input Data Collection, Analysis and Treatment

4.1. Input Data Collection Process

In order to perform the present research, it's necessary a large amount of financial data related with thousands of companies. The data used in this study was obtained from one of the biggest financial database in the world. This database is widely used by private institutions, governments and financial worldwide, seeking to assess the credit risk, having access to financial information about others companies, that could be a future business partner. In this way, this platform allows corporations to obtain financial data about thousands of companies, facilitating the decision-making process about potential clients, partners and other type of business decisions.

The regulating entity of this database Bureau van Dijk, a Moody's analytics, and is responsible for the capture and treatment of the data. Bureau van Dijk has more than 160 independent information providers, connecting customers with data that addresses a wide range of business challenges. This global data provider establishes solutions to support credit analysis, investment research, tax risk, among others, in all sectors of business (Schwartz, 2017).

To have access to this database it's necessary to pay a subscription fee. The access to this data set only was possible due to a company's license, containing thousands of companies which are potential business associates or even clients. It should be stressed the fact that the data contains financial information of different companies in different years. In other words, there are cases where the last actualization was in 2015 for a company, for example, but others companies could have latest updates, or could even have older updates. For this reason, it cannot be said that this data corresponds to a specific fiscal year. This aspect has no impact on the credibility of the results, given that, even if the case is very old, all the data of this specific case corresponds to the same year, both financial information and the credit limit. The data set contains financial information of three years, the year of the last update, and the two before, excluding the cases which correspond to companies that were created in the year of the last update.

In the data are included companies headquartered in Spain or Portugal, containing several financial ratios and information (such as total assets, net income, credit limit, among others), risk class and main business sector. In addition to these, the status of the corporation is mentioned, having six hypotheses: active, bankruptcy, dissolved, in liquidation, inactive or status unknown. Generally, only active corporations have an associated credit limit, which means that this type of companies are the main focus of this research.

It's essential to access to the maximum financial data possible, aiming to determine the influence of each variable on the credit limit. One less variable, could mean the loss of an important piece for the credit limit calculation. In this way, the greater the number of the variables, the greater the probability of reaching a model capable of calculating the credit limit. The initial data set, without any kind of treatment, displays about 21700 corporations, with 43 financial indicators. As mentioned previously, for each financial indicator there are information about the three last years that the information was available. The data set was provided through an Excel file, that can be converted in a SPSS document, where the studies will be performed.

4.2. Input Data Analysis

4.2.1. Data Definition

As already referred in section 4.1 of the present research, the data set includes 43 variables: raw financial data, profitability ratios, operational ratios, structural ratios, and others. In the following table are listed in detail all the variables that were extracted from the database.

Type of indicator	Variable	Units
Raw Financial	Credit limit	€
	Fixed Assets	
	Current Assets	
	Shareholders' Funds	
	Non-current Liabilities	
	Current Liabilities	
	Operating Revenue	
	Operating P/L [EBIT]	
	Profit or Loss after tax	
	Depreciation & Amortization	
	Added value	
	Long-term Debt	
	Loans	
	Other Non-current Liabilities	
	Total Assets	
	EBITDA	
	Stock	
	Cash Flow	
Structural	Liquidity ratio	-
	Shareholders Liquidity ratio	
	Current ratio	
	Probability of default	%
	Gearing ratio	
	Solvency ratio	
	Credit period	
Profitability	Interest coverage ratio	-
	EBIT margin	%
	Profit margin	
	EBITDA margin	
	Cash Flow / Operating Revenue	
	ROE using P/L before tax	
	ROCE using P/L before tax	
	ROA using P/L before tax	
	ROE using Net Income	
	ROCE using Net Income	
	ROA using Net Income	

Operational	Net assets tunover	-
	Stock Turnover	
	Collection period	days
	Credit period	
Qualitative	Risk Class	-
	Main business sector	
	Status	

Table 3- Variables exported from database, and corresponding indicators.

As it can be seen in table 3, the data set contains types of financial information, which ensures a more diversified study, fully exploiting the corporations financial situation. Some indicators don't have units due to their own nature, or because they are ratios or because they are qualitative or categorical variables.

With so many types of indicators it's important to characterize them, in order to understand what is the role of each one:

- Raw financial: contains, essentially, information taken from the companies' accounting records, and is easy to interpret. This type of indicator can be tricky, once companies in different positions may present similar results, which can lead to a poor analysis;
- Structural indicators: aims to assess the capability of an entity to pay off its obligations, with its own resources. The variables included in this type are very relevant, once they describe the company's financial health;
- Operational indicators: this indicator encompasses information about the way that the companies conduct the business, measuring some performance indicators, such as turnovers, average time these companies take to pay their obligations, and others;
- Profitability indicators: measures the financial returns of the companies. Generally, the greater the measure, the more profitable the business is;
- Qualitative indicators: this type of indicator gives information about companies' environment and details, which can be preponderant when assessing a company.

In order to perform a good research, it's required a good understanding of the data, given that without this it's impossible to assess if the results make sense. In this way, it's crucial to understand which variables are more likely to be important to build the model that define the credit limit. For example, presumably the variable "Depreciation & Amortization" is not an important variable, once it does not tell anything about the financial situation of the company, so it's expected that this one will not be significant in the model, unless there are strong evidences to the contrary. Despite this, all variables will be tested equally, with no discrimination, searching for indications that support the theory or not. Using this example, was highlighted the importance of a good understanding of each variable.

4.2.2. Formulas for financial ratios

In the previous section are mentioned several financial ratios. Below are listed the formulas necessary to calculate these ratios.

$$ROE\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Equity} \times 100 \quad (20)$$

$$ROA\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Total\ assets} \times 100 \quad (21)$$

$$ROE\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Equity} \times 100 \quad (22)$$

$$ROA\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Total\ assets} \times 100 \quad (23)$$

$$ROCE\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Capital\ employed} \times 100 \quad (24)$$

$$ROE\ using\ net\ income = \frac{Net\ income}{Equity} \times 100 \quad (25)$$

$$ROA\ using\ net\ income = \frac{Net\ income}{Total\ assets} \times 100 \quad (26)$$

$$ROCE\ using\ net\ income = \frac{Net\ income}{Capital\ employed} \times 100 \quad (27)$$

$$Profit\ margin = \frac{Profit\ before\ tax}{Sales} \times 100 \quad (28)$$

$$EBITDA\ margin = \frac{EBITDA}{Operating\ revenue} \times 100 \quad (29)$$

$$EBIT\ margin = \frac{EBIT}{Operating\ revenue} \times 100 \quad (30)$$

$$Net\ assets\ turnover = \frac{Sales}{Equity + Non\ current\ liabilities} \times 100 \quad (31)$$

$$Current\ ratio = \frac{Current\ assets}{Current\ liabilities} \quad (32)$$

$$Liquidity\ ratio = \frac{Current\ assets - Stocks}{Current\ liabilities} \quad (33)$$

$$Gearing = \frac{Non\ current\ liabilities - Loans}{Equity} \times 100 \quad (34)$$

$$Interest\ coverage\ ratio = \frac{EBIT}{Interest\ Expenses} \quad (35)$$

$$Shareholders\ liquidity\ ratio = \frac{Total\ Shareholders\ Equity}{Total\ Assets} \quad (36)$$

$$Solvency\ ratio = \frac{Net\ After - Tax\ Income + Non\ cash\ Expenses}{Short\ Term\ Liabilities + Long\ Term\ Liabilities} \quad (37)$$

4.3. Input Data Treatment

4.3.1. Removal of invalid cases

Before performing any type of study, it's important to analyze the data set, searching for cases that may jeopardize all research. The first data set that was removed from the analysis is related with the cases that didn't have an associated credit limit, once without the dependent variable, cases are not valuable for the research. There are three possible causes for missing values in the credit limit:

- Unavailable data to calculate the credit limit of a certain company;
- Mathematical error when calculating the credit limit, such as divisions by zero;
- Conversion errors between different softwares, as it can happen when transferring data from database software to Excel.

Companies with an associated credit limit of 0 were also excluded from this analysis, given that these cases may jeopardize the statistical models, because it's the minimum threshold of the monetary unit, and there is no differentiation between cases that may have completely different financial data.

Another analysis of the data that had to be carried was the main business sector of the corporations. Entities that are financed by governments were excluded from the study, given that, these types of companies don't usually become insolvent because they have State aids, something that doesn't happen with private companies. So, private corporations behave in a completely different way than public institutions. For this reason, companies which have the following main business sectors were not considered in this research:

- O- Public administration and defense; compulsory social security;
- P- Education;
- U- Activities of extraterritorial organizations and bodies.

Among the cases present in the sectors mentioned above are: schools, embassies, states, non-profit foundations, and others.

It should be noted the fact that some variables were soon excluded of this research and are not even mentioned in section 4.2 of the present research, due to the fact that are empty variables, which means that for all cases these variables don't contain any valid value. Are the cases of the following variables: Gross margin, Enterprise value / EBITDA, Market cap / Cash flow from operations, Export revenue / Operating revenue, and R&D expenses / Operating revenue.

After this data treatment, 9237 available cases were counted for the study.

4.3.2. Correlation Analysis

Correlation analysis is a statistical method which aims to evaluate the strength of the relationship between two quantitative variables. The higher the correlation between two variables, the stronger is the relation between these two, while a weak correlation indicates that there is no relationship with each other (Franzese & Iuliano, 2019).

When performing a regression, it's desirable that all variables be independent of each other, what doesn't happen when there is a high correlation between variables. One evidence of multicollinearity is when changing one variable, the other one also changes, which indicates a relationship between them.

The introduction of highly correlated variables may affect the model, once it's difficult for the model to estimate the relationship between each independent variable and the dependent one independently because the independent variables tend to act in accordance with each other. So, if two highly correlated variables are inserted in the model, at least one of them isn't adding value to the model, so it must be excluded (Rogers & Boyd Enders, 2013).

This analysis isn't so important for MLR models, once during the building of these models, a collinearity analysis is performed with the variables that are included in the model. On the other hand, through the selection method Stepwise, the multicollinearity is taking into account when choosing which variables are included in the model.

By using SPSS software, there are different approaches to search for evidences of correlation between independent variables. The simplest and fastest technique is to analyze the variance inflation factor (VIF) for each variable. If there are signals of collinearity, this measure can detect the degree of the multicollinearity between the variable in question and the remaining independent variables. The VIF may be calculated by the following expression:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad (38)$$

In this formula, R_j^2 is the multiple correlation coefficient, which gives the proportion of variance in the independent variable j associated with the remaining independent variables. Values of VIF exceeding 10 are often source of concern, indicating multicollinearity. A VIF value of 10 implies a R_j^2 equal to 0.9, which means that 90% of the variability of the variable j is explained by the remaining independent variables. Examining only the correlations coefficients may be helpful but not sufficient, given that it's possible to have a set of variables with no high correlation, but several variables together may be highly interdependent (Midi, Sarkar, & Rana, 2010).

In the table below are listed all the VIF values of all variables present in the data set of this research.

Variable	VIF	Variable	VIF
Fixed Assets	84 128 801	Shareholders Liquidity ratio	1.1
Current Assets	21.2	Current ratio	1.9
Shareholders' Funds	16.5	Gearing ratio	2.2
Non-current Liabilities	$1.611 \cdot 10^{10}$	Solvency ratio	2.1
Current Liabilities	25.2	Net assets turnover	1.4
Operating Revenue	8	Stock Turnover	1.1
Operating P/L [EBIT]	7.2	Collection period	1.4
Profit or Loss after tax	773	Credit period	1.4
Depreciation & Amortization	219	Interest coverage ratio	1.2
Added value	17.3	EBIT margin	116
Long-term Debt	6.1	Profit margin	104
Loans	2.1	EBITDA margin	137
Other Non-current Liabilities	2.6	ROE using P/L before tax	161
Total Assets	43 912	ROCE using P/L before tax	174
EBITDA	-	ROA using P/L before tax	171

Variable	VIF	Variable	VIF
Stock	15	ROE using Net Income	158
Cash Flow	1 561	ROCE using Net Income	141
Probability of Default	1.7	ROA using Net Income	165
Liquidity ratio	1.9	Cash Flow / Operating Revenue	113

Table 4- VIF values of all variables present in the original data set.

As it can be seen in table 4, there are variables that show clear evidences of multicollinearity in the original data, containing several variables with a VIF much greater that any threshold that can be applied. In order to solve collinearity problems, it was adopted the following methodology: the variables were removed iteratively until no VIF values were over 8. By adopting a limit of 8, a comfortable margin is being given for variables that have similar behaviors but don't depend on others. For example, it's normal that the higher the revenue of a company, the higher the assets will be, although these variables are different, but may behave in a similar way.

Without any type of analysis, Fixed Assets, EBITDA, Total Assets, Cash Flow and Non-current Liabilities were removed from the data set, once these variables display very high VIF values. After removing these variables, it was necessary to perform a deeper analysis about the remaining variables, seeking to determine which variables may be correlated.

In this step, in addition to analyzing VIF values, it was also analyzed the Pearson correlation coefficients between variables, to search for strong relationships between them. Through this analysis it becomes clear the strength of the relationship between some variables. The following table presents the highest correlation coefficients detected and the respective pairs of variables.

Pair of variables	Pearson correlation coefficient
ROE using P/L before tax and ROE using Net Income	0.984
ROCE using P/L before tax and ROCE using Net Income	0.982
ROA using P/L before tax and ROA using Net Income	0.986
EBITDA margin and Cash Flow / Operating Revenue	0.939
Profit margin and EBIT margin	0.933
Stock and Shareholders funds	0.924
Shareholders' funds and Added Value	0.898
Current Assets and Current Liabilities	0.936
Cash Flow and Profit or Loss after tax	0.968
Current Liabilities and Added Value	0.869
Depreciation & Amortization and Shareholders' funds	0.850

Table 5- Highest correlation coefficients and the respective pairs of variables.

Considering table 5, some variables were excluded from the present research, in order to eliminate the strongest relationships between independent variables, and thus solve multicollinearity problems. Consequently, the following variables were removed from the research:

- Current Liabilities;
- Depreciation & Amortization;

- Added Value;
- Stock;
- EBIT margin;
- ROE using P/L before tax;
- ROCE using P/L before tax;
- ROA using P/L before tax;
- Cash Flow / Operating Revenue.

Through the removal of these variables there is no more evidence of multicollinearity between them, as it can be seen in the table below, by analyzing the VIF values of variables after the removal of these variables.

Variable	VIF	Variable	VIF
Current Assets	5.7	Gearing ratio	2
Shareholders' Funds	7.4	Solvency ratio	2
Operating Revenue	7.2	Net assets turnover	1.3
Operating P/L [EBIT]	5.9	Stock Turnover	1.1
Profit or Loss after tax	5.2	Collection period	1.4
Long-term Debt	4.8	Credit period	1.4
Loans	1.6	Interest coverage ratio	1.1
Other Non-current Liabilities	1.9	Profit margin	3.4
Probability of Default	1.7	EBITDA margin	2.4
Liquidity ratio	1.9	ROE using Net Income	3.7
Shareholders Liquidity ratio	1	ROCE using Net Income	4.5
Current ratio	1.9	ROA using Net Income	4.4

Table 6- VIF values after removal of some variables.

As intended, no VIF value is above the established limit of 8, mitigating correlation problems present in the initial data set.

4.3.3. Outlier Analysis

Outlier detection belongs to one of the most important tasks in data analysis and treatment. An outlier is an element of data set that distinctly stands out from the rest of the data. In statistics, is an observation that is distant from other observations, not corresponding to the possible overall pattern that the other data respect.

There are two types of outliers:

- Univariate: it's a data point that consists of an extreme value of only one variable;
- Multivariate: it's a combination of unusual data of at least two variables (Govindaraj, 2018).

Outliers can also be characterized according to the respective environment:

- Point outliers: data point that is distant from the other observations;
- Contextual outliers: it's a data point or subset of points that significantly deviates from the rest of the data points in the same context, and these observations are not outside of the normal range;
- Collective outliers: a subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set (Cohen, 2018).

There are two basic origins of outliers: they can arise from errors (errors can occur in data entry, measurement, and others) or they can be natural, and only be extreme and isolated cases.

As it's known, the credit limit depends on more than one variable, and for that reason it's necessary to perform a multivariate outlier analysis. However, the greater the number of variables that the dependent variable depends on, more sophisticated is to perform an outlier analysis, and so, is more susceptible to misinterpretations. In this way, it's recommended to perform a multivariate outlier analysis when building the models, bearing in mind that in that step there will be fewer variables inserted into the model, simplifying the analysis and making it more rigorous. If the multivariate outlier analysis was conducted by considering all available variables, cases could be excluded due to the impact of variables that are not even important for the credit limit calculation, which could jeopardize the rigor of the present study, by removing cases that can be essential for the performance of this.

At this stage of this research, the main focus is to detect credit limits that are largely distant from the rest of the data, and don't guarantee robustness and consistency of the results. Only one point or a subset of points may grossly distort the analysis, influencing the whole model. For example, if the highest credit limit present in the data set is equal to 100 million euros, and the second highest is 1 million euros. So, it's clear that there is a huge difference between one case, and the rest of the cases (equal or below of 1 million euros). And this extreme case can have repercussions on the model, as the model will adapt to this case, which doesn't guarantee robustness of the results.

	Percentiles						
	5	10	25	50	75	90	95
Credit limit	2 000	4 000	13 000	55 000	220 000	670 000	1 300 000

Table 7- Distribution of the credit limit by percentiles.

In table 7 is described the distribution of the companies' credit limits present in the data set, by using an SPSS descriptive analysis. As it can be seen, 25% of the total companies have an associated credit limit equal or lower than 13 thousand euros, whereas 50% of the companies have a credit limit equal or lower than 55 thousand euros. It's notorious the fact that the higher the credit limit, the lower the concentration of the companies, and for this reason, if there are outliers, they are outliers with high credit limit.

By analyzing the boxplot of the credit limit present in the data set (shown in Appendix B), it's clear that there is no case extremely isolated from the others. Due to this fact, it's not necessary to remove any case, without a deeper analysis. The highest associated credit limit to a company in the data set is 20 million euros, and from credit limits of more than 10 million euros it's evident a greater dispersion but there are no evidences that these cases will prejudice the results.

4.3.4. Risk Class Classification

As mentioned before, the Risk Class, Status and Main business sector are qualitative data, so they need to be treated differently than others variables. The Main business sector and Status only will be considered after building the model, aiming to improve its results, such as separate cases by business sector or status probably will improve results. The Risk Class is encoded for the purpose of to be able to enter in the model as an independent variable, since to perform a statistical analysis it's required that all variables be the same type, it cannot be used scale with nominal variables.

Code	Risk Class
10	AAA
9	AA
8	A
7	BBB
6	BB
5	B
4	CCC
3	CC
2	C
1	D

Table 8- Risk Class: a cataloging scale for probabilities of default to categorize each company.

In table 8 is described the code which will define each Risk Class to perform statistical analysis, being that the best class (AAA) is represented by a 10 and the worst class (D) by a 1. Through numerical encoding, the variable Risk Class switches from nominal to scale, enabling its possible participation in the model.

4.3.5. Sampling Method

As already mentioned in section 3.3.2 of the present research, the data set must be divided mainly in two groups: training and test samples. Regarding the cases, these two samples must be equal for all approaches and methods performed in this research, in order to guarantee the same conditions for all techniques. For example, if two different approaches are assessed with different samples, the comparison between them it's not rigorous, once these methods are not being tested under the same conditions, given that the performance of the model depends largely on the cases which are submitted.

In this step the validation data set it's ignored once this type of data it's only necessary for ANN methods, not being practiced in statistical techniques, and since that in IBM SPSS Statistics 25 there are not all hyperparameters available, it's not worth to considerer this type of data.

To ensure a fair assessment and comparison, it's also necessary to diversify the two samples, to cover all type of cases. In this way, it's mandatory to include several classes of cases in these samples, seeking to create balanced samples. For instance, if the training sample contains mainly cases of companies with an associated credit limit higher than 100 thousand euros, and the test sample only includes companies with an associated credit limit lower than 100 thousand euros, the assessment of this method would be a bit pointless once samples are not balanced. In this specific example, probably the model would be accurate for companies with an associated credit limit higher than 100 thousand euros, but the test sample would only consider companies with credit limit below 100 thousand euros, and for this reason, the assessment would not be representative of the all cases, once the samples don't include all classes possible.

The main limitation when defining the two types of samples is the possibility of some cases may only be available for certain models, in accordance with the variables which are included in these models. The cases available depend largely on the variables that are considered in the respective model, once if the model includes a certain independent variable, all cases that don't have a valid value for this variable are not considered. Therefore, it's expected that the greater the number of variables inserted in the model, the smaller the number of cases available for study.

In this way, it's complex to define the test sample (which is the sample responsible to perform the comparison and are not used in the model building phase) compatible with all models. For this reason, the methodology adopted was to define about 15% of the data set as the test sample, and the remaining 85% the training sample, and after building models, determine which cases of the test sample were available for all methods. The cases that are part of the test sample, and are available for all methods will be considered as the comparison sample between different models and techniques.

Classes	Training cases		Test cases		Number of Total cases per class
	Number	Percentage	Number	Percentage	
≤ 30 000 €	3 092	85.5%	525	14.5%	3 617
≤ 100 000 €	1 759	84.7%	317	15.3%	2 076
≤ 250 000 €	1 243	84.8%	222	15.2%	1 465
≤ 500 000 €	730	84.4%	135	15.6%	865
≤ 1 000 000 €	536	85.0%	94	15.0%	630
≥ 1 000 000 €	522	89.3%	62	10.7%	584
Total	7 882	85.3%	1 355	14.7%	9 237

Table 9- Number of cases for each sample and their percentage by class.

Table 9 describes the number and percentage of cases for each sample considering the respective class. It should be noted the fact that these classes are not a financial indicator representative of the financial health of the company, since credit limit itself is not representative of the corporation financial situation. These classes were only created taking into account the credit limit of companies, to show that the proportion of the test sample was constant across all classes.

As it can be seen in the table, and as expected all classes have a test sample size a round of 15% of the total cases of the respective class, except the class that contains the highest credit limits, which its test sample is about 11% of the total class. There could be classes with a big percentage of test sample, and others with a small percentage. In this case, consequently there would be the possibility, of the total number of cases present in the test sample would also be equal to 15% of the total cases, being this sample unbalanced. In this way, it's ensured a balanced sample test, which allows to perform a fair comparison between all models.

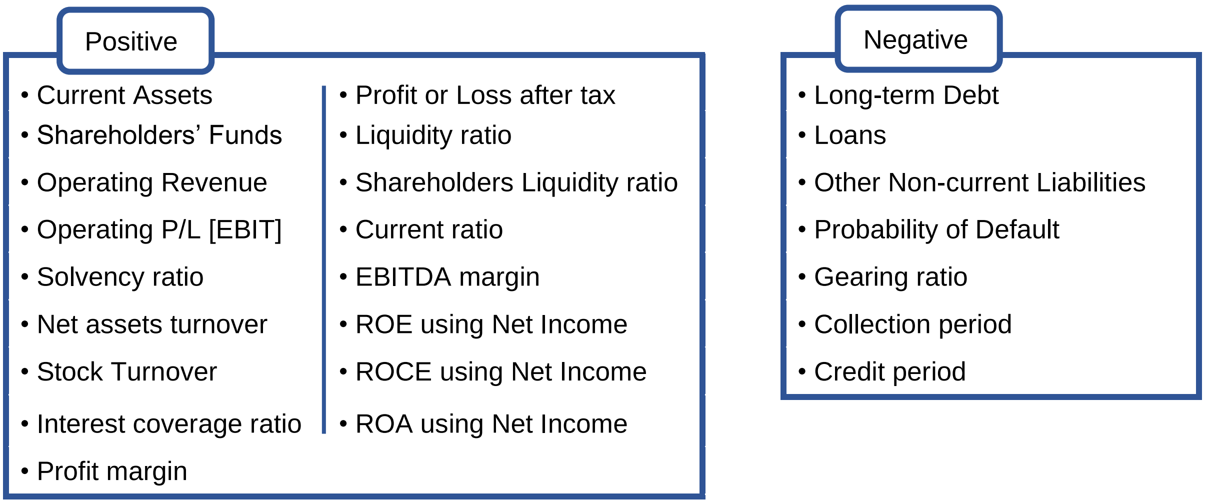
5. Model Building

5.1. Multiple Linear Regression

5.1.1. Relationship between Credit Limit and explanatory variables

One of the most important prerequisites before applying MLR study is to have a detailed knowledge of each variable that may enter in the model. In addition to this, it's necessary to have a notion about the potential relationship between the dependent variable and the explanatory ones, such as if certain variable has a negative or positive effect in the credit limit. For example, there are 2 different companies in identical financial situation, and the only factor that distinguishes them in financial terms is the value of the debt. As expected, if the debt is a preponderant factor in calculating the credit limit, the company with the lowest debt will have a higher associated credit limit than the one with the biggest debt (if the debt is not used in calculating the credit limit, both companies will have the same credit limit). In this specific example, the credit limit is negatively correlated with debt, which means that if this independent variable enters the model, its sign will be negative.

This step is crucial to perform a rigorous MLR analysis, once one of the critical tasks is to assess whether the variables that entered in the model have the right signal. Given that the model is developed in a computational mode, some independent variables may be included in the model only to improve the model results but the presence of these variables makes no sense. Some variables which are not even important for calculating the credit limit can be included in the model with the wrong signal of correlation, and in this type of cases there are evidences that the variable is not useful in the model, once the signal of correlation must be respected, and consequently, the model is excluded from the analysis, once it contains wrongly defined variables.



Positive

- Current Assets
- Shareholders' Funds
- Operating Revenue
- Operating P/L [EBIT]
- Solvency ratio
- Net assets turnover
- Stock Turnover
- Interest coverage ratio
- Profit margin
- Profit or Loss after tax
- Liquidity ratio
- Shareholders Liquidity ratio
- Current ratio
- EBITDA margin
- ROE using Net Income
- ROCE using Net Income
- ROA using Net Income

Negative

- Long-term Debt
- Loans
- Other Non-current Liabilities
- Probability of Default
- Gearing ratio
- Collection period
- Credit period

Figure 12- List of variables according to its type of correlation with credit limit.

In figure 12 are discriminated the type of correlation (positive or negative) between all the independent variables that are considered in this study and the credit limit variable. As previously stated, in order to be considered a credible model, all the signals of the correlation between independent variables and credit limit must be respected, variables which are positive correlated must have its term positive in the model, while variables which are negative correlated must have its term negative in the model. In this way, for each model developed must be verified all variables included in the model, to prove the validity of its presence in the model.

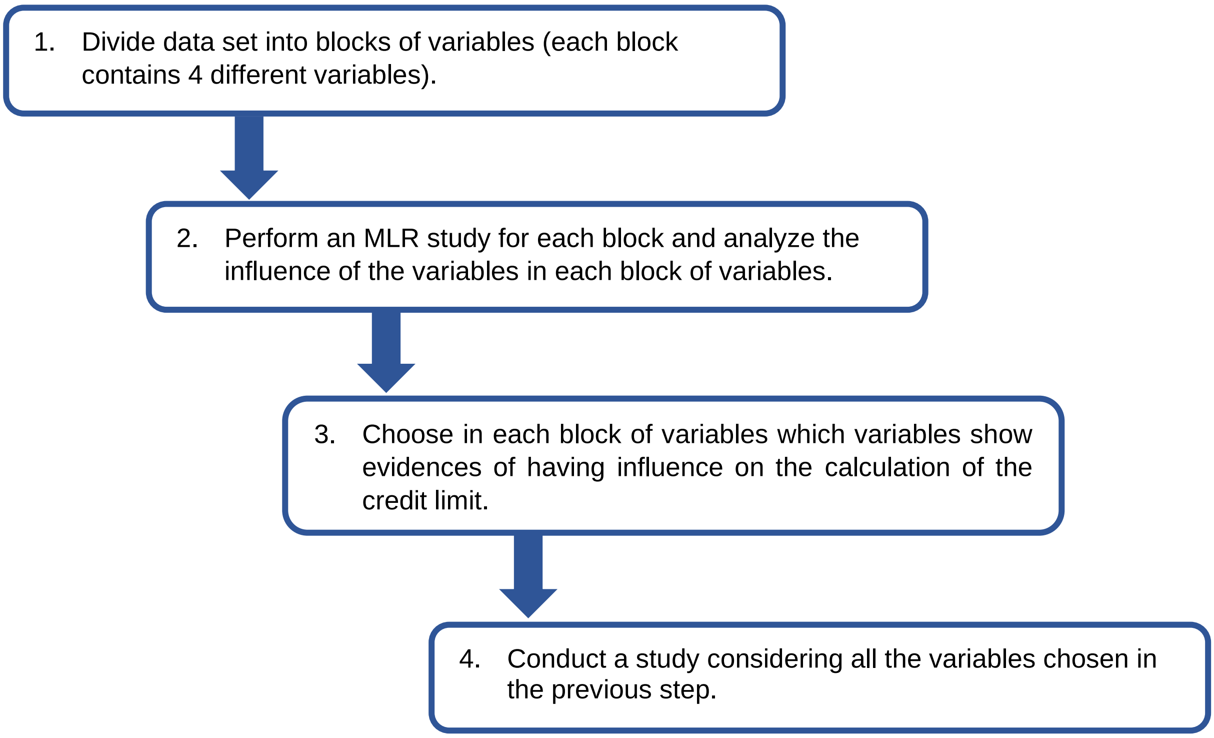
5.1.2. Methodology for models' analysis

The first approach that was adopted, with the aim of detecting which variables are important for credit limit calculation, was to divide the data set into small blocks of variables. Through this approach, instead of performing a study with all variables together (which wouldn't be worth, once some useless variables for credit limit calculating would be included in the model), it allowed to analyze each variable more carefully, observing the impact of each one. In this way, by using this technique there is a more detailed study about each variable, having a greater notion about the influence of each variable.

The basic principle of this approach is to perform MLR study for each block of variables, and due to the Stepwise selection method, variables that are useless for credit limit calculation are excluded from the model created by each block of variables. It's also important to analyze each variable that is included in these models, checking if their signals correspond to those specified in figure 12, since if with few variables in consideration, a variable entering with the wrong signal is an evidence that this variable has no strong relationship with the dependent variable.

On the other hand, the R^2_{Adj} of each model developed from each block of variables must be a focus of analysis, to understand how strong the relationship between the credit limit and the independent variables included in this model is. The greater the value of the R^2_{Adj} of each model created from each block of variables, considering that all variables included in the model have the right signal of correlation, the more likely it is that at least one of the variables included in the model is really important for calculating the credit limit.

After the assessment of each model developed from each block of variables, an MLR study is carried out considering only the variables given as important for the credit limit, being that the previous steps aim to filter out the variables are not necessary for the credit limit calculation.



1. Divide data set into blocks of variables (each block contains 4 different variables).
2. Perform an MLR study for each block and analyze the influence of the variables in each block of variables.
3. Choose in each block of variables which variables show evidences of having influence on the calculation of the credit limit.
4. Conduct a study considering all the variables chosen in the previous step.

Figure 13- Representative scheme of the first procedure adopted for the analysis of variables.

Figure 13 summarizes the first procedure that was taken to analyze each variable, in order to perform a rigorous MLR study. It should be noted the fact that, only in step 4 is developed the model to calculate the credit limit, in other steps are created models just to assess the importance of each variable.

After performing this procedure, it was quickly concluded that it is not an appropriate approach when conducting an MLR study. The main limitation of the previous method is that when performing an MLR analysis for each block of variables, the results generated from the block consider the model itself, not the variables. For example, one given variable when belonging to a block of variables may not even be included in the model due to the Stepwise selection method, but when belonging to other block of variables may be included in the model and show evidences of being an important variable for the credit limit calculation.

On the other hand, the R^2_{Adj} of each block of variables is not a robust indicator of the presence of an important variable for credit limit. The R^2_{Adj} largely depends on the conjunction of variables that are included in the model, and for this reason models from each block of variables may present a low R^2_{Adj} because the right variables are not being combined. For instance, when developing a model with 2 variables, the model may present a low R^2_{Adj}, but if one of the variables is combined with other variables, there may be a large increase in R^2_{Adj} because this coefficient assesses the variables as whole. A variable alone or combined with useless variables may present low R^2_{Adj} values, but when combined with useful ones can be responsible for a large increase of the R^2_{Adj} of the model that calculates the dependent variable. In this way, the R^2_{Adj} cannot be an indicator of the presence of important variables, once the block of variables are made randomly.

For the reasons mentioned above, the methodology of the analysis of the variables had to be changed, ensuring that no variable is disregarded. The procedure that was adopted is based on a trial-error process, and the robustness of the study is only achieved when many attempts are made, with the aim of testing a large number of variable combinations.

This study begins by considering only one random independent variable, and runs an MLR analysis with this variable. Then, if the variable previously considered has any influence on the credit limit, and its signal respects the corresponding signal of correlation, is added one more independent variable to the study, and an MLR analysis runs again. At this step, it’s important to verify if both variables were included in the model, and if there was an improvement of the R^2_{Adj} over the previous model. If in fact, there wasn’t a decrease of the R^2_{Adj}, and if one of the variables was excluded from the model and the variable that entered in the model had the right signal of correlation, the excluded variable must be excluded from the current analysis. Alternatively, if the two variables were included in the model, and if there was an improvement of the R^2_{Adj} and their signals of correlation are correct, these two variables should be kept in the study.

This process of including and excluding variables from the study should be repeated until there are no new variables to be inserted into the current analysis. So, although it is a very repetitive and exhausting process, being the only solution to develop a reliable model for credit limit calculation, after several attempts.

Another very important detail when performing this process, it's the importance of each variable in the model. There is a possibility that there are variables to be included in the model but they have minimal impact on it, and may even be a random relationship with the dependent variable. From the need to perceive the significance of each variable in the model, arises the p-value of each variable, as explained previously in section 3.1.5 of the present research. The analysis of this value becomes mandatory in the occurrence of the following situation: when one or more variables arise with the wrong signal of correlation, it's necessary to check whether or not the variables that have the wrong signal, are significant in the model, and if the variables are not, then they can be removed from the model.

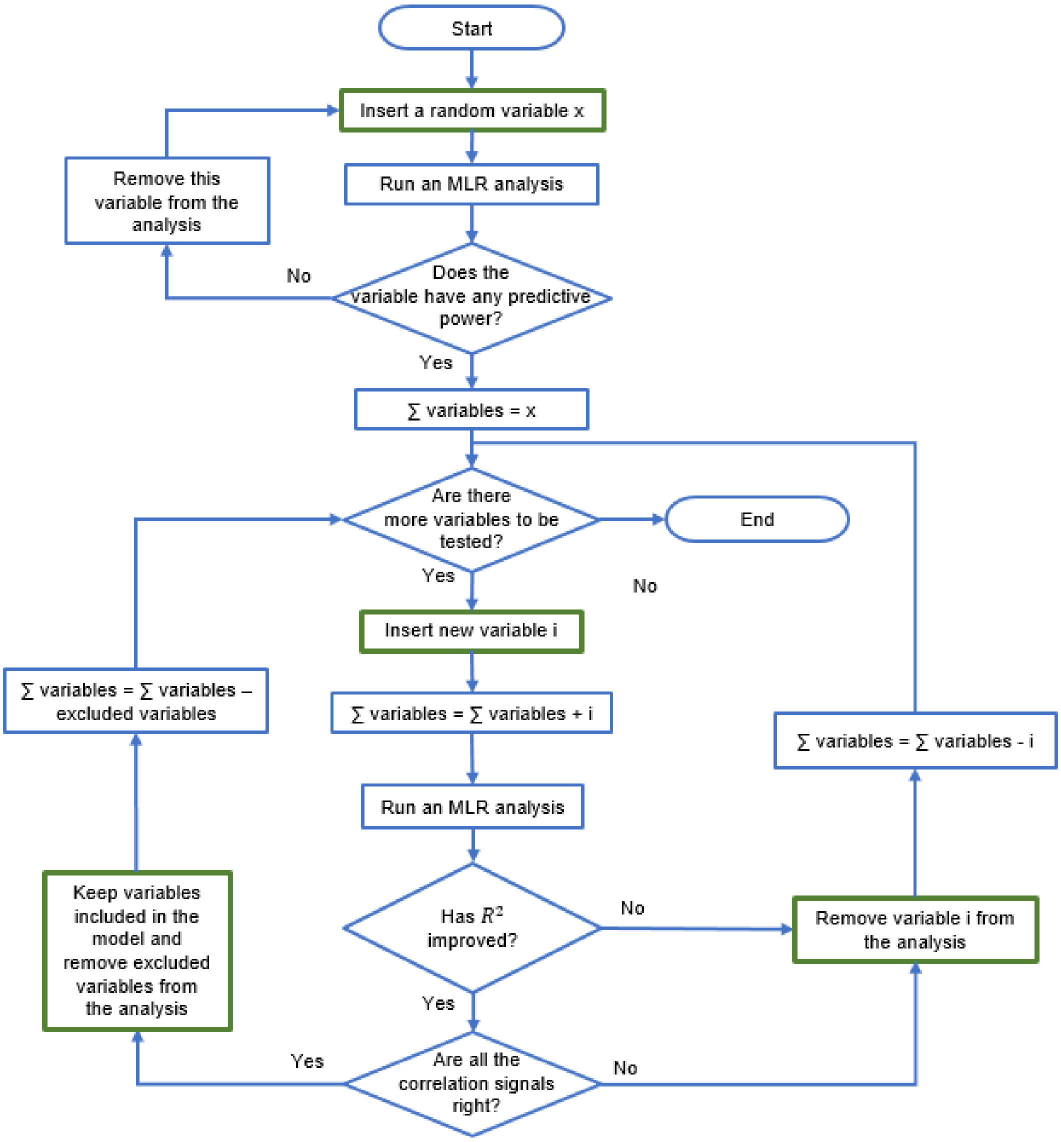

Start
Insert a random variable x
Remove this variable from the analysis
Run an MLR analysis
No
Does the variable have any predictive power?
Yes
∑ variables = x
Are there more variables to be tested?
End
Yes
No
Insert new variable i
∑ variables = ∑ variables – excluded variables
∑ variables = ∑ variables + i
∑ variables = ∑ variables - i
Run an MLR analysis
Keep variables included in the model and remove excluded variables from the analysis
Has R^2 improved?
No
Remove variable i from the analysis
Yes
Yes
Are all the correlation signals right?
No

Figure 14- Flowchart representative of the variables' assessment approach.

In figure 14 is described the adopted process to understand and analyze the impact of each variable present in the data set. It should be highlighted the fact that the previous flowchart it's related to only one attempt, and it's required several ones to find a reliable and robust model, and the sequence of chosen variables must be always different, to ensure that this approach tests various scenarios about variables combinations. An example of a measure to take to guarantee that different scenarios are tested is the choice of different variables at the initial stage of the process, as well as, to change the sequence of the variables inserted in the study. The green boxes represent the steps where one variable or a set of them are inserted or removed from the model.

Another concept that is important to define is the minimum threshold of an acceptable adjusted coefficient of determination. In the present research was assumed that the minimum acceptable R^2_{Adj} for the model to be considered in further analysis and comparisons, is about 75%, unless after several attempts no model achieves this value. Final models with an associated value of R^2_{Adj} below 75% are not considered, once models with values below this don't show a strong relationship with the dependent variable. In this way, in order to be considered a strong model for credit limit calculation it's required its R^2_{Adj} to have, at least, 75%.

On the other hand, if no model with a R^2_{Adj} of 75% or more is developed, there is evidence that: the regression being used (in this case MLR) doesn't fit in the data, the available variables in the data set are not important in the calculation of the credit limit, or even both. Once again, it should be noted the fact that the occurrence of these phenomena only can be concluded after the development of a large number of models.

One important aspect that must be considered, is the fact that there are missing values in some observations. In the data set, which contains 9 237 observations (each observation corresponds to one company), there are missing values in some variables, which means that there are few variables that have values for all observations. As it can be seen in Appendix C, only two independent variables (Current Assets and Shareholders' Funds) contain values for all observations, whereas, in the opposite situation, the variable "Solvency ratio" has only values for 4 946 observations. Missing values in the observations may impair the study, given that if there is an important variable which has many missing values, and due to this the regression could not be able to detect the influence of this variable. For this reason, variables which show many missing values can be overlooked by the study, since for the study in question only can be considered observations that contain values for all variables inserted in the study, which can cause a maladjustment of the regression parameters. It should be also mentioned that the greater the number of variables included in the model, the lower is the number of available observations responsible for developing the model.

5.1.3. MLR models

By using IBM SPSS Statistics 25, the methodology described in figure 14 was put into practice. As already mentioned in section 3.1 of the present thesis, it was necessary to customize some configurations to perform the desired analysis:

- Variable selection method was changed for Stepwise;
- In the section of Statistics, which is the section where are included different analysis that can be shown when the study is conducted, were selected Model fit, Descriptives, Durbin-Watson ratio and Normal P-P plot;
- When applying the Stepwise selection method is mandatory to define the stepping method criteria. Variables can be entered or removed from the model depending on the significance of the F value, which is an arbitrary value, depending on the specifications defined for the inclusion or not of variables in the model (Pope & Webster, 1972). A variable is entered into the model if the

significance level of its F value is less than the Entry value and is removed if the significance level is greater than Removal value. The entry and removal value of F value were the defined ones by default in IBM SPSS Statistics 25, 3.84 and 2.71, respectively. First the software calculates the F value for each variable in the model. If the model contains j variables and being $SS_{E(j-X_R)}$ the Sum Squared Error for the model that does not contain X_R, $SS_{E(j)}$ the Sum Squared Error for the model that contains X_R and $MS_{R(j)}$ the Mean Squared Error for model that contains X_R, then F value for any variable (X_R) is calculated through the following equation:

$$F = \frac{\frac{SS_{E(j-X_R)} - SS_{E(j)}}{DF_{X_R}}}{MS_{R(j)}} \qquad (39)$$

If this value is greater than the one specified to remove, the software removes the variables with an associated F value higher than the one specified. After that the software initiates the next step that seeks to add a variable to the model, through the calculation of its F value. If the model contains j variables, and X_a is the variable that is being tested to be included in the model, the expression of the F value is given by:

$$F = \frac{\frac{SS_{E(j)} - SS_{E(j+X_a)}}{DF_{X_a}}}{MS_{E(j+X_a)}} \qquad (40)$$

Where $SS_{E(j)}$ is the Sum Squared error for model that does not contain X_a, $SS_{E(j+X_a)}$ the Sum Squared Error after X_a being added to the model, DF_{X_a} is the degree of freedom for variable X_a, and $MS_{E(j+X_a)}$ is the Mean Squared Residual after the variable X_a being added to the model. If this value is lower than the one specified to enter, then the variable is added to the model (Minitab, 2019).

After the development of several models, two models emerged that distinguished from the others:

		MLR-1 model			MLR-2 model		
		Unstand.	Stand.	p-value	Unstand.	Stand.	p-value
Regression coefficients	Constant	536 474	-	-	26 251	-	-
	Probability of default (x_1)	-260 608	-0.12	0.00	-	-	-
	Current Assets (x_2)	2.5	0.39	0.00	2.4	0.5	0.00
	Shareholders' funds (x_3)	3.4	0.61	0.00	0,4	0.1	0.00
	Operating Revenue (x_4)	-	-	-	0.5	0.2	0.00
	Operating P/L [=EBIT] (x_5)	18.7	0.29	0.00	15.8	0.3	0.00
	Profit margin (x_6)	6 529	0.05	0.00	5 207	0.04	0.00
	Solvency ratio (x_7)	-	-	-	2 458	0.04	0.00
	Long term debt (x_8)	-2.9	-0.42	0.00	-0,9	-0.2	0.00
	Other non-current liabilities (x_9)	-1	-0.2	0.00	-	-	-
	R^2_{Adj}	65.5%			72.3%		
	Probability of F	0.00			0.00		
	N	5 895			3 562		

Table 10- Independent variables selected, corresponding unstandardized and standardized coefficients and statistics for each model.

As shown in table 10, were considered more than one final model, given that their R^2_{Adj} are low (once don't even reach 75%), and each one presents its own particularities, not being evident which one is the most accurate. Given that, no model has an R^2_{Adj} greater than 75%, there are strong evidences that the credit limit calculation model is not defined by a Multiple Linear Regression, unless the credit limit is defined by different regression according to the class of each company. For example, companies with less than 5% of probability of default are defined by regression x, while the others by regression y.

It is worth noting that in table 10 are only listed variables that were included in at least one model. The remaining variables showed no evidence of importance for the calculation of the credit limit by applying an MLR analysis. It should be noted that the number of observations displayed in the previous table is only considering the number of observations used to develop each model, excluding the observations used for comparison between models.

In addition to presenting only the unstandardized coefficients of each variable, it's important to express the respecting standardized coefficients, once these coefficients allow a direct comparison between all variables included in the model, considering different scales (according to each variable), like if all variables were converted to the same unit (Bhalla, 2016). By observing the standardized coefficients of both models, it was concluded that the variable which contributes most to the calculation of the credit limit in MLR-1 model is "Shareholders' funds", while in MLR-2 model is "Current Assets".

As it can be seen, the model MLR-2 is the one with the largest R^2_{Adj}, however this model contains significantly less observations than the other model. Having a very small number of observations compared to the total sample of observations is a major limitation, once the R^2_{Adj} value may be due to the fact that the model really fits the regression that calculates the credit limit, or due to the reduction of the number of observations which may decrease the variability of the sample, increasing the R^2_{Adj} of the model. In this model, the significant reduction in observations is possibly caused by the inclusion of the independent variable "Solvency ratio", which contains many missing values, which results in a reduction of the number of observations.

Regarding the t-value of the independent variables included in both models, it should be noted that all variables are significant in the model. Once the assumed significance level α, is 0.05, and the t-values are quite high, the respective p-value is 0, which indicates that variables actually have significance in the models.

Considering their unstandardized coefficients, the two models are described by the following expressions, accordingly to the coding present in table 10, and X being the input vector:

$$\begin{aligned} MLR1(X) = 536\ 474 - 260\ 608x_1 + 2{,}5x_2 + 3.4x_3 + 18.7x_5 \\ + 6\ 529x_6 - 2.9x_8 - x_9 \end{aligned} \tag{41}$$

$$\begin{aligned} MLR2(X) = 26251 + 2.4x_2 + 0.4x_3 + 0.5x_4 + 15.8x_5 + 5\ 207x_6 \\ + 2\ 458x_7 - 0.9x_8 \end{aligned} \tag{42}$$

By analyzing the results of the models and considering the properties of the MLR study there is one major conclusion that has to be made. This regression has an additive character, so doesn't make much sense to sum different ratios, once this type of information doesn't take into consideration the dimension of the company. In contrast, to calculate the credit limit by applying an MLR, it makes more sense to sum variables that are in the units of the credit limit, in euros. As evidenced in table 10, the presence of ratios in the models was not privileged, given that in MLR-1 model only was included one ratio (Profit margin), while in MLR-2 model were included two ratios (Profit margin and Solvency ratio). The rest of the variables included in these two models are raw financial data, which probably confirms the theory that ratio itself doesn't have much power to predict credit limit, only if combined with raw financial variables.

5.1.4. Residuals analysis

The residuals analysis has an important role in the validation of a regression model. To be considered as valid, the residuals of the model must satisfy four conditions:

- Be approximately normally distributed;
- Have a mean of 0;
- Have a constant variance (homoscedasticity);
- Be independent of one another (Anderson, 2020).

In order to check whether the residuals are approximately normally distributed, it can be used the normal probability plot (Normal P-P plot). The standardized residuals are plotted against a theoretical normal distribution in such way that if this plot of the residuals is approximately linear, it is verified that the residuals are normally distributed (Pardoe, 2019). The standardized residual is given by:

$$St.Residual(i) = \frac{e_i}{St.Deviation\ of\ Residual(i)} \qquad (43)$$

Once one of the defined outputs in IBM SPSS Statistics 25 was Normal P-P plot, this graphic was automatically generated. Below, are described the Normal P-P plots of both MLR models.

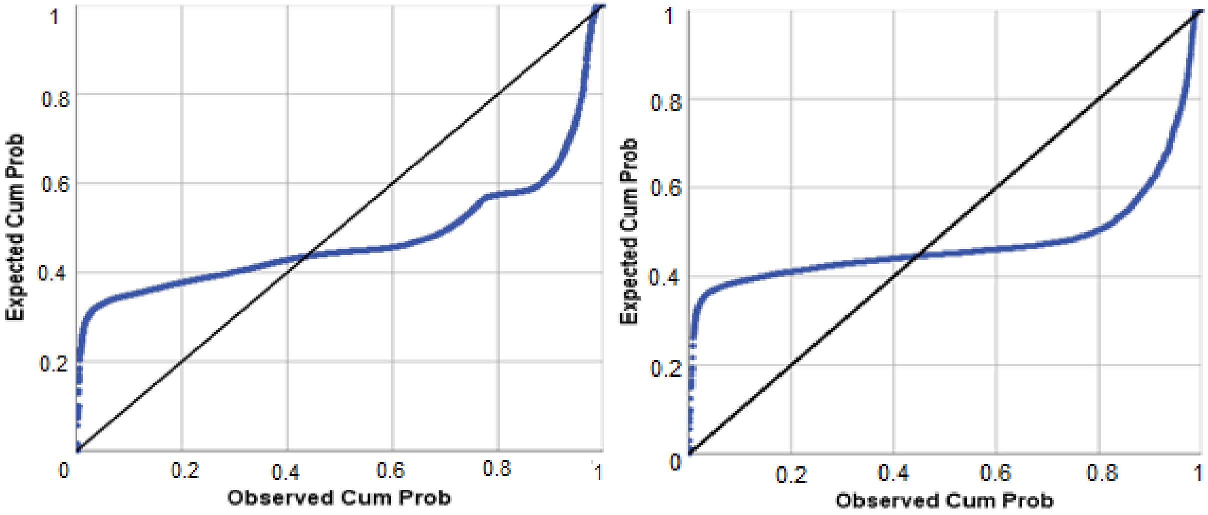


Figure 15- Normal P-P plot of regression residuals of MLR-1 (on the left) and MLR-2 (on the right) models.

As it can be seen in figure 15, in both plots there are significant deviations from the line, which indicate the non-normality of the residuals. It can also be noted that despite the two different models, the distribution of the residuals is quite similar. In this way, the assumption of normality of the residuals is not verified in both models, putting into question the validation of the MLR models.

Regarding the homoscedasticity of the residuals, it can be proven by applying the Breusch-Pagan test. This test aims to accept or reject the hypothesis of homoscedasticity of the residuals (Su, 2017). The process contains four main steps:

1. It's necessary to square all residuals provided by each observation of each model.
2. Regress squared residuals on independent variables included in each model. In other words, develop an auxiliary model where squared residual is the dependent variable, and the independent variables are the ones included in the primary model about credit limit calculation.
3. Multiply the R^2 of this auxiliary model by the number of observations, in order to calculate the Lagrange Multiplier.

$$LM = R^2_{aux} \bullet N \tag{44}$$

4. Under the hypothesis the test statistic χ^2_{BP} follows a χ^2 distribution with k degrees of freedom, where k is the number of independent variables included in the model. If p-value of $\chi^2(k, LM)$ is lower the significance level (in this case α= 0.05), then the heteroscedasticity of the residuals is verified, which means that the homoscedasticity of the residuals is not verified (Zach, 2020).

Considering this procedure, the analysis was conducted for both MLR models, as described in the table below.

Model	N	R^2_{aux}	LM	k	p-value
MLR-1	5 895	0.373	2 198.8	7	0.00
MLR-2	3 562	0.221	787.2	7	0.00

Table 11- Summary of the homoscedasticity verification of the residuals for MLR models.

Through the analysis of the table 11 it becomes clear that the heteroscedasticity hypothesis is not rejected, once the p-values of both models are lower than the significance level of 0.05. So, it can't be assumed that residuals of both models don't have a constant variance.

About the mean of the distribution of the residuals is very simple to prove it, since when running an MLR study in IBM SPSS Statistics 25 it's shown the statistical parameters related with standardized and non-standardized residuals. In this way, by analyzing the information provided by the software, the mean of residuals of both models is 0. Thus, the assumption of the residuals has a mean of 0 in both models is confirmed.

The method adopted to verify the independence of the residuals is based on the analysis of the Durbin-Watson ratio. This ratio measures the autocorrelation in residuals from regression analysis, and it will always assume a value between 0 and 4. A value of 2 indicates that there is no autocorrelation detected in the data set. Values less than 2 mean a positive autocorrelation, while values higher than 2 mean negative correlation. Despite this, it is generally assumed that values between 1.5 and 2.5 are relatively normal, and are not cause for concern in this subject (Glen, 2016). Durbin-Watson ratio is calculated through the following expression:

$$DW = \frac{\sum_{i=2}(e_i - e_{i-1})^2}{SS_E} \tag{45}$$

The fraction numerator represents the squared difference between the residuals of each observation with the one immediately next.

Through the analysis of the output provided by IBM SPSS Statistics 25, the Durbin-Watson ratio is 1.912 for MLR-1 and 1.864 for MLR-2. Both values are very close to 2, which means that there is no autocorrelation in residuals, so it can be stated that the residuals are independent. Therefore, the four assumptions are confirmed, validating both models.

By analyzing appendixes D and E it can be concluded that there is no independent variable that has the role of separating observations by classes, and where each class could be defined by a different credit limit calculation regression. If there was a separation of cases by classes, in the residual plots over an independent variable there would have to be different agglomerations of cases, however this is not the case, as there is no pattern in any graph. It should be noted that, this analysis only was performed for variables that were included in model MLR-1 and MLR-2, respectively, once are those with the greatest influence on the credit limit.

In this way, it's concluded that both models don't respect all the assumptions made, given that the assumption of the normality and the homoscedasticity of the residuals is not verified, putting into question the validity of the MLR models.

5.2. Multiplicative Models

5.2.1. Adopted Methodology and Models developed

Regarding Multiplicative Models, the adopted approach to find the best model to calculate the credit limit is very similar to the approach taken in MLR, described in figure 14. The big difference between MM and MLR is the interpretation of the results, once the studies are conducted in the same way by using IBM SPSS Statistics 25.

While in MLR, the coefficients of the regressions relate to the parameters of the linear regression, in MM these coefficients are interpreted as elasticities as explained in section 3.2 of the present research, however the requisites are the same when analyzing each model. Each variable included in the model must have an associated elasticity that respects the respective signal, mentioned in figure 12, which translates the relationship between the independent and dependent variable. It should be noted that elasticities are equal regarding the signal of correlation between variables in MLR. In other words, variables that have a positive correlation in MLR, in MM have the same signal of correlation, and for this reason, the correlations outlined in figure 12 remain equal in this method.

Although this is another method, all the used options and configurations are the same that were used in MLR. About R^2_{Adj}, the objective to reach a model with a value of, at least 75%, also remains the same as the previous technique.

Once again, one of the major concerns is the inclusion of variables in the model that may drastically decrease the sample of cases available to develop the model, which may consequently reduce the variability of the observations, and lead to an increase of R^2_{Adj}. So, it's necessary to carefully analyze each model, searching for evidences of its validity.

With all the requirements and options established, two models were developed with promising results, as shown in the following table.

		MM-1 model			MM-2 model		
		Unstand.	Stand.	p-value	Unstand.	Stand.	p-value
Regression elasticities	Constant	0.755	-	-	1.652	-	-
	Current Assets (x_2)	0.368	0.382	0.00	0.108	0.109	0.00
	Shareholders' funds (x_3)	0.184	0.198	0.00	0.675	0.710	0.00
	Operating Revenue (x_4)	0.324	0.323	0.00	0.182	0.177	0.00
	Operating P/L [=EBIT] (x_5)	0.063	0.072	0.00	-	-	-
	Solvency ratio (x_7)	0.495	0.189	0.00	-	-	-
	R^2_{Adj}	93.1%			92.2%		
	Probability of F	0.00			0.00		
	N	3 913			7 099		

Table 12- Independent variables selected, corresponding unstandardized and standardized elasticities and statistics for each model.

As it can be seen in table 12, the values of R^2_{Adj} in both models are quite promising, once the threshold of 75% for this coefficient is largely exceeded. Contrary to what happened in the MLR study, applying the MM method there are strong indications that the model responsible for calculating the credit limit is defined by a multiplicative model.

The difference between MM-1 and MM-2 models is the exclusion of the independent variable "Solvency ratio", given that this variable has the fewest observations (the number of observations shown in the table are only the ones which were used to develop the model), and as already explained in MLR study, the decrease in observations may increase R^2_{Adj}, because the variability of the sample used to develop each model may also decrease. For this reason, in model MM-2, the variable "Solvency ratio" was not considered, which consequently led to the exclusion of the variable "Operating P/L [=EBIT]", compared with the MM-1 model. With these two variables not being included in the model, the number of observations increased by about 3 100 cases, and the value of R^2_{Adj} only decreased 1.1%, which could mean that the remaining three variables included in the model may have a greater impact on the credit limit calculation.

By analyzing the standardized elasticities of both models, it can be concluded that in MM-1 model the variable that contributes the most to predict the credit limit is "Current Assets", while in MM-2 model is "Shareholders' funds".

Once again, it should be stated that the variables present in table 12 are the ones that were included in, at least one MM model. So, the variables which are not listed in the table have not shown to be important in any MM model.

Considering their unstandardized elasticities, the two models are described by the following expressions, being X the input vector:

$$MM1(X) = 0{,}755 \bullet {x_2}^{0.368} \bullet {x_3}^{0.184} \bullet {x_4}^{0.324} \bullet {x_5}^{0.063} \bullet {x_7}^{0.495} \qquad (46)$$

$$MM2(X) = 1.652 \bullet {x_2}^{0.108} \bullet {x_3}^{0.675} \bullet {x_4}^{0.182} \qquad (47)$$

The big advantage of this method when compared to MLR is the possibility of including ratios and raw financial data in the same model, given that makes sense to multiply these two types of financial information, however, doesn't make any sense to sum them up.

5.2.2. Residual Analysis

As performed in MLR study, in this method it's also important to analyze the residuals of each model developed, searching for evidences that confirm the four assumptions that are made when applying a Linear Regression. The residuals must respect the four conditions already listed in section 5.1.4: be approximately normally distributed and independent of one another, have a mean of 0 and a constant variance (homoscedasticity).

As made in the previous method, in order to verify whether the residuals are approximately normally distributed, the Normal P-P plots of both models were analyzed

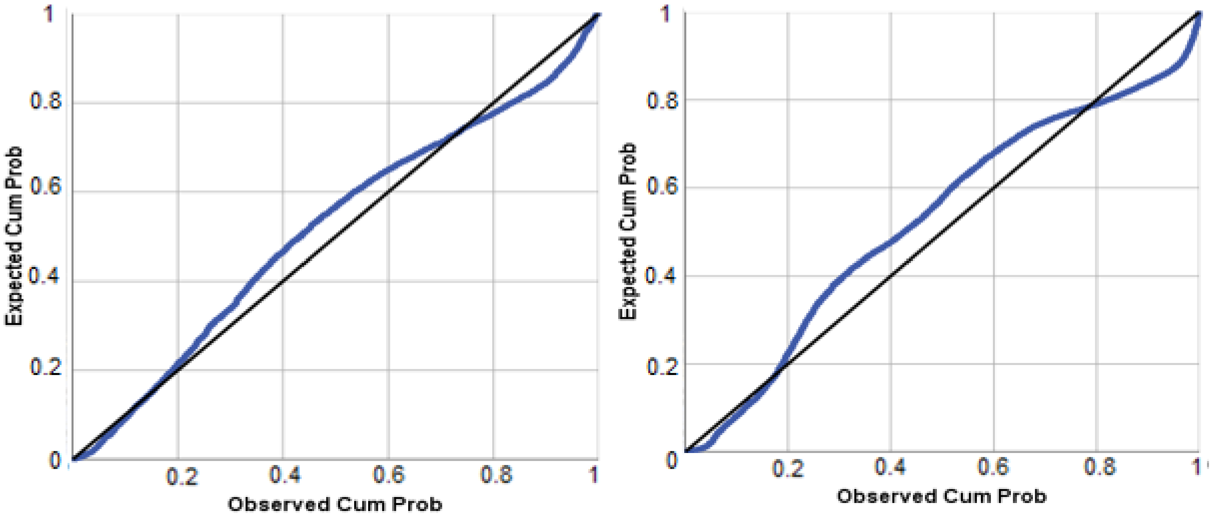


Figure 16- Normal P-P plot of regression residuals of MM-1 (on the left) and MM-2 (on the right) models.

As it shown in the previous figure (figure 16), in both plots is evident that there are only small deviations from the line, and as it is known the closer the points are to the line, the greater the normality of the residuals. It can also be noted that despite the two different models, the distribution of the residuals is very similar. In this way, it can be stated that the normality of the residuals in both models is verified, being that the points fall right on the line, which means that normality has been met. As a comparison between MLR and MM models, it can be observed that the latter show less deviations from the line that defines the normality of the residuals. So, residuals from MM models fits more to the normal distribution.

Concerning the homoscedasticity of the residuals, once again the analysis is made through the application of Breusch-Pagan test, as performed in MLR models. The explanation of how this test is performed, and its theoretical considerations are present in section 5.1.4 of the present research.

Model	N	R^2_{aux}	LM	k	p-value
MM-1	3 913	0.001	5.87	5	0.31
MM-2	7 099	$2.8•10^{-4}$	1.99	3	0.57

Table 13- Summary of the homoscedasticity verification of the residuals for MM models.

As it can be seen in the previous table (table 13), it’s very evident that the heteroscedasticity of the residuals in both MM models is rejected, since their p-values are much higher than the defined significance level of 0.05. In this way, the homoscedasticity of the residuals of the MM models is verified, which means a constant variance.

About the mean of the distribution of residuals, it’s only necessary to read the output generated by IBM SPSS Statistics 25 for each model, as already done for MRL models. Once again, as happened in the previous method, the mean of the residuals of MM models is 0, being confirmed another assumption.

Lastly, only the independence of the residuals remains to be confirmed. In order to check the independence of the residuals, is analyzed the Durbin-Watson ratio of each model. As mentioned in section 5.1.4 of the present research, the values of this ratio are between 0 and 4, and generally values between 1.5 and 2.5 indicate that residuals are independent one another and the calculation of this ratio is given by equation 44. By observing the output given by the software, MM-1 model has a ratio of 1.887, while MM-2 model exhibits a value of 1.892. Given that both values are very close to 2, it can be assured the independence of the residuals of the both models.

All the four assumptions were verified in MM models, contrary to what happened in MLR models, MM models are validated.

By observing appendixes F and G, there isn’t any independent variable included in the MM-1 and MM-2 model which has the role of separate observations by classes defined by different regressions. The analysis performed in this stage of this method are very similar to the ones made in MLR study, so this process is repetitive for each model, and given that there are no patterns present in the charts, the conclusions are the same in both MLR and MM models.

5.3. Multilayer Perceptron

5.3.1. Configurations and options

When applying an AI method, the big concern is to adjust all the configurations and options, to be programmed in the context of the study to be conducted. If these conditions are not meet the study may be impaired, giving completely unsatisfactory results.

Being AI technology autonomous, its methods have their own ways of thinking, and may be constantly changing according to the iteration of the model. In other words, even if the conditions are the same, the method always generates different results, because it constantly changes its problem methodology. It is therefore imperative to ensure that for the same conditions the results are equal. To solve this problem arises the Random Number Generator (RNG), in IBM SPSS Statistics 25. The RNG allows to set the starting sequence value, aiming to reproduce a sequence of random numbers. To put this technique into practice is necessary to select an Active Generator Initialization, to replicate a sequence of random numbers, by setting the initialization starting point value prior to each analysis. The fixed number that is set has no requirement, it just has to be the same in each analysis, and it must be a positive integer. So, the fixed number that was set in the RNG to perform this analysis was the default setting by the software, 2000000.

As already mentioned in section 4.3.5 of the present research, the data set was divided into two groups (training and test group) through the creation of a binary variable, where observations having the value 0 in this variable belong to the test group, and observations having the value 1 belong to the training group. Where the training group is responsible for adjusting the models parameters to make them as accurate as possible, while test set aims to evaluate impartially the model fit on the training dataset, and it's only used when the model is completely trained. So, instead of selecting the option defined by default, which is to associate a percentage to each type of data (whether training or testing), the option of having a variable that makes this distinction was selected.

About the variables, it's necessary to highlight the fact that the order of the selected variables interferes with the analysis' course in IBM SPSS Statistics 25. Which means that even if selected variables are the same but in a different order, the generated results will be different. So, to promote a fair comparison between models, is mandatory to respect the order of the selected variables in all developed models.

Regarding the output when the study is performed it were selected the following options: Description, Synaptic weights, Model summary, Case processing summary and Independent variable importance analysis.

About the observations in the data that have missing values, all these cases will be excluded from the analysis. So, an observation that doesn't contain a value in a variable that is considered in the study will be automatically excluded from the study.

Another configuration that is mandatory when performing an AI technique, is the definition of the stopping rules. The stopping rule that was considered was the maximum number of steps without a decrease in error, 8 in specific. In this way, each analysis can have no more than 8 iterations without lowering the error rate, given that when this condition is achieved the analysis stops and is considered the best result.

One very important aspect that must be considered in MLP is the type of training, which determines how the network processes the records. In IBM SPSS Statistics 25 there are three training types:

- Batch: Updates the synaptic weights only after passing all training data records. It's useful for small datasets (less than 1 000 observations);
- Online: Updates the synaptic weights after every single training data record, that is, uses information from one record at a time. This type of training continuously gets a record and updates the weights until one of the stopping rules is met, and if all the records are used once and none of the stopping rule is met, then the process continues by recycling the data. Online training type ensures a better performance when there are several inputs (independent variables);
- Mini-batch: Divides the training data records into groups of approximately equal size, then updates the synaptic weights after passing one group. This type of training offers a compromise between batch and online training, and fits for "medium-size" datasets (IBM, 2019).

By analyzing these three training types it becomes clear that the most appropriate to the study to be conducted is the Online training. Besides having a better performance when the study includes several inputs and guarantees constant case-by-case updating. Online training type also provides a chance to reuse the available data which represents a big advantage compared with the others training types, once the same observation may be adjusted more than one time, improving the adjustment of the model.

Regarding the architecture of the MLP, there are several options to select. First, the number of hidden layers, which in the software only can be one or two. As it is initially not known which of these options is the best, it's important to test both through the development of two models (one with one hidden layer, and other with two), and in the end assess which one guarantees the best accuracy. The number of units in each hidden layer is set automatically by the software, which means that the software calculates the ideal number of units for each layer.

Beyond the number of hidden layers, it's necessary to define the activation function in these layers and there are two options: Hyperbolic tangent or Sigmoid. Once again, there is no certainty about what's the best option, which leads to the development of two models, each with one option.

About the output layer it's only necessary to define the activation function, being that this choice it's simpler comparing with the hidden layers. As already explained in section 3.3 of the present research, the most appropriate activation function considering the problem, is the linear/identity function, once this function is the only one that doesn't limit its final value, which is required when the answer to the problem is in euros, which is not restricted to a range of values. So, all MLP models have as activation function in the output layer the Linear/Identity.

Therefore, it will be developed the models describe in the next table with different characteristics.

Model	Number of hidden layers	Activation function in hidden layer	Activation function in output layer
MLP-1HT	1	Hyperbolic tangent	Linear/Identity
MLP-1S	1	Sigmoid	Linear/Identity
MLP-2HT	2	Hyperbolic tangent	Linear/Identity
MLP-2S	2	Sigmoid	Linear/Identity

Table 14- MLP models' details.

As listed in table 14, it's necessary to develop 4 models to test all the scenarios and analyze which one is the most accurate.

5.3.2. MLP models

Considering all the configurations and options mentioned in the previous section, all the conditions for the development of models are met.

	MLP Models			
	MLP-1HT	MLP-1S	MLP-2HT	MLP-2S
Number of units (input layer)	24			
Number of units (hidden layers)	9		9 (1st layer) / 7 (2nd layer)	
Training Sample	2 505			
Testing Sample	438			

Table 15- Summary of the four MLP models.

As expected, in all models there are 24 units in input layer once there are 24 independent variables inserted in the study (one unit for each independent variable). As shown in table 15, each model with one hidden layer has 9 units in this layer, while models that have two hidden layers have 9 units in the first layer, and 7 units in the second.

The number of observations available to develop the model is a cause of concern, because, as it can be seen in the table, there are only about 3 000 observations to train and test the model. When applying an AI method, it's important to take as much data as possible in order to "teach" the model well. So, it's crucial to take actions to increase the number of available observations.

The analysis of the importance of each variable in the model arises as a method to reduce the number of variables included in the study, being that variables with a low importance in the model can be excluded from the study. The significance or importance of each variable in the data reflects its effect on the generated model, being a rank based on the contribution that predictors make to the model. Basically is a measure of how much the dependent variables changes for different values of the independent variable (Alabi, Issa, & Afolayan, 2013). The non-normalized importance of a variable is calculated by the sum of the decrease in error when split by a variable through a complex computational process made by IBM SPSS Statistics 25, and then the normalized importance is the non-normalized variable importance divided by the highest variable importance value (Chauhan, 2017).

Through the independent variable importance analysis generated in the model development report, all the variables were analyzed aiming to search for variables that were not important in the four models. Therefore, the criterion was defined that variables with a normalized importance below 10% in these 4 models must be removed from the study. In this way, by setting a minimum threshold for normalized importance of 10%, variables which have an insignificant impact will be removed from the generated model.

By analyzing Appendix H it becomes clear that the following 8 independent variables have a normalized importance lower than 10% in all models:

- Probability of default;
- Shareholders Liquidity ratio;
- Gearing ratio;
- Solvency ratio;
- Stock Turnover;
- Collection period;
- Credit period;
- Interest coverage ratio.

With the removal of these variables is expected that the sample for model development increase, once the higher the number of variables inserted in the study the lower the number of observations is, due to the missing values present in the data.

As a result of the implementation of this measure is necessary to develop again more 4 MLP models, and analyze whether the option of privileging the number of observations by removing independent variables that don't show any evidence of being important in the model is beneficial for the accuracy of the model, although it's difficult to prove it, given that the sample of testing will also be different. In other words, the first four MLP models (MLP-1HT, MLP-1S, MLP-2HT and MLP-2S) can be more accurate only because these models have fewer observations, which reduces the sample variability, something that does not happen with updated models (MLP-UPD-1HT, MLP-1S, MLP-UPD-2HT and MLP-UPD-2S).

	Updated MLP Models			
	MLP-UPD-1HT	MLP-UPD-1S	MLP-UPD-2HT	MLP-UPD-2S
Number of units (input layer)	16			
Number of units (hidden layers)	8		8 (1st layer) / 6 (2nd layer)	
Training Sample	5 369			
Testing Sample	925			

Table 16- Summary of the four MLP models after exclusion of some independent variables.

With the removal of the 8 independent variables mentioned previously, the number of units in the input layer has decreased to 16. About units in hidden layers, in the models with one hidden layer there are 8 units, while in models with two layers there are 8 in the first layer and 6 in the second. Compared to the models that consider all the variables, the reduction in the units in the hidden layers of the models is evident, being that each layer has lost one unit.

Regarding the number of observations, it's notorious the increase of the observations when compared to preceding models. These updated models contain over double the observations of previous models, which is a big difference.

Through the examination of the Appendix I, there are strong indications that the variable "Shareholders' Funds" stands out as the most important in all MLP updated models, with a normalized importance of 100%. On the other hand, the variable "Net assets turnover" appears as the least important in all these models, having a normalized importance of less than 10% in all models. However, as it contains many observations (few missing values), its presence in the study is not problematic, given that its presence does not reduce the number of available observations.

5.4. Radial Basis Function

5.4.1. Configurations and options

As made with MLP method, when applying an RBF model, it's necessary to define several running settings. The following configurations are the ones that are used in both RBF and MLP techniques, that were already described in section 5.3.1 of the present research:

- RNG: this option remains with the same value that was defined in MLP (2000000), since this is an AI method, it's mandatory to define this value and to keep the same in all AI methods, in order to guarantee the rigor of the study;
- The training and testing sample are also the same, defined by the same variable, where observations having the value 0 in this variable belong to the test group, and those having the value 1 belong to the training group;
- Observations that have missing values are excluded from this analysis;
- Selected output: Description, Synaptic weights, Model summary, Case processing summary and Independent variable importance analysis.

Regarding the architecture of the RBF model, it’s necessary to define the activation function for the hidden layer. There are two options for this setting in IBM SPSS Statistics 25: Normalized radial basis function and Ordinary radial basis function. As it happened with the MLP method, there is no certainty about which of these is the best option, so two models will be developed, one with each option, to test both scenarios, and find out which one ensures a better accuracy. The remaining settings related with the architecture of the model is the number of units in the hidden layer and the amount of overlap to allow among these units, and both are automatically defined by the software in search of the best value. The overlapping factor is a multiplier applied to the width of the RBF’s.

Model	Number of hidden units	Activation function in hidden layer	Overlap among hidden units
RBF-N	Set automatically	Normalized RBF	Set automatically
RBF-O	Set automatically	Ordinary RBF	Set automatically

Table 17- RBF models’ details.

In table 17 is described the details that define each model, being that the only difference between these two is the activation function in hidden layer.

5.4.2. RBF models

Given that all the customizable settings were defined, the models were created, producing the results described in the following table.

	RBF Models
	RBF-N and RBF-O
Number of units (input layer)	24
Number of units (hidden layers)	10
Training Sample	2 505
Testing Sample	438

Table 18- Summary of the RBF models.

By analyzing the previous table (table 18), and as expected both models have 24 units in the input layer (there are 24 independent variables) and 10 units in the hidden layer.

Once again, as in MLP models, the major limitation of including several independent variables is the reduction of available observations to develop the models due to missing values in the data set. Given this, it would be beneficial to remove some variables from the study to increase the number of available observations. Since, according to appendix J, there are no variables with a normalized importance of less than 10% in both models, the criterion applied to remove “less important variables” aiming to increase the number of available observations, must be consistent from that one used in the MLP method.

In this way, it was necessary to establish a new criterion to remove independent variables. The established criterion was to remove the eight variables with less importance in both models, once in MLP study it were also removed eight independent variables. Therefore, the variables excluded on the RBF updated models were the ones also excluded in MLP updated models, which resulted in the removal of the following variables:

- Probability of default;
- Shareholders Liquidity ratio;
- Gearing ratio;
- Solvency ratio;
- Stock Turnover;
- Collection period;
- Credit period;
- Interest coverage ratio.

The decision to remove eight variables from the study to increase the number of observations was taken considering the number of variables that were excluded in the MLP technique, aiming to provide similar conditions in both AI methods. Curiously, the variables removed from the RBF method are the same as the ones removed from the MLP method, which means that the updated RBF models consider the same variables as the updated MPL models.

	Updated RBF Models	
	RBF-UPD-N	RBF-UPD-O
Number of units (input layer)	16	
Number of units (hidden layers)	9	10
Training Sample	5 369	
Testing Sample	925	

Table 19- Summary of the two RBF models after exclusion of some independent variables.

After the removal of eight independent variables, the number of observations more than doubled, as it happened in updated MLP models as it can be seen in table 19.

Obviously, the number of units in input layer decreased to 16, since 16 independent variables are being considered. Regarding the number of units in the hidden layer, it stands out the fact that the models contain a number of different units in these models.

It should be noted that, according to appendix K, the RBF-UPD-N model considers that a large part of the variables is very important for credit limit calculation, while RBF-UPD-O displays a greater difference in importance between some variables. In other words, in the second model there are evidences that there is a group of variables more important than others, however this is not the case in the first model, where it is not so easy to distinguish between important and non-important variables, once their importance is very similar, separated by small intervals.

6. Comparison between models

In this section it’s performed a comparison, first, between models of the same method, aiming to find out which one is the most accurate, and then compare models developed by different techniques, and thus, analyze and define which method guarantees the best predictive power to calculate the credit limit for enterprises.

It’s recommended to make an analysis of each method first, and only then to make an analysis between different methods, once the larger the number of models included in the comparison, the smaller the comparison sample size. The comparison should preferably be made on the basis of observations which are common to all methods involved in the comparison. On the other hand, this approach may undermine the models that have the most observations, because, models that have few observations may show less variability just because it contains fewer cases, and consequently the model adjustment is better for these observations. However, in a larger sample it can fail completely, given that the model is only well adjusted for a small set of cases.

It should be noted that the sample that is used to compare models was defined in section 4.3.5 of the present research, representing 15% of the total data set, where the remaining 85% are used to develop the models.

6.1. Most accurate MLR model

Regarding MLR models, it’s important to determine the most accurate model through clear evidences, by analyzing the R^2 (being the most important indicator) and the error rate of each model.

Before comparing the models, it was important to adjust the values of each model. Since the model translates a regression, this can have negative values, something that is not possible when the units of the regression are in euros. For this reason, negative credit limit values calculated by the models were transformed into 0, once it would not make sense to consider negative values, being the unit of the values in euros.

Firstly, an analysis is made considering the total number of test cases available for each model. In other words, the models involved in the comparison are compared using the same data sample, however many cases may be available in one model but not in the other due to missing values present in the variables that are included in the model. After this analysis, and only if the model with the fewest test cases available shows evidence of being more accurate, further analysis should be carried out, but the observations considered will be only those contained in both models. This second analysis aims to understand whether the model with the fewest test observations available presents a better accuracy due to its regression adjustment, or whether due to the decrease in observations, which leads to a reduction in variability. In other hand, if in the first analysis the model that has the most observations shows a better accuracy, the second analysis is not performed, since it is proven that even containing more observations (increasing the variability of the sample), the model displays a better accuracy in the credit limit prediction.

In order to calculate the error associated to each observation it was necessary to implement a formula expressing the difference between the real credit limit value and the one calculated by the developed model, in percentage. The error associated to each model may be useful for cases where R^2 of both models are almost the same, and it's hard to choose the best. Being $\widehat{Y_i}$ the credit limit given by the model, and Y_i the real credit limit defined from the data, the relative error associated for each sentence is calculated through the following equation:

$$E_i = \frac{|Y_i - \widehat{Y_i}|}{Y_i} \times 100 \qquad (48)$$

Table 18 describes the distribution of the error (in percentage) associated to each model, the coefficient of determination (which is the most important parameter in the comparison), and the coefficient of variation (CV), which is a statistical ratio of the dispersion of the data around the mean, and it's useful for comparing the degree of variation from one model to another, being calculated by dividing the standard deviation by the mean (Hayes, 2020).

The coefficient of determination for all models was calculated analytically, considering the test set, through de application of equations 3 and 6, and considering the following equation:

$$R^2 = 1 - \frac{SS_E}{SS_T} \qquad (49)$$

Since the test sample is being considered in this section, it's normal that the value of the R^2 will be not the same as the one shown in section 5 of this research for statistical methods, besides de fact that at this stage of comparison it's being analyzed the R^2 and not R^2_{Adj}, and as already explained in section 3.1, they are not the same. In this way, the coefficient of determination is the main parameter of comparison among all models developed in this research, being a robust indicator.

	MLR Models	
Statistics	MLR-1	MLR-2
Sample size	1 019	616
R^2	72	75.8
Mean Error	918%	1 282%
Median Error	127%	225%
Std. Deviation Error	2 885%	2 903%
Coef. of Variation Error	3.14	2.26

Table 20- Statistical data about MLR models' performance.

First aspect that should be mentioned, is the inaccuracy of both MLR models, once both models don't contain an R^2 acceptable to be considered as reliable in credit limit calculation.

The difference between R^2 of each model is not significant, being that MLR-2 model has almost half of the observations than MLR-1 model. In this way, it's important to test MLR-1 model under the same conditions than MLR-2, aiming to understand whether the larger number of MLR-2 observations increases the variability of the sample, which may cause a worst adjustment of the model. By analyzing the mean, median and the standard deviation of the error of both models it becomes very clear that the MLR-1 model has lower values in these factors than MLR-2 model, which indicates a lower associated error. This may suggest that effectively, although R^2 is lower in the MLR-1 model, it may be due to its high number of observations compared to MLR-2.

For this reason, it becomes mandatory to perform a study with MLR-1 considering only the observations that are considered in MLR-2 model, in order to check whether the smallest number of observations is the origin for the lower variability of the sample, increasing the R^2. By considering the same observations in MLR-1 model that were considered in MLR-2, the R^2 has increased to 83.6%, which confirms the theory that MLR-1 is hampered due to its greater number of observations. Since the difference between R^2 values of the two models is large, there is no need to study the distribution of the error, since given the improvement in the R^2 of the MLR-1, the error is also expected to decrease.

Therefore, it can be concluded that MLR-1 model is the most accurate MLR model. Although initially the MLR-1 model had a slightly lower R2 (72%) than MLR-2 (75.8%), this was due to the higher number of observation available for the first model, since when considering the same observations as the MLR-2, the R^2 of the MLR-1 rose significantly (83.6%), being higher that the one of the MLR-2.

It is emphasized again that even the MLR-1 model (which is more accurate than MLR-2) does not show the desired error rates, having high error values, which means that the model is not reliable for credit limit calculation. With such high error rate, it's unacceptable to consider that an MLR fits the regression that calculates the credit limit.

6.2. Most accurate MM model

MM models in section 5.2.2 of the present thesis have shown evidence that can be reliable in the calculation of the credit limit. Compared to MLR models, the MM models produced promising results, which suggests that these models fit better with the regression that calculates the credit limit, however it's necessary to assess the results.

The methodology applied to conduct the comparison of the MM models was the same as the one used to compare MLR models. As is already known, the MM-1 model has fewer observations than the MM-2 model, which may put into question the validity of the study. Given this, in the first phase of the comparison are considered all the available test cases, and if the model with the most cases is the most accurate, the comparison process stops. On the other hand, if the model with fewer cases presents evidence that it's the most accurate, then a new comparison has to be made considering only the cases that both have in common.

The table below indicates the coefficient of determination (the most important indicator), descriptive statistics of the error associated to each model calculated by equation 46 and their respective CV, just in case the R^2 of the models are very similar.

	MM Models	
Statistics	MM-1	MM-2
Sample size	690	1 348
R^2	94.1	90.9
Mean Error	84%	95%
Median Error	88%	96%
Std. Deviation Error	10%	3%
Coef. of Variation Error	0.12	0.03

Table 21- Statistical data about MM models' performance.

Although the error associated with MM models is much lower than MLR models, the distribution of the error remains high enough to consider a reliable model for the calculation of the credit limit.

In this case, the model that shows evidence of being the most accurate is the one that has fewer cases available, the MM-1 model. This argument is derived from the fact that the MM-1 model has an higher R^2, as well, lower mean and median values than MM-2, despite having a higher standard deviation and CV.

Given that in this comparison the model that has shown evidence of being the most accurate was the one with that has fewer cases available it's necessary to perform the second phase of the comparison, to ensure that the model is indeed the most accurate. Therefore, in order to perform the second phase of the comparison it will only be considered those cases which are available for both models. Once the variables present in MM-2 model are present in MM-1, the observations considered in the next analysis will be the same as considered by MM-2 in the first phase of the comparison.

	MM Model
Statistics	MM-2
Sample size	690
R^2	91.2
Mean Error	94%
Median Error	96%
Std. Deviation Error	4.1%
Coef. of Variation Error	0,04

Table 22- Statistical data about MM-2 model's performance considering the same observations as MM-1.

In the table above (table 22) it's only listed the MM-2 model because, the MM-1 model remains equal as described in table 19, once the observations considered in this phase are the same for MM-2. On the other hand, MM-2 model, in this phase considered the same observations as MM-1 in the previous one.

About the analysis of the table, it can be concluded once again that the MM-2 shows evidence of having a worse accuracy than MM-1 model, considering the R^2. The reality is that even considering only the cases available for MM-1 model, MM-2 has not changed significantly in its associated error. The only parameters that present a better result in MM-2 model are the standard deviation and CV, which indicates that in this model the error is more concentrated than in MM-1 model, but its mean is located in an higher error level, while MM-1 model has a greater dispersion for error, but its mean is located in a lower error level.

It can be stated that MM-1 model guarantees a better accuracy to calculate the credit limit than MM-2 model. The factor that also contributed to this decision, behind the fact of the coefficient of determination was higher in MM-1 model, was the median, since having the lowest median indicates that for at least 50% of the observations. It should be noted that there no big difference between these two models, given that both have similar values in parameters, but R^2 being the most solid indicator, the decision was taken on the basis of this factor.

6.3. Most accurate MLP model

After finding in MLR and MM methods which are the most accurate models, in each method, it’s necessary to perform the same analysis to detect which MLP model is the most accurate. As previously developed, while in statistical methods two models have been developed for each, in MLP method were developed eight models. The methodology of the comparison remains the same as in previous methods, which consists in the two phases of the comparison, depending on which method is the most accurate, as already explained.

	Statistics					
MLP models	Sample size	R^2	Mean Error	Median Error	Std. Deviation Error	Coef. of Variation Error
MLP-1HT	438	92.8	592%	100%	2 011%	3.4
MLP-1S	438	92.2	424%	94%	2 026%	4.7
MLP-2HT	438	89.0	687%	100%	2 563%	3.7
MLP-2S	438	95.7	652%	100%	2 448%	3.7
MLP-UPD-1HT	925	84.8	737%	100%	2 415%	3.2
MLP-UPD-1S	925	87.1	674%	100%	1 774%	2.6
MLP-UPD-2HT	925	83.2	748%	100%	2 170%	2.9
MLP-UPD-2S	925	87.1	506%	88%	1 452%	2.9

Table 23- Statistical data about MLP models’ performance.

By analyzing table 23, the selection of the most accurate model is not evident. All the first models being developed show higher values for R^2 than any other updated model. However, once again it has to be analyzed whether these results are due to the fact that these models have fewer observations, having less variability on the sample. So, it’s necessary to assess the updated models with the same observations used in the others models.

	Statistics					
MLP models	Sample size	R^2	Mean Error	Median Error	Std. Deviation Error	Coef. of Variation Error
MLP-UPD-1HT	438	90.5	498%	100%	1 287%	2.6
MLP-UPD-1S	438	94.9	513%	100%	1 278%	2.5
MLP-UPD-2HT	438	90.8	522%	100%	1 381%	2.6
MLP-UPD-2S	438	95.4	390%	95%	1 156%	3.0

Table 24- Statistical data about MLP models’ performance considering the same observations as others MLP models.

As it can be seen in the table above (table 24), all updated models have been greatly improved, when considering the same observations as the ones considered for the others MLP models. Considering these two last tables, there are three models that stand out from the others: MLP-2S, MLP-UPD-1S e MLP-UPD-2S. So, initially it can be concluded that when using the MLP technology the Activation function in hidden layer must be Sigmoid.

These three models show very close R^2 values, so the decision of which is the most accurate MLP model depends on the error distribution associated to each model. The MLP-UPD-2S model shows better values for all parameters excluding the CV, where MLP-UPD-1S has a lower CV, but the difference between both is almost insignificant. So, it can be stated that the MLP-UPD-2S model is the most accurate MLP model, although it has a slightly lower R^2 than MLP-2S model, not being significant, containing much better values for error rate than any other model, even considering more observations, as shown in table 21.

It should be noted that the error rates analyzed in this method are high and there is no model that shows evidence of being reliable in credit limit definition, given than these rates are not acceptable for an accurate model.

6.4. Most accurate RBF model

Finally, it's necessary to compare RBF models. Once again, the adopted methodology was the same as that used in the previous methods, composed by two phases.

	RBF models			
Statistics	RBF-N	RBF-O	RBF-UPD-N	RBF-UPD-O
Sample size	438	438	925	925
R^2	55.8	27.4	13.8	27.6
Mean Error	1 066%	1 954%	2 827%	1 901%
Median Error	100%	100%	114%	118%
Std. Deviation Error	4 211%	6 852%	19 086%	6 777%
Coef. of Variation Error	3.95	3.51	6.75	3.56

Table 25- Description of the distribution of the error associated to each RBF model, in percentage.

As it can be observed in table 25, the choice of the most accurate model in this technique is not very difficult at first sight, since one model stands out clearly from the others. The RBF-N model has all the statistical parameters better than any other RBF model, except for CV, since RBF-O model contains a slightly lower CV, having no impact on this comparison, because the R^2 of the RBF-N is much higher than any other RBF model, although it's still very low.

Given that RBF-N model contains about half of the observations of the models updated, it's mandatory to perform an analysis, in order to understand if these two models are being impaired because these models contain more observations than the other two models, which may increase the variability of the sample. Given the big difference in R^2 between RBF-N and the others, even with less observations, it's expected that no model is expected to improve to the point of having an R^2 higher than RBF-N model.

In this way, is this stage of the analysis is only considered only those observations which have a credit limit associated in RBF-N and RBF-O models, once the difference between these two models and the updated ones is the exclusion of some variables, to increase the number of available observations. Therefore, all the observations available in the first models are also available for the updated models.

	RBF updated models	
Statistics	RBF-UPD-N	RBF-UPD-O
Sample size	438	438
R^2	30.2	29.3
Mean Error	1 305%	1 256%
Median Error	120%	125%
Std. Deviation Error	3 820%	3 381%
Coef. of Variation Error	2.93	2.69

Table 26- Description of the distribution of the error associated to RBF updated models, considering the observations that has in common with RBF-N, in percentage.

With the significant reduction of the observation, the improvement experienced by the RBF updated models is evident, once all values of statistical parameters have improved, except its median, as shown in the previous table (table 26).

Despite this considerable improvement in these updated models, RBF-N model continues to show a better R^2 than any other RBF model. So, RBF-N is the most accurate model, despite the fact that it's still too low in relation to what was expected by an IA model.

6.5. Most accurate model over all methods

In this section the main goal is to define which model is the most accurate over all methods. In other words, a comparison will be made between the most accurate model of each technique, in order to understand which one is the most reliable for credit limit calculation. Once again, it should be highlighted the fact that the error rates of all model are unacceptable for a credible calculation, however, the objective that has been proposed in the present thesis is to define the most accurate model.

The following table shows the statistical parameters of the most accurate model of each method in order to perform a comparison between them.

	Models			
Statistics	MLR-1	MM-1	MLP-UPD-2S	RBF-N
Sample size	1 019	690	925	438
R^2	72	94.1	95.4	55.8
Mean Error	918%	84%	506%	1 066%
Median Error	127%	88%	88%	100%
Std. Deviation Error	2 885%	10%	1 452%	4 211%
Coef. of Variation Error	3.14	0.12	2.87	3.95

Table 27- Description of the distribution of the error associated to the most accurate model of each method, in percentage.

By analyzing table 27, it can be concluded through a direct comparison that RBF-N and MLR-1 models have worse values for all parameters compared to MM-1 and MLP-UPD-2S models, which means that these two models can be out of equation for the most accurate model over all methods. The RBF-N model contains substantially fewer observations than MM-1 and MLP-UPD-1S models, so the latter models are susceptible to greater variability in the sample, and still show evidence of being more accurate than RBF-N model. In the case of the MLR-1 model, despite having more observation than MM1 and MLP-UPD-2S models, its R^2 is much lower than those presented by the last two models. For these reasons, both MLR-1 and RBF-N models are certainly not more accurate in the credit limit calculation than the others two models.

In this way, the models which show better statistical values for model adjustment are MM-1 and MLP-UPD-2S models. On the one hand, The MLP-UPD-2S model has a slightly higher R^2 (about 1.3%) even though it has more observations, being that MM-1 model contains about less 25% observations. On the other hand, by assessing the distribution of the error associated to each model, it's notorious that the MM-1 model has an error rate concentrated in lower levels than MLP-UPD-2S (which has a greater dispersion of the error), having lower values for all parameters, except for median, which is equal in both models.

Since the MLP-UPD-2S model has a higher R^2 than the MM-1, having a higher number of observations, keeping the most accurate model selection criterion verified in the previous considerations, the MLP-UPD-2S was considered the most accurate model. It should also be mentioned that, by analyzing the error rates of these two models it becomes clear that, despite the good results for R^2, these models still don't present conditions for the autonomous and accurate calculation of the credit limit, as desired.

7. Conclusions

This section aims to synthesise the conclusions drawn from the present thesis, pointing out the future work that can be carried out with the purpose of further study. It's also important to list the limitations that have been experienced during the research, given that these factors have a direct impact on the thesis' development.

The first phase of the practical part of the present research was based on the treatment of the data used in the study. This phase is crucial to ensure a credible and fair research, given than the data are the foundation of the study. If data problems are not solved (for example, removal of invalid observations, multicollinearity problems, among others), the credibility of the study may be at stake, so correct processing of data is necessary in order to avoid compromising the quality of the research.

Regarding MLR models, it became very evident that these models don't have power of prediction of the credit limit, and there is no evidence that the MLR method is useful for calculating the credit limit, by observing their coefficient of determination and error rate. Once these models were not validated, given that don't respect all assumptions made when applying a Multiple Linear Regression, it was expected that these models would not be accurate for credit limit calculation. In view of these evidences, it's very unlikely that the regression that defines the credit limit is a linear one.

About AI models it's important to mentioned that the final results were not the expected ones. As AI technologies, it was expected that both techniques would be able to develop models that had the ability to calculate the credit limit accurately. This means that the process of calculating the credit limit, or the model(s) that defines the credit limit is highly complex, being that none of the AI methods have discovered a pattern in the data that would indicate the calculation process. An MLP model (MLP-UPD-2S) was considered the most accurate considering all models developed in this thesis. However, as already mentioned, it cannot be considered as a reliable model in the calculation of the credit limit, once it still presents high error levels. Concerning the RBF method, it's clear that this method is not suitable for the credit limit investigation, having R^2 values extremely low, which means that models developed by this technology don't fit in the credit limit definition.

It should be stressed that removing variables that were shown to be less important in the updated models in AI methods was beneficial (due to missing values), as there was a notable improvement compared to root models. This proved the importance of ensuring as many observations as possible aiming to improve the adjustment of the model.

Lastly, the MM models, although not achieving the desired results, left excellent indications. It can be stated that it's unlikely to be just one model to define the credit limit for all companies, being that companies are defined according to one or more indicators in order to classify the class of the company. This means that even if the credit limit calculation is MM, and the variables are those considered previously, the model will always be adjusted according to the class to which the company belongs, once there is probably no model that accurately defines the credit limit for all companies. The complexity of the credit limit calculation is proven by the lack of accuracy of AI models, which may mean that there are several patterns in the database, where each pattern may indicate one regression for each class of companies. In this way, with a deeper study, by using the MM method it's possible to reach more reliable models (with lower errors) for the definition of the credit limit. Therefore, it’s concluded that although the models developed in this research are not accurate, the method may be the one used for the definition of the credit limit.

In this way, in alignment with the academic literature of this thesis, an MLP model was considered the most accurate model developed in this research, but MM-1 model left good evidences about the credit limit definition, which may be the analytical method to calculate it. So, MM and MLP methods stood out as the most suitable for defining the credit limit, while in the opposite direction appear the MLR and RBF methods, which are not suitable for calculating the credit limit.

7.1. Limitations

During the present dissertation there were several limitations that made it difficult to perform, or had a direct impact on the final results. These limitations have different origins, some are related to the data available for the study, others with particularities of each method, among others. Below are described and explained each limitation experienced in this study:

- First, one of the major limitations found was the presence of many missing values in the data. This occurrence has a great impact on the performance of the study, since it influences more than one aspect of this research. During the development of the models, special attention had to be paid to the trade-off between the variables considered in each model and the respective available observations. Given that there were variables in the database that were more susceptible to having many missing values, many of these variables may have been poorly exploited, as they reduce the available observations which are essential for model adjustment. The impact of this limitation is aggravated in the AI methods, because in these methods all variables should be considered, unlike statistical methods, where it is chosen which variables should enter in the model. The presence of missing values also affects the rigor of the comparison of models, since this should ideally be done by considering the same observations in all models;
- Although the database contains many variables, there is no guarantee that all variables necessary to calculate the credit limit are present in this set. For this reason, there is a possibility that other variables may be required for the definition of the credit limit;
- As is well known, the database contains the values of financial data for the last three years of companies, since the last date they were made available, however, due to lack of knowledge about the effect of these on statistical models, they were not used in this study. This limitation is due to the lack of certainty regarding the correlations between these variables and the independent variable. For example, there is no certainty whether the signals of correlation are maintained compared to the last year, or change and how it changes;
- Most of the observations have credit limits below 500 thousand euros, so observations with higher credit limits have a major impact on the model adjustment. As there is a large dispersion of cases with high credit limits, the model may not adjust properly, and may jeopardize the adjustment in the lower credit limits;
- Regarding the statistical methods, and as previously mentioned, it’s very likely that is needed more than one model to calculate the credit limit of companies, and a regression for each class of

companies may be required, being that each class of companies may be defined according to one or more variables. Besides the fact that there is the possibility that several models may be needed to calculate the credit limit (depending on the class of each company), each model associated with a class may contain different variables. In other words, each regression may have different variables according to the respective class of the company, which means that companies in different classes may have different variables for calculation of their credit limits;

- About the methodology adopted in statistical methods, described in figure 14, a limitation can be also mentioned, once it only considers the entry of one variable at a time, and therefore the impact of certain variables may not even be considered together. There are some variables which alone have no importance, but when combined with other variables, they can become significant in the model. Thus, there is a lack of consideration for variables that only have impact on the model when certain variables are also being included in the model;
- Although it was not an obstacle in the course of the development of MLP models, once the model given as the most accurate was a model with only one hidden layer (which indicates that one hidden layer should be appropriate for credit limit calculation), IBM SPSS Statistics 25 software limits the number of hidden layers in MLP method to a maximum of 2. As is well known, an MLP model has no limits on the number of hidden layers, so it’s a limitation derived from the software used.

As it can be seen, there were a considerable number of limitations in this research, standing out the presence of missing values as the one with the greatest impact on the development of models, influencing several stages of the study.

7.2. Further work

Given that the final results in this research were not the desired ones (which was to develop a reliable model for credit limit calculation), there are complementary studies that can be carried out in order to achieve more accurate and reliable models. Thus, it’s important to indicate the direction in which studies based on this research should be conducted to enrich the academic literature on definition of credit limit for companies.

Since one of the major constraints in the study is the presence of many missing values, additional effort should be made to ensure as many cases as possible (without missing values). The more complete the database, the better the quality of the study, once due to the missing values, some cases were automatically discarded by the predictive methods, which reduces the size of the sample used. To deal with this situation, financial information from companies that are not present in the data set can be searched in other digital platforms.

Regarding the pre-processing of the data, which is a crucial element in the development of this type of studies, although several measures have been taken to ensure the quality of the database, in further studies in can be adopted other methodology to address some limitations experienced in the current thesis. Specifically, the multivariate outlier’s detection method could be improved, since no strict method has been implemented in this study, only the analysis of dispersion charts and boxplots, not being a robust and accurate technique. As demonstrated in section 3.1.6, the calculation of the

Mahalanobis Distance is highly limited to guarantee with precision the presence or not of outliers, so it would be useful to implement a more sophisticated and accurate method.

One limitation that was pointed out in the last section was the limitation of the number of hidden layers in the development of MLP models by the software IBM SPSS Statistics 25. Since this limitation only involves the software used, would be beneficial to develop MLP models in other software that has no technical constraints on this method, given that, as it’s known, IBM SPSS Statistics 25 it’s not the best software to develop MLP models.

Lastly, and as already mentioned many times above, it’s highly likely that there is a model for calculating the credit limit for each class of companies, as opposed to a model which is common to all companies, which was considered in this study. The class of companies would be defined by one or more of their characteristics, such as operation sector, raw financial data, financial ratios, and others. Thus, in a further study it would be advantageous to create several classes of companies, and define a model for each one, which would involve a lot of work and effort. There are two different approaches that can be adopted to define the classes of enterprises: in an arbitrary way, by analyzing the data, without a theoretical foundation or by using AI technology to detect patterns in data. It should be noted that this work only applies to the development of models by statistical methods, since AI techniques should be able to detect the presence of classes without any type of treatment.

8. References

365 DataScience. (2020). *Sum of Squares Total, Sum of Squares Regression and Sum of Squares Error*. Retrieved from 365 DataScience: https://365datascience.com/sum-squares/

A. Marill, K. (2004, January). Multiple Linear Regression. *Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression*, pp. 94-102.

Abu Bakar, N. M., & Mohd Tahir, I. (2009). Applying Multiple Linear Regression and Neural Network to Predict Bank Performance. *Internationa Business Research*, 176-183.

Ahmadian, A. S. (2016). *Numerical Modeling and Simulation*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/engineering/radial-basis-function-network

Akhter, N. (2019, May 20). *Linear Discriminant Analysis in the Financial Market*. Retrieved from Data Driven Investor: https://www.datadriveninvestor.com/2019/05/20/linear-discriminant-analysis-in-the-financial-market/

Alabi, M., Issa, S., & Afolayan, R. (2013). An Application of Artificial Intelligent Neural Network and Discriminant Analyses On Credit Scoring. *Mathematical Theory and Modeling, Vol. 3, Nº11*, 20-29.

Anderson, D. R. (2020, October 20). *Statistics*. Retrieved from Britannica: https://www.britannica.com/science/statistics/Numerical-measures

Bazzi, M., & Hasna, C. (2015). Rating models and its applications: Setting Credit Limits. *Journal of Applied Financial & Banking, Vol. 5, no. 5*, 201-216.

BBC History Magazine. (2019). *The 2008 crisis explained*. Retrieved from History Extra: https://www.historyextra.com/period/modern/financial-crisis-crash-explained-facts-causes/?fbclid=IwAR3YmYB3axWkM7cnwf7BWdw1fjocM6gePghD88j5hz5_6qyxkHHuQAfFeo8

BBC News Business. (2013, April 23). *In graphics: Eurozone crisis*. Retrieved from BBC News Business: https://www.bbc.com/news/business-13366011

Benoit, K. (2011, March 17). Linear Regression Models with Logarithmic Transformations.

Bhalla, D. (2016, December). *Standardized vs Unstandardized Regression Coefficient*. Retrieved from Listen Data: https://www.listendata.com/2015/04/standardized-vs-unstandardized.html

Bremer, M. (2012). *Multiple Linear Regression.* Retrieved from http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf

Brett, D. (2017, August 9). *The global financial crisis 10 years on: six charts that tell the story*. Retrieved from Schroders: https://www.schroders.com/en/insights/economics/the-global-financial-crisis-10-years-on-six-charts-that-tell-the-story/

Brownlee, J. (2017, July 14). *What is the Difference Between Test and Validation Datasets?* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/difference-test-validation-datasets/

Chandradevan, R. (2017, 18 August). *Radial Basis Function Neural Networks- All we need to know*. Retrieved from Towards Data Science: https://towardsdatascience.com/radial-basis-functions-neural-networks-all-we-need-to-know-9a88cc053448

Chauhan, A. (2017, June 15). *What is variable importance and how is it calculated?* Retrieved from DZone- AI Zone: https://dzone.com/articles/variable-importance-and-how-it-is-calculated?fbclid=IwAR01sbVjSOoTtCVCAeVEutuJ_16VVwVHewWxtWpFNGWBm6kygo7YCXdMFtA

Chiang, J.-T. (2007). The Masking and Swamping Effects: Using the Planted Mean-Shift Outliers Models. *Int. J. Contemp. Math. Sciences, Vol.2, No.7*, 297-307.

Cohen, I. (2018). *A Quick Guide to different Types of Outliers*. Retrieved from Anodot: https://www.anodot.com/blog/quick-guide-different-types-outliers/

Corporate Finance Institute. (2019). *Rating Agency: Evaluating the credit-worthiness of debt-issuing companies and organizations*. Retrieved from CFI: https://corporatefinanceinstitute.com/resources/knowledge/finance/rating-agency/

Danenas, P., & Garsva, G. (2018, March 24). *Support Vector Machines and Their Aplication in Credit Risk Evaluation Process*.

Dotai, J. (2015, December 28). *Everything you need to know about Artificial Neural Networks*. Retrieved from Medium: https://medium.com/technology-invention-and-more/everything-you-need-to-know-about-artificial-neural-networks-57fac18245a1

Faris, H., & Mirjalili, S. (2017). *Evolving Radial Basis Function Networks using Moth-Flame Optimizer*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/engineering/radial-basis-function-network

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models, Thrid Edition.* Canada: Sage Publications.

Franzese, M., & Iuliano, A. (2019). *Correlation Analysis*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/medicine-and-dentistry/correlation-analysis

Freeman, K. M. (2018, December 22). The Economics of Trade Credit: Risk and Power.

Frost, J. (2017, September). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Frost, J. (2018, January). *How to interpret P-values and Coefficients in Regression Analysis*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Ghodselahi, A., & Amirmadhi, A. (2011). Application of Artificial Intelligence Techniques for Credit Risk Evatuation. *International Journal of Modeling and Optimization*, 243-249.

Ghorbani, H. (2019). Mahalanobis Distance and its application for detecting multivariate outliers. *Ser. Math. Inform. Vol.34, No 3*, 583-595.

Glen, S. (2016, June 20). *Durbin Watson Test & Test Statistic*. Retrieved from Geometry How To: https://www.statisticshowto.com/durbin-watson-test-coefficient/

Govindaraj, P. (2018, May 26). *Outliers- What it say during data analysis*. Retrieved from Medium: https://medium.com/@praveengovi.analytics/outliers-what-it-say-during-data-analysis-75d664dcce04

Gutpa, T. (2017, January 5). *Deep Learning: Feedforward Network*. Retrieved from Towards Data Science: https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7

Hayes, A. (2020, September 13). *Corporate Finance & Accounting- Financial Analysis (Coefficient of Variation)*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/coefficientofvariation.asp

IBM. (2019). *Training (Multilayer Perceptron)*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/neural_network/idh_idd_mlp_training.html

Jiang, Y., Yang, C., Na, J., Li, G., Li, Y., & Zhong, J. (2017). A brief review of Neural Networks based learning and control and their applications for robots. *Hindawi*.

Kang, N. (2017, June 27). *Multi-Layer Neural Networks with Sigmoid Function- Deep Learning*. Retrieved from Towars Data Science: https://towardsdatascience.com/multi-layer-neural-networks-with-sigmoid-function-deep-learning-for-rookies-2-bf464f09eb7f

Kenton, W. (2019, June 4). *Credit & Debt*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/credit.asp

Kenton, W. (2019, April 14). *Multiple Linear Regression – MLR Definition*. Retrieved from Investopedia: https://www.investopedia.com/terms/m/mlr.asp

Kopf, D. (2015, November 6). *The discovery of statistical Regression*. Retrieved from Priceonomics: https://priceonomics.com/the-discovery-of-statistical-regression/

Kotu, V., & Deshpande, B. (2019). *Deep Learning*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

Kumar Kain, N. (2018, November 21). *Understanding of Multilayer Perceptron (MLP)*. Retrieved from Medium: https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f

Lang, W. W., & Jagtiani, J. (2010, February 9). The Mortgage and Financial Crises: The role of Credit Risk Management and Corporate Governance.

Lewinson, E. (2019, April 16). *Explaining probability plots*. Retrieved from Towards Data Science: https://towardsdatascience.com/explaining-probability-plots-9e5c5d304703

Lou, K.-R., & Wang, W.-C. (2017). Optimal trade credit and order quantity when trade credit impacts on both demand rate and default risk. *Journal of the Operational Research Society*.

Marr, B. (2018, September 24). *What are Artificial Neural Networks- A simple explanation for absolutely anyone*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/#1836f4e11245

McCaffrey, J. (2015, May 13). *Neural Network Train-Validate-Test Stopping*. Retrieved from Visual Studio Magazine: https://visualstudiomagazine.com/articles/2015/05/01/train-validate-test-stopping.aspx

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics, Vol. 13, No.3*, 253-267.

Militký, J. (2011). *Fundamentals of soft models in textiles*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/chemical-engineering/radial-basis-function-networks

Minitab. (2019). *Methods and formulas for stepwise in Fit Regression Model*. Retrieved from Minitab 18 Support: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/methods-and-formulas/stepwise/

NCSS Statistical Software. (2016). *Stepwise Regression.* Retrieved from NCSS Statistical Software: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

Nicholson, C. (2017). *A beginner's guide to Multilayer perceptrons (MLP)*. Retrieved from Pathmind: https://pathmind.com/wiki/multilayer-perceptron

Osborne, J. (2002). Pratical Assessment, Research, and Evaluation: Vol.8, Article 6. *Notes on the use of data transformations*. Retrieved from https://scholarworks.umass.edu/pare/vol8/iss1/6

Pardoe, I. (2019). *Normal Probability Plot of Residuals*. Retrieved from PennState Eberly College of Science: https://online.stat.psu.edu/stat501/lesson/4/4.6

Parient, R. (2017, May 4). *What are rating agencies?* Retrieved from BBVA: https://www.bbva.com/en/what-are-rating-agencies/

Paul E., A., Dan Dan, E., & I. Sidney, O. (2015). Mathematical Theory and Modeling. *A Review of the Limitations of Some Discriminant Analysis: Procedures in Multi-Group Classification*, pp. 199-203.

Pope, P., & Webster, J. (1972). The Use of an F-statistic in Stepwise Regression Porcedures. *Technometrics, Vol.14, No2*, 327-340.

Renaud, O., & Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 1852-1862.

Riani, M., & Zani, S. (1995). An iterative method for the detection of multivariate outliers. *Giomate di analisi dei dati multidimensionali* (pp. 101-116). Naples: Ministry of University (MURST).

Riley, G. (2017, March 15). *Consequences of a Financial Crises.* Retrieved from SlideShare: https://pt.slideshare.net/tutor2u/consequences-of-financial-crises

Rogers, K., & Boyd Enders, F. (2013, December). *Collinearity*. Retrieved from Britannica: https://www.britannica.com/topic/collinearity-statistics

Rouse, M. (2019, September). *Supervised Learning*. Retrieved from SearchEnterprisedAI: https://searchenterpriseai.techtarget.com/definition/supervised-learning

Sadeghkhani, I., Ketabi, A., & Feuillet, R. (2012, May 1). Radial Basis Function Neural Network Application to Power System Restoration Studies. *Computacional Intelligence and Neuroscience*, pp. 1-10. Retrieved from Hindawi.

Saunders, A., & Allen, L. (2010). *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms (Second Edition).* John Wiley & Sons, Inc.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression Analysis. *Deutsches Arzteblatt International*, 776-782.

Schwartz, S. (2017, October 8). *Moody's Completes Acquisition of Bureau van Dijk*. Retrieved from Moody's: https://ir.moodys.com/news-and-financials/press-releases/press-release-details/2017/Moodys-Completes-Acquisition-of-Bureau-van-Dijk/default.aspx

Shah, T. (2017, December 6). *About Train, Validation and Test Sets in Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Sharma, A. (2017, March 30). *Understanding Activation Function in Neural Networks*. Retrieved from Medium: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0

Shilpi, G. (2014). Credit Risk Modelling: A wheel of Risk Management. *International Journal of Research (IJR)*.

Siegel, A. F. (2016). *Individual Regression Coefficient.* Retrieved from ScienceDirect: https://www.sciencedirect.com/topics/mathematics/individual-regression-coefficient

Singh Chauhan, N. (2019, October 3). *Introduction to Artificial Neural Networks (ANN)*. Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9

Soman, D., & Cheema, A. (2002). Marketing Science. *The Effect of Credit on Spending Decisions: The Role of the Credit Limit and Credibility*, pp. 32-53.

Stata. (2020, August 18). *Checking Homoscedasticity of Residuals*. Retrieved from Libraries| Research Guides- The University of Utah: https://campusguides.lib.utah.edu/c.php?g=160853&p=1054158

Statistics Solution. (2019). *Assumptions of Multiple Linear Regression*. Retrieved from Statistics Solution: https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

Stephanie. (2015, September 22). *Multicollinearity: Definition, Causes, Examples*. Retrieved from Statistics How to: https://www.statisticshowto.com/multicollinearity/

Stephanie. (2015, September 24). *Stepwise Regression*. Retrieved from Statistics How to: https://www.statisticshowto.datasciencecentral.com/stepwise-regression/

Su, H. L. (2017). Comparing tests of homoscedasticity in Simple Linear Regression. *JSM Mathematics and Statistics 4 (1)*, 1017.

Teekens, R., & Koerts, J. (1972). Some statistical implications of the log transformation of multiplicative models. *Econometrica Vol. 40, No. 5*, 793-819.

Vairava Subramanian, G., & Nehru, S. (2012). Implementation of credit rating for SMEs. *International Journal of Scientific and Research Publications*, 205-211.

Van Zandt, T. (2012, August). Firms, Prices and Markets. *Elasticity of Demand*.

Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & M. Tu, X. (2014). Log-transformation and its implications for data analysis. *Shangai Archives of Psychiatry*, 105-109.

Witten, I. H., & Pal, C. J. (2017). *Extending instance-based and linear models*. Retrieved from Science DIrect: https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

Woodruff, J. (2019, January 25). *The Advantages & Disadvantages of offering credit*. Retrieved from Chron: https://smallbusiness.chron.com/advantages-disadvantages-offering-credit-30773.html

Zach. (2020, March 26). *How to Perform a Breusch-Pagan Test in Excel*. Retrieved from Statology: https://www.statology.org/breusch-pagan-test-excel/

Zanders. (2016, January 19). *Public credit ratings: Navigating on faith*. Retrieved from Zanders: https://zanders.eu/en/latest-insights/public-credit-ratings-navigating-on-faith/

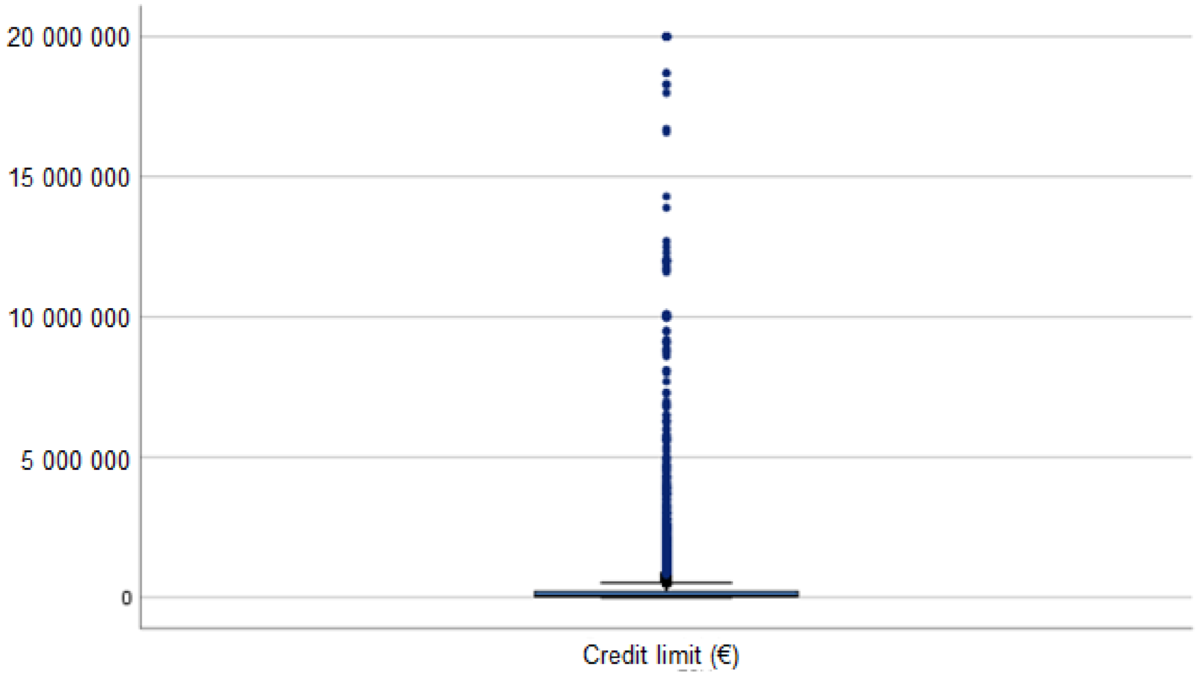
Appendices

Appendix A- Individual Description of each variable

Variables	Description
Risk Class	Defines in which risk class each company is inserted.
Credit Limit	Maximum amount of credit that can be granted to the company.
Probability of default	Measures the probability of default taking into account financial records.
Fixed Assets	Is a long-term tangible piece of property or equipment that a company owns and uses in its operations to generate income.
Current Assets	Are company's cash and its other assets that are expected to be converted to cash within one year.
Shareholders Funds	Refers to the amount of equity in a company which belongs to the shareholders.
Non-Current Liabilities	Are long-term financial obligations.
Current Liabilities	Are short-term debt with maturity dates within one year.
Operating Revenue (Turnover)	Indicates the revenue generated from the company's primary business activities.
EBITDA margin	Indicates the company's earnings before interest, taxes, depreciation and amortization (EBITDA) as a percentage of the Revenue, and it's obtained by dividing the EBITDA by the Revenue.
Operating P/L [=EBIT]	Is the profit that a company generates from its core business.
P/L after tax	Indicates the profit or loss in a given period time after tax.
Depreciation & Amortization	Represents the loss of assets' value over the time.
Added value	Describes which a company gives its products or services before offering it to costumers.
Cash Flow/Operating Revenue	Measures the Cash Flow as a percentage of the Total Revenue, and it's calculated by dividing the Cash Flow by the Revenue.
ROE using P/L before tax	Return on Equity (ROE) ratio indicates the financial performance of the company, by dividing the profit or loss before tax by the equity.
ROCE using P/L before tax	Return on Capital Employed (ROCE) measures the efficiency of the company to generate profits from the capital employed, by dividing the profit or loss before tax by the subtraction of the total assets by the current liabilities.
ROA using P/L before tax	Return on Assets (ROA) measures the efficiency of the company to generate returns with its assets, by dividing the profit or loss before tax by the assets.
ROE using Net Income	Variant of ROE using Net Income, and it's calculated by dividing Net Income by the equity.
ROCE using Net Income	Variant of ROCE using Net Income, and it's calculated by dividing the Net Income by the subtraction of the total assets by the current liabilities.

ROA using Net Income	Variant of ROA using Net Income, and it's calculated by dividing the Net Income by the assets.
Profit margin	Measures the profitability of the company, by dividing the Net Profit by the Revenue.
EBIT margin	Indicates the company's earnings before interest and taxes (EBIT) as a percentage of the Revenue, and it's obtained by dividing the EBIT by the Revenue.
Net Assets Turnover	Indicates the company's ability to manage its assets to generate sales, and it's obtained by dividing the Net Sales Revenue by the Total Sales.
Interest Coverage ratio	Is a debt and profitability ratio used to determine how easily a company can pay interest on its outstanding debt.
Stock Turnover	Is a ratio that shows how many times a company has sold and replaced stock during a given period, and it’s obtained by dividing sales by the stock.
Collection period days	Average time taken to receive payments owed.
Credit period days	Time period in which the debtor must meet all the financial obligations previously agreed with the creditor.
Current ratio	Indicates the company's ability to meet its obligations, and it's calculated by dividing the Current Assets by the Current Liabilities.
Liquidity ratio	It's an acid-test ratio that measures the ability of the company to meet short and long-term obligations, and it's obtained by the subtraction of the Current Assets by the Inventory over the Current Liabilities.
Shareholders Liquidity ratio	Indicates how much of the company’s assets are funded by equity shares.
Solvency ratio	Measures the company's ability to meet its debt obligations.
Gearing ratio	Measures the company's leverage, which demonstrates the degree of activities are funded by shareholders funds versus creditor's funds.
Long-term Debt	Is debt that maturates in more than one year.
Loans	Is money, property, or other material goods given to another party in exchange for future payment of the loan value.
Other Non-current Liabilities	Are liabilities that don’t maturate within one year.
EBITDA	Earnings before interest, taxes, depreciation, and amortization.
Stock	Refers to the foundation of many individual investors' portfolios, or private sales as well.
Total Assets	Refers to the total amount of assets owned by the company.
Cash Flow	Net amount of cash and cash-equivalents that is transferred in or out of each company's accounts.
Status	Describe the company’s status.
Main business sector	Refers in which sector the company mainly operates.

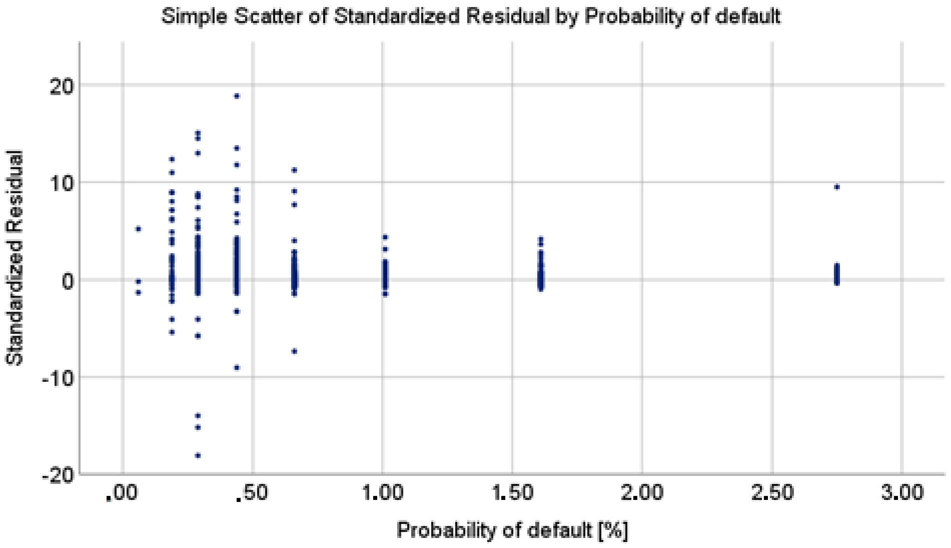
Appendix B- Boxplot of the credit limit

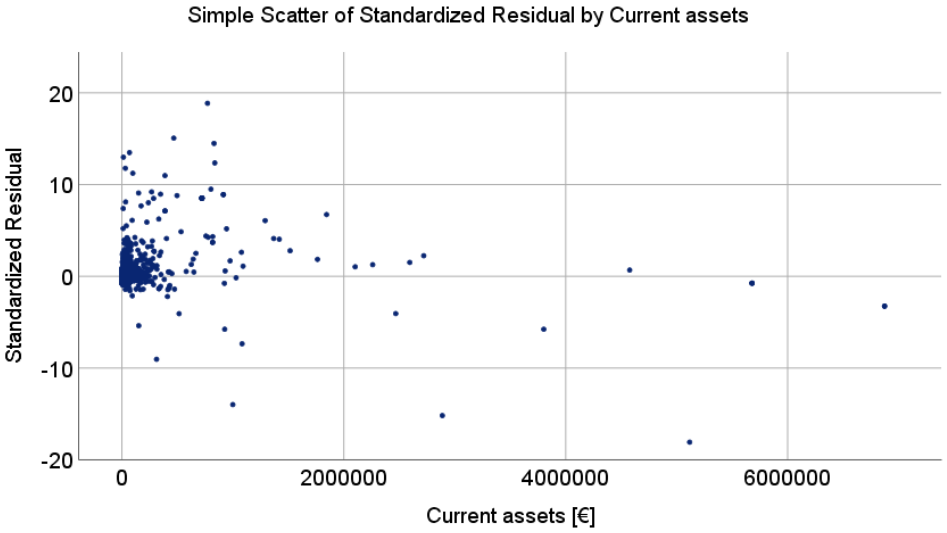


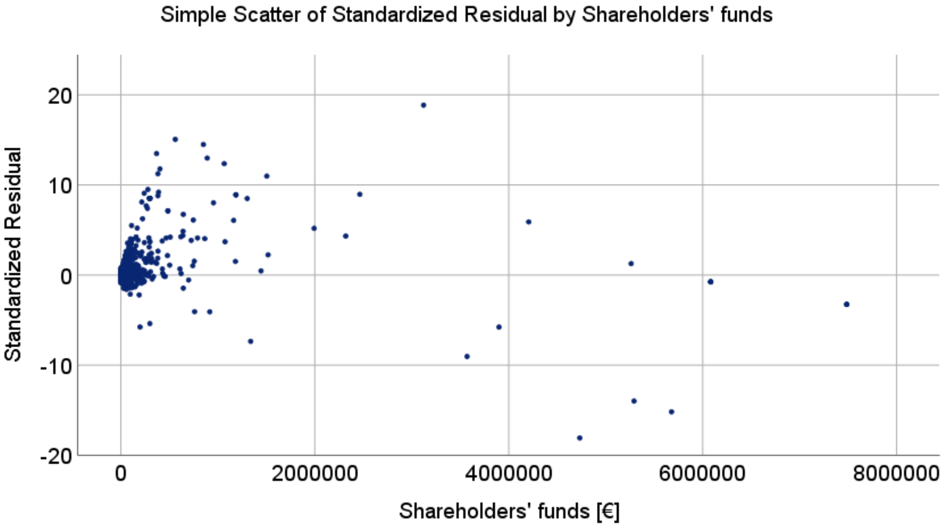
Appendix C- Variables' descriptive statistics

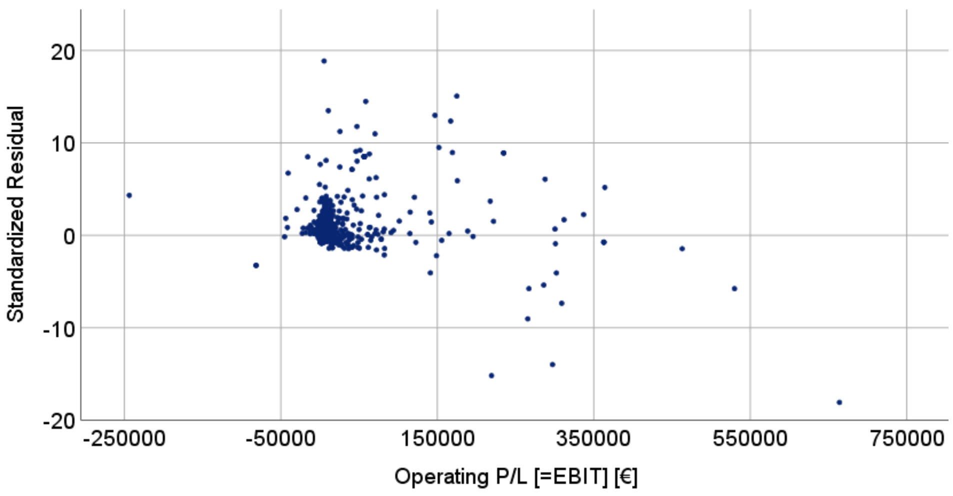
Variables	N	Minimum	Maximum	Mean	Std. Deviation
Credit limit	9 237	1 000	20 000 000	354 495	1 379 851
Probability of default	9 046	0.06	2.75	1.26	0.7
Current Assets	9 237	0	6 872 897	21 198	195 631
Shareholders' funds	9 237	-20 084	20 100 155	24 403	306 698
Operating Revenue	9 194	-2 527	22 595 593	46 799	513 234
EBITDA margin	9 101	-98	99	11.5	15.6
Operating P/L [=EBIT]	9 227	-888 950	663 943	2 549	23 414
P/L after tax	9 225	-1 037 097	1 010 464	2 301	28 068
ROE using Net income	9 220	-521	743	13.2	31.6
ROCE using Net income	7 681	-494	452	10.5	19.8
ROA using Net income	9 217	-65	99	5.1	9.1
Profit margin	9 067	-98	99	6.4	13.6
Net assets Turnover	9 185	0	996	4.4	16.9
Interest coverage	7 022	-97	998	52	132
Stock Turnover	6 959	0	994	53	116
Collection period	9 118	0	953	80	89
Credit period	9 118	0	967	47	59
Current ratio	9 177	0,001	98	3.3	6.6
Liquidity ratio	9 178	0	98	2.7	6.3
Shareholders liquidity ratio	7 044	0	999	25	83
Solvency ratio	4 946	0.1	99	47	25
Gearing ratio	9 121	0	996	74	117
Long term debt	7 163	-0.4	10 014 872	16 005	217 101
Other non-current Liabilities	7 163	-1	13 356 402	8 614	275 728
Loans	9 204	0	2 988 068	4 479	74 212

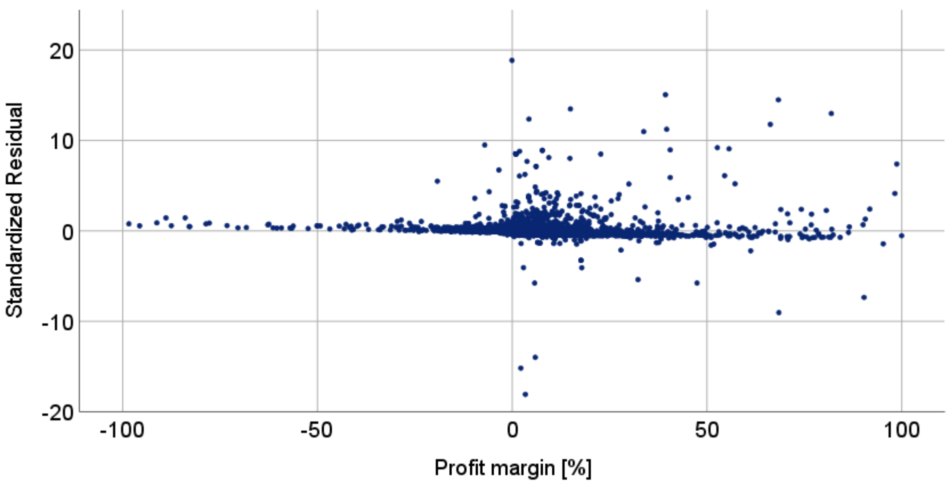
Appendix D- Dispersion charts of residuals over independent variables of MLR-1

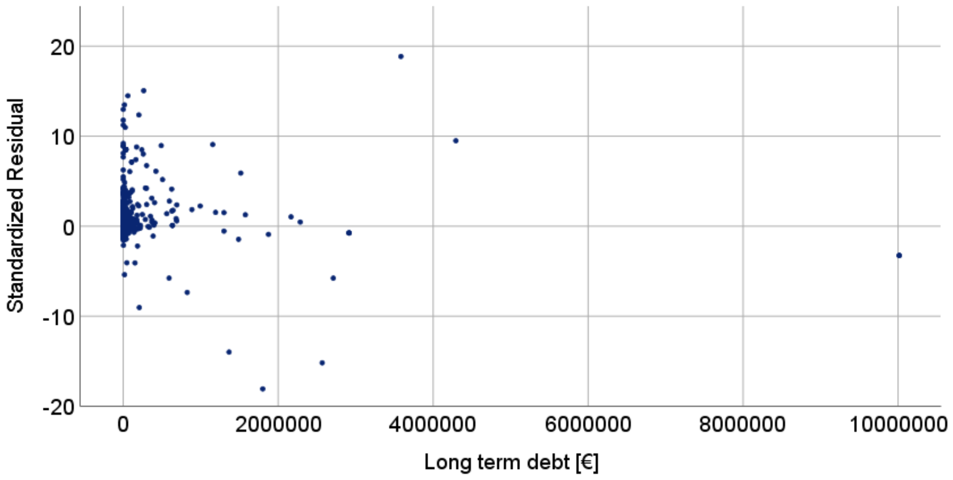


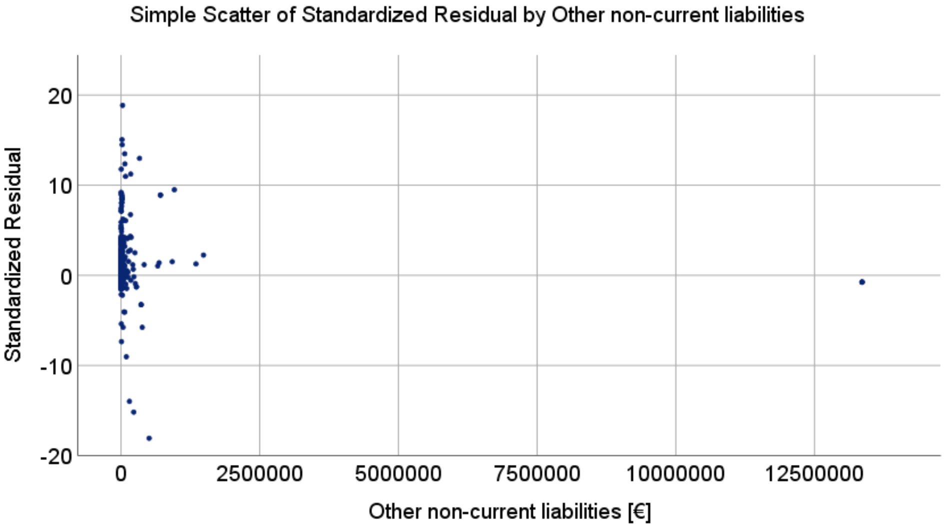




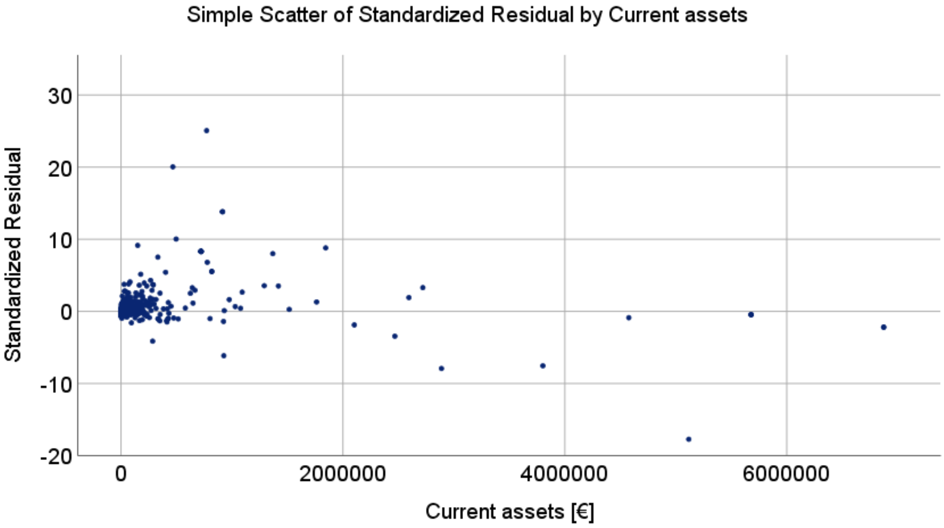

Simple Scatter of Standardized Residual by Operating P/L [=EBIT]

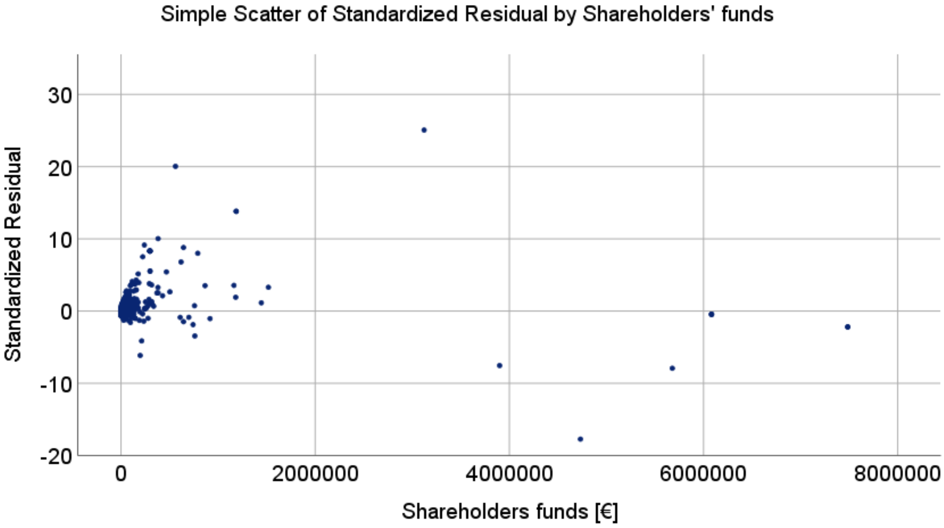

Simple Scatter of Standardized Residual by Profit margin

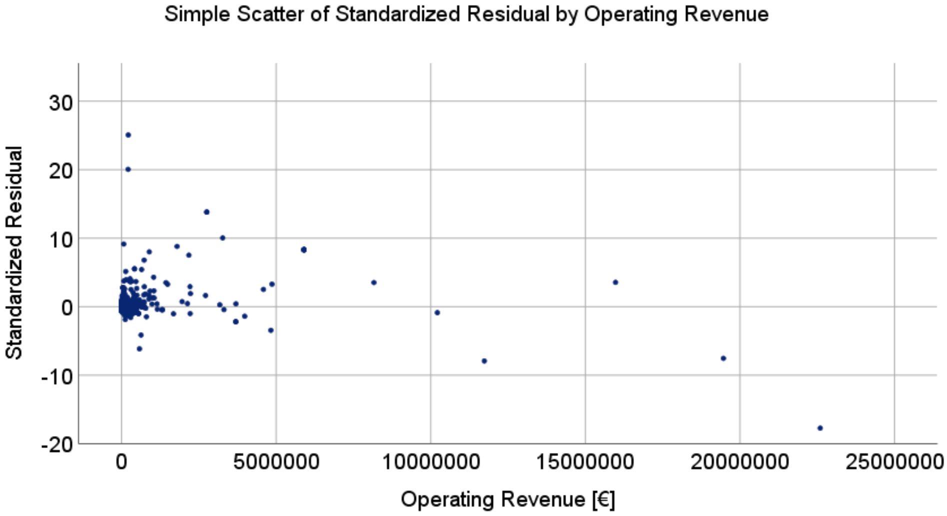

Simple Scatter of Standardized Residual by Long term debt

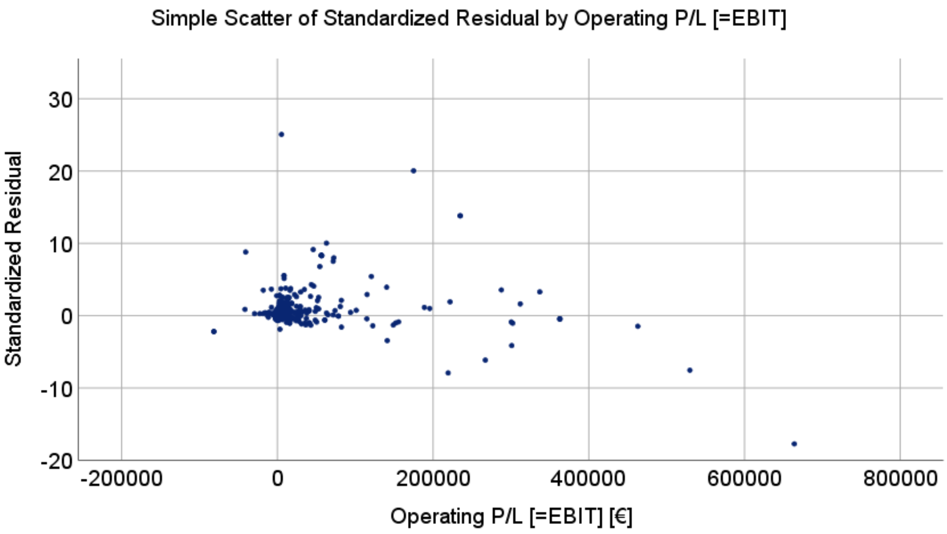

Simple Scatter of Standardized Residual by Other non-current liabilities
Standardized Residual
20
10
0
-10
-20
0
2500000
5000000
7500000
10000000
12500000
Other non-current liabilities [€]

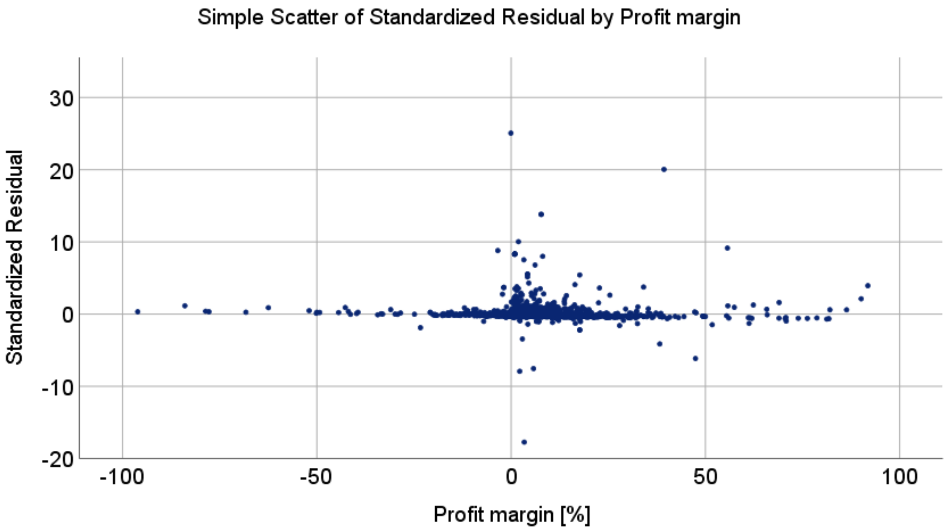
Appendix E- Dispersion charts of residuals over independent variables of MLR-2

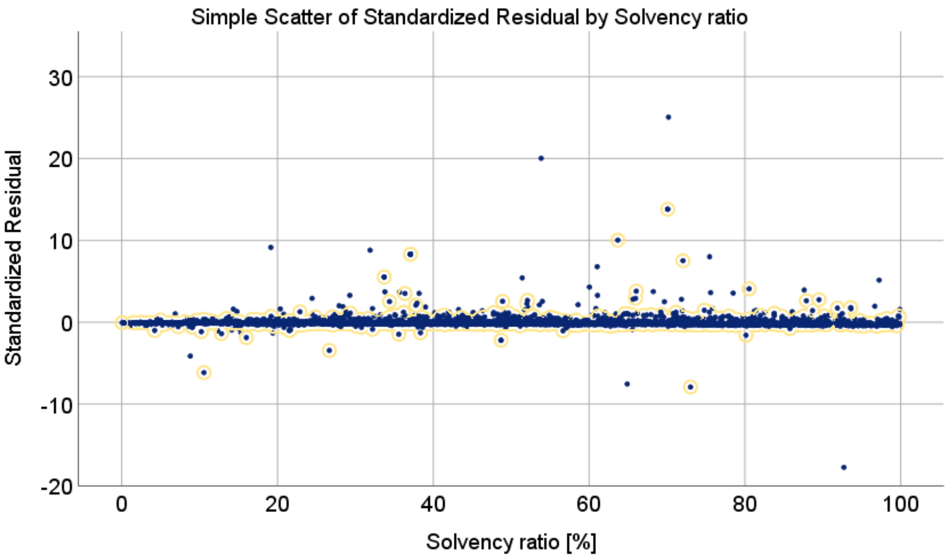


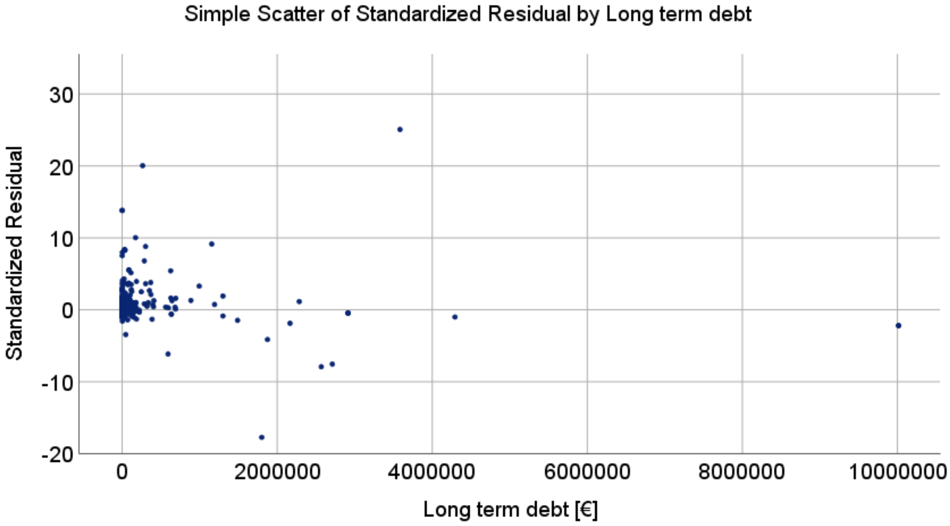




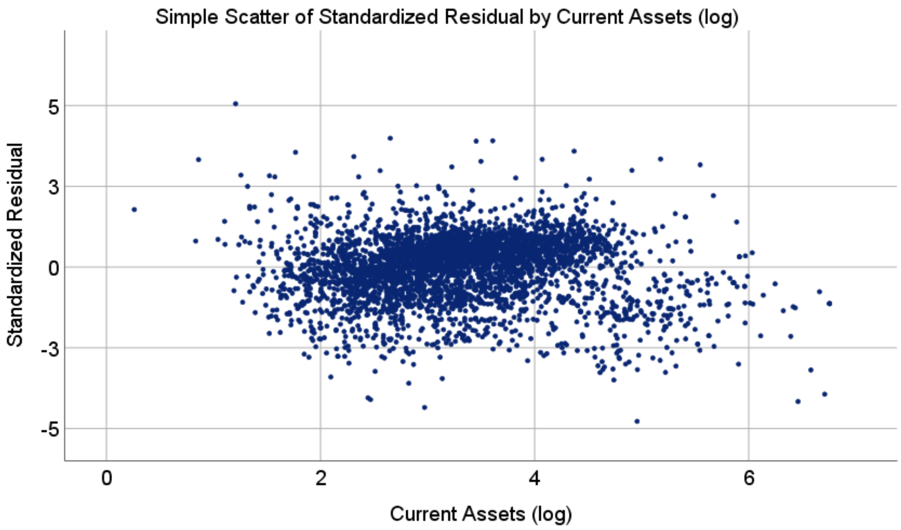

Simple Scatter of Standardized Residual by Operating P/L [=EBIT]
Standardized Residual
30
20
10
0
-10
-20
-200000
0
200000
400000
600000
800000
Operating P/L [=EBIT] [€]

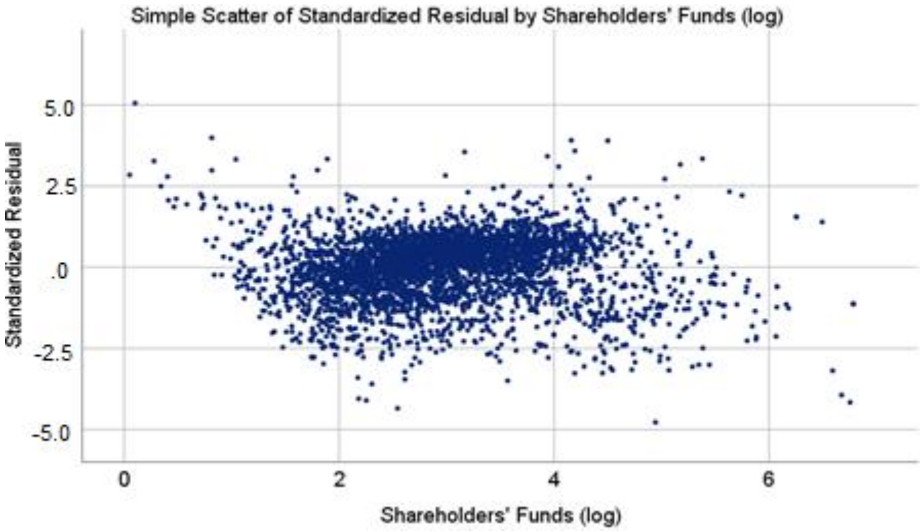

Simple Scatter of Standardized Residual by Profit margin
Standardized Residual
30
20
10
0
-10
-20
-100
-50
0
50
100
Profit margin [%]

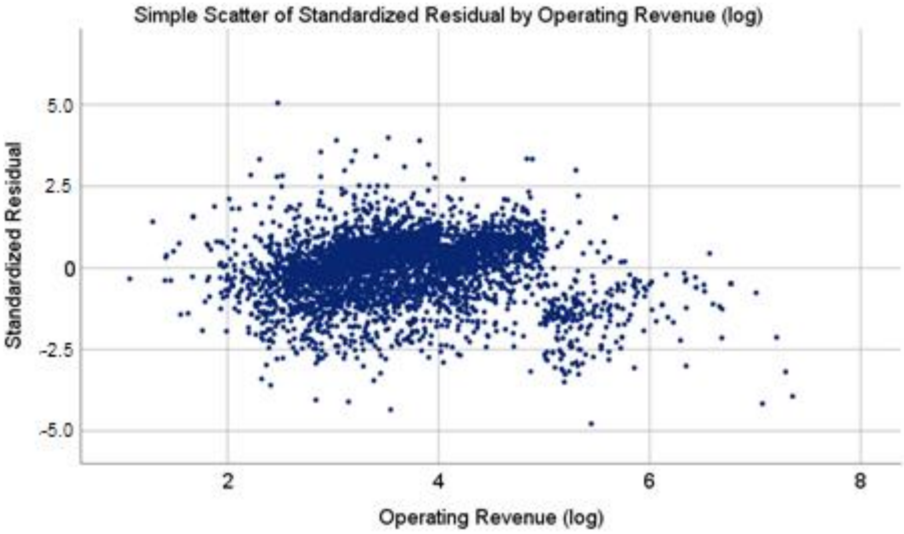

Simple Scatter of Standardized Residual by Solvency ratio
Standardized Residual
30
20
10
0
-10
-20
0
20
40
60
80
100
Solvency ratio [%]

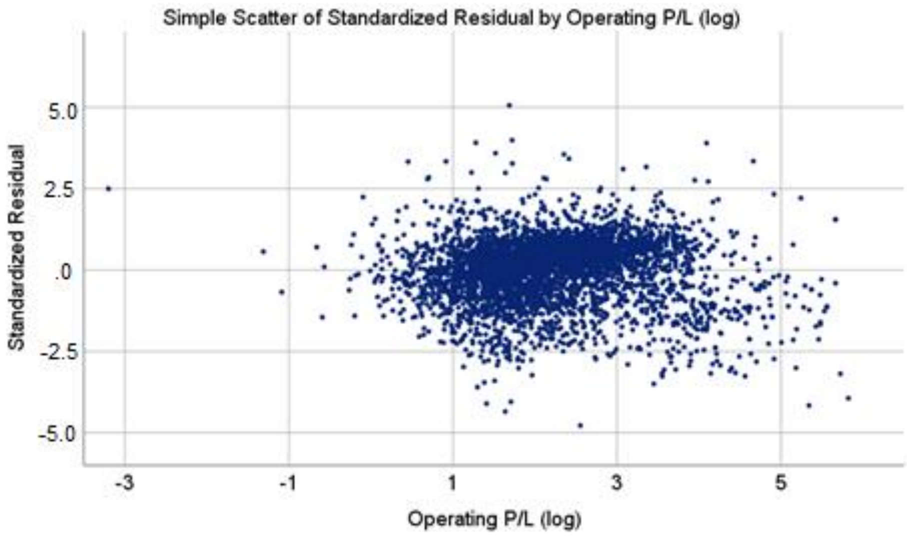

Simple Scatter of Standardized Residual by Long term debt
Standardized Residual
30
20
10
0
-10
-20
0
2000000
4000000
6000000
8000000
10000000
Long term debt [€]

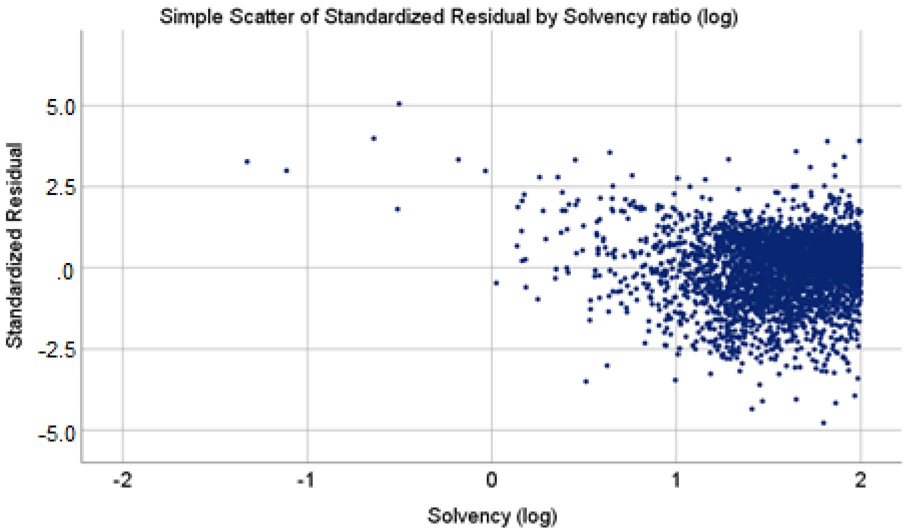
Appendix F- Dispersion charts of residuals over independent variables of MM-1



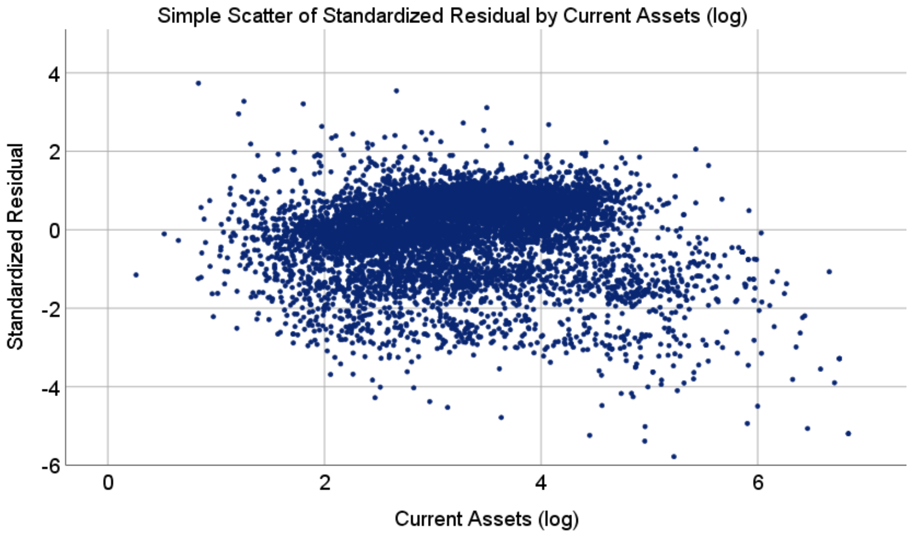


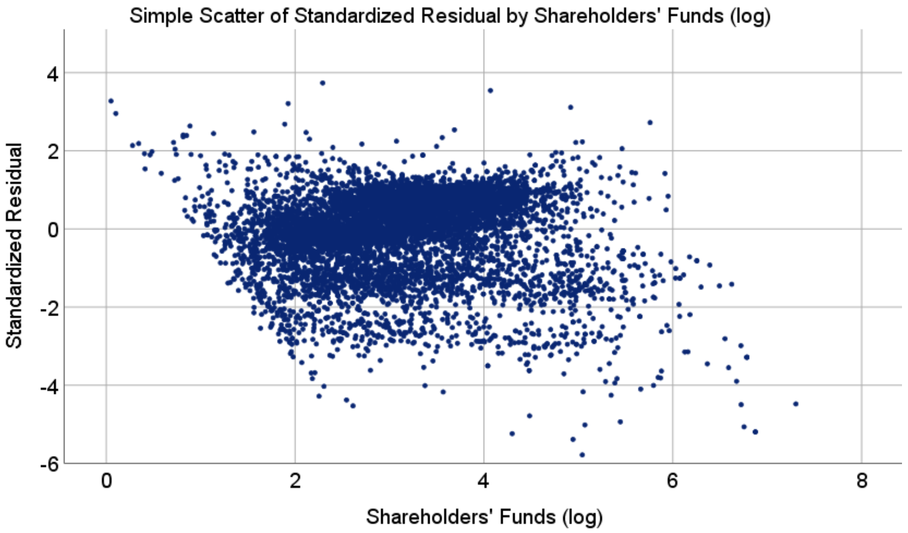


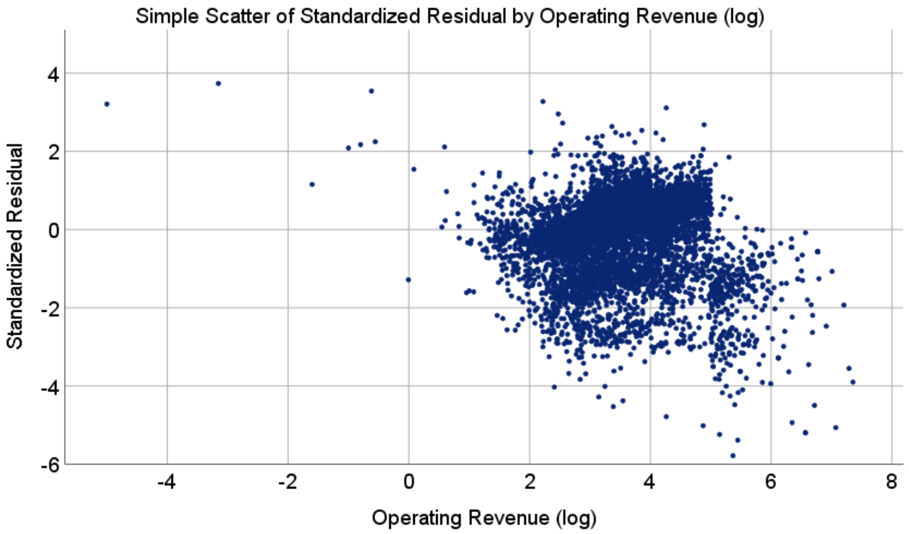

Simple Scatter of Standardized Residual by Operating P/L (log)
Standardized Residual
5.0
2.5
.0
-2.5
-5.0
-3
-1
1
3
5
Operating P/L (log)


Simple Scatter of Standardized Residual by Solvency ratio (log)
Standardized Residual
5.0
2.5
.0
-2.5
-5.0
-2
-1
0
1
2
Solvency (log)

Appendix G- Dispersion charts of residuals over independent variables of MM-2







Appendix H- Independent variables’ normalized importance in MLP models

Variable	MLP-1HT	MLP-1S	MLP-2HT	MLP-2S
Current Assets	100%	65.0%	46.2%	100%
Shareholders’ Funds	97.5%	100.0%	80.9%	82.2%
Operating Revenue	51.0%	86.1%	31.6%	48.8%
Operating P/L [EBIT]	53.8%	22.5%	76.6%	84.0%
Profit or Loss after tax	9.7%	11.9%	27.9%	24.5%
Long-term Debt	27.9%	21.2%	45.2%	48.5%
Loans	11.3%	6.1%	19.8%	31.5%
Other Non-current Liabilities	22.1%	15.0%	31.7%	51.9%
Probability of Default	2.5%	3.0%	5.6%	6.2%
Liquidity ratio	26.4%	4.8%	51.9%	16.9%
Shareholders Liquidity ratio	6.6%	3.6%	8.6%	5.2%
Current ratio	15.0%	5.3%	8.5%	4.9%
Gearing ratio	8.3%	6.5%	5.8%	3.7%
Solvency ratio	4.1%	2.7%	3.3%	2.4%
Net assets turnover	8.7%	3.7%	12.8%	6.0%
Stock Turnover	4.0%	2.2%	4.5%	4.7%
Collection period	6.0%	4.0%	6.3%	3.7%
Credit period	7.8%	4.6%	5.0%	8.2%
Interest coverage ratio	5.4%	5.2%	4.9%	5.6%
Profit margin	9.7%	11.9%	27.9%	24.5%
EBITDA margin	20.7%	12.3%	100.0%	18.1%
ROE using Net Income	29.2%	22.3%	13.1%	31.6%
ROCE using Net Income	16.6%	11.0%	19.0%	14.7%
ROA using Net Income	23.8%	12.0%	35.7%	24.1%

Appendix I- Independent variables' normalized importance in MLP updated models

Variable	MLP-UPD-1HT	MLP-UPD-1S	MLP-UPD-2HT	MLP-UPD-2S
Current Assets	44.1%	40.3%	75.6%	59.6%
Shareholders' Funds	100.0%	100.0%	100.0%	100.0%
Operating Revenue	36.9%	45.6%	66.9%	32.9%
Operating P/L [EBIT]	71.0%	50.1%	25.9%	34.7%
Profit or Loss after tax	15.2%	31.5%	27.9%	29.9%
Long-term Debt	15.2%	12.3%	30.3%	7.1%
Loans	30.2%	8.1%	18.4%	9.9%
Other Non-current Liabilities	12.2%	40.3%	6.7%	12.7%
Liquidity ratio	9.9%	20.8%	18.6%	20.5%
Current ratio	12.7%	11.6%	11.6%	4.3%
Net assets turnover	4.9%	3.5%	5.0%	3.4%
Profit margin	10.3%	10.1%	7.7%	7.0%
EBITDA margin	6.2%	6.8%	32.3%	6.6%
ROE using Net Income	5.2%	12.3%	26.3%	16.0%
ROCE using Net Income	5.9%	4.8%	20.9%	6.2%
ROA using Net Income	7.1%	16.2%	15.5%	11.5%

Appendix J- Independent variables’ normalized importance in RBF models

Variable	RBF-N	RBF-O
Current Assets	100%	100%
Shareholders’ Funds	78.4%	85.1%
Operating Revenue	89.7%	94.4%
Operating P/L [EBIT]	98.2%	96.6%
Profit or Loss after tax	81.5%	91.3%
Long-term Debt	77.0%	83.9%
Loans	69.3%	72.2%
Other Non-current Liabilities	71.9%	78.9%
Probability of Default	6.0%	20.8%
Liquidity ratio	67.7%	56.6%
Shareholders Liquidity ratio	23.1%	40.5%
Current ratio	69.7%	85.4%
Gearing ratio	15.1%	35.1%
Solvency ratio	2.7%	13.6%
Net assets turnover	53.5%	30.5%
Stock Turnover	18.6%	35.4%
Collection period	28.4%	40.8%
Credit period	25.8%	35.3%
Interest coverage ratio	21.8%	36.1%
Profit margin	33.7%	45.4%
EBITDA margin	26.0%	53.1%
ROE using Net Income	86.9%	38.5%
ROCE using Net Income	57.9%	37.2%
ROA using Net Income	35.1%	40.9%

Appendix K- Independent variables’ normalized importance in updated RBF models

Variable	RBF-UPD-N	RBF-UPD-O
Current Assets	90.5%	93.7%
Shareholders’ Funds	90.4%	87.6%
Operating Revenue	90.4%	100%
Operating P/L [EBIT]	94.1%	72%
Profit or Loss after tax	63.8%	81.9%
Long-term Debt	90.2%	94.2%
Loans	90.3%	86.9%
Other Non-current Liabilities	90.3%	97.5%
Liquidity ratio	89.7%	49.3%
Current ratio	89.6%	42.5%
Net assets turnover	100%	46.8%
Profit margin	88.6%	15.9%
EBITDA margin	65.8%	16.5%
ROE using Net Income	90.8%	58.2%
ROCE using Net Income	84.8%	22.8%
ROA using Net Income	84.9%	16.9%