Evaluating Classical and Artificial Intelligence Methods for Credit Risk Analysis

An experimental comparison using a database for B2B clients

Bruno Miguel Querido dos Reis

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 Comittee: Prof. João Carlos da Cruz Lourenço

November 2019

Abstract

Credit scoring remains one of the most important subjects in financial risk management. Although the methods in this field have grown in sophistication, further improvements are necessary. These could translate into major gains for financial institutions and other companies that extend credit by diminishing the potential for losses in this process. This research seeks to compare statistical and artificial intelligence predictors in a credit risk analysis setting. In order to perform this comparison, a credit scoring experiment is conducted with a sample of companies that corresponds to the business-to-business clients of Galp in 2016.

This dataset contains a variety of financial information and other relevant data regarding these companies, which allows for the development of several distinct credit scoring models. Pre-processing procedures are established, namely in the form of a proper sampling technique to assure the balance of the sample. Additionally, multicollinearity in the dataset is assessed via an analysis of the variance inflation factors and the presence of outliers is addressed with a detection technique based on robust Mahalanobis distances. This phase of the research allows for non-robust models to perform better, namely the statistical models that would be particularly affected by these issues.

Several alternative architectures and/or settings are examined for each category of predictors considered. Following these experimentations, the best performing models are selected to be included in the benchmarking study. The results obtained reveal that the best predictive performance was obtained by artificial intelligence methods, confirming previous findings in the academic literature.

Keywords: *credit risk, artificial intelligence, discriminant analysis, logistic regression, artificial neural networks, random forest.*

Resumo

O risco na concessão de crédito continua a ser um tópico de suprema importância na área da gestão do risco em finanças. Apesar dos métodos utilizados nesta área se terem tornado gradualmente mais sofisticados, existe ainda algum espaço para melhorias. Estes avanços podem-se traduzir em enormes ganhos para instituições financeiras e outras organizações que concedam crédito através da redução do potencial para perdas neste processo. Este trabalho de pesquisa procura comparar métodos estatísticos e de inteligência artificial num contexto de análise de risco de crédito. Com este intuito, é realizada uma experiência de *credit scoring* com uma amostra de empresas que correspondem aos clientes *business-to-business* da Galp para o ano de 2016.

A amostra obtida contém uma variedade de informação financeira e de outros tipos a respeito destas empresas, possibilitando o desenvolvimento e implementação de diversos modelos. Os procedimentos de pré-processamento dos dados são estabelecidos, nomeadamente na forma de uma técnica de amostragem adequada para a obtenção de uma amostra equilibrada. Adicionalmente, problemas de multicolinearidade são estudados e a presença de valores discrepantes é abordada. Esta fase da pesquisa permite que modelos não-robustos a estas questões tenham desempenhos superiores, nomeadamente os métodos estatísticos que são tendencialmente mais afetados.

Diversos modelos alternativos são examinados para cada um dos métodos de *credit scoring* considerados. Após esta fase de experimentação, os melhores modelos são selecionados para integrarem o estudo comparativo. Esta análise revela que o melhor desempenho é obtido pelos métodos de inteligência artificial, confirmando os resultados de outros estudos comparativos.

Palavras-chave: *risco de crédito, inteligência artificial, análise discriminante, regressão logística, redes neuronais artificiais, árvores de decisão.*

Acknowledgements

I would like to thank my parents for their unconditional support throughout my academic life that culminated in this dissertation. I leave also a note of appreciation to my brother for his advice in this project and to CJ for being such an overwhelming positive force in my life.

Regarding the risk and credit management team at Galp that took me in these last few months, I would like to thank my colleagues for their support and for making me feel welcome during this experience.

Finally, a special thank you is due to Professor António Quintino for his part in this project. As my supervisor, he demonstrated continuous commitment to this research, being available around the clock and contributing with the necessary guidance to reach meaningful results.

Table of Contents

List of Figures

List of Tables

Glossary of Acronyms

AI – Artificial intelligence

AID – Automatic interaction detector

ANN – Artificial neural network

ANNs – Artificial neural networks

AUC – Area under the curve

B2B – Business-to-business

Bagging – Bootstrap aggregating

BdP – *Banco de Portugal*

BP – Backpropagation

BvD – Bureau van Dijk

CART – Classification and regression tree

CARTs – Classification and regression trees

CHAID – Chi-square automatic interaction detector

CHAIDs – Chi-square automatic interaction detectors

DT – Decision tree

DTs – Decision trees

EBIT – Earnings before interest and tax

EBITDA – Earnings before interest, tax, depreciation and amortization

FN – False negatives

FP – False positives

FY – Fiscal year

GM – Geometric median

GMs – Geometric medians

LDA – Linear discriminant analysis

LR – Logistic Regression

MCMC – Markov chain Monte Carlo

MD – Mahalanobis distance

MDs – Mahalanobis distances

ML – Maximum likelihood

NA – Not available

NS – Not significant

KPIs – Key Performance Indicators

PCC – Percentage Correctly Classified

PER – *Processo Especial de Revitalização*

RBF – Radial Basis Function

ROA – Return on assets

ROC – Receiver Operating Characteristic

ROCE – Return on capital employed

ROE – Return on equity

SPR – Special Revitalization Process

TN – True negatives

TP – True positives

VIF – Variance inflation factor

1. Introduction

1.1. Contextualization of the Problem

Companies acquire funds not only from specialized financial intermediaries but also from the respective suppliers (Fabbri & Menichini, 2010). This practice is denominated trade credit and occurs frequently in the business-to-business (B2B) market when buyers receive credit from suppliers in the form of merchandise and/or services. If credit is approved by a seller for a certain client, there is always the possibility that this client will not honor the agreement to repay the amount in question. On the other hand, if credit is denied, it is possible that a potentially profitable client was handed over to rival companies. Therefore, one must carefully weigh these two factors when deciding on how to proceed regarding credit decisions, since a poor evaluation can cause significant losses (Gouvêa & Bacconi, 2007).

Credit risk, in general, is a topic of the utmost importance in financial risk management, being a major source of concern for financial and banking institutions (Khashman, 2010). However, the strategies in place to manage credit risk have seldom been able to predict when non-compliance will occur with the desired efficacy, which leads to a rise in toxic credits (Batista, 2012). As companies face the possibility of going out of business if insufficient measures are taken to manage this risk, the history of developing models to ascertain the ability of debtors to repay the respective credits is extensive.

In the last decades, quantitative methods to manage credit risk have grown in sophistication. The end-goal is to separate good credit applicants from bad ones. The criterion used in this classification is the ability of the applicants to repay the full amount of the loan plus the interest. Usually, this is achieved by feeding a predictive model with past customer data, thus finding the relationships between the clients' characteristics and the potential for default (Huang, Liu, & Ren, 2018). There is substantial research material on this topic, as only a small improvement in prediction accuracy may result in large gains in profitability (Kvamme, Sellereite, Aas, & Sjursen, 2018). Until recently, to build these credit scoring models, the sole solution was to employ statistical models. The linear discriminant analysis and logistic regression are among the statistical techniques widely used for this purpose (Baesens, Setiono, Mues, & Vanthienen, 2003).

However, technological advances have allowed for enhanced computational capabilities, paving the way for new and more efficient techniques. Such is the case of artificial intelligence (AI) methods. There are numerous studies showing that machine learning tools like artificial neural networks (ANNs), decision trees (DTs) and support vector machines, present an opportunity to improve on the prediction accuracy of statistical models with regards to credit risk (Vellido, Lisboa, & Vaughan, 1999; Huang et al., 2004; Ong, Huang, & Tzeng, 2005).

Despite significant developments in terms of newer classifiers, the literature on credit risk has not kept pace with the breakthroughs in predictive learning (Lessmann, Baesens, Seow, & Thomas, 2015; Jones, Johnstone, & Wilson, 2015). Indeed, more recent techniques such as random forests and generalized boosting have been explored by a limited number of studies, although some sources report them as superior to previous methods (Jones et al., 2015). It is therefore imperative to further study these new techniques to understand how these compare to older and more established methods of credit scoring with respect to performance and applicability.

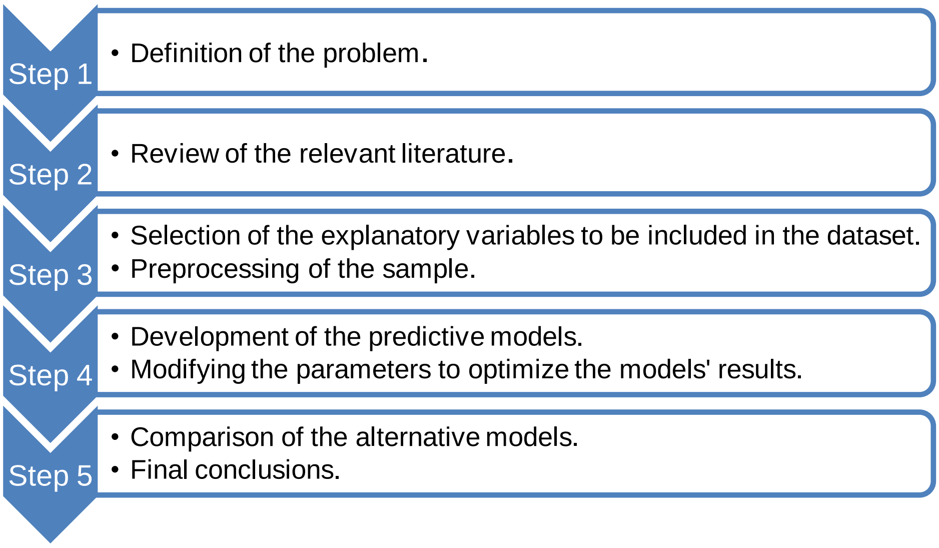
1.2. Thesis Goals and Scope

The objective of this thesis is to complement the academic literature on credit risk analysis by comparing traditional methods for credit scoring with artificial intelligence alternatives. This research should help determine which techniques offer a superior prediction performance. It is known that the current literature is poor in terms of studies comparing different classifiers (Tsai & Wu, 2008), which only emphasizes the importance of the research to be conducted. As there are numerous statistical and AI methods used for the purposes of credit scoring, this dissertation focuses only on these specific techniques: discriminant analysis, logistic regression, artificial neural networks and random forests. The detailed explanation regarding why these methods were selected is presented in section 2.2., along with a brief overview of each of these models.

In order to be able to benchmark these methods, this dissertation includes a credit scoring experiment. This practical component comprises the implementation of the different predictive models in the credit risk analysis problem faced by Galp when dealing with trade credit extended to its corporate clients. The dataset used as input in these predictive models corresponds to a selection of financial and non-financial indicators for the B2B clients of Galp in 2016. After feeding the models with this information, these make predictions on what businesses present a default risk in the following year of 2017. By comparing these predictions with the known outcomes of the companies, it is possible to proceed with the computation of several performance metrics to assess the correctness of each model. After these results have been obtained, conclusions are drawn regarding the suitability of statistical and AI approaches.

1.3. Research Methodology

In order to solve the problem defined in section 2, there must be a methodological approach in place that assures the scientific rigor and completeness of the dissertation. Figure 1 lists the steps to be followed to reach an appropriate conclusion for the problem.



Step 1
- Definition of the problem.

Step 2
- Review of the relevant literature.

Step 3
- Selection of the explanatory variables to be included in the dataset.
- Preprocessing of the sample.

Step 4
- Development of the predictive models.
- Modifying the parameters to optimize the models' results.

Step 5
- Comparison of the alternative models.
- Final conclusions.

Figure 1 - List of the steps to perform the research.

Step 1 regards the definition of the problem tackled in this thesis. In this stage, the crucial importance of credit risk analysis is stressed, while also providing an overview of the methods used. Additionally, the shortcomings of the most popular techniques in this area are detailed, along with the consequences of poor credit risk management.

Step 2 includes a review of the relevant academic literature. This stage allows for an understanding of the state-of-the-art practices, serving as the theoretical basis for the research to be conducted subsequently. This review comprises an extensive analysis of the models' structures and determines which performance measures ought to be used when evaluating the results.

Step 3 encompasses the selection of the indicators to be included in the sample for the credit scoring experiment. Additionally, this stage addresses the preprocessing of the data, which ensures that the dataset is apt to be used as input in the statistical and artificial intelligence methods. This procedure includes an examination into potential multicollinearity problems, an analysis concerning the presence of outlier instances and the development of a conversion procedure to ensure the coding of categorical attributes as numerical ones.

Step 4 includes the construction of various alternative models for each category of predictors considered in the scope of the research. By testing various settings and/or architectures, it is possible to discover which are the most advantageous in the problem at hand. The models are evaluated in terms of the respective predictive performances, which allows for the most suitable model of each category to be included in the final benchmarking.

The objective of step 5 is to benchmark the quality of the predictions obtained with the discriminant analysis, logistic regression, artificial neural network (ANN) and random forest (RF) methods. In order to do this, the correctness of these techniques is assessed through various key performance indicators. There is also a review of the relevant issues regarding the development of these models. Finally, the consequences of this research to the credit risk analysis problem are stated, while also delineating any further work that may be done.

1.4. Thesis Structure

In order to define the structure of this dissertation, Table 1 displays the chapter list of this document, along with a brief description of the contents of these sections.

Table 1 - Structure and contents of the thesis.

Section	Contents
1. Introduction	Definition and proper contextualization of the problem tackled in the thesis. Listing of the research goals. Definition of the project's scope and the research methodology pursued.
2. Problem Definition	Characterization of the credit risk in the Portuguese market. Brief introduction of the models to be studied and their strengths and weaknesses. Description of the experimental setup to evaluate these models.
3. Theoretical Framework	Review of the relevant academic literature concerning the statistical and AI methods to be tested. Analysis of the structure, assumptions and implementation of these techniques. Listing of the main performance indicators used in the evaluation of credit scoring models.
4. Input Data Collection, Analysis and Treatment	Description of the method used to obtain the input data. Explanation of the logic behind the types of explanatory variables used and how these may be computed. Application of the proper pre-processing procedures regarding the sampling technique, multicollinearity issues and the presence of outliers in the data.
5. Model Development	Characterization of the procedures employed in the selection of the independent variables in each model. Definition of the alternative architectures and/or settings evaluated for each method. Interpretation of the relevant parameters relating to the models. Presentation of the relevant performance indicators and selection of the best alternative for each category of predictor contemplated in the research.
6. Comparing AI and statistical methods	Comparison of the final statistical and AI models in terms of the development process. Benchmarking of the various methods considering a selection of key performance indicators.
7. Conclusions and Further Work	Definition of further work that may be done based on this research. Presentation of the conclusions reached.

2. Problem Definition

2.1. Trade Credit Risk in Portugal

As this research concerns trade credit, understanding the level of financial solidity of the Portuguese companies is relevant to the ensuing analyses. When a corporation assesses the possibility to concede credit in a B2B deal, it is fundamental to evaluate the past financial performance of the applicant, as this is indicative of the probability of future non-compliance with payment. One should take into consideration that the risk of default may be higher in trade credit than in other forms of financing. In fact, more robust corporations tend to use relatively less trade credit in comparison with financially debilitated companies (Hill, Kelly, Preve, & Sarria-Allende, 2017). Potentially due to this factor, Jacobson and von Schedvin (2015) reported that the losses incurred by trade creditors are significantly higher than those of banks.

Extensive academic research has examined the importance of financial statement ratios in the prediction of credit failure (Altman, 1968; Ohlson, 1980; Henry, Robinson, & van Greuning, 2011). The companies' operating performance is a major determinant of the respective credit risk (Demerjian, 2007). Analyzing the annual statistics published by *Banco de Portugal* (BdP), it is notorious that a significant fraction of Portuguese companies are in financial distress and hence at a higher risk of default.

Figure 2 contains a chart with the percentage of companies that fit in each category of financial distress. It should be noted that there is significant overlap between these categories, as is expected because these are not independent events. A company displaying one poor financial indicator has a greater probability to have other poor financial indicators.

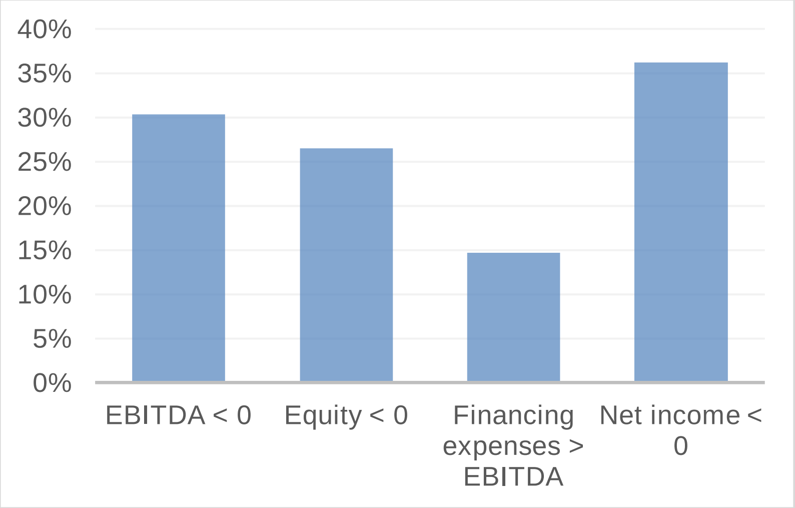


Figure 2 - Portuguese companies displaying poor financial indicators in 2017 (Source: BdP).

Analyzing the chart, it becomes clear that the fraction of companies with negative results is significant. Over 35% of all companies are not profitable, exhibiting negative net incomes. For almost 15% of all companies, the earnings before interest, tax, depreciation and amortization (EBITDA) are not enough to cover the financing expenses. These cases represent a risk in trade credit scenarios, as there is a high probability that these companies will not be capable to provide payment for any services and/or products provided.

In order to further examine the potential for default of Portuguese companies, the records provided by the Ministry of Justice for new insolvency proceedings and revitalization processes were analyzed. These revitalization procedures have been introduced in several EU countries to attempt to save financially distressed, but viable businesses (Eidenmuller, 2018). The Portuguese legislation established the special revitalization process (SPR) for this purpose. This concept will prove especially important in the latter credit scoring experiment.

Figure 3 displays the progression of the numbers of new insolvency proceedings and revitalization processes from 2014 to 2017, along with the total count of companies registered in Portugal for each of those years.

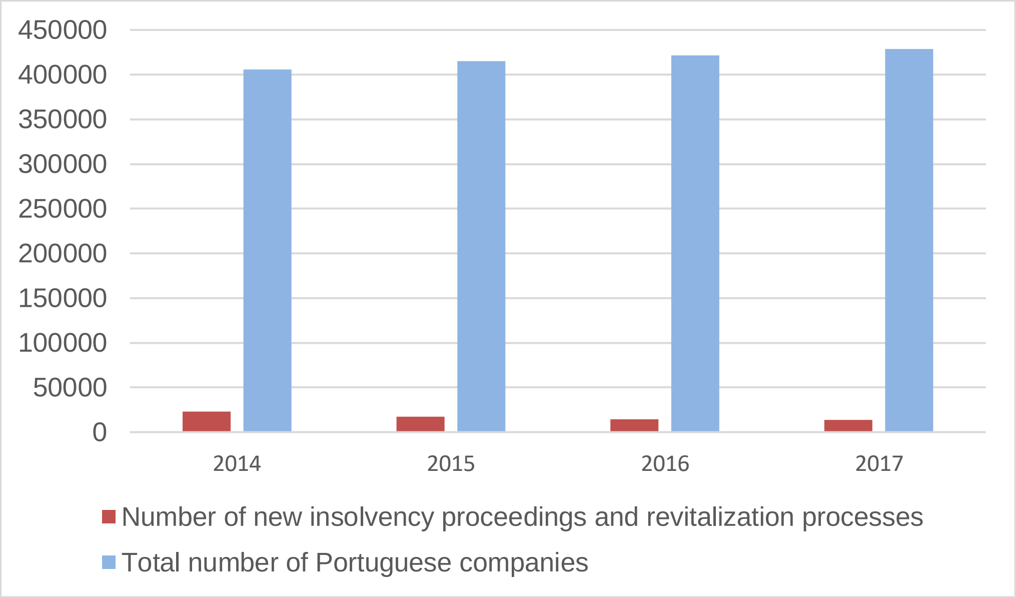


Figure 3 - Evolution of the number of insolvencies and revitalization processes in comparison with the total number of Portuguese companies (Sources: BdP and the Portuguese ministry of Justice).

Analyzing the information in Figure 3, one can conclude that only a small percentage of companies declare bankruptcy or enter revitalization processes each year. On average, these cases amount to just 4.2% of the total number of Portuguese companies in the period considered. This value contrasts with the high percentage of businesses with poor financials that was previously discussed, but these findings are not necessarily contradictory.

Taking into consideration the data utilized in this research, it was notorious that only a limited number of the companies displaying negative results end up applying for revitalization processes or declaring bankruptcy. A business may operate at a loss for several years and default on payments without any formal request for insolvency.

Although the information presented so far reflects the reality of corporations in Portugal, there is the need to compare the characteristics of these companies with the ones of foreign equivalents. This will provide a point of reference, which is needed to assess the relative performance of Portuguese businesses. In order to make this comparison, the temporal evolution (2013-2016) of two relevant financial ratios, the shareholder equity ratio and return on equity (ROE), was analyzed. These indicators are particularly relevant, as both are found to be highly correlated with the company outcomes in the later stages of this research.

The plot in Figure 4 displays the trajectory of the shareholder equity ratio for companies in Portugal, France, Italy, Germany and Austria. This ratio is computed by dividing the equity by the total assets of a company. It is a measure of financial autonomy and can be interpreted as how much capital the shareholders would receive if the corporation shut its doors. Observing the values for the Portuguese companies, these stand out as the worst performing in this selection, although there is a trend of convergence with most of the other countries. This means that Portuguese companies are highly leveraged in comparison with these European counterparts

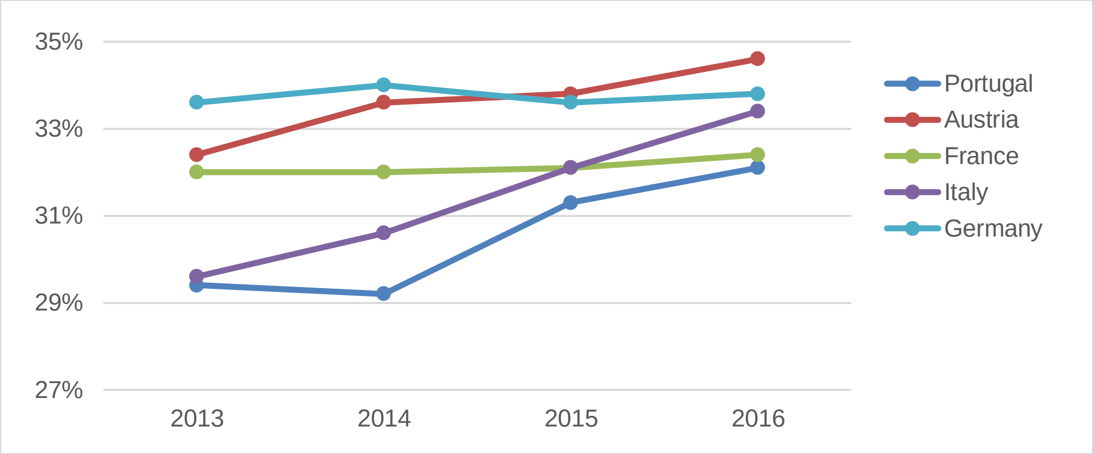


Figure 4 - Shareholder equity ratio in companies from different European countries (Source: BdP).

Figure 5 shows the evolution of a different indicator, the return on equity, taking into consideration the same selection of EU countries. This variable belongs to the profitability measures category of financial indicators. It essentially assesses how capable the companies are to generate returns according to the equity available. Analyzing the results of the Portuguese companies, it is clear again that these are among the worst performing. Hence, after observing both charts, one may say that Portuguese companies appear to be not only the most indebted ones but also among the worst prepared to meet the respective financial obligations.

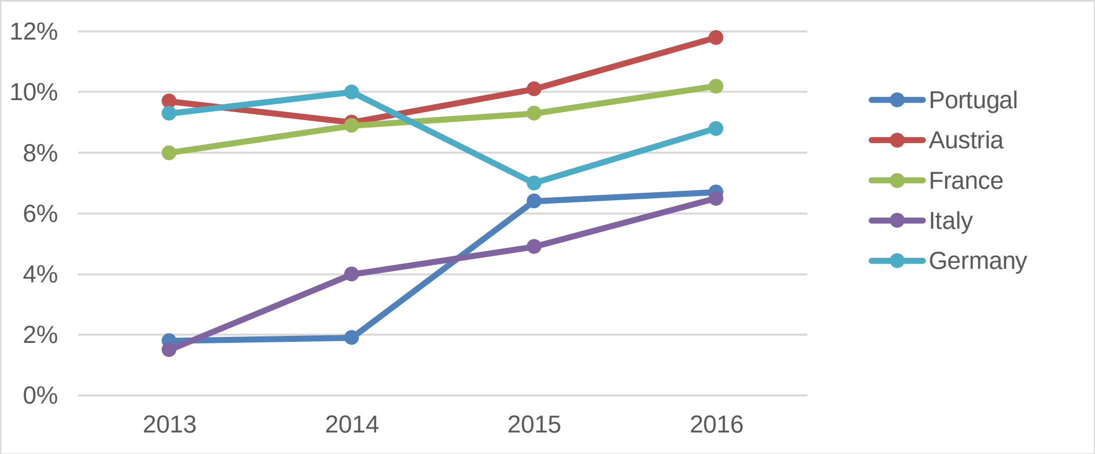


Figure 5 – Return on equity of companies from different European countries (Source: BdP).

What this means in terms of credit risk analysis is that any creditor is expected to be especially careful when pondering credit applications from Portuguese businesses. This is also a motivating factor to research new models or improve the current ones, in order to facilitate the concession of credit to robust companies and reduce the number of corporations not able to meet the respective financial obligations.

2.2. Methods for Credit Risk Analysis

The linear discriminant analysis (LDA) model is among the first statistical techniques utilized for credit scoring (West, 2000). LDA had the advantage over previous techniques, such as ratio analysis, that it could consider multiple characteristics of the credit applicants, as well as the interactions between them (Altman, 1968). However, there are some limitations regarding its validity. It is dependent on stringent assumptions, namely that all variables must present a normal distribution and be mutually independent (Huang, Chen, Hsu, Chen & Wu, 2004; Sustersic, Mramor, & Zupan 2009). These conditions have been proven difficult to meet when dealing with real-world scenarios.

Considering the limitations of the discriminant analysis, researchers started experimenting with logistic regression (LR) models in credit risk problems (Altman & Sabato, 2008). This technique offered some improvements over the LDA, namely that its output is equal to the probability of a given instance belonging to a certain category and that the results could be easily interpreted. Albeit these theoretical differences, studies demonstrate that the empirical results are similar in terms of classification accuracy (Lo, 1985; Altman & Sabato, 2008).

These two techniques, discriminant analysis and logistic regression, are amongst the most commonly applied linear statistical tools in credit scoring (Pacelli & Azzollini, 2011). However, the significant differences between these methods may impact their respective applicability and the quality of the results achieved. Due to the dissimilarities between LDA and LR and the relative popularity of both methods, these techniques are selected to be studied and tested in this thesis, representing the statistical methods for credit scoring.

Nevertheless, it should be stressed that these tools assume linear relationships between the models' outputs and the corresponding explanatory variables. In many situations, that is not the case and the performance of such techniques may be severely hampered.

Other methods have since been developed to deal with complex non-linear relationships. Most prominently, artificial neural networks, a machine learning technique that is now well-established as a credit scoring method. The potential of this technique is confirmed by comparative studies either showing this tool outperforming discriminant analysis (Khemakhem & Boujelbène, 2015; Wójcicka-Wójtowicz & Piasecki, 2017) or suggesting the use of a hybrid model as the best alternative (Lee, Chiu, Lu, & Chen, 2002; Lai, Yu, Wang, & Zhou, 2006). Additionally, previous research demonstrates that ANNs handle particularly well datasets with noise and incorrect entries (Tollo, 2006; Angelini, di Tollo, Roli, 2008; Wójcicka-Wójtowicz & Piasecki, 2017).

Artificial neural networks may be currently the most used individual classifier in credit scoring (Lessmann et al., 2015; Louzada, Ara, & Fernandes, 2016). This means that other methods involving hybrid or combined approaches may be more frequent, but ANNs are the most common technique in terms of stand-alone models. Therefore, this model has been deemed as a critical tool to be analyzed in this thesis.

Despite the great promise of ANNs, there are some disadvantages that should be noted. The black-box nature of neural networks, which basically means that it is very difficult to interpret how the results are achieved (Abdou & Pointon, 2011), is a major flaw. There is also a propensity to become stuck in local minima (Pacelli & Azzollini, 2011), although this limitation is difficult to surpass with any type of model. This is due to the non-linear nature of the problem at hand, which makes the computation of global minima burdensome. Additionally, artificial neural networks display a limited ability to deal with large datasets, which means this technique becomes more time consuming to process the data (Abdou & Pointon, 2011).

As mentioned in section 1.1., although the research on credit scoring is very extensive, it does not reflect recent advances in predictive learning. One of the most prominent AI approaches that has recently started being used in classification problems respects random forests. This technique has been explored by few studies, although some sources report it as superior to earlier methods (Jones et al., 2015). Hence, the random forest method is selected as a technique to be researched in this thesis because of both its potential and the lack of comparative studies including this type of model.

2.3. Experimental Setup

In order to compare statistical and AI models in assessing credit risk, this project includes an experiment using a novel dataset to check which method achieves the best results in the prediction of defaults. The dataset fed to the models contains information about 1994 companies operating in Portugal during 2016. It consists of 24 financial and non-financial indicators for these corporations concerning the fiscal year (FY) of 2016, along with the financial status (insolvent, under special revitalization process or non-compliant) of these businesses in 2017. This dataset lacks the values for some entries due mainly to lapses in the database, but some of the cases may be caused instead by computational errors in the calculation of some indicators (e.g. attempting to calculate a ratio by dividing a value by zero).

This thesis should offer a comprehensive analysis of how these methods compare at predicting the default risk of these companies. Additionally, this research must determine the optimal architecture and/or parameters that allow each technique to reach the best results and if there are any shortcomings in implementing these solutions in a business setting.

2.4. Chapter Conclusions

The analysis of the financial situation of the Portuguese companies stresses the importance of credit scoring in B2B trading. It was observed that a significant portion of companies display poor financial results, which is indicative of an increased risk of default.

The credit scoring models included in the scope of this dissertation are diverse and display different strengths and weaknesses. Although AI predictors may achieve better results in terms of accuracy, this comes at the cost of increased complexity, which requires more computational power and leads to a reduced interpretability of the results due to the black-box syndrome. These factors must be taken into consideration when assessing the potential of the models.

The experiment conducted in this project with the sample of Galp’s B2B clients allows for a thorough comparison between the performances of statistical and artificial intelligence methods. This practical aspect of the dissertation makes the current research valuable and a significant contribution to the academic literature on credit scoring models.

3. Theoretical Framework

3.1. Linear Discriminant Analysis

The linear discriminant analysis method has been in use for a long time now, ever since being introduced by Sir Ronald Fisher in 1936. Initially, it was mainly utilized in biological and behavioral sciences (Altman, 1968). Researchers would use physical measurements of organisms as inputs of the LDA for taxonomic purposes. It was only later that this method began being applied to financial problems.

The LDA may be defined as a statistical technique utilized to classify an observation into one of several a priori groupings depending on the observation's individual characteristics (Altman, 1968). It should be stressed that each instance belongs to only one of the groups previously defined. It is one of the earliest statistical classifiers, sharing some characteristics with the regression analysis and the analysis of variance (ANOVA) methods.

Assuming a certain feature vector X containing values for s variables, $x_1, \dots, x_s$, there is an interest in knowing what linear function of these measurements best discriminates the groups in question (Fisher, 1936). This function can be defined as:

$$X = \lambda_1 x_1 + \cdots + \lambda_s x_s \tag{1}$$

In this expression, λ_i represents the discriminant coefficient for explanatory variable i. The goal for the LDA is to estimate the values for these coefficients that maximize the differences between the groups as measured by a given objective function.

3.1.1. F-ratio criterion

The original method defined by Fisher in his pioneering paper sought to find the coefficients that maximized the ratio of the difference between the means of the variables to the standard deviations within groups (Fisher, 1936). This ratio become later known as the F-ratio in honor of the statistician. It can be obtained by computing the following expression:

$$F = \frac{\sum_{g=1}^{G} N_g (\bar{y}_g - \bar{y})^2}{\sum_{g=1}^{G} \sum_{p=1}^{N_g} (y_{pg} - \bar{y}_g)^2} \tag{2}$$

This formula assumes the following notation:

G – Number of groups;

g – Group g (g = 1, …, G);

N_g – Number of instances in group g;

y_{pg} – Instance p in group g (p = 1, …, N_g);

$\bar{y}_g$ – Group mean;

$\bar{y}$ – Overall sample mean.

The numerator of expression (2) corresponds to the sums-of-squares between groups and the denominator to the within-groups sums-of-squares (Altman, 1968). This is equivalent to divide the explained variance observed in the dataset (the differences due to distinct group membership) by the unexplained variance in the dataset (the differences due to chance).

In order to get a practical perspective of the F-ratio, it is a good exercise to observe what output distributions provide higher or lower values for this measure. Figures 6 and 7 display plots of possible probability density functions for the output of two discriminant functions. The curves in red correspond to instances belonging to category I and the ones in blue to instances belonging to category II. Analyzing the plots, both the positioning and the slope of the curves provide clues into the F-ratio which can be obtained for these two cases.

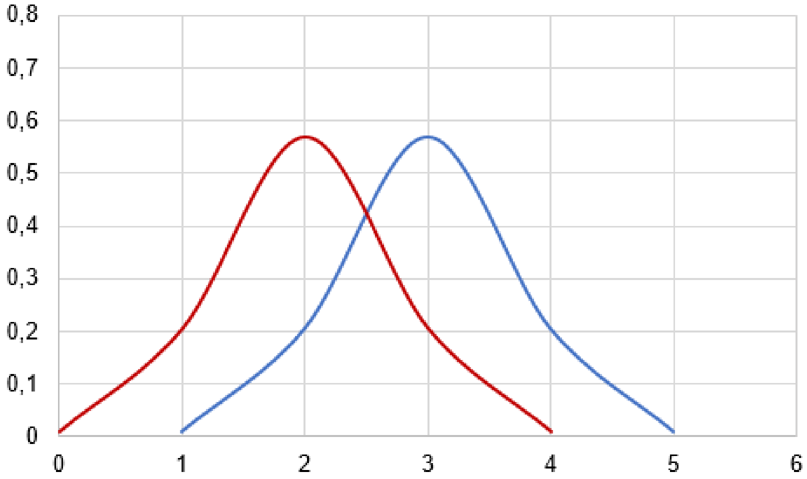


Figure 6 - Plots of the probability density functions for pair 1.

Figure 6 displays the curves for two normal distributions with means 2 and 3 and a variance of 0.7. The fact that the means are close to each other and the high variance contribute to a significant intersection between the distributions. This means that it is difficult to distinguish which category the instances belong to by the respective features. There are significant portions of the two categories displaying the same characteristics, hence it is hard to classify them.

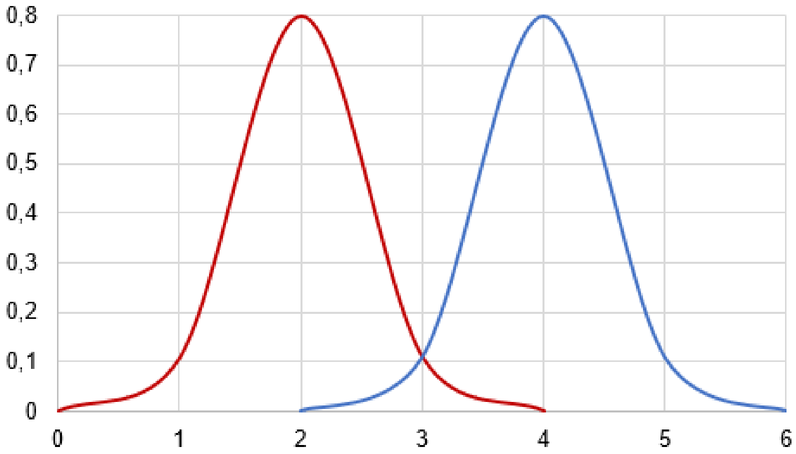


Figure 7 - Plots of the probability density functions for pair 2.

Figure 7 on the other hand, displays the curves for two normal distributions with means 2 and 4 and a variance of 0.5. Comparing with the previous case, one can observe the functions are steeper both going upwards and downwards and the peaks for the two distributions further apart. This means that there are less instances belonging to different categories displaying the same features. Considering expression (2), the numerator should be greater because of an increased distance between the means of the different groups and the denominator should be smaller due to a reduced variance within each group. Therefore, it is easier to discriminate between the two groups, which means the F-ratio for the second case will be greater than for the first.

3.1.2. Altman's Z-score

Once the coefficients have been computed to maximize the discriminant power of the function, it is possible to calculate the score for each observation in the sample and assign it to a certain group accordingly. This technique was first applied to credit scoring by Edward Altman in his 1968 paper "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". This approach is designated by Altman's Z-score and served as the basis for the future applications of discriminant analysis in credit scoring. Altman's method implies assigning each instance to the group it resembles the most. The comparisons are measured by a chi-square value and classifications are made based upon the relative proximity of the instance's score to the various group centroids (Altman,1968).

It should be noted that there are other variants of the discriminant analysis, differing from the Z-score namely in the objective function that is utilized to measure the discriminant power of the possible coefficients. This matter may generate some ambiguity when defining this method, however, this report details the original method as described in the works of Fisher and Altman.

3.2. Logistic Regression

The logistic regression (also known as logit model) is one of the most widespread statistical tools for classification problems in general (Ong, Huang, & Tzeng, 2005). The LR started being used early in the twentieth century, mostly in the area of the biological sciences. Much as the discriminant analysis, it is a technique utilized in problems with categorical dependent variables displaying linear relationships with the corresponding explanatory variables. Despite the similarities, it should be stressed that the logistic regression model does not assume the populations in classification problems to be normally distributed. Unlike the discriminant analysis, the LR can deal with various distribution functions (Press & Wilson, 1978; Ong, Huang, & Tzeng, 2005), and is thus, arguably, a better option in credit scoring tasks.

This technique may be applied to classification problems with a dichotomous outcome or ones displaying multiple classes. Assuming the case of a binary logistic regression that is used to determine if an event E will happen (e.g. company bankruptcy), then $\pi(x)$ may be defined as the probability of E occurring given the n-dimensional input vector X. As there are only two possible outcomes, 1 - $\pi(x)$ is equal to the probability of the event E not happening. The odds ratio may then be obtained by computing the following expression:

$$Odds\ Ratio = \frac{P(E|X)}{P(\bar{E}|X)} \tag{3}$$

The natural logarithm of the odds ratio is equal to the logit of $\pi(x)$ and corresponds to the linear form of the logistic regression (Agresti, 2007). This leads to the subsequent mathematical formulation:

$$logit(\pi(x)) = log\frac{\pi(x)}{1-\ \pi(x)}\ =\ \alpha + \beta X \tag{4}$$

The logit transformation ensures the linearization of the probability estimates and guarantees that the output of a LR is always restricted between 0 and 1 (Batista, 2012). On the other hand, a linear regression could produce negative probability estimates, which constitutes a deficiency of that method.

A different formulation of the logistic regression is usually obtained by relating the probability of a given event, E, happening, conditional on a vector X of observed explanatory variables, to the vector X (Press & Wilson, 1978). This corresponds to expression (5), which may be also obtained by manipulating the formulation (4).

$$\pi(x) =\ \mathrm{P}(E|x) =\ \frac{1}{1+e^{-\alpha-\beta X}} \tag{5}$$

The output of this expression describes a sigmoid curve taking values between zero and one. After the parameter α and the vector of coefficients β are calculated, it may be used as a predictor. The maximum likelihood (ML) method that is commonly used in statistics can be applied to estimate these parameters. This technique may be described as having two steps. First, there is the definition of both the distribution function of the dependent variable and the functional form that relates it to the values of the explanatory variables (Allison, 2012). In the case of the LR, the dependent variable displays a binomial distribution and the expression (5) can explicit the relation between X and the probability $\pi(x)$. The second step is to maximize the likelihood of achieving a correct prediction, usually by employing an iterative numerical method (Allison, 2012).

The logistic regression has seen a rise in popularity in recent years, with more research focusing on the applications of this method in credit scoring problems (Louzada et al., 2016). This contrasts with the trend to favor more sophisticated alternatives, namely AI methods. The ease of interpretability and the accuracy of the results obtained by the LR may explain this phenomenon.

3.2.1 Input Variable Selection Procedures for the LR

There are three techniques commonly used to define which input variables should be included when building a logistic regression model:

- Forward selection;
- Backward elimination;
- Stepwise methods.

In forward selection, the first step is to pick the best input variable at distinguishing between the dependent categories. Subsequently, other variables are added to obtain the sets of the two best variables, three best variables and so on, until there are no indicators left that meet the condition to be considered as input (Batista, 2012). In backward elimination, the starting point is to include all the explanatory variables as input. Afterward, indicators are removed one by one according to a pre-defined criterium, until all the variables left satisfy the necessary conditions. Only the indicators with the most predictive potential should be left in the final set.

Stepwise techniques are a combination of the previous two. These include criteria for the entry and removal of variables. Therefore, variables may be included in the set of inputs and also be excluded, which happens until the best set is determined.

The criteria used to evaluate the predictive potential of the variables may be the respective significance levels, likelihood ratios, Wald statistics, etc.

3.3. Artificial Neural Networks

An artificial neural network is an AI method that gets its name from its components which resemble biological neurons. Much in the same way as a biological neural structure modifies itself to perform cognitive tasks, this model adapts by changing its parameters in order to carry out a certain computational task (Angelini et al., 2008).

This method was first introduced by Warren McCullock and Walter Pitts in 1943. They showed this type of network could, in theory, perform any arithmetic or logic function (Khemakhem & Boujelbène, 2015). Ever since, this tool has been used for several goals, usually pattern recognition, classification or forecasting (Wójcicka, 2017). Artificial neural networks started being studied as a possible credit risk predictor in the nineties (Tang et al., 2018) and since then have become a mainstream tool utilized by several financial institutions and other companies.

Neural networks are composed of several artificial neurons, which can be regarded as processing units. The outline of the respective structure is provided in Figure 8. These elements are interconnected via synapses that convey values, with each one of these connections having an assigned weight. When a neuron performs a computation, the first step is to do a weighted sum of the inputs x_{ij} (this corresponds to the operation in the left-hand part of the ellipse in Figure 8), afterward, the result is used in the transfer function that will calculate the neuron's output y_j.

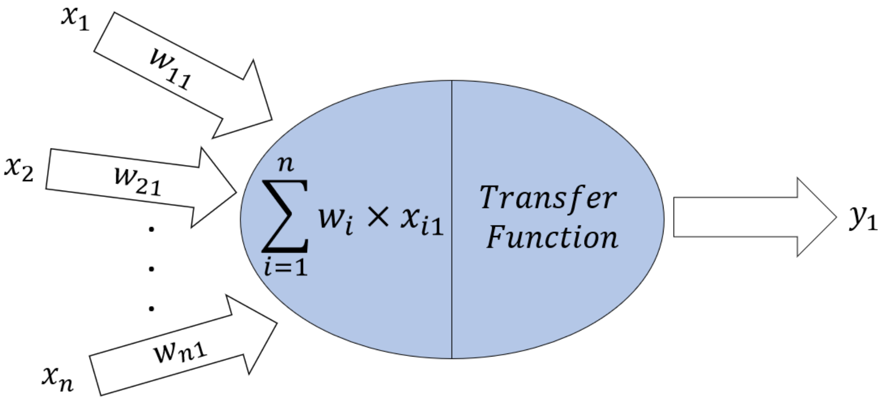


Figure 8 - Outline of an artificial neuron and its synapses.

Sigmoid, linear and step functions are common transfer functions (Angelini et al., 2008). These are plotted in Figure 9. There are no restrictions on the type of function to be used, but it is usually dependent on the type of ANN. The designer of the network should be careful as this choice will influence the quality of the results.

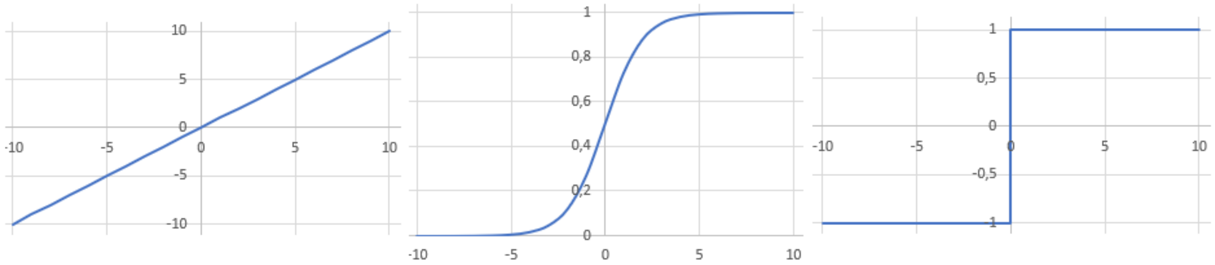


Figure 9 - Plots of a linear function (left), a sigmoid function (center) and a step function (right).

The way neurons are connected depends on the type of ANN. When each neuron is connected to all the neurons in the following layer, the network is called fully connected. If the networks allow loops in the flow of data, then these are called recurrent (Angelini et al., 2008). On the other hand, the feedforward networks correspond to structures where the data always flows from the preceding layers to the ones ahead, with no values being fed back to earlier layers (Pacelli & Azzollini, 2011).

Figure 10 shows an example of a quite simple feedforward type ANN with a minimum of layers. From left to right, the first layer is the input layer and its incoming synapses contain the original explanatory variables of the model. The response of this segment of the network is passed onto the hidden layers (in this example a single one for simplicity). These get their name from being in-between the inner and outer layers and come in an adjustable number. The output layer comes last, being composed by a single artificial neuron, element #7. After this neuron computes the result of the respective transfer function, the end-result y of the network is achieved.

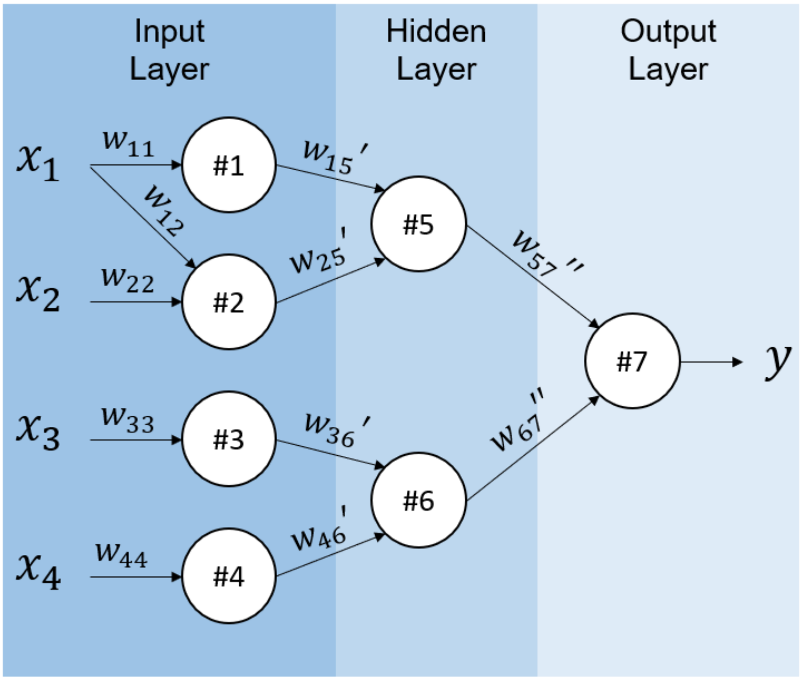


Figure 10 - Structure of a simple feedforward ANN.

As previously stated, artificial neural networks are dynamic systems. This means that some parameters change as it learns and improves results. The network checks the input and output values, so it can change the weights of the links to reduce the difference between the current result and the target (Agatonovic-Kustrin & Beresford, 2000). Thus, for each new entry of values for the input variables and the corresponding output that is fed into the network, the ANN adapts by modifying its weights and hence improve the accuracy of its predictions. This phase corresponds to the training of the ANN.

3.3.1. Learning Mechanisms

There are different types of learning mechanisms for neural networks (Angelini et al., 2008). These methods can be divided into three categories:

- Supervised learning;
- Unsupervised learning;
- Reinforcement learning.

Supervised learning consists in feeding an ANN with a training set which is composed of correct examples (Pacelli & Azzollini, 2011). Methods of this category are faster than the others because the weight adjustment is made directly through the error (Khemakhem & Boujelbène, 2015). This kind of learning procedure is used when the network must learn to generalize the given examples (Angelini et al., 2008).

In unsupervised learning, the training set only contains unlabeled data (i.e. there are inputs, but no outcome to group them). It is commonly utilized to search for patterns in big data in tasks such as data mining (Angelini et al., 2008).

Reinforcement learning, similarly to supervised learning, has a clear goal for the network (Reed & Marks, 1999). However, it is not the error function to drive the weight updates. There is a system of bonuses and penalties that evaluates the output for the ANN and guides the values attributed to the weights. It is frequently used in models that must complete a sequence of actions. In these cases, the outcome is dependent on the sequence of steps and, as such, each step must be assessed taking into consideration the wider chain of steps (Angelini et al., 2008).

3.3.2. Training, Validation and Testing Phases

There are different stages in the learning process of an ANN. These stages correspond to the training, validation and testing phases. The training and validation occur simultaneously, whereas the testing only takes place after the other two have been completed. Each one of these steps requires a distinct fraction of the original dataset. Essentially, the original input dataset must be divided into three parts (training set, validation set and test set), each with a different purpose. Consequently, the training-testing-validation ratio must be defined by the network designer. This parameter is extremely important because inadequate ratios do not allow for meaningful learning (Khashman, 2010).

The training set, as the name foreshadows, will be used in the training procedure described in the previous section. As this phase starts and the output error decreases, the validation set serves to control a phenomenon called over-fitting (Zhao et al., 2015). Over-fitting happens when the ANN begins to model noise in the training set. When modeling noise, although the accuracy rate is supposedly improving, this learning cannot be generalized for other data entries not part of the training set. As such, the predictive ability of the network is actually being degraded. In order to prevent this, with each iteration of the learning algorithm, the weights are only updated if this will mean a better success rate of the predictions in the validation set that is kept separate from the training. This procedure is commonly referred to as the early stopping technique (Kvamme et al., 2018). It serves as a backstop by ensuring that each new weight update truly betters the model and stops the training before the onset of over-fitting.

Finally, the goal of the testing set is to evaluate the predictive ability of the model. It serves as an independent way to check if the model generalizes well outside the sets used for training and validation. It serves as a final measurement of the quality of the ANN model.

3.3.3. Multilayer Perceptron (MLP) Neural Networks

The MLP is the most frequently used ANN in credit risk assessment (West, 2000) and has been tested in numerous studies for this purpose. It is a type of fully connected feedforward artificial neural network. The architecture of such networks usually displays several layers, with the first being the input layer which serves no computational role. The sole purpose of this layer is to pass the input to the following layers (Gardner & Dorling, 1998).

For the purpose of updating the weights in MLP networks, the most commonly used algorithm follows the backpropagation rule (Huang et al., 2018). Referred to as the backpropagation (BP) algorithm, it is a type of supervised learning model. It begins by initializing the weights with small random values (West, 2000). Subsequently, the gradient (i.e. the vector of derivatives of the error with respect to the weights) is computed and the weights are modified accordingly, in the direction which reduces the overall error of the network.

Table 2 shows the application of the BP algorithm by detailing the necessary steps. The term epoch is used to describe going through a full cycle (executing steps 2 through 5).

Table 2 - List of steps for an epoch of the backpropagation algorithm.

Step	Procedure
1st Step	Initialize the weights with small random values.
2nd Step	Feed a set of input values to the network.
3rd Step	Propagate the values though the ANN to obtain a result.
4th Step	Calculate the difference between desired and network outputs.
5th Step	Propagate the error backward and calculate the gradient.
6th Step	Adjust the weights according to the gradient.
7th Step	Repeat steps 2-7 for the following input entries until the overall error is acceptable.

In order to present a mathematical formulation of the backpropagation algorithm, the following notation is considered:

x_k – inputs for the network;

w_{kj} – weight for links between the input and hidden layer;

w_{ji} – weight for links between the hidden and output layer;

I_k – activation function for the input layer;

H_j – activation function for the hidden layer;

O_i – activation function for the output layer;

T_i – desired output;

$Error_i$ – error of the output result ($Error_i = T_i - O_i$);

η – learning rate.

Most of these parameters are self-explanatory, except for the learning rate. This rate is set by the network designer and corresponds to the speed at which the weights are updated. It is important to carefully tune the learning rate, so the adjustment can occur smoothly and without big jumps in the weights' values that can lead to instability.

The formula for updating the weights for links between the hidden and output layer (assuming a given activation function f) is as follows:

$$w_{ji}(t+1) = w_{ji}(t) + \eta H_j Error_i \frac{\partial f\left(\sum_j w_{ji} H_j\right)}{\partial w_{ji}(t)} \tag{6}$$

The expression for the weights concerning links between the input and hidden layers (assuming a given activation function f) is similar:

$$w_{kj}(t+1) = w_{kj}(t) + \eta I_k \frac{\partial f(\sum_k w_{kj} I_k)}{\partial w_{kj}(t)} \sum_i w_{ji}\, Error_i \frac{\partial f(\sum_j w_{ji} H_j)}{\partial w_{ji}(t)} \tag{7}$$

3.3.4. Radial Basis Function (RBF) Neural Networks

RBF networks are another common type of feedforward ANN that has been studied thoroughly in the credit risk analysis field. This model is comparatively quicker to learn than MLP type networks, but slower in computing an output and more demanding in terms of memory (Wójcicka, 2017). When comparing RBF and MLP networks, it is easy to conclude that these display an analogous structure, however, there are a few distinctive features.

Radial basis function ANNs are also composed of input, hidden and output layers. The first layer just carries the data directly to the hidden layer which is entirely composed of neurons with radial basis transfer functions, such as Gaussian functions (Ayala & Coelho, 2016). The output layer then performs a weighted linear combination of the results of these functions (West, 2000).

The outcome of a radial basis function is dependent on three parameters: the received input vector X, the center of the respective neuron c_j and the spread σ_j. The center corresponds to a point with as many dimensions as the input vector X (i.e. one dimension for each explanatory variable). The function evaluates the distance between X and c_j (the Euclidean distance is usually considered) to assess the similarity of the two vectors. The spread is set to control the smoothness of the drop seen in the function for greater distances. For each neuron j, assuming a Gaussian radial basis function as the transfer function, the output will be:

$$\varphi_j(X) = \exp\left(-\frac{\|X - c_j\|^2}{2\sigma_j^2}\right) \tag{8}$$

The training that RBF networks undergo allows for the determination of the appropriate number of hidden layers and also the best centers and widths for each hidden neuron (Chen, Wang, Liu, & Wu, 2018). These parameters will be the ones that allow for a minimization of the overall error of the network's output. This output may be computed by the following expression:

$$Y(X) = \sum_j w_j \exp\left(-\frac{\|X - c_j\|^2}{2\sigma_j^2}\right) \tag{9}$$

The estimation of the centers can be done via a clustering algorithm, as these are usually easy to apply and offer robust results (Ríha, 2016). The k-means clustering, for example, is one of the common and intuitive methods of this type. This algorithm considers a set of initial centers and then iteratively changes the centers to minimize the total within-cluster variance (Hastie, Tibshirani, & Friedman, 2009). All the input data points are attributed to the closest center, which effectively is dividing the data into separate subsets. Then each center is recalculated to correspond to the vector of the means for the features of the data points composing the respective subset.

After obtaining these parameters, a possible strategy regarding the spreads σ_j is to define them as the average distance between the respective center and the two closest neighboring centers. This is done in order to minimize any gaps or overlaps between clusters (West, 2000).

The remaining step in the training phase of RBF networks is to determine the optimal values for the weights of each link in-between layers. This can be achieved through the least-squares method for example.

3.4. Decision Trees

A sole decision tree (DT) model is a weak classifier for the purposes of credit scoring. However, DTs form the building block of the random forests' structure, being therefore imperative at this point to clarify how this method works.

A DT is composed of multiple pathways originating from a common starting point and ending at the final nodes, also called leaves. These pathways present several nodes, which function as branching or splitting points. Each instance in the data that passes through a node is assigned a path according to some pre-established criterion (e.g. Total Assets > 100 000 €), being ultimately directed to a certain leave and corresponding output. An example of a decision tree is provided in Figure 11.

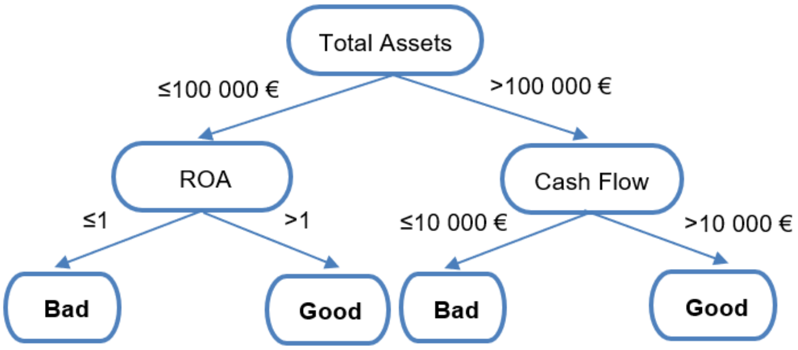


Figure 11 - Example of a decision tree used for a rudimentary credit scoring.

There are different algorithms to guide the construction of decision trees. These usually fall into three categories (Louzada et al., 2016): Chi-square automatic interaction detectors (CHAIDs), Classification and regression trees (CARTs) and C5.0/C4.5 decision trees.

The Chi-square automatic interaction detector (CHAID) method was introduced by G. Kass in 1980 as an improvement over the automatic interaction detector (AID) technique (van Diepen & Franses, 2006). The first stage of the algorithm is to discover the best split for each predictor and then select the one that is most advantageous overall. Subsequently, the data is split according to the chosen predictor and the subgroups present in the tree are re-analyzed independently, in order to produce further subdivisions for analysis (Kass, 1980). The significance level tests performed in this algorithm correspond to chi-squared tests, hence the name of the method. CHAID may be used for prediction and classification purposes.

The Classification and regression tree (CART) methodology was first presented by Breiman, Freidman, Olshen and Stone in their paper “Classification and Regression Trees” in 1984 (Timofeev, 2004). As well as CHAID, this method evolved from the original automatic interaction technique. The partitioning procedure of this algorithm is based on the Gini index. CARTs may be used for classification and regression purposes.

The C5.0 and C4.5 methods were developed by R. Quinlan, with C5.0 being a more recent technique based on the C4.5 algorithm. These algorithms are used mainly for classification purposes and work by splitting the data according to the entropy of the partitions. Entropy essentially measures the purity or homogeneity of the data segments and may be computed with the following expression:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2(p_i) \tag{10}$$

In this formula, the following notation is considered:

$S -$ Data segment or partition;

$i -$ Category index;

$p_i -$ Fraction of instances in a segment that belongs to category i;

$c -$ Total number of categories.

If the partitioning of the data were to perfectly separate the different categories, then each data segment would be assigned a fraction of 1 for a certain category, amounting to a total entropy of zero. In this case, each segment would be perfectly homogenous, displaying instances for a single category. If the partitioning is not perfect, then the entropy value increases, reaching one for the case of maximum chaos.

Regardless of the purpose or type of decision tree, it is important to decide whether to allow it to grow to its full extent or, on the other hand, limit its size. The procedure of limiting the size of decision trees is called pruning. This procedure is desirable because it helps prevent overfitting (Bradford, Kunz, Kohavi, Brunk, & Brodley, 1998). The pruning approach consists of breaking ties in a decision tree's structure until one finds the model with the best performance on previously unobserved cases (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993). Therefore, the key advantage of this procedure is that it allows the model to distinguish the noisy instances from the predictive patterns present in the data (Frank, 2000).

Additionally, it should be noted that decision trees are generally robust against the presence of outliers and missing data, as the node splitting seldom occurs at extreme values (Batista, 2012). DTs can also use a certain indicator repeatedly in the respective pathways, taking into consideration the interactions between different predictors (Espahbodi & Espahbodi, 2003). These characteristics represent major strengths of the DTs, but as it will be discussed in the next section, there is still room for improvement by considering several different decision trees, instead of a single DT. The random forest method capitalizes on the predictive power of several DTs, while averaging out any potential deficiencies of individual decision trees.

3.5. Random Forests

A random forest is an AI technique that belongs to the category of homogenous ensemble classifiers. It is an ensemble method because it relies on the output of multiple individual classifiers. As all these classifiers are of the same type, it is then defined as a homogenous ensemble. A random forest may be defined as a collection of decision trees.

The aggregation of the various outputs obtained into a single outcome (i.e. the random forest's prediction) may be done by averaging over all the output values when predicting a numerical outcome or by performing a vote when predicting a class (Breiman, 1996). There is much evidence that this procedure of model combination can lead to increased prediction accuracy (Paleologo, Elisseeff, & Antonini, 2010; Finlay, 2011; Lessmann et al., 2015).

In the context of classification problems, a random forest is analogous to a voting committee. Each decision tree reaches a prediction or classification and then the results of all trees are checked to find what is the output of the majority. Therefore, it is implied in this logic that the decision trees reach different results and consequently display distinct structures. A fundamental challenge when building a RF is thus to ensure the diversity of the decision trees.

The diversification of decision trees is achieved via two mechanisms:

- Bootstrap aggregating (bagging): this procedure allows for each tree to use a different sample as input without partitioning the data. These replicate datasets, each consisting of a given number of cases, are drawn at random, but with replacement, from the original dataset (Breiman, 1996). This means that an instance may be sampled multiple times or not be present at all in the data used to feed a given decision tree;
- Random feature selection: this mechanism dictates that each node is assigned a random subset of predictors that it may use in the node splitting procedure. Therefore, any explanatory variables that are not included in this subset may not be used in the splitting. This random selection of features at each node decreases the correlation between the decision trees, causing a reduction in the random forest error rate (Bryll, Gutierrez-Osuna, & Quek, 2003; Archer & Kimes, 2008).

Random feature selection has been demonstrated to perform better than bagging alone (Dietterich, 2000), namely in problems with several redundant features (Archer & Kimes, 2008). This strategy has also proven to perform well in comparison to other predictive methods, including discriminant analysis and neural networks, helping prevent the overfitting phenomenon.

When implementing this type of model, certain parameters must be defined in advance that will shape how the random forest is constructed. These are the total number of trees and the number of attributes that may be used to grow each tree (Brown & Mues, 2012). After the random forest is constructed, its results are not easily interpretable, which is inconvenient when it is critical to understand the interactions between the variables of the problem (Breiman, 2001). The black-box nature of this method contrasts with the ease of interpretation of decision trees, being a downside of its inherent increased complexity.

3.6. Key Performance Indicators (KPIs) in Credit Scoring

There are several metrics that may be used when comparing the performance of alternative predictive models (Addo, Guegan, & Hassani, 2018). The key performance indicators selected to evaluate the models in this research were picked by their popularity in the academic literature and ease of interpretation. A brief explanation is provided for each of these indicators in the following sub-sections.

3.6.1. Percentage Correctly Classified (PCC)

In general, the percentage of correctly classified instances is the most common quantitative measure utilized in the evaluation of the predictive models' results. This statistic can be obtained by the following expression:

$$PCC = \frac{Number\ of\ correct\ predictions}{Number\ of\ predictions\ performed} \tag{11}$$

This metric corresponds to the accuracy of the models and is the most intuitive of all the key performance indicators selected.

3.6.2. Error Types, Sensitivity and Specificity

Other important measures include the type I and type II error rates used in statistical hypothesis testing. In order to understand these errors, first it is beneficial to define a null hypothesis. Assuming the null hypothesis is that the company applying for credit will not default next year, then these errors are defined as:

- Type I error: An incorrect rejection of the null hypothesis. The model predicted that the company would default, when in fact this did not happen. This is commonly referred to as a false positive (FP);
- Type II error: An incorrect acceptance of the null hypothesis as true. The model predicted that the company would not default, but it defaulted the following year. This is commonly referred to as a false negative (FN).

Additionally, true positives (TP) and true negatives (TN) correspond to correctly rejecting and correctly accepting the null hypothesis, respectively. Considering this information, as defined by Huang et al. (2018), the average error rates are computed by:

$$Type\ I\ error\ rate = \frac{FP}{TN + FP} \tag{12}$$

$$Type\ II\ error\ rate = \frac{FN}{TP + FN} \tag{13}$$

These statistics are important in order to determine what type of error the models are most prone to. Generally, type II errors are considered more damaging to creditors than the type I errors (Huang et al., 2018), because the total amount of the credit may be lost in such situations. Type I errors do not mean a financial loss, just an opportunity loss because a profitable client was turned away. In spite of this, the majority of studies on credit risk analysis do not use these metrics to evaluate the performance of the respective models, which constitutes a major flaw in the literature on this topic (Tsai & Wu, 2008).

In alternative to these rates, one can assess the performance of a credit scoring model through the sensitivity and specificity. These parameters correspond to the true positives rate and true negatives rate, respectively (Batista, 2012). The formulas for the computation of these rates are found below.

$$Sensitivity\ (True\ Positives\ Rate) = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity\ (True\ Negatives\ Rate) = \frac{TN}{TN + FP} \tag{15}$$

3.6.3. Area under the ROC Curve (AUC)

The AUC is a commonly used performance indicator to benchmark different credit scoring models. In order to understand this concept, it is fundamental to define the receiver operating characteristic (ROC) curve. The ROC curve is obtained by plotting, for each classification threshold, the rate of true positives against the rate of false positives (Swets, Dawes, & Monahan, 2000).

The threshold in these models corresponds to the cutoff value separating the categorical outcomes, which means that, assuming a binary classifier, outputs under a given threshold indicate a certain category and the ones over it indicate the other category.

Figure 12 displays how distinct threshold levels influence the results obtained with a predictive model. Assuming the red and blue curves to be representative of the probability distributions of the model's outputs for classes "Bad" and "Good", respectively, then it is clear there is a tradeoff between type I and type II errors. For threshold 1, there is a high risk of predicting "Good", when in fact the correct class was "Bad", and for threshold 2, there is a high risk of predicting "Bad" incorrectly, because the case belonged to the "Good" class. The optimal cutoff value will depend on the problem at hand, namely on which type of error is most detrimental.

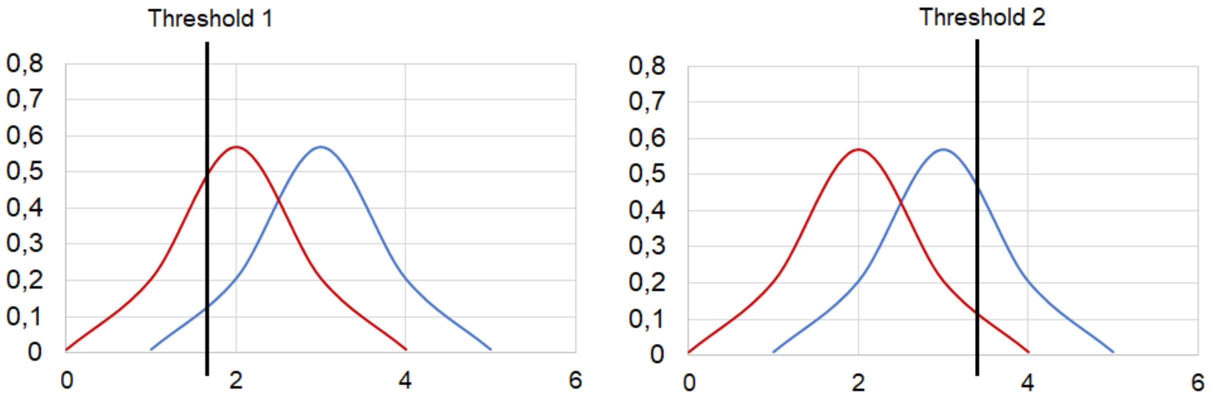


Figure 12 - Examples of different threshold levels.

Figure 13 displays an example of a ROC curve (in red), which may be obtained by checking, for each possible threshold, what are the corresponding true positive and true negative rates. Using these values as coordinates in a graphic, one can plot the ROC curve. The area under this curve corresponds to the AUC. It provides a way to understand how well a given model differentiates between the distinct data classes, regardless of the threshold selected. This is a fundamental concern when comparing various models, as different cutoff values may be established according to the aversion to type I and type II errors.

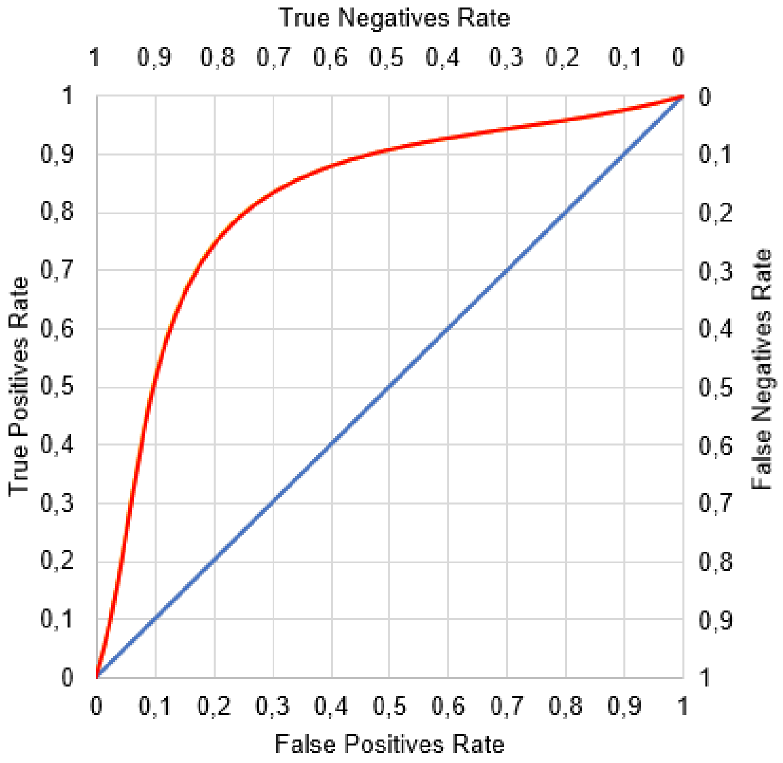


Figure 13 - Example of a ROC curve along with a line representing chance accuracy (adapted from Swets et al., 2000).

The blue line in Figure 13 corresponds to the results of a prediction with chance accuracy. This represents the case of a perfectly random classifier, which scores an AUC of 0.5 (Hand, 2009). In this figure, it is noticeable that the area under the blue line corresponds to 50% of the total. Better classifiers will obtain ROC curves that bow more to the left, denoting superior accuracy. The Gini coefficient, another commonly used performance indicator in benchmarking studies of credit scoring methods, is a chance standardized alternative of the AUC. According to Anderson (2007), the relationship between the AUC and the Gini coefficient may be expressed by the following expression:

$$AUC \approx \frac{Gini\ coefficient + 1}{2} \qquad (16)$$

It is also relevant to explain an alternative interpretation of the AUC. Assuming, for instance, that bad companies are coded as 0 and good companies as 1, then it may be said that the AUC equals the probability that a random bad company will obtain an output in a credit scoring model lower than a random good company.

3.6.4. Limitations of the Performance Indicators and Alternatives

Although the PCC and AUC are the most widely used performance indicators for evaluating credit scoring models, there are several other measures that may be used. Furthermore, some authors have questioned the suitability of the AUC measure to compare different classifiers (e.g. Hand, 2009). It has been argued that the AUC assumes different cost distributions for different classifiers (Lessmann et al., 2015). This distribution of misclassification costs should be dependent on the classification problem, not on the type of classifier (Hand & Anagnostopoulos, 2013). Alternatives such as the H-measure have been proposed to address the limitations of the AUC.

However, Lessmann et al. (2015) have suggested that the classifier rankings obtained by using the AUC and the H-measure do not differ much. The empirical results for these two measures display a correlation value of 0.93. Several other popular performance indicators display pairwise correlations values of around 0.9, pointing to a high similarity between the classifier ranks obtained using these scores. These results suggest that a limited number of performance indicators is enough to compare different classifier methods, while also indicating that the AUC limitations presented by Hand (2009) do not cause major discrepancies between the ranks obtained with this measure and with other alternative indicators.

As the AUC remains the most popular performance indicator in the credit scoring literature and its potential shortcomings do not seem to produce a major impact in the ranks obtained, it was deemed appropriate to employ it to evaluate the predictive models of this research. Furthermore, this choice allows for the comparison between these models and others previously applied in the academic literature. The high correlation observed between the ranks obtained with the different performance indicators, as described by Lessmann et al. (2015), justifies the choice to only use a reduced number of these measures in this research.

The PCC was included in this selection of performance indicators as it provides a more intuitive global view of the accuracy of the models. The sensitivity and specificity of the results are also relevant indicators, especially in cases of unbalanced datasets in which the PCC and AUC measures may mask a propensity of the models to a certain error type. These two measures will, therefore, allow to differentiate between type I and type II errors, which is of paramount importance.

3.7. Chapter Conclusions

This chapter allowed for the analysis of the main credit scoring techniques, which in turn led to a comprehension of the necessary concepts for the development of such models. Furthermore, the study of the performance indicators used in credit scoring permitted the definition of the appropriate measures to be used in the later stages of the research.

By examining the academic literature relating to the linear discriminant analysis, it became clear that this method displays several variants. However, all these versions suffer from some shortcomings that ultimately led to the disuse of this technique. During the early stages of the research, it was noticeable that the recent credit scoring research articles seldom discussed the application of the discriminant analysis method. This perception was subsequently confirmed by checking benchmarking studies that pointed to the near absence of research in recent years regarding the LDA (e.g. Louzada et al., 2016).

This contrasts with the use of the logistic regression and AI methods. The popularity of the LR technique is particularly surprising, as other statistical techniques are clearly being abandoned in favor of artificial intelligence models. As the literature indicates that the logistic regression performs better than other statistical methods, namely the discriminant analysis, it is relevant to understand if this is confirmed by the results of the credit scoring experiment of this dissertation. Lastly, the study of the procedures used in the LR for the selection of input variables was important, as this aspect impacts the regression obtained and must be taken into consideration when implementing this model.

Regarding the ANNs, the comprehension of the underlying structure of these networks is vital to grasp the black-box phenomenon. Additionally, this analysis is necessary to the proper definition of several parameters during the model development stages (such as the number of layers to be used, partitioning of the dataset, possible transfer functions, etc.).

The research regarding the random forest method allowed for an understanding of how the decision trees function and how these may be combined to produce a significantly better model. The study of the different types of DTs also permitted a comprehension of several techniques (namely pruning procedures) that will aid in the development of the random forest model and the critical analysis to be performed of the results obtained.

Finally, the analysis of previous credit scoring models in the literature stressed the need to differentiate between error types. It also became clear that the AUC indicator is the most widely used performance metric in research settings. Although some authors have recently brought into question some potential shortcomings of this indicator, it remains prevalent in state-of-the-art studies. Additionally, later articles pointed out that the rankings produced with the AUC and other popular indicators do not differ significantly.

Therefore, it was concluded that only a small selection of KPIs will suffice for the evaluation of the models' quality and that these former limitations are of limited importance (or may be shared by all major performance indicators). Furthermore, by using the AUC, it is possible to compare the models developed in this dissertation with others present in the literature.

4. Input Data Collection, Analysis and Treatment

4.1. Input Data Collection Process

The data used in the models was obtained from the Orbis financial database. This service is widely used by private companies, governments and financial institutions worldwide. Bureau van Dijk (BvD), a Moody's Analytics Company, is responsible for the capture (from regulatory and other sources) and treatment of the data present in this database. The company website lists credit risk analysis as one of the main purposes for the use of Orbis. The access to the database is provided in exchange for a subscription fee, albeit there is a free trial version available online at: https://www.bvdinfo.com/en-gb/contact-us/free-trial?product=orbis.

The financial information used in this research was extracted for a list of Galp's clients and concerns the fiscal year of 2016. It was obtained by searching for these companies' respective BvD ID numbers and subsequently selecting the relevant financial and non-financial information. This data was then extracted via an Excel file which may be utilized as input in the software packages that are used to develop the models.

Additionally, the information regarding the clients' financial status (active, insolvent, under special revitalization processes or non-compliant) in the FY of 2017 was retrieved from the internal data kept by Galp. This last indicator, which corresponds to the dependent variable of the models, was then appended to the Excel file exported from Orbis.

It should be noted that the initial version of the dataset included fewer cases and was extremely unbalanced (the active companies vastly outnumbered the ones of the remaining categories). This is commonly regarded as a problem in prediction and classification algorithms (Kvamme et al., 2018). The performance of predictive models may substantially deteriorate under these circumstances.

Taking this situation into account, a new dataset was sought that included a higher number of financially distressed companies. After obtaining additional cases for this type of business, the respective information was also extracted from the Orbis financial database and merged with the previous instances to create a larger dataset. By assembling this wider sample, it was then possible to apply a sampling procedure to assure its balance. This step is detailed further in this report in section 4.4.2.

The wider sample displays 6378 instances for 24 financial indicators in addition to the variable for the financial status of the companies in 2017. It should be noted that all the data contained in the dataset corresponds to information which these corporations must publish in accordance with Portuguese fiscal law. These variables are listed and briefly explained with regards to their meaning and computation method in the following section.

4.2. Description of the Input Variables

In order to obtain the most explanatory input variables, several financial and non-financial indicators were gathered from the data exported from the Orbis database. Table 3 details the types of variables that were tested, along with the indicators that pertain to each of these categories.

Table 3 - Categories of explanatory variables tested in the models and corresponding indicators.

Type of Indicator	Variables	Units
Raw Financial	ln(Total assets)[1]	ln(€)[1]
Equity Ratios	Shareholder equity ratio	-
Growth Trends	Cash flow variation (2015 – 2016)	Percentage
	Total assets variation (2015 – 2016)	Percentage
	Equity variation (2015 – 2016)	Percentage
Sector of Activity	BvD major sector	-
Company Maturity	Number of years active	Years
Profitability Ratios	ROE using profit or loss before tax	Percentage
	ROA using profit or loss before tax	Percentage
	ROCE using profit or loss before tax	Percentage
	ROE using net income before tax	Percentage
	ROA using net income before tax	Percentage
	ROCE using net income before tax	Percentage
	Profit margin	Percentage
	EBITDA margin	Percentage
	EBIT margin	Percentage
	Cash flow / Total assets	Percentage
	Profit per employee	Thousands of € per employee
Operational Ratios	Net assets turnover	-
	Credit period	Days
Structural Ratios	Liquidity ratio	-
	Current ratio	-
	Gearing	-
	Debt / EBITDA	-

[1] ln stands for the natural logarithm

It should be noted that the units are absent in certain indicators present in Table 3. These cases correspond to the financial ratios, which are dimensionless, and the BvD major sector that is a categorical variable.

The raw financials type of variable was among the first tested in the predictive models. This data is essentially taken from the companies' accounting books and presents a very straightforward interpretation. However, one must be prudent when drawing conclusions from these indicators. Companies with very different risk profiles may present similar metrics (e.g. the same value for net income).

Considering, for example, two companies, one has a slightly better net income, but it is also much larger than its counterpart, then the smaller company is a lot more profitable even though it produces a reduced net profit. These financials metrics fail to account for the size of the companies. In order to somehow mitigate this undesirable size factor, the natural algorithm was used when considering these indicators. A logarithmical scale reduces the differences seen between distinct companies in a certain variable, while also keeping the original ranking unchanged. Additionally, it may be beneficial not to altogether ignore the size of the companies, as this feature may also hold explanatory potential if larger companies display reduced chances of default. Indeed, some studies indicate that a company's size is positively related to the provision of trade credit (Andrieu, Staglianò, & van der Zwan, 2018).

Regarding the equity ratios, these are fundamental to get a perspective of how the companies are structured in terms of the capital employed. These are essentially metrics that evaluate how much capital corresponds to the companies' own resources. It is intuitive that a self-funded business has a decreased probability of default because there are fewer financial obligations.

Another issue that may potentially dictate the outcome of a given company is whether there is a clear tendency of growth in the past years. A business that has a negative net income but has been able to cut its losses significantly in the past, may have an increased probability to reach positive results in the following periods. In order to capture this momentum, some indicators for 2015 were also exported from the Orbis financial database, allowing for a crude evaluation of any potential growth trends.

In order to test if companies from different sectors of activity have distinct risk profiles, the variable BvD major sector was extracted from the Orbis financial database. This variable groups together companies in broad sectors of activity, such as the construction or the transportation sectors. A more detailed classification was not possible, as more restrictive categories would severely diminish the number of companies per sector, therefore reducing the model's ability to generalize from this information.

The maturity of the company was also a factor considered to have explanatory potential regarding the companies' probability of default. Older companies with more experienced and/or determined owners may fare better than upstarting businesses. Petersen and Rajan (1997) have found that the age of a company may be an important proxy for the respective robustness and the reputation it has among potential lenders. Indeed, after partitioning the dataset into good and bad companies, it was observed that bad companies are on average almost nine years newer than the good ones.

The profitability measures are crucial to understand the financial returns of these companies. A higher return on equity, for example, is indicative of a good corporate performance and hence a decreased probability of default. Regarding the cash flow type indicators, these are also profitability measures, but only consider the net amount of cash and cash-equivalents that is transferred in or out of a company's accounts.

Operational ratios encompass various indicators conveying information about the way a business is run. This category may include variables measuring the turnover rate of the companies' assets, the average time these businesses take to pay the respective debts or even the level of commitment in research and development activities. Therefore, this category includes several seemingly disparate indicators that relate to the operational aspects of the businesses.

Lastly, the structure ratios are a vital type of financial indicator used to assess the capability of an entity to pay off the current debt obligations with its own resources. This information is very relevant to understand how financially robust a company is, which may in turn indicate the probability of default.

The final indicator, the company status in the following year, corresponds to the dependent variable in all the models. In this variable, all companies are assigned to the mutually excluding categories:

- Active: The company remains in operation in 2017;
- Insolvent: The company has filed for bankruptcy in 2017;
- Undergoing a Special Revitalization Process (under Portuguese law, *Processo Especial de Revitalização - PER*): The company has been given a protection against creditors status, which allows it to continue operating while limiting creditors requests, which prevents an imminent insolvency;
- Non-compliant: The company has not fulfilled its financial obligations to Galp.

Although the definitions of active and insolvent companies are straightforward, a clarification on the Special Revitalization Process (SRP) and on the non-compliant companies is warranted.

Portuguese companies that are in a very difficult financial situation with a high probability of default in the future may apply for an SRP status. This process is only triggered after a vote by all creditors on whether they are in favor or against SRP. It should be noted that the immediate effect of this process is a ban on any coercive debt collection. Indeed, the revitalization plan might include debt pardons, extended deadlines for the repayments and reductions on the interest rates. In practical terms, this process may worsen the creditors' situation, since it may be just delaying an inevitable insolvency.

Regarding the last category, companies are considered non-compliant when these entities have failed to pay for the products and/or services of Galp as was agreed under the terms of the deal. These cases will lead to contentious debt settlements or reimbursements from insurance companies if the amount in question was insured.

4.3. Formulas for the Financial Ratios

Below are listed the formulas necessary to calculate the financial ratios mentioned in the previous section. It should be noted that in the cases of companies displaying negative results, the profit before tax is substituted by the loss in the computation of these ratios.

$$ROE\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Equity} \times 100 \tag{17}$$

$$ROA\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Total\ assets} \times 100 \tag{18}$$

$$ROCE\ using\ profit\ before\ tax = \frac{Profit\ before\ tax}{Capital\ employed} \times 100 \tag{19}$$

$$ROE\ using\ net\ income = \frac{Net\ income}{Equity} \times 100 \tag{20}$$

$$ROA\ using\ net\ income = \frac{Net\ income}{Total\ assets} \times 100 \tag{21}$$

$$ROCE\ using\ net\ income = \frac{Net\ income}{Capital\ employed} \times 100 \tag{22}$$

$$Profit\ margin = \frac{Profit\ before\ tax}{Sales} \times 100 \tag{23}$$

$$EBITDA\ margin = \frac{EBITDA}{Operating\ revenue} \times 100 \tag{24}$$

$$EBIT\ margin = \frac{EBIT}{Operating\ revenue} \times 100 \tag{25}$$

$$Profit\ per\ employee = \frac{Profit\ before\ tax}{Number\ of\ employees} \tag{26}$$

$$Credit\ period = \frac{Creditors}{Sales} \times 100 \tag{27}$$

$$Net\ assets\ turnover = \frac{Sales}{Equity + Non\ current\ liabilities} \tag{28}$$

$$Current\ ratio = \frac{Current\ assets}{Current\ liabilities} \tag{29}$$

$$Liquidity\ ratio = \frac{Current\ assets - Stocks}{Current\ liabilities} \tag{30}$$

$$Gearing = \frac{Non\ current\ liabilities - Loans}{Equity} \tag{31}$$

$$\frac{Debt}{EBITDA} = \frac{Long\ term\ debt + Other\ non\ current\ liabilities + Loans}{EBITDA} \tag{32}$$

$$Equity\ variation = \frac{Equity(2017) - Equity(2016)}{Equity(2016)} \times 100 \tag{33}$$

$$Total\ assets\ variation = \frac{Total\ assets(2017) - Total\ assets(2016)}{Total\ assets(2016)} \times 100 \tag{34}$$

$$Cash\ flow\ variation = \frac{Cash\ flow(2017) - Cash\ flow(2016)}{Cash\ flow(2016)} \times 100 \tag{35}$$

4.4. Preprocessing of Inputs

4.4.1. Preprocessing of Categorical Attributes

Some predictive models are not capable of using the categorical variables directly as input. The dataset to be used in this research contains two such variables, the BvD major sector and the company status. In the case of the company status, this indicator acts as a dependent variable in the models, for which reason it is not necessary to code it into a numerical variable.

In order to code the BvD major sector variable into numerical attributes that may be used as input in all the predictor models, there is a need to create auxiliary binary variables that hold this information. Thus, a binary variable is created for each possible sector. These assume a value of one when the instances belong to the respective sectors and zero otherwise. As such, each binary variable may assume values 0 or 1, which signal “true” or “false”, respectively, regarding the presence of the specific category that is being coded.

It should be noted that this coding may lead to the removal of certain auxiliary variables from the models. Some sectors are poorly represented in the dataset (i.e. display a low number of cases), which may prevent them from reaching an appropriate significance level. Therefore, by breaking down the BvD major sector into several variables that indicate the presence or absence of each individual sector, a group of non-significant sectors may be excluded from the analyses.

Although the company status variable does not require such numerical coding, it still poses a challenge, as it must be decided whether to aggregate the negative categories under broader classes or, instead, keep all the categories separate. This is a unique problem of this research because the current literature on credit risk prediction uses mostly datasets containing only two options regarding the debtors’ risk profile, usually denoting it as “good” or “bad”. This matter must be carefully analyzed in order to discover what is more beneficial, keeping two possible outcomes by merging the insolvent, SRP and non-compliant categories into a single class (as all denote an impossibility to receive the credits conceded or at least a serious delay), merging only two of these categories, or having four distinct outcomes by keeping all categories separate. These combinations are displayed in Table 4.

Table 4 - Possible combinations for the aggregating classes “Good” and “Bad”.

Number of Outcomes	Active class	Insolvent class	SRP class	Non-compliant class
2 Outcomes	Good	Bad	Bad	Bad
3 Outcomes	Good	Bad	Bad	Non-compliant
	Good	Bad	SPR	Bad
	Good	Insolvent	Bad	Bad
4 Outcomes	Good	Insolvent	SPR	Non-compliant

Although there are several possible combinations for each distinct strategy, a preliminary analysis is enough to understand that some seem counterintuitive. The discriminant analysis, as well as the artificial neural networks and other predictive models, offer similar predictions for close inputs, as such, it is disadvantageous to merge classes that are characterized by very dissimilar inputs. Therefore, one must take this factor into consideration when deciding on the best course of action regarding the aggregation of classes.

Both insolvent and SRP companies display similar very poor financial indicators, as may be confirmed in Appendix A. Therefore, this pair of classes is the most logical choice to undergo merging. Non-compliant companies display better financial indicators in comparison with the other two negative categories, although these indicators remain deteriorated in relation to active companies. Consequently, the second and third combinations for a 3 outcomes strategy are excluded.

Upon experimenting with the remaining aggregation strategies, it became evident that it is beneficial to keep only two possible outcomes. It is difficult for credit scoring models to differentiate between the different categories of financially distressed companies. This is due to the similarity of the inputs obtained for insolvent, SPR and non-compliant classes. Furthermore, from the perspective of the entity issuing the credit, it is irrelevant to know which of these categories applies to a given applicant. The main goal of a creditor is to understand if there is a significant risk of default for any given potential debtor, and it is notorious that the applicants included in these three classes present such a risk. Considering these factors, it was ultimately decided to pursue the two-outcome aggregation strategy that is presented in Table 4.

4.4.2. Sampling Procedure

Regarding the sampling process, there are two key aspects to consider:

- Sample size: the total number of cases listed in the dataset;
- Sample balance: proportions of good and bad cases in relation to the total number of instances.

Although the majority of credit scoring research has not focused on the input samples' characteristics, the size and balance of such datasets have a tremendous potential to affect, for better or worse, the performance of the predictive models. Some methods are more sensitive than others to changes in the input data's size and structure, but both statistical and AI techniques are affected by these features to varying degrees.

The standard procedure for sampling credit applicants is to retrieve around 1500 cases of bad companies and another 1500 of good companies, with this number of cases being considered sufficient (Crone & Finlay, 2012). This means that a perfectly balanced sample of 3000 cases is recommended for the construction of robust credit scoring models.

The dataset extracted from Orbis proved very unbalanced, as is presented in the following table.

Table 5 - Original distribution of the cases by the dependent variable's categories.

Category	Number of cases	Percentage of total instances
Active	5372	84,2%
Insolvent	701	11%
SPR	265	4,2%
Non-compliant	40	0,6%
All categories	**6378**	**100%**

There are two options to deal with unbalanced datasets, under-sampling and over-sampling techniques. Under-sampling is used to reduce the number of instances in the majority class, while oversampling increases the number of cases pertaining to the minority class.

In this research, it was decided to under-sample the majority class, which encompasses the cases of good companies. Although over-sampling may produce better results according to Crone & Finlay (2012), this dataset proved extremely unbalanced, which makes it difficult to employ this technique. Considering that the minority class is much smaller, the over-sampling would cause certain cases in the minority class to be repeated several times. This repetition may, in turn, cause the models to overfit, therefore degrading the results. Additionally, while under-sampling the majority class, there is the possibility of removal of non-valid instances (due to missing values), which will increase the integrity of the sample.

After selecting a random subset of valid instances from the good companies' class which displayed a total number of cases equal to the one of bad companies, the near perfectly balanced dataset described in Table 6 was obtained.

Table 6 - Distribution of the cases by the dependent variable's categories in the balanced dataset.

Category	Number of cases	Percentage of total instances
Active	1001	50.2%
Insolvent	701	35.2%
SPR	265	13.3%
Non-compliant	27	1.4%
All categories	**1994**	**100%**

The slightly bigger number of active companies in relation to the total instances of negative categories is due to a few detected cases of duplicated companies in the data. These occurrences were caused by two factors. Firstly, some instances correspond to non-compliant companies that declared insolvency or entered an SRP during 2017, therefore being present in both categories. Secondly, some of the active companies did not provide payment for products/services, which explains the remaining duplicated cases.

In order to solve this issue, these cases were attributed to a single category. The active companies that did not provide payment were exclusively assigned to the non-compliant class, as it is counterintuitive to consider such companies as good clients. The non-compliant companies that also entered an SPR or declared insolvency in 2017 were removed from the non-compliant category.

After this procedure, there is no longer a 1:1 ratio, but the difference is considered irrelevant (the proportion is 1.0081 good companies for each bad company).

4.4.3. Missing and Invalid Data

Another important aspect to be addressed by the data preprocessing relates to the presence of missing values in the dataset. When building predictive models, there should be an exhaustive search for any missing values. After all instances have been identified, one must then investigate the causes of the missing data entries.

The usual reasons for missing values in credit scoring problems are that those values were already missing in source data (e.g. the source database) or were out of the theoretical allowed range. The latter motive is quite common in these situations due to typos or transcription errors (Angelini et al., 2006). On the other hand, these lapses may be due to computational errors (e.g. attempting to compute a ratio which has zero as the denominator).

After analyzing the dataset, two main types of invalid cases were detected, NA and NS instances. The first one corresponds to the instances that are truly missing, NA being an acronym for not available in the Orbis financial database. In order to understand the meaning of the second term, there was a need to contact Galp's client manager regarding the Orbis services. This query clarified that NS stands for not significant and is used in situations where indicators expressed as percentages take values near zero.

As NS instances do not truly represent missing data, these were replaced by null values in the dataset. This is an approximation that allows for the use of such instances in the predictive models. Although the real values for these indicators might not be exactly zero, there is no way to export this information. Therefore, it was deemed beneficial to set them as zero, since there was a significant number of cases in this situation that would otherwise be excluded from the analyses.

There were a few cases detected of instances with invalid values due to divisions by zero. These errors were present in financial ratios computed from the data present in Orbis in which the denominator corresponded to an indicator with a null value. The reduced number of such instances means that this situation does not affect the models significantly, consequently requiring no further treatment or analysis.

Finally, some instances were detected of public entities in the data. These cases are not relevant for the credit scoring exercise, as such organizations are largely different from private companies regarding the financial indicators displayed and may not become insolvent even in extreme situations. In order to prevent these instances from skewing the results, there was an exhaustive search for all the public entities in the dataset. Subsequently, these cases were eliminated from the data. In order to maintain the sample balance, the active companies that were found to be public organizations were replaced by an equal number of valid instances belonging to the same category.

4.4.4. Correlation Analysis

The multicollinearity problem refers to the existence of strong correlations between independent variables in a dataset. This phenomenon becomes a problem when fitting regression models to the data. Many authors have stated before that the logistic model becomes unstable when there exists strong dependence among explanatory variables, as it seems that no single variable is important when all the others are in the model (e.g. Hosmer & Lemeshow, 1989; Ryan, 1997; Aguilera, Escabias, & Valderrama, 2006). This weakness is shared with the linear regression and discriminant analysis methods.

Multicollinearity can cause slope parameter estimates to have magnitudes or signs that are not consistent with expectations and, in some situations, lead independent variables in a regression model not to demonstrate statistical significance, despite large individual predictor-outcome correlations and a large coefficient of determination, R^2 (Thompson, Kim, Aloe, & Becker, 2017).

A common technique to detect multicollinearity issues involves the computation of the variance inflation factor (VIF) measure. As specified by Wasserman & Kutner (1983), the VIF may be computed by the following expression:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{36}$$

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 (Thompson et al., 2017).

There are researchers who consider that values of VIF over 10 are indicative of multicollinearity (e.g. Chatterjee & Price, 1991 and Midi & Bagheri, 2010). However, other authors point out that this threshold is very lenient. A VIF of 10 implies a R_j^2 equal to 0.9, which is the same as saying that 90% of the variability in the independent variable j is explained by the remainder independent variables (Thompson et al., 2017). Another typical threshold is a maximum VIF of 5 (Craney & Surles, 2002). This is a more conservative approach that was deemed adequate, as certain variables displayed VIF values nearing 10 and would not be excluded with the former criterium. Table 7 presents the VIF values computed for the variables in the original and after removal datasets.

Table 7 - VIF values for the original and after removal datasets.

Variable	VIF (original set)	VIF (after removal)
ROE using profit or loss before tax	9.213	Removed
ROCE using profit or loss before tax	271.180	Removed
ROA using profit or loss before tax	38.077	Removed
ROE using net income before tax	9.223	1.172
ROCE using net income before tax	269.722	1.329
ROA using net income before tax	36.189	1.707
Profit margin	7.320	3.146
EBITDA margin	4.731	3.183
EBIT margin	10.439	Removed
Net assets turnover	1.092	1.080
Credit period	1.197	1.193
Current ratio	1.583	1.577
Liquidity ratio	1.519	1.518
Debt / EBITDA	1.001	1.001
Gearing	1.105	1.102
Cash flow / Total assets	3.239	3.234
Equity / Total assets	3.058	3.057
Profit per employee	1.110	1.093
Equity variation 2015-2016	1.184	1.172
Total assets variation 2015-2016	1.094	1.093
Cash flow variation 2015-2016	1.182	1.180
Number of years active	1.130	1.119
ln(Total assets)	1.370	1.356

As may be seen in the table above, there are clear signs of multicollinearity in the original data, with several VIF values exceeding the threshold defined. In order to solve this problem, the variables were removed iteratively until no VIF values were over 5. This removal procedure was performed giving preference to the variables that are more correlated (the ones displaying the largest VIF values).

The final dataset obtained displays no indications of multicollinearity, as may be verified in the rightmost column of Table 7. Four variables were removed, and it is interesting to note that these indicators were closely related to other variables in the dataset. The ROCE, ROE and ROA using profit or loss were related to ROCE, ROE, and ROA using net income, while the EBIT margin depends largely on the EBITDA and profit margins. Therefore, it seems logical that the removal of these particular indicators will decrease the dependencies between the variables.

This reasoning is also supported by the correlation matrix obtained for these variables. Table 8 presents the highest correlation coefficients detected and the respective pairs of variables.

Table 8 - Listing of the highest Pearson correlation coefficients.

Pair of variables	Pearson correlation coefficient
ROCE using profit before tax - ROCE using net income	0.996
ROA using profit before tax - ROA using net income	0.983
ROE using profit before tax - ROE using net income	0.963
EBIT margin - Profit margin	0.914

Furthermore, a strong argument in favor of the multicollinearity analysis performed in this project is the absence of such high correlations after the removal of the variables displaying the largest VIF values. The correlation matrix for the final selection of variables is presented in Appendix B.

4.4.5. Outlier Analysis

Although there is no universally accepted definition, several authors refer to outlier instances as observations that appear to deviate markedly from other members of the sample in which these occur (e.g. Grubbs, 1969; Barnett & Lewis, 1994; Hodge & Austin, 2004).

Outliers may be the result of errors, fraudulent activity, novelties in the data, among other reasons. It is important to address this phenomenon, as outlier instances pose a challenge to the development of the predictive models. When fitting a model to the data, outliers need to be identified and eliminated, or, alternatively, examined closely if these cases are the focus of the analysis (Beliakov, Kelarev, & Yearwood, 2011). In credit scoring, these instances are of limited interest, but the potential to negatively affect the results of the models must be eliminated.

Although some models have a structure that is inherently robust to the presence of extreme values, such as decision tree type predictors, other methods, namely regression techniques, are particularly susceptible to parameter skewing as a result of using datasets containing outliers. The maximum likelihood principle that is frequently used in logistic regressions is not robust against outliers (Beliakov, et al., 2011). Moreover, parameters such as the empirical mean and the covariance matrix that are frequently used in the classical discriminant rules are highly influenced by outlying observations, which will cause these rules to be inappropriate (Hubert & van Driessen, 2004).

The detection of outliers is simple when dealing with univariate data or even two-dimensional data. In these situations, the existence of aberrant data instances can be checked using simple boxplots for example. However, when there is a greater number of variables, it is not possible to rely solely on visual perception and it becomes necessary to employ an algorithm to detect these instances (Rousseeuw & van Zomeren, 1990).

According to Filzmoser (2004), the basis for multivariate outlier detection is the Mahalanobis distance (MD). The MD is an alternative to the Euclidean distance, with both being measures of distance in the multivariate space. The key feature of the Mahalanobis distance is that it considers the correlations between variables, as well as the respective scales (Brereton & Lloyd, 2016). The MD may be computed with following expression:

$$MD = \sqrt{(x_i - \bar{x})S^{-1}(x_i - \bar{x})^T} \tag{37}$$

This formula uses the following notation:

x_i – Vector for a given data instance;

$\bar{x}$ – Arithmetic mean of the dataset;

S – Sample covariance matrix.

The Mahalanobis distance for a dataset with n explanatory variables that display normal distributions follows a chi-squared (χ^2) distribution with n degrees of freedom. In these cases of multivariate normality, a common procedure is to compare the MD with a critical value of the chi-squared distribution. The instances with a MD over the value of the χ^2 distribution for a given quantile (e.g. 95%) are then labeled as outliers.

However, this procedure suffers from some shortcomings previously identified by certain authors. The computation of the MD for a given instance relies on the sample mean, which is not a robust indicator in the presence of several outliers. These extreme values may distort the observed mean by moving it closer in multivariate space to the outlier points, which may in turn cause two undesirable phenomena. First, a small cluster of outliers may impact the mean in such a way that these are no longer detected as aberrant instances. Secondly, the distortion brought on by the outliers may be so high that normal instances are wrongly labeled as outliers. These occurrences are commonly referred to in the literature as masking and swamping, respectively.

Some studies in various fields of research have proposed alternative procedures for outlier detection that seek to minimize the masking and swamping effects. Some examples of these are techniques with the computation of Mahalanobis distances (MDs) with robust indicators, such as the method proposed by Leys, Klein, Dominicy & Ley (2018) with a minimum covariance determinant approach, or entirely distinct approaches using projection pursuit strategies.

In this research, it was decided to examine the presence of outliers via the computation of the Mahalanobis distances with the geometric medians (GMs). This indicator is one of the most common robust estimators of centrality in Euclidean spaces (Fletcher, Venkatasubramanian, Joshi, 2008).

The geometric median (GM) follows an intuitive concept, although its computation presents a reasonable challenge. Considering a multivariate sample, the geometric median corresponds to the point that minimizes the sum of the Euclidean distances to all the instances present in the dataset. As defined by Aftab, Hartley and Trumpf (2015), considering a sample with k cases, the GM can be obtained by solving the minimization problem found below.

$$m = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{k} d(x, y_i) \tag{38}$$

This formula uses the following notation:

$d(x, y_i)$ – Operator for the Euclidean distance between x and y_i;

$\underset{x}{\operatorname{argmin}}$ – The argument x that minimizes the sum of the GMs.

In order to address this problem, the Weiszfeld algorithm is employed. This method is frequently used in the computation of the GM. It is an iterative procedure that with the appropriate initialization values converges to the point that presents the lowest sum of Euclidean distances for all the sample instances. In accordance with Fritz, Filzmoser and Croux (2012), the Weiszfeld algorithm consists of the iterative application of the expression that follows.

$$T_0(m) = \frac{\sum_{i=1}^{k} \frac{y_i}{\|y_i - m\|}}{\sum_{i=1}^{k} \frac{1}{\|y_i - m\|}} \tag{39}$$

As long as the estimate T_0 does not coincide with any of the cases present in the sample, this algorithm will converge to the geometric median. This condition may be represented mathematically by the expression below.

$$\widehat{m}_{l+1} = \begin{cases} T_0(\widehat{m}_l) \; if \; \widehat{m}_l \notin \{y_i, \dots, y_k\} \\ \widehat{m}_l \; if \; \widehat{m}_l \in \{y_i, \dots, y_k\} \end{cases} \tag{40}$$

The computation of the geometric medians does not tolerate missing or otherwise invalid instances. As such, there is a need for a method that replaces these invalid cases by usable data. The techniques used for this purpose are called imputation procedures. These may be divided in two categories, single and multiple imputation techniques. The fundamental difference between these categories is that the first implies the generation of a single estimate for each missing value, while the latter dictates that several estimates are produced for each lapse in the data. Taking into consideration that the objective of the current analysis is the computation and comparison of the robust Mahalanobis distances, it is advantageous to implement a single imputation technique. These techniques are inherently simpler and, more importantly, allow for a straightforward comparison of the robust MDs.

There are several ways to perform a single imputation procedure, however, one must acknowledge that some of the more commonly used techniques present fundamental flaws. For example, the mean imputation method is perhaps the most frequently utilized procedure and is defined as the replacement of the missing/invalid instances using the observed mean for the variable in question. This procedure will underestimate the variance, disturb the relations between variables and bias almost any estimate (van Buuren, 2018). Thus, this technique is unacceptable in the current problem due to its potential to skew the estimates of the robust MDs.

In order to choose the imputation method that best fits this problem, it is also important to clarify the type of missing data that is at hand. Missing data are known to be completely at random (MCAR) when their absence is not related to both observed and unobserved data (Twisk, de Boer, de Vente, & Heymans, 2013). The data may also be missing at random (MAR), which means that the probability of a particular set of values being missing for an instance does not depend on the values themselves, conditional on the observed values of other variables (White & Carlin, 2010). Finally, often the most problematic situation is when the data is missing not at random (MNAR). In this case, the probability of a certain value being absent depends on the missing value itself (Dong & Peng, 2013).

SPSS offers several options in terms of both single and multiple imputation procedures. By defining the number of imputations as one in the multiple imputation techniques, it is also possible to use these methods for the generation of a single estimate for each missing/invalid value. This is beneficial as a sole stochastic regression imputation may be superior to the deterministic single imputation procedures available. A certain degree of variability in the estimates is vital, as deterministic methods lead to the underestimation of the data's variance. Nevertheless, the generation of multiple estimates for each lapse in the data would be preferable to further reduce the biasing of the variance towards zero, but this is not compatible with the goal of directly comparing the robust MDs. Thus, it was decided to use one of the stochastic multiple imputation procedures available in SPSS to compute the single estimates.

This category of imputation methods has two options, fully conditional specification, which is an iterative Markov chain Monte Carlo (MCMC) technique, and a monotone procedure. The SPSS software is capable of automatically choosing one of these procedures in accordance with the pattern of missing values observed. Assuming an incomplete rectangular data table, the dataset is said to be monotone if any variable is either at least or at most as observed as any other variable (Liu, 1995). As the scan of the dataset indicates that it is non-monotone, SPSS automatically selects the MCMC method, as this technique does not assume any specific missing data pattern.

Most multiple imputation procedures, including fully conditional specification, generally assume that the data is MAR or MCAR (Liu & De, 2015). In order to understand which type of missing data is displayed in the credit scoring dataset, Little's MCAR test was performed. Assuming the null hypothesis is that the data is missing completely at random, then a p-value inferior to 0.05 is interpreted as indicating that the data is not MCAR. As the output of this test is a significance value of 0.00, the null hypothesis is rejected.

Considering this result, it may be concluded with a high degree of certainty that the data is not MCAR. This is also coherent with what was perceived from the data so far. For example, bad companies display a much higher percentage of missing/invalid instances than the remaining corporations. This would not happen if the probability of a certain value being missing was independent of the observed values of the other indicators. However, it is still necessary to understand if the data displays a MAR or MNAR pattern.

The distinction between MAR and MNAR is based on a non-testable assumption (Harel & Zhou, 2007). Therefore, the only option to distinguish between these patterns of missing data is to seek an explanation for this phenomenon. Upon enquiring an Orbis representative about the causes behind the missing data, it was discovered that these lapses are most likely due to situations in which the companies fail to disclose their full financial information. This tends to occur in cases of businesses that, although are undergoing insolvency proceedings or entering special revitalization processes, remain in operation in the period in question. As there is no indication that the lapses are due to the missing values themselves, the data is assumed to be MAR, which allows for the application of multiple imputation procedures.

After separating the sample into two groups, dataset A and dataset B, which contain exclusively good and bad companies, respectively, it was possible to impute the values using the fully conditional specification method. Since the whole sample contains two distinct populations with very different characteristics, this splitting procedure is fundamental to assure that the MD is computed with the GMs of the class (good or bad) to which each instance belongs. The geometric medians obtained are detailed in the following table.

Table 9 - Geometric medians for the two classes of companies.

Independent Variable	GM for the good companies	GM for the bad companies
ROE using net income	5.0842	-11.9968
ROCE using net income	5.1136	-12.6036
ROA using net income	2.5344	-11.8482
Profit margin	3.4549	-11.2633
EBITDA margin	9.5170	-6.0160
Net assets turnover	2.1344	2.4427
Credit period	40.9809	102.3660
Current ratio	3.7691	1.9431
Debt / EBITDA	3.1521	-0.4057
Liquidity ratio	3.3115	1.4996
Gearing	0.6767	0.3353
ln(Total assets)	13.0830	13.2194
Cash flow / Total assets	7.1446	-0.3541
Number of years active	23.5109	22.1655
Equity variation (2015-2016)	0.0705	-1.1444
Total assets variation (2015-2016)	0.0532	0.1114
Shareholder equity ratio	0.5173	-1.4001
Profit per employee	3.3685	-7.8761
Cash flow variation (2015-2016)	0.0470	-0.0507

As can be observed by comparing the results displayed in Table 9 with the average values for each indicator presented in Appendix A, there are significant differences. This was expected due to the non-robust nature of the arithmetic mean in the presence of outliers.

The next step was to perform normality tests in SPSS for the explanatory variables. These tests demonstrated, with a high degree of certainty, that several indicators do not follow normal distributions. Consequently, it was decided to use an alternative exclusion criterion to the comparison of the MDs with a quantile of chi-squared distribution, as there is no guarantee that the Mahalanobis distances follow this particular distribution in the absence of multivariate normality. By constructing scatter plots with the sample identification numbers and the robust MDs, it is possible, via visual inspection, to detect any potential outliers.

Regarding the good companies, as can be perceived by analyzing the scatter plot in Figure 14, there are a couple of instances that stand out for being anomalous. These are marked in red for easier identification. Company 126 presents a robust MD nearing 6000, a much higher value than those of the remaining corporations. Additionally, company 1 also stands out with a robust MD of almost 1500. These are the only instances that present robust Mahalanobis distances over the 1000 mark. As such, these companies are flagged as potential outliers.

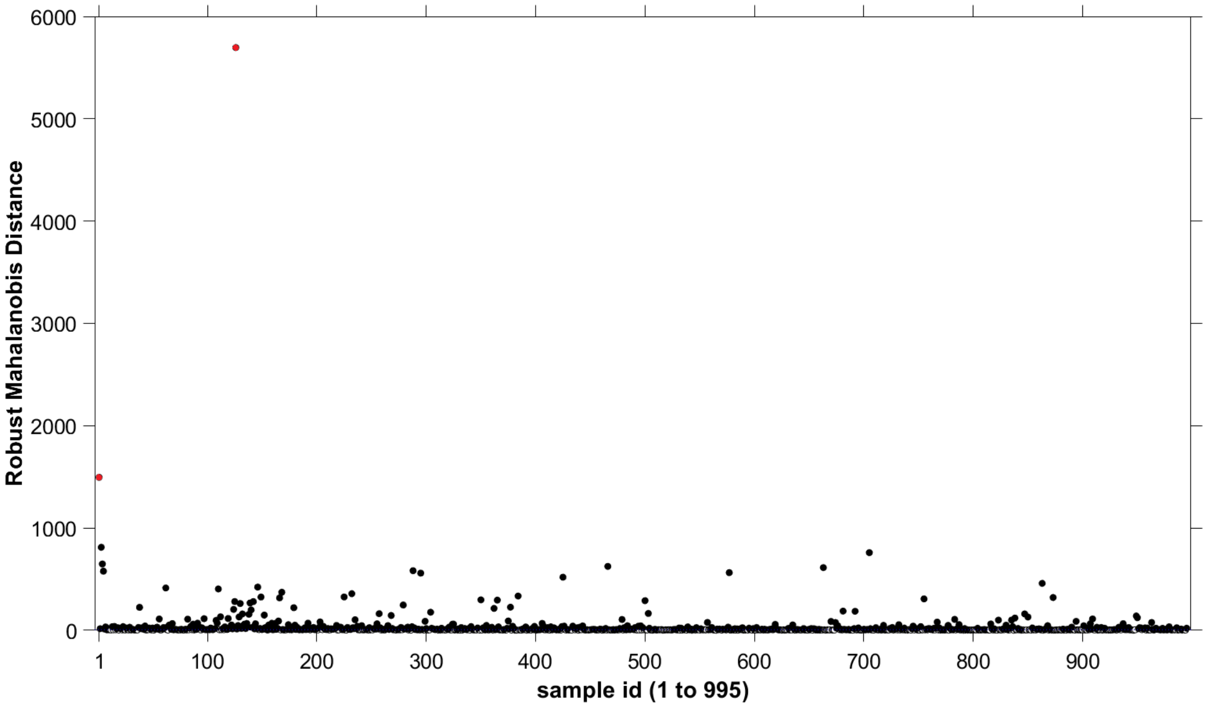


Figure 14 - Scatter plot of the robust MDs for the good companies.

As for the bad companies, the scatter plot in Figure 15 allows for the identification of the outliers in this subset, which are also marked in red. In comparison with the good companies, this group presents a much more erratic distribution of the robust MDs. In order to isolate the most aberrant instances, all companies with robust Mahalanobis distances over 1000 were flagged as potential outliers.

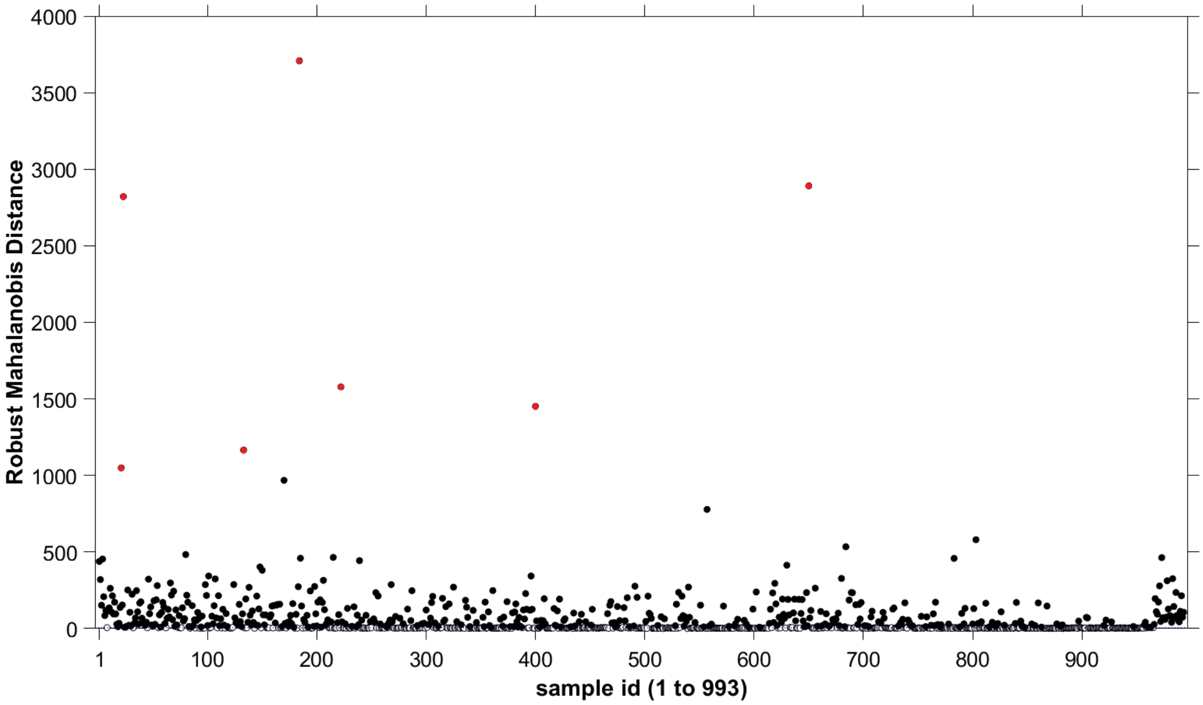


Figure 15 - Scatter plot of the robust MDs for the bad companies.

In order to comprehend to what extent these flagged instances are aberrant, there was an analysis of the indicators presented by these companies. This study reinforced the idea that such companies display altered values for several indicators.

Considering that the results of the robust MDs analysis were confirmed in both datasets A and B by the subsequent findings of extreme values for several indicators in the flagged cases, the decision was taken to label these nine instances as outliers and remove them from the sample.

It should be noted that the outlier detection procedure implemented in this project is partially based on the work of Semechko (2019). Further details are provided in the reference section.

4.5. Chapter Conclusions

This chapter allowed for a detailed description of the sample to be used as input in the credit scoring models. It described how the companies were selected from the portfolio of Galp's B2B clients and the procedure of extracting the financial information from the Orbis financial database. The various financial indicators considered in this research are meant to allow for different aspects of these companies to be captured in the models. These variables reflect different characteristics, such as the profitability of the companies, the financial autonomy of these businesses, operational metrics, sector of activity, among others.

As for the pre-processing of the data, this procedure was fundamental to treat the input and assure that the models can reach better results. Regarding the sampling of the instances, a balanced dataset was sought in accordance with what is recommended in credit scoring settings. This was done by under-sampling the majority class, which also allowed for the removal of non-valid instances. In terms of the categorical variables, a method was established to assure their conversion into numerical attributes that could be used in all the models.

There was also an exhaustive search for any instances with missing or otherwise invalid values. After detecting these cases, there was an effort to replace these values with valid entries or, in certain situations, remove them from the dataset entirely.

A correlation analysis was performed to assess any multicollinearity issues in the data. After detecting these dependencies among variables, the most affected indicators were removed from the dataset. The correlation analysis is particularly important to the logistic regression, which would be severely impacted if this issue was not addressed before its implementation.

Finally, there was an investigation into whether the data contained outliers. These extreme cases can severely impact the results of the statistical models. In order to tackle this issue, an analysis with the computation of robust Mahalanobis distances was performed. As it was necessary to consider all the cases in this procedure, even the ones with incomplete information, there was an imputation technique in place to fill in any missing data. This study led to the identification of 9 potential outliers. As these instances were confirmed to display aberrant values for several indicators, it was decided to exclude them from the dataset.

5. Model Development

5.1. Discriminant Analysis

The discriminant analysis model was applied to the data with IBM SPSS Statistics 25.

This software offers various stepwise techniques for the selection of inputs, more specifically:

- Method 1 (Wilks' lambda): Selection is based on how much each variable lowers the overall Wilks' lambda;
- Method 2 (unexplained variance): The variables are entered in order to minimize the sum of the unexplained variation between groups;
- Method 3 (Mahalanobis distance): The variables are entered in order to boost the MD between groups;
- Method 4 (lowest F ratio): Selection is done in order to maximize an F ratio computed from the Mahalanobis distance between groups;
- Method 5 (Rao's V): At each step, the variable that maximizes the increase in Rao's V is entered.

Additionally, the conditions that may be applied in each method for the entry and removal of variables are the following:

- F value thresholds: Variables are inserted in the model if the corresponding F values are over a given entry threshold and removed if the F value falls below a certain minimum.
- Probability of F thresholds: Variables enter the model if the significance levels of the respective F values are under a defined minimum and removed if the significance levels exceed a maximum cutoff value.

Considering the capabilities of the software, alternative discriminant analysis models were computed using the different combinations of stepwise techniques and entry/removal criteria. Regarding the maximum and minimum thresholds used in the selection of variables, it was decided to use the default settings of SPSS for the F value method and custom limits for the probability of F method. This decision was taken after observing that by using the default settings in both methods, the results produced were ultimately the same. Therefore, the thresholds for the probability of F method were set manually to be more conservative than the pre-defined values of SPSS. These thresholds are specified in Table 10.

Table 10 - Thresholds for the entry and removal of variables in the models.

Method	Entry	Removal
F value	3.00	1.00
Probability of F	0.05	0.10

After applying the different discriminant analyses, several key performance indicators were calculated in order to facilitate a robust comparison of the models. These results are displayed in Table 11.

Table 11 - KPIs obtained for the different combinations of stepwise methods and entry/removal criteria.

Stepwise method	Entry/removal criteria	PCC (%)	Sensitivity (%)	Specificity (%)
Method 1	F value	80.0	88.9	67.7
	Probability of F	79.8	88.9	67.1
Method 2	F value	80.0	88.9	67.7
	Probability of F	79.8	88.9	67.1
Method 3	F value	80.0	88.9	67.7
	Probability of F	79.8	88.9	67.1
Method 4	F value	80.0	88.9	67.7
	Probability of F	79.8	88.9	67.1
Method 5	F value	80.0	88.9	67.7
	Probability of F	79.8	88.9	67.1

By observing the values for the different KPIs, one can conclude that the best results are achieved by the methods with the F value thresholds. However, it is also notorious that the stepwise technique used does not impact the performance metrics. This situation arises because, for each entry/removal criteria, all the stepwise methods have selected the same indicators to be used as inputs. Considering these results, it was decided to apply the stepwise techniques using the F value rules, which led to the determination of the indicators with the most predictive potential. This selection of indicators is presented in Table 12, along with the unstandardized and standardized canonical coefficients.

Table 12 - Independent variables selected and corresponding coefficients.

Independent variables	Unstandardized coefficients	Standardized coefficients
ROE using net income	0.002	0.125
ROCE using net income	0.003	0.188
ROA using net income	0.024	0.379
Profit margin	0.010	0.182
EBITDA margin	0.008	0.168
Credit period	-0.003	-0.316
Current ratio	0.014	0.094
ln(Total assets)	-0.116	-0.186
Cash flow / Total assets	-0.690	-0.391
Food, beverages and tobacco sector	0.440	0.101
Hotels and restaurants sector	0.311	0.082
Post and telecommunications sector	-2.993	-0.085
Primary sector	0.769	0.182
Textiles, wearing apparel and leather sector	-1.040	-0.225
Transport sector	0.581	0.176
Number of years active	0.006	0.085
Shareholder equity ratio	0.005	0.730
Constant	1.590	-

The standardized coefficients are important to assess the discriminating ability of the explanatory variables. The standardization allows for the comparison of variables expressed in distinct scales. Positive and negative values are indicative of the direction of change in the LDA model's output when the variables increase. The five variables with the most predictive potential are thus, the shareholder equity ratio, Cash flow / Total assets, ROA using net income, credit period and the textiles sector binary variable, by descending order of discriminating ability. On the other hand, the unstandardized canonical coefficients may be used for the direct calculation of the discriminant function, which corresponds to the expression below:

$$DA(X) = 0.002 \times x_1 + 0.003 \times x_2 + 0.024 \times x_3 + 0.010 \times x_4 + 0.008 \times x_5 - 0.003 \times x_6 + 0.014 \times x_7 - 0.116 \times x_8 - 0.690 \times x_9 + 0.440 \times x_{10} + 0.311 \times x_{11} - 2.993 \times x_{12} + 0.769 \times x_{13} - 1.040 \times x_{14} + 0.581 \times x_{15} + 0.006 \times x_{16} + 0.005 \times x_{17} + 1.590 \quad (41)$$

In this formula, the following coding is considered:

- X- Input data vector;
- x_1- ROE using net income;
- x_2- ROCE using net income;
- x_3- ROA using net income;
- x_4- Profit margin;
- x_5- EBITDA margin;
- x_6- Credit period;
- x_7- Current ratio;
- x_8- ln(Total assets);
- x_9- Cash flow / Total assets;
- x_{10}- Food, beverages and tobacco sector;
- x_{11}- Hotels and restaurants sector;
- x_{12}- Post and telecommunications sector;
- x_{13}- Primary sector;
- x_{14}- Textiles, wearing apparel and leather sector;
- x_{15}- Transport sector;
- x_{16}- Number of years active;
- x_{17}- Shareholder equity ratio.

This expression may be used to calculate the discriminant scores for each instance in the data, which are indicative of the predicted group membership. The average scores for good and bad companies are 0.658 and -1.090, respectively. If a given instance obtains a discriminant score that is close to the score of a group centroid, then it probably belongs to that group. However, the prediction always depends on a cutoff value that is previously defined which segregates the classes.

Based on these scores, it is possible to plot a receiver operating characteristic curve that studies the model's performance for the range of possible cutoff values. Subsequently, it is also possible to compute the AUC and Gini index. The ROC curve plotted in Figure 16 yields an AUC of 0.863, which corresponds approximately to a Gini index of 0.726. These KPIs are utilized in the latter performance comparison between the different credit scoring models.

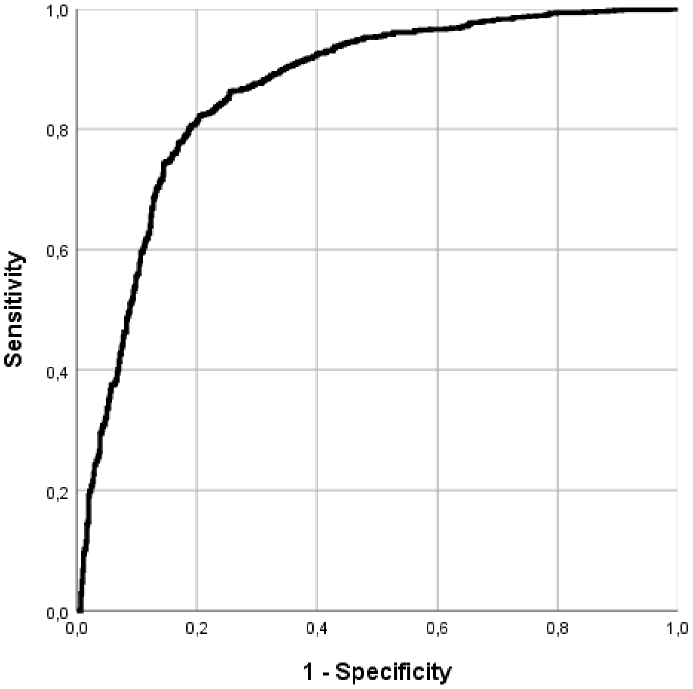


Figure 16 - ROC curve for the LDA model.

5.2. Logistic Regression

The binary logistic regression model was applied to the data with the IBM SPSS Statistics 25 software. There is no need to apply a multinomial logistic regression, as the considered output of the model is dichotomous.

The first step in the development of this model is picking the input variable selection procedure. As described in IBM's documentation for SPSS, this software has six options regarding these methods:

- Method 1: Entry testing based on the significance of the score statistic and removal dependent on the probability of a likelihood-ratio statistic based on conditional parameter estimates;
- Method 2: Entry testing based on the significance of the score statistic and removal dependent on the probability of a likelihood-ratio statistic based on maximum partial likelihood estimates;
- Method 3: Entry testing based on the significance of the score statistic and removal dependent on the probability of the Wald statistic;
- Method 4: Backward elimination with removal dependent on the probability of the likelihood-ratio statistic based on conditional parameter estimates;
- Method 5: Backward elimination with removal dependent on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates;
- Method 6: Backward elimination with removal dependent on the probability of the Wald statistic.

Additionally, there is also the option to manually enter the input variables desired, regardless of the respective contribution to the LR model. After experimenting with all the selection procedures above, it was concluded that the best results were obtained with the forward selection techniques (methods 1, 2 and 3). The variables chosen by these techniques to be included as inputs of the LR were the same, which ultimately led to identical regression models. It should be noted that the comparison between selection procedures was done by considering only the sensitivity, specificity and overall PCC. In this stage, the AUC and Gini index were not readily available in SPSS.

The maximum number of iterations before model termination was kept at 20, the default setting in SPSS, as overriding this configuration did not improve the results. In terms of the thresholds used in the stepwise methods, the best results were obtained when the probability for the score statistic must be less than 0.01 for entry and over 0.03 for removal. The option to include a constant in the LR model remained selected. Furthermore, the SPSS user interface allows for the definition of the classification cutoff directly, which was kept at 0.5. Although it is relevant to study the model's performance under different thresholds, this is addressed later with the computation of the remaining KPIs.

Table 13 displays the output obtained for the best logistic regression model, which achieved a percentage of correctly classified cases of 89.9%, a sensitivity of 93.8% and a specificity of 83.5%. From left to right, the first column contains the variables selected as input, the second displays the coefficients for the regression and the third presents the statistical significance of each variable. Additionally, the exponentials of the coefficients B are also included, as well as the respective 95% confidence intervals.

Table 13 - Logistic regression model obtained with methods 1, 2 and 3.

				95% CI for exp(B)	
	B	Sig.	exp(B)	Lower	Upper
Cash flow / Total assets	0.120	0.000	1.128	1.101	1.156
Shareholder equity ratio	0.072	0.000	1.075	1.064	1.085
Textiles sector	-1.873	0.000	0.154	0.060	0.393
Total assets variation (2015-2016)	0.003	0.002	1.003	1.001	1.004
ROA using net income	-0.068	0.000	0.934	0.907	0.963
ln(Total Assets)	-0.216	0.001	0.806	0.710	0.914
Constant	1.452	0.110	4.271	-	-

The statistical significance is used in the context of hypothesis testing. The null hypothesis for each variable is that the respective coefficient is zero. When the significance is under a certain p-value defined beforehand (usually 0.05), then it can be assumed that the coefficient is not zero with a high degree of certainty. All variables included in the LR are statistically significant, even when considering a more conservative p-value of 0.01. This is a strong argument in favor of the predictive power of the indicators used.

In terms of the coefficients themselves, it is more intuitive to interpret the respective exponentials, which are odds ratios. One can interpret these values as a measure of the effect that an increase of one unit in the independent value would have on the probability of predicting that a case corresponds to a good company. This reasoning is valid assuming that the other independent variables remain unchanged. Values above zero denote that this event has become more likely (e.g. if exp(B) is equal to 2, then it is twice as likely that a given case corresponds to a good company), on the other hand, if the value is less than one, then it becomes less likely (e.g. if exp(B) is equal to 0.5, then the probability that the company is good is halved).

Analyzing the values for exp(B) in Table 13, it can be verified that most variables behave in the way it was expected. The Cash flow / Total assets ratio shows an odds ratio that is quite high in comparison with the others, indicating that a company with only a slight increase in this indicator is much likelier to be a good company. This reflects the inherent superior predictive power of this variable.

The shareholder equity ratio and the variation of the total assets also display odds ratios above one, which is an intuitive result. The companies that are self-financed will have less debt, thus being at a reduced risk of default. Additionally, an increase in the total assets of a company may be a consequence of business growth, which is beneficial to its finances.

On the other hand, the binary variable for the textiles, wearing apparel and leather sector, which signals if a company belongs to this particular sector, presents an odds ratio near zero. This is coherent with the histogram presented in Appendix C. Looking across the different categories in this graph, the sampled companies in this specific sector are predominantly bad. Although discrepancies like this are present in other sectors, the textiles, wearing apparel and leather sector was better represented in the sample. This justifies the inclusion of this variable, as a high number of cases is necessary to reach statistical significance in the model.

As for the other variables, the values displayed for the exponentials of the coefficients are more difficult to interpret. Both the ROA using net income and the logarithm of the total assets display odds ratios below one. Larger values for these variables should be indicative of more robust companies. The ROA measures profitability and it was hypothesized that the logarithm of the assets would account for a decreased risk of insolvency in bigger companies. However, these results do not necessarily contradict these assumptions. This simplistic interpretation of the odds ratios only remains valid if the values of the different input variables included in the logistic regression are truly independent of each other. For instance, the ROA is not independent of the Cash flow / Total assets. These dependencies could not be completely eliminated in the multicollinearity analysis and jeopardize the interpretation of the results for these particular variables.

The output of the final logistic regression model may then be computed with the following expression:

$$LR(X) = \frac{1}{1 + e^{-1.452+0.068x_1+0.216x_2-0.120x_3-0.072x_4+1.873x_5-0.003x_6}} \tag{42}$$

This formula considers the following notation:

- X- Input data vector;
- x_1- ROA using net income;
- x_2- ln(Total assets);
- x_3- Cash flow / Total assets;
- x_4- Shareholder equity ratio;
- x_5- Textiles sector;
- x_6- Total assets variation (2015-2016).

The SPSS software does not allow for the direct calculation of the AUC and Gini Index for logistic regression models. Consequently, the results of the LR were saved as a new variable in the worksheet and a receiver operating curve was later generated with this information. Figure 17 displays the ROC curve in blue, along with the reference diagonal in red. The value obtained for the area under the curve was 0.926, which corresponds approximately to a Gini index of 0.852.

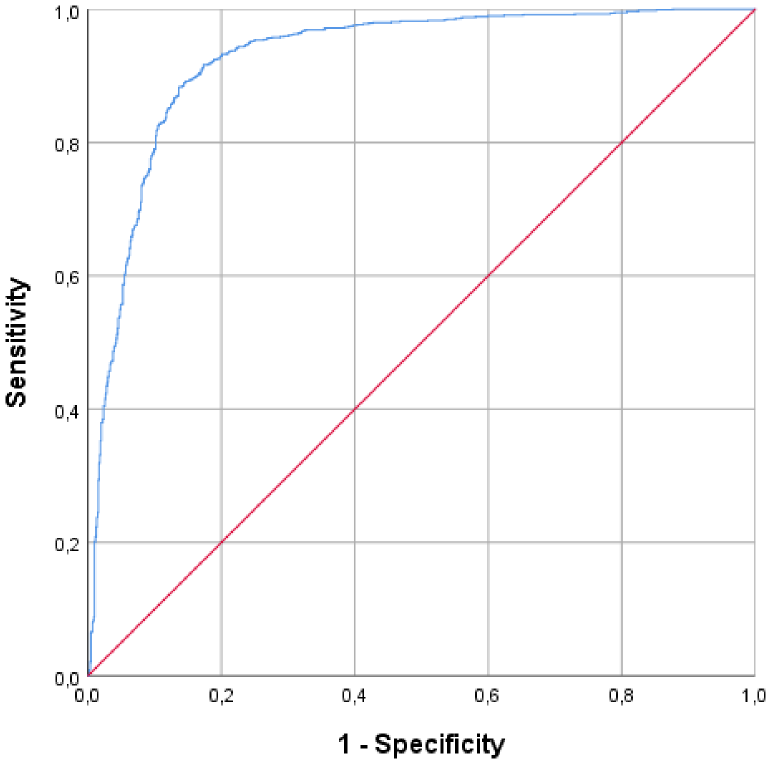


Figure 17 - ROC curve for the output of the logistic regression model.

5.3. MLP Artificial Neural Network

The MLP model was applied in the neural networks' module of SPSS Statistics 25. This software offers the user various options regarding the way the ANNs are structured and the methods through which the results are computed.

First, the partitioning of the sample may be set. This involves specifying the fraction of instances that is allocated to the training, validation and testing datasets. Secondly, the structure of the MLP network may be stipulated in terms of the number of hidden layers, the activation function to be used in these layers and the transfer function of the output layer. Finally, there are different options for the learning algorithm to be employed in the networks' development. Considering these possibilities, four different MLP neural networks are proposed. The characteristics of these ANNs are presented in Table 14.

Table 14 - Features of the MLP networks tested.

Classifiers	MLP 1	MLP 2	MLP 3	MLP 4
Number of hidden layers	1	2	1	2
Number of units per hidden layer	Set automatically	Set automatically	Set automatically	Set automatically
Activation function for the hidden layers	Sigmoid	Sigmoid	Hyperbolic tangent	Hyperbolic tangent
Activation function for the output layer	Identity function	Identity function	Identity function	Identity function
Training algorithm	Scaled conjugate gradient	Scaled conjugate gradient	Scaled conjugate gradient	Scaled conjugate gradient

As the process of developing these models and calculating the respective key performance indicators is computationally demanding, it was opted to restrict the number of MLP models to be tested to four. The implementation of the ANNs is dependent on a time-consuming iterative procedure that is detailed further in this section. Consequently, by considering additional architectures in this phase, there would be a sharp increase in the number of runs needed to evaluate all the alternatives. Nevertheless, the settings of the four MLP models tested were established to ensure, within the limitations present, a diversity of architectures and parameters.

Regarding the partitioning of the data, several combinations were selected, also in accordance with the best practices in the literature. The first combination corresponds to a training-testing-validation ratio of 700:300:0. This is the default setting in SPSS and the most popular partition, being used by numerous authors (e.g. Angelini et al., 2008 and Pacelli & Azzollini, 2010). The second option uses a training-testing-validation ratio of 600:150:250 and is used by Lai et al., 2006. Lastly, the third partitioning, which varies only slightly in relation to the second alternative, corresponds to a ratio of 600:200:200 and is based on the work of Addo et. al, 2018.

For the comparison between methods to be fair, one must be careful when setting the partitioning strategy in SPSS. The percentage of cases that are attributed to each set may be defined directly in the software's user interface for a given network. However, this introduces the potential for chance to influence the results. As the cases are randomly sampled from the dataset to build the training, testing and validation sets, the results obtained will be strongly influenced by this arbitrary selection. By not guaranteeing the replicability of the partition, the comparison between results cannot yield meaningful results.

This issue essentially arises because some companies are more difficult to classify than others. Not all instances present overwhelmingly positive or negative indicators. These cases are the ones that contribute the most to the errors committed by the models. If a given partition randomly samples more of these instances than the others, then the models would tend to display poorer results, but this does not mean that the partitioning strategy is inferior to the others. The same reasoning applies to comparisons between different models that use the same partitioning strategy. A given model may perform better solely because it was evaluated with a test set containing instances that are easier to sort.

In order to overcome this flaw, a strategy is employed that mitigates the potential for the models' results to be influenced by chance. First, three partitioning variables are defined beforehand. These variables contain values that determine the placement of each instance. Positive, null and negative values attribute the cases to the training, testing and validation sets, respectively. The variables' values are generated in accordance with the partitioning strategy desired and used for the testing of all the MLP models for that given strategy. This assures that the networks are comparable if the results were obtained for the same partitioning option, as these models are generated with similar initial conditions.

However, when comparing networks that used different partitioning strategies, which correspond to different auxiliary partitioning variables, there is still the potential for chance to affect the analysis. Therefore, it was deemed necessary to do multiple runs of the algorithm that generates these variables and then compute the average values for the KPIs.

The pseudo-code for the algorithms utilized in the generation of the partitioning variables is presented in Appendix D. Additionally, the KPIs obtained for each specific iteration and partitioning scheme are listed in Appendix E. By averaging out all the performance metrics across the iterations (according to the MLP model and partitioning option considered in each iteration), it was possible to compute the results that are presented in the following table.

Table 15 - Average values of the KPIs after 5 runs for each combination of MLP model and partition.

Partitioning	ANN model	PCC (%)	Sensitivity (%)	Specificity (%)	AUC	Gini Index
700:300:0	MLP 1	88.10	93.68	79.32	0.9404	0.8808
	MLP 2	88.14	93.44	79.92	0.9422	0.8844
	MLP 3	88.32	94.08	79.28	0.9492	0.8984
	MLP 4	89.12	93.74	81.74	0.9508	0.9016
600:150:250	MLP 1	88.02	91.86	81.22	0.9468	0.8936
	MLP 2	88.04	91.98	81.10	0.9392	0.8784
	MLP 3	89.74	93.28	83.28	0.9570	0.9140
	MLP 4	90.74	94.54	84.04	0.9586	0.9172
600:200:200	MLP 1	89.02	93.88	81.00	0.9462	0.8924
	MLP 2	88.52	93.38	80.38	0.9374	0.8748
	MLP 3	89.44	94.44	81.08	0.9496	0.8992
	MLP 4	90.20	94.42	83.14	0.9544	0.9088

Analyzing the values of the KPIs displayed in Table 15, which are all relative to the testing set, it can be understood how each MLP network performs for all the partitioning strategies considered. After comparing the models, it was considered that the best network is MLP 4 trained with 60% of instances in the training set, 15% in the testing set and the remaining 25% in the validation set. The diagram depicting this network is presented in Appendix F.

This ANN displays the best value for the area under the ROC curve, as well as the greatest Gini index. The AUC and Gini index are given priority over the remaining metrics in this evaluation and all the subsequent testing of AI models. Ling, Huang and Zhang (2003) have shown that the AUC is statistically consistent and more discriminating than a simple accuracy measure, thus being better at evaluating the performance of learning algorithms. Nevertheless, in the specific case of MLP 4 under the second partitioning scheme, this option remarkably outperforms the alternative architectures in all the remainder metrics considered.

SPSS also has the option to perform a sensitivity analysis to compute the importance of each independent variable in the MLP artificial neural networks. These estimates are computed by assessing how much the network's prediction value variates in response to changes in the explanatory variables. This procedure is based on the training and testing sets and outputs a chart containing simple and normalized importance estimates. After selecting this option for MLP 4 under the second partitioning scheme, the graph displayed in Figure 18 was obtained.

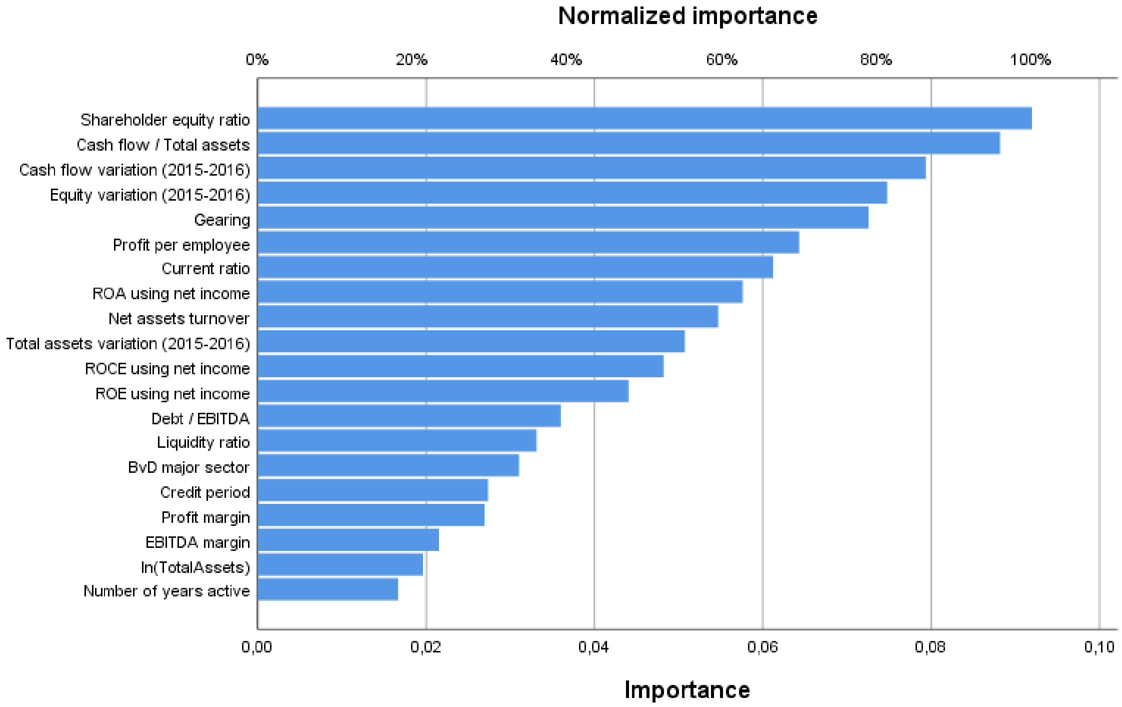


Figure 18 - Importance estimates of the explanatory variables for the MLP model.

Analyzing these results, it becomes evident that the equity of the companies is highly influential on the respective outcomes. The most important indicator is the shareholder equity ratio, a measure of financial autonomy, while the fourth most predictive variable is the variation observed in the companies' equity from 2015 to 2016. Additionally, the profitability indicators, such as Cash flow / Total assets and the profit per employee, are also very important to this credit scoring model.

5.4. Radial Basis Function Artificial Neural Network

The radial basis function ANN model was also applied in the neural networks' module of SPSS Statistics 25. In the same way as the MLP models, there is the option to directly define the percentages that are assigned to the training, validation and testing sets. Additionally, there are two alternatives for the activation function used in the hidden layers, which are ordinary and normalized radial basis functions. The remaining customizable settings are the number of elements in the hidden layers and the overlap among hidden units. The overlapping factor is a multiplier applied to the width of the radial basis functions.

As SPSS offers algorithms that define the optimal number of units in the hidden layers and the best values for the overlapping factors, these features were not set manually. Thus, the software automatically defined the most advantageous architecture regarding these characteristics. Considering that there is no mechanism in place to select the transfer function in the hidden layers that achieves the best results, two alternative RBF networks are studied that differ solely in this aspect. These are detailed in Table 16.

Table 16 - Features of the RBF networks tested.

RBF network	RBF 1	RBF 2
Number of elements in the hidden layers	Set automatically	Set automatically
Overlapping factor	Set automatically	Set automatically
Activation function for the hidden layers	Normalized RBF	Ordinary RBF

The partitioning schemes used in the MLP networks are also employed in the development of the RBF models. Additionally, similarly to what was done in the previous section to mitigate the variability in the results caused by the random sampling procedure used to build the various sets, three partitioning variables are computed and used iteratively to build the networks and collect the KPIs. The results for these performance metrics are found in the table below.

Table 17 - Average values of the KPIs after 5 runs for each combination of RBF model and partition.

Partitioning	ANN model	PCC (%)	Sensitivity (%)	Specificity (%)	AUC	Gini Index
700:300:0	RBF 1	83.54	87.36	77.68	0.8918	0.7836
	RBF 2	81.64	88.02	72.00	0.8842	0.7684
600:150:250	RBF 1	81.02	84.54	75.12	0.8888	0.7776
	RBF 2	81.28	86.76	72.02	0.8924	0.7848
600:200:200	RBF 1	81.88	84.84	76.94	0.8900	0.7800
	RBF 2	82.38	85.94	76.22	0.8910	0.7820

The AUC and Gini Index remain the most robust metrics for the evaluation of these models and are thus prioritized over the others, similarly to what was done in the previous section. Considering this, it may be concluded from the results displayed in Table 17 that RBF model 2 under the second partitioning option (60% Training, 15% Testing and 25% Validation) outperforms the remaining alternatives.

5.5. Random Forest

The random forest method was applied in MATLAB R2018b. This model can be obtained by using the TreeBagger function available in this software, which builds an ensemble of bootstrapped decision trees for either classification or regression purposes. Additionally, TreeBagger selects a random subset of explanatory variables to use at each decision split as is described by Breiman (2001) in the original random forest algorithm. The code used in MATLAB to generate and analyze the random forest is presented in Appendix G.

In order to build the RF model, the dataset is imported into MATLAB as a table type object. The dependent variable is imported subsequently as a column vector, which allows for it to be used as a separate argument in the function. This data allows for the construction of the model, which is set for classification purposes, as the outcome considered is categorical.

The surrogate splits option is activated to handle cases of missing data. In situations with a missing value for the best split, this technique calculates to what extent alternative splits resemble the best split in terms of the cases that are sent down each path in the decision trees (Feelders, 2000). Afterward, the most similar split is used, instead of the original optimal partition.

Additionally, the optional arguments 'OOBPredictorImportance' and 'OOBPrediction' were included in the function to allow for the assessment of the variables' explanatory power and the computation of the predicted class probabilities, respectively. The probabilities are especially important, as these are used in the latter plotting of the ROC curve.

The TreeBagger function offers two possibilities for the algorithm that selects the best split predictor at each node, curvature and interaction-curvature tests. A curvature test selects the split predictor that minimizes the p-value of chi-square tests of independence between each explanatory variable and the dependent variable (Loh & Shih, 1997). An interaction-curvature algorithm chooses the split predictor that minimizes the p-value of chi-square tests of independence between each explanatory variable and the output, while also minimizing the p-value of a chi-square test of independence between each pair of independent variables and the dependent variable (Loh, 2002).

In order to understand which of these algorithms provides the best results, two distinct random forest models are applied differing in the splitting techniques. The relevant KPIs obtained for both models are displayed in the following table.

Table 18 - KPIs for the different splitting algorithm options.

RF splitting algorithm	PCC (%)	Sensitivity (%)	Specificity (%)	AUC	Gini Index
Curvature test	96.46	98.59	94.32	0.997	0.994
Interaction-curvature test	96.61	98.49	94.73	0.996	0.992

The random forest using the curvature tests provided the best predictions in terms of AUC, Gini Index and sensitivity. Although the percentage of correctly classified cases is slightly inferior to the one presented by the model trained with the interaction-curvature tests, a higher AUC is prioritized.

A critical parameter that must be defined prior to the implementation of these models is the number of decision trees contained in the ensembles. The results displayed so far were obtained with models composed of 50 decision trees. However, it must be analyzed if there are gains to be had by adding more trees or, on the other hand, there is an excess of DTs that does not translate into a reduction of the prediction error and increases the computation time unnecessarily. In order to do this, the out-of-bag prediction error is plotted for a variable number of decision trees in the graph below.

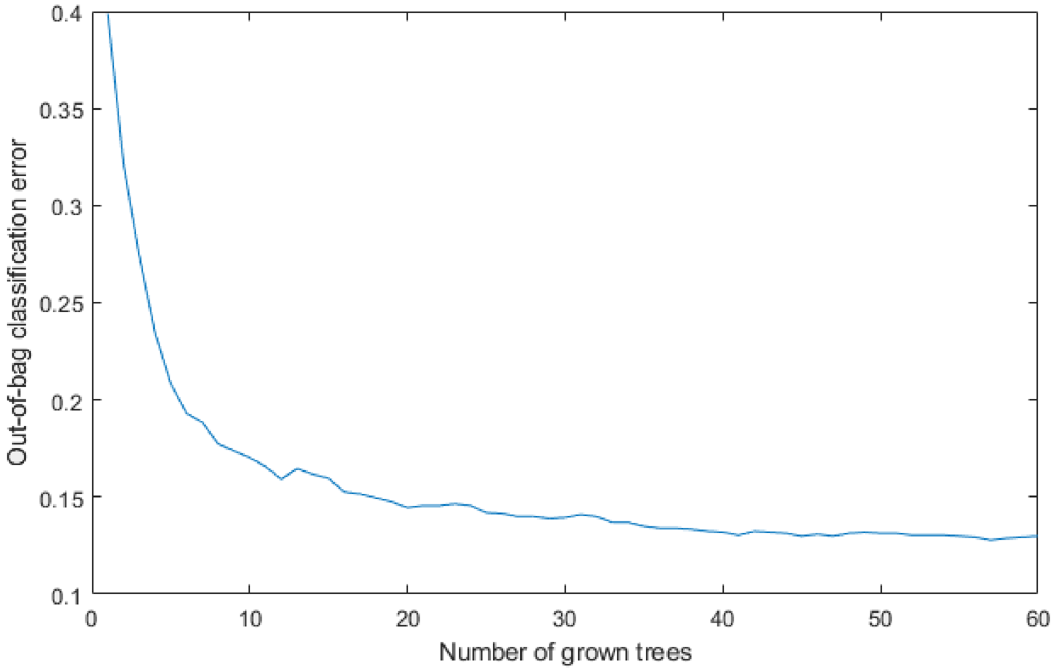


Figure 19 - Out-of-bag prediction error obtained for a variable number of decision trees.

Analyzing Figure 19, one can observe that, when the total number of grown trees is small, there is a rapid decrease of the out-of-bag prediction error with additional DTs in the ensemble. However, these gains in accuracy are progressively smaller, which causes the out-of-bag prediction error to stabilize around an ensemble of 50 trees. The error rate observed for a RF containing 50 decision trees is 0.1314, whereas an ensemble of 60 DTs displays a rate of 0.1299. As this reduction is hardly significant, it was opted to keep the number of decision trees at 50.

Using the data stored in 'OOBPredictorImportance', one can obtain the importance estimates for the explanatory variables. Analyzing these results, it is relevant to point out that the variable with the most explanatory power is the shareholder equity ratio, which displays a remarkable score in comparison with the other indicators. The credit period and the Cash flow / Total assets indicators display the second and third highest importance estimates, respectively. Certain measures, namely the profit per employee and gearing, are also important to the results of the model. Other variables, such as the major sector of activity, some temporal trends and company maturity, prove to be reasonably exploratory of the companies' risk profiles, but to a lesser degree than the ones that were previously mentioned.

5.6. Chapter Conclusions

In this chapter, several models were tested for each of the credit scoring methods considered. Subsequently, the best ones were selected to represent the respective techniques in the latter benchmarking.

The statistical models were implemented in SPSS 25. The first predictor to be employed was the discriminant analysis. Numerous stepwise methods were applied in the selection of inputs. Additionally, the parameters used as thresholds for the entry and removal of variables were tuned to increase the diversity of the models and thus increase the chances of finding a good alternative. Secondly, the logistic regression model was applied to the data. As before, various stepwise techniques were tested along with different thresholds. In this case, it was opted to use the default settings for the limits conditioning the entry and removal of indicators, as these produced the best results.

The artificial neural networks were also implemented in SPSS 25. Several architectures were tried out for the multilayer perceptron and radial basis function models. These alternative models differed in the number of hidden layers, the activation functions and the learning algorithms employed. Besides the distinct structures, three separate partitioning schemes were considered. Additional measures were taken to mitigate the detrimental effects of random sampling procedures. These consisted in the use of partitioning variables and iterative techniques for the computation of the relevant KPIs.

Finally, the random forest model was developed using MATLAB R2018b. The function TreeBagger that is available in this software package allows for the construction of ensembles of bootstrapped decision trees that follow the RF algorithm. As the data still contains various instances with missing data, the surrogate splits option was activated to handle such cases. In terms of the settings tested, two different splitting procedures were applied, curvature and interaction-curvature tests. Most importantly, there was also an analysis of the out-of-bag prediction error for a range of possible numbers of decision trees to be included in the random forest. It was concluded that this error stabilized around an ensemble of 50 trees.

6. Comparing AI and statistical methods

6.1. Development Process

Considering the results presented in the previous sections, several observations may be made about the final AI and statistical models obtained. Firstly, it is relevant to point out that the indicators that were chosen as the most explicative in the variable selection processes are largely the same across all the models. For instance, most of the variables designated to be used as inputs in the logistic regression are also the ones displaying the highest predictive power in the linear discriminant analysis. Furthermore, these same variables are also among the most predictive indicators of the MLP artificial neural network and random forest methods. This uniformity is a strong argument in favor of the selection procedures in place, demonstrating that the following variables are the most predictive of the companies' outcomes:

- Shareholder equity ratio;
- Cash flow / Total assets;
- Credit period;
- ROA using net income;
- BvD major sector;
- Gearing.

All the variables above were found to be highly indicative of the companies' risk profiles in at least two distinct credit scoring methods. Nevertheless, the remaining indicators were still relevant for credit risk analysis purposes, although not to the extent of the previously mentioned ones. It is also noteworthy that the shareholder equity ratio was the most important independent variable in the majority of the models implemented.

The statistical methods proved more difficult to implement overall, as some of the limitations of these models warranted additional pre-processing procedures. The vulnerability to the presence of outliers and the potential impact of multicollinearity had to be addressed before proceeding with the development of these models, as skipping these steps would risk undermining the respective performances. Nevertheless, the AI methods were more demanding in other aspects, namely in terms of defining the parameters necessary for the learning processes and the experimentation of different architectures.

6.2. Benchmarking the models

By compiling the results obtained so far in terms of the relevant KPIs, it is now possible to compare the credit scoring approaches. For each category of predictive methods, the best model in the developmental stage is considered for benchmarking purposes. Table 19 exhibits the values for the performance metrics, as well as a ranking (from best to worst performing) based on the AUC and Gini Index displayed. This pair of KPIs is again prioritized over the remaining measures, as these demonstrate increased robustness to potential distortions brought on by sample imbalances that could not be addressed in the sampling process.

Table 19 - KPIs for all the credit scoring models implemented.

Model	PCC (%)	Sensitivity (%)	Specificity (%)	AUC	Gini Index	Rank
Discriminant analysis	80.00	88.90	67.70	0.8630	0.7260	5
Logistic regression	89.90	93.80	83.50	0.9260	0.8520	3
MLP neural network	90.74	94.54	84.04	0.9586	0.9172	2
RBF neural network	81.28	86.76	72.02	0.8924	0.7848	4
Random forest	96.46	98.59	94.32	0.9970	0.9940	1

Analyzing Table 19, it can be observed that the random forest model is ranked as the best credit scoring model, displaying the highest AUC and Gini Index, while also presenting a remarkable overall accuracy. Over 95% of all instances are assigned correct predictions, with 98.59% of all future good companies being classified as such. In second place, the MLP neural network displayed robust KPIs, although not up to par with the ones obtained with the random forest. On the other hand, the RBF neural network was the overall worst AI model considered, being even outranked by the logistic regression model.

Regarding the statistical methods, the results fall in line with what was observed in other benchmarking studies. The discriminant analysis proved to be the least predictive model of all the credit scoring methods tested, which may be a result of the violation of this models' assumptions in terms of normality and mutual independence regarding the explanatory variables. The logistic regression is ranked as the third best predictor, behind the MLP neural network and the random forest. This model provides accurate predictions in almost 90% of the cases and demonstrates good sensitivity and specificity, which translate into low rates of type I and type II errors. Despite this, the LR model fell short on the more robust KPIs, namely the AUC and Gini Index, which caused it to be ranked behind some of the AI models.

Considering these results, it can be concluded that the MLP neural network and the random forest outperformed the statistical approaches in the credit scoring experiment. However, the logistic regression proved to be a quality predictor, displaying a high level of accuracy and presenting values for other performance measures that come close to the results of the AI alternatives. This is coherent with the recent rise in popularity of the LR method, which is a solid compromise in terms of prediction performance and ease of implementation. Furthermore, the logistic regression also permits an intuitive interpretation of the model's parameters, overcoming the black-box syndrome of AI predictors.

7. Conclusions and Further Work

7.1. Further Work

In order to delineate the direction of further research based on this work, it is important to state some issues that could be addressed to reach new findings and further enrich the academic literature on credit scoring.

Regarding the pre-processing of the input dataset, several measures were taken to assure the quality of the data, which necessarily impacts the performance of the predictive models. However, posterior studies may adopt distinct methodological approaches to address some limitations of the current research. Specifically, the detection of the multivariate outliers could be improved in terms of the rule utilized in the labeling of these instances.

As the variables in the input data failed the normality tests, it was not possible to proceed with the typical criterium of labeling as outliers any observations with robust Mahalanobis distances beyond a given quantile of the chi-squared distribution. The detection of the multivariate outliers relied then upon the visual examination of the scatterplots with the robust MDs for each observation in the dataset. Consequently, the labeling process lacked objectivity. Therefore, it would be beneficial to develop a more sophisticated outlier labeling rule that is applicable to multivariate non-normal data.

Additionally, there could be an effort to procure additional cases to include in the dataset. The number of instances in the sample was constrained by the limited observations of bad companies and the necessity to preserve class balance. This could be achieved by searching for new cases of non-compliant businesses, insolvencies and special revitalization processes. As the current research considered only company outcomes for the year of 2016, it would be advantageous to study the possibility to include observations for other years. This was done to a limited extent to validate some of the predictors, but such instances were never included during the development of the credit scoring models.

Further research could also attempt to mitigate the detrimental effects of the missing values in the dataset. Some of the predictor methods applied in this study simply discard such cases, which reduces the size of the sample utilized. In order to deal with this situation in the context of the computation of the robust MDs, an imputation procedure was put in place based on a Markov chain Monte Carlo technique. However, the imputed dataset could not be used in the development of some of the models, which limited the applicability of this sample to the pre-processing stage of this project. Thus, additional studies could attempt to employ multiple imputation procedures that are compatible with the implementation of the credit scoring methods.

7.2. Conclusions

Credit risk remains one of the biggest risks for financial institutions and corporations alike. The methods utilized in this field have increased in sophistication considerably throughout the years. Nevertheless, any improvements in the accuracy rates of the current models are extremely important, as even small advances can mean significant savings for creditors by preventing defaults. This served as the main motivation for the current study.

This research allowed for the comparison of statistical and AI predictors, adding significantly to the academic literature by designing a credit scoring experiment using a novel dataset with financial and other relevant data for a selection of Portuguese companies. Credit scoring methods were successfully implemented based on this information and used to distinguish between good and bad applicants in the timespan of a year. The statistical methods considered were the discriminant analysis and logistic regression. As for the artificial intelligence models, this study focused on the MLP and RBF artificial neural networks and the random forest method.

As the statistical predictors are particularly susceptible to multicollinearity in the data and to the presence of outlier instances, there was a thorough pre-processing of the dataset prior to the implementation of the models. This procedure included a correlation analysis to remove certain indicators that displayed high VIF values, which corresponded necessarily to the ones displaying the highest dependencies upon the remaining independent variables. Regarding the outlier issue, there was a detection technique in place based on robust Mahalanobis distances that allowed for the identification of certain aberrant instances in the multivariate space. Additionally, a proper sampling technique was defined in order to build a balanced dataset, as the base data was extremely unbalanced. However, this issue could not be fully addressed due to the missing instances being more prevalent in a class than in the other.

After the careful preparation of the dataset, it was possible to implement the credit scoring methods mentioned above. In order to obtain the best performing models in each category of predictors, several alternatives were employed and subsequently compared. The performance evaluation was based on numerous metrics, which included the percentage of correctly classified instances, sensitivity, specificity, the area under the ROC curve and the Gini Index. The diversity of the statistical models was ensured by experimenting with several distinct stepwise techniques and thresholds for the entry and removal of variables. Regarding the artificial intelligence methods, numerous architectures were tested by manually setting some of the structural parameters.

Following the selection of the best models for each category of credit scoring methods, it was possible to compare the KPIs of the statistical and AI alternatives. The benchmarking study completed found that the artificial intelligence methods outperformed the more conventional statistical approaches. The random forest model demonstrated the most potential, followed by the MLP neural network. The RBF neural network and the logistic regression were considered to be the fourth and third most adequate models respectively, whereas the discriminant analysis was the worst-performing model overall.

Regarding the statistical approaches, the results are coherent with the findings of previously published benchmarking research articles. The discriminant analysis is dependent on strict assumptions in terms of normality and mutual independence regarding the input variables, which was a contributing factor to its disuse among credit risk professionals and may explain the poor performance obtained in this experiment. The logistic regression proved to be a quality predictor, displaying a high level of accuracy and presenting values for other performance measures that come close to the results of the AI alternatives. This is consistent with the recent rise in popularity of the LR method, which demonstrated to be a solid compromise in terms of prediction performance and ease of implementation.

The random forest models, along with the MLP artificial neural networks, display tremendous potential in the credit scoring field. In contrast with the statistical techniques, these methods can model hidden non-linear relationships between the explanatory variables and the dependent variable, being also more robust to multicollinearity and to the presence of outliers. Besides these advantages, these methods do not make assumptions regarding the probability distributions of the input data. These factors may have contributed to the observed superiority of the AI approaches. The major drawback of these alternatives continues to be the black-box syndrome, which makes the interpretation of the results almost impossible. This may restrict the use of such models in certain settings due to regulatory requirements.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2–3), 59–88. https://doi.org/10.1002/isaf.325

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, 6(2):38. https://doi.org/10.2139/ssrn.3155047

Aftab, K., Hartley, R., & Trumpf, J. (2015). Generalized Weiszfeld Algorithms for Lq Optimization, 37(4), 728-745. 10.1109/TPAMI.2014.2353625.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5), 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

Aguilera, A., Escabias, M., & Valderrama, M. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8), 1905-1924. https://doi.org/10.1016/j.csda.2005.03.011

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*. The Journal of Finance*, 23(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E., & Sabato, G. (2008). Modelling Credit Risk for SMEs: Evidence from the U.S. Market. *Abacus*, 43 (3), 332-357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation* (3rd ed.). Oxford, UK: Oxford University Press.

Andrieu, G., Staglianò, R., & van der Zwan, P. (2018). Bank debt and trade credit for SMEs in Europe: firm-, industry-, and country-level determinants. *Small Business Economics*, 51(1), 245-264. https://doi.org/10.1007/s11187-017-9926-y

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *Quarterly Review of Economics and Finance*, 48(4), 733–755. https://doi.org/10.1016/j.qref.2007.04.001

Anton Semechko (2019). Detect outliers in multivariate datasets. (https://www.github.com/AntonSemechko/Multivariate-Outliers), GitHub. Retrieved September 24, 2019. [2]

Archer, K., & Kimes, R. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249-2260. https://doi.org/10.1016/j.csda.2007.08.015

[2] The author kindly asks to be referenced in this format

Ayala, H., & Coelho, L. (2016). Cascaded evolutionary algorithm for nonlinear system identification based on correlation functions and radial basis functions neural networks. *Mechanical Systems and Signal Processing*, 68–69, 378–393. https://doi.org/10.1016/j.ymssp.2015.05.022

Baesens, B., Setiono, R., Mues, C., Vanthienen, J. (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science*, 49(3), 312-329. https://doi.org/10.1287/mnsc.49.3.312.12739

Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). Chichester, UK: Wiley

Batista, A. (2012). *Credit Scoring – Uma ferramenta de gestão financeira*. Porto, Portugal: Vida Económica.

Beliakov, G., Kelarev, A., & Yearwood, J. (2011). Robust artificial neural networks and outlier detection. *Technical report*. 10.1080/02331934.2012.674946

Bradford, J., Kunz, C., Brunk, C., & Brodley, C. (1998). Pruning Decision Trees with Misclassification Costs. *ECML '98 Proceedings of the 10th European Conference on Machine Learning*, 131-136. https://link.springer.com/chapter/10.1007/BFb0026682

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), 123-140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees* (1st ed.). Boca Raton, USA: Chapman and Hall/CRC.

Brereton, R., & Lloyd (2016). Re-evaluating the role of the Mahalanobis distance measure. Journal of Chemometrics, 30(4), 134-143. https://doi.org/10.1002/cem.2779

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6), 1291-1302. https://doi.org/10.1016/S0031-3203(02)00121-8

Chatterjee, S., & Price, B. (1991). *Regression analysis by example* (2nd ed.). New York, United States of America: Wiley.

Chen, X. & Wang, D. & Liu, Z. & Wu, Y. (2018). A Fast Direct Position Determination for Multiple Sources Based on Radial Basis Function Neural Network. *10th International Conference on Communication Software and Networks (ICCSN)*, 381-385. http://dx.doi.org/10.1109/CONTROLO.2018.8439740

Chih-Wei Hsu, Chih-Chung Chang, & Lin, C.-J. (2008). A Practical Guide to Support Vector Classification. *BJU International*, 101(1), 1396–1400. https://doi.org/10.1177/02632760022050997

Craney, T., & Surles, J. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14(3), 391-403. https://doi.org/10.1081/QEN-120001878

Crone, S., & Finlay, F. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1), 224-238. https://doi.org/10.1016/j.ijforecast.2011.07.006

Dietterich, T. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2), 139-157. https://doi.org/10.1023/A:1007607513941

Dong, Y., Peng, C. (2013). Principled missing data methods for researchers. *SpringerPlus 2013*, 2:222. https://doi.org/10.1186/2193-1801-2-222

Eidenmuller, H. (2018). What is an insolvency proceeding? *American Bankruptcy Law Journal*, 92(1), 53-71. http://dx.doi.org/10.2139/ssrn.2712628

Espahbodi, H., & Espahbodi, P. (2003). Binary choice models and corporate takeover. *Journal of Banking and Finance*, 27(4), 549-574.https://doi.org/10.1016/S0378-4266(01)00258-8

Fabbri, D., & Menichini, A. (2010). Trade credit, collateral liquidation and borrowing constraints. *Journal of Financial Economics*, 96(3), 413-432. https://doi.org/10.1016/j.jfineco.2010.02.010

Feelders, A. (2000). Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? *Lecture Notes on Computer Science*, 1704. 10.1007/978-3-540-48247-5_38

Filzmoser, P. (2004). A multivariate outlier detection method. *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, 1, 18-22.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368-378. https://doi.org/10.1016/j.ejor.2010.09.029

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fletcher, P., Venkatasubramanian, S., & Joshi, S. (2008). *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 10.1109/CVPR.2008.4587747

Frank, E. (2000). Pruning Decision Trees and Lists. Doctorate Thesis developed in the University of Waikato.

Fritz, H., Filzmoser, P., & Croux, C. (2012). A comparison of algorithms for the multivariate L1-median. *Computational Statistics*, 27(3), 393-410. https://doi.org/10.1007/s00180-011-0262-4

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447

Gouvêa, A., & Bacconi, E. (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. *POMS 18th Annual Conference, Dallas, Texas, USA*, 4-7 https://doi.org/10.1.1.626.8493

Grubbs, F. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics* ,11(1), 1-21. 10.1080/00401706.1969.10490657

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5

Hand, D., & Anagnostopoulos (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34(5), 492-495. https://doi.org/10.1007/s10994-009-5119-5

Harel, O., & Zhou, X. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16), 3957-3077. https://doi.org/10.1002/sim.2787

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, USA: Springer-Verlag New York.

Henry, E., Robinson, T., & van Greuning, J. (2011). *CFA Program Curriculum 2012 Level I*. Boston, USA: Pearson

Hill, M., Kelly, G., Preve, L., & Sarria-Allende (2017). Trade Credit or Financial Credit? An International Study of the Choice and Its Influences. *Emerging Markets Finance & Trade*, 53(10), 2318-2332. https://doi.org/10.1080/1540496X.2017.1319355

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), 85-126. https://doi.org/10.1007/s10462-004-4304-y

Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. New York, United States of America: Wiley.

Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research*, 52, 317–324. https://doi.org/10.1016/j.cogsys.2018.07.023

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37(4), 543–558. https://doi.org/10.1016/S0167-9236(03)00086-1

Hubert, M., & van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45(2), 301-320. https://doi.org/10.1016/S0167-9473(02)00299-2

IBM. (2017). IBM SPSS Statistics V25.0 documentation. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_25.0.0/statistics_kc_ddita/spss/product_landing.html

Jacobson, T., von Schedvin, E. (2015). Trade credit and the propagation of corporate failure: An empirical analysis. *Econometrica*, 83(4), 1315–1371. https://doi.org/10.3982/ECTA12148

Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking and Finance*, 56, 72–85. https://doi.org/10.1016/j.jbankfin.2015.02.006

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2), 119-127. https://doi.org/10.2307/2986296

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), 6233–6239. https://doi.org/10.1016/j.eswa.2010.02.101

Khemakhem, S., & Boujelbènea, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach. *Accounting and Management Information Systems*, 14(1), 60–78. http://www.cig.ase.ro/repec/ami/articles/14_1_3.pdf

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102, 207–217. https://doi.org/10.1016/j.eswa.2018.02.029

Lai, K., Yu, L., Wang, S., & Zhou, L. (2006). Credit risk analysis using a reliability-based neural network ensemble model. *Artificial Neural Networks – ICANN 2006*, 682–690. https://doi.org/10.1007/11840930_71

Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Leys, C., Klein, O., Dominicy, Y., & Ley, C (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, 74, 150-156. https://doi.org/10.1016/j.jesp.2017.09.011

Ling, C., Huang, J., & Zhang, H. (2003). AUC: a statistically consistent and more discriminating measure than accuracy. *Proceedings of the 18th international joint conference on Artificial intelligence*, 519-524. https://cling.csd.uwo.ca/papers/ijcai03.pdf

Liu, C. (1995). Missing Data Imputation Using the Multivariate t Distribution. *Journal of Multivariate Analysis*, 53(1), 139-158. https://doi.org/10.1006/jmva.1995.1029

Liu, Y., & De, A. (2015). Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study. *International Journal of Statistics in Medical Research*, 4(3), 287-295. 10.6000/1929-6029.2015.04.03.7

Lo, A. (1985). Logit Versus Discriminant Analysis. *Journal of Econometrics*, 31 (3), 151-178. https://doi.org/10.1016/0304-4076(86)90046-1

Loh, W. (2002). Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, 12(2), 361-386. https://doi.org/10.1002/sim.6454

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7(4), 815-840. https://www.jstor.org/stable/24306157

MathWorks. (2019). MATLAB Documentation. Retrieved from https://www.mathworks.com/help/matlab/index.html

Midi, H., & Bagheri, A. (2010). Robust multicollinearity diagnostic measure in collinear data set. *4th international conference on applied mathematics*, simulation, modeling, 138–142.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109-131. doi:10.2307/2490395

Ong, C., Huang, J., & Tzeng, G. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, 29(1), 41-47. https://doi.org/10.1016/j.eswa.2005.01.003

Pacelli, V., & Azzollini, M. (2011). An Artificial Neural Network Approach for Credit Risk Management. *Journal of Intelligent Learning Systems and Applications*, 03(02), 103–112. https://doi.org/10.4236/jilsa.2011.32012

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, 201(2), 490-499. https://doi.org/10.1016/j.ejor.2009.03.008

Petersen, M., & Rajan, R. (1997). Trade Credit: Theory and Evidence. *Review of Financial Studies*, 10(3), 661-691. https://doi.org/10.1093/rfs/10.3.661

Piasecki, K., & Wójcicka-Wójtowicz, A. (2017). Capacity of Neural Networks and Discriminant Analysis in Classifying Potential Debtors. *Folia Oeconomica Stetinensia*, 17 (2), 129-143. 10.1515/foli-2017-0023

Press, S., & Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364), 699-705. 10.1080/01621459.1978.10480080

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, USA: Morgan Kaufmann Publishers.

Reed, R. D., & Marks, R. J. (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Cambridge, USA: A Bradford Book.

Ríha, J. (2016). Artificial Intelligence Approach to Credit Risk. Master thesis developed in Charles University Prague - Institute of Economic Studies.

Ryan, T. (1997). *Modern Regression Methods*. New York, United States of America: Wiley.

Sustersic, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications*, 36(3), 4736–4744. https://doi.org/10.1016/j.eswa.2008.06.016

Swets, J., Dawes, R., & Monahan, J. (2000). Better decisions through science. *Scientific American*, 283(4), 82–87. 10.1038/scientificamerican1000-82

Tang, Y., Ji, J., Gao, S., Dai, H., Yu, Y., & Todo, Y. (2018). A Pruning Neural Network Model in Credit Classification Analysis. *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 9390410, 22 pages, 2018. https://doi.org/10.1155/2018/9390410

Thompson, C., Kim, R., Aloe, A., & Becker, B. (2017). Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. *Basic and Applied Social Psychology*, 39(2), 81-90. https://doi.org/10.1080/01973533.2016.1277529

Timofeev, R. (2004). Classification and Regression Trees (CART) Theory and Applications. Master Thesis developed in Humboldt University - Center of Applied Statistics and Economics.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4), 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixes-model analysis. *Journal of Clinical Epidemiology*, 66(9), 1022-1028. 10.1016/j.jclinepi.2013.03.017

Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton, United States of America: Chapman and Hall/CRC.

Van Diepen, M., & Franses, P. (2006). Evaluating chi-squared automatic interaction detection. *Information Systems*, 31(8), 814-831. https://doi.org/10.1016/j.is.2005.03.002

Vellido, A., Lisboa, P. J. G. & Vaughan, J. (1999). Neural networks in business: A survey of applications (1992-1998). *Expert Systems with Applications*, 17(1), 51-70. http://doi.org/10.1016/s0957-4174(99)00016-0

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, 27(11–12), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

White, I., Carlin, J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28), 2920-2931. 10.1002/sim.3944

Wójcicka, A. (2017). Neural Networks in Credit Risk Classification of Companies in the Construction Sector. *Econometric Research in Finance*, 2(2), 63–77. http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171483796

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications*, 42(7), 3508–3516. https://doi.org/10.1016/j.eswa.2014.12.006

Appendices

Appendix A – Descriptive statistics and instance counts for the sample.

	Mean		Std. Deviation		Valid Instances Count	
	Good	Bad	Good	Bad	Good	Bad
ROE using net income before tax	2.03	-23.44	42.20	97.35	988	940
ROCE using net income before tax	4.39	-21.89	50.13	87.76	988	735
ROA using net income before tax	1.97	-11.99	11.35	23.06	988	952
Profit margin	3.18	-12.33	16.11	25.57	988	883
EBITDA margin	9.89	-8.00	18.23	28.23	988	886
Net assets turnover	2.22	3.09	4.01	11.86	988	884
Credit period	59.51	137.86	87.83	174.42	987	883
Current ratio	4.30	1.90	7.98	5.75	988	984
Debt / EBITDA	3.21	-398.60	17.05	7215.87	815	699
Liquidity ratio	4.44	3.70	21.45	63.68	988	973
Gearing	0.69	1.80	2.22	39.08	984	962
ln(Total assets)	13.19	12.81	1.34	2.28	988	984
Cash Flow / Total Assets	6.00	-59.49	14.00	240.99	988	949
Number of years active	24.21	21.22	13.93	15.01	988	989
Equity variation 2015-2016	4.31	56.06	50.69	4574.67	988	933
Total assets variation 2015-2016	4.72	68.78	37.31	1602.15	988	931
Shareholder equity ratio	48.62	-4.86	30.15	6510.26	988	984
Profit per employee	6.94	-8.64	49.30	25.17	934	753
Cash Flow variation 15-16	-36.46	420.41	1735.07	15650.14	972	863

	Mean				Std. Deviation				Valid Instances Count			
	Active	Insolvent	SPR	Non-compliant	Active	Insolvent	SPR	Non-compliant	Active	Insolvent	SPR	Non-compliant
ROE using net income before tax	2.03	-22.27	-27.46	-5.40	42.20	99.82	93.18	43.36	988	663	262	15
ROCE using net income before tax	4.39	-27.54	-11.09	-5.42	50.13	100.35	55.05	19.68	988	488	231	16
ROA using net income before tax	1.97	-12.73	-10.21	-11.00	11.35	24.06	20.14	24.33	988	663	262	27
Profit margin	3.18	-12.56	-12.46	-3.82	16.11	26.17	24.72	14.85	988	611	252	20
EBITDA margin	9.89	-8.36	-7.82	0.18	18.23	28.61	27.39	27.05	988	612	252	22
Net assets turnover	2.22	3.32	2.26	6.49	4.01	11.91	11.37	15.73	988	612	252	20
Credit period	59.51	131.46	150.61	169.13	87.83	176.49	164.47	219.87	987	609	252	22
Current ratio	4.30	1.93	1.80	2.13	7.98	6.42	3.52	5.58	988	692	265	27
Debt / EBITDA	3.21	-214.57	-755.05	-	17.05	3197.80	11545.95	-	815	461	238	0
Liquidity ratio	4.44	4.78	1.11	1.93	21.45	76.03	1.74	5.60	988	682	264	27
Gearing	0.69	1.29	0.09	32.41	2.22	17.50	40.41	179.87	984	673	263	26
ln(Total assets)	13.19	12.23	14.25	13.41	1.34	2.17	1.88	2.06	988	692	265	27
Cash Flow / Total Assets	6.00	-79.00	-17.28	-0.08	14.00	282.00	92.23	26.00	988	660	262	27
Number of years active	24.21	19.92	24.72	20.41	13.93	14.30	16.40	13.653	988	697	265	27
Equity variation 2015-2016	4.31	-61.87	353.33	-	50.69	3054.78	7083.81	-	988	668	265	0
Total assets variation 2015-2016	4.72	98.35	-5.53	-	37.31	1893.71	37.15	-	988	666	265	0
Shareholder equity ratio	48.62	-671.51	-47.24	-38.72	30.15	7756.66	167.94	168.66	988	692	265	27
Profit per employee	6.94	-8.00	-10.83	-3.73	49.30	22.47	31.97	8.69	934	523	206	24
Cash Flow variation 15-16	-36.46	-91.45	1751.94	-708.41	1735.07	6269.47	27504.37	4547.75	972	590	248	25

Appendix B – Correlation matrix for the dataset after the multicollinearity analysis.

Pearson correlation coefficients matrix

	ROE using Net income	ROCE using Net income	ROA using Net income	Profit margin	EBITDA margin	Net assets turnover	Credit period days	Current ratio	Debt / EBITDA	Liquidity ratio	Gearing	ln(Total Assets)	Cash flow / Total assets	Number of years active	Equity variation 15-16	Total assets variation 15-16	Equity / Total Assets	Profit per employee	Cash flow variation 15-16
ROE using Net income	1																		
ROCE using Net income	0.280**	1																	
ROA using Net income	0.276**	0.377**	1																
Profit margin	0.277**	0.252**	0.525**	1															
EBITDA margin	0.233**	0.258**	0.531**	0.817**	1														
Net assets turnover	0.006	-0.169**	-0.004	-0.002	-0.016	1													
Credit period days	-0.087**	-0.090**	-0.181**	-0.217**	-0.230**	-0.017	1												
Current ratio	0.047*	0.038	0.126**	0.100**	0.116**	-0.029	-0.140**	1											
Debt / EBITDA	-0.005	-0.003	-0.009	0.002	0.021	0.001	-0.006	0.014	1										
Liquidity ratio	0.011	0.017	0.023	0.058*	0.073**	-0.008	-0.040	0.110**	0.003	1									
Gearing	-0.129**	-0.011	-0.010	-0.052*	-0.030	-0.005	0.022	-0.010	0.002	-0.002	1								
ln(Total Assets)	-0.062**	-0.013	0.088**	0.068**	0.154**	-0.031	0.073**	-0.004	-0.085**	-0.001	0.043	1							
Cash flow / Total assets	0.026	0.113**	0.073**	0.182**	0.197**	0.001	-0.037	0.058*	-0.005	0.012	0.008	0.300**	1						
Number of years active	-0.029	-0.021	0.037	-0.002	-0.026	-0.004	0.036	0.030	0.026	0.008	0.045*	0.299**	0.092**	1					
Equity variation 15-16	0.003	0.003	-0.026	0.076**	0.062**	0.001	-0.024	0.002	-0.003	0.000	0.000	0.027	-0.012	0.002	1				
Total assets variation 15-16	-0.138**	-0.003	-0.002	-0.081**	-0.063**	-0.005	0.001	-0.011	0.002	-0.003	0.037	-0.021	-0.002	-0.052*	-0.217**	1			
Equity / Total Assets	-0.006	-0.001	-0.002	0.057*	0.037	0.006	-0.010	0.024	-0.000	0.005	0.004	0.177**	0.251**	-0.002	-0.003	0.003	1		
Profit per employee	0.043	0.032	0.055*	0.155**	0.188**	-0.005	-0.039	0.008	0.005	0.002	0.030	0.155**	0.028	0.002	0.003	-0.002	0.001	1	
Cash flow variation 15-16	-0.001	0.006	0.000	0.100**	0.089**	-0.003	-0.027	-0.004	0.001	-0.001	0.033	0.044	-0.037	-0.012	0.119**	0.000	-0.000	0.007	1

* Correlation is significant at the 0.05 level (two tailed).

** Correlation is significant at the 0.01 level (two tailed).

Appendix C – Histograms for the explanatory variables (discriminated by class).

The following histograms detail the distributions of the observed values for each potential explanatory variable. The plots are divided into two parts. The distributions of the values for the good companies are displayed on the left and the ones for the bad companies on the right. In order to facilitate the interpretation of the graphs, the width of each column corresponds to half of the intervals seen in the vertical axes. The minimum and maximum values displayed in the axes were adjusted to provide a better view of the data. Consequently, some univariate extreme instances may not be represented.

These plots were generated for a balanced sample. However, there are lapses in the data which affect the instances of bad companies to a greater degree than the other ones. This may lead to different total counts of good and bad companies in the same plot. It is advised not to compare absolute values between categories due to these discrepancies.

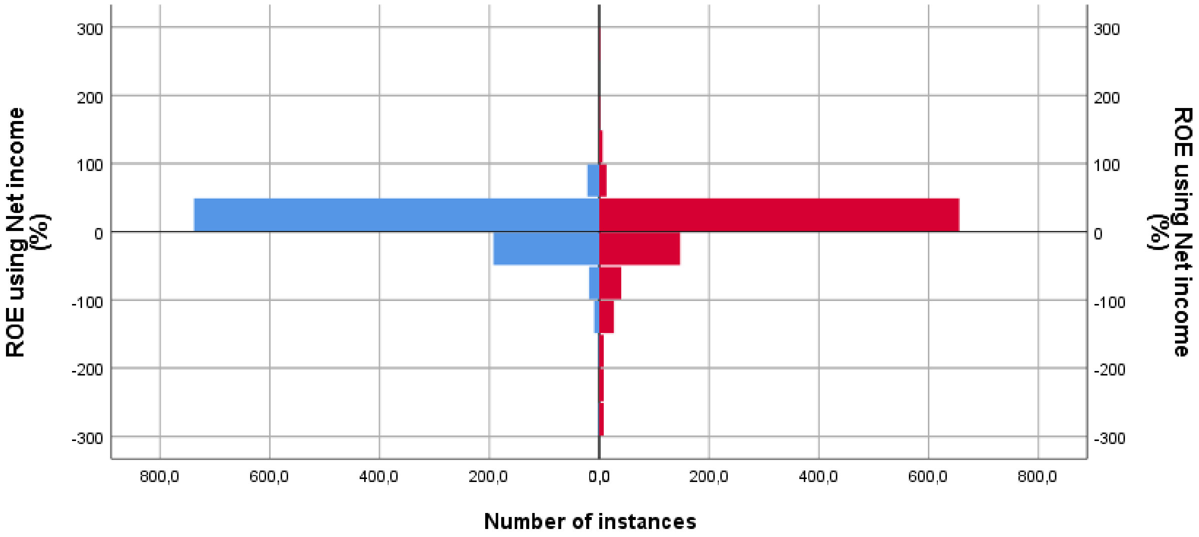


Figure 20 - Histograms of the ROE using net income for good and bad companies.

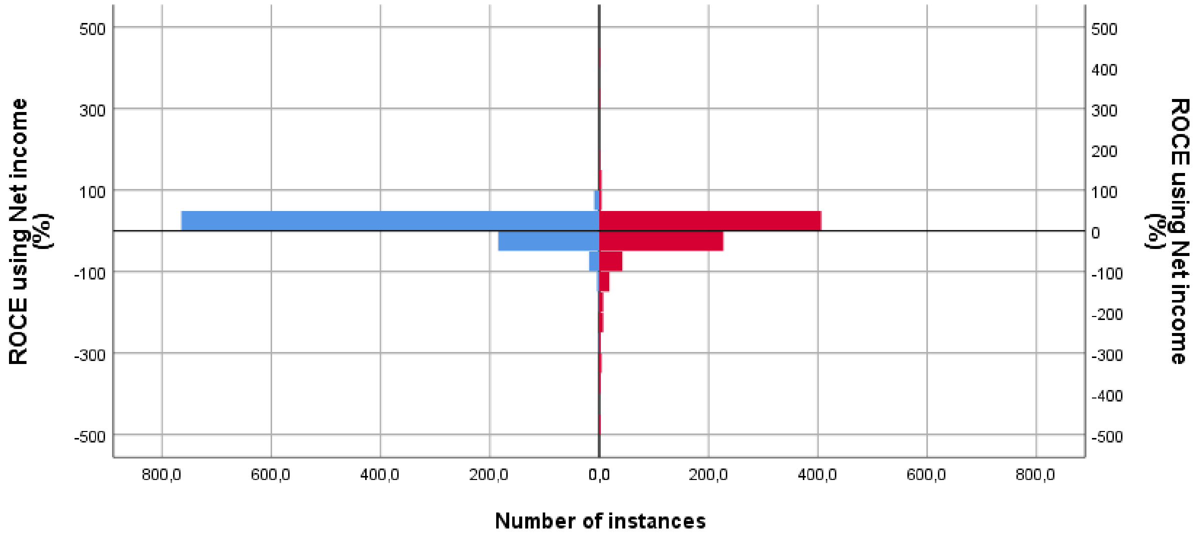


Figure 21 - Histograms of the ROCE using net income for good and bad companies.

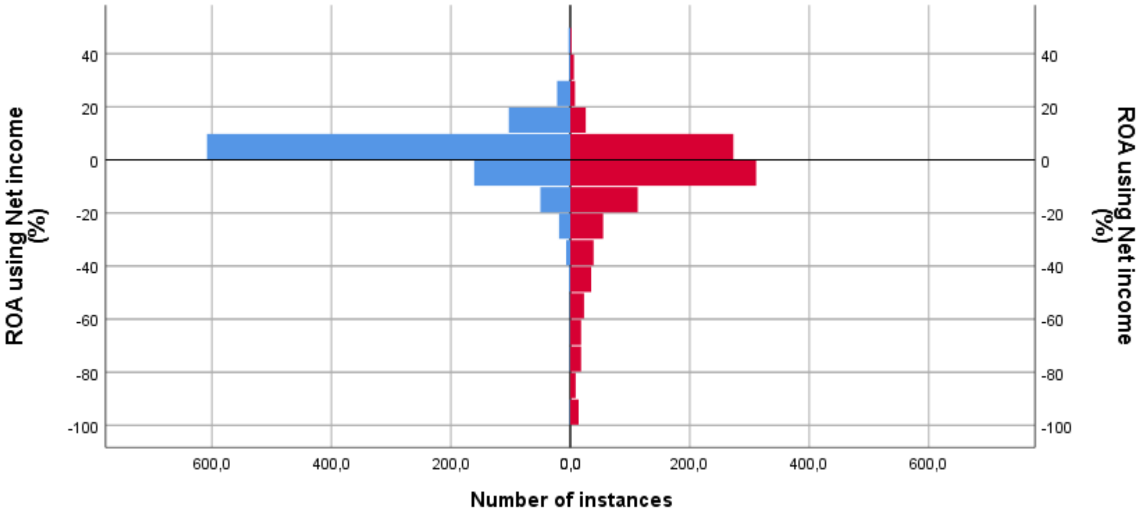


Figure 22 - Histograms of the ROA using net income for good and bad companies.

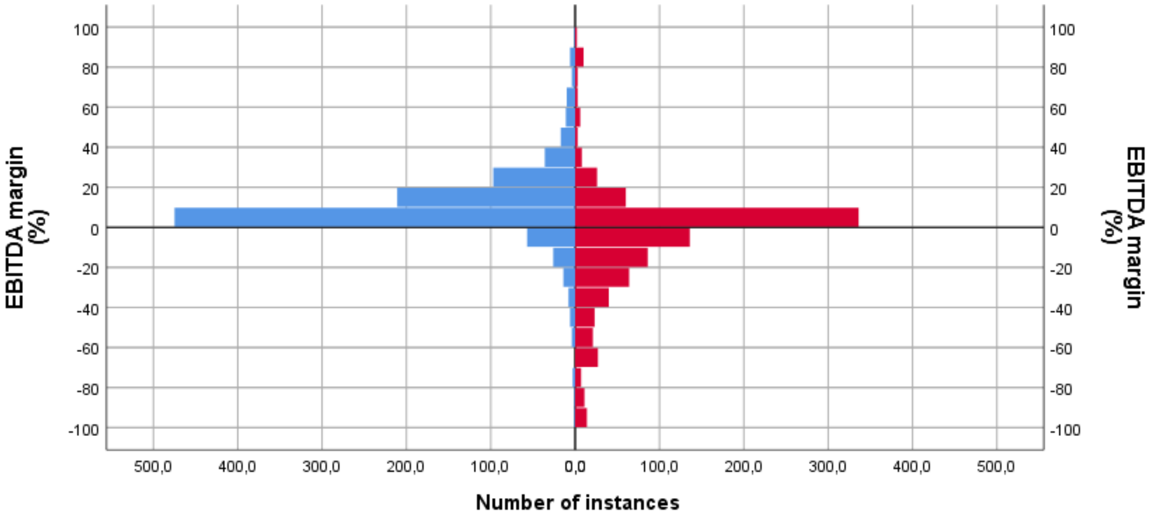


Figure 23 - Histograms of the EBITDA margin for good and bad companies.

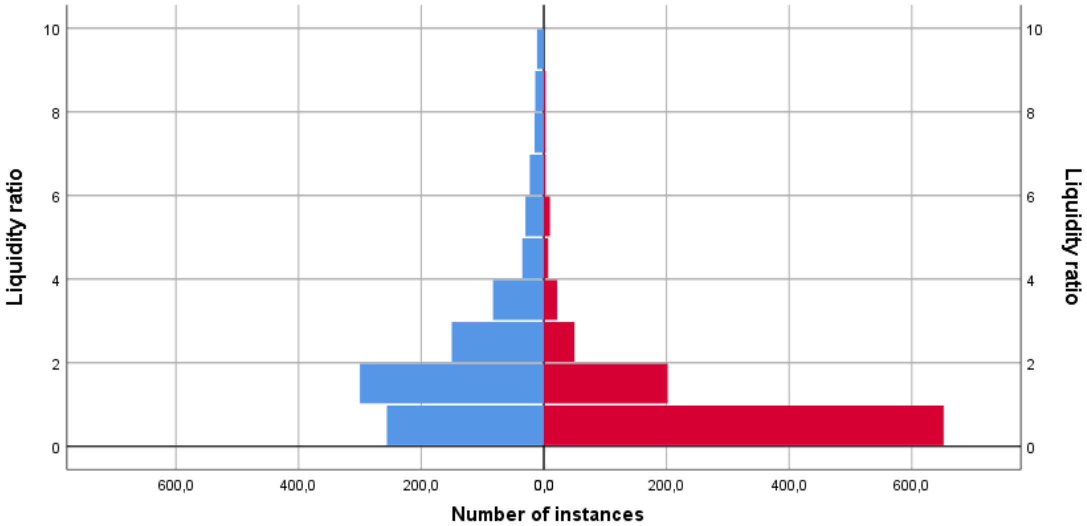


Figure 24 - Histograms of the liquidity ratio for good and bad companies.

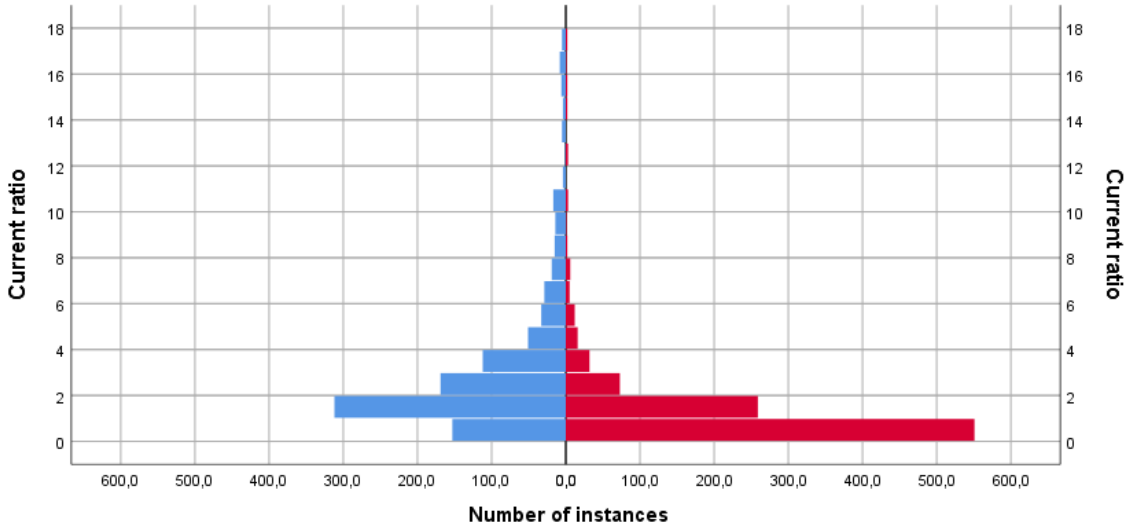


Figure 25 - Histograms of the current ratio for good and bad companies.

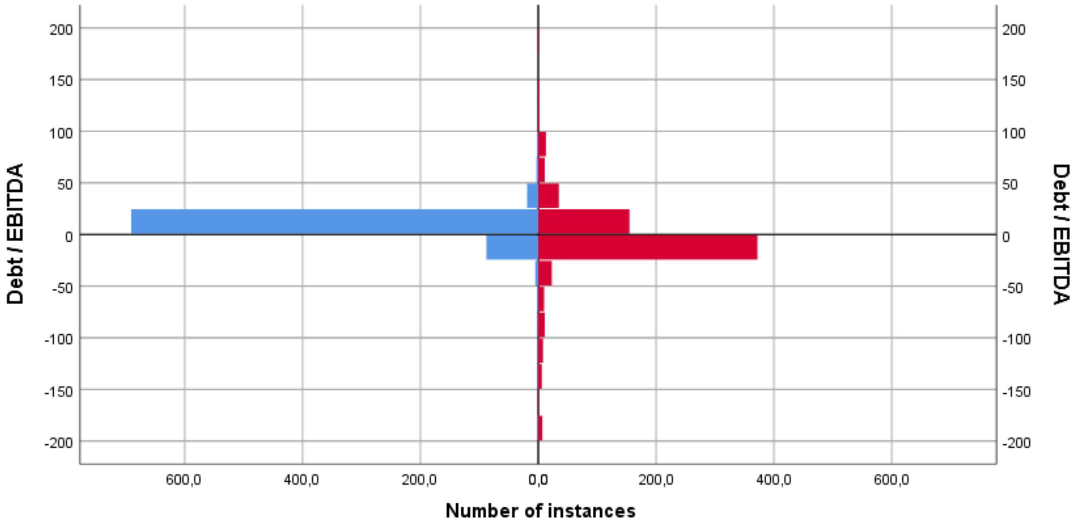


Figure 26 - Histograms of Debt / EBITDA for good and bad companies.

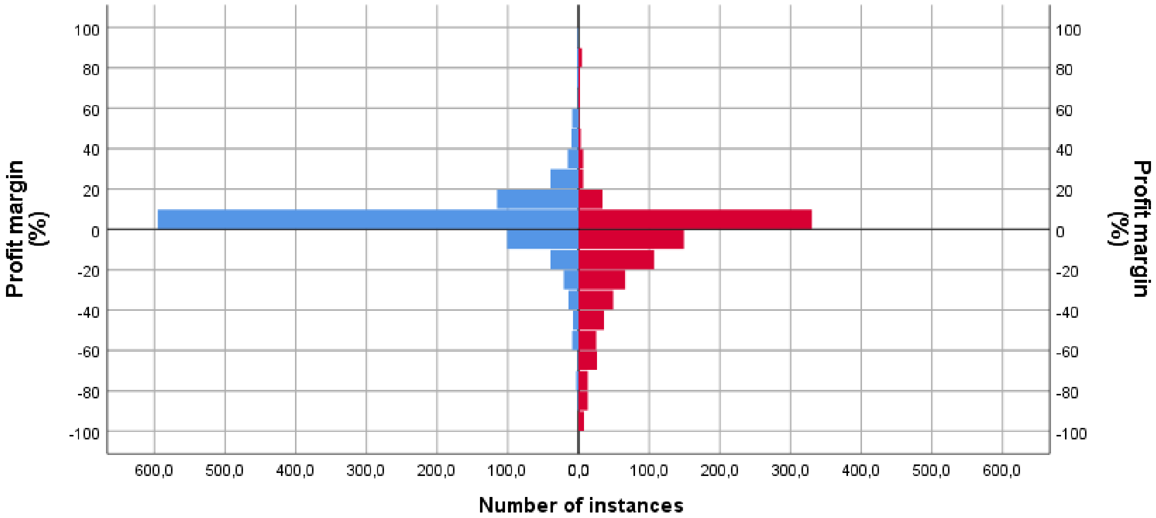


Figure 27 - Histograms of the profit margin for good and bad companies.

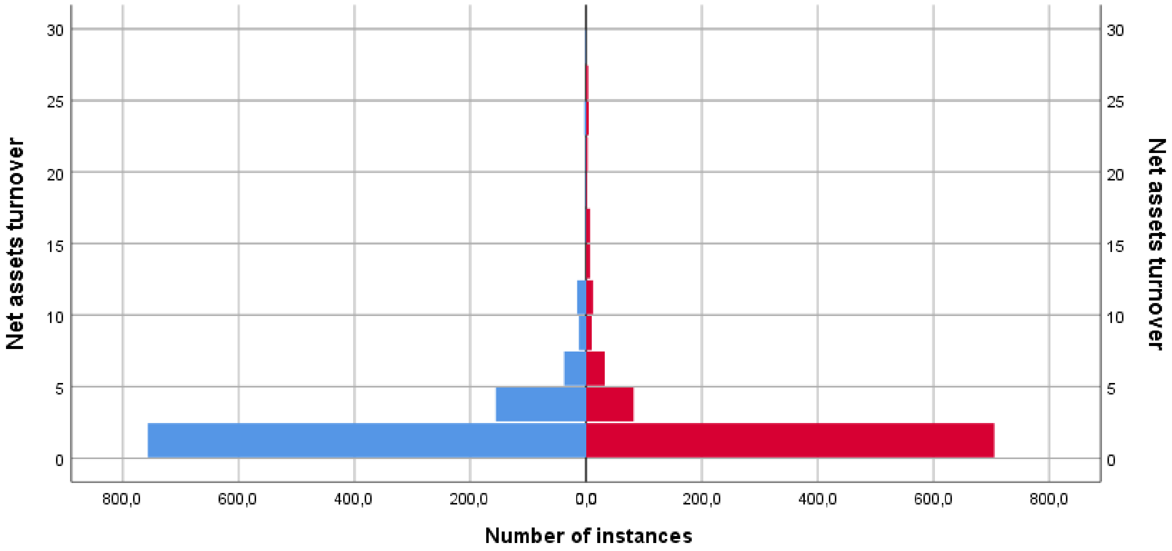


Figure 28 - Histograms of the net assets turnover for good and bad companies.

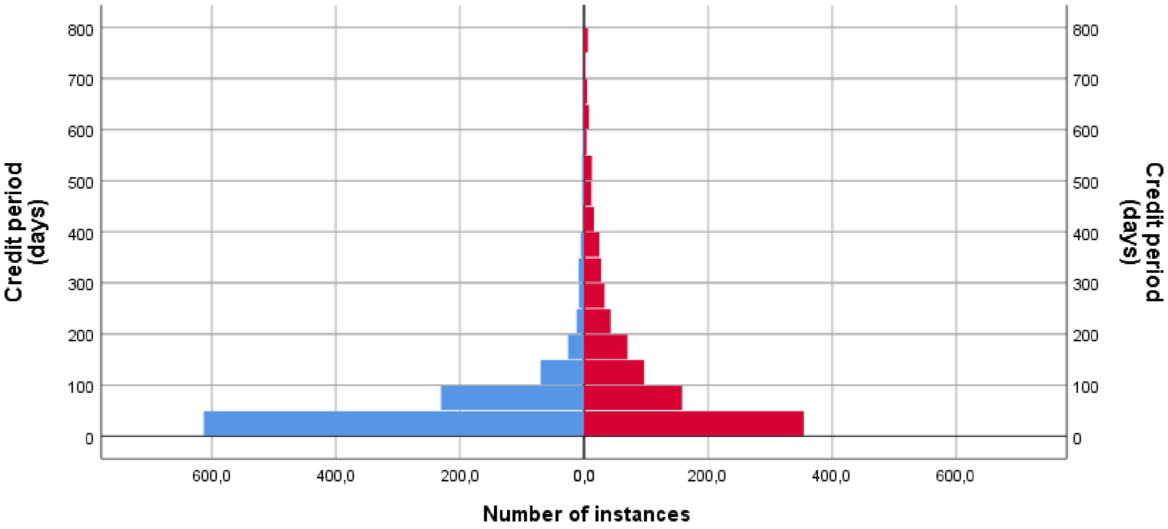


Figure 29 - Histograms of the credit period for good and bad companies.

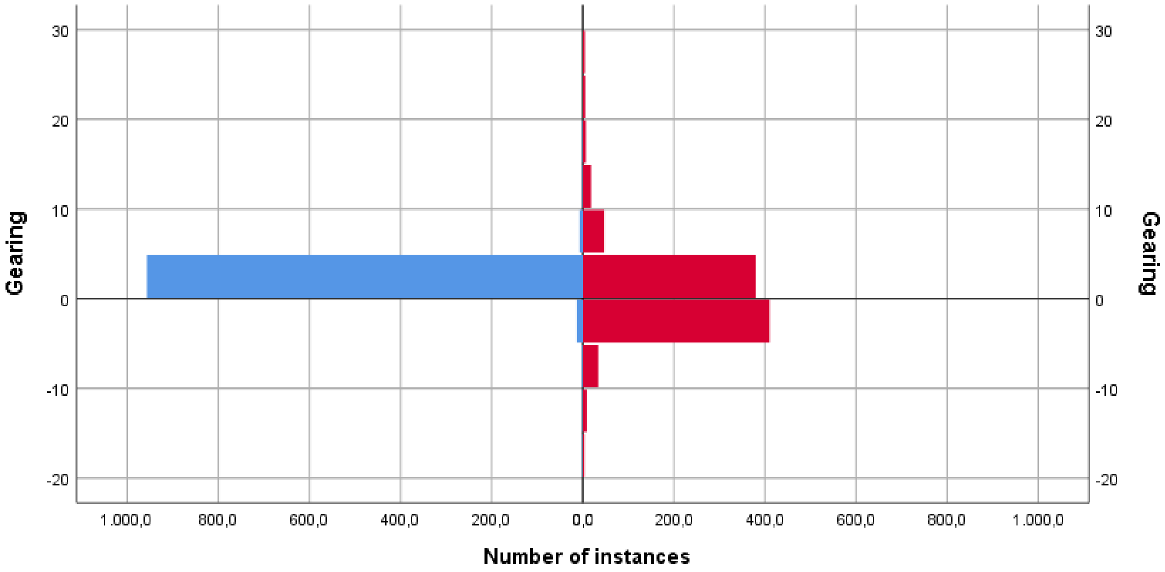


Figure 30 - Histograms of the gearing values for good and bad companies.

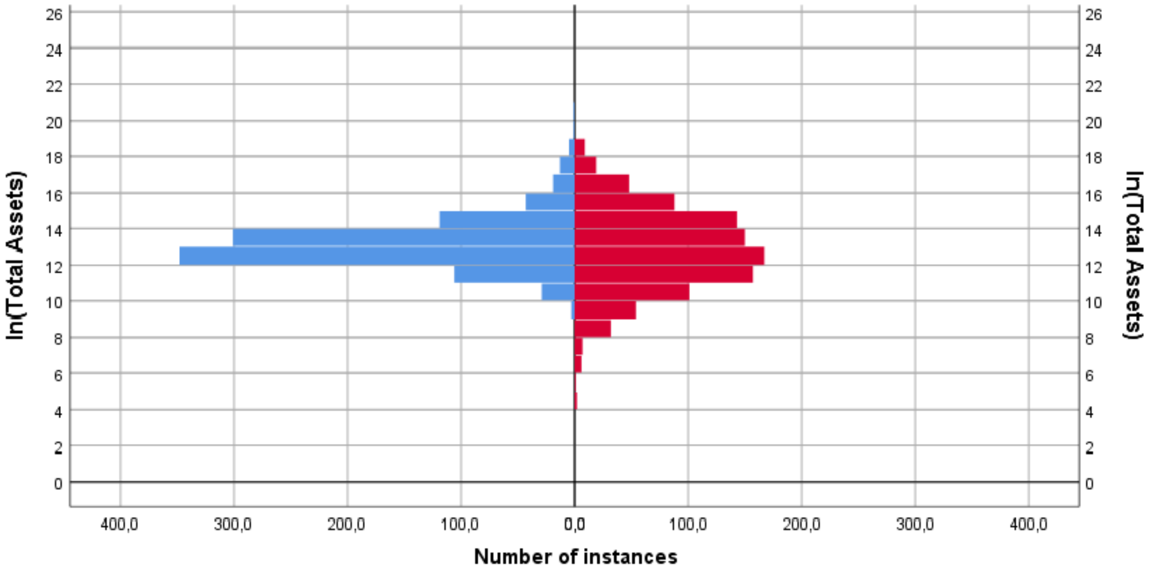


Figure 31 - Histograms of the natural logarithm of total assets for good and bad companies.

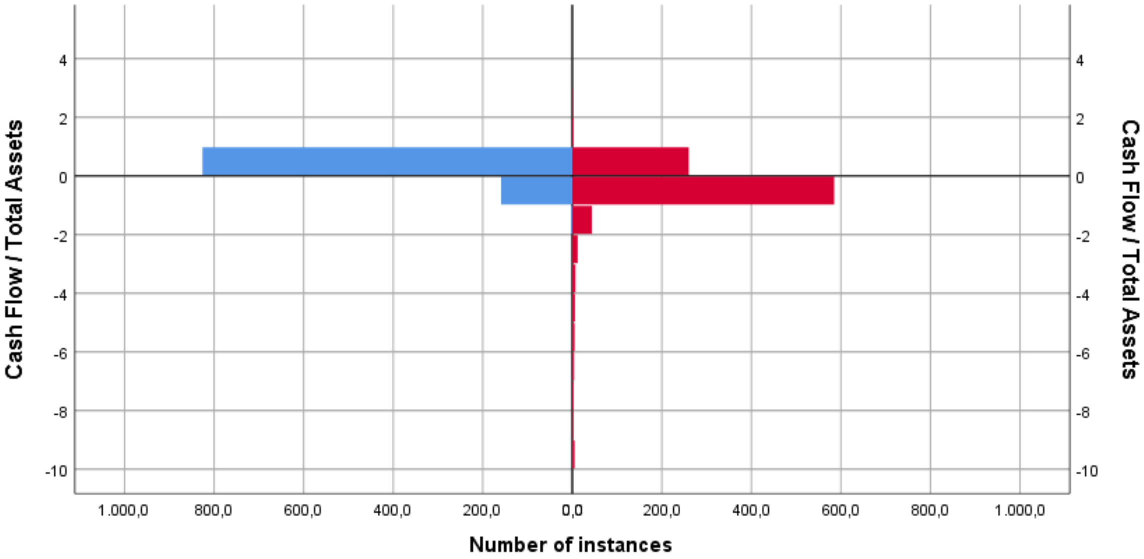


Figure 32 - Histograms of the Cash flow / Total assets for good and bad companies.

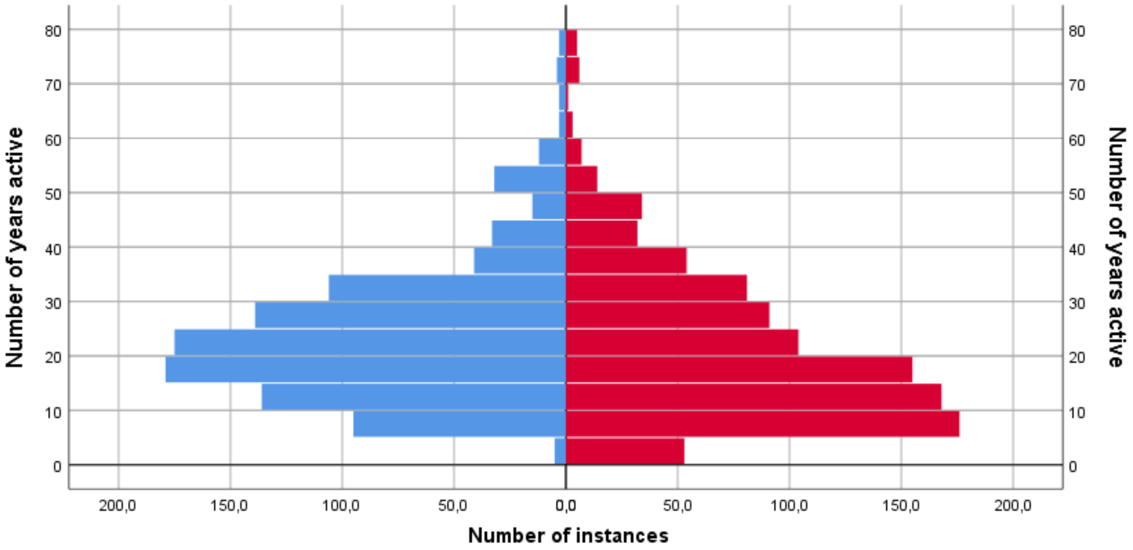


Figure 33 - Histograms of the number of years active for good and bad companies.

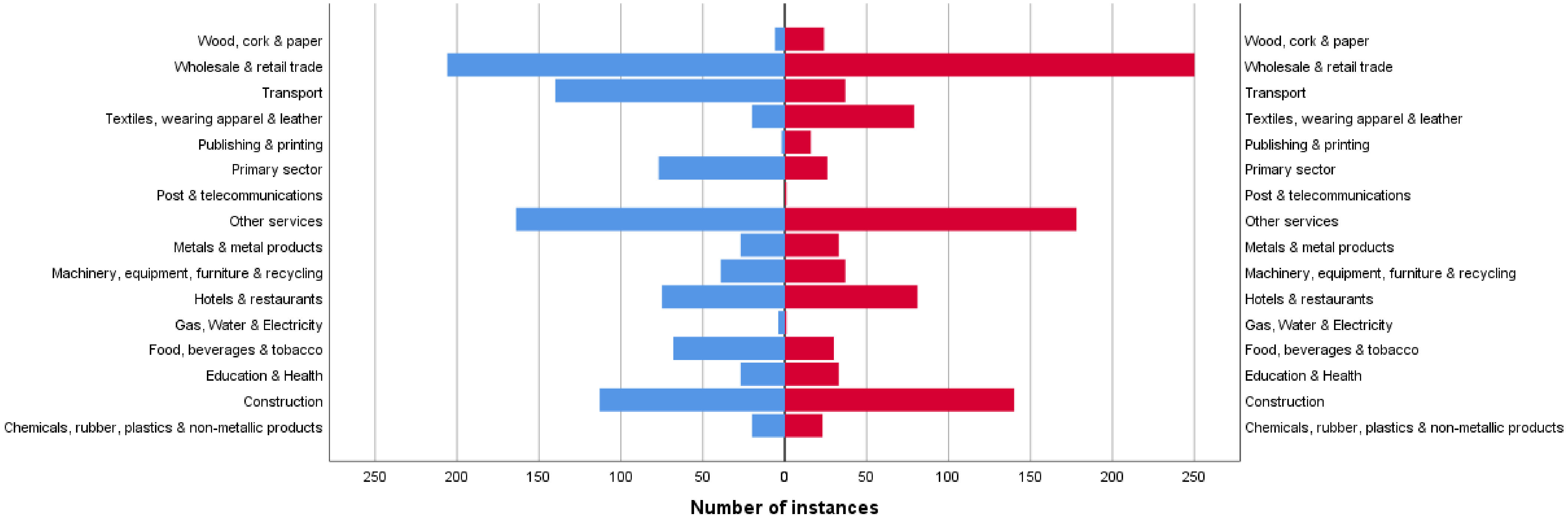


Figure 34 - Histograms of the BvD major sector for good and bad companies.

Appendix D – Pseudo-code for the partitioning algorithms.

Algorithm 1 (70% Training - 30% Testing - 0% Validation):

```
$i$ = 1;
while $i$ ≤ total number of instances in the sample {
        Generate a random integer n between 1 and 10;
                if n ≤ 7 then $x_i$= 1;
                else $x_i$= 0;
        $i$ = $i$ + 1;
  }
```

Algorithm 2 (60% Training - 15% Testing - 25% Validation):

```
$i$ = 1;
while $i$ ≤ total number of instances in the sample {
        Generate a random integer n between 1 and 20;
                if n ≤ 12 then $x_i$= 1;
                if 12 < n ≤ 15 then $x_i$= 0;
                else $x_i$= -1;
        $i$ = $i$ + 1;
}
```

Algorithm 3 (60% Training - 20% Testing - 20% Validation):

```
$i$ = 1;
while $i$ ≤ total number of instances in the sample {
        Generate a random integer n between 1 and 10;
                if n ≤ 6 then $x_i$= 1;
                if 6 < n ≤ 8 then $x_i$= 0;
                else $x_i$= -1;
        $i$ = $i$ + 1;
}
```

Appendix E – Results for all the iterations of the partitioning algorithm.

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 1	Partition 1	MLP1	87.2	93.5	77.8	0.937	0.874
		MLP2	86.6	92.6	77.8	0.934	0.868
		MLP3	85.5	93.5	73.6	0.946	0.892
		MLP4	88	90.7	84	0.954	0.908
	Partition 2	MLP1	91.1	92.2	89.1	0.956	0.912
		MLP2	89.9	92.2	85.9	0.945	0.89
		MLP3	90.5	92.2	87.5	0.956	0.912
		MLP4	92.7	95.7	87.5	0.958	0.916
	Partition 3	MLP1	89.3	93.5	82.5	0.938	0.876
		MLP2	86.7	91.1	79.6	0.932	0.864
		MLP3	88.6	92.3	82.5	0.94	0.88
		MLP4	88.2	92.9	80.6	0.954	0.908

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 2	Partition 1	MLP1	89.3	92.4	83.9	0.945	0.89
		MLP2	88.5	91.1	83.9	0.947	0.894
		MLP3	91.4	94.1	86.9	0.954	0.908
		MLP4	89.6	94.1	81.8	0.952	0.904
	Partition 2	MLP1	87.8	89	86.1	0.955	0.91
		MLP2	88.8	89	88.6	0.961	0.922
		MLP3	89.3	88.1	91.1	0.968	0.936
		MLP4	90.4	92.4	87.3	0.954	0.908
	Partition 3	MLP1	85.5	94.5	71.6	0.941	0.882
		MLP2	87.1	96.6	72.6	0.937	0.874
		MLP3	88	95.2	76.8	0.948	0.896
		MLP4	88	94.5	77.9	0.953	0.906

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 3	Partition 1	MLP1	87.3	92.9	78.4	0.938	0.876
		MLP2	88.4	92	82.7	0.948	0.896
		MLP3	88.2	94.6	77.7	0.95	0.9
		MLP4	89	92.9	82.7	0.946	0.892
	Partition 2	MLP1	87.7	93.2	77.8	0.934	0.868
		MLP2	86.3	94.7	70.8	0.891	0.782
		MLP3	92.2	96.2	84.7	0.961	0.922
		MLP4	91.2	95.5	83.3	0.96	0.92
	Partition 3	MLP1	88.8	94	80.2	0.958	0.916
		MLP2	88.8	93.3	81.3	0.948	0.896
		MLP3	89.6	96.6	78	0.963	0.926
		MLP4	91.7	97.3	82.4	0.96	0.92

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 4	Partition 1	MLP1	86.4	93.9	74	0.941	0.882
		MLP2	87.5	95.3	74.8	0.937	0.874
		MLP3	87	93	77.1	0.947	0.894
		MLP4	87.8	95.3	75.6	0.955	0.91
	Partition 2	MLP1	84	89.2	76.3	0.949	0.898
		MLP2	84.5	90	76.3	0.945	0.89
		MLP3	86	92.5	76.3	0.948	0.896
		MLP4	87.5	92.5	80	0.953	0.906
	Partition 3	MLP1	86.9	91	79.8	0.94	0.88
		MLP2	87.8	91	82	0.933	0.866
		MLP3	86.9	91.7	78.7	0.943	0.886
		MLP4	89	91	85.4	0.955	0.91

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 5	Partition 1	MLP1	90.3	95.7	82.5	0.941	0.882
		MLP2	89.7	96.2	80.4	0.945	0.89
		MLP3	89.5	95.2	81.1	0.949	0.898
		MLP4	91.2	95.7	84.6	0.947	0.894
	Partition 2	MLP1	89.5	95.7	76.8	0.94	0.88
		MLP2	90.7	94	83.9	0.954	0.908
		MLP3	90.7	97.4	76.8	0.952	0.904
		MLP4	91.9	96.6	82.1	0.968	0.936
	Partition 3	MLP1	94.6	96.4	90.9	0.954	0.908
		MLP2	92.2	94.9	86.4	0.937	0.874
		MLP3	94.1	96.4	89.4	0.954	0.908
		MLP4	94.1	96.4	89.4	0.95	0.9

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 1	Partition 1	RBF1	87.50	89.40	84.60	0.893	0.786
		RBF2	84.60	95.70	68.50	0.883	0.766
	Partition 2	RBF1	82.40	85.80	77.10	0.879	0.758
		RBF2	86.90	93.40	77.10	0.902	0.804
	Partition 3	RBF1	84.00	85.50	81.60	0.895	0.79
		RBF2	82.90	85.50	78.60	0.891	0.782

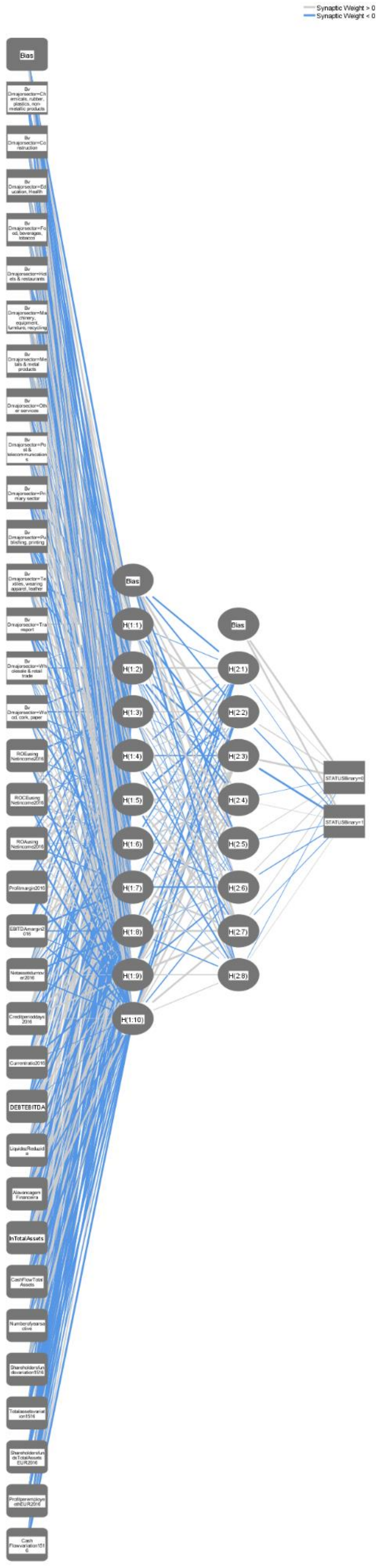
			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 2	Partition 1	RBF1	86.00	91.70	76.40	0.895	0.79
		RBF2	85.50	90.10	77.70	0.89	0.78
	Partition 2	RBF1	77.10	83.00	66.70	0.887	0.774
		RBF2	76.50	84.00	63.30	0.889	0.778
	Partition 3	RBF1	83.70	90.50	73.20	0.888	0.776
		RBF2	83.70	86.50	79.40	0.889	0.778

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 3	Partition 1	RBF1	83.10	86.10	78.40	0.892	0.784
		RBF2	78.90	84.80	69.90	0.864	0.728
	Partition 2	RBF1	80.00	82.50	76.50	0.891	0.782
		RBF2	80.00	85.10	72.80	0.889	0.778
	Partition 3	RBF1	82.50	83.00	81.60	0.893	0.786
		RBF2	82.90	87.90	74.70	0.896	0.792

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 4	Partition 1	RBF1	82.40	81.00	84.80	0.894	0.788
		RBF2	81.60	83.10	79.00	0.896	0.792
	Partition 2	RBF1	84.60	85.90	81.70	0.893	0.786
		RBF2	82.60	86.70	73.30	0.889	0.778
	Partition 3	RBF1	80.40	83.40	74.40	0.89	0.78
		RBF2	81.50	87.40	70.00	0.889	0.778

			PCC (%)	Sens. (%)	Spec. (%)	AUC	Gini Index
Iteration 5	Partition 1	RBF1	78.70	88.60	64.20	0.885	0.77
		RBF2	77.60	86.40	64.90	0.888	0.776
	Partition 2	RBF1	81.00	85.50	73.60	0.894	0.788
		RBF2	80.40	84.60	73.60	0.893	0.786
	Partition 3	RBF1	78.80	81.80	73.90	0.884	0.768
		RBF2	80.90	82.40	78.40	0.89	0.78

Appendix F – Diagram for the MLP artificial neural network.


Synaptic Weight > 0
Synaptic Weight < 0
Bias
Bv Dmajorsector=Chemicals, rubber, plastics, non-metallic products
Bv Dmajorsector=Construction
Bv Dmajorsector=Education, Health
Bv Dmajorsector=Food, beverages, tobacco
Bv Dmajorsector=Hotels & restaurants
Bv Dmajorsector=Machinery, equipment, furniture, recycling
Bv Dmajorsector=Metals & metal products
Bv Dmajorsector=Other services
Bv Dmajorsector=Post & telecommunications
Bv Dmajorsector=Primary sector
Bv Dmajorsector=Publishing, printing
Bv Dmajorsector=Textiles, wearing apparel, leather
Bv Dmajorsector=Transport
Bv Dmajorsector=Wholesale & retail trade
Bv Dmajorsector=Wood, cork, paper
ROEusing Netincome2016
ROCEusing Netincome2016
ROAusing Netincome2016
Profitmargin2016
EBITDAmargin2016
Netassetsturnover2016
Creditperioddays2016
Currentratio2016
DEBTEBITDA
LiquidezReduzida
Alavancagem Financeira
lnTotalAssets
CashFlowTotalAssets
Numberofyearsactive
Shareholdersfundsvariation1516
Totalassetsvariation1516
ShareholdersfundsTotalAssetsEUR2016
ProfitperemployeethEUR2016
Cash Flowvariation1516
Bias
H(1:1)
H(1:2)
H(1:3)
H(1:4)
H(1:5)
H(1:6)
H(1:7)
H(1:8)
H(1:9)
H(1:10)
Bias
H(2:1)
H(2:2)
H(2:3)
H(2:4)
H(2:5)
H(2:6)
H(2:7)
H(2:8)
STATUSBinary=0
STATUSBinary=1

Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

Appendix G – MATLAB script for the generation and analysis of the random forest model.

```
s = rng % Saves the current state of the random stream, ensuring the
reproducibility of the results

forest = TreeBagger(50, Dataset, 'Company_Status', 'PredictorSelection',
'curvature', 'OOBPrediction', 'On', 'OOBPredictorImportance', 'On', 'Surrogate',
'On', 'Method', 'classification') % Generates 50 bagged trees using the columns of
Dataset as the independent variables to predict Company_Status. Defines curvature
tests as the splitting method. Uses surrogate splits to deal with missing values.

figure; % Creates a new figure window with default settings
oobErrorBaggedEnsemble = oobError(forest); % Calculates the out-of-bag
classification error by testing instances in each tree that were not used in the
training process
plot(oobErrorBaggedEnsemble) % Plots the error of the ensemble of trees
xlabel 'Number of trees'; % Names the x axis as the number of grown trees
ylabel 'Out-of-bag classification error'; % Names the y axis as the out-of-bag
classification error


importance = forest.OOBPermutedPredictorDeltaError; % Stores the importance
estimates in a vector called importance
figure; % Creates a new figure window with default settings
bar(importance);
title('Curvature Test'); % Names the plot
ylabel('Predictor importance estimates'); % Names the vertical axis
xlabel('Predictors'); % Names the horizontal axis
h = gca;
h.XTickLabel = forest.PredictorNames; % Selects the names of the independent
variables to include in as labels in the horizontal axis
h.XTickLabelRotation = 60; % Sets the inclination of the labels in the horizontal
axis
h.TickLabelInterpreter = 'none';


[yfit, sfit] = oobPredict(forest) % Stores the class probabilities in sfit
Probability_of_being1 = sfit(:,end) % Extracts the last column of sfit, which
contains the probability of each instance belonging to the good companies class
(coded as 1) that the RF computes

[X, Y, T, AUC] = perfcurve(Company_Status, Probability_of_being1, '1') % Calculates
the AUC from the Probability_of_being1
```