NOVA IMS
Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

# The efficiency of bankruptcy predictive models

## Genetic Algorithms Approach

Mário Manuel Neto Antão

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# THE EFFICIENCY OF BANKRUPTCY PREDICTIVE MODELS

**-**

# GENETIC ALGORITHMS APPROACH

by

Mário Manuel Neto Antão

Dissertation presented as a partial requirement for obtaining the Master's degree in Information Management, specialization in Knowledge Management and Business Intelligence

**Advisor:** Mauro Castelli

May 2020

# ACKNOWLEDGMENTS

This dissertation marks the end of a journey of learning and challenge, after five years that allowed me to grow both on a personal level, and professional.

In the course of this dissertation, I relied on the direct or indirect support of many people to whom I am profoundly thankful. At the risk of not mentioning any of the contributions unfairly, I would like to express my thanks to the advisor of this dissertation Professor Mauro Castelli, for the guidance given for his encouragement, availability, suggestions, and support that he has always shown. Here I express my gratitude to him.

To all friends and colleagues who directly or indirectly contributed or helped in the preparation of this study, for their patience, attention, and strength in less easy moments. In order to avoid the risk of not mentioning any of them, I am not going to identify anyone, those to whom this gratitude is addressed will know it from now on, my sincere thanks.

I could not forget to thank my family for all the financial support, strength, and affection they have always given me throughout my academic life, as well as for the preparation of this dissertation ,which would have been impossible without their support.

To my girlfriend for having walked beside me, for her patience, understanding, and help given during the elaboration of this dissertation, especially for always presenting a smile when she sacrificed the days, weekends and holidays for the accomplishment of this study.

Also, I would like to thank Professor Dr. Cândido Jorge Peres Moreira, Professor Doctor Engineer Jaime Raúl Seixas Fonseca, for all guidance given in the collection of data and information as well as in its statistical and financial analysis and vision.

Moreover, I could not forget to thank Eng. Hélder Godinho Guerreiro, for all the lost hours and days helping me with problems and doubts related to the technical level of programming.

# ABSTRACT

The present dissertation evaluates the contribution of genetic algorithms to improve the performance of bankruptcy prediction models.

The state-of-the-art points to a better performance of MDA (Multiple Discriminant Analysis)-based models, which, since 1968, are the most applied in the field of bankruptcy prediction. These models usually recur to ratios commonly used in financial analysis.

From the comparative study of (1) logistic regression-based models with the forward stepwise method for feature selection, (2) Altman's Z-Score model (Edward I. Altman, 1983) based on MDA and (3) logistic regression with the contribution of genetic algorithms for variable selection, a clear predominance of the efficiency revealed by the former models can be observed. These new models were developed using 1887 ratios generated a posteriori from 66 known variables, derived from the accounting, financial, operating, and macroeconomic analysis of firms.

New models are thus presented, which are very promising for predicting bankruptcy in the medium to long term, in the context of increasing instability surrounding firms for different countries and sectors.

# KEYWORDS

Genetic algorithm; Logit; Bankruptcy; MDA; Z-Score; Logistic Regression

# RESUMO

A dissertação realizada avalia a contribuição dos algoritmos genéticos para melhorar a *performance* dos modelos de previsão de falência.

O estado da arte aponta para uma melhor *performance* dos modelos baseados em *MDA* (Análise descriminante multivariada) que por isso, desde de 1968, são os mais aplicados no âmbito da previsão de falência. Estes modelos recorrem habitualmente a rácios comumente utlizados em análise financeira.

A partir do estudo comparado de modelos baseados em (1) regressão logística com o método *forward stepwise* para escolha variáveis, (2) o modelo *Z-Score* de Edward Altman (1983) baseado em *MDA* e (3) regressão logística com o contributo de algoritmos genéticos para escolha variáveis, observa-se um claro predomínio da eficácia revelada por estes últimos. Estes novos modelos, agora propostos, foram desenvolvidos com recurso a 1887 rácios gerados a posteriori a partir de 66 variáveis conhecidas, oriundas da análise contabilística, financeira, de funcionamento e de enquadramento macroeconómico das empresas.

São assim apresentados novos modelos, muito promissores, para a previsão de falência a médio longo prazo em contexto de crescente instabilidade na envolvente das empresas, para diferentes países e sectores.

# PALAVRAS-CHAVE

Algoritmos Genéticos; *Logit*; Falência; *MDA*; *Z-Score*; Regressão Logística

# INDEX

1. Introduction ........................................................................................ 1
1.1. Project Genesis........................................................................................ 1
1.2. Objectives........................................................................................ 1
1.3. Dissertation Structures ........................................................................ 2
1.4. Research Questions and Hypothesis .................................................. 3
1.4.1. Research Questions and Hypothesis ........................................... 3
2. Literature review ........................................................................................ 4
2.1. Bankruptcy Problematic ........................................................................ 4
2.1.1. Bankruptcy Definition ................................................................... 4
2.1.2. Evolution of the Bankruptcy concept ........................................... 4
2.1.3. Definition and Evolution of the International and National Law of Bankruptcy 6
2.1.4. Number of Bankruptcy cases in Portugal and France. ................................ 7
2.2. Bankruptcy Predictive Models ........................................................... 10
2.2.1. Introduction ................................................................................ 10
2.2.2. The different types of Bankruptcy Predictive Models............................... 11
2.2.3. Model Based on MDA ................................................................... 13
2.2.4. Logistic Regression ...................................................................... 20
2.2.5. Genetic Algorithms ...................................................................... 23
3. Methodology........................................................................................ 35
4. Proposed Models ........................................................................................ 37
4.1. Sample and Population ........................................................................ 37
4.1.1. Population ................................................................................ 37
4.1.2. Sample Selection Criteria ............................................................. 37
4.1.3. Sample and Sample Control ........................................................ 40
4.2. Applications of the proposed models ................................................ 43
4.2.1. Code Development ...................................................................... 43
4.2.2. Adoption ................................................................................ 43
5. Results and discussion........................................................................................ 52
5.1. Introduction ........................................................................................ 52
5.2. MDA models........................................................................................ 53
5.2.1. Short-Term Performance ............................................................. 54
5.2.2. Long-term Performance ............................................................... 54

5.3. Logit Model with Forward Stepwise ........ 55
5.3.1. Short-Term Performance ........ 55
5.3.2. Long-term Performance ........ 56
5.4. Logit Model with GA ........ 56
5.4.1. Short-Term Performance ........ 57
5.4.2. Long-Term Performance ........ 57
5.5. Comparative Analysis of the Model's Performance ........ 58
6. Conclusions ........ 63
6.1. Research Questions and Hypothesis evaluation ........ 63
6.2. Conclusions ........ 65
7. Limitations and recommendations for future works ........ 67
7.1. Limitations ........ 67
7.2. Recommendations for Future Works ........ 67
8. Bibliography ........ 68
9. Annexes ........ 74

# LIST OF FIGURES

Figure 2.1 Percentage of Bankruptcies in Portugal in 2010 per Sector .................................. 8
Figure 2.2 Percentage of Bankruptcies in Portugal in 2016 per Sector .................................. 8
Figure 2.3 Number of Bankruptcies in Portugal n 2010 vs 2016 per sector (CAE).................... 9
Figure 2.4 Percentage of Bankrupt Companies per Number of Companies in Portugal........... 9
Figure 2.5 Percentage of Bankrupt Companies per Number of Companies in France ........... 10
Figure 2.6 Proportion of model categories employed by past studies. Adapted from “Predicting corporate bankruptcy: where we stand?” by (Adnan Aziz & Dar, 2006) ........................ 12
Figure 2.7 Proportion of models employed by past studies. Adapted from “Predicting corporate bankruptcy: where we stand?” by (Adnan Aziz & Dar, 2006) ........................ 12
Figure 2.8 Cut Off Scores for listed companies ...................................................................... 17
Figure 2.9 Cut Off Scores for non-listed companies ............................................................... 19
Figure 2.10 Objective Function and Parameter Value (“Genetic Algorithms - Introduction - Tutorialspoint,” n.d.).............................................................................................. 24
Figure 2.11 Gene, Chromosome and Population Example ..................................................... 25
Figure 2.12 Genotype and Phenotype transformation ........................................................... 25
Figure 2.13 Basic steps of Gas, Adapted [reprinted] from (K.-S. Shin & Lee, 2002)................ 26
Figure 2.14 Roulette Wheel Selection, (“Genetic Algorithms - Parent Selection - Tutorialspoint,” n.d.).............................................................................................. 27
Figure 2.15 Tournament Selection Example,(“Genetic Algorithms - Parent Selection - Tutorialspoint,” n.d.).............................................................................................. 28
Figure 2.16 Selection process using an elitism function. Adapted [reprinted] from (Romero-Hdz, Aranda, Toledo-Ramirez, Segura, & Saha, 2016) ........................................ 28
Figure 2.17 Crossover Operation in Python Genetic Algorithms, (Gour, n.d.)........................ 29
Figure 2.18 Mutation Operator in Python Genetic Algorithms (Gour, n.d.) ........................... 29
Figure 2.19 Binary Representation ......................................................................................... 29
Figure 2.20 Permutation Representation ............................................................................... 29
Figure 2.21 Floating-Point Value Representation ................................................................... 30
Figure 2.22 Integer Value Representation .............................................................................. 30
Figure 3.1 – Dissertation Methodology .................................................................................. 35
Figure 4.1 Identification of normal vs. anomalous observations with Isolation Forest (Liu et al.,2008) ................................................................................................................. 41
Figure 4.2 Outliers Detection and Treatment Code ................................................................ 43
Figure 4.3 Correlation Analysis and Treatment Code ............................................................. 44
Figure 4.4 Data Standardization Code .................................................................................... 45

Figure 4.5 Logit with GA Process ........................................................................................ 46
Figure 4.6 Logit with GA process for France and Portugal Sector G and C ............................ 47
Figure 4.7 K-fold Cross-Validation Process ............................................................................ 48
Figure 4.8 Selection, Mutation and Crossover settings code................................................. 49
Figure 4.9 Z-Score Model Code............................................................................................. 49
Figure 5.1 ROC curve and AUC for Logistic Model with GA applied to France and Portugal, sector G and C........................................................................................................... 53
Figure 5.2 Models Performance (Portugal and France - Sectors G and C)............................. 59
Figure 5.3 Model Performance (Portugal - Sector G) ........................................................... 60
Figure 5.4 Model Performance (France - Sector G) ............................................................... 60

# LIST OF TABLES


Table 2.1 Comparison of definitions of bankruptcy in benchmark investigations.................. 5
Table 2.2 Average of Variables and Significance Test (listed companies) ............................ 15
Table 2.3 Relative explanatory contribution of the variables.............................................. 16
Table 2.4 Model Efficiency..................................................................................................... 17
Table 2.5 Model Efficiency two year before bankruptcy ..................................................... 18
Table 2.6 Summary of Model Results .................................................................................... 18
Table 4.1 Initial Population Summary.................................................................................... 37
Table 5.1 Confusion Matrix..................................................................................................... 52
Table 5.2 MDA model short-term performance .................................................................. 54
Table 5.3 MDA model long-term performance..................................................................... 54
Table 5.4 Logit with Forward Stepwise model short-term performance.............................. 55
Table 5.5 Logit with Forward Stepwise model long-term performance ............................... 56
Table 5.6 Logit with GA model short-term performance...................................................... 57
Table 5.7 Logit with GA model long-term performance ...................................................... 57
Table 5.8 Performance Comparison between the models presented in this dissertation ..... 58
Table 5.9 Logit with GA vs MDA for France and Portugal Sector C and G, one year before the bankruptcy ..................................................................................................... 61
Table 5.10 Logit with GA vs MDA for France and Portugal Sector C and G, four years before the bankruptcy ..................................................................................................... 61
Table 5.11 Logit with GA vs MDA for Portugal Sector G, one year before the bankruptcy .... 62
Table 5.12 Logit with GA vs MDA for Portugal Sector G, four years before the bankruptcy .. 62
Table 6.1 Models Comparative Analysis ............................................................................... 64
Table 6.2 Models Results achieved with the smallest sample (Portugal Sector C)................ 64
Table 9.1 Initial Variables Description, represented in Table 9.2 ......................................... 74
Table 9.2 A 1 - Initial Variables.............................................................................................. 74
Table 9.3 A 2 - Outlier Treatment Summary Results ........................................................... 76
Table 9.4 A 3 - Logit with Forward Stepwise Variables........................................................ 77
Table 9.5 A 4 - Logit with GA Variables................................................................................. 78
Table 9.6 A 5 - Variables (used in the models) Description .................................................. 83

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **GA** | Genetic Algorithm |
| **MDA** | Multiple Discriminant Analysis |
| **CAE** | Portuguese Classification of Economic Activities |
| **SVM** | Support Vector Machine |
| **NN** | Neural Network |
| **AI** | Artificial Intelligence |
| **AIES** | Artificially intelligent expert system models |
| **LDA** | Linear Discriminant Analysis |
| **CPEREF** | Code of Special Company Recovery and Bankruptcy Procedures |
| **PLS-DA** | Partial Least Square Discriminant Analysis |
| **GSR** | GA Genetic Score |
| **ANN** | Artificial Neural Network |
| **GP** | Genetic Programing |
| **BK** | Bankrupt |
| **NBK** | Non-Bankrupt |

# 1. INTRODUCTION

This dissertation aims to contribute to the improvement of the performance of bankruptcy prediction models.

Since the early 1960s, several authors have focused on this issue by producing models of different genesis, always with the aim of improving the level of certainty of the prediction and its antecedence.

The growing number of bankruptcies and their significant impact on the economic competitivity of the various sectors, countries, or regions, justify the increase in resources allocated to research in this area of knowledge.

## 1.1. PROJECT GENESIS

It is widely known that the number of bankruptcies is directly correlated with the economic situation of countries, regions, common areas, and the world. The instability experienced in the recent past and the present justifies the increased concern in the study of this subject

In fact, during recent years, the world economy has become very different from that which had been since the recovery from the Great Depression.

Approximately 20 years ago, a financial crisis affected the world economy in 2007, wherein one of the origins of this crisis, "the subprime crisis," was permitted to the financial institutions approve low-quality loans, such as the NINJA-type loans. Due to these indiscriminate practices, the financial word paid a heavy price, being one of the principal causes, by many, for the worst crises in the history of capitalism since 1929, affecting, directly or indirectly, all sectors of activity and countries.

Moreover, nowadays we are presented with "a slower growth abroad and the U.S.-China trade war, the US Federal Reserve cut interest rates for the first time since the financial crisis and a slowing global economy is pressuring central banks abroad to lower borrowing rates at unprecedented levels and a tit-for-tat tariff war between Washington and Beijing is weighing on business sentiment," as well "an inversion of the yield curve. The bond market phenomenon is historically a good signal of an eventual recession: It has preceded the seven last recessions. A recession occurs about 22 months after an inversion on average, according to Credit Suisse." (Hanson, n.d.)

In addition, due to the Brexit phenomenon in a scenario of a no-deal Brexit that can impact the world economy, "the shock in the U.K. hits the ailing European economy hard, and the impact reverberates around the globe. Things could be made a lot worse if a messy departure adds to the already elevated levels of uncertainty that have been created by the trade war." ("How a No-Deal Brexit May Become a Problem for the World Economy - Bloomberg," n.d.)

## 1.2. OBJECTIVES

Due to the actual economic scenario, the world has highlighted the need to anticipate and predict these situations in order to allow timely contingency measures to be taken, or at least to make it possible to mitigate the adverse effects.

During the recent decades, was made preliminary work by Beaver (1966) in the application of the univariate analysis to the prediction of "bankruptcy," followed by Altman (1968) and its multivariate discriminant analysis, as well, several authors have developed different techniques and models for this purpose. From all the techniques applied for the last 60 years of study and prediction of "bankruptcy," techniques such as multiple discriminant analysis (MDA) (Edward I. Altman, 1968, 1983), logit (Ohlson, 1980)and probit (Zmijewski, 1984), we highlight the Multivariate Discriminant Analysis, due to its long-lasting applicability, simplicity, and effectiveness.

Furthermore, several studies have recently shown that artificial intelligence such as neural networks ( NNs) may be an effective approach for classification problems to which conventional statistical methods have previously been applied (Barniv, Agarwal, & Leach, 1997; Beaver, 1966; Bell, 1997; Chung & Tam, 1993; Efrim Boritz & Kennedy, 1995; Etheridge & Sriram, 1997; Fletcher & Goss, 1993; Jo, Han, & Lee, 1997; Odom & Sharda, 1990; Salchenberger, Cinar, & Lash, 1992);

While numerous theoretical and experimental studies have shown the value of NNs in classification studies, exposing several cons in developing and exploiting the model. First, due to the difficulties of finding an appropriate NN model, which can reflect problematic in the cause because the network architectures, learning methods, and parameters are varied. Secondly, the user cannot fully grasp and comprehend the final rules acquired by the NN models, also referred to as black boxes.

In this study, we highlight and propose genetic algorithms (Ga's) application to corporate failure prediction modeling, having as an advantage capability of extracting rules that are easy to understand for users like expert systems.

It should be noted that, in this dissertation, the word bankruptcy due to a lack of consensus in the literature on the meaning of the term is associated with the inability of a company to comply with its commitments up to a simple calculation of Assets < Liabilities.

## 1.3. Dissertation Structures

This dissertation is organized into nine chapters.

In the first chapter, the introduction has presented a summary of the objectives addressed in this dissertation, and the dissertation structure is explained.

In the second chapter is presented the State of Art of the models applied to predict the corporate bankruptcy. Moreover, is also explored the bankruptcy problematic, the most referred Bankruptcy Predictive Models, and also a more focused exploration of the Models based on MDA, along with the potential of Genetic Algorithms when applied to the prediction of corporate bankruptcy.

In the third chapter, the methodology of this dissertation is presented.

In the fourth chapter, Proposed Models, the structure of the predictive models (MDA and GA) is explained as well as the logic behind. Also introduced is the sample and population that was used to the creation of the models along with the application of the proposed models, including the Code Development, Adoption, and results.

In the fifth chapter, a Comparative analysis of the model's performance is made a comparison between both models (MDA and GA) into different perspectives, Short-Term and Long-Term Performance. In conjunction with this comparison, the restrictions and limitations of the model's applications are also scrutinized.

In the sixth chapter, Conclusions, the Research Questions and Hypothesis are evaluated.

In the seventh chapter, it is explored the Limitations and Further Developments that can be performed in this thematic.

In the eighth chapter covers all the Bibliography and References utilized in this dissertation.

Lastly, in the ninth chapter, the Annexes of this dissertation are presented.

## 1.4. RESEARCH QUESTIONS AND HYPOTHESIS

### 1.4.1. Research Questions and Hypothesis

#### 1.4.1.1. Research Questions

1. Will the application of genetic algorithms for predicting bankruptcies be promising?

2. Will the size of the sample be a limitation to the application of GA models?

3. Does the data from different sectors and countries influence the performance of the models?

4. Does the performance of GA predictive models have different efficiency from those based on MDA in the short and long term?

#### 1.4.1.2. Hypothesis

H1. GA predictive models of bankruptcy are more effective than MDA predictive models.

H2. The GA models, even with relatively small samples, maintain a good performance in supporting the prediction of bankruptcy.

H3. Isolation Forest is a promising method in the identification and elimination of outliers, taking into account the significant volume of economic variables involved.

# 2. LITERATURE REVIEW

## 2.1. BANKRUPTCY PROBLEMATIC

### 2.1.1. Bankruptcy Definition

Although there are multiple notions of corporate bankruptcy, this issue has been of concern to several authors, inducing an extensive set of definitions proposed in the published international bibliography on the subject.

Since it is the crucial reference in this area of research related to this study, reference should be made to Altman's broad definition (Edward I. Altman, 1993), which defined bankruptcy as a situation in which a business fails to pay its debt or other claims, is unable to meet current liquidity obligations because of a persistent lack of liquidity or simply induces long-term rates of return lower than company capital. In addition, Altman often classifies all the companies listed as legally bankrupt in liquidation proceedings or under court supervision, as well as in the recovery process, as bankrupt.

### 2.1.2. Evolution of the Bankruptcy concept

Bankruptcy can be defined by three vectors from different authors: legal, economic, and financial.

The objective legal viewpoint, which as a rule organizes the concept according to two aspects-that of the company in distress and that of the bankrupt company-provides for the majority of cases that companies with financial problems, i.e., unable to guarantee the settlement of outstanding and ultimately fragile and diminished liabilities, which are assisted in a judicial recovery process. Some scholars, such as Malécot (1991), do not differentiate between financial difficulties and bankruptcy, revealing other (Lee, 1985) significant concerns about the distinction between difficult financial situation and bankruptcy, especially when it comes to situations involving credit institutions ' decisions on how to recover their credits.

Furthermore, in the economic and financial field, the authors identify several types of risks observed in the companies about to bankrupt. Thus, depending on the set of ratios involved, Casta and Zerbib (CASTA, JF, & JP, 1979) mention the liquidity risk (associated with the inability to solve short-term liabilities), the asset or over-indebtedness risk, associated with the "credit-men" method, which will be developed in the following sections, the risk of non-reimbursement proposed by Altman (1968), the economic risk and the non-liquidity risk proposed by Collongues (1977) and finally the asset risk and the non-repayment risk identified by Conan and Holder (1979).

Moreover, the correlation of bankruptcies with insufficient environmental suitability, i.e., from a strategic point of view, is well explained in the Bescos research (Bescos, 1987) which mention that as the primary cause of the problematic situations/bankruptcy experienced by companies, their apparent inability to adapt to the environment, this situation at an advanced stage of maladjustment can, therefore, be seen in its advanced stage of maladjustment.

Lastly, since this is the crucial reference in this area of research, reference should be made to the broad definition of Altman (1993), which described bankruptcy as circumstances in which corporations are unable to liquidate debt service or other receivables, are unable to meet current obligations due to chronic liquidity shortages or simply cause long-term rates of return below the cost of the company. Altman also classifies all firms that are or are recovering, legally bankrupt, in liquidation or under court supervision as bankrupt. In summary, we present several concepts of insolvency used in comparison to predictive models in the table below. Such concepts are organized according to two significant vectors, one defined between the perception of the high probability of bankruptcy and the actual death of the company, and the other, ranging from the concept's legal independence to its complete allocation of the legal requirements stated by the supervisory bodies.

Table 2.1 Comparison of definitions of bankruptcy in benchmark investigations

| Year | Author | Criteria |
|---|---|---|
| 1966 | Beaver (Beaver, 1966) | “Liquidation of assets, inability to pay shareholders or bondholders.” |
| 1968 | Altman (Edward I. Altman, 1968) | “Deposit of the balance sheet in the form provided by the in Chapter 10 of the National Bankruptcy Act (USA).” |
| 1972 | Edminster (Edmister, 1972) | “Companies that do not make full reimbursement of the amounts agreed by the SBA (Small Business Administration), the body responsible for regulating the recovery processes.” |
| 1975 | Sinkey (Sinkey Jr, 1975) | “Violation of laws and regulations jeopardizing solvency.” |
| 1979 | Conan and Holder (Conan & Holder, 1979) | “Companies in difficulty, whose financial statements have already been analyzed by DATAR (the official bankruptcy supervisory body in France).” |
| 1980 | Pettway and Sinkey (Pettway & Sinkey, 1980) | “Declaration of insolvency by the rating or restructuring agency.” |
| 1983 | El Hennawy and Morris (El Hennawy & Morris, 1983) | “Liquidated, suspended by court order or under controlled management.” |
| 1985 | Frydman et al. (Frydman, Altman, & Kao, 1985) | “Effective bankruptcy and application for bankruptcy by chapter XI.” |
| 1988 | Aziz et al. (Aziz, Emanuel, & Lawson, 1988) | “Excluded from COMPTUSAT for bankruptcy and who interrupted trading.” |
| 1990 | Koh and Killough (Koh & Killough, 1990) | “Reported in the Wall Street Journal Index as broke.” |
| 1993 | Theodossiou (Theodossiou, 1993) | “Application for bankruptcy or controlled management.” |
| 1993 | Altman (Edward I. Altman, 1993) | “Not to settle the debt service chronically or simply induce long-term rates of return lower than their cost of capital.” |
| 1996 | Begley et al. (Begley, Ming, & Watts, 1996) | “Excluded from COMPTUSAT by request of chapter XI.” |
| 1999 | Kahya and Theodossiou (Kahya & Theodossiou, 1999) | “Failure to perform obligations or attempt to negotiate with creditors.” |
| 2000 | Altman (E I Altman, 2000) | “Application for bankruptcy under Chapter X of the National Bankruptcy Act.” |

| Year | Author | Criteria |
|---|---|---|
| 2001 | Grice e Ingram (Grice & Ingram, 2001) | “Request for Chapter XI, Chapter VII (liquidation, vulnerability to default, or low share rating).” |
| 2002 | Ross et al. (Ross, Westerfield, & Jaffe, 2002) | “Difficulty in meeting obligations, Assets not sufficient to settle Liabilities.” |
| 2005 | Pompe e Bilderbeek (Pompe & Bilderbeek, 2005) | “Legal classification as bankrupt.” |
| 2007 | Boritz et al. (Boritz, Kennedy, & Sun, 2007) | “Bankruptcy, controlled management, or liquidation request.” |
| 2008 | Agarwal & Taffler (Agarwal & Taffler, 2008) | “Controlled management or liquidation.” |
| 2009 | Xu & Zhang (Xu & Zhang, 2009) | “List of companies removed from the stock exchange (delisted).” |
| 2012 | Brealey &Myers (Brealey, Myers, Allen, & Mohanty, 2012) | “Moment from which business is worth more dead than alive. Take control of the company by the creditors for breaking promises or shareholder use of default rights.” |

### 2.1.3. Definition and Evolution of the International and National Law of Bankruptcy

Since this is a long-standing issue, it will be important to examine the Bankruptcy Law established by the Belgian Government on April 18, 1851, according to which a bankrupt company's main characteristics are: it is a company on its own behalf or in a joint capacity that ceases to fulfill its obligations towards its creditors, leading to the suspension of the credit. The bankruptcy filing action may come from the corporation itself, by depositing the balance sheet, creditors, or the commercial court itself.

The procedure envisaged for the bankruptcy proceedings included a declaratory bankruptcy judgment, the appointment of a commissioner judge, a trustee who would provisionally guarantee the management and ultimately the liquidation and consequent distribution of the bankrupt estate among the creditors. Therefore, the bankruptcy proceedings could have three forms of conclusion; the composition before bankruptcy, the liquidation, or the commercial court's closure of the company.

There are two clear legal trends in bankruptcy. One is embodied in French and Italian law, which considers bankruptcy to be a private institute of commercial undertakings, with debtors who are not commercial undertakings being remitted for the individual application scheme. A second group, involving German and Anglo-Saxon law, which does not differentiate between the various debtor entities as regards the institution of bankruptcy. Portuguese legislation, as from Decree-Law no. 21758 of 1932, establishes distinct procedural procedures for commercial and non-commercial entities.

The current trend is to stop discriminating against commercial and non-commercial users, with the legislation becoming generally applicable.

CPEREF put an end to the separation of processes observed under Portuguese law, which regarded bankruptcy as a private institution of commercial traders, reserving insolvency to non-traders unable to fulfill their duties. Accordingly, Decree-Law 21758 of 22/10/1932, which implemented the special insolvency procedures, modified the definition when the CPEREF took effect. Under Article 3, insolvency is supposed to refer to the restructuring or bankruptcy proceedings of the company, so we will use both terms; insolvency and bankruptcy.

In view of the fact that the present analysis is formulated within the framework of the forecast of insolvency from an economic and financial perspective, it is important to point out that, although it is different from a legal point of view to talk of bankruptcy or insolvency, the latter being only one of the reasons for bankruptcy, we shall use both terms in an undifferentiated manner, except for legal aspect. In Portugal, based on the CPEREF, a company is considered insolvent if, due to a lack of own resources and a lack of credit, it is unable to fulfill its obligations on time.

#### 2.1.4. Number of Bankruptcy cases in Portugal and France.

The problem of bankruptcy, a subject highly correlated with environmental instability - a growing change in size and pace of the context in which companies operate - has assumed growing importance over the last few decades.

Regardless of the influence on the pace of bankruptcies caused by economic cycles, an increase in the pace of bankruptcies and their impact on competition rules in the various sectors of activity can be observed.

The volume of bankruptcies recorded in each area of activity, weighted by the number of firms in the sector, is conditioned by its level of competitiveness, its cost structure and its critical point, and the pressure of substitute products, among other things.

The following sections describe the sectoral distribution of the number of bankruptcies in Portugal and France, as well as the evolution over time of this indicator between 2010 and 2016.

### 2.1.4.1. Portugal

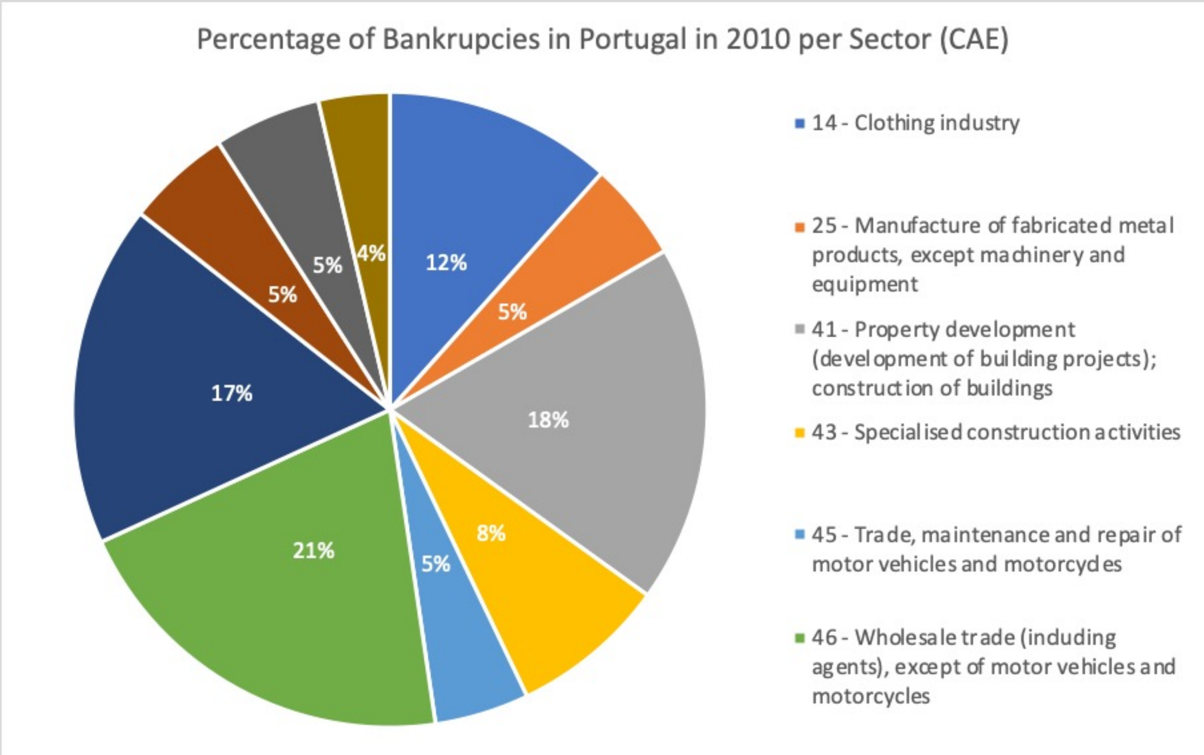



Figure 2.1 Percentage of Bankruptcies in Portugal in 2010 per Sector

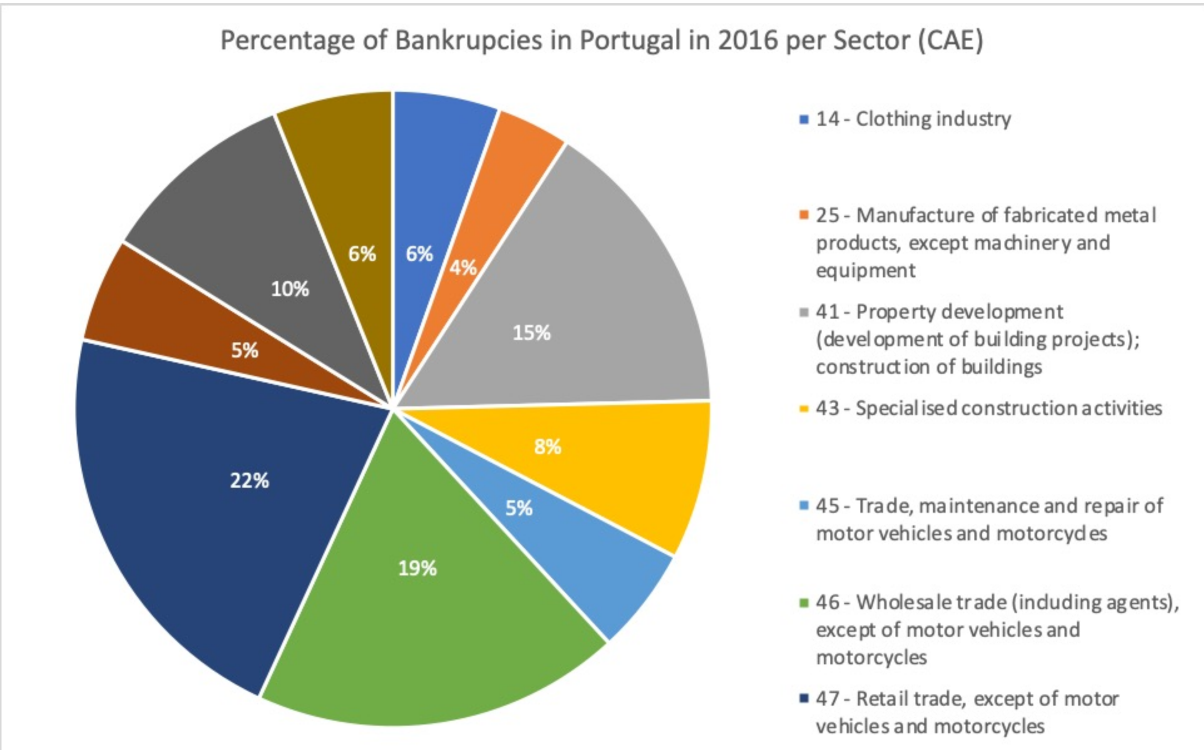



Figure 2.2 Percentage of Bankruptcies in Portugal in 2016 per Sector

From the analysis of the above figures, it can be concluded that between the beginning and the end of the period under analysis 2010 to 2016, there is a tendency to maintain the sectors with the highest incidence of bankruptcies.

The sectors with CAE 47, 46, and 41 in both 2010 and 2016 show the highest occurrence of bankruptcies.

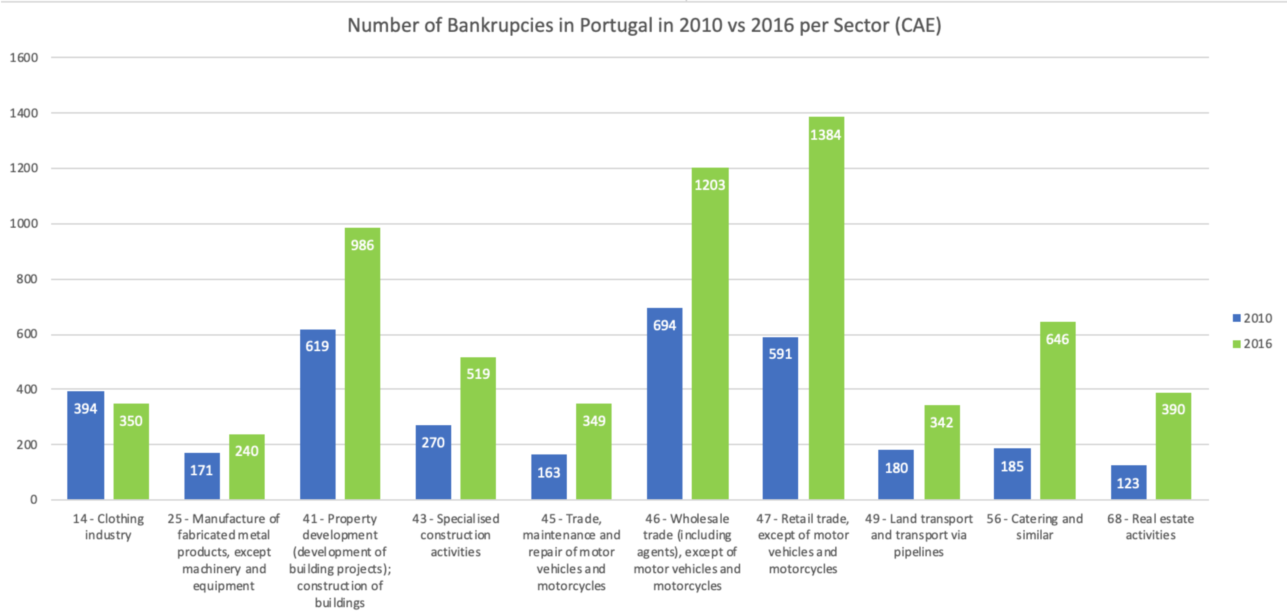



Figure 2.3 Number of Bankruptcies in Portugal n 2010 vs 2016 per sector (CAE)

The graph above shows an increase in the number of bankruptcies, with the exception of CAE 14. As this graph is expressed in absolute values, it is important to understand whether this increase stems from an increase in the incidence of bankruptcies, or whether it results from an increase in the number of companies in each sector, a situation that we evaluate in the graphic below.

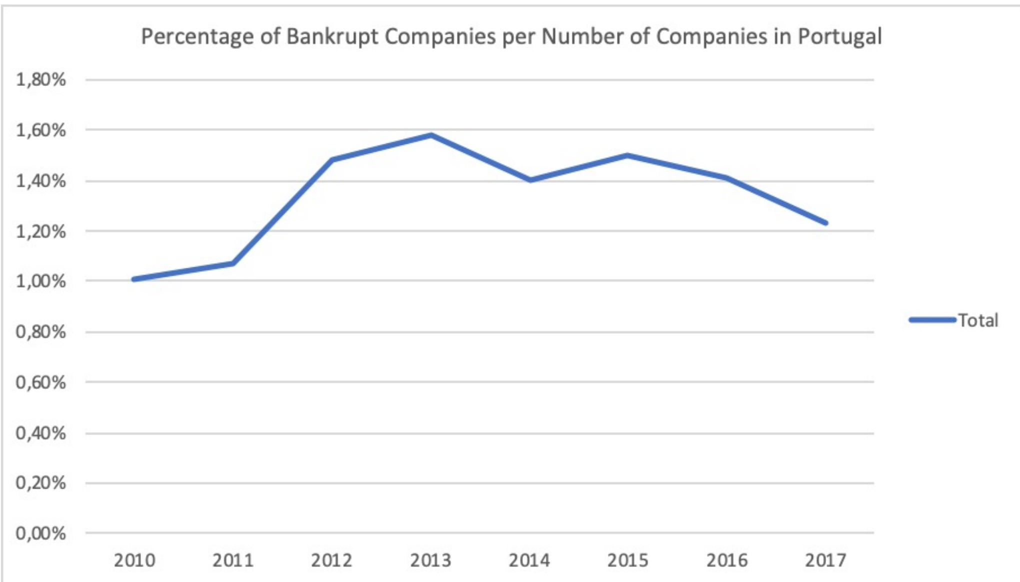



Figure 2.4 Percentage of Bankrupt Companies per Number of Companies in Portugal

In fact, evaluating the occurrence of bankruptcies weighted by the number of active firms in Portugal, it is concluded that the evolution of the number of active firms is significantly contributing to the increase in the number of bankruptcies. Even so, the fact that the number of bankruptcies increases in absolute value, having an unfavorable impact on the activity of the various sectors and consequently on the economy as a whole, justifies an increasing investment in the search for more efficient bankruptcy prediction models.

#### 2.1.4.2. France

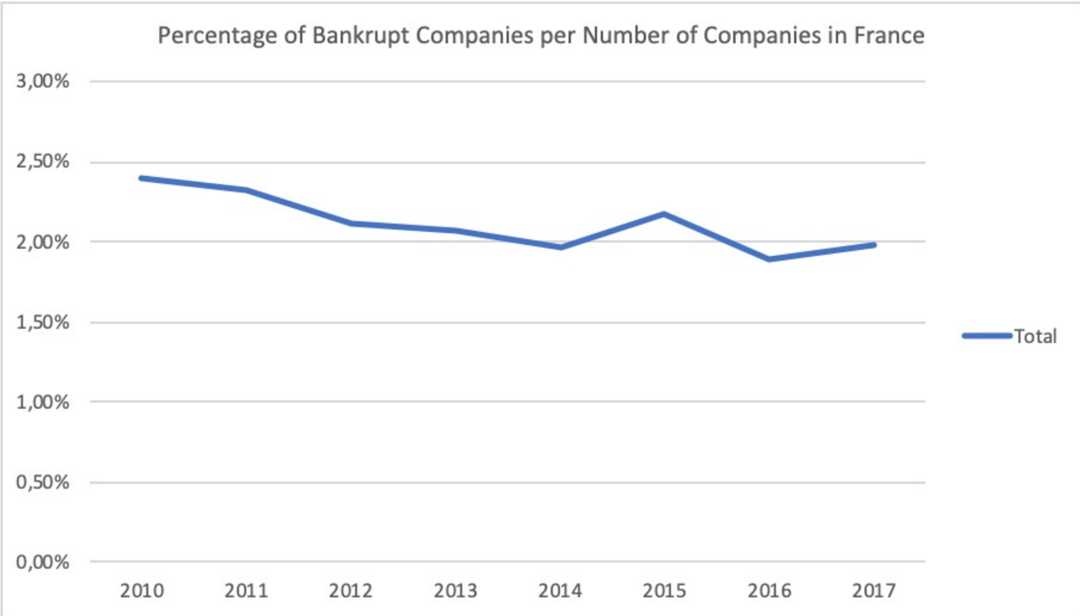



Figure 2.5 Percentage of Bankrupt Companies per Number of Companies in France

The same situation as in Portugal is observed in France, where there is an increase in the number of bankruptcies, although this is largely due to the larger number of active companies.

## 2.2. Bankruptcy Predictive Models

### 2.2.1. Introduction

The growing need to predict bankruptcy risk situations as a means of reducing the probability of default, in particular by creditors–including credit institutions has stimulated the development of a series of theories and models to predict the likelihood of bankruptcy between other institutions in a timely manner. With this, the risk of bankruptcy is one of the major topics for business and financial institutions in recent decades. With respect to classical theory, market imperfections and inefficient allocation of resources should be taken into account, which may result in economic regulation playing a major role in reducing bankruptcies.

In 1987 (and later revised in 1998), the Basel Committee proposed various steps, which became known as the International Integration of Capital Measurement and Capital Requirements or as Basel I.

Subsequently, the discovery of certain shortcomings in this and a later agreement (Basel II), combined with the start of the financial crisis (related to the well-known bankruptcies of the main US banks), stimulated the implementation of new steps, creating Basel III in December 2010: a global regulatory framework for ensuring stronger banks and banking systems.

Through the implementation of these reforms, financial institutions are projected to pursue, according to Gaspar (2014)", a rigorous credit policy that will allow them to mitigate the risk assumed against their clients throughout the life cycle of operations". Achieved within the Internal Rating Base (IRB) developed by Basel II, in conjunction with the objective of evaluating the probability of default and the estimated loss, allowing the provision of potential credit portfolio losses to be made with the last.

Regarding these issues, important contributions were made to the methods of predicting corporate bankruptcy.

### 2.2.2. The different types of Bankruptcy Predictive Models

The techniques used to extend bankruptcy prediction models are subdivided into three groups: Statistical models, Artificially intelligent expert system models (AIES), and theoretical models (Mehrazin, Taghipour, Ghabdian, & Soleimani, 2013). Statistical inference approaches have both a univariate and multivariate analysis, focusing on symptoms of failure. The main multivariate techniques are MDA, Altman's Z score, multidimensional scaling, logit analysis (Ohlson, 1980), probit analysis ((Zmijewski, 1984), Fischer's LDA (Fisher, 1936), cluster analysis, factor analysis, and logit–probit (D. Zhang & Zhou, 2004; G. Zhang, Hu, Patuwo, & Indro, 1999).

Among these contributions, we highlight the work developed by Beaver (1966) and Altman (1968) with, correspondingly, the univariate and multivariate discriminant analysis models that would later be re-tuned by the authors themselves as well as by many other researchers. These models had strict assumptions of linearity, normality, independence between predictor variables, and pre-existing functional types linked to criterion and predictor variables. These strict theoretical mathematical premises have kept their implementation limited to the real world. The availability of computers and technological advances motivated the development of technology-oriented models, especially since 1980.

Even then, long ago, AIES remerged as an alternative to the traditional statistical models in use. It was concluded with technical development that computers could mimic human-like cognitive intelligence behaviors in problem-solving. Accordingly, it sparked the search for programs that could fairly replicate such capabilities. And the field of information related to this issue started to emerge in 1950, having been called the computational "intelligence" of Artificial Intelligence (AI).

Consequently, where humans are able to use their intellect to solve problems by applying their knowledge and experience-based logic and reasoning. To approach human intelligence, AI must take advantage of common expertise in applying logic and reasoning to the presented problem, and Expert Systems (ES) have been developed to solve this problem. This category includes the models: Vector Machines (SVM), Neural Networks (NN), Case-based Reasoning, Decision Trees (DT), Random Forest (RF), among others (Bryant, 1997; Buta, 1994; Han, Chandler, & Liang, 1996; Kumar & Ravi, 2007; Laitinen & Kankaanpaa, 1999; Min, Lee, & Han, 2006; Odom & Sharda, 1990; Shaw & Gentry, 1988; K.-S. Shin, Lee, & Kim, 2005). In addition, Genetic Programming and Genetic Algorithms were studied as an approach to bankruptcy prediction, Varetto (1998) was the first person to present a model for the classification of bankruptcies based on a GA. Varetto proposed two distinct GA-based models, one of which is a linear model estimating the constant and variable coefficients of the discriminating function with a view to optimizing its discriminating power. The other is a model focused on rules that classify companies according to rules (GSR) that apply GA according to their respective discriminatory scores called genetic score.

In the following figures reprinted from the paper and research developed by M. Adnan Aziz and Humayon A. Dar (2006), it is possible to visualize and analyze the proportion of model categories and models developed for bankruptcy prediction by past studies until the paper publication.

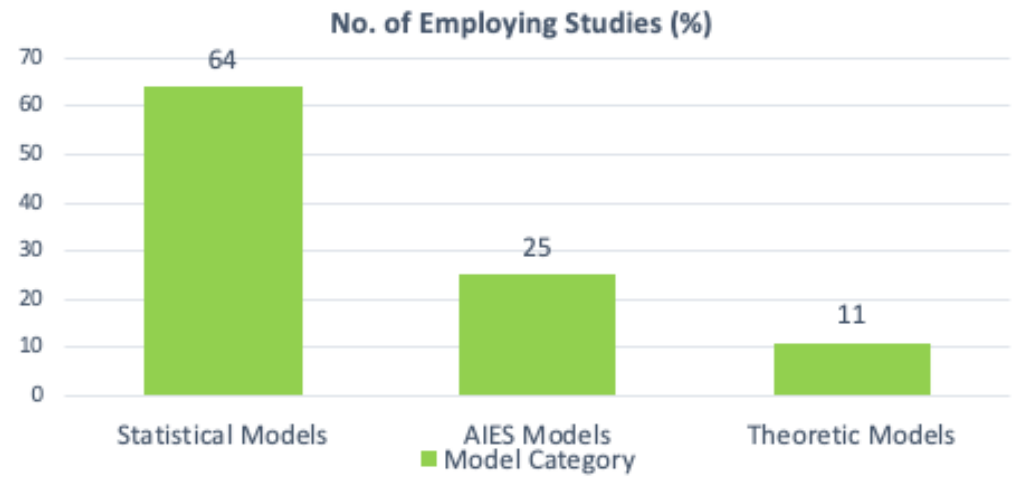



Figure 2.6 Proportion of model categories employed by past studies. Adapted from "Predicting corporate bankruptcy: where we stand?" by (Adnan Aziz & Dar, 2006)

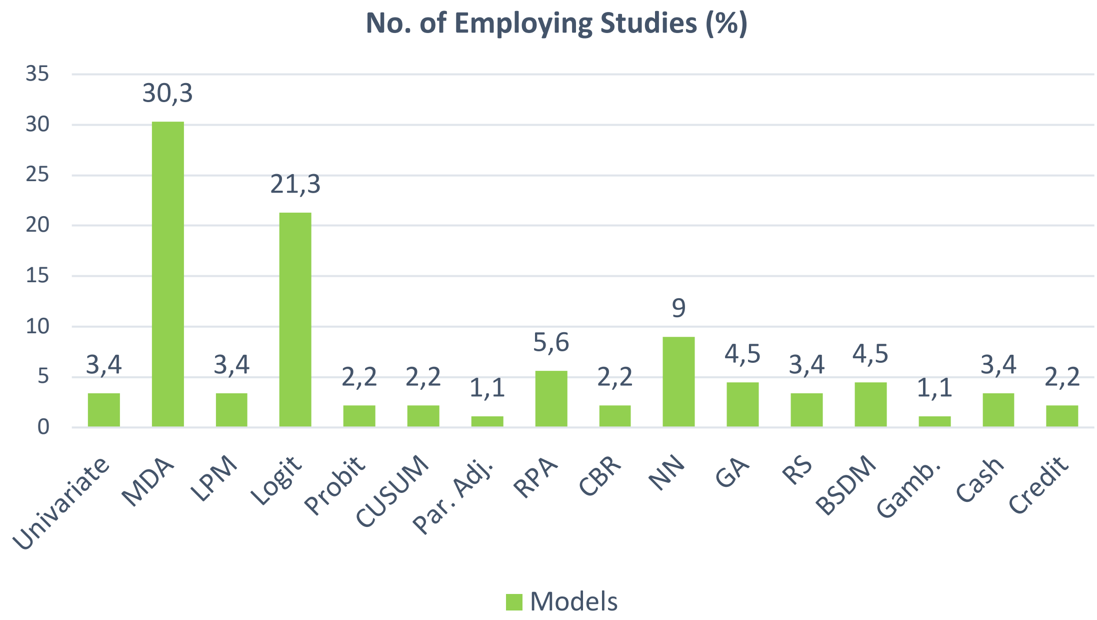



Figure 2.7 Proportion of models employed by past studies. Adapted from "Predicting corporate bankruptcy: where we stand?" by (Adnan Aziz & Dar, 2006)

Furthermore, after examining the figures above, it can be observed that Statistical Models are the predominant type of model used in past research. Furthermore, according to Figure 2.7, it can be discerned that the most studied models for prediction of bankruptcy are MDA and Logit, respectively, where both belong to the category of Statistical Models.

In addition, the value of NN, Recursive Partitioning (Decision Tree) Analysis (RPA), and GA models in past studies, which belong to the category of AIES models, can also be identified.

Subsequently, the importance of the models represented in the above figures can support the models studied in this dissertation, MDA, Logit, and GA.

### 2.2.3. Model Based on MDA

#### 2.2.3.1. Theoretical Context

Multi Discriminant Analysis (MDA) was among the first statistical techniques to be employed. It appears in an effort to statistically separate two or more classes of items using simultaneously a combination of several variables, where its use for financial analysis is based on the logical evolution of the Univariate analysis. Since its emergence in 1968, along with Edward Altman (1968), many researchers have focused and continue to focus on creating models to this day, making it the most studied technique in the corporate bankruptcy framework.

<u>a) Characteristics:</u>

Looking at the issue in a simple way, it simply consists of the "aggregation" of several univariate analyzes, each of which leads to the final assessment in a different way. In other words, from a pre-calculated set of indicators, we try to select, through regression, those that better capture the characteristics of the companies under review when combined and with the means to create a score or cut-off point that discriminates better against companies with reduced financial health from others.

<u>b) Historical Evolution:</u>

As already stated, it has been one of the most researched techniques since its introduction in 1968 and, over time, it has found multiple variants to its initial structure, from the Probit and Logit functions which transform the dependent variable into a continuous one and as such adapted to linear regression, not being subject to the assumption that the independent variables follow the normal distribution.

However, in addition to the multicollinearity problem, i.e., the presence of a relationship or connection between the independent variables, the variables also need a functional relationship between them.

This problem has also been solved recently by the use of the Partial Least Square Discriminant Analysis (PLS-DA) technique, which does not aim for the total variance between dependent and independent variables, but projects both in a new space, known as a bilinear factor model.

<u>c) Operational Mechanics</u>

The model is a linear process where discriminatory variables are combined in the following form: $Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + ...$ To $\beta_n X_n$. Where Z is a value transformed into a score used to classify the object, $\alpha$

is a constant, βs are discriminating coefficients or weights, and Xs are the values of independent discriminating variables (which correspond, in our case, to financial indicators).

#### 2.2.3.2. Edward Altman Model (Z-Score)

This model was developed by Altman (Edward I. Altman, 1968) and is one of the most significant models, recognized and used to this day, which combined multiple productivity and risk measures. After publishing his article "Financial Ratios, Discriminant Analysis, and Corporate Bankruptcy Prediction", Altman became the primary influence in the probability of bankruptcy. This position is still valid today, although there are significant recent contributions to the development of techniques for predicting bankruptcy, based on much more elaborate computational media and techniques, such as the application of neural networks to this area of research, among others.

In several contexts and markets, this model has proven to have a high predictive potential for bankruptcies.

**Sample Selection**

Altman used a selection of 66 listed industrial companies from each group (Group 1-bankrupt and Group 2-non-bankrupt) and was collected from "Moody's Industrial Manuals" and annual reports. The bankruptcies (Group 1) were registered at the National Bankruptcy Act under Chapter X, with the bankruptcy registration that took place between 1946 and 1965. Obviously, a sample representing a 20-year period does not equate to the best sample as it is such a period of evolution of the average ratio value that may have influenced the results obtained. Ideally, data from a t-period would have been used for the model collected, predicting the company's behavior in the t+1 cycle, but this was not feasible since the sample was difficult to obtain. The sample companies had assets ranging from $0.7 to $25.9 million, with an average value of $6.4 million.

Being aware that the set of bankrupt firms was not completely homogeneous given the differences in size and volume (asset value) of both the industry market, Altman carefully selected the non-bankrupt firms (Group 2) to be included in the study. The criterion corresponded to the collection of a paired set of industrial enterprises, chosen for stratification. The stratification was carried out by sector and by the size of the company, with a size between $1 and $25 million, with an average value of $9.6 million. Accordingly, the average assets of Group 2 companies were higher than in Group 1, but Altman considered that the differences in Average asset values were not a factor in model development.

**Variable Selection (Ratios)**

Prior to bankruptcy, the data were extracted from the financial statements, and the variables were grouped into five categories: liquidity, profitability, debt, solvency, and activity. Initially, 22 ratios were selected due essentially to two factors: prominence of literature and possible relevance to the research. To evaluate the ability to independently break down variables, these 22 ratios were performed several "F" tests (significance test). In other words, if, for example, the variable indicates that there are significant differences in that ratio between the a priori specified two groups of companies. Altman (1968) selected five ratios in his study:

- $X_1$ = working capital / total assets;

The value of this ratio is often used in businesses with financial difficulties, because according to Altman (Edward I. Altman, 1968), a corporation that is continuously incurring operating losses allows its current assets to be replaced by accumulated liabilities, decreasing the proportion of current assets over total assets, thus reflecting the liquidity of the company.

- $X_2$ = retained earnings / total assets;

$X_{2'}$ ratio value is an indicator that represents profit accumulation. For Altman, this measure is based on the company's age, i.e., a younger firm will have lower retained earnings than an older firm.

Thus, it implies that there may be companies that are wrongly classified just because they are younger when all else is unchanged.

- $X_3$ = earnings before interest and taxes / total assets;

According to Altman, the value of $X_3$ is the one that presents the true profitability of the company's assets, thus excluding tax and tax impact, and should be considered in studies related to bankruptcy prediction.

- $X_4$ = market value of equity / total liabilities;

The market value of equity is market capitalization (assuming this value reflects the company's correct value), and since this is the difference between Total Assets and Total Liabilities, this ratio tells us how much the company's assets will decline before the company goes bankrupt (i.e., its liabilities surpass their assets).

- $X_5$ = sales / total assets;

Finally, we have the ratio that indicates the company's capacity to generate sales based on their inventory, being a very useful indicator that the company's managers can accomplish their goals strategically.

Through the results of the significance tests performed by applying the F Test, comparing the difference between the mean values of each ratio in each group and the respective group variations, which the results are available in Table 2.2, it has been found that the ratio that best discriminates between business groups is the $X_2$ variable, i.e., the variable that most varies in value between bankrupt and non-bankrupt companies. On the other hand, the variable showing a much lower significance level than the others was $X_5$, which shows that it does not reflect very different values between the two business realities.

Table 2.2 Average of Variables and Significance Test (listed companies)

| Variable | Mean of Bankrupt Sample | Mean of Not Bankrupt Sample | F-Test |
|---|---|---|---|
| $X_1$ | -0,061 | 0,414 | 32,6* |
| $X_2$ | -0,626 | 0,353 | 58,86* |
| $X_3$ | -0,318 | 0,153 | 26,56* |

| Variable | Mean of Bankrupt Sample | Mean of Not Bankrupt Sample | F-Test |
|---|---|---|---|
| $X_4$ | 0,401 | 2,477 | 33,26* |
| $X_5$ | 1,503 | 1,939 | 2,84 |

* Significance at 0.001 level

Nonetheless, Altman sought to determine each variable's relative explanatory contribution in relation to the total capacity of discrimination provided by the feature and the interconnection between them. To this end, the variables have been modified to nullify the bias induced by the various units in which they are being expressed. According to the table below, there is a list of variables contributing the most to the function's capacity for discrimination:

Table 2.3 Relative explanatory contribution of the variables

| Variable | Scaled Vector | Ranking |
|---|---|---|
| $X_3$ | 9,89 | 1 |
| $X_4$ | 7,42 | 2 |
| $X_5$ | 8,41 | 3 |
| $X_2$ | 6,04 | 4 |
| $X_1$ | 3,29 | 5 |

Accordingly, Altman concludes that, contrary to the initial analysis, the variables $X_3$ e $X_4$ , and $X_5$ are the ones that contribute the most to discriminate between the different groups of companies. Consequently, according to Altman's original study (1968), the discriminating function for the companies listed is as follows:

$$\mathrm{Z} = 0{,}012\, X_1 + 0{,}014\, X_2 + 0{,}033\, X_3 + 0{,}006\, X_4 + 0{,}999\, X_5 \qquad (1)$$

Once the function was defined, the value of (Z) was set (at 2.675) according to the classification error minimization criterion. The value was thus set as the boundary of classification between bankrupt and non-bankrupt companies. Values above 2,675 correspond to a classification of a' non-bankrupt' company, and the lower values correspond to a classification of a' bankrupt' company.

In addition, Altman stated three rating categories where the value Z has lower and upper limits, i.e., if the value Z is below 1,81, then it is highly likely that the firm would fail. But if the Z value reaches 2,99 then the company has low bankruptcy risk rates, suggesting that the company is in good financial' health' If the value Z is between these two limits then the company does not have a well-defined tendency, i.e., it does not have an exactly lower or higher likelihood of bankruptcy, but there are points to be improved it is the so-called zone of ignorance."

Therefore, Altman decided to identify a region of ignorance, and the following "cut-off" scores were generated to ensure a level of safety for the ratings of healthy and unhealthy companies. With this interpretation of the cut-off scores created by Altman, the model's users ' confidence was strengthened.

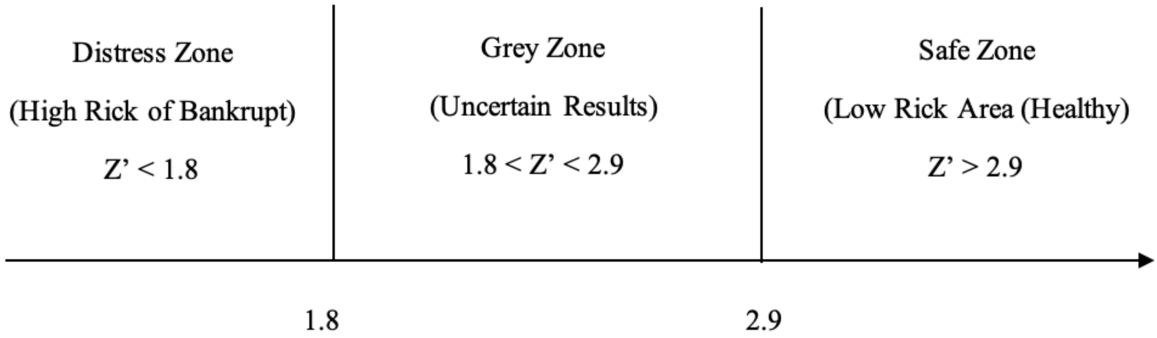



Figure 2.8 Cut Off Scores for listed companies

The results of the Altman (1968) study for one year before the bankruptcy revealed the following:

Table 2.4 Model Efficiency

| Actual Value | Predicted Value | |
|---|---|---|
| | Non-Bankrupted | Bankrupted |
| Non-Bankrupted | 97% (32 Companies) | 3% (1 Company)<br>(Type II Error) |
| Bankrupted | 6% (2 Companies)<br>(Type I Error) | 94% (31 Companies) |
| Accuracy: $\frac{(31+32)}{(33+33)} = 95\%$ | | |

The table 2.4 shows the performance of the model for the first year before bankruptcy. This has resulted in a success rate of 95s% of correct classifications. Nevertheless, it is important to note that there were 2/33 type I classification errors (6%), i.e., firms classified with a low level of insolvency risk that would actually bankrupt. This type of error is more serious than the type II errors, where the opposite is true.

Lastly, the results of the Altman study (1968) for two years before the bankruptcy revealed the following:

Table 2.5 Model Efficiency two year before bankruptcy

| Actual Value | Predicted Value | |
|---|---|---|
| | Non-Bankrupted | Bankrupted |
| Non-Bankrupted | 94% (31 Companies) | 6% (2 Company) (Type II Error) |
| Bankrupted | 28% (9 Companies) (Type I Error) | 72% (23 Companies) |
| Accuracy: $\frac{(23+31)}{(33+32)} = 83\%$ | | |

As for the model's effectiveness in the 2nd year prior to the bankruptcy, the results were also satisfactory, achieving an 83% success rate of correct rankings. It should be noted, however, that in this case, the most serious Type I classification errors correspond to 18%, being much higher than the less serious Type II errors, which amount to 6%.

Altman (1968) tested the model prediction up to five years before the bankruptcy, showing that as the number of years increases, the model's predictive ability has declined.

Therefore, it can be concluded that the model makes clear the potential bankruptcy that is approaching and, not being oriented towards an accurate analysis of the company's behavior in the future, it is a model that responds very accurately to us regarding the future behavior of the company, within a short time span (between 1 and 2 years).

With this research, other important insights were obtained cumulatively, namely and among others of less relevance the fact that all the observed ratios show a tendency to deteriorate with the approach of bankruptcy, most of the major changes in these ratios occur between the second and third years prior to the failure if the degree of severity is measured by annual changes in the values of the ratios.

Table 2.6 Summary of Model Results

| Nº of Years Before Bankruptcy | Nº of Companies | Nº of Correct Predictions | Nº of Incorrect Predictions | Accuracy |
|---|---|---|---|---|
| 1 | 33 | 31 | 2 | 95 % |
| 2 | 32 | 23 | 9 | 72 % |
| 3 | 29 | 14 | 15 | 48 % |
| 4 | 28 | 8 | 20 | 29 % |
| 5 | 25 | 9 | 16 | 36 % |

The Altman research (1968) had some limitations. One of the most important was that the model was only prepared to be conducted on listed companies, i.e., that had market values of their own equity. Therefore, Altman (Altman, 1983) modified the ratio $X_4$, where the market value of equity was replaced by the book value of equity as the market value is often not easily obtained.

The model is similar to the one developed previously (for listed companies), showing identical precision with regard to Type I and Type II errors. The function obtained was the following:

$$\mathrm{Z} = 0{,}717\, X_1 + 0{,}847\, X_2 + 3{,}107\, X_3 + 0{,}42\, X_4 + 0{,}998\, X_5 \qquad (2)$$

- $X_4$ = book value of equity / total liabilities

The results reveal three new groups of rankings in relation to the values for the Z-Score function. The financial solvency zone is now for Z-values above 2,90 and the financial insolvency zone for Z-values below 1,23. The "zone of ignorance" is now between 1,23 and 2,90, as it is possible to observe in the following figure:

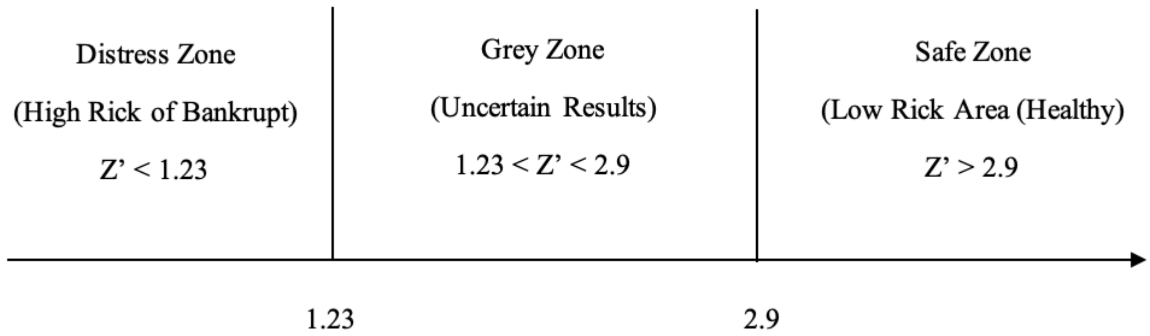



Figure 2.9 Cut Off Scores for non-listed companies

### 2.2.3.3. Other MDA models

**Univariate Models**

Beaver (1966) presented the first modern method to distress prediction. Using a matched sample (by sector and asset size) approach, using a curated sample of nonbankrupt companies.

Natural evolution contributed to the expansion of the univariate analysis when taking into account a variety of indicators. According to Bellovary, Beaver, in their suggestions for future investigations, "indicated the probability that multiple ratios considered simultaneously might have higher predictive potential than single ratios-and thus the development of predictive models of bankruptcy began" (Bellovary, Giacomino, & Akers, 2007, p. 4).

Thus, in 1968, Altman merged multiple indicators into a binary method, showing a powerful model. With the introduction of these approaches, which marked the beginning of the statistical study of "bankruptcy," many researchers have discussed and answered these questions. According to Sun, Li, Huang and He (2014), among others, the predictive capacity of the MDA over the previous year to "bankruptcy" is substantially higher than the univariate analytical model.

#### 2.2.3.4. MDA model's performance (Efficiency advantages and Disadvantages)

In a given study, applying multiple discriminant analysis involves certain assumptions, that is, there are requirements that must be met.

These requirements include the normality of the variables, the homogeneity of the variance-covariance matrices, the existence of significant differences between the groups, and the removal of outliers.

We know that the existence of normal economic and financial variables is rare, and therefore hard to "manipulate." As regards the parity of covariance variance matrices, this must be fulfilled, as it is a consequence of the nature of the linear function. It is important to choose the most significant variables, and it is a process that can be performed using various statistical techniques, as it will influence the results obtained. A statistical test can often not be validated in the presence of outliers due to abnormal data that misrepresent the mean value, and their existence can hinder the interpretation of the results obtained from a given sample.

Another disadvantage is that groups need to be defined a priori, which means that it is necessary to know which year precedes the bankruptcy in order to apply the discriminating analysis, because if we want to study the company's future with regard to its possible failure, we cannot do so unless we know the year preceding it.

Consequently, it should be noted that Altman's Z-Score model (1968) has its drawbacks, such as the fact that qualitative data cannot be included, i.e., the financial data considered do not represent unexpected events which may occur in the company's operations and which are often not reflected in the financial statements and also should be noted that the ratios were chosen on the basis of statistical significance and literature popularity and not by a correct correspondence with the reality of the country and companies.

However, different economic agents still use the model the most. Some of the key features contribute to Z-Score are: simplicity, a methodology that is statistically robust, simple to understand, and an efficient trade-off between data volume vs performance.

### 2.2.4. Logistic Regression

#### 2.2.4.1. Introduction

The first studies on logistic regression models emerged at the end of the decade of 70 and presented themselves as an attempt to overcome the known limitations that affect the discriminant analysis. In particular, this analysis assumes that the explanatory variables have a normal multivariate distribution with different means but matrices of equal dispersion. However, if all variables do not present a normal distribution, the methods employed may result in an inappropriate selection of all predictors

Sheppard (1994, p. 10) and Ohlson (1980) recognizes that financial ratios rarely follow a normal distribution, believing therefore, that one should abandon the presumption of normality of error distribution underlying bankruptcy predictive models based on this methodology. Accordingly, Ohlson (1980) supports the use of Logit models to the detriment of multivariate discriminant analysis models taking into account the limitations of the former methodology.

#### 2.2.4.2. Logit

The Logit methodology uses the estimation by the Maximum Likelihood Method, in other words, it is an algorithm that allows the model's coefficients β to be estimated, maximizing the natural logarithm of the likelihood function. Lo (1986) compares this method to the discriminant analysis, stating that the parameters estimated by the Logit model are more robust than those estimated by the multivariate discriminant model based on the analysis. With regard to this methodology, the dependent variable is defined as a binary variable that takes zero or one of the values.

In the case of predictive bankruptcy model estimation, typically zero is associated with companies that do not enter into a bankruptcy situation, and value one is assigned to bankrupt businesses.

In regression model logistics, therefore, the relationship between a company's probability of bankruptcy (P) and the value of the ratios in a given year (X) is an S-shaped curve that varies between zero and one, from which the following expression is obtained:

$$P(y_t = 1) = \frac{1}{1 + e^{-z}i} \tag{3}$$

Where Z is represented by the above linear relationship:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m \tag{4}$$

According to the following representation:

P = probability of bankruptcy;

X = financial ratios (vector with the values of the explanatory features);

i = number of years observed;

β = coefficients to estimate (vector of unknown parameters that reflects the impact of the explanatory variables on the probability of the company being "Healthy" or "Unhealthy")

In other words, we can also determine:

$$P(y_t = 1) = \frac{1}{1 + e^{-x}i\beta} \tag{5}$$

The probability of bankruptcy is determined by their coefficients, which are obtained by linear regression, as the function of the economic and financial ratios, and an index Z can be calculated, which, transformed using the previous expression, provides a certain likelihood of bankruptcy (P).

#### 2.2.4.3. Past Logit Models Research regarding Bankruptcy Prediction

**Ohlson Model**

Ohlson (1980, pp. 109–119) uses a logistic regression model to examine the effect of four basic factors in calculating the probability of failure: size, financial structure, performance, and liquidity.

This study is similar to others, where "...the methodology is the maximum likelihood estimator of the model designated” as "the conditional logit model" (Begley et al., 1996, p. 273).

It was possible to identify four basic factors as being statistically significant for the determination of the probability of bankruptcy of firms (one year before). These factors are size, financial structure, the performance, and current liquidity, hence validating the author's initial hypothesis.

The sample evolved 105 bankrupt firms and 2058 firms in the situation of non-bankruptcy with regard to accounting data for the years 1970 to 1976.

Three models have been developed: a first model that predicts bankruptcy one year before bankruptcy, a second model that predicts bankruptcy two years before, and a third model that predicts bankruptcy one year or two years before the bankruptcy.

These models do not consider any market variables (e.g., market capitalization) but include the size of the company.

As they measure the probability of bankruptcy (conditioned by the economic and financial ratios), their values can only range between 0 and 1; [0.1], with low values indicating financial strength and high values indicating weakness and a consequently higher likelihood of bankruptcy.

Ohlson (1980) assumed that the errors of an incorrect classification are the same for both groups of firms (bankrupt vs. non-bankrupt). In this context, it defined a cut-off point of 0,5, where a bankrupt firm would be classified as having $P(X_i, \beta) > 0{,}5$ and non-bankrupt if $P(X_i, \beta) < 0{,}5$. By implementing this principle, Ohlson (1980) achieved the correct classification of 96,12% of firms in Model 1, 95,55% for Model 2, and 92,84% for Model 3.

Furthermore, the author concludes that the model's predictive power depends on the timing of the data obtained regarding the bankruptcy event and that further explanatory variables would be desired for a significant improvement of the model.

**Zavgren Model**

Zavrgen (1985) initially criticized Ohlson (1980), particularly in terms of the definition of the model and the variables selection. Moreover, this author is skeptical about the non-use of the paired sample method and the fact that the model error rate was calculated from the sample used for its estimation.

For this research, Zavgren (1985) used the Logit methodology for a period of five years before the bankruptcy. The sampled companies involved 45 bankrupted and 45 non-bankrupted companies of a comparable sector and size belonging to the New York Stock Exchange and over the counter (OTC) market for the period 1972 to 1978. The variables considered for the model essentially involve liquidity, investment, and financial ratios.

The model has managed to be quite significant in terms of the probability of detecting firms in financial distress up to five years before the bankruptcy, resulting in the correct classification of 82% of firms one year before bankruptcy, 83% of firms two years before, 72% of firms three years before, 73% of firms four years before and 80% five years before.

### 2.2.4.4. Advantages and Disadvantages of Logit

In this regard, Ooghe and Balcaen (2004) present the advantages and limitations of the Logit model. As far as the advantages of the model are concerned, these are briefly summarized below:

- does not assume a linear relationship between the dependent and independent variables;
- does not require the variables to follow a normal distribution;
- accepts qualitative and quantitative variables as explanatory, that is, non-financial information can be used in the model;
- is more robust than discriminant analysis, since other than the normal distribution is also applicable;
- the dependent variable can be interpreted as the probability of the firm going into insolvency;

On the other hand, Ooghe and Balcaen (2004) identify the following disadvantages in the application of Logit models:

- it is mandatory that the groups are separated since this technique does not define that threshold, only possible in the discriminant analysis (dichotomous dependent variable);
- the probability of failure follows the logistic distribution and ranges between [0, 1];
- it is mandatory that the explanatory variables are independent;
- there could exist multicollinearity among variables;

### 2.2.5. Genetic Algorithms

#### 2.2.5.1. Definition and Application of Genetic Algorithms

A Genetic Algorithm (GA) is a random-based classical evolutionary algorithm by natural selection, which is a search heuristic that mimics the process of natural evolution (Davis, 1991; Goldberg, 1989; Holland, 1975) being a part of the class of Evolutionary Algorithms (EA). Having as has its core idea from Charles Darwin's theory of natural evolution "survival of the fittest," spired by the biological evolution principle of survival of the fittest.

Occasionally GA can be called Simple GA (SGA) because of its simplicity compared to other EAs. In 1992, Jonh Koza used GA to develop programs in order to perform a certain task, which is called Genetic Programming (Koza, 1992)

GAs are distinct from many conventional search algorithms in the following ways (Karr, 1995):

1. GAs do not find a single point but several search space points that concurrently minimize the probability of converging to local optima;
2. GAs work directly with strings of characters representing the parameter set, not the parameters themselves;
3. GAs use probabilistic rules, not deterministic rules, to guide their search.

While traditional calculus-based methods start at a random point and move in the gradient direction until the top of the hill is reached, this is an effective technique that works very well for single-peaked objective functions such as cost function in linear regression.

For most real-world situations, it is possible to observe very complex problems called landscapes, composed of many peaks and many valleys, which cause these methods to fail, thus suffering from an intrinsic propensity to become stuck in the local optimum (Shalev-Shwartz, Shamir, & Shammah, 2017) as shown in the following figure:

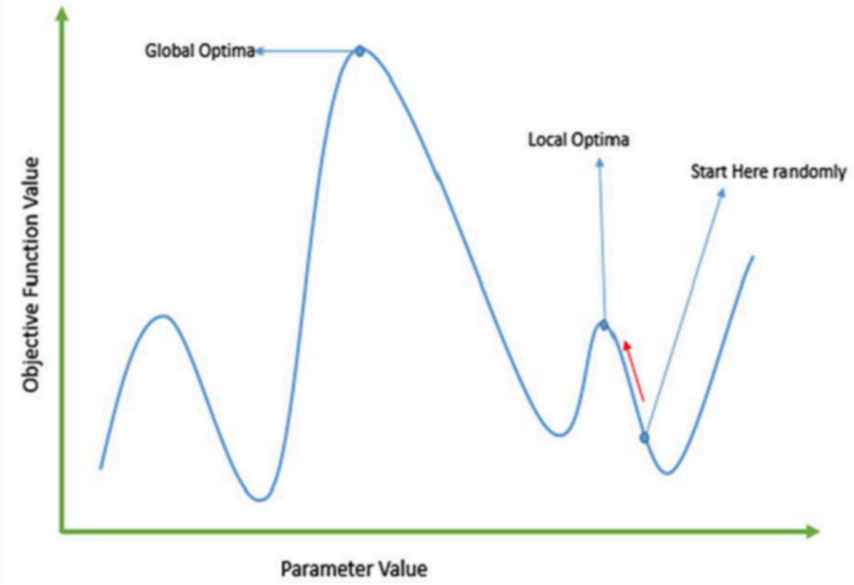



Figure 2.10 Objective Function and Parameter Value ("Genetic Algorithms - Introduction - Tutorialspoint," n.d.)

GAs prove to be an efficient tool for providing near-optimal, usable solutions in a short time and have been demonstrated, as well effective and robust in searching very large spaces in a wide range of applications (Colin, 1994; Koza, 1992; K. Shin & Han, 1999).

Consequently, this advantage has been demonstrated in several numbers of applications in the trading system (Colin, 1994; Deboeck, 1994), stock selection (Mahfoud & Mani, 1995), portfolio selection (Rutan, 1993), bankruptcy prediction (Kingdon & Feldman, 1995; K.-S. Shin & Lee, 2002) and credit evaluation (K. Shin & Han, 1999).

**Basic Terminology**

There is a particular basic terminology which is used throughout this dissertation:

- **Individual** – Any possible solution
- **Population** – A subset of all the possible solutions (encoded) for the given problem. The population for a GA is equivalent to the population of humans except that there are Candidate Solutions of the entities instead of human beings.
- **Chromosomes** – A chromosome is one of the solutions to the given problem.
- **Gene** – One element position of a chromosome.
- **Allele** – Value that a gene takes for a particular chromosome.

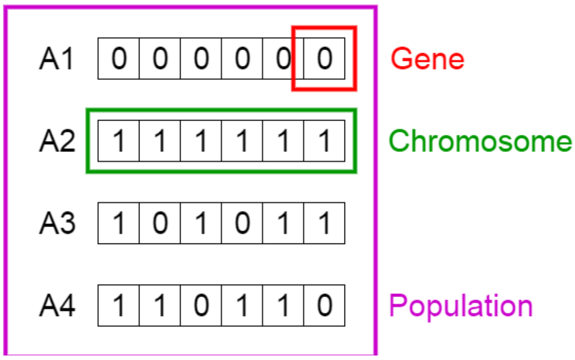



Figure 2.11 Gene, Chromosome and Population Example

- **Genotype** – Population in the computational space, where the solutions are interpreted in a way that can be easily understood and manipulated using a computing device.
- **Phenotype** – Population in the real-world solution space in which solutions are described in a way that is demonstrated in circumstances of the real world.
- **Decoding and Encoding** – In simple problems, the spaces of the phenotype and genotype are the same. Though, the spaces of phenotype and genotype are distinctive in many cases. Decoding is the process of transforming a solution from the genotype to the space of the phenotype, whereas encoding is a process of transformation from the phenotype to the space of genotype. Note that decoding has to be prompt as it is done repeatedly in a GA during the calculation of the fitness value.

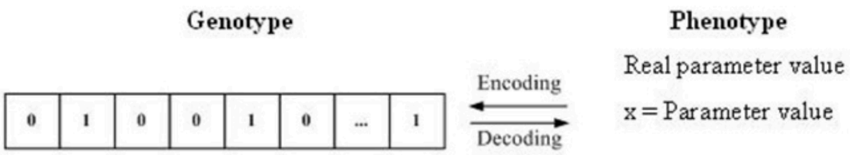



Figure 2.12 Genotype and Phenotype transformation

**Basic Structure**

GA is used to produce high-quality solutions for optimizing and searching being suitable for multi-parameter optimization, with hard and soft constrains problems for which they use bio-inspired operators such as mutation and crossover.

Furthermore, GA enables to have a number of potential solutions for any problem, perform the search process in four stages: initialization, selection, crossover, and mutation (Davis, 1991; Wong & Tan, 1994).

Search space is the set of all potential solutions and values that can be taken from the input. In optimization, it is pursued to find the point or set of points inside this search space, which will give the optimal solution. Each individual is like a string of characters/integers/floats, and the strings are like chromosomes.

For each individual (candidate), the fitness value (from a fitness function) indicates how close it is to the optimal solution. This is in line with Darwin's 'Survival of the Fittest' principle, which is how to continue to generate better (evolving) individuals/solutions over generations until is reached a threshold where to stop. The algorithm works, taking into account four major rules:

1. Individuals in population compete for resources, mate;
2. Fittest individuals' mate to reproduce more offspring than others;
3. Fittest parent propagates genes through generation;
4. Each successive generation evolves to suit its ambiance;

Since the size of the population is constant, some people will die to make room for the younger ones. With this, a convergence situation is reached where the gap between offspring created by current populations and ancestral ones is no longer significant. Then, the algorithm converges to a set of problem solutions.

There are several representations available for the chromosome, although the selection of the proper representation is problem-specific since it has been observed that improper representation can lead to the poor performance of the GA.

The best representation and having an appropriate definition of the mappings between the phenotype and genotype spaces is crucial for the success of a GA, making the search space reduced and thus easier to search.

In the following figure, it is possible to observe the basic steps of GAs:

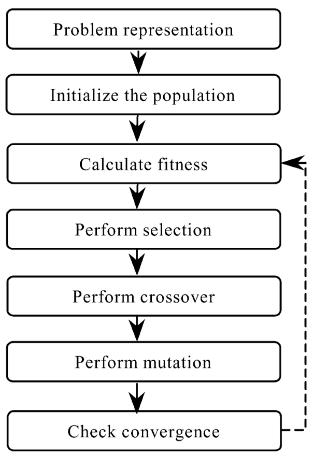



Figure 2.13 Basic steps of Gas, Adapted [reprinted] from (K.-S. Shin & Lee, 2002)

A population of genetic structures (chromosomes), randomly distributed in the solution space, is chosen as the starting point of the search during the initialization process.

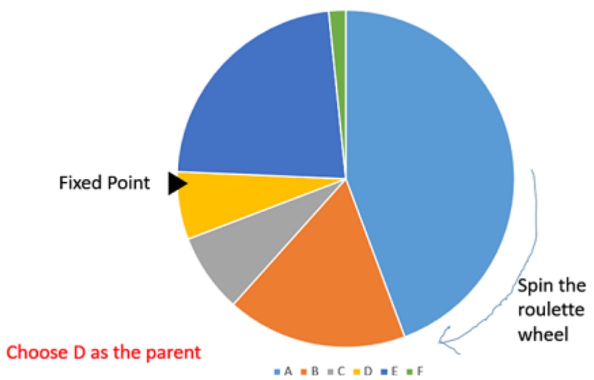
Each chromosome is evaluated following the initialization stage according to the user-defined fitness function. For each solution, the value of the fitness function in the genotype universe is determined. This function has the role of numerically encoding the chromosome 's performance. Choosing the fitness function is the most critical step for real-world problems of optimization methods such as GAs.

According to the mating properties of reproduction, only the high scoring members will preserve and propagate the worthy characteristics from generation to generation, thus helping to sustain the search for an optimal solution, so that, for example, binary strings with higher fitness values are more likely to be selected as parents.

Consequently, high-performance chromosomes may be selected several times for replication, while low performing structures will not be selected at all. This selective process originates from the population's best-performing chromosome for conquering an increasingly large proportion of the population over time. Some of the existent selection operators are:

### a. **Roulette Wheel Selection**

In this selection operator, the circular wheel is separated, and a static point is chosen on the wheel circumference as presented. After this process, the wheel is rotated, and the section of the wheel, which occurs in front of the fixed point is chosen as the parent. For the second parent, the equivalent process is reiterated.




| Chromosome | Fitness Value |
|---|---|
| A | 8.2 |
| B | 3.2 |
| C | 1.4 |
| D | 1.2 |
| E | 4.2 |
| F | 0.3 |

Figure 2.14 Roulette Wheel Selection, (“Genetic Algorithms - Parent Selection - Tutorialspoint,” n.d.)

Besides the probability of selecting an individual depends directly on his fitness, as seen in the figure, it is obvious that a fitter individual has a larger pie on the wheel and therefore a greater chance of landing before the fixed point after spinning the wheel.

### b. **Rank Selection**

Rank Selection is often used when individuals within the population have very close fitness values (this usually happens at the end of the run). This also works with negative fitness values and leads to an almost equal share of the pie being taken by each individual. Consequently, no matter how fit they are compared to each other, each individual has almost the same probability of being chosen as a parent. Consequently, turning leads to a loss of the selection pressure towards fitter individuals, which makes the GA poorly selected parents in such situations.

### c. **Random Selection**

There is no selection pressure regarding fitter individuals in this strategy, and therefore this strategy is usually avoided since the existing population selects parents randomly.

### d. Tournament Selection

K individuals are randomly selected from the population in the selection of tournaments and the best out of these to become a parent. Repeat the same procedure for selecting the subsequent parent.

In literature, Tournament Selection is extremely popular as it has the advantage of working with negative fitness values.

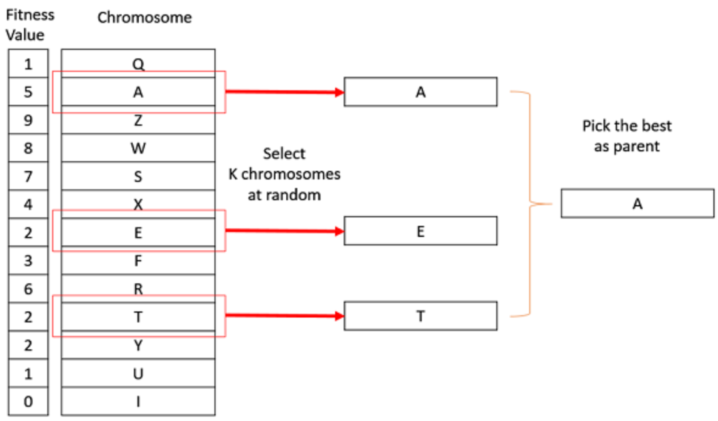



Figure 2.15 Tournament Selection Example,(“Genetic Algorithms - Parent Selection - Tutorialspoint,” n.d.)

### e. Elitism Selection

A small proportion of the fittest applicants in the Elitism Selection are recycled, unchanged, into the next generation. Occasionally, this selection can impact the performance by certifying that the GA is not wasting time re-discovering partial solutions previously discarded.

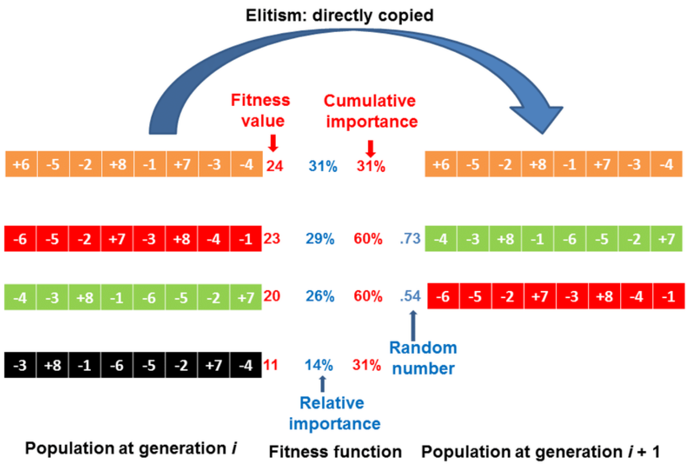



Figure 2.16 Selection process using an elitism function. Adapted [reprinted] from (Romero-Hdz, Aranda, Toledo-Ramirez, Segura, & Saha, 2016)

The crossover produces a new offspring of two successful parents who have been randomly picked. The crossover operates by swapping related segments of the parents' string representation and extends the search for a new solution, thus producing a completely new individual. The crossover occurs with a certain likelihood, called the crossover rate. From the single point, the two-point, to the uniform type, several different forms of crossover can be carried out (Syswerda, 1989)

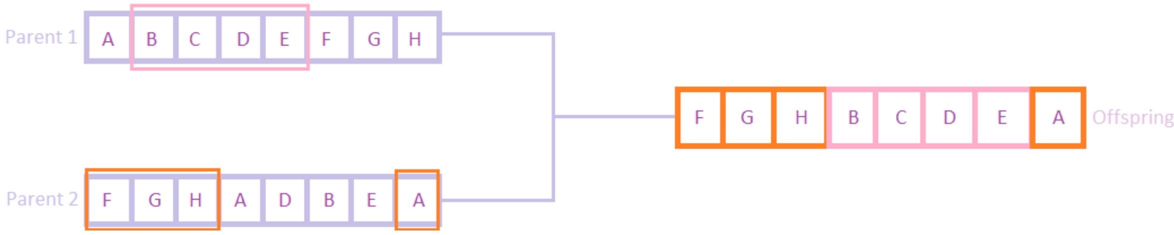



Figure 2.17 Crossover Operation in Python Genetic Algorithms, (Gour, n.d.)

The mutation is a GA operation that selects a random member of the population and alters a randomly selected bit in its representation of the bit string. Although the reproduction and crossover generate numerous new strings, crossover operations do not introduce new information about the bit level into the population. If the mutant member can be attained, it swaps the member who has been mutated in the population. The presence of mutation guarantees that the likelihood of reaching any point within the search space is certainly not zero and that diversity is maintained to avoid premature convergence.


Figure 2.18 Mutation Operator in Python Genetic Algorithms (Gour, n.d.)

**Genotype Representation**

Some of the representations available for the chromosome are:

- **Binary**: One of GA 's simplest and most frequently used representations, where each chromosome is represented as a string of zeros and zeros.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |

Figure 2.19 Binary Representation

- **Permutation**: Advantageous for ordering issues such as traveling salesman problem, where an order of elements represents the solution.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 9 | 8 | 7 | 4 | 2 | 3 | 6 | 0 |

Figure 2.20 Permutation Representation

- **Value**: The actual value is encoded as it is, where a valued number can be real or represented as an integer. Genes use continuous rather than discrete variables for real valued numbers, where the actual valued representation is the most common, but the accuracy of these real-valued or floating-point numbers is restricted to the machine. Often, we can't always restrict solution space to binary 'yes' or 'no' for discrete-valued genes. For example, if the four distances-North, South East and West-need to be encoded, they can be encoded as {0,1,2,3}. In such cases, the representation of the integer is advantageous.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 0.2 | 0.6 | 0.8 | 0.7 | 0.4 | 0.3 | 0.2 | 0.1 | 0.9 |

Figure 2.21 Floating-Point Value Representation

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 3 | 2 | 4 | 1 | 2 | 1 |

Figure 2.22 Integer Value Representation

**Limitations of Genetic Algorithms**

With all those benefits, Genetic Algorithms also have certain limitations:

- GAs are not suitable for mere derivative information problems
- Stochastic; no guarantee of an optimal result solution
- Repeated fitness calculation is computationally expensive for certain problems
- If not implemented properly, there is no guarantee of convergence to an optimal solution.

**Applications of Genetic Algorithms**

Genetic Algorithms have several other applications, such as:

- Recurrent Neural Network
- Mutation testing
- Codebreaking
- Filtering and signal processing
- DNA Analysis
- Economics

### 2.2.5.2. Past Genetic Algorithms applications for Bankruptcy prediction

Regarding the Genetic Algorithms applications for Bankruptcy prediction, Varetto (1998) was the first person to study this topic by suggesting a GA-based bankruptcy classification model. Varetto has proposed two separate GA-based models, one of which is a linear model that calculates the discriminating function's constant and variable coefficients in order to maximize its discriminating power using GA. The other is a rule-based model, named GA Genetic Score (GSR), which categorizes companies according to their respective unequal scores.

In 2002, Shin & Lee (2002) Applied GAs to derive rules that could predict financial collapse, where they made the first attempt to explore the ability of genetic-based systems to systematically resolve predictive bankruptcy problems. As a result, the rule extraction method using GAs for model-based prediction of bankruptcy was demonstrated to be successful.

Their study, nevertheless, had several limitations: initially, while several rules were extracted using conventional GAs, expanding the GAs using the nesting method was appropriate (Mahfoud & Mani, 1995); Second, the current structure of the rule had been quite small. They proposed that their system would be substantially expanded as a next research step by adding additional features, with more detailed features that would likely lead to improved results.

Continuing the research of GA and GP application on bankruptcy prediction, Lensberg, Eilifsen andMcKee (2006) studied Norwegian bankrupt and non-bankrupt firms and appointed six variables extracted from various analyses of past bankruptcies and fraud. The selected model based on genetic programming was presenting some interesting new features in this research. One such is that an unfavorable audit report has a more negative impact on the insolvency status of a big firm than a small one. The model also shows that the willingness to pay interest has a more favorable impact on the bankruptcy status of big companies than that of small businesses.

This could be interpreted as meaning that the model suggests that accounting information (including auditors' assessment) is more relevant to bigger firms than smaller ones. It also indicates that information relating to liquidity and non-accounting is the most important information for small companies.

Sikora and Shaw (1994) created an ANN-GA hybrid model, which derived if-then rules for predicting bankruptcy. Moreover, GA also has been used to optimize the parameters of support vector machine for predicting. Consequently, Min, Lee, & Han (2006) using a real data set containing bankrupt and non-bankrupt Korean firms, proposed methods for improving the performance of SVM (GASVM model), on the prediction of bankruptcy in two respects: the selection of subset features and optimization of the parameters. Additionally, GA was used to simultaneously optimize both the subset of features and SVM parameters for bankruptcy prediction.

Furthermore, Wu, Tzeng, Goo, & Fang (2007) used GA to optimize two support vector machine parameters (C and r) for bankruptcy prediction by constructing the most powerful support vector machine (SVM) model. Accordingly, an additional genetic-based SVM (GA-SVM) model was created that could automatically evaluate SVM's optimal parameters, C and r, achieving at the same time the highest predictive accuracy and generalization efficiency.

Extending the research of GA application on bankruptcy prediction, Ahn & Kim (2009) published a new case-based hybrid (CBR) and GA model, called GA – GOCBR. The proposed model simultaneously optimizes the weighting of the features and selection of instances. It can reduce noises or skewed cases that lead to erroneous prediction by selecting optimal instances. In addition, our model can also find suitable nearest neighbors for CBR by adding optimal feature weights to similarity calculation, which can improve predictive accuracy.

Another research study the use of GA to generate a set of rules based on the tests derived from the signs and cut-off values of the selected ratios and, in this regard, Garkaz and Abdollahi (2010), recommended a rule that would induce the model to optimize its predictive power by applying GA, in which each person indicates a potential solution to the previously identified problem.

Equally important, Kim & Kang (2012) Studied ensemble selection of classifiers using genetic algorithms to predict bankruptcy, which was proposed as a coverage optimization algorithm to solve multicollinearity problems and improve the stable efficiency of the ensemble. The proposed algorithm utilized GA to construct an ensemble that includes different classifiers in the optimization coverage process.

In this research, the optimal (or near-optimal) classifier subset was selected based on predictive accuracy, and the measurement of diversity defined as a statistical value of variance was used as a measure to improve the performance of an ensemble, and also to measure multicollinearity as a degree of diversity to select different classifiers, which is the objective of optimizing coverage. Therefore, predictive accuracy is used as a fitness function and as a GA search constraint to eliminate the high correlation between the classifiers to ensure the diversification of classifiers.

Furthermore, for the period 2006–2014, Bateni & Asghari (2020) Used 174 bankrupt and non-bankrupt Iranian companies listed on the Tehran stock exchange to set a standard for the GA classification. Where predictive bankruptcy models' efficacy was contrasted by identifying conditions under which a model performs better. In summary, it was concluded that data from the financial statements had a high predictive capacity, and both logit and GA models recommended sales to total assets and EBIT as the most significant variables in the prediction of bankruptcy. Moreover, it was found that the models can be used on the Tehran Stock Exchange, and the classification performance of the GA model is substantially higher than that of the logit model.

According to experimental findings, it has been proposed that the super-efficiency logit model and overall GA models are lower in forecasting bankrupt firms correctly than non-bankrupt firms. Furthermore, the most important external, uncontrollable factors (by companies) that contribute to financial distress in Iran are economic instability and political variables. High production costs, interest charges, and development bureaucracy are the key factors that cause bankruptcy in the country (Hsieh, 1993).

#### 2.2.5.3. Potential of Genetic Algorithms for Bankruptcy prediction

Because of the difficulty and the variety of solutions regarding the prediction of bankruptcy, genetic algorithms are a common method for solving these problems with local search operations, helping to avoid the existence of a single local optima solution using cross-over and mutation operators.

Consequently, using probabilistic search techniques, they can easily search a wide and complicated search space for an optimal or near-optimal solution.

In several studies and research on this issue, the potential of the genetic algorithm has been reflected: "The results show that rule extraction approach using GAs for bankruptcy prediction modeling is promising" (K.-S. Shin & Lee, 2002) ;

"Recent research using data from US companies has shown that genetic programming is extremely powerful and can be used to produce a simple but feature-rich model that provides new insights into the prediction of bankruptcy and, thus, the creation of bankruptcy theory" (McKee & Lensberg, 2002);

"Genetic algorithms (GAs) are popularly used as an effective tool to solve such local search operations. GAs can prevent local optima by using cross-over and mutation operators and can search a vast and complicated search space rapidly to find an optimal or near-optimal solution using probabilistic search methods" (Kim & Kang, 2012).

### 2.2.5.4. Genetic Algorithms performance for Bankruptcy prediction

Regarding the performance of GA on bankruptcy prediction, the genetic programming developed by Lensberg, Eilifsen, & McKee (2006) can be considered to be a highly accurate predictive model of bankruptcy, taking into account both the characteristics of the predominantly non-public companies in the study and the predictive period of up to 18 months. In this study, genetic programming based on six variables was generated from a large set of 28 variables that had been important in previous multiple prior studies linked to bankruptcy and fraud. This model was more accurate than a conventional logit model using the same variables, achieving 82% and 81% accuracy on the 900 firm training samples and 236 firm validation samples, respectively, while the two logit models built using the same six variables were only 77% and 76% accurate on the 900 firm training samples and 236 firm validation samples. The most important variable in the final model was the auditor's prior opinion, thus validating the auditors ' report knowledge meaning.

Furthermore, the GA-SVM model proposed by Wu et al. (2007), which was tested on the prediction of the financial crisis in Taiwan, compared the accuracy of the proposed GA-SVM model to that of other models in multivariate statistics (logit and probit) and artificial intelligence (NN and SVM). Accordingly, experimental results showed that the GA-SVM model achieved the best predictive accuracy, suggesting a very successful integration of the RGA with the conventional SVM model.

Moreover, Kim & Kang (2012) evaluated their proposed model using a benchmark data set obtained from one major commercial bank in Korea. The benchmark data set comprises 1,200 manufacturing firms independently audited, half of which went bankrupt between 2002–2005, while healthy firms were chosen from active firms by the end of 2005. The first 31 financial ratios were investigated through literature review and categorized them as profitability, debt coverage, leverage, capital structure, liquidity, activity, and size. Then, final input variables were selected by evaluating the value of each variable based on the analysis of the receiver operating characteristic (ROC) curve.

All ensembles showed greater performance than individual classifiers. Furthermore, results showed that DT ensembles ' predictive accuracy (75.10%, 75.78%) was higher than both NN (73,10%, 73,97%) and SVM (73,07%, 72,85%) ensembles. Both ensembles registered marginal improvements for validation data over a single classifier, with approximately 4,8% and 5,48% for DT, 2,08% and 2,95% for NN, and 0,62% and 0,4% for SVM, respectively.

The results mean that DT ensembles containing a variety of classifiers minimize the generalization error and thus generate prominent performance improvement, whereas stable NN / SVM ensembles have the performance improvement limit due to multicollinearity problems. CO-SVM (77.53 percent, 77.23 percent) has more accurate results than CO-DT (76.00 percent, 76.20 percent) and CO-NN (76.52 percent, 76.92 percent) in the comparison of configured ensembles. The increases in coverage output for DT ensembles are as high as around 0.9 percent and 0.42 percent, for NN ensembles 3.42 percent and 2.95 percent, and for SVM ensembles 4.46 percent and 4.38 percent. In summary, optimized classifiers have fewer classifiers as shown compared to ensemble classifiers, but their accuracies are higher than those of the original ensemble classifiers.

More recently, regarding the Bateni & Asghari proposed model (Bateni & Asghari, 2020), the GA model achieved accuracy rates of 95 and 93.5 percent in training and test samples, respectively. While the logit model achieved accuracy rates of just 77 and 75 percent in training and test samples, respectively.

The results suggest that the two models can predict the bankruptcy, and the GA model is more accurate in this regard than the logit model.

# 3. METHODOLOGY

The methodology was chosen with the main objective of identifying and compare the efficiency between MDA, Logit with stepwise and Logit with GA models, in order to predict the bankruptcy in Portuguese and French companies.

In the following figure, it is presented the methodology model, which is composed of three different phases, with the propose of illustrating and synthesize the process followed in the study:

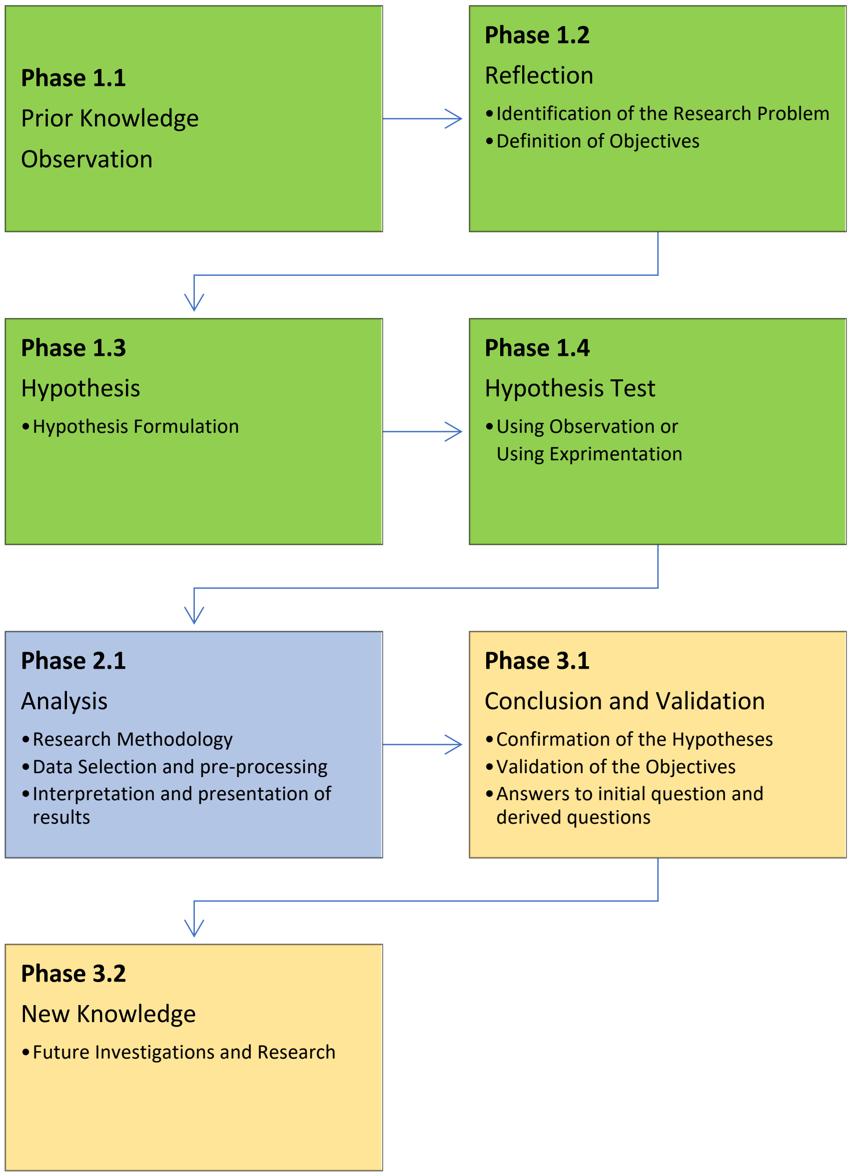



Figure 3.1 – Dissertation Methodology

In the first phase (**Exploratory**), with the green color, it's included the questions related to the conception of the plan of the investigation to be developed.

In this activity, we also studied the most relevant articles, and books published related to bankruptcy prediction, genetic algorithms, criteria for identification of outliers and statistical methodologies for variable selection.

In the second phase (**Analytic**), with blue color, is covered the collection, record, and analysis of the data, as well the result interpretation.

Specific models have been developed here for each sector, group of sectors, country, group of countries and group of sectors and countries. This procedure allowed us to understand the impact of the samples on the structure of the models as well as the differences in their performance.

Still, at this stage, the approach to genetic algorithms was applied as an alternative methodology to forward stepwise, thus introducing, also at this level, an innovative character that proved to be an important contribution to the efficiency of the models.

Isolation Forest was also used in the detection and treatment of outliers.

In the third and last phase (**Conclusive**), colored in yellow, is presented the conclusion of the work developed, after the validation of hypothesis, substantiation of the results, verification of the proposed objectives and answered to the research questions.

Moreover, in this phase, the limitations of the work and investigation are considered.

# 4. PROPOSED MODELS

This chapter corresponds to the development of the models, including sample selection, data pre-processing, model development, and result from the analysis.

The universe of analysis was randomly selected both for countries and sectors from a database involving all the countries and sectors of activity.

## 4.1. SAMPLE AND POPULATION

### 4.1.1. Population

From the Amadeus Database ("Amadeus - European business information | Bureau van Dijk") were extracted data related with 567 companies from two different countries, Portugal and France, as well from two different sectors C – Manufacturing and G - Wholesale and Retail, Repair of Motor Vehicles, and Motorcycles, from 2010 to 2016. In addition, for each of the companies were extracted 70 variables (Annex A1).

In the following that table we can observe a summary of the initial population per year:

Table 4.1 Initial Population Summary

| Country | Sector | BK or NBK | Initial Nº of Companies |
|---|---|---|---|
| PT | C | Bankrupt | 25 |
| PT | G | Bankrupt | 41 |
| PT | C | Not Bankrupt | 28 |
| PT | G | Not Bankrupt | 46 |
| FR | C | Bankrupt | 67 |
| FR | G | Bankrupt | 137 |
| FR | C | Not Bankrupt | 67 |
| FR | G | Not Bankrupt | 156 |

### 4.1.2. Sample Selection Criteria

The methodology proposed for the selection of the samples to be used during the dissertation is materialized in the study of events and is based on eight distinct phases:

1) Selection of relevant information.

2) Selection of the countries to be analyzed:
    a. In addition to Portugal, another country belonging to the Euro Zone, France, was randomly selected from which the several macroeconomic indicators were collected, as described in point 5.a.

3) Selection of CAE Rev. two (Sectors) to be analyzed:
    a. Public entities, holding companies (SGPS's), sports entities, and those of the financial sector (banking and insurance), were excluded. Two aggregates were selected, CAE

Sections (Sectors): C - Manufacturing; and G - Wholesale and Retail, Repair of Motor Vehicles, and Motorcycles.

4) Selection of cut-off points to be applied to the sample:
    a. The cut-off point selected for the study was the accounting cut-off point, which covers companies with Equity < 0.

5) Selection of companies by Country and CAE Section:
    a. Not individual, with this, companies, with a size equal to or greater than small enterprises and subject to Statutory Audit according to the rules in force in the country of origin;
    b. Have complete and consistent financial information (Balance Sheet, Income Statement and complementary information) for all years from 2010 to 2016;
    c. Following the cut-off point identified in 4. and with the criteria indicated in and the companies will be considered, in the sample of this thesis, as:
        i. bankruptcies (F): if they meet one of the criteria in 4. in 2016 and cumulatively fail to meet it from 2010 to 2015;
        ii. non-bankruptcies (NF): if they do not meet any of the criteria in 4 cumulatively from 2010 to 2016;

6) Sample Construction:
    a. For each Country, Section CAE and Cutting Point, the companies were grouped into two related sub-samples:
        i. Sub-Sample F: Classified as Bankrupted as indicated in each of 4 and 5.c.i;
        ii. Sub-sample NF: Classified as Healthy, as indicated in 5.c.ii, paired with the sub-sample in the matched sample, by the Mean of Total Assets and Turnover in the period under analysis with a deviation of [±0.25std] from the mean.

7) Collection of economic and financial indicators:
    a. The most present in the economic-financial analysis as well as in the 123 models studied.

8) Generation of economic and financial indicators by a combination of macroeconomic, balance sheet, and income statement variables.

In summary, after the process, composed of the eight phases previously described, were selected a total of 1887 indicators (ratios). All the indicators have the propriety of being based on economic and financial ratios since Altman's (1968) methodology was adopted, ensuring that all 22 ratios selected by Altman would be included in the inputs of our model, being an obligatory requirement in the selection criteria.

**Sample and Data Processing**

Since the primary purpose of this dissertation is to address the different deficiencies and sensitivities attributed to the models mentioned above, we attempted to create a multinational sample in the first phase, concentrating on the Euro Zone, where one more country was randomly selected apart from Portugal and also belonging to the Euro Zone, France.

Subsequently, using the Bureau Van Dijk's Amadeus database ("Amadeus - European business information | Bureau van Dijk,"), we selected a list of all Portuguese and French companies with data available for the period 2010-2016, which meet the selection criteria set out in subparagraph 5 of the preceding section.

As a result of their unique characteristics concerning accounting principles specific to their accounting, entities from the financial sector (banking and insurance), public corporations, holding companies (SGPS's) and Sports Public Limited Companies (SAD's) were, therefore, excluded at the very beginning

Therefore, having been excluded at the very beginning, entities from the Financial Sector (banking and insurance), Public Companies, Holding Companies (SGPS's) and Sports Public Limited Companies (SAD's), as a result of their specific characteristics concerning accounting standards applicable to their accounting, were randomly selected from the remaining two CAE aggregates (sectors most frequently used in the study and derivation of these models internationally: C — Manufacturing and G — Wholesale and Retail Trade; Motor Vehicle and Motorcycle Repair.

Among these, and according to Yang's suggestion (Yang, 2014),'[...] auditing is a valuable [...] external monitoring tool. High-quality auditing can increase the quality of financial information,' to ensure a higher standard of financial information, companies have been chosen, subject to legislative audit under the country of origin regulations.

Two sub-samples were then developed for each country and CAE, with those of the sub-samples of companies classified as not bankrupt being established by matched sample (matching Total Assets and Turnover), i.e., considering that for each company of bankrupt sub-samples one or more companies are directly comparable in non-bankrupt samples, being the most representative of the reality.

**Variables Selection**

For this dissertation were initially selected 70 variables belonging to 4 different categories:

- Variables from the Financial statement also known as Balance Sheet statement;
- Profit & Loss (P&L) statement variables;
- Classification variables of the company (for example Age or Number of employees); and
- Macroeconomic variables;

All variables were taken from the Amadeus database except the macroeconomic variables. In contrast, the macroeconomic variables have the PORDATA Database (“PORDATA - Statistics, charts, and indicators on Municipalities, Portugal and Europe,” ) as their source. These initial variables can be observed in the annex **A 1**.

After the selection of these variables, by the combination of these variables was created additional ones. These additional variables were calculated according to the following process and criteria:

1. All the variables used as a basis for the process are expressed in euros;
2. The macroeconomic variables were used as numerator or denominator with another, not macroeconomic variable;
3. Each variable was divided by the remaining variables, were one of the variables (ratio) is the numerator and the other the denominator since all the variables are divided by the other. This operation was only performed one time by a pair of variables avoiding the existence of a division plus the inverse of this division. In other words, if we imagine a matrix with the combination of these divisions, only the first part was calculated.
4. Eliminate the columns (variables) with a mean less than 0,0000, in order to have variables with an insignificant weight.

After this process, 1887 variables(ratio) were created, being the variables used in the models studied in this dissertation. It is essential to mention that these 1887 ratios include the ratios used in the 123 most relevant studies regarding bankruptcy prediction (Peres & Antão, 2017) and also the most relevant in the corporate finance literature.

Since these ratios are calculated using the first 66 variables referred initially, which are standard and common among most of the companies in Europe, it is possible to conclude that these models can be applied and tested for the majority of European companies.

### 4.1.3. Sample and Sample Control

Some data pre-processing techniques were implemented after the sample collection in order to prepare the data for model creation. Since the models implemented in this dissertation are prone to outliers and correlation, the implementation of these pre-processing techniques was important in order to ensure data consistency and not bias the model results, ensuring the best possible performance of the models.

**Outliers Detection and Treatment**

Since we have a large number of variables (ratios) in this study, a technique that would classify the outliers in a multidimensional space was required. The approach that seemed ideally suited to this problem after some analysis and experiments was the Isolation Forest algorithm, an approach that is in theory close to the well-known and common Random Forest proposed by Liu et al. (2008), publishing a profoundly different model-based method which specifically isolates anomalies rather than usual profile points.

Isolation Forest directly detects anomalies rather than profiling normal data points, which is an approach other than other common methods of outlier detection. Isolation Forest, like any tree ensemble system, is constructed on the basis of the decision trees. Partitions are generated in these trees by selecting a feature at random first, and then selecting a random split value between the minimum and maximum value of the selected feature.

Outliers are usually less frequent than usual observations and vary in terms of values (they are further away from standard feature space observations). Therefore, by using such random partitioning (shorter average path length, i.e., the number of edges that an observation can cross in the tree going from the root to the terminal node), they can be detected nearer to the root of the tree, with fewer splits needed.

The definition of defining a natural or anomalous event can be seen in the figure below. A normal point (on the left) needs more partition detection than an anomalous point (right).

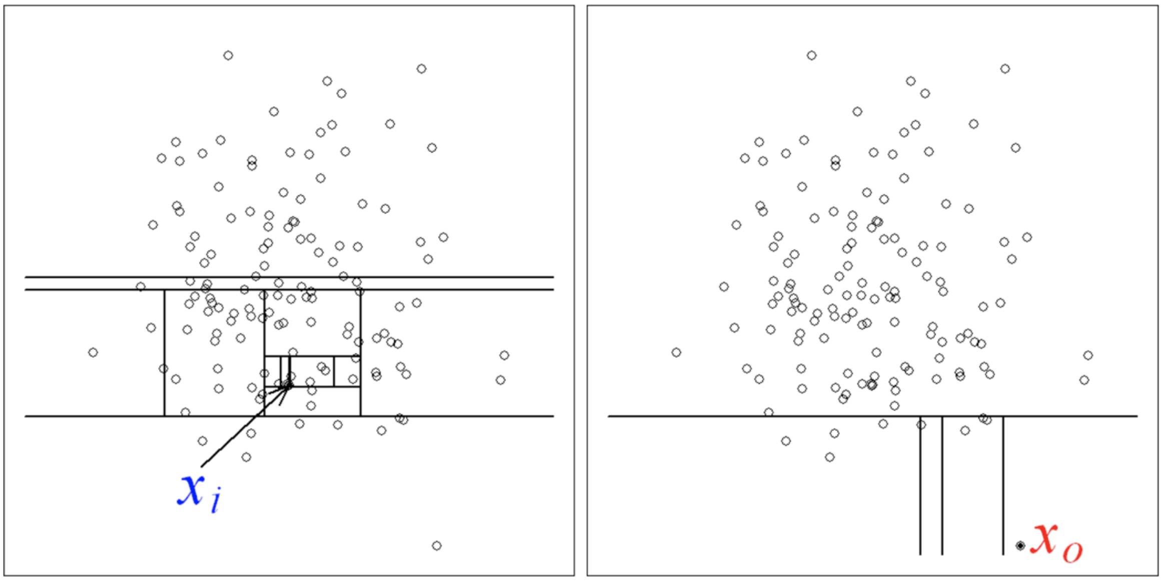



Figure 4.1 Identification of normal vs. anomalous observations with Isolation Forest (Liu et al.,2008)

As with other methods of outlier identification, is required an anomaly score. For Isolation Forest, is represented as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{6}$$

In this function, h(x) is the observation x path length, c(n) is the average path length of unsuccessful search in the Binary Search Tree, and n is the number of external nodes. In the research of Liu et al. (2008) is possible to read with more detail the anomalies score and its components.

Each observation has an anomaly and can be taken on the basis of the following:

- Anomalies are represented by a score close to 1;
- Normal observations are represented by a score less than 0,5;
- When all scores are close to 0,5 then the entire sample does not reveal clearly distinct anomalies.

**Correlation Treatment**

Because the sample may have complex and uncertain relationships between variables, the degree to which variables depend on one another in the data sample needs to be discovered and calculated. Such knowledge would help better prepare data for the demands of this dissertation's machine learning algorithms (MDA, GA, and Logit), whose efficiency will degrade with the presence of these interdependencies.

For several purposes, variables within a data sample may be related, for example, one variable could trigger or depend on the values of another variable, or one variable could be related significantly to another variable, or two variables might even depend on an unknown third variable.

Consequently, in data analysis and modeling, a deeper understanding of the relationships between variables can be useful, and the statistical association between two variables is called correlation.

A correlation may be positive, meaning that both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values go down. Correlation can also be neural or zero, so it's independent of variables.

For this dissertation, which can be used to describe the strength of the linear relationship between two data sets, the Pearson correlation coefficient named after Karl Pearson was used.

The Pearson coefficient of correlation is defined as the covariance of the two variables separated by the sum of the standard deviation of each data set.

Given a pair of two random variables (X, Y) the formula for ρ (population Pearson correlation coefficient) is:

$$\rho\, X, Y = \frac{cov(X, Y)}{\sigma X \sigma Y} \tag{7}$$

Where in (7):

- $cov$ is the covariance
- $\sigma X$is the standard deviation of X
- $\sigma Y$is the standard deviation of Y

The use of mean and standard deviation in the calculation suggests the need to distribute the two data samples in a Gaussian distribution. The result of the equation can be interpreted as the coefficient of correlation for understanding the relation.

The coefficient returns a value between -1 and 1 that represents the correlation 's limits from a total negative correlation to a full positive correlation. There is no correlation suggested by a value of 0. The value needs to be perceived where the value near-0.5 or above 0.5 frequently suggests a significant correlation and the values below these values show a less significant correlation.

To apply this form, Pandas in Python's corr () functions were used, returning a matrix with 1 along the diagonals and symmetric, regardless of the behavior of the callable. The limit value specified for the maximum agreed correlation for this dissertation was 0.85.

Therefore, variables with a correlation greater than 0.85 are omitted after the correlation matrix has been established, thereby reducing the number of variables(ratios).

## 4.2. APPLICATIONS OF THE PROPOSED MODELS

### 4.2.1. Code Development

The code in this dissertation was developed in Python 3.6 and R. The data preprocessing, MDA and the Logistic with GA models were developed in Python. Both environment for Python and R were saved, in order to be reproduced for future research.

For Logistic with GA, the genetic Algorithm was developed using Python's Distributed Evolutionary Algorithms (DEAP). DEAP has been developed at *Université Laval* and is an evolutionary computing framework for fast prototyping and testing of ideas, which integrates the data structures and methods required to apply the most common evolutionary computing techniques, such as genetic algorithms, genetic programming, evolution strategies, differential evolution and distribution algorithm estimation.

### 4.2.2. Adoption

#### 4.2.2.1. Data preprocessing

As referred before, all the Data preprocessing was developed in Python using Pycharm IDE.

Data Reduction

Outliers Detection and Treatment

The Isolation Forest function from the scikit-learn library was used for Outliers Detection and Treatment, all the parameters assigned were the default ones expect the "warm start" parameter assigned as "True" to use the previous call solution to fit and add more estimators to the ensemble, otherwise a whole new forest would just fit.

Furthermore, it is important to point out that the default contamination value of 0.1 in the "contamination parameter" means that a value of 10% of contamination is expected in the data set, i.e. 10% of outliers in the data set.

The code developed can be observed in the figure below:

```
iForest=IsolationForest(n_estimators=100, max_samples='auto', contamination=0.1, \
                        max_features=1.0, bootstrap=False, n_jobs=-1, random_state=42, verbose=0,warm_start=True)

first_ratio = list(data).index('R1')

dataset = data.iloc[:,first_ratio:]

iForest.fit(dataset)
pred = iForest.predict(dataset)

dataset['anomaly']=pred

outliers=dataset.loc[dataset['anomaly']==-1]

data_without_outliers = dataset.loc[dataset['anomaly']==1]
```

Figure 4.2 Outliers Detection and Treatment Code

In the annex **A 2** it is possible to observe the summary results of the outlier treatment.

### Correlation Analysis and Treatment

With respect to correlation analysis and treatment, the pandas library corr() function was used, with all the default parameters. With this, the Pearson correlation coefficient was used to detect the outliers. After detection, the outliers were removed based on a 0.85 threshold of correlation between the variables.

A section of the code that was created for this process can be seen in the following figure:

```python
#Create Correlation Function
def remove_correlated_vars(data,threshold):
    """
    :param data:
    :param threshold:
    :return:
    """

    first_ratio = list(data).index('R1')
    corr_matrix = data.iloc[:,first_ratio:].corr().abs()

    # Select upper triangle of correlation matrix
    upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))

    # Find index of feature columns with correlation greater than 0.85
    to_drop = [column for column in upper.columns if any(upper[column] > threshold)]

    # Drop features
    X1 = data.drop(data[to_drop], axis=1)
    return X1
```

Figure 4.3 Correlation Analysis and Treatment Code

## Data Transformation

### Standardize data

Since the measured features are in different scales, they do not contribute equally to the analysis and may result in a bias. Accordingly, the standardization (Standard Scalar) of the data should be performed, which means that the variable is centered at zero and the variance at 1. This procedure involves subtracting the mean of each observation and dividing it by standard deviation afterward.

The standardization consists in rescaling the features so that they'll have the properties of a standard normal distribution with where $\mu$ is the mean (average) and $\sigma$ is the standard deviation from the mean, which we can be seen in the following formula (7):

$$z = \frac{x - \mu}{\sigma} \tag{7}$$

For example, by applying this method, a variable ranging from 0 to 1000000 would outweigh a variable ranging from 0 to 1. The use of these variables without standardization would give the variable with the greater range weight of 1000000 in the study more significance. This problem can be avoided by transforming the data into comparable scales.

It should be noted that since the Train / Test split is used to evaluate the performance of the models, we will perform feature normalization over the training data, then still perform standardization on test instances, but this time using the mean and variance of the explanatory variables of the training. In this way, we can test and evaluate whether our model can well generalize to new, unseen data points.

This process was developed in Python using the StandardScaler () function from scikit-learn package, as can be observed in the following figure:

```
std_scale = StandardScaler().fit(X_train)
scaled_train = std_scale.fit_transform(X_train.values)
scaled_test = std_scale.fit_transform(X_test.values)
X_train = pd.DataFrame(scaled_train, index=X_train.index, columns=X_train.columns)
X_test = pd.DataFrame(scaled_test, index=X_test.index, columns=X_test.columns)
```

Figure 4.4 Data Standardization Code

#### 4.2.2.2. Models Development

In this chapter, we will explain how the models were developed and the structure of the process and code for each model studied in this thesis. Starting from a general overview of the initialization and different phases of the applied steps and methods, to a detailed summary of the settings used in each model.

GA Model with logit

A combination of Genetic Algorithms and Genetic Programming with Logistic Regression was developed with respect to this model. Given that we have a high-dimensional dataset for this study, composed of a large number of variables (1887 ratios before correlation treatment), the Genetic Algorithms were used to select the feature. As already mentioned, we built this process for this research using python programming language and DEAP library. With this tool, we can have a lot of flexibility and even the code inside the package gives the user the possibility to easily configure and enhance the GP run settings.

While the selection of the feature is performed by the Genetic Algorithm, the model is developed with a Logistic Regression using the features previously selected by the GA. With this, the main objective is to take the predictions of the logistic model, for each of the observation check if the prediction is the same as target label and train the Genetic Algorithm to in the order to choose the best optimal set of features.

After the model is developed, an evaluation is performed using K-fold cross-validation where the assigned K is 5, meaning that the data sample is divided into five groups where the group of each sample is given the opportunity to do so. With this it is possible to have the minimum, maximum and average value of the 5-fold cross-validation, which will determine the fitness value of the fitness function, process which will be explained afterwards, and then the GA adjusts the variable selection process depending on that value. All this process will be cyclical and repeated n times according to the number of generations specified.

At the end of the process the subset with the highest fitness value is chosen and will be evaluated to 1,2,3, and 4 years before the bankruptcy.

The process described above can be illustrated in the following figure:

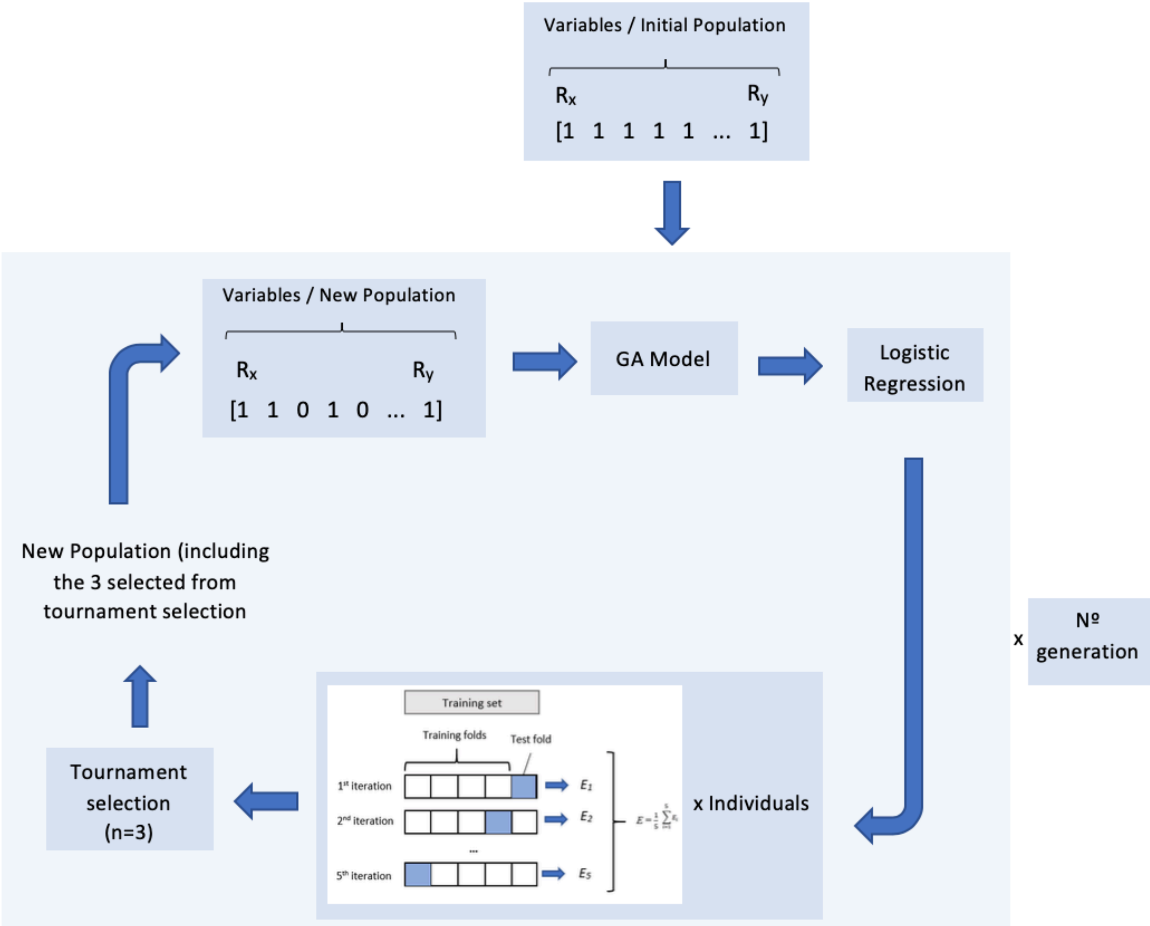



Figure 4.5 Logit with GA Process

**Fitness Function**

After the creation of the population, the evolution occurs. In each generation, the individuals are evaluated using the fitness function implemented for this study. In the Genetic algorithm program, the selection of the fitness function is crucial to guarantee a good performance. In this study allows to a set of features which is not only well-fitted to the data but also provides good generalization abilities that can improve the performance of the logistic regression.

Since accuracy is the most conventional method for evaluation of machine learning models. This metric also summarizes the percentage of correctly classified observations.

Although it provides good overview of the predictions, it is highly biased in some of the cases, but since the data set is balanced, we use accuracy for evaluate the model.

As mentioned, before we use K-cross-validation to evaluate the performance of the logistic regression, and the average accuracy was selected as the fitness value, where the objective of the fitness function is to maximize this value. The evolution of the fitness value as well the minimum and maximum values of accuracy during the Logit with GA process can be observed in the following figure:

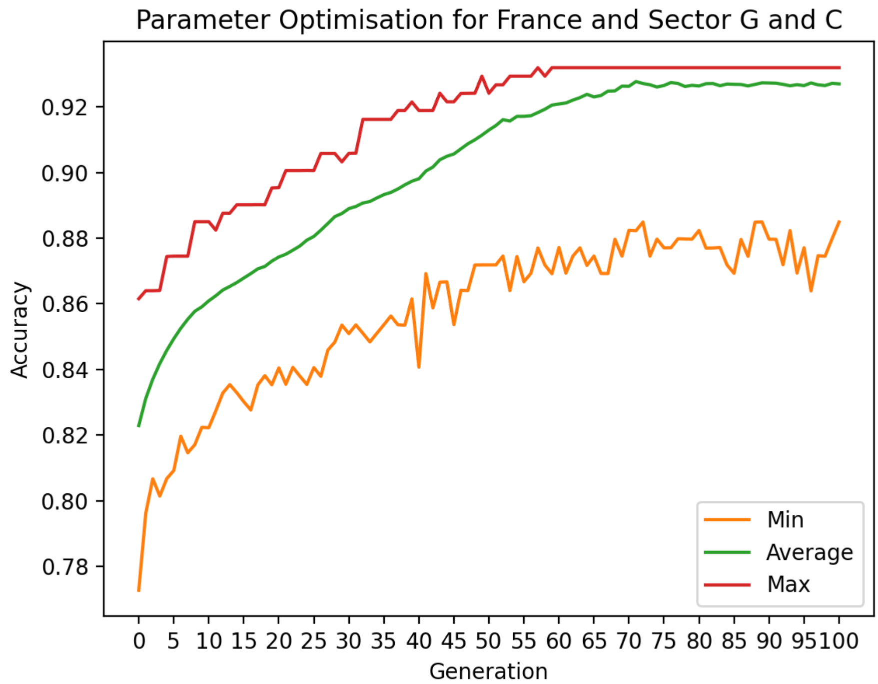



Figure 4.6 Logit with GA process for France and Portugal Sector G and C

Cross-validation is a resampling procedure used on a limited data sample to evaluate the machine learning models. This procedure has a single parameter named k which refers to the number of groups to be divided into for a given data sample. This method is very popular since is simple to understand and generally provides a less biased or a less optimistic compared with other methods.

For this study, the Cross-validation process can be represented as follows:

1. The dataset randomly shuffled;
2. The dataset is sliced into 5 groups;
3. For each unique group
    a. One of the groups is selected as a test data set
    b. While the remaining groups are selected as a training data set
4. The logistic regression is fitted on the training set and evaluate it on the test set
5. Retain the evaluation score and discard the model.
6. In the end the result of the model is summarized using the sample of model evaluation scores.

With this process each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure, where each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

In the following figure it is possible to observe this process:

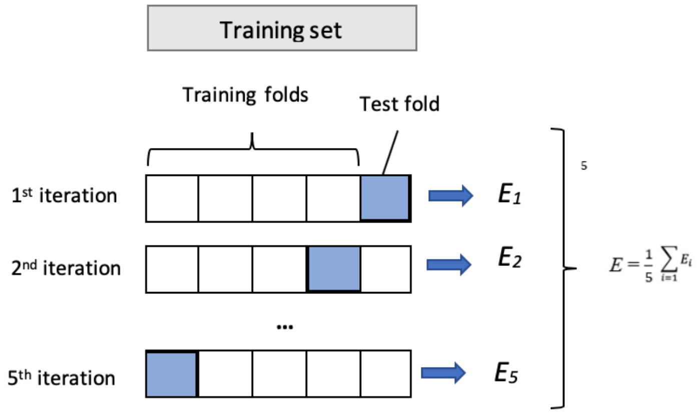



Figure 4.7 K-fold Cross-Validation Process

**GA settings**

With the DEAP library multiple methods for initialization and variation (cross-over and mutation) can be selected.

First, all the variables are transformed into a binary representation as input to the Genetic Algorithm, assigning the value 1 to each of the variables and creating a float list. Thus, an array with 1887 (ratios before correlation treatment) values of 1 values is created as an initial input, where all the variables are represented and used by the logistic regression, composing the initial population of the GA.

After first population of individuals is evaluated, the genetic algorithm created a new population for the next generation. In this dissertation we decided to use Tournament Selection, using the tournament size of 3, which means that for n individuals the tournament is performed n times, and for each time that the tournament is performed , three individuals are randomly selected from the population and the best out of these to become the parent, being saved for the next population.

Additionally, we use a feature offered by the DEAP package called Hall of Fame to prevent that best program is not selected. This feature enables the implementation of a concept known as Elitism, where the best individuals in the population are stored through the process of evolution.

Upon selection, the Genetic Operators must be applied according to the user-defined settings. In this study we use for Crossover the Two Point Crossover with a probability of 0,5. In addition, a feature offered by the DEAP package named muFlipBit was used where the value of the individual input attributes was flipped and the mutant returned, which means that the value 1 was flipped to 0 or vice versa , determining the subsets of variables for the logistics regression, we used a probability of 0,2 for the Mutation. The definition of these settings can be observed in the following figure:

```python
# create toolbox
toolbox = base.Toolbox()
toolbox.register("attr_bool", random.randint, 0, 1)
toolbox.register("individual", tools.initRepeat,
                 creator.Individual, toolbox.attr_bool, len(X.columns))
toolbox.register("population", tools.initRepeat, list,
                 toolbox.individual)
toolbox.register("evaluate", getFitness, X=X, y=y)
toolbox.register("mate", tools.cxTwoPoint)
toolbox.register("mutate", tools.mutFlipBit, indpb=0.05)
toolbox.register("select", tools.selTournament, tournsize=3)
```

Figure 4.8 Selection, Mutation and Crossover settings code

Moreover, for this study we started with a population of 1000 individuals and we used 100 generations in order to have a large diversity, but consequently with trade-off of extending the time of training.

MDA Model

As mentioned before Altman presented in 1968 the Z-score model based on MDA (Edward I. Altman, 1968). Since Altman research (1968) had some limitations, where one of the most important was the fact of the model was only prepared to be conducted on listed companies, i.e. that had market values of their own equity, so in order to solve and after in 1983 Altman(1983) modified the ratio $X_4$ where the market value of equity was replaced by the book value of equity as the market value is often not easily obtained.

Consequently, since not all the companies in our sample are listed companies, we applied the modified Z-Score developed in 1983.

The MDA model was developed in Python as it is possible to observe in the following figure:

```python
def z_score(dt,X1, X2, X3, X4, X5):


    Z = 0.717 * (1 / dt.loc[:, X1]) + 0.847 * (1 / dt.loc[:, X2]) + \
    3.107 * (1 / dt.loc[:, X3]) + 0.42 * dt.loc[:, X4] + 0.998 * (1 / dt.loc[:, X5])

    y = []
    for i in Z:
        if i < 1.23:
            y.append(1)
        else:
            y.append(0)

    return y
```

Figure 4.9 Z-Score Model Code

As mentioned before, the calculation process of the ratios, was only applied once for each pair of variables, consequently for this model it was necessary to perform transformations to the variables in order to have the inverse.

Following ratios (variables) were used as an input in the model:

- $\frac{1}{R1227}$ = $X_1$ - working capital / total assets;
- $\frac{1}{R1199}$ = $X_2$ - retained earnings / total assets;
- $\frac{1}{R1216}$ = $X_3$ - earnings before interest and taxes / total assets;
- R1310 = $X_4$ - book value of equity / total liabilities;
- $\frac{1}{R1209}$ = $X_5$ = sales / total assets;

## Logit Model

The last model developed in this dissertation was the Logit Model with forward stepwise for feature selection, and it was developed in R using the glm() function. Stepwise selection for linear regression models was originally developed as a feature selection technique. The forward stepwise regression method was used for this dissertation, using a sequence of steps that allows variables to enter the one-on-a-time regression model before it converges to a subset of features.

In order to enter a variable, the model is usually based on a threshold of p-value. A common entry criterion is generally that a p-value must be less than 0.15 for a function to enter the model.

This process starts with the development of n logistic regression models, where each one uses precisely one of the features. In addition, their individual explanatory potential variability in the outcome is then ranked in the value of the features. For simplicity the amount of variance described can be expressed by the p-value. If no functions have a p-value of less than 0.15, the cycle will end.

This method begins with the creation of n logistic regression models, where each uses one of the features specifically. Each of the additional features is evaluated and the chosen feature set is applied to the best function that meets the requirements for inclusion. Therefore, in the presence of the other feature, the sum of variance defined by each function is determined with respect to the p-value If the p-values do not meet the cutoff criterion, both will be maintained, and a third attribute will be verified by the search process. This loop continues until it reaches convergence criteria, where it cannot add new variables.

However, the accumulation of false positive results is a primary fault of the stepwise selection, as step by step selection uses many repetitive hypothesis tests to make decisions about the inclusion or exclusion of individual predictors. The resulting p-values are thus unadjusted, leading to an over-selection of features (i.e., false positive findings). Moreover, this issue gets compounded when there are strongly correlated predictors.

In order to mitigate this problem and using a statistic other than p-values to select a feature, we used the Akaike information criterion (AIC) (Akaike, 1974). The AIC statistics are applied to models that use the likelihood as the goal (i.e. linear or logistic regression) and penalize the probability by the number of parameters used in the model. Models that optimize probability and have fewer parameters are preferred, however. Functionally, the AIC statistical value for each sub-model which includes a new feature can be determined after fitting an initial model. The next model is the one which has the best AIC statistics. The method repeats until it produces the best AIC statistics in the current iteration.

The maximum value of the log-likelihood function of a model could be defined as:

$$AIC = 2K + n\,Log\left(\frac{RSS}{n}\right) \tag{8}$$

Where RSS is then the residual sum of squares, and then K the number of independent variables, and n the number of observations. Consequently, if all models have the same k, selecting the model with minimum AIC is equivalent to selecting the model with minimum RSS, which is the usual objective of selecting the model based on the minimum squares.

# 5. RESULTS AND DISCUSSION

## 5.1. INTRODUCTION

In this chapter it will be analyzed and compared the results of the models developed in this dissertation, using Accuracy (9) and Area under ROC (Receiver Operating Characteristic) value as performance measures. Since the usual accuracy measure only summarizes true negatives and true positives, using the AUC from the ROC curve we can observe how true positive rate and false positive rates are changing for different threshold values. The Area under ROC was studied by Bradley (1997) where it is concluded in the paper that this method can be used for a broad number of machine learning, improving the perceptibility of the predictions as well as the performance of the models. The calculation of the area under ROC is also based in the confusion matrix (Table 5.1), where the data that are summarized there in 4 categories: true negatives, false negatives, false positives and true positives.

Table 5.1 Confusion Matrix

| Confusion Matrix | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | True Negative | False Positive |
| | Positive | False Negative | True Positive |

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \qquad (9)$$

As mentioned before the Area Under ROC, abbreviated as AUC, is a single scalar value that measures the overall performance of a binary classifier (Hanley & McNeil, 1982) in this case if a company will bankrupt or if is healthy. The AUC value ranges between 0.5 and 1, where the minimum value represents the performance of a random classifier and the maximum value would correspond to a perfect classifier. In the following figure is possible to observe the ROC curve and the AUC for Logistic model with GA applied for Portugal and France for Sector G and C:

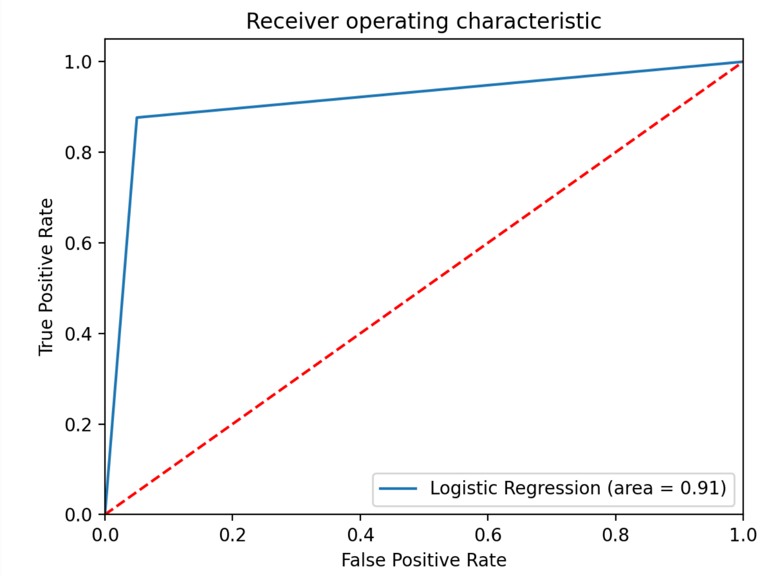



Figure 5.1 ROC curve and AUC for Logistic Model with GA applied to France and Portugal, sector G and C

In the example above it is possible to observe an AUC of 0,91, which is an excellent result. In addition, as we can visualize in the figure the ROC curve is composed of the False Positive Rate (FPR) in the X-axis and True Positive Rate (TPR) on the Y-axis.

**Sensitivity** is the proportion of bankrupted companies which were predicted to bankrupt, in probability notation:

$$TPR = Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \tag{10}$$

**Specificity** is the proportion of healthy companies that were predicted not to bankrupt. In probability notation:

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)} \tag{11}$$

$$FPR = 1 - Specificity = 1 - \frac{True\ Negative}{(True\ Negative + False\ Positive)} \tag{12}$$

## 5.2. MDA MODELS

On this section it is presented the short-term and long-term performance analysis for each modified Z-Score model based on each sub sample.

### 5.2.1. Short-Term Performance

Table 5.2 MDA model short-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 1 | 61,60% | 58,99% |
| France | G and C | 1 | 62,66% | 61,02% |
| Portugal | G | 1 | 73,08% | 71,75% |
| Portugal | C | 1 | 89,36% | 88,91% |
| Portugal | G and C | 1 | 79,20% | 78,15% |
| France and Portugal | G | 1 | 64,22% | 61,95% |
| France and Portugal | G and C | 1 | 66,73% | 65,19% |

In the table above is possible to observe a good short-term performance of the modified Z-Score (MDA model) when applied in Portugal Sector C with an accuracy of 89,36% and a fair performance in Portugal Sector G and C with an accuracy of 88,91%. However, regarding the remaining samples the model presented a poor short-term performance, where for France Sector G resulted in an accuracy of 61,60% on the year before the bankruptcy or not bankruptcy occurrence.

### 5.2.2. Long-term Performance

Table 5.3 MDA model long-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 2 | 57,41% | 54,47% |
| France | G | 3 | 58,94% | 56,15% |
| France | G | 4 | 57,41% | 54,47% |
| France | G and C | 2 | 58,49% | 56,60% |
| France | G and C | 3 | 57,70% | 55,83% |
| France | G and C | 4 | 55,87% | 53,89% |
| Portugal | G | 2 | 62,82% | 61,21% |
| Portugal | G | 3 | 61,54% | 59,85% |
| Portugal | G | 4 | 61,54% | 59,85% |
| Portugal | C | 2 | 85,11% | 84,09% |
| Portugal | C | 3 | 74,47% | 73,55% |
| Portugal | C | 4 | 70,21% | 69,27% |
| Portugal | G and C | 2 | 71,20% | 69,76% |
| Portugal | G and C | 3 | 66,40% | 64,95% |
| Portugal | G and C | 4 | 64,80% | 63,34% |
| France and Portugal | G | 2 | 58,65% | 56,05% |
| France and Portugal | G | 3 | 59,53% | 57,02% |
| France and Portugal | G | 4 | 58,36% | 55,73% |
| France and Portugal | G and C | 2 | 61,61% | 59,80% |

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France and Portugal | G and C | 3 | 59,84% | 58,04% |
| France and Portugal | G and C | 4 | 58,07% | 56,18% |

Observing the table above is possible to observe that the modified Z-score model continues to present a good performance for Portugal sector C with an accuracy of 85,11% and 70,21% two years and four years before the bankruptcy or not bankruptcy occurrence, respectively.

In addition, the long-term performance of the model for Portugal in sector G and C drops significantly compared with the short-term performance, decreasing from 79,20% to 64,80% two years and four years before the bankruptcy, respectively.

## 5.3. Logit Model with Forward Stepwise

On this section, it is presented the short-term and long-term performance analysis for each modified Logit Model with feature selection using Forward Stepwise model based on the same subsamples of the models presented before.

### 5.3.1. Short-Term Performance

Table 5.4 Logit with Forward Stepwise model short-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 1 | 82,28% | 82,37% |
| France | G and C | 1 | 84,35% | 84,47% |
| Portugal | G | 1 | 82,61% | 81,82% |
| Portugal | C | 1 | 78,57% | 78,57% |
| Portugal | G and C | 1 | 100,00% | 100,00% |
| France and Portugal | G | 1 | 83,33% | 83,10% |
| France and Portugal | G and C | 1 | 81,05% | 80,98% |

In the table above is possible to observe that the model with the best short-term performance is Portugal G and C achieving 100% accuracy. In another hand, the model with the worst short-term performance is Portugal C achieving an accuracy of 78,57%.

Moreover, it is possible to observe that all the remaining models present a good performance of around 80% of accuracy where the performance is better when we combine two different sectors are combined, except with the combination of the two countries.

### 5.3.2. Long-term Performance

Table 5.5 Logit with Forward Stepwise model long-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 2 | 78,33% | 76,83% |
| France | G | 3 | 77,19% | 75,61% |
| France | G | 4 | 77,19% | 75,61% |
| France | G and C | 2 | 74,67% | 73,50% |
| France | G and C | 3 | 72,85% | 71,58% |
| France | G and C | 4 | 72,32% | 71,06% |
| Portugal | G | 2 | 91,03% | 90,80% |
| Portugal | G | 3 | 79,49% | 79,70% |
| Portugal | G | 4 | 48,72% | 51,09% |
| Portugal | C | 2 | 74,47% | 75,18% |
| Portugal | C | 3 | 76,60% | 77,45% |
| Portugal | C | 4 | 76,60% | 77,18% |
| Portugal | G and C | 2 | 85,60% | 84,84% |
| Portugal | G and C | 3 | 80,80% | 81,01% |
| Portugal | G and C | 4 | 55,20% | 57,49% |
| France and Portugal | G | 2 | 74,78% | 73,12% |
| France and Portugal | G | 3 | 73,61% | 71,88% |
| France and Portugal | G | 4 | 73,61% | 71,91% |
| France and Portugal | G and C | 2 | 77,36% | 77,34% |
| France and Portugal | G and C | 3 | 77,36% | 76,97% |
| France and Portugal | G and C | 4 | 74,41% | 73,83% |

- We can observe two different situations:
  - The model performance remains average along the years
  - A significant decrease in the performance of the model four years before the bankruptcy, Portugal G and C and Portugal G
  - But for Portugal G the model is better two years before the bankruptcy than one year, 82,61% vs 91,03% of accuracy, respectively.

In addition, in table **A 3** in the annexes is possible to observe the variable used for each model developed with the application of Logit with Forward Stepwise.

## 5.4. Logit Model with GA

In this section, it is presented the short-term and long-term performance analysis for each modified Logit Model with feature selection using Genetic Algorithms model based on the same subsamples of the models presented before.

### 5.4.1. Short-Term Performance

Table 5.6 Logit with GA model short-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 1 | 94,94% | 94,76% |
| France | G and C | 1 | 92,11% | 92,22% |
| Portugal | G | 1 | 100,00% | 100,00% |
| Portugal | C | 1 | 100,00% | 100,00% |
| Portugal | G and C | 1 | 100,00% | 100,00% |
| France and Portugal | G | 1 | 93,20% | 93,24% |
| France and Portugal | G and C | 1 | 91,50% | 91,34% |

Observing the table above we can conclude that the Short-Term performance is always higher than 91,5%, wherein three models achieve 100% of accuracy.

### 5.4.2. Long-Term Performance

Table 5.7 Logit with GA model long-term performance

| Country | Sector | Year Before Bankruptcy | Accuracy | AUC |
|---|---|---|---|---|
| France | G | 2 | 86,31% | 86,16% |
| France | G | 3 | 80,99% | 80,56% |
| France | G | 4 | 82,13% | 81,98% |
| France | G and C | 2 | 86,40% | 86,40% |
| France | G and C | 3 | 86,40% | 86,13% |
| France | G and C | 4 | 82,40% | 82,43% |
| Portugal | G | 2 | 96,15% | 96,08% |
| Portugal | G | 3 | 92,31% | 92,02% |
| Portugal | G | 4 | 87,18% | 87,15% |
| Portugal | C | 2 | 91,49% | 90,91% |
| Portugal | C | 3 | 82,98% | 82,36% |
| Portugal | C | 4 | 89,36% | 88,91% |
| Portugal | G and C | 2 | 89,60% | 89,43% |
| Portugal | G and C | 3 | 84,80% | 84,44% |
| Portugal | G and C | 4 | 79,20% | 79,31% |
| France and Portugal | G | 2 | 86,22% | 86,15% |
| France and Portugal | G | 3 | 83,58% | 83,55% |
| France and Portugal | G | 4 | 82,70% | 82,65% |
| France and Portugal | G and C | 2 | 81,69% | 81,68% |
| France and Portugal | G and C | 3 | 81,30% | 81,25% |
| France and Portugal | G and C | 4 | 76,18% | 76,17% |

In a Long-Term perspective including four years before the bankruptcy, after analyzing the accuracy value for all the models presented in the table above, it is possible to observe an accuracy higher than 81% except in two models, Portugal G and C (79,20%) and France and Portugal G and C (76,18%).

In addition, in table **A 4** in the annexes is possible to observe the variable used for each model developed with the application of Logit with GA for feature selection.

## 5.5. Comparative Analysis of the Model's Performance

On this subchapter it compared the performance of the three models studied in this dissertation, modified Z-Score (MDA model), Logit Model with feature selection using Forward Stepwise and Logit Model with feature selection using Genetic Algorithms, based on the same subsamples of the models presented before.

Table 5.8 Performance Comparison between the models presented in this dissertation

| | | | GA with Logit | | Logit | | MDA | |
|---|---|---|---|---|---|---|---|---|
| Country | Sector | Year Before Bankruptcy | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC |
| France | G | 1 | 94,94% | 94,76% | 82,28% | 82,37% | 61,60% | 58,99% |
| France | G | 2 | 86,31% | 86,16% | 78,33% | 76,83% | 57,41% | 54,47% |
| France | G | 3 | 80,99% | 80,56% | 77,19% | 75,61% | 58,94% | 56,15% |
| France | G | 4 | 82,13% | 81,98% | 77,19% | 75,61% | 57,41% | 54,47% |
| France | G and C | 1 | 92,11% | 92,22% | 84,35% | 84,47% | 62,66% | 61,02% |
| France | G and C | 2 | 86,40% | 86,40% | 74,67% | 73,50% | 58,49% | 56,60% |
| France | G and C | 3 | 86,40% | 86,13% | 72,85% | 71,58% | 57,70% | 55,83% |
| France | G and C | 4 | 82,40% | 82,43% | 72,32% | 71,06% | 55,87% | 53,89% |
| Portugal | G | 1 | 100,00% | 100,00% | 82,61% | 81,82% | 73,08% | 71,75% |
| Portugal | G | 2 | 96,15% | 96,08% | 91,03% | 90,80% | 62,82% | 61,21% |
| Portugal | G | 3 | 92,31% | 92,02% | 79,49% | 79,70% | 61,54% | 59,85% |
| Portugal | G | 4 | 87,18% | 87,15% | 48,72% | 51,09% | 61,54% | 59,85% |
| Portugal | C | 1 | 100,00% | 100,00% | 78,57% | 78,57% | 89,36% | 88,91% |
| Portugal | C | 2 | 91,49% | 90,91% | 74,47% | 75,18% | 85,11% | 84,09% |
| Portugal | C | 3 | 82,98% | 82,36% | 76,60% | 77,45% | 74,47% | 73,55% |
| Portugal | C | 4 | 89,36% | 88,91% | 76,60% | 77,18% | 70,21% | 69,27% |
| Portugal | G and C | 1 | 100,00% | 100,00% | 100,00% | 100,00% | 79,20% | 78,15% |
| Portugal | G and C | 2 | 89,60% | 89,43% | 85,60% | 84,84% | 71,20% | 69,76% |
| Portugal | G and C | 3 | 84,80% | 84,44% | 80,80% | 81,01% | 66,40% | 64,95% |
| Portugal | G and C | 4 | 79,20% | 79,31% | 55,20% | 57,49% | 64,80% | 63,34% |
| France and Portugal | G | 1 | 93,20% | 93,24% | 83,33% | 83,10% | 64,22% | 61,95% |
| France and Portugal | G | 2 | 86,22% | 86,15% | 74,78% | 73,12% | 58,65% | 56,05% |

| | | | GA with Logit | | Logit | | MDA | |
|---|---|---|---|---|---|---|---|---|
| Country | Sector | Year Before Bankruptcy | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC |
| France and Portugal | G | 3 | 83,58% | 83,55% | 73,61% | 71,88% | 59,53% | 57,02% |
| France and Portugal | G | 4 | 82,70% | 82,65% | 73,61% | 71,91% | 58,36% | 55,73% |
| France and Portugal | G and C | 1 | 91,50% | 91,34% | 81,05% | 80,98% | 66,73% | 65,19% |
| France and Portugal | G and C | 2 | 81,69% | 81,68% | 77,36% | 77,34% | 61,61% | 59,80% |
| France and Portugal | G and C | 3 | 81,30% | 81,25% | 77,36% | 76,97% | 59,84% | 58,04% |
| France and Portugal | G and C | 4 | 76,18% | 76,17% | 74,41% | 73,83% | 58,07% | 56,18% |

Observing the table above we can conclude that Logit with GA model has a better performance than all the other models, except in the sub sample Portugal G and C where both models (Logit Model with GA and Logit model with forward stepwise) achieve 100% of accuracy.

An equally important aspect is the long-term efficiency of Logit models with GA, which, in almost all the situations analyzed, achieves a percentage of more than 80% of hits four years before bankruptcy. This situation is especially remarkable given the uncertainty observed in the recent past in all sectors of activity.

The following graphs present the evolution of the performance of the different models generated for each of the sectors and/or set of studied sectors.

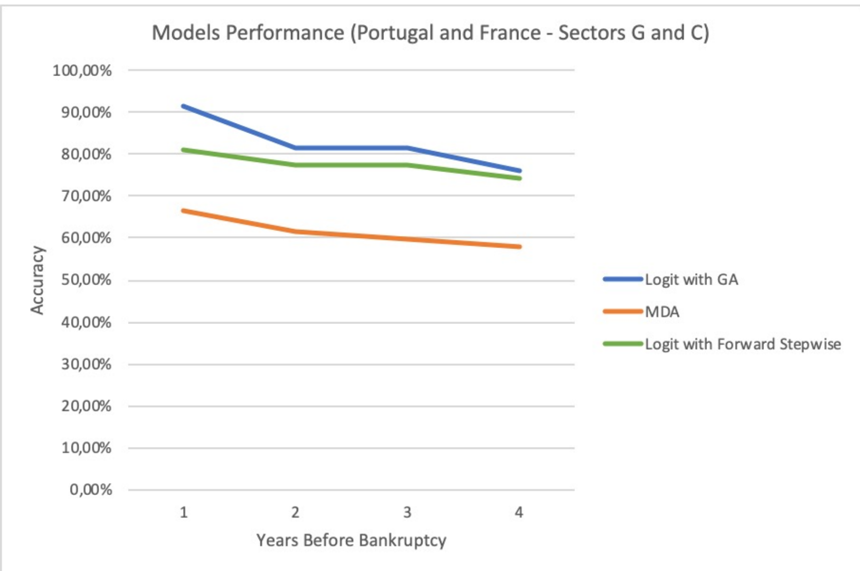



Figure 5.2 Models Performance (Portugal and France - Sectors G and C)

Moreover, from the analysis of the above plot it can be concluded that the Logit with GA model is more promising every year.

In the long term, the efficiency of Logit models with forward stepwise, approximates the performance of the logit model with GA, and also, we can conclude that the MDA model is the least efficient model regarding this sector.

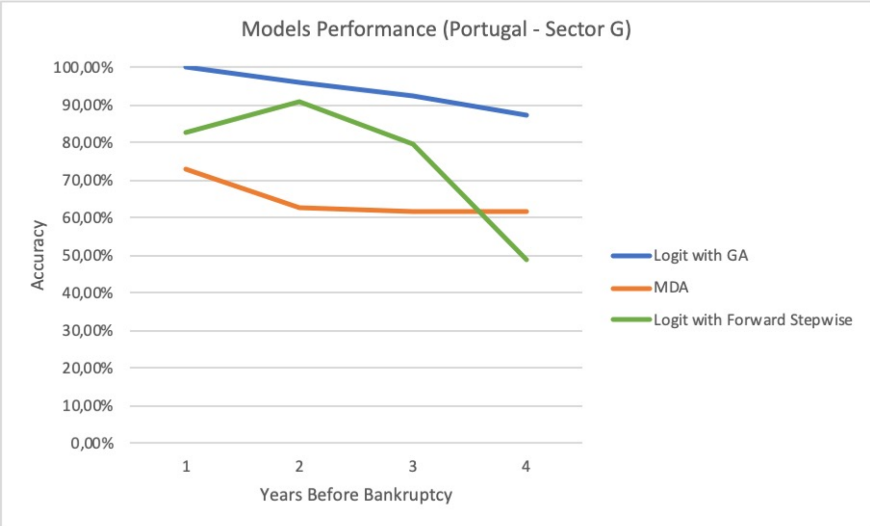



Figure 5.3 Model Performance (Portugal - Sector G)

The preceding graph presents an atypical scenario, considering what is happening in the other sectors. Logit models with forward stepwise are quite promising until the third year before bankruptcy and then lose most of its efficiency in the fourth year.

Throughout the whole period the Logit model with GA is the most efficient model, having 100% accuracy one year before the bankruptcy occurrence.

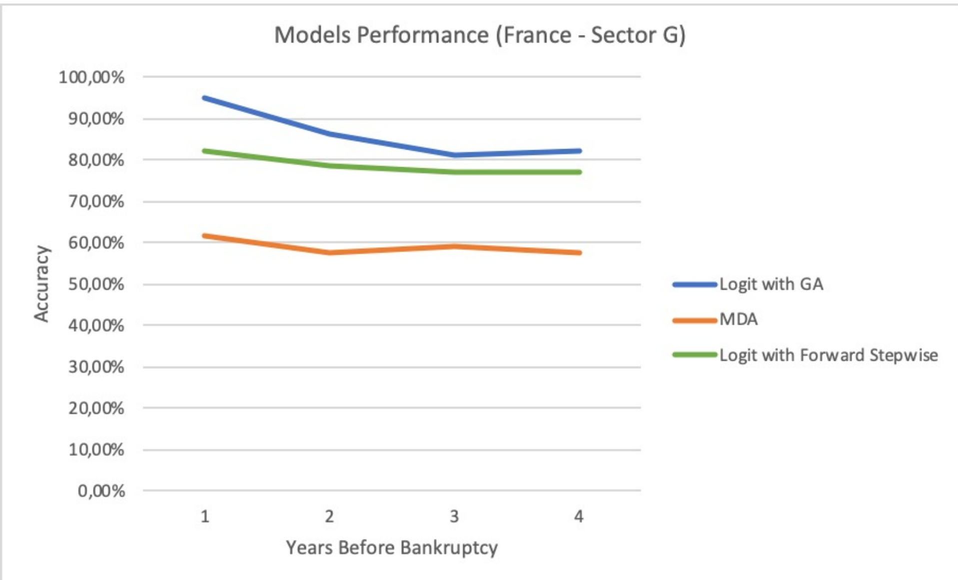



Figure 5.4 Model Performance (France - Sector G)

According to the analyze of the above plot is possible to observe that the performance of the Logit with GA model is significantly better on short-term compared with the other models regarding the data relative to France sector G. In another hand, the performance of Logit with GA approximates with the performance of Logit with Forwards Stepwise in long-term, which demonstrate that for France Sector G Logit with GA contribution continues to be the model with the best performance.

In addition, it is possible to observe that MDA continues to be the model with the worst performance achieving a constant performance of around 60% along the four years period.

Furthermore, in table **A 5** in the annexes is possible to observe all the variables used in all the models developed in this dissertation, as their description.

In the performance evaluation of the models we can also observe the type of errors generated by them. Type 1 errors - firms in which the model gives an indication that they would not bankrupt but in fact did, and Type 2 errors - firms in which the model gives an indication of a high risk of bankruptcy but in fact did not bankrupt.

Despite the fact that we consider both types of errors to be important, the respective types were calculated for all the samples studied. In the following tables we present the results that, in line with the general trend, we consider illustrative:

Table 5.9 Logit with GA vs MDA for France and Portugal Sector C and G, one year before the bankruptcy

| | | Predicted Value | |
|---|---|---|---|
| | | Non-Bankrupted | Bankrupted |
| Actual Value | | Logit With GA | |
| | Non-Bankrupted | 76 | 4 |
| | Bankrupted | 9 | 64 |
| | | MDA | |
| | Non-Bankrupted | 260 | 6 |
| | Bankrupted | 163 | 79 |

Table 5.10 Logit with GA vs MDA for France and Portugal Sector C and G, four years before the bankruptcy

| | | Predicted Value | |
|---|---|---|---|
| | | Non-Bankrupted | Bankrupted |
| Actual Value | | Logit With GA | |
| | Non-Bankrupted | 203 | 63 |
| | Bankrupted | 58 | 184 |
| | | MDA | |
| | Non-Bankrupted | 256 | 10 |
| | Bankrupted | 203 | 39 |

Table 5.11 Logit with GA vs MDA for Portugal Sector G, one year before the bankruptcy

| | | Predicted Value | |
|---|---|---|---|
| | | Non-Bankrupted | Bankrupted |
| Actual Value | | Logit With GA | |
| | Non-Bankrupted | 13 | 0 |
| | Bankrupted | 0 | 11 |
| | | MDA | |
| | Non-Bankrupted | 40 | 1 |
| | Bankrupted | 20 | 17 |

Table 5.12 Logit with GA vs MDA for Portugal Sector G, four years before the bankruptcy

| | | Predicted Value | |
|---|---|---|---|
| | | Non-Bankrupted | Bankrupted |
| Actual Value | | Logit With GA | |
| | Non-Bankrupted | 36 | 5 |
| | Bankrupted | 5 | 32 |
| | | MDA | |
| | Non-Bankrupted | 38 | 3 |
| | Bankrupted | 27 | 10 |

From the analysis of tables 5.9 to 5.12 it can be concluded that, with few exceptions, also in terms of error type classification, Logit models with GA are more promising compared to MDA. This situation can be observed both in short as well as in the long term.

# 6. CONCLUSIONS

This dissertation addressed the topic of bankruptcy prediction models, studying the MDA models, which are more valued in the state of the art compared with the proposed logit models, supported by GA and forward stepwise.

In this chapter, we present the main conclusions of the study, highlighting the results obtained by each model, as well as the comparison of the performance of the various models.

## 6.1. Research Questions and Hypothesis evaluation

Having formulated a set of research questions and hypotheses. In this section we evaluate the hypothesis and their validation or rejection.

#### 6.1.1.1. Research Questions

1. Will the application of genetic algorithms for predicting bankruptcies be promising?

   After the development and application of the models we can observe that the models where the genetic algorithms were applied, presented the best results in all the sub-samples relative to the other models, demonstrating the great potential of applying genetic algorithms for bankruptcy prediction, according to the population in analysis.

2. Will the size of the sample be a limitation to the application of GA models?

   Since the results presented by the models, developed with population of this dissertation, were promising and showed a vast diversification, were the minimum accuracy value presented was 76,18% and maximum 100% we can conclude that there is no evidence that the size of the sample or of each sub sample is a limitation for the performance and application of GA models in this dissertation.

3. Does the data from different sectors and countries will be a limitation and affect the performance of the models?

   By analyzing the results of each of the models applied to each sub-sample composed of different sectors and countries, it can be observed that the results of the models do not present a significant variation when applied to different sectors and countries, which shows that there is no evidence that data from different sectors and countries are a limitation and affect the performance.

4. Does the performance of GA predictive models have different efficiency from those based on MDA in short and long term?

   After analyzing and comparing the results of the GA (Logit with GA) and MDA (Z-score) models studied in this dissertation, we can observe that the Logit with GA models always presented better results than the MDA models in the short-term period as well in the long-term.

In addition, since the results of the Logit with GA models were always the most efficient (highest accuracy value) in short-term, there is evidence that GA have different efficiency for bankruptcy prediction from those based on MDA, as the prediction is made earlier, according with the models developed in this dissertation.

#### 6.1.1.2. Hypothesis Evaluation

H1. GA predictive models of bankruptcy are more effective than MDA predictive models.

Since the results of the GA (Logit with GA) and MDA (Z-score) models studied in this dissertation display that the Logit with GA models always presents better results than the MDA models in the short-term period as well in the long-term.

Table 6.1 Models Comparative Analysis

| | MDA (Accuracy) | | Logit with GA (Accuracy) | |
|---|---|---|---|---|
| Years Before Bankruptcy | Best Model | Worst Model | Best Model | Worst Model |
| 1 Year | 89,36% | 61,60% | 100,00% | 91,50% |
| 4 Years | 70,21% | 55,87% | 89,36% | 76,18% |

Observing the table above is possible to conclude that exists evidence that the GA predictive models of bankruptcy are more effective than MDA predictive models.

The hypothesis 1 is validated.

H2. The GA models, even with relatively small samples, maintain a good performance in supporting the prediction of bankruptcy.

Based on the research, and despite the small number of observations contained in the sample, Logit models with the contribution of GA always showed a better performance, according to the population in analysis., as we can observe in the following table:

Table 6.2 Models Results achieved with the smallest sample (Portugal Sector C)

| | MDA | | Logit with GA | |
|---|---|---|---|---|
| Years Before Bankruptcy | Accuracy | AUC | Accuracy | AUC |
| 1 Year | 73,08% | 71,75% | 100,00% | 100,00% |
| 4 Years | 61,54% | 59,85% | 87,18% | 87,15% |

The hypothesis 2 is validated.

H3. Isolation Forest is a promising method in the identification and elimination of outliers, taking into account the significant volume of economic variables involved.

In fact, the Isolation Forest outlier detection method applied in this study has shown the ability to identify and eliminate outliers, thus contributing significantly to improving the performance of the models. Although the results obtained without the use of this method were not included in this document, they proved to be quite inferior.

The hypothesis 3 is validated.

## 6.2. Conclusions

This dissertation aimed to evaluate the contribution of GA to improve the performance of bankruptcy prediction models. In parallel, new ratios were tested on the basis of accounting, financial, operating and macroeconomic framework information.

In the present work fourteen models were generated and seven more applied. The models developed corresponded to seven Logistic models with the use of stepwise forward for variable selection (Portugal sector G, Portugal sector C, Portugal sector G and C, France sector G, France and Portugal sector G, France and Portugal sector G and C), and seven Logistic models with the contribution of GA for variable selection developed through the same universe. The application of Altman's Z-Score model supported the efficiency of the models, is also applied to that very same universe of countries and sectors.

From 66 variables based on a priori knowledge were generated 1887 ratios that allowed to build logistical models with GA support on a posteriori knowledge, presenting performances with accuracy values, one year from bankruptcy ranging from 100% to 91,5% and 4 years before bankruptcy ranging from 89.36% to 76.18%. One year before bankruptcy the value of 100% accuracy is achieved for all models generated for Portugal, and the value of 91,5% through the model generated for France and Portugal G and C.

At four years the value of 89,36% accuracy is achieved for the models generated for Portugal sector C, and the value of 76,18% through the model generated for France and Portugal G and C.

However, it is normal that a model like Z-Score being developed a few years ago loses some performance compared to other models that use more variables, due to the appearance of new data and more sources of information.

From the analysis of the frequency of the ratios presented in the models regardless of their type, it is concluded that when evaluating the type of ratios included in the first quartile of frequency, the indebtedness ratios are more used around 30%, followed by the profitability ratios that correspond to around 25% and surprisingly those associated with the size of the company that correspond to a little less than 20% of the selected ratios.

According to our study, the application of GA in the selection of variables to be included in the logistic models for bankruptcy prediction results in an efficiency gain vis-à-vis logistic models without GA support and the MDA-based reference model (Edward I. Altman, 1983). It should be noted that in the feature selection based on GA, ratios that were never used in past models were selected.

Isolation Forest (Liu et al., 2008) was successfully tested in the identification of outliers to support the development of bankruptcy prediction models.

From the findings obtained, it can be concluded that both in short as well as in the long term, logistic models with the contribution of GA are more promising than the others studied, regardless of the sector and country.

Contrary to the trend manifested in the state of the art by several authors, when they analyzed different types of failure prediction models, we conclude that the models proposed in this dissertation, when generated from information from different countries at the same time and/or from different sectors at the same time, also achieve very promising results.

Based on our research it can be concluded that logistic models with GA's contribution to the selection of variables - designed for all samples, representing different countries and sectors - achieve the best performance compared to the models presented in the state of the art.

# 7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Limitations and recommendations for future work will be addressed in this chapter. The first section would display which factors have restricted the execution of this dissertation and which could have influenced the final performance.

Later, the second section address which improvements should be made in future applications in order to try to produce better results as well as models are more aligned, specific and in accordance with the financial environment.

## 7.1. LIMITATIONS

Since the models studied along with this dissertation were developed with financial data, several limitations arose that were common to other studies concerning the subject of the prediction of bankruptcy.

The main limitation was the lack of confidence we have in the data, as some stressed companies are taking action to present more favorably, such as by choosing income-increasing accounting methods or switching auditors.

Furthermore, some difficulties in searching and processing financial data have been found, as the databases used are not so user-friendly and are not well organized. Moreover, since some of the financial data in the sample are from private companies, finding and storing this data was quite difficult and costly.

Finally, since one of the primary objectives of this dissertation is to interpret the variables of the models, it was a limitation on the use of some machine and deep learning models, such as decision trees, random forests, neural networks and support vector machine (SVM) models.

## 7.2. RECOMMENDATIONS FOR FUTURE WORKS

There are indeed improvements that need to be applied in future research and development of this dissertation. The first set of improvements to consider is, naturally, to work on previously identified limitations, either to erase them or to reduce their impact.

Moreover, another study that can be carried out is to apply the models developed in this dissertation to more recent data and from other sectors and countries, in order to study their performance with even more distinct data.

Furthermore, new Logit with GA models can be developed specifically for countries and sectors other than those studied in this dissertation, in order to increase the range of economic environments and to further individualize the models to the country and sector of study.

In addition, could be promising to develop and apply logit models with GA contribution for each year before bankruptcy and sector in order to assess which are the most important variables for each year and compare short-term variables with long-term variables.

Finally, could be interesting to combine and apply each of the models developed to each of the sub-samples, i.e. for each year and sector, in order to choose the model with the best universal performance, regardless of the sector and year to be applied.

## 8. BIBLIOGRAPHY


Adnan Aziz, M., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand? *Corporate Governance: The International Journal of Business in Society*, *6*(1), 18–33. https://doi.org/10.1108/14720700610649436

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, *32*(8), 1541–1551.

Ahn, H., & Kim, K. jae. (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing Journal*, *9*(2), 599–607. https://doi.org/10.1016/j.asoc.2008.08.002

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Altman, E I. (2000). others.(2000). Predicting financial distress of companies: revisiting the Z-score and ZETA models. *Stern School of Business, New York University*.

Altman, Edward I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589. https://doi.org/10.2307/2978933

Altman, Edward I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*.

Altman, Edward I. (1993). *Corporate financial distress and bankruptcy : a complete guide to predicting & avoiding distress and profiting from bankruptcy*. Wiley.

Amadeus - European business information | Bureau van Dijk. (n.d.). Retrieved July 14, 2019, from https://www.bvdinfo.com/en-gb/our-products/data/international/amadeus

Aziz, A., Emanuel, D. C., & Lawson, G. H. (1988). Bankruptcy prediction-an investigation of cash flow based models [1]. *Journal of Management Studies*, *25*(5), 419–437.

Barniv, R., Agarwal, A., & Leach, R. (1997). Predicting the outcome following bankruptcy filing: a three-state classification using neural networks. *Intelligent Systems in Accounting, Finance & Management*, *6*(3), 177–194. https://doi.org/10.1002/(SICI)1099-1174(199709)6:3<177::AID-ISAF134>3.0.CO;2-D

Bateni, L., & Asghari, F. (2020). Bankruptcy prediction using logit and genetic algorithm models: A comparative analysis. *Computational Economics*, *55*(1), 335–348.

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

Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman’s and Ohlson’s models. *Review of Accounting Studies*, *1*(4), 267–284.

Bell, T. B. (1997). Neural nets or the logit model? A comparison of each model’s ability to predict commercial bank failures. *Intelligent Systems in Accounting, Finance & Management*, *6*(3), 249–264. https://doi.org/10.1002/(SICI)1099-1174(199709)6:3<249::AID-ISAF125>3.0.CO;2-H

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, 1–42.

Bescos, P. L. (1987). Défaillance et redressement des PMI: Recherche des indices et des causes de défaillance. *Cahier de Recherche Du CEREG*, *8701*.

Boritz, J. E., Kennedy, D. B., & Sun, J. Y. (2007). Predicting business failures in Canada. *Accounting Perspectives*, *6*(2), 141–165.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Brealey, R. A., Myers, S. C., Allen, F., & Mohanty, P. (2012). *Principles of corporate finance*. Tata McGraw-Hill Education.

Bryant, S. M. (1997). A case-based reasoning approach to bankruptcy prediction modeling. *Intelligent Systems in Accounting, Finance & Management*, *6*(3), 195–214.

Buta, P. (1994). Mining for financial knowledge with CBR. *Ai Expert*, *9*(2), 34–41.

CASTA, J., JF, C., & JP, Z. (1979). PREVOIR LA DEFAILLANCE DES ENTREPRISES. *PREVOIR LA DEFAILLANCE DES ENTREPRISES.*, 506–527.

Chung, H.-M. M., & Tam, K. Y. (1993). A Comparative Analysis of Inductive-Learning Algorithms. *Intelligent Systems in Accounting, Finance and Management*, *2*(1), 3–18. https://doi.org/10.1002/j.1099-1174.1993.tb00031.x

Colin, A. M. (1994). Genetic algorithms for financial modeling. In *Trading on the Edge* (pp. 148–173). Wiley New York.

Collongues, Y. (1977). Ratios financiers et prévision des faillites des petites et moyennes entreprises. *Revue Banque*, *365*, 963–970.

Conan, J., & Holder, M. (1979). *Variables Explicatives de Performances et Contrôle de Gestion dans les PMI Paris-Dauphine*. Thèse DEtat-Université.

Davis, L. (1991). *Handbook of genetic algorithms*.

Deboeck, G. J. (1994). Using GAs to optimize a trading system. *Trading on the Edge*, 174–188.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis*, *7*(2), 1477–1493.

Efrim Boritz, J., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems With Applications*, *9*(4), 503–512. https://doi.org/10.1016/0957-4174(95)00020-8

El Hennawy, R. H. A., & Morris, R. C. (1983). The significance of base year in developing failure prediction models. *Journal of Business Finance & Accounting*, *10*(2), 209–223.

Etheridge, H. L., & Sriram, R. S. (1997). A comparison of the relative costs of financial distress models: artificial neural networks, logit and multivariate discriminant analysis. *Intelligent Systems in Accounting, Finance & Management*, *6*(3), 235–248. https://doi.org/10.1002/(SICI)1099-1174(199709)6:3<235::AID-ISAF135>3.0.CO;2-N

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

Fletcher, D., & Goss, E. (1993). Forecasting with neural networks. An application using bankruptcy data. *Information and Management*, *24*(3), 159–167. https://doi.org/10.1016/0378-

7206(93)90064-Z

Frydman, H., Altman, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance*, *40*(1), 269–291.

Garkaz, M., & Abdollahi, A. (2010). The Investigation of Possibility of the Use of Genetic Algorithm in Predicting Companies' Bankruptcy. *Proceeding of the IEEE International Conference on Business Economic Research*, 282–285.

Gaspar, C. (2014). Risco de Crédito: A importância da Gestão de Carteiras de Crédito. *Inforbanca*, *100*, 41–43.

Genetic Algorithms - Introduction - Tutorialspoint. (n.d.). Retrieved April 7, 2020, from https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_introduction.htm

Genetic Algorithms - Parent Selection - Tutorialspoint. (n.d.). Retrieved April 7, 2020, from https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_parent_selection.htm

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Retrieved from https://books.google.pt/books?id=3_RQAAAAMAAJ

Gour, R. (n.d.). Python Genetic Algorithms With Artificial Intelligence. Retrieved April 7, 2020, from https://medium.com/@rinu.gour123/python-genetic-algorithms-with-artificial-intelligence-b8d0c7db60ac

Grice, J. S., & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. *Journal of Business Research*, *54*(1), 53–61.

Han, I., Chandler, J. S., & Liang, T.-P. (1996). The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods. *Expert Systems with Applications*, *10*(2), 209–221.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hanson, D. (n.d.). Here's a list of recession signals that are flashing red. Retrieved November 17, 2019, from https://www.cnbc.com/2019/09/02/heres-a-list-of-recession-signals-that-are-flashing-red.html

Holland, J. H. (1975). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. In *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* Oxford, England: U Michigan Press.

How a No-Deal Brexit May Become a Problem for the World Economy - Bloomberg. (n.d.). Retrieved November 17, 2019, from https://www.bloomberg.com/news/articles/2019-10-04/how-a-no-deal-brexit-may-become-a-problem-for-the-world-economy

Hsieh, S. (1993). A note on the optimal cutoff point in bankruptcy prediction models. *Journal of Business Finance & Accounting*, *20*(3), 457–464.

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, *13*(2), 97–108. https://doi.org/10.1016/S0957-4174(97)00011-0

Kahya, E., & Theodossiou, P. (1999). Predicting corporate financial distress: A time-series CUSUM methodology. *Review of Quantitative Finance and Accounting*, *13*(4), 323–345.

Kim, M.-J., & Kang, D.-K. (2012). Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction. *Expert Systems with Applications*, *39*(10), 9308–9314. https://doi.org/10.1016/J.ESWA.2012.02.072

Kingdon, J., & Feldman, K. (1995). *Genetic algorithms for bankruptcy prediction*. SearchSpace Research Report 1-95, SearchSpace Ltd, London.

Koh, H. C., & Killough, L. N. (1990). The use of multiple discriminant analysis in the assessment of the going-concern status of an audit client. *Journal of Business Finance & Accounting*, *17*(2), 179–192.

Koza, J. R. (1992). *Genetic programming : on the programming of computers by means of natural selection*. MIT Press.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research*, *180*(1), 1–28.

Laitinen, T., & Kankaanpaa, M. (1999). Comparative analysis of failure prediction methods: the Finnish case. *European Accounting Review*, *8*(1), 67–92.

Lee, C. F. (1985). Financial analysis and planning : theory and application. In *Published in **1985** in Reading Mass) by Addison-Wesley*. Addison-Wesley.

Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research*, *169*(2), 677–697. https://doi.org/10.1016/J.EJOR.2004.06.013

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. IEEE.

Mahfoud, S., & Mani, G. (1995). Genetic algorithms for predicting individual stock performance. *Proceedings of the Third International Conference on Artificial Intelligence Applications on Wall Street*, 174–181.

Malécot, J.-F. (1991). Analyses théoriques des défaillances d’entreprises : Une revue de la littérature. *Revue d’économie Financière*, *19*(4), 205–227. https://doi.org/10.3406/ecofi.1991.1746

McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research*, *138*(2), 436–451.

Mehrazin, A., Taghipour, M., Ghabdian, B., & Soleimani, H. (2013). Radial basis function in artificial neural network for prediction of bankruptcy. *International Business Research*, *6*(8), 121.

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

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *IJCNN. International Joint Conference on Neural Networks*, 163–168. https://doi.org/10.1109/ijcnn.1990.137710

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

Ooghe, H., & Balcaen, S. (2004). 35 years of studies on business failure: an overview of the classical statistical methodologies and their related problems. *Working Paper Series Faculteit Economie En Bedrijfskunde*, *2004*, 1–56.

Peres, C., & Antão, M. (2017). The use of multivariate discriminant analysis to predict corporate bankruptcy: A review. *Aestimatio: The IEB International Journal of Finance*, (14), 108–131.

Pettway, R. H., & Sinkey, J. F. (1980). Establishing on-site bank examination priorities: An early-warning system using accounting and market information. *The Journal of Finance*, *35*(1), 137–150.

Pompe, P. P. M., & Bilderbeek, J. (2005). Bankruptcy prediction: The influence of the year prior to failure selected for model building and the effects in a period of economic decline. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(2), 95–112.

PORDATA - Statistics, charts and indicators on Municipalities, Portugal and Europe. (n.d.). Retrieved January 30, 2019, from https://www.pordata.pt/en/Home

Romero-Hdz, J., Aranda, S., Toledo-Ramirez, G., Segura, J., & Saha, B. (2016). An Elitism Based Genetic Algorithm for Welding Sequence Optimization to Reduce Deformation. *Research in Computing Science*, *121*, 17–36.

Ross, S. A., Westerfield, R. W., & Jaffe, J. (2002). Corporate finance. McGraw-Hill Irwin. *New York*.

Rutan, E. (1993). Experiments with optimal stock screens. *Proceedings of the 3rd International Conference on Artificial Intelligence Applications on Wall Street*, 269–273.

Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992). Neural Networks: A New Tool for Predicting Thrift Failures. *Decision Sciences*, *23*(4), 899–916. https://doi.org/10.1111/j.1540-5915.1992.tb00425.x

Shalev-Shwartz, S., Shamir, O., & Shammah, S. (2017). Failures of gradient-based deep learning. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3067–3075. JMLR. org.

Shaw, M. J., & Gentry, J. A. (1988). Using an expert system with inductive learning to evaluate business loans. *Financial Management*, 45–56.

Sheppard, J. P. (1994). The dilemma of matched pairs and diversified firms in bankruptcy prediction models. *The Mid-Atlantic Journal of Business*, *30*(1), 9.

Shin, K.-S., Lee, T. S., & Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, *28*(1), 127–135.

Shin, K.-S., & Lee, Y.-J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, *23*(3), 321–328. https://doi.org/10.1016/S0957-4174(02)00051-9

Shin, K., & Han, I. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications*, *16*(2), 85–95.

Sikora, R., & Shaw, M. (1994). A double-layered learning approach to acquiring rules for classification: Integrating genetic algorithms with similarity-based learning. *ORSA Journal on Computing*, *6*(2), 174–187.

Sinkey Jr, J. F. (1975). A multivariate statistical analysis of the characteristics of problem banks. *The Journal of Finance*, *30*(1), 21–36.

Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41–56.

Syswerda, G. (1989). Uniform crossover in genetic algorithms. *Proceedings of the 3rd International Conference on Genetic Algorithms*, 2–9.

Theodossiou, P. T. (1993). Predicting shifts in the mean of a multivariate time series process: an application in predicting business failures. *Journal of the American Statistical Association*, *88*(422), 441–449.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking and Finance*, *22*(10–11), 1421–1439. https://doi.org/10.1016/S0378-4266(98)00059-4

Wong, F., & Tan, C. (1994). Hybrid neural, genetic, and fuzzy systems. *Trading on the Edge*, 243–261.

Wu, C.-H., Tzeng, G.-H., Goo, Y.-J., & Fang, W.-C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications*, *32*(2), 397–408.

Xu, M., & Zhang, C. (2009). Bankruptcy prediction: the case of Japanese listed companies. *Review of Accounting Studies*, *14*(4), 534–558.

Yang, Y. (2014). Does high-quality auditing decrease the use of collateral? Analysis from the perspective of lenders' self-protection. *China Journal of Accounting Research*, *7*(3), 203–221.

Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: a logistic analysis. *Journal of Business Finance & Accounting*, *12*(1), 19–45.

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(4), 513–522.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*, *116*(1), 16–32.

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59. https://doi.org/10.2307/2490859

# 9. ANNEXES

Table 9.1 Initial Variables Description, represented in Table 9.2

| Description | |
|---|---|
| | Classification Variable of the Company |
| | Macroeconomic Variable |
| | Non Macroeconomic Variable |
| | Ratio |

Table 9.2 A 1 - Initial Variables

| Var ID | Variable Name |
|---|---|
| | Age |
| | Number of Employees |
| A1 | Standard VAT Rate |
| A2 | Yield on 10-year Treasury bonds |
| A3 | Effective Remaining Tax Rate |
| A4 | Effective Financing Rate |
| A5 | Sector Business Volume |
| A6 | Gross Operating Surplus |
| A7 | Sector Investment |
| A8 | Resident Population |
| A9 | GDP per Capita |
| A10 | Public Deficit / Surplus |
| A11 | PA Expenditure |
| A12 | General Government Gross Debt |
| A13 | Net External Debt |
| A14 | Exports |
| A15 | Gross Fixed Capital Formation |
| A16 | Imports |
| A17 | PA Revenues |
| A18 | Balance of Trade |
| A19 | Current Account Balance |
| A20 | Capital Account Balance |
| A21 | Financial Balance |
| A22 | Closing Market Index |
| A23 | Intangible Assets |
| A24 | Tangible Fixed Assets |
| A25 | Other Non-Current Assets |
| A26 | Non Current Assets |

| Var ID | Variable Name |
|---|---|
| A27 | Inventories |
| A28 | Third Party Debts-C |
| A29 | Other Current Assets |
| A30 | Bank Deposits and Cash |
| A31 | Current Assets |
| A32 | Total Assets |
| A33 | Share Capital |
| A34 | Retained Earnings or Other Equity |
| A35 | Equity |
| A36 | Debts to Third Parties-NC |
| A37 | Other Liabilities-NC |
| A38 | Non-Current Liabilities |
| A39 | Financial Debts-C |
| A40 | Debts to Third Parties-C |
| A41 | Other liability-C |
| A42 | Current Liabilities |
| A43 | Total Liabilities |
| A44 | Operational Income |
| A45 | Cost of Goods Sold and Materials Cons. |
| A46 | Gross Margin |
| A47 | Personnel Expenditure |
| A48 | Other Operational Items -FSE and Others |
| A49 | Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) |
| A50 | Depreciation and Amortization |
| A51 | Operational Result |
| A52 | Financial Results |
| A53 | Current Result |
| A54 | Income Tax |
| A55 | Net Profit for the Year |
| A56 | Turnover |
| A57 | Interest Expenses |
| A58 | Cash Flows |
| A59 | Gross Value Added |
| A60 | Operating expenses: (Sales - EBT - Adjustments) |
| A61 | Permanent Capitals: Equity + Non Current Liabilities |
| A62 | Working Capital |
| A63 | Working Capital Requirements |
| A64 | Net Cashflow |
| A65 | Self-financing: NR + Adjustments |
| A66 | Capital Invested: NCA + WCN + AT |
| A67 | Net Operating Profit After Tax |

| Var ID | Variable Name |
|---|---|
| A68 | Earnings Before Interest, Depreciation, Amortization and after Tax |

1

2 Table 9.3 A 2 - Outlier Treatment Summary Results

| Country | Sector | BK or NBK | Year before Bankruptcy | Initial Nº of Companies | Outliers | Nº of Companies Without Outliers |
|---|---|---|---|---|---|---|
| PT | C | Bankrupt | 2 | 25 | 3 | 22 |
| PT | C | Bankrupt | 3 | 25 | 3 | 22 |
| PT | C | Bankrupt | 4 | 25 | 3 | 22 |
| PT | C | Bankrupt | 5 | 25 | 3 | 22 |
| PT | G | Bankrupt | 2 | 41 | 4 | 37 |
| PT | G | Bankrupt | 3 | 41 | 4 | 37 |
| PT | G | Bankrupt | 4 | 41 | 4 | 37 |
| PT | G | Bankrupt | 5 | 41 | 4 | 37 |
| PT | C | Not Bankrupt | 2 | 28 | 3 | 25 |
| PT | C | Not Bankrupt | 3 | 28 | 3 | 25 |
| PT | C | Not Bankrupt | 4 | 28 | 3 | 25 |
| PT | C | Not Bankrupt | 5 | 28 | 3 | 25 |
| PT | G | Not Bankrupt | 2 | 46 | 5 | 41 |
| PT | G | Not Bankrupt | 3 | 46 | 5 | 41 |
| PT | G | Not Bankrupt | 4 | 46 | 5 | 41 |
| PT | G | Not Bankrupt | 5 | 46 | 5 | 41 |
| FR | C | Bankrupt | 2 | 67 | 7 | 60 |
| FR | C | Bankrupt | 3 | 67 | 7 | 60 |
| FR | C | Bankrupt | 4 | 67 | 7 | 60 |
| FR | C | Bankrupt | 5 | 67 | 7 | 60 |
| FR | G | Bankrupt | 2 | 137 | 14 | 123 |
| FR | G | Bankrupt | 3 | 137 | 14 | 123 |
| FR | G | Bankrupt | 4 | 137 | 14 | 123 |
| FR | G | Bankrupt | 5 | 137 | 14 | 123 |
| FR | C | Not Bankrupt | 2 | 67 | 7 | 60 |
| FR | C | Not Bankrupt | 3 | 67 | 7 | 60 |
| FR | C | Not Bankrupt | 4 | 67 | 7 | 60 |
| FR | C | Not Bankrupt | 5 | 67 | 7 | 60 |
| FR | G | Not Bankrupt | 2 | 156 | 16 | 140 |
| FR | G | Not Bankrupt | 3 | 156 | 16 | 140 |
| FR | G | Not Bankrupt | 4 | 156 | 16 | 140 |
| FR | G | Not Bankrupt | 5 | 156 | 16 | 140 |

3

Table 9.4 A 3 - Logit with Forward Stepwise Variables

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R13 | R13 | R13 | R1514 | R1333 | R1333 | R1333 |
| R882 | R10 | R854 | R1057 | R13 | R13 | R19 |
| R10 | R35 | | | R1322 | R1869 | R1690 |
| R1734 | R850 | | | R1013 | R1865 | R962 |
| R1103 | R1801 | | | R1630 | R10 | R1454 |
| R945 | R1401 | | | | R843 | R39 |
| R829 | R1463 | | | | R1025 | R1458 |
| R1674 | R1846 | | | | R1154 | R1641 |
| R1275 | R886 | | | | R1125 | R1869 |
| R1005 | R1743 | | | | R1766 | R1030 |
| | R1249 | | | | R1760 | R1373 |
| | R1116 | | | | R1329 | R1155 |
| | | | | | R1017 | R1003 |
| | | | | | R1005 | R1595 |
| | | | | | | R927 |
| | | | | | | R5 |
| | | | | | | R1296 |
| | | | | | | R14 |
| | | | | | | R934 |
| | | | | | | R968 |
| | | | | | | R1352 |
| | | | | | | R13R1005 |
| | | | | | | R1258 |
| | | | | | | R930 |
| | | | | | | R1317 |
| | | | | | | R1195 |
| | | | | | | R1303 |
| | | | | | | R3 |
| | | | | | | R1051 |
| | | | | | | R1241 |
| | | | | | | R1311 |
| | | | | | | R1058 |
| | | | | | | R966 |

1 Table 9.5 A 4 - Logit with GA Variables

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R2 | R1 | R1 | R1 | R2 | R1 | R1 |
| R5 | R2 | R6 | R2 | R7 | R5 | R3 |
| R6 | R3 | R9 | R4 | R11 | R6 | R4 |
| R9 | R6 | R14 | R5 | R20 | R8 | R9 |
| R10 | R10 | R17 | R8 | R21 | R9 | R11 |
| R11 | R11 | R18 | R10 | R22 | R10 | R13 |
| R16 | R13 | R20 | R11 | R24 | R11 | R14 |
| R17 | R23 | R22 | R13 | R26 | R12 | R15 |
| R18 | R28 | R24 | R14 | R28 | R14 | R16 |
| R19 | R31 | R27 | R16 | R32 | R15 | R17 |
| R27 | R35 | R28 | R17 | R33 | R18 | R19 |
| R28 | R37 | R32 | R19 | R35 | R19 | R23 |
| R31 | R39 | R33 | R22 | R41 | R23 | R27 |
| R32 | R40 | R37 | R24 | R42 | R27 | R29 |
| R33 | R41 | R40 | R27 | R43 | R28 | R33 |
| R37 | R42 | R41 | R28 | R831 | R29 | R39 |
| R39 | R43 | R43 | R33 | R832 | R32 | R42 |
| R40 | R830 | R829 | R35 | R833 | R33 | R43 |
| R41 | R837 | R832 | R36 | R839 | R35 | R829 |
| R45 | R838 | R835 | R39 | R843 | R39 | R830 |
| R829 | R842 | R846 | R44 | R846 | R40 | R833 |
| R831 | R846 | R854 | R844 | R857 | R41 | R834 |
| R835 | R855 | R877 | R862 | R858 | R829 | R838 |
| R838 | R859 | R882 | R870 | R869 | R830 | R839 |
| R839 | R866 | R883 | R874 | R875 | R832 | R841 |
| R846 | R869 | R892 | R877 | R876 | R839 | R842 |
| R850 | R870 | R895 | R879 | R877 | R842 | R845 |
| R854 | R884 | R899 | R883 | R880 | R845 | R846 |
| R855 | R886 | R902 | R889 | R882 | R846 | R849 |
| R859 | R898 | R908 | R894 | R884 | R849 | R851 |
| R864 | R900 | R921 | R895 | R894 | R850 | R855 |
| R866 | R901 | R922 | R896 | R895 | R854 | R864 |
| R867 | R902 | R925 | R903 | R903 | R857 | R869 |
| R870 | R903 | R927 | R904 | R908 | R859 | R870 |
| R874 | R904 | R928 | R911 | R913 | R867 | R876 |
| R878 | R911 | R936 | R912 | R922 | R875 | R880 |
| R882 | R912 | R942 | R915 | R923 | R877 | R881 |
| R885 | R914 | R943 | R920 | R925 | R884 | R885 |

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R886 | R923 | R944 | R923 | R928 | R888 | R886 |
| R889 | R925 | R945 | R924 | R930 | R889 | R891 |
| R893 | R927 | R955 | R931 | R934 | R893 | R895 |
| R901 | R933 | R969 | R936 | R942 | R898 | R897 |
| R902 | R937 | R975 | R938 | R943 | R899 | R898 |
| R903 | R944 | R992 | R940 | R955 | R900 | R900 |
| R910 | R947 | R1006 | R942 | R969 | R902 | R901 |
| R914 | R951 | R1013 | R946 | R981 | R908 | R902 |
| R918 | R953 | R1014 | R955 | R992 | R911 | R903 |
| R925 | R955 | R1016 | R956 | R1003 | R912 | R911 |
| R927 | R957 | R1017 | R959 | R1005 | R913 | R913 |
| R929 | R962 | R1021 | R984 | R1006 | R920 | R919 |
| R930 | R966 | R1023 | R1006 | R1007 | R923 | R920 |
| R932 | R975 | R1036 | R1010 | R1015 | R926 | R923 |
| R933 | R976 | R1038 | R1013 | R1016 | R933 | R930 |
| R950 | R977 | R1056 | R1016 | R1017 | R936 | R934 |
| R953 | R983 | R1059 | R1019 | R1019 | R937 | R942 |
| R954 | R986 | R1061 | R1022 | R1021 | R938 | R946 |
| R966 | R988 | R1062 | R1025 | R1022 | R942 | R947 |
| R970 | R993 | R1078 | R1034 | R1023 | R946 | R953 |
| R975 | R997 | R1090 | R1040 | R1033 | R947 | R955 |
| R976 | R1004 | R1092 | R1045 | R1038 | R955 | R956 |
| R980 | R1005 | R1103 | R1047 | R1039 | R959 | R959 |
| R983 | R1008 | R1107 | R1053 | R1041 | R962 | R961 |
| R1004 | R1009 | R1125 | R1057 | R1047 | R966 | R966 |
| R1005 | R1010 | R1127 | R1059 | R1050 | R975 | R968 |
| R1006 | R1022 | R1138 | R1061 | R1056 | R976 | R976 |
| R1007 | R1023 | R1140 | R1076 | R1058 | R977 | R978 |
| R1015 | R1025 | R1151 | R1084 | R1061 | R978 | R979 |
| R1016 | R1033 | R1157 | R1085 | R1062 | R982 | R981 |
| R1019 | R1034 | R1160 | R1090 | R1078 | R992 | R1003 |
| R1025 | R1035 | R1173 | R1093 | R1091 | R1006 | R1006 |
| R1028 | R1038 | R1234 | R1095 | R1094 | R1010 | R1009 |
| R1029 | R1043 | R1238 | R1104 | R1103 | R1016 | R1016 |
| R1030 | R1050 | R1241 | R1105 | R1105 | R1018 | R1017 |
| R1043 | R1055 | R1242 | R1113 | R1109 | R1019 | R1023 |
| R1044 | R1056 | R1249 | R1128 | R1119 | R1022 | R1024 |
| R1046 | R1058 | R1250 | R1153 | R1133 | R1027 | R1027 |
| R1048 | R1060 | R1252 | R1157 | R1136 | R1030 | R1029 |
| R1050 | R1062 | R1261 | R1173 | R1138 | R1033 | R1033 |

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R1051 | R1063 | R1270 | R1177 | R1149 | R1035 | R1035 |
| R1054 | R1067 | R1271 | R1211 | R1151 | R1036 | R1039 |
| R1055 | R1069 | R1274 | R1212 | R1157 | R1040 | R1041 |
| R1056 | R1074 | R1283 | R1240 | R1171 | R1041 | R1046 |
| R1058 | R1075 | R1297 | R1244 | R1173 | R1043 | R1048 |
| R1062 | R1076 | R1298 | R1249 | R1176 | R1046 | R1049 |
| R1063 | R1080 | R1307 | R1250 | R1195 | R1048 | R1051 |
| R1068 | R1094 | R1311 | R1254 | R1231 | R1049 | R1056 |
| R1072 | R1096 | R1320 | R1261 | R1234 | R1050 | R1057 |
| R1074 | R1101 | R1334 | R1265 | R1238 | R1051 | R1058 |
| R1075 | R1102 | R1347 | R1279 | R1240 | R1053 | R1062 |
| R1080 | R1103 | R1369 | R1283 | R1241 | R1054 | R1063 |
| R1083 | R1106 | R1389 | R1284 | R1242 | R1056 | R1080 |
| R1085 | R1107 | R1431 | R1289 | R1246 | R1057 | R1081 |
| R1088 | R1113 | R1451 | R1292 | R1247 | R1058 | R1083 |
| R1089 | R1114 | R1452 | R1293 | R1249 | R1060 | R1094 |
| R1095 | R1123 | R1458 | R1296 | R1252 | R1062 | R1096 |
| R1103 | R1126 | R1460 | R1297 | R1258 | R1064 | R1103 |
| R1106 | R1127 | R1461 | R1299 | R1259 | R1067 | R1105 |
| R1113 | R1130 | R1464 | R1312 | R1268 | R1069 | R1114 |
| R1116 | R1132 | R1465 | R1316 | R1275 | R1070 | R1115 |
| R1117 | R1133 | R1468 | R1326 | R1278 | R1075 | R1120 |
| R1126 | R1137 | R1486 | R1332 | R1282 | R1081 | R1123 |
| R1128 | R1145 | R1493 | R1337 | R1283 | R1085 | R1128 |
| R1130 | R1147 | R1494 | R1341 | R1284 | R1095 | R1130 |
| R1134 | R1153 | R1502 | R1348 | R1293 | R1096 | R1131 |
| R1151 | R1155 | R1507 | R1353 | R1297 | R1101 | R1134 |
| R1153 | R1156 | R1510 | R1364 | R1298 | R1103 | R1136 |
| R1154 | R1157 | R1520 | R1368 | R1299 | R1106 | R1140 |
| R1157 | R1163 | R1579 | R1370 | R1308 | R1115 | R1142 |
| R1170 | R1188 | R1598 | R1380 | R1311 | R1116 | R1147 |
| R1171 | R1194 | R1620 | R1384 | R1316 | R1123 | R1156 |
| R1172 | R1235 | R1622 | R1391 | R1339 | R1126 | R1157 |
| R1188 | R1240 | R1627 | R1392 | R1341 | R1133 | R1162 |
| R1208 | R1245 | R1630 | R1394 | R1343 | R1138 | R1163 |
| R1235 | R1249 | R1661 | R1429 | R1351 | R1139 | R1171 |
| R1240 | R1253 | R1680 | R1430 | R1352 | R1140 | R1194 |
| R1242 | R1259 | R1698 | R1446 | R1358 | R1151 | R1240 |
| R1245 | R1261 | R1708 | R1461 | R1363 | R1153 | R1241 |
| R1247 | R1262 | R1729 | R1464 | R1382 | R1155 | R1247 |

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R1248 | R1264 | R1733 | R1486 | R1389 | R1156 | R1250 |
| R1251 | R1265 | R1737 | R1488 | R1391 | R1164 | R1252 |
| R1253 | R1266 | R1738 | R1505 | R1393 | R1171 | R1255 |
| R1257 | R1274 | R1757 | R1509 | R1401 | R1172 | R1258 |
| R1266 | R1281 | R1762 | R1517 | R1408 | R1191 | R1260 |
| R1271 | R1283 | R1766 | R1543 | R1413 | R1195 | R1265 |
| R1272 | R1285 | R1768 | R1565 | R1431 | R1235 | R1268 |
| R1273 | R1288 | R1802 | R1575 | R1433 | R1240 | R1275 |
| R1278 | R1289 | R1803 | R1579 | R1439 | R1241 | R1288 |
| R1284 | R1297 | R1808 | R1611 | R1440 | R1242 | R1300 |
| R1285 | R1298 | R1819 | R1620 | R1451 | R1245 | R1302 |
| R1286 | R1321 | R1846 | R1635 | R1452 | R1250 | R1303 |
| R1287 | R1333 | R1848 | R1641 | R1463 | R1251 | R1313 |
| R1289 | R1340 | R1868 | R1666 | R1465 | R1252 | R1316 |
| R1296 | R1341 | R1869 | R1686 | R1478 | R1253 | R1317 |
| R1298 | R1349 | R1884 | R1687 | R1486 | R1260 | R1321 |
| R1305 | R1357 | | R1690 | R1493 | R1262 | R1326 |
| R1313 | R1362 | | R1731 | R1498 | R1264 | R1328 |
| R1317 | R1363 | | R1741 | R1502 | R1266 | R1337 |
| R1327 | R1370 | | R1789 | R1507 | R1267 | R1344 |
| R1328 | R1371 | | R1820 | R1510 | R1271 | R1345 |
| R1329 | R1378 | | R1848 | R1565 | R1278 | R1346 |
| R1333 | R1382 | | R1860 | R1575 | R1280 | R1350 |
| R1340 | R1383 | | R1868 | R1579 | R1284 | R1351 |
| R1349 | R1384 | | R1874 | R1620 | R1287 | R1352 |
| R1351 | R1400 | | R1880 | R1622 | R1297 | R1353 |
| R1352 | R1401 | | | R1625 | R1298 | R1356 |
| R1368 | R1403 | | | R1630 | R1305 | R1358 |
| R1371 | R1408 | | | R1641 | R1317 | R1362 |
| R1373 | R1427 | | | R1661 | R1327 | R1370 |
| R1378 | R1431 | | | R1671 | R1329 | R1379 |
| R1379 | R1434 | | | R1685 | R1333 | R1380 |
| R1380 | R1441 | | | R1686 | R1339 | R1381 |
| R1387 | R1442 | | | R1687 | R1341 | R1383 |
| R1389 | R1451 | | | R1717 | R1345 | R1384 |
| R1391 | R1459 | | | R1729 | R1353 | R1387 |
| R1392 | R1464 | | | R1731 | R1363 | R1391 |
| R1401 | R1465 | | | R1737 | R1370 | R1395 |
| R1429 | R1477 | | | R1741 | R1371 | R1400 |
| R1430 | R1489 | | | R1760 | R1373 | R1401 |

| France - Sector G | France - Sector G and C | Portugal - Sector G | Portugal - Sector C | Portugal - Sector G and C | France and Portugal - Sector G | France and Portugal - Sector G and C |
|---|---|---|---|---|---|---|
| R1431 | R1495 | | | R1762 | R1380 | R1403 |
| R1434 | R1502 | | | R1763 | R1384 | R1407 |
| R1435 | R1516 | | | R1771 | R1389 | R1413 |
| R1441 | R1517 | | | R1772 | R1391 | R1431 |
| R1442 | R1565 | | | R1798 | R1395 | R1433 |
| R1464 | R1575 | | | R1802 | R1400 | R1436 |
| R1477 | R1579 | | | R1843 | R1401 | R1439 |
| R1489 | R1611 | | | R1846 | R1403 | R1443 |
| R1508 | R1625 | | | R1860 | R1414 | R1458 |
| R1565 | R1651 | | | R1880 | R1430 | R1459 |
| R1575 | R1674 | | | R1884 | R1431 | R1461 |
| R1578 | R1689 | | | | R1433 | R1464 |
| R1579 | R1690 | | | | R1434 | R1465 |
| R1585 | R1692 | | | | R1442 | R1469 |
| R1598 | R1708 | | | | R1446 | R1477 |
| R1644 | R1731 | | | | R1451 | R1478 |
| R1645 | R1732 | | | | R1459 | R1493 |
| R1646 | R1733 | | | | R1463 | R1508 |
| R1676 | R1735 | | | | R1464 | R1510 |
| R1682 | R1738 | | | | R1486 | R1512 |
| R1684 | R1741 | | | | R1488 | R1566 |
| R1686 | R1760 | | | | R1508 | R1578 |
| R1690 | R1762 | | | | R1510 | R1611 |
| R1701 | R1765 | | | | R1512 | R1613 |
| R1731 | R1766 | | | | R1566 | R1622 |
| R1734 | R1767 | | | | R1579 | R1625 |
| R1737 | R1770 | | | | R1580 | R1630 |
| R1739 | R1772 | | | | R1585 | R1633 |
| R1762 | R1799 | | | | R1611 | R1641 |
| R1766 | R1808 | | | | R1630 | R1646 |
| R1768 | R1845 | | | | R1633 | R1651 |
| R1770 | R1846 | | | | R1645 | R1685 |
| R1798 | R1847 | | | | R1665 | R1689 |
| R1799 | R1868 | | | | R1684 | R1690 |
| R1818 | R1880 | | | | R1685 | R1692 |
| R1847 | | | | | R1686 | R1701 |
| R1874 | | | | | R1690 | R1703 |

1

1 Table 9.6 A 5 - Variables (used in the models) Description

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R1 | (Sector Turnover) // (Intangible Assets) | R1036 | (Inventories) // (Permanent Capitals: Equity + Non Current Liabilities) | R1347 | (Debts to Third Parties-NC) // (Other Operational Items - FSE and Others) |
| R2 | (Sector Turnover) // (Tangible Fixed Assets) | R1038 | (Inventories) // (Working Capital Requirements ) | R1348 | (Debts to Third Parties-NC) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) |
| R3 | (Sector Turnover) // (Other Non-Current Assets) | R1039 | (Inventories) // (Net Cashflow ) | R1349 | (Debts to Third Parties-NC) // (Depreciation and Amortization) |
| R4 | (Sector Turnover) // (Non Current Assets) | R1040 | (Inventories) // (Self-financing: NR + Adjustments) | R1350 | (Debts to Third Parties-NC) // (Operational Result (EBIT) |
| R5 | (Sector Turnover) // (Inventories) | R1041 | (Inventories) // (Invested Capital: NCA + WC + TA) | R1351 | (Debts to Third Parties-NC) // (Financial Results) |
| R6 | (Sector Turnover) // (Third Party Debts-C) | R1043 | (Inventories) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) | R1352 | (Debts to Third Parties-NC) // (Current Result) |
| R7 | (Sector Turnover) // (Other Current Assets) | R1044 | (Third Party Debts-C) // (Other Current Assets) | R1353 | (Debts to Third Parties-NC) // (Income Tax) |
| R8 | (Sector Turnover) // (Cash) | R1045 | (Third Party Debts-C) // (Cash) | R1356 | (Debts to Third Parties-NC) // (Interest Expenses) |
| R9 | (Sector Turnover) // (Current Assets) | R1046 | (Third Party Debts-C) // (Current Assets) | R1357 | (Debts to Third Parties-NC) // (Cash Flows) |
| R10 | (Sector Turnover) // (Total Assets) | R1047 | (Third Party Debts-C) // (Total Assets) | R1358 | (Debts to Third Parties-NC) // (Gross Value Added) |
| R11 | (Sector Turnover) // (Share Capital) | R1048 | (Third Party Debts-C) // (Share Capital) | R1362 | (Debts to Third Parties-NC) // (Working Capital Requirements ) |
| R12 | (Sector Turnover) // (Retained Earnings or Other Equity) | R1049 | (Third Party Debts-C) // (Retained Earnings or Other Equity) | R1363 | (Debts to Third Parties-NC) // (Net Cashflow ) |
| R13 | (Sector Turnover) // (Equity) | R1050 | (Third Party Debts-C) // (Equity) | R1364 | (Debts to Third Parties-NC) // (Self-financing: NR + Adjustments) |
| R14 | (Sector Turnover) // (Debts to Third Parties-NC) | R1051 | (Third Party Debts-C) // (Debts to Third Parties-NC) | R1368 | (Other Liabilities-NC) // (Non-Current Liabilities) |
| R15 | (Sector Turnover) // (Other Liabilities-NC) | R1053 | (Third Party Debts-C) // (Non-Current Liabilities) | R1369 | (Other Liabilities-NC) // (Financial Debts-C) |
| R16 | (Sector Turnover) // (Non-Current Liabilities) | R1054 | (Third Party Debts-C) // (Financial Debts-C) | R1370 | (Other Liabilities-NC) // (Debts to Third Parties-C) |
| R17 | (Sector Turnover) // (Financial Debts-C) | R1055 | (Third Party Debts-C) // (Debts to Third Parties-C) | R1371 | (Other Liabilities-NC) // (Other liability-C) |
| R18 | (Sector Turnover) // (Debts to Third Parties-C) | R1056 | (Third Party Debts-C) // (Other liability-C) | R1373 | (Other Liabilities-NC) // (Total Liabilities) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R19 | (Sector Turnover) // (Other liability-C) | R1057 | (Third Party Debts-C) // (Current Liabilities) | R1378 | (Other Liabilities-NC) // (Other Operational Items - FSE and Others) |
| R20 | (Sector Turnover) // (Current Liabilities) | R1058 | (Third Party Debts-C) // (Total Liabilities) | R1379 | (Other Liabilities-NC) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) |
| R21 | (Sector Turnover) // (Total Liabilities) | R1059 | (Third Party Debts-C) // (Operational Income) | R1380 | (Other Liabilities-NC) // (Depreciation and Amortization) |
| R22 | (Sector Turnover) // (Operational Income) | R1060 | (Third Party Debts-C) // (Cost of Goods Sold ) | R1381 | (Other Liabilities-NC) // (Operational Result (EBIT) |
| R23 | (Sector Turnover) // (Cost of Goods Sold ) | R1061 | (Third Party Debts-C) // (Gross Margin) | R1382 | (Other Liabilities-NC) // (Financial Results) |
| R24 | (Sector Turnover) // (Gross Margin) | R1062 | (Third Party Debts-C) // (Personnel Expenditure) | R1383 | (Other Liabilities-NC) // (Current Result) |
| R26 | (Sector Turnover) // (Other Operational Items -FSE and Others) | R1063 | (Third Party Debts-C) // (Other Operational Items -FSE and Others) | R1384 | (Other Liabilities-NC) // (Income Tax) |
| R27 | (Sector Turnover) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1064 | (Third Party Debts-C) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1387 | (Other Liabilities-NC) // (Interest Expenses) |
| R28 | (Sector Turnover) // (Depreciation and Amortization) | R1067 | (Third Party Debts-C) // (Financial Results) | R1389 | (Other Liabilities-NC) // (Gross Value Added) |
| R29 | (Sector Turnover) // (Operational Result (EBIT) | R1068 | (Third Party Debts-C) // (Current Result) | R1391 | (Other Liabilities-NC) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R31 | (Sector Turnover) // (Current Result) | R1069 | (Third Party Debts-C) // (Income Tax) | R1392 | (Other Liabilities-NC) // (Working Capital) |
| R32 | (Sector Turnover) // (Income Tax) | R1070 | (Third Party Debts-C) // (Net Profit for the Year) | R1393 | (Other Liabilities-NC) // (Working Capital Requirements ) |
| R33 | (Sector Turnover) // (Net Profit for the Year) | R1072 | (Third Party Debts-C) // (Interest Expenses) | R1394 | (Other Liabilities-NC) // (Net Cashflow ) |
| R35 | (Sector Turnover) // (Interest Expenses) | R1074 | (Third Party Debts-C) // (Gross Value Added) | R1395 | (Other Liabilities-NC) // (Self-financing: NR + Adjustments) |
| R36 | (Sector Turnover) // (Cash Flows) | R1075 | (Third Party Debts-C) // (Operating expenses: (Sales - EBT - Adjustments) | R1400 | (Non-Current Liabilities) // (Debts to Third Parties-C) |
| R37 | (Sector Turnover) // (Gross Value Added) | R1076 | (Third Party Debts-C) // (Permanent Capitals: Equity + Non Current Liabilities) | R1401 | (Non-Current Liabilities) // (Other liability-C) |
| R39 | (Sector Turnover) // (Permanent Capitals: Equity + Non Current Liabilities) | R1078 | (Third Party Debts-C) // (Working Capital Requirements ) | R1403 | (Non-Current Liabilities) // (Total Liabilities) |
| R40 | (Sector Turnover) // (Working Capital) | R1080 | (Third Party Debts-C) // (Self-financing: NR + Adjustments) | R1407 | (Non-Current Liabilities) // (Personnel Expenditure) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R41 | (Sector Turnover) // (Working Capital Requirements ) | R1081 | (Third Party Debts-C) // (Invested Capital: NCA + WC + TA) | R1408 | (Non-Current Liabilities) // (Other Operational Items - FSE and Others) |
| R42 | (Sector Turnover) // (Net Cashflow ) | R1083 | (Third Party Debts-C) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) | R1413 | (Non-Current Liabilities) // (Current Result) |
| R43 | (Sector Turnover) // (Self-financing: NR + Adjustments) | R1084 | (Other Current Assets) // (Cash) | R1414 | (Non-Current Liabilities) // (Income Tax) |
| R44 | (Sector Turnover) // (Invested Capital: NCA + WC + TA) | R1085 | (Other Current Assets) // (Current Assets) | R1427 | (Non-Current Liabilities) // (Net Operating Profit After Tax) |
| R45 | (Sector Turnover) // (Net Operating Profit After Tax) | R1088 | (Other Current Assets) // (Retained Earnings or Other Equity) | R1429 | (Financial Debts-C) // (Debts to Third Parties-C) |
| R829 | (Intangible Assets) // (Tangible Fixed Assets) | R1089 | (Other Current Assets) // (Equity) | R1430 | (Financial Debts-C) // (Other liability-C) |
| R830 | (Intangible Assets) // (Other Non-Current Assets) | R1090 | (Other Current Assets) // (Debts to Third Parties-NC) | R1431 | (Financial Debts-C) // (Current Liabilities) |
| R831 | (Intangible Assets) // (Non Current Assets) | R1091 | (Other Current Assets) // (Other Liabilities-NC) | R1433 | (Financial Debts-C) // (Operational Income) |
| R832 | (Intangible Assets) // (Inventories) | R1092 | (Other Current Assets) // (Non-Current Liabilities) | R1434 | (Financial Debts-C) // (Cost of Goods Sold ) |
| R833 | (Intangible Assets) // (Third Party Debts-C) | R1093 | (Other Current Assets) // (Financial Debts-C) | R1435 | (Financial Debts-C) // (Gross Margin) |
| R834 | (Intangible Assets) // (Other Current Assets) | R1094 | (Other Current Assets) // (Debts to Third Parties-C) | R1436 | (Financial Debts-C) // (Personnel Expenditure) |
| R835 | (Intangible Assets) // (Cash) | R1095 | (Other Current Assets) // (Other liability-C) | R1439 | (Financial Debts-C) // (Depreciation and Amortization) |
| R837 | (Intangible Assets) // (Total Assets) | R1096 | (Other Current Assets) // (Current Liabilities) | R1440 | (Financial Debts-C) // (Operational Result (EBIT) |
| R838 | (Intangible Assets) // (Share Capital) | R1101 | (Other Current Assets) // (Personnel Expenditure) | R1441 | (Financial Debts-C) // (Financial Results) |
| R839 | (Intangible Assets) // (Retained Earnings or Other Equity) | R1102 | (Other Current Assets) // (Other Operational Items -FSE and Others) | R1442 | (Financial Debts-C) // (Current Result) |
| R841 | (Intangible Assets) // (Debts to Third Parties-NC) | R1103 | (Other Current Assets) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1443 | (Financial Debts-C) // (Income Tax) |
| R842 | (Intangible Assets) // (Other Liabilities-NC) | R1104 | (Other Current Assets) // (Depreciation and Amortization) | R1446 | (Financial Debts-C) // (Interest Expenses) |
| R843 | (Intangible Assets) // (Non-Current Liabilities) | R1105 | (Other Current Assets) // (Operational Result (EBIT) | R1451 | (Financial Debts-C) // (Working Capital) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R844 | (Intangible Assets) // (Financial Debts-C) | R1106 | (Other Current Assets) // (Financial Results) | R1452 | (Financial Debts-C) // (Working Capital Requirements ) |
| R845 | (Intangible Assets) // (Debts to Third Parties-C) | R1107 | (Other Current Assets) // (Current Result) | R1454 | (Financial Debts-C) // (Self-financing: NR + Adjustments) |
| R846 | (Intangible Assets) // (Other liability-C) | R1109 | (Other Current Assets) // (Net Profit for the Year) | R1458 | (Debts to Third Parties-C) // (Other liability-C) |
| R849 | (Intangible Assets) // (Operational Income) | R1113 | (Other Current Assets) // (Gross Value Added) | R1459 | (Debts to Third Parties-C) // (Current Liabilities) |
| R850 | (Intangible Assets) // (Cost of Goods Sold ) | R1114 | (Other Current Assets) // (Operating expenses: (Sales - EBT - Adjustments) | R1460 | (Debts to Third Parties-C) // (Total Liabilities) |
| R851 | (Intangible Assets) // (Gross Margin) | R1115 | (Other Current Assets) // (Permanent Capitals: Equity + Non Current Liabilities) | R1461 | (Debts to Third Parties-C) // (Operational Income) |
| R854 | (Intangible Assets) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1116 | (Other Current Assets) // (Working Capital) | R1463 | (Debts to Third Parties-C) // (Gross Margin) |
| R855 | (Intangible Assets) // (Depreciation and Amortization) | R1117 | (Other Current Assets) // (Working Capital Requirements ) | R1464 | (Debts to Third Parties-C) // (Personnel Expenditure) |
| R857 | (Intangible Assets) // (Financial Results) | R1119 | (Other Current Assets) // (Self-financing: NR + Adjustments) | R1465 | (Debts to Third Parties-C) // (Other Operational Items - FSE and Others) |
| R858 | (Intangible Assets) // (Current Result) | R1120 | (Other Current Assets) // (Invested Capital: NCA + WC + TA) | R1468 | (Debts to Third Parties-C) // (Operational Result (EBIT) |
| R859 | (Intangible Assets) // (Income Tax) | R1123 | (Cash) // (Current Assets) | R1469 | (Debts to Third Parties-C) // (Financial Results) |
| R862 | (Intangible Assets) // (Interest Expenses) | R1125 | (Cash) // (Share Capital) | R1477 | (Debts to Third Parties-C) // (Operating expenses: (Sales - EBT - Adjustments) |
| R864 | (Intangible Assets) // (Gross Value Added) | R1126 | (Cash) // (Retained Earnings or Other Equity) | R1478 | (Debts to Third Parties-C) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R866 | (Intangible Assets) // (Permanent Capitals: Equity + Non Current Liabilities) | R1127 | (Cash) // (Equity) | R1486 | (Other liability-C) // (Current Liabilities) |
| R867 | (Intangible Assets) // (Working Capital) | R1128 | (Cash) // (Debts to Third Parties-NC) | R1488 | (Other liability-C) // (Operational Income) |
| R869 | (Intangible Assets) // (Net Cashflow ) | R1130 | (Cash) // (Non-Current Liabilities) | R1489 | (Other liability-C) // (Cost of Goods Sold ) |
| R870 | (Intangible Assets) // (Self-financing: NR + Adjustments) | R1131 | (Cash) // (Financial Debts-C) | R1493 | (Other liability-C) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R874 | (Tangible Fixed Assets) // (Other Non-Current Assets) | R1132 | (Cash) // (Debts to Third Parties-C) | R1494 | (Other liability-C) // (Depreciation and Amortization) |
| R875 | (Tangible Fixed Assets) // (Non Current Assets) | R1133 | (Cash) // (Other liability-C) | R1495 | (Other liability-C) // (Operational Result (EBIT) |
| R876 | (Tangible Fixed Assets) // (Inventories) | R1134 | (Cash) // (Current Liabilities) | R1498 | (Other liability-C) // (Income Tax) |
| R877 | (Tangible Fixed Assets) // (Third Party Debts-C) | R1136 | (Cash) // (Operational Income) | R1502 | (Other liability-C) // (Cash Flows) |
| R878 | (Tangible Fixed Assets) // (Other Current Assets) | R1137 | (Cash) // (Cost of Goods Sold ) | R1505 | (Other liability-C) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R879 | (Tangible Fixed Assets) // (Cash) | R1138 | (Cash) // (Gross Margin) | R1507 | (Other liability-C) // (Working Capital Requirements ) |
| R880 | (Tangible Fixed Assets) // (Current Assets) | R1139 | (Cash) // (Personnel Expenditure) | R1508 | (Other liability-C) // (Net Cashflow ) |
| R881 | (Tangible Fixed Assets) // (Total Assets) | R1140 | (Cash) // (Other Operational Items -FSE and Others) | R1509 | (Other liability-C) // (Self-financing: NR + Adjustments) |
| R882 | (Tangible Fixed Assets) // (Share Capital) | R1142 | (Cash) // (Depreciation and Amortization) | R1510 | (Other liability-C) // (Invested Capital: NCA + WC + TA) |
| R883 | (Tangible Fixed Assets) // (Retained Earnings or Other Equity) | R1145 | (Cash) // (Current Result) | R1512 | (Other liability-C) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R884 | (Tangible Fixed Assets) // (Equity) | R1147 | (Cash) // (Net Profit for the Year) | R1514 | (Current Liabilities) // (Operational Income) |
| R885 | (Tangible Fixed Assets) // (Debts to Third Parties-NC) | R1149 | (Cash) // (Interest Expenses) | R1516 | (Current Liabilities) // (Gross Margin) |
| R886 | (Tangible Fixed Assets) // (Other Liabilities-NC) | R1151 | (Cash) // (Gross Value Added) | R1517 | (Current Liabilities) // (Personnel Expenditure) |
| R888 | (Tangible Fixed Assets) // (Financial Debts-C) | R1153 | (Cash) // (Permanent Capitals: Equity + Non Current Liabilities) | R1520 | (Current Liabilities) // (Depreciation and Amortization) |
| R889 | (Tangible Fixed Assets) // (Debts to Third Parties-C) | R1154 | (Cash) // (Working Capital) | R1543 | (Total Liabilities) // (Other Operational Items -FSE and Others) |
| R891 | (Tangible Fixed Assets) // (Current Liabilities) | R1155 | (Cash) // (Working Capital Requirements ) | R1565 | (Operational Income) // (Gross Margin) |
| R892 | (Tangible Fixed Assets) // (Total Liabilities) | R1156 | (Cash) // (Net Cashflow ) | R1566 | (Operational Income) // (Personnel Expenditure) |
| R893 | (Tangible Fixed Assets) // (Operational Income) | R1157 | (Cash) // (Self-financing: NR + Adjustments) | R1575 | (Operational Income) // (Turnover) |
| R894 | (Tangible Fixed Assets) // (Cost of Goods Sold ) | R1160 | (Cash) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) | R1578 | (Operational Income) // (Gross Value Added) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R895 | (Tangible Fixed Assets) // (Gross Margin) | R1162 | (Current Assets) // (Share Capital) | R1579 | (Operational Income) // (Operating expenses: (Sales - EBT - Adjustments) |
| R896 | (Tangible Fixed Assets) // (Personnel Expenditure) | R1163 | (Current Assets) // (Retained Earnings or Other Equity) | R1580 | (Operational Income) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R897 | (Tangible Fixed Assets) // (Other Operational Items -FSE and Others) | R1164 | (Current Assets) // (Equity) | R1585 | (Operational Income) // (Invested Capital: NCA + WC + TA) |
| R898 | (Tangible Fixed Assets) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1170 | (Current Assets) // (Other liability-C) | R1595 | (Cost of Goods Sold) // (Current Result) |
| R899 | (Tangible Fixed Assets) // (Depreciation and Amortization) | R1171 | (Current Assets) // (Current Liabilities) | R1598 | (Cost of Goods Sold) // (Turnover) |
| R900 | (Tangible Fixed Assets) // (Operational Result (EBIT) | R1172 | (Current Assets) // (Total Liabilities) | R1611 | (Gross Margin) // (Personnel Expenditure) |
| R901 | (Tangible Fixed Assets) // (Financial Results) | R1173 | (Current Assets) // (Operational Income) | R1613 | (Gross Margin) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) |
| R902 | (Tangible Fixed Assets) // (Current Result) | R1176 | (Current Assets) // (Personnel Expenditure) | R1620 | (Gross Margin) // (Turnover) |
| R903 | (Tangible Fixed Assets) // (Income Tax) | R1177 | (Current Assets) // (Other Operational Items -FSE and Others) | R1622 | (Gross Margin) // (Cash Flows) |
| R904 | (Tangible Fixed Assets) // (Net Profit for the Year) | R1188 | (Current Assets) // (Gross Value Added) | R1625 | (Gross Margin) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R908 | (Tangible Fixed Assets) // (Gross Value Added) | R1191 | (Current Assets) // (Working Capital) | R1627 | (Gross Margin) // (Working Capital Requirements ) |
| R910 | (Tangible Fixed Assets) // (Permanent Capitals: Equity + Non Current Liabilities) | R1194 | (Current Assets) // (Self-financing: NR + Adjustments) | R1630 | (Gross Margin) // (Invested Capital: NCA + WC + TA) |
| R911 | (Tangible Fixed Assets) // (Working Capital) | R1195 | (Current Assets) // (Invested Capital: NCA + WC + TA) | R1633 | (Personnel Expenditure) // (Other Operational Items - FSE and Others) |
| R912 | (Tangible Fixed Assets) // (Working Capital Requirements ) | R1208 | (Total Assets) // (Total Liabilities) | R1635 | (Personnel Expenditure) // (Depreciation and Amortization) |
| R913 | (Tangible Fixed Assets) // (Net Cashflow ) | R1211 | (Total Assets) // (Gross Margin) | R1641 | (Personnel Expenditure) // (Turnover) |
| R914 | (Tangible Fixed Assets) // (Self-financing: NR + Adjustments) | R1212 | (Total Assets) // (Personnel Expenditure) | R1644 | (Personnel Expenditure) // (Gross Value Added) |
| R915 | (Tangible Fixed Assets) // (Invested Capital: NCA + WC + TA) | R1231 | (Total Assets) // (Invested Capital: NCA + WC + TA) | R1645 | (Personnel Expenditure) // (Operating expenses: (Sales - EBT - Adjustments) |
| R918 | (Other Non-Current Assets) // (Non Current Assets) | R1234 | (Share Capital) // (Retained Earnings or Other Equity) | R1646 | (Personnel Expenditure) // (Permanent Capitals: Equity + Non Current Liabilities) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R919 | (Other Non-Current Assets) // (Inventories) | R1235 | (Share Capital) // (Equity) | R1651 | (Personnel Expenditure) // (Invested Capital: NCA + WC + TA) |
| R920 | (Other Non-Current Assets) // (Third Party Debts-C) | R1238 | (Share Capital) // (Non-Current Liabilities) | R1661 | (Other Operational Items - FSE and Others) // (Turnover) |
| R921 | (Other Non-Current Assets) // (Other Current Assets) | R1240 | (Share Capital) // (Debts to Third Parties-C) | R1665 | (Other Operational Items - FSE and Others) // (Operating expenses: (Sales - EBT - Adjustments) |
| R922 | (Other Non-Current Assets) // (Cash) | R1241 | (Share Capital) // (Other liability-C) | R1666 | (Other Operational Items - FSE and Others) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R923 | (Other Non-Current Assets) // (Current Assets) | R1242 | (Share Capital) // (Current Liabilities) | R1671 | (Other Operational Items - FSE and Others) // (Invested Capital: NCA + WC + TA) |
| R924 | (Other Non-Current Assets) // (Total Assets) | R1244 | (Share Capital) // (Operational Income) | R1674 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Depreciation and Amortization) |
| R925 | (Other Non-Current Assets) // (Share Capital) | R1245 | (Share Capital) // (Cost of Goods Sold ) | R1676 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Financial Results) |
| R926 | (Other Non-Current Assets) // (Retained Earnings or Other Equity) | R1246 | (Share Capital) // (Gross Margin) | R1680 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Turnover) |
| R927 | (Other Non-Current Assets) // (Equity) | R1247 | (Share Capital) // (Personnel Expenditure) | R1682 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Cash Flows) |
| R928 | (Other Non-Current Assets) // (Debts to Third Parties-NC) | R1248 | (Share Capital) // (Other Operational Items -FSE and Others) | R1684 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Operating expenses: (Sales - EBT - Adjustments) |
| R929 | (Other Non-Current Assets) // (Other Liabilities-NC) | R1249 | (Share Capital) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1685 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R930 | (Other Non-Current Assets) // (Non-Current Liabilities) | R1250 | (Share Capital) // (Depreciation and Amortization) | R1686 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Working Capital) |
| R931 | (Other Non-Current Assets) // (Financial Debts-C) | R1251 | (Share Capital) // (Operational Result (EBIT) | R1687 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Working Capital Requirements ) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R932 | (Other Non-Current Assets) // (Debts to Third Parties-C) | R1252 | (Share Capital) // (Financial Results) | R1689 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Self-financing: NR + Adjustments) |
| R933 | (Other Non-Current Assets) // (Other liability-C) | R1253 | (Share Capital) // (Current Result) | R1690 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Invested Capital: NCA + WC + TA) |
| R934 | (Other Non-Current Assets) // (Current Liabilities) | R1254 | (Share Capital) // (Income Tax) | R1692 | (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R936 | (Other Non-Current Assets) // (Operational Income) | R1255 | (Share Capital) // (Net Profit for the Year) | R1698 | (Depreciation and Amortization) // (Turnover) |
| R937 | (Other Non-Current Assets) // (Cost of Goods Sold ) | R1257 | (Share Capital) // (Interest Expenses) | R1701 | (Depreciation and Amortization) // (Gross Value Added) |
| R938 | (Other Non-Current Assets) // (Gross Margin) | R1258 | (Share Capital) // (Cash Flows) | R1703 | (Depreciation and Amortization) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R940 | (Other Non-Current Assets) // (Other Operational Items -FSE and Others) | R1259 | (Share Capital) // (Gross Value Added) | R1705 | (Depreciation and Amortization) // (Working Capital Requirements ) |
| R942 | (Other Non-Current Assets) // (Depreciation and Amortization) | R1260 | (Share Capital) // (Operating expenses: (Sales - EBT - Adjustments) | R1706 | (Depreciation and Amortization) // (Net Cashflow ) |
| R943 | (Other Non-Current Assets) // (Operational Result (EBIT) | R1261 | (Share Capital) // (Permanent Capitals: Equity + Non Current Liabilities) | R1708 | (Depreciation and Amortization) // (Invested Capital: NCA + WC + TA) |
| R944 | (Other Non-Current Assets) // (Financial Results) | R1262 | (Share Capital) // (Working Capital) | R1710 | (Depreciation and Amortization) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R945 | (Other Non-Current Assets) // (Current Result) | R1264 | (Share Capital) // (Net Cashflow ) | R1713 | (Operational Result (EBIT) // (Income Tax) |
| R946 | (Other Non-Current Assets) // (Income Tax) | R1265 | (Share Capital) // (Self-financing: NR + Adjustments) | R1717 | (Operational Result (EBIT) // (Cash Flows) |
| R947 | (Other Non-Current Assets) // (Net Profit for the Year) | R1266 | (Share Capital) // (Invested Capital: NCA + WC + TA) | R1724 | (Operational Result (EBIT) // (Self-financing: NR + Adjustments) |
| R950 | (Other Non-Current Assets) // (Cash Flows) | R1267 | (Share Capital) // (Net Operating Profit After Tax) | R1727 | (Operational Result (EBIT) // (Earnings Before Interest, |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| | | | | | Depreciation, Amortization and after Tax ) |
| R951 | (Other Non-Current Assets) // (Gross Value Added) | R1268 | (Share Capital) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) | R1729 | (Financial Results) // (Income Tax) |
| R953 | (Other Non-Current Assets) // (Permanent Capitals: Equity + Non Current Liabilities) | R1270 | (Retained Earnings or Other Equity) // (Debts to Third Parties-NC) | R1731 | (Financial Results) // (Turnover) |
| R954 | (Other Non-Current Assets) // (Working Capital) | R1271 | (Retained Earnings or Other Equity) // (Other Liabilities-NC) | R1732 | (Financial Results) // (Interest Expenses) |
| R955 | (Other Non-Current Assets) // (Working Capital Requirements ) | R1272 | (Retained Earnings or Other Equity) // (Non-Current Liabilities) | R1733 | (Financial Results) // (Cash Flows) |
| R956 | (Other Non-Current Assets) // (Net Cashflow ) | R1273 | (Retained Earnings or Other Equity) // (Financial Debts-C) | R1734 | (Financial Results) // (Gross Value Added) |
| R957 | (Other Non-Current Assets) // (Self-financing: NR + Adjustments) | R1274 | (Retained Earnings or Other Equity) // (Debts to Third Parties-C) | R1735 | (Financial Results) // (Operating expenses: (Sales - EBT - Adjustments) |
| R959 | (Other Non-Current Assets) // (Net Operating Profit After Tax) | R1275 | (Retained Earnings or Other Equity) // (Other liability-C) | R1737 | (Financial Results) // (Working Capital) |
| R961 | (Non Current Assets) // (Inventories) | R1278 | (Retained Earnings or Other Equity) // (Operational Income) | R1738 | (Financial Results) // (Working Capital Requirements ) |
| R962 | (Non Current Assets) // (Third Party Debts-C) | R1279 | (Retained Earnings or Other Equity) // (Cost of Goods Sold ) | R1739 | (Financial Results) // (Net Cashflow ) |
| R966 | (Non Current Assets) // (Total Assets) | R1280 | (Retained Earnings or Other Equity) // (Gross Margin) | R1740 | (Financial Results) // (Self-financing: NR + Adjustments) |
| R968 | (Non Current Assets) // (Retained Earnings or Other Equity) | R1281 | (Retained Earnings or Other Equity) // (Personnel Expenditure) | R1741 | (Financial Results) // (Invested Capital: NCA + WC + TA) |
| R969 | (Non Current Assets) // (Equity) | R1282 | (Retained Earnings or Other Equity) // (Other Operational Items -FSE and Others) | R1743 | (Financial Results) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R970 | (Non Current Assets) // (Debts to Third Parties-NC) | R1283 | (Retained Earnings or Other Equity) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1748 | (Current Result) // (Cash Flows) |
| R975 | (Non Current Assets) // (Other liability-C) | R1284 | (Retained Earnings or Other Equity) // (Depreciation and Amortization) | R1757 | (Current Result) // (Net Operating Profit After Tax) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R976 | (Non Current Assets) // (Current Liabilities) | R1285 | (Retained Earnings or Other Equity) // (Operational Result (EBIT) | R1760 | (Income Tax) // (Turnover) |
| R977 | (Non Current Assets) // (Total Liabilities) | R1286 | (Retained Earnings or Other Equity) // (Financial Results) | R1762 | (Income Tax) // (Cash Flows) |
| R978 | (Non Current Assets) // (Operational Income) | R1287 | (Retained Earnings or Other Equity) // (Current Result) | R1763 | (Income Tax) // (Gross Value Added) |
| R979 | (Non Current Assets) // (Cost of Goods Sold ) | R1288 | (Retained Earnings or Other Equity) // (Income Tax) | R1765 | (Income Tax) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R980 | (Non Current Assets) // (Gross Margin) | R1289 | (Retained Earnings or Other Equity) // (Net Profit for the Year) | R1766 | (Income Tax) // (Working Capital) |
| R981 | (Non Current Assets) // (Personnel Expenditure) | R1292 | (Retained Earnings or Other Equity) // (Cash Flows) | R1767 | (Income Tax) // (Working Capital Requirements ) |
| R982 | (Non Current Assets) // (Other Operational Items -FSE and Others) | R1293 | (Retained Earnings or Other Equity) // (Gross Value Added) | R1768 | (Income Tax) // (Net Cashflow ) |
| R983 | (Non Current Assets) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1296 | (Retained Earnings or Other Equity) // (Working Capital) | R1770 | (Income Tax) // (Invested Capital: NCA + WC + TA) |
| R984 | (Non Current Assets) // (Depreciation and Amortization) | R1297 | (Retained Earnings or Other Equity) // (Working Capital Requirements ) | R1771 | (Income Tax) // (Net Operating Profit After Tax) |
| R986 | (Non Current Assets) // (Financial Results) | R1298 | (Retained Earnings or Other Equity) // (Net Cashflow ) | R1772 | (Income Tax) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R988 | (Non Current Assets) // (Income Tax) | R1299 | (Retained Earnings or Other Equity) // (Self-financing: NR + Adjustments) | R1774 | (Net Profit for the Year) // (Interest Expenses) |
| R992 | (Non Current Assets) // (Cash Flows) | R1300 | (Retained Earnings or Other Equity) // (Invested Capital: NCA + WC + TA) | R1789 | (Turnover) // (Operating expenses: (Sales - EBT - Adjustments) |
| R993 | (Non Current Assets) // (Gross Value Added) | R1302 | (Retained Earnings or Other Equity) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) | R1798 | (Interest Expenses) // (Cash Flows) |
| R997 | (Non Current Assets) // (Working Capital Requirements ) | R1303 | (Equity) // (Debts to Third Parties-NC) | R1799 | (Interest Expenses) // (Gross Value Added) |
| R1003 | (Inventories) // (Third Party Debts-C) | R1305 | (Equity) // (Non-Current Liabilities) | R1800 | (Interest Expenses) // (Operating expenses: (Sales - EBT - Adjustments) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R1004 | (Inventories) // (Other Current Assets) | R1307 | (Equity) // (Debts to Third Parties-C) | R1801 | (Interest Expenses) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R1005 | (Inventories) // (Cash) | R1308 | (Equity) // (Other liability-C) | R1802 | (Interest Expenses) // (Working Capital) |
| R1006 | (Inventories) // (Current Assets) | R1311 | (Equity) // (Operational Income) | R1803 | (Interest Expenses) // (Working Capital Requirements ) |
| R1007 | (Inventories) // (Total Assets) | R1312 | (Equity) // (Cost of Goods Sold ) | R1808 | (Interest Expenses) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R1008 | (Inventories) // (Share Capital) | R1313 | (Equity) // (Gross Margin) | R1818 | (Cash Flows) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R1009 | (Inventories) // (Retained Earnings or Other Equity) | R1316 | (Equity) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1819 | (Gross Value Added) // (Operating expenses: (Sales - EBT - Adjustments) |
| R1010 | (Inventories) // (Equity) | R1317 | (Equity) // (Depreciation and Amortization) | R1820 | (Gross Value Added) // (Permanent Capitals: Equity + Non Current Liabilities) |
| R1013 | (Inventories) // (Non-Current Liabilities) | R1320 | (Equity) // (Current Result) | R1825 | (Gross Value Added) // (Invested Capital: NCA + WC + TA) |
| R1014 | (Inventories) // (Financial Debts-C) | R1321 | (Equity) // (Income Tax) | R1843 | (Working Capital) // (Working Capital Requirements ) |
| R1015 | (Inventories) // (Debts to Third Parties-C) | R1322 | (Equity) // (Net Profit for the Year) | R1845 | (Working Capital) // (Self-financing: NR + Adjustments) |
| R1016 | (Inventories) // (Other liability-C) | R1326 | (Equity) // (Gross Value Added) | R1846 | (Working Capital) // (Invested Capital: NCA + WC + TA) |
| R1017 | (Inventories) // (Current Liabilities) | R1327 | (Equity) // (Operating expenses: (Sales - EBT - Adjustments) | R1847 | (Working Capital) // (Net Operating Profit After Tax) |
| R1018 | (Inventories) // (Total Liabilities) | R1328 | (Equity) // (Permanent Capitals: Equity + Non Current Liabilities) | R1848 | (Working Capital) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R1019 | (Inventories) // (Operational Income) | R1329 | (Equity) // (Working Capital) | R1851 | (Working Capital Requirements ) // (Invested Capital: NCA + WC + TA) |
| R1021 | (Inventories) // (Gross Margin) | R1332 | (Equity) // (Self-financing: NR + Adjustments) | R1853 | (Working Capital Requirements ) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R1022 | (Inventories) // (Personnel Expenditure) | R1333 | (Equity) // (Invested Capital: NCA + WC + TA) | R1857 | (Net Cashflow ) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |

| Var ID | Ratio | Var ID | Ratio | Var ID | Ratio |
|---|---|---|---|---|---|
| R1023 | (Inventories) // (Other Operational Items -FSE and Others) | R1334 | (Equity) // (Net Operating Profit After Tax) | R1860 | (Self-financing: NR + Adjustments) // (Earnings Before Interest, Depreciation, Amortization and after Tax ) |
| R1024 | (Inventories) // (Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) | R1337 | (Debts to Third Parties-NC) // (Non-Current Liabilities) | R1865 | (Premium(ROE-TBY) // () |
| R1025 | (Inventories) // (Depreciation and Amortization) | R1339 | (Debts to Third Parties-NC) // (Debts to Third Parties-C) | R1868 | (Current Assets - Inventories - Current Liabilities) // (Operating expenses: (Sales - EBT - Adjustments) |
| R1027 | (Inventories) // (Financial Results) | R1340 | (Debts to Third Parties-NC) // (Other liability-C) | R1869 | (Current Assets - Total Liabilities) // (Total Assets) |
| R1028 | (Inventories) // (Current Result) | R1341 | (Debts to Third Parties-NC) // (Current Liabilities) | R1871 | (Current Assets - Inventories) // (Total Assets) |
| R1029 | (Inventories) // (Income Tax) | R1343 | (Debts to Third Parties-NC) // (Operational Income) | R1874 | (Net Profit - Current Assets + Cash) // (Total Assets) |
| R1030 | (Inventories) // (Net Profit for the Year) | R1344 | (Debts to Third Parties-NC) // (Cost of Goods Sold ) | R1880 | (Bank Loans) // (Current Assets) |
| R1033 | (Inventories) // (Cash Flows) | R1345 | (Debts to Third Parties-NC) // (Gross Margin) | R1881 | (Bank Loans) // (Total Liabilities) |
| R1034 | (Inventories) // (Gross Value Added) | R1346 | (Debts to Third Parties-NC) // (Personnel Expenditure) | R1882 | (Bank Loans) // (Cash Flow) |
| R1035 | (Inventories) // (Operating expenses: (Sales - EBT - Adjustments) | | | R1884 | (Bank Loans) // (Total Assets) |

1