UNIVERSITY OF THESSALY

BACHELOR THESIS

Bankruptcy prediction with machine learning

Πρόβλεψη πτώχευσης με χρήση μηχανικής μάθησης

Author:
Antonis
EVMORFOPOULOS

Supervisor:
Prof. Emmanouil
VAVALIS

A thesis submitted in fulfillment of the requirements for the degree of Diploma in Engineering

in the

Department of Electrical and Computer Engineering

July 2, 2020

Declaration of Authorship

I, Antonis EVMORFOPOULOS, declare that this thesis titled, "Bankruptcy prediction with machine learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

"I am not a visionary. I'm an engineer. I'm happy with the people who are wandering around looking at the stars but I am looking at the ground and I want to fix the pothole before I fall in."

Linus Torvalds

UNIVERSITY OF THESSALY

Abstract

Department of Electrical and Computer Engineering

Diploma in Engineering

Bankruptcy prediction with machine learning

by Antonis EVMORFOPOULOS

Η πρόβλεψη της εταιρικής πτώχευσης είναι ένα θέμα ενδιαφέροντος για τους επενδυτές, τους πιστωτές, τις τράπεζες καθώς και για τις κυβερνήσεις τις ίδιες. Στη διπλωματική αυτή, έχει χρησιμοποιηθεί πληθώρα προβλεπτικών μοντέλων βασισμένα σε αλγόριθμους μηχανικής μάθησης, σε ένα σύνολο δεδομένων που αποτελείται από 145 Ελληνικές εταιρίες όπου οι πτωχεύσεις παρατηρούνται μεταξύ του 2003 και του 2004, και αξιολογεί την αποτελεσματικότητά τους. Τα μοντέλα αυτά προβλέπουν την πιθανότητα πτώχευσης μιας εταιρείας λαμβάνοντας υπόψη τους ένα σύνολο λογιστικών δεδομένων και επεξηγηματικών μεταβλητών. Τα πειράματα που πραγματοποιήθηκαν δείχνουν ότι τα Δέντρα Αποφάσεων, τα Νευρωνικά Δίκτυα καθώς και οι βιβλιοθήκες H20 και AutoSklearn παρουσιάζουν θετικά δείγματα για αυτού του τύπου προβλήματος. Εν αντιθέσει, η Λογιστική Παλινδρόμηση και ο αλγόριθμος Naive Bayes, δεν είχαν την ίδια επιτυχία. Επιπλέον οι αλγόριθμοι XG-Boost, CatBoost και Gradient Boosting Machine, δεν ήταν ιδιαίτερα αποτελεσματικοί καθώς δεν ύπηρχε αρκετός όγκος δεδομένων ωστέ να δικαιολογεί τη χρήση τους. Τα πειραματικά αποτελέσματα είναι σύμφωνα με παλαιότερες έρευνες και δείχνουν ότι τα προτεινόμενα μοντέλα δίνουν αρκετά υποσχόμενα αποτελέσματα.

Predicting corporate bankruptcy is of interest to creditors, investors, borrowing organizations, and governments. This thesis applies various business prediction models based on several machine learning algorithms to a set of 145 financial distress Greek companies between the years 2003 and 2004 and estimates their effectiveness. These models estimate the bankruptcy probability considering a variety of accounting ratios and explanatory variables. The experiments contacted have shown that Decision trees, Neural Networks, and the libraries H2O and AutoSklearn can be effective in this type of problem. Naive Bayes and Logistic Regression on the other hand did not have the same success. Furthermore, the Boosting algorithms such as XG-Boost, CatBoost and Gradient Boosting Machine, did not handle the problem effectively since the dataset is not large enough to justify their use. The experimental results are in line with previously published findings and showed that the suggested models give promising outcomes.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Prof. Elias Houstis and Prof. Manolis Vavalis who gave me great help and guidance during the development of this thesis. I would like to thank Prof. Michael Vassilakopoulos and Prof. Georgios Petrakos as well for their support and faith in me and my ideas. Working with them gave me the opportunity to learn a lot about research, allowed me to gain confidence in my knowledge and capabilities as an engineer and helped me mature as a person.

I also wish to thank my father, Panos, my mother, Eva and my siblings, Dimitra and Konstantina for their unceasing encouragement throughout the entirety of my studies. This accomplishment would not have been possible without their great love and support.

Contents

Dedicated to my family and friends.

Chapter 1

Introduction

The economy of a country is shown by its development through trade, the transactions with other countries and the progress of industry and services within its own borders. As a result, companies in Greece have a great impact to the economy of the country. Thus, academic researchers have tried to capture statistically the levels of sustainability of Greek's firms which may be unsuccessful and have developed business models to predict corporate bankruptcy. Corporate failure is considered avoidable if the danger is detected soon enough for the appropriate measures to be taken.

Bankruptcy has been given throughout the years, many different terms such as failure, financial distress and insolvency. For researchers though, bankruptcy has to be judicial, meaning that the company has to declare bankruptcy with court order of the country it is operating in[1][2]. It has also been stated in published reports, that a company is considered bankrupt when any of the following events have occurred: "bankruptcy, bond default on an overdrawn account or non payment of a preferred stock dividend".[3]

For a long time, corporate bankruptcy prediction has been one of the utmost significance parts in evaluating the corporate prospects. Lenders, investors, governments and all kinds of stakeholders are eager to seek an efficient way to understand the financial abilities of the company so that they can choose the suitable decision making. Thus, researchers, have tried to develop machine learning and statistical models to predict financial risks. Some of the most important studies to note are the ones of Altman in 1968 [4] who used the multivariate discriminant analysis to predict financial failure, as well as, the original study in bankruptcy prediction which can be dated back to the early 20th century when Fitzpatrick (1932) used the economic index to describe predictive capacity of default business [5]. After that, more and more researchers focused on the bankruptcy prediction. The turning point in the survey of the business failure, was happened around 60 years ago when statistical models to make financial forecasts were initiated[3]. Logistic regression to forecast financial status was also one of the dominant models some years later[2]. In 2006 Aziz and Dar attempted at collecting and reporting the models that have been used throughout literature, consistently, and to highlight their strenths and/or flaws, trying to set a modern standpoint in predicting corporate failure[6]. More contemporary researchers have tried to use more modern methods as well, such as genetic algorithms and variations of neural networks among others [7][8][9][10].

The ability of predicting bankruptcy as early as possible is a very important matter, both from the aspect of the management but also from the point of view of the investors and creditors. That's the reason why so many models have been developed to perform this task as it is a way to measure the financial stability of the company itself and its financial health in the market.[11]

This thesis is structured as follows. In chapter 2, we define the various terms used and present a literature review . In chapter 3, we briefly describe the machine learning algorithms considered for implementing various bankruptcy models, list the various metrics considered for evaluating the bankruptcy models, and review various datasets used in training such models. Chapter 4 discusses the dataset used for the evaluation of the bankruptcy models considered in this thesis and derive the various covariates parameters of the bankruptcy models. Chapter 5 attempts to summarize the contribution of this thesis. Finally, we describe the main observations and outcomes of this research effort and discuss its possible future expansions.

Chapter 2

Related Work

2.1 Bankruptcy-Definitions

In the early years of predicting financial distress[3], the functionality of a company was compared with the one of a tank of water which depends on the flow of water coming in and out. A corporation that is dealing with financial problems and declares bankruptcy, can have a lot of common features with a tank of water that is completely drained. Using the rule of income and outflow as a benchmark, bankruptcy refers to a situation where the company is unable to meet its financial obligations or pay its dividends to the stockholders. This results to the liquidation of its assets or even the declaration of bankruptcy.

Corporation failure has been described in the literature as financial distress, insolvency, bankruptcy and default[3].The different terms given to this phenomenon can actually mean different stages of the same practice, thus resulting in a confusion on the holistic approach of this matter. For that reason, many researchers have tried to give a definition. Thus, bankruptcy is connected definitively with the financial problems a company might be facing. Being unable to pay its debts,shortage of liquidity and assets constitute financial problems that would result in a bankrupt company. Furthermore, Doumpos and Zompounidis in 1999, [12], reasoned that for a company to declare bankruptcy, its assets should be lower than its liabilities. The negative net income, gives away the company's debts and has a direct impact in the development of this phenomenon.

However, for the researchers what is considered the most dominant definition, is the prediction of bankruptcy which has been declared under judicial terms. More particularly, bankruptcy has to be declared according to the law of the country in which the company has been operating in. [3][4]

In Greece, in order for a company to declare bankruptcy, a request to the appropriate insolvency court has to be filed. If the company is a natural person, the request has to be filed by him/herself while if it is a legal person then the management of the company is required to do so, if that decision is part of the management quota, according to the article 96 of the greek insolvency law. For the bankruptcy claim to be legal, the claim has to include the financial statements of the company, the reasons that led to this decision and possible negotiations that have taken place with the buyer and the creditors.[13]

According to the paragraph 101 of the greek insolvency law, if the insolvency court considers that an agreement between the creditors and the person filing for bankruptcy is imminent, and also that the creditors stand no harm, deliberately caused by the latter, then the procedure of declaration of bankruptcy begins. If that happens, the judge is responsible to evaluate if the negotiations taking place are real or if the debtor is trying to commit fraud. Moreover, he is responsible to decide whether the settlement between the two sides has a chance of realization and if the interests of the creditors will remain intact. Furthermore, the insolvency law states that if the creditors do not agree with the declaration of bankruptcy that has been lawfully started by the insolvency court, then the debtor has no financial obligation towards them.[14]

In order for a settlement to be reached, information needs to be gathered for the company's actions and client list, its assets and liabilities. The creditors can decide on an appropriate settlement either each by themselves or with each other. However, for their decision to be considered valid, the creditors should represent 60% of the company's liabilities. The rest of the creditors, which are considered creditors of lower financial risk, could be part of this agreement if they are granted this privilege, or else the rest of the creditors sign a mortgage agreement for them. The number of the creditors that sign the agreement is not of importance. Even one creditor can sign the settlement by himself, if he represents the majority of the debt of the company.[13]

For a holistic approach on the matter though, apart from the definition of bankruptcy, we also have to visit the reasons that lead to such a problematic situation. In general, a company's financial situation is impacted both by endogenous and extrinsic factors.

Endogenous are considered the factors that are linked with the management of the company. More particularly, a company's policy is regarded by their financial, natural, human and technological means it uses to reach their goals. Inefficient management, disadvantageous accounting information, disanalogy between debt and assets are some of the reasons that can lead to financial distress.[14]

On the other hand, extrinsic factors are the ones that can not be controlled by the company. Such factors that can lead to financial failure are sudden changes in the industry or the ebbing of the financial and political situation of the country the company is operating in. Other examples, are law changes that have a direct impact on the policy of the company or even trend changes that make the product not desirable by the market.[14]

Most of the times, these factors combine or even success each other, leading the company to eventually declare bankruptcy. It is important to notice that financial distress is a dynamic procedure during which the company is presented with problems to meet its financial obligations, until and if this situation becomes permanent.[14]

2.2 Predicting Corporate Bankruptcy-Historical Review

In the past, many researchers have created many models and samples that can be used in predicting corporate failure. In order to do that, the researchers have categorized companies in to two samples, the healthy and the bankrupt ones. In the literature there are many ways that one can divide companies into such categories according to their models. The first studies were about financial ratios and the analysis was univariate. That means that the studies were focusing on each ratio on its own and compared it across the companies in the two categories.

One of the first models that were used in predicting financial distress was developed by Fitzpatrick in the 1930's. According to Fitzpatrick there are five stages of financial failure. These stages are incubation, financial embarrassment, financial insolvency, total insolvency and confirmed insolvency. His research was focused on finding the differences in the financial ratios of the healthy and the bankrupt firms. He used 13 ratios and tested 38 companies out of which half were bankrupt. The comparison between the ratios has two important points to note, the cutoff point and their trend in the years prior to bankruptcy. In his research, Fitzpatrick concluded that Net Income/Net Position and Net Position/Total Liabilities are the ratios that are most viable in performing this task.[5]

Another important model that was built and was used in predicting corporate bankruptcy, was published in the 1960's[3] and was built upon univariate analysis. In the sample, a number of ratios were tested, each on its own, and based on the comparison between the companies, they were then categorized as healthy or bankrupt. The cutoff point for each ratio was calculated unrelated to each other and was used to minimize the classification errors. According to the author, the ratios resulting from the assets and the liabilities can assist in predicting if the company will go bankrupt the next few years. In this research 30 ratios and a sample of 79 companies were used and tested univariate analysis. Out of these 30, 6 were chosen as the most appropriate in fulfilling this task. Furthermore, the ratio Cash Flow to Debt is emphasized, which gave indications of possible bankruptcy many years prior to bankruptcy.

However, predicting bankruptcy is a multidimensional problem, therefore it is not easy to say that a sole ratio can perform successfully the task. This becomes even more apparent, if someone considers the fact that accounting ratios can be meddled with from the companies or even how hard is to collect a substantial sample. As a result, many researches that used univariate analysis, didn't agree with each other, and multi-variate analysis unearthed the following years.

Edward I. Altman is a Professor of Finance, Emeritus, at New York University's Stern School of Business. He is best known for the development of the Altman Z-score for predicting bankruptcy which he published in 1968. Professor Altman is a leading academic on the High-Yield and Distressed Debt markets and is the pioneer in the building of models for credit risk management and bankruptcy prediction. His most notable work will be analysed in this chapter.[15]

In 1968, Altman published a study where he tests the effectiveness of multivariate analysis in predicting financial distress. In his sample there were 66 companies, half of which were bankrupt and the rest were healthy. The 33 bankrupt companies, declared so in between 1946 and 1965 and had an average of assets at $6.4 million with a range of $0.7 million up to $25.9 million. The healthy companies were chosen at random according to the industry of the bankrupt ones and the asset range was in between $1 million and $25 million.[4]

Altman chose 22 financial ratios that can assist in the task of predicting financial failure. Their choice was decided by their appearances in the literature and how effective they were considered to be in performing this task. In these 22 ratios he included some, which empirically was decided that they were important. These ratios were then divided into 5 categories such as ratios of liquidity, of leverage etc. Out of the 22 ratios, 5 made it to the final model which was then called the Z score. The formula is described below:

$$Z = 0.012 * X_1 + 0.014 * X_2 + 0.033 * X_3 + 0.006 * X_4 + 0.999 * X_5$$

where

- X1: working capital/total assets
- X2: retained earnings/total assets

- X3: earnings before interest and tax/total assets
- X4: market value of equity/total liabilities
- X5: sales/total assets

When the Z score was greater than 2.67 the company was considered healthy and there is no danger of financial distress for the year. If the Z score was lower than 1.81 then the company would be lead to bankruptcy within the year and if the score was in between, then the result was not clear as it was considered to be a grey area.[4]

In addition to Altman's Z-score model, there were two other models that were developed during the 1960s. After 1960 several other models have been developed. 28 studies were published in the 1970s, 53 studies were published in the 1980s and 70 studies were published in the 1990s. In the period 2000 – 2004 there were 11 studies that were published [16]. These studies are on different research area and therefore different number of ratios is included in the models.

There were many problems in Discriminant Analysis methods. In 1970's a report was published[17] that provided a summary of these problems. According to the author there were 7 problems related to the applications of DA like violation of assumptions, usage of linear function, no interpretation of variables separately, less dimensions, no group definition, unsuitable choice of prior probabilities and estimation of classification errors. Due to these errors researchers started to introduce many other statistical models like linear probability, factor analysis, logit and probability analysis.

Around the same time, Altman et al, developed another model for predicting financial distress, in which they have introduced quadratic analysis [1]. The development of the model was done both with the help of quadratic analysis but also with linear discriminant analysis in order to compare the two models and their effectiveness. This model was named the "ZETA" model.

In the beginnings of the 1980's, the concept of conditional probability model was proposed [2]. The data set that was used, derived from 10-K (Annual report of a firm that gives a comprehensive summary of firms' financial performance) financial statements. In this study, the author of the article, elaborated on the following four statistically important factors for bankruptcy prediction:

- The size of the company.
- Financial structure of the company.

- Performance of the company.
- Current liquidity of the company.

He also [2],criticised the MDA technique because of the three problems associated with it. (i) Matched samples. (ii) MDA behaves like a Discriminating device and does not provide any statistical importance of variables. (iii) MDA model gives output in the form of a score which is difficult to understand. Conditional logistic model keeps away all of the problems related to the MDA. The accuracy of this logistic prediction model was 96.12%, 95.55% for one year and two years respectively.

Since bankruptcy prediction has become a hot topic for the researchers and they have started to use different techniques to get better and more reliable results. Many researchers started to use different models to improve the results of the Altman's technique. Data mining techniques were not used until 1980. The use of data mining techniques like support vector machines, neural networks, decision trees was started in late 1980's for bankruptcy prediction.

In the more recent years, many researchers has tried to compare the results of Neural networks especially, with those of MDA and logit analysis. These studies concluded that neural network are more capable than conventional techniques in performing this task than the statistical methods. Moreover, neural networks have other advantages over the statistical methods, such as that they don't have to make certain assumptions, which has been noted as problematic in previous studies.[18][10]

However, in comparison with statistical methods, there are other disadvantages that need to be considered when developing a neural network model. For example, larger datasets are necessary for building an effective model. This causes concern since, in this field, collecting financial statements especially from bankrupt firms, is considered a tedious and difficult task. Furthermore, a point of controversy is the necessity of understanding the model. Neural networks, while effective, produce models that seem to be like black box, which is not desirable in every case.

Support vector machines is a novel method that belongs to the machine learning techniques. This particular method depends on a structured method of reducing risk rather than a theoretical one. The first time this method was used was by Shin et al in 2004 [19] who tried to predict bankruptcy in a dataset of North Korean companies. In their research they found that they had better results than the MDA, Logit analysis and Neural Networks.

It is important to note that, even though the vast majority of the literature of bankruptcy prediction has focused the research on the US and the UK, there are also models that have been developed for Greek firms. Some examples of that are Gloubos and Grammatikos [20] that focused their research on Greek firms and created a set of linear probability, probit, logit and multi-discriminate analysis models. The most accurate of the developed models were the probit and the linear probability models which both had a 70.8% accuracy. Also, Zopounidis and Doumpos [12] created a utilities additives discriminant model, which used twelve different ratios with accuracy varying from 47.37% to 84.21% for bankrupt companies. However, the models discussed in this section have been developed and used before the financial crisis of 2007.

In most researches that have tried to predict corporate failure, the samples of bankrupt and healthy companies are almost equal eg.Altman in 1968 [4]. However, a report in the 1980's [21] deducted that the analogy of bankrupt to healthy firms that exist in a sample, if it is too different than the one existing in the current state of the market, it can lead to false results and the model is not reliable enough.

As according to how the ratios in all of these researches are chosen, there are many opinions on what is the best possible method of choosing them according to literature. In one of the most prominent studies[3], the researcher chose the ratios used in his studies based on four criteria. Firstly, the ratios that were initially chosen, were chosen based on how much they have been used previously in the literature and how effective they were considered to be. Moreover, he decided that cashflow is of a great impact in the financial health of the company so he decided that he would include it in his research. Finally, he decided that the ratios included should be independent from the rest of the ratios.

Altman in 1968 split the choosing of ratios in two stages. In the first stage, the ratios were chosen by how many times they have been used in previous research but also empirically, based on his intuition, if the ratios will help him achieve his target. In the second stage, choosing the ratios that would make the final cut, was decided by four criteria, their statistical importance, their correlation, their impact on the accuracy of the model and again, his intuition. Thus, he decided on 5 ratios that would constitute the Z-score.[4]

It is clear, that the most seen reasons of deciding which ratios will be included in a study, is dependent on the intuition of the researcher and also, their popularity in the literature.

More particularly, in the mid 2000's a report studied the models used in bankruptcy prediction from 1930 to 2007[16]. A total of 752 different ratios have been used. However only 42 of them have been used in 5 or more models. Specifically, the ratio that has been used the most extensively, with a total of 54 appearances, is net earnings to total assets. It is also important to note that most studies on average include 8 to 10 ratios and the best results were achieved by studies that have used from 2 up to 21 independent variables in total.

Chapter 3

Algorithms and Data Availability

In this chapter, we present the machine learning algorithms used to implement various bankruptcy models, the metrics for their evaluation. Furthermore, we exhibit a list of sources that have been used in the literature to create datasets.

3.1 Logistic Regression

In classification problems we have a binary output Y and we want to model the conditional probability $Pr(Y = 1|X = x)$ as a function of x. The logistic transformation of p,which is $\frac{p}{1-p}$, gives a linear function of x without fear of nonsensical results as the probability p is required to be between 0 and 1.

Formally the model of logistic regression is that
$\log \frac{p(x)}{1-p(x)} = \beta_0 + x * \beta$
Which solved for p, gives
$p(x) = \frac{1}{1+_{exp}^{-(\beta_0+x*\beta)}}$

To minimize the misclasiffication rate, we should predict Y=1 whenever $p \geq 1$ and Y=0 otherwise. As a result logistic regression gives us a linear classifier. The decision boundary separating the two classes is the solution of $\beta_0 + x * \beta = 0$, which is a point if x is one dimensional, a line if two dimensional etc. Moreover logistic regression not only says where the boundary between the classes is but also that the class probabilities depend on the distance from the boundary. This trait allows us to gauge on the confidence of our prediction and tune it with class decision thresholds. Furthermore, logistic regression is able to explain the impact of each feature to the model through their coefficients, therefore allowing the researcher to further examine the performance of the model.

In supervised machine learning, models are trained in a subset of data. The goal is to compute the target of each training data. Overfitting happens when the model learns the noise of our data, therefore not having the ability to generalize on unseen data. Regularization adds a penalty as model's complexity increases so it won't overfit.

There are two widely used ways of regularization, the lasso regression ("L1") and the ridge regression ("L2").

Ridge regression adds squared magnitude of coefficient as a penalty term to the loss function.

$$\sum_{i=1}^{n}(y_i - \sum_{j=i}^{p} x_ij * \beta_j)^2 + \lambda * \sum_{j=1}^{p} \beta_j^2$$

If lambda is 0 then we have ordinary logistic regression(OLS). However if lambda is very large then it adds huge weight and it will lead to underfitting. Thus, it is important to choose our lambda wisely.

Lasso regression adds absolute value of magnitude of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}(y_i - \sum_{j=i}^{p} x_ij * \beta_j)^2 + \lambda * \sum_{j=1}^{p} |\beta_j|$$

If lambda is zero then we get again the OLS whereas large values of lambda will make coefficients zero, thus it would underfit.

The key difference between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.

Apart from the linear approximation described above, there is also the quadratic approximation. This method works when we are dealing with a curve and that we can not point to 0 by using the tangent line. Therefore we use a parabola. In order to fit a good parabola, both parabola and quadratic function should have same value, same first derivative, AND second derivative etc.

As it is logical, there are many methods to perform logistic regression and here we will make a note of those that were used in the research.

Newton's method is an iterative equation solver. It is an algorithm to find the roots of a polynomial function. Moreover, the geometric interpretation of Newton's method is that at each iteration one approximates f(x) by a quadratic function around x, and then takes a step towards the maximum/ minimum of that quadratic function.

The Liblinear method is a linear classification that supports logistic regression and linear support vector machines. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics i.e feature value. The solver uses a coordinate descent (CD) algorithm that solves optimization problems by successively performing approximate minimization along coordinate directions or coordinate hyperplanes.

The Stochastic Average Gradient method optimizes the sum of a finite number of smooth convex functions. Like stochastic gradient (SG) methods, the SAG method's iteration cost is independent of the number of terms in the sum. However, by incorporating a memory of previous gradient values the SAG method achieves a faster convergence rate than black-box SG methods.

3.2 Naive Bayes

Naive Bayes is a family of probabilistic algorithms that take advantage of the probabilty and the Bayes theorem to predict the tag of an instance. They are probabilistic, which means that they calculate the probability of each tag and then output the tag with the highest one. The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

The fundamental assumption that Naives Bayes makes is that each features makes an independent and an equal contribution to the outcome. The assumptions made by Naive Bayes are not generally correct in real-world situations. In-fact, the independence assumption is never correct but often works well in practice.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(h|d) = \frac{P(d|h)*P(h)}{P(d)}$$

where

- P(h | d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d | h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

- P(d) is the probability of the data (regardless of the hypothesis).

We are interested in calculating the posterior probability of P(h | d) from the prior probability p(h) with P(D) and P(d | h).

After calculating the posterior probability for a number of different hypotheses, the hypothesis with the highest probability is selected. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.

This can be written as:

$$MAP(h) = max(P(h|d))$$

or

$$MAP(h) = max(\frac{(P(d|h) * P(h))}{P(d)})$$

or

$$MAP(h) = max(P(d|h) * P(h))$$

The P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.

The class probabilities are simply the frequency of instances that belong to each class divided by the total number of instances. The conditional probabilities are the frequency of each attribute value for a given class value divided by the frequency of instances with that class value. Given a naive Bayes model, we can make predictions for new data using Bayes theorem.

$$MAP(h) = max(P(d|h) * P(h))$$

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of the data. The most commonly used one is the Gaussian Naive Bayes which assume that the data follow the Gaussian (normal) distribution. This means that in addition to the probabilities for each class, we must also store the mean and standard deviations for each input variable for each class.

The likelihood of the features is assumed to be Gaussian, hence, conditional probability is given by

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_{y2}}} *_{exp}^{-} \frac{(x_i - \mu_y)^2}{2\sigma_y^2}$$

When making predictions these parameters can be plugged into the Gaussian PDF with a new input for the variable, and in return the Gaussian PDF will provide an estimate of the probability of that new input value for that class.

3.3 Decision Trees

Decision trees are an algorithm that works through recursive partitioning the dataset that are as pure as possible to a given target class. Each node of the tree is associated to a particular set of records T that is splitted by a specific test on a feature. For example, a split on a continuous attribute A can be induced by the test $A \leq x$. The set of records T is then partitioned in two subsets that leads to the left branch of the tree and the right one.

$$T_l = t \epsilon T : t(A) \leq x$$

and

$$T_r = t \epsilon T : t(A) > x$$

The divide step of the recursive algorithm to induce decision tree takes into account all possible splits for each feature and tries to find the best one according to a chosen quality measure: the splitting criterion. If the dataset is induced on the following scheme

$$A_1, \ldots, A_m, C$$

where A_j are attributes and C is the target class, all candidates splits are generated and evaluated by the splitting criterion. Splits on continuous attributes and categorical ones are generated as described above. The selection of the best split is usually carried out by impurity measures. The impurity of the parent node has to be decreased by the split. Let $(E_1, E_2, \ldots, E_k)$ be a split induced on the set of records E, a splitting criterion that makes used of the impurity measure I is:

$$\delta = I(E) - \sum_{i=1}^{k} \frac{|E_i|}{E} * I(E_I)$$

At each node, we first select the best possible split for each variable, and then select the variable which minimizes the total (weighted) impurity of the daughter nodes. Note that all the variables are visited (including those which have been split before). Then all nodes are recursively split (including daughter nodes) until there is only one observation at each node. Then the tree has been saturated. For a saturated tree, the class assignment for future observations in each leaf node is just the class of the only training observation on this node. For other trees, the class assignment for future observations in each leaf node is the class with the majority presence at the node.

Standard impurity measures are the Shannon entropy or the Gini index. Below are the formulas of both:

$$Gini : G(E) = 1 - \sum_{j=1}^{c} p_j^2$$

$$Entropy : E(E) = 1 - \sum_{j=1}^{c} p_j * \log p_j$$

As per parsimony principal Gini outperform entropy as of computation ease (log is obvious has more computations involved rather that plain multiplication at processor/Machine level). However entropy definitely has an edge in some data cases involving high imbalance as entropy makes use of logistic probabilities and multiplies them with probabilities of event, resulting in the upscaling value of lower probabilities.

3.4 Artificial Neural Networks

Neural Networks are learning algorithms inspired from the neurons in human brain. The network comprises of interconnected neurons as a function of input data. Basically, NNs are a collection of software 'neurons' arranged in layers, connected together in a way that allows communication. Each neuron receives a set of x-values (numbered from 1 to n) as an input and compute the predicted y-hat value. Vector x contains the values of the features in one of m examples from the training set. Each of units has its own set of parameters, usually referred to as w (column vector of weights) and b (bias) which changes during the learning process. In each iteration, the neuron calculates a weighted average of the values of the vector x, based on its current weight vector w and adds bias. Finally, the result of this calculation is passed through a non-linear activation function g. All these result in the following formula:

$$z = w_1 * x_1 + w_2 * x_2 + w_n * x_n = w^T * x$$

For a single neuron m with activation a we have:

$$z_i = w_i^T * a_{i-1} + b_i$$

and

$$a_i = g_i * z_i$$

Each of the layers have to perform a number of very similar operations. In order to speed up the calculation vectorization is used. First of all, by stacking together horizontal vectors of weights w (transposed), matrix W is built. Similarly, bias is stacked together creating vertical vector b.

Activation functions are one of the key elements of the neural network. Without them, the neural network would become a combination of linear functions, so it would be just a linear function itself. The model would have limited expansiveness, no greater than logistic regression. The non-linearity element allows for greater flexibility and creation of complex functions during the learning process. The activation function also has a significant impact on the speed of learning, which is one of the main criteria for their selection. Currently, the most popular one for hidden layers is probably ReLU. Sometimes sigmoid is used as well, especially in the output layer, when we are dealing with a binary classification and we want the values returned from the model to be in the range from 0 to 1.

The learning process is about changing the values of the W and b parameters so that the loss function is minimized. In order to achieve this goal, the gradient descent method is used to find a function minimum. In each iteration the values of the loss function partial derivatives with respect to each of the parameters of our neural network is calculated. The neural network in order to compute such complicated gradients makes use of the backpropagation algorithm. The parameters of the neural network are adjusted according to the following formulas:

$$W_i = W_i + a * dW_i \; b_i + b_i + a * db_i$$

where a is the learning rate, a hyperparameter which allows you to control the value of performed adjustment. dW and db are calculated using the chain rule, partial derivatives of loss function with respect to W and b. The size of dW and db are the same as that of W and b respectively. The sequence of operations within the neural network is as follows:

$$dW_i = \frac{\theta L}{\theta W_i} = \frac{1}{m} * dZ_i * A_{i-1}^T$$

$$db_i = \frac{\theta L}{\theta b_i} = \frac{1}{m} * \sum_{i=1}^{n} dZ$$

$$dA_{i-1} = \frac{\theta L}{\theta A_{i-1}} = W_i^T * dZ_i$$

$$dZ_i = dA_i * g'(Z_i)$$

Basically, backpropagation allows the information to go back from the cost backward through the network in order to compute the gradient. Therefore, loop over the nodes starting at the final node in reverse topological order to compute the derivative of the final node output with respect to each edge's node tail. Doing so will help the network know which node is responsible for the most error and change the parameters in that direction and therefore, learning.

3.5 Gradient Boosting Classifier

Like Random Forest, Gradient Boosting is another technique for performing supervised machine learning tasks, like classification and regression. Similar to Random Forests, Gradient Boosting is an ensemble learner. This means it will create a final model based on a collection of individual models. The predictive power of these individual models is weak and prone to overfitting but combining many such weak models in an ensemble will lead to an overall much improved result. In Gradient Boosting machines, the most common type of weak model used is decision trees.

Boosting builds models from individual so called "weak learners" in an iterative way. In boosting, the individual models are not built on completely random subsets of data and features but sequentially by putting more weight on instances with wrong predictions and high errors. The general idea behind this is that instances, which are hard to predict correctly ("difficult" cases) will be focused on during learning, so that the model learns from past mistakes.

The gradient is used to minimize a loss function, similar to how Neural Nets utilize gradient descent to optimize ("learn") weights. In each round of training, the weak learner is built and its predictions are compared to the correct outcome that we expect. The distance between prediction ('residuals') and truth represents the error rate of our model. These errors can now be used to calculate the gradient. The residuals are treated as a negative gradient, giving us the direction that we want the model to learn.

3.6 XGBoost

XGBoost is one of the most popular and efficient implementations of the Gradient Boosted Trees algorithm, a supervised learning method that is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques.

The objective function (loss function and regularization) at iteration t that we need to minimize is the following:

$$L^{(}t) = \sum_{i=1}^{n}(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \omega(f_t)$$

It is easy to see that the XGBoost objective is a function of functions.

Both XGBoost and GBM follows the principle of gradient boosting. There are however, the difference in modeling details. Specifically, xgboost used a more regularized model formalization to control over-fitting, which gives it better performance.

Standard GBM implementation has no regularization like XGBoost, therefore it also helps to reduce overfitting. In fact, XGBoost is also known as a 'regularized boosting' technique.

Moreover GBM would stop splitting a node when it encounters a negative loss in the split. Thus, it is more of a greedy algorithm. XGBoost on the other hand make splits up to the depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.

3.7 CatBoost

CatBoost is another gradient boosting algorithm which introduces some innovations to combat the problems that the rest of the "family" faces. CatBoost essentially uses symmetric or oblivious trees. It uses the same features to split learning instances into the left and the right partitions for each level of the tree, in order to address the issue of target leakage. In simple words, the rest of the boosting algorithms use the residuals obtained in the training set to train the model, creating an optimistic bias of the learning quality.

To fight this prediction shift CatBoost uses a more effective strategy. It relies on the ordering principle and is inspired by online learning algorithms which get training examples sequentially in time. In this setting, the values of the target statistic for each example rely only on the observed history. To adapt this idea, Catboost introduces an artificial "time"— a random permutation $\sigma 1$ of the training examples. Then, for each example, it uses all the available "history" to compute its Target Statistic. Note that, using only one random permutation, results in preceding examples with higher variance in Target Statistic than subsequent ones. To this end, CatBoost uses different permutations for different steps of gradient boosting.

Moreover, CatBoost has two modes for choosing the tree structure, Ordered and Plain. Plain mode corresponds to a combination of the standard GBC algorithm with an ordered Target Statistic. In Ordered mode boosting we perform a random permutation of the training examples $\sigma 2$, and maintain n different supporting models $-M_1..., M_n$ such that the model M_i is trained using only the first i samples in the permutation. At each step, in order to obtain the residual for j-th sample, we use the model M_{j-1}.

Unfortunately, this algorithm is not feasible in most practical tasks due to the need of maintaining n different models, which increase the complexity and memory requirements by n times. Catboost implements a modification of this algorithm, on the basis of the gradient boosting algorithm, using one tree structure shared by all the models to be built.In order to avoid prediction shift, Catboost uses permutations such that $\sigma 1 = \sigma 2$. This guarantees that the target-y_i is not used for training M_i neither for the Target Statistic calculation nor for the gradient estimation.

3.8 H2O AutoML

Automated machine learning can be thought of as the standard machine learning process with the automation of some of the steps involved. AutoML very broadly includes:

- Automating certain parts of data preparation, e.g. imputation, standardization, feature selection, etc.
- Being able to generate various models automatically, e.g. random grid search, Bayesian Hyperparameter Optimization, etc.
- Getting the best model out of all the generated models, which most of the time is an Ensemble, e.g. ensemble selection, stacking, etc.

All in all, automated machine learning (AutoML) is the process of automating the end-to-end process of applying machine learning to real-world problems. AutoML tends to automate the maximum number of steps in an ML pipeline — with a minimum amount of human effort — without compromising the model's performance.

H2O has an industry-leading AutoML functionality (available in H2O 3.14) that automates the process of building a large number of models, to find the "best" model without any prior knowledge or effort by the Data Scientist. H2O AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit.

The current version of AutoML trains and cross-validates a default Random Forest, an Extremely-Randomized Forest, a random grid of Gradient Boosting Machines (GBMs), a random grid of Deep Neural Nets, a fixed grid of GLMs, and then trains two Stacked Ensemble models at the end. One ensemble contains all the models (optimized for model performance), and the second ensemble contains just the best performing model from each algorithm class/family (optimized for production use).[22]

3.9 AutoSklearn

Auto-sklearn provides out-of-the-box supervised machine learning. Built around the scikit-learn machine learning library, auto-sklearn automatically searches for the right learning algorithm for a new machine learning dataset and optimizes its hyperparameters. Thus, it frees the machine learning practitioner from these tedious tasks and allows her to focus on the real problem.

Auto-sklearn extends the idea of configuring a general machine learning framework with efficient global optimization which was introduced with Auto-WEKA. To improve generalization, auto-sklearn builds an ensemble of all models tested during the global optimization process. In order to speed up the optimization process, auto-sklearn uses meta-learning to identify similar datasets and use knowledge gathered in the past. Auto-sklearn wraps a total of 15 classification algorithms, 14 feature preprocessing algorithms, yielding a total of 110 hyperparameters, and takes care about data scaling, encoding of categorical parameters and missing values.

Optimizing performance in Auto-sklearn's space of 110 hyperparameters can of course be slow, and to jumpstart this process it uses meta-learning to start from good hyperparameter settings for previous similar datasets. Specifically, Auto-sklearn comes with a database of previous optimization runs on 140 diverse datasets from OpenML. For a new dataset, it first identifies the most similar datasets and starts from the saved best settings for those.

A second improvement was to automatically construct ensembles: instead of returning a single hyperparameter setting (as standard Bayesian optimization would), it automatically constructs ensembles from the models trained during the Bayesian optimization. Specifically, Ensemble Selection was used to create small, powerful ensembles with increased predictive power and robustness.

This system was developed for the ChaLearn AutoML challenge and won six out of ten phases of the first ChaLearn AutoML challenge, and the comprehensive analysis on over 100 diverse datasets shows that it substantially outperforms the previous state of the art models in AutoML.[23]

3.10 Metrics of evaluation of Machine Learning Models

Choice of metrics influences how the performance of machine learning algorithms is measured and compared. They influence how the importance of different characteristics is weighted in the results and ultimately impact the choice of which algorithm to use as a final model. In this section, some of the most used metrics will be explained in order to justify the metric of our choosing to measure the performance of our models.

3.10.1 Classification Accuracy

Classification accuracy is the number of correct predictions made as a ratio of all predictions made.

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP=True Positives, TN=True Negatives, FP=False Positives, FN=False Negatives.

This is the most common evaluation metric for classification problems, but accuracy alone doesn’t tell the full story when there is a class-imbalanced data set. It is really only suitable when there are an equal number of observations in each class (which is rarely the case) and that all predictions and prediction errors are equally important.

3.10.2 Precision and Recall

Precision attempts to identify the proportion of positive predictions that were correct.

$$Precision = \frac{TP}{TP+FP}$$

Recall attempts to identify the proportion of actual predictions that were predicted correctly.

$$Recall = \frac{TP}{TP+FN}$$

To fully evaluate the effectiveness of a model, both precision and recall must be examined. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa. It is up to the researcher to decide which measure matters the most in every case, in order to make the best possible choise for a model.

3.10.3 Area Under the ROC Curve

A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

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

False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP+TN}$$

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information, called AUC. AUC stands for "Area under the ROC Curve". That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).

The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

The ROC curve is a useful tool for a few reasons:

- The curves of different models can be compared directly in general or for different thresholds.
- The area under the curve (AUC) can be used as a summary of the model skill.

The shape of the curve contains a lot of information, including what we might care about most for a problem, the expected false positive rate, and the false negative rate.

Our task is to predict corporate bankruptcy. Therefore it is inherent that our dataset is imbalanced, as it will be described in the next section. Since we are more interested in the general skill of our classifier rather than the highest possible accuracy (which could be achieved even randomly), AUC will be used to evaluate the performance of the algorithms implemented.

3.11 Data Availability

Our dataset has been provided from the ICAP database. It includes 49 bankrupt and 96 healthy firms for a total of 145. In the sample, the years of bankruptcy are 2003 and 2004 and we have a total of 3 years worth of financial data prior to bankruptcy. This research is not the only one that uses this dataset as it has been used in previous studies as well[24][25][26]. In this chapter, we will try to document other popular sources that have been used in the literature.

3.11.1 Finanacial Databases

Similarly to our source, ICAP, there are other financial databases that have been used extensively by researchers to find financial statements and data in order to do their study in the field of bankruptcy prediction. Such examples are the Standard and Poor Capital IQ Service [4][27], the Wind Financial Terminal Database & CCER Economic and Financial database[28] and the Korea Credit Guarantee Fund[19]

3.11.2 The 10-K and the Annual Report

A 10-K is a comprehensive report filed annually by a publicly-traded company about its financial performance and is required by the U.S. Securities and Exchange Commission (SEC). Some of the information a company is required to document in the 10-K includes its history, organizational structure, financial statements, earnings per share, subsidiaries, executive compensation, and any other relevant data. The SEC requires this report to keep investors aware of a company's financial condition and to allow them to have enough information before they buy or sell shares in the corporation, or before investing in the firm's corporate bonds.[29]

Because of the depth and nature of the information they contain, 10-Ks are fairly long and tend to be complicated. But investors need to understand that this is one of the most comprehensive and most important documents a public company can publish on a yearly basis. The more information they can gather from the 10-K, the more they can understand about the company.[29]

The government requires companies to publish 10-K forms so investors have fundamental information about companies so they can make informed investment decisions. This form gives a clearer picture of everything a company does and what kinds of risks it faces.[29]

The 10-K includes five distinct sections[29]:

- Business: This provides an overview of the company's main operations, including its products and services (i.e., how it makes money).
- Risk factors: These outline any and all risks the company faces or may face in the future. The risks are typically listed in order of importance.
- Selected financial data: This section details specific financial information about the company over the last five years. This section presents more of a near-term view of the company's recent performance.
- Management's discussion and analysis of financial condition and results of operations: Also known as MDA, this gives the company an opportunity to explain its business results from the previous fiscal year. This section is where the company can tell its story in its own words.

- Financial statements and supplementary data: This includes the company's audited financial statements including the income statement, balance sheets, and statement of cash flows. A letter from the company's independent auditor certifying the scope of their review is also included in this section.

An annual report is a publication that public corporations must provide annually to shareholders to describe their operations and financial conditions. The front part of the report often contains an impressive combination of graphics, photos, and an accompanying narrative, all of which chronicle the company's activities over the past year. The back part of the report contains detailed financial and operational information.[30]

It was not until legislation was enacted after the stock market crash of 1929 that the annual report became a regular component of corporate financial reporting. The intent of the required annual report is to provide public disclosure of a company's corporate activities over the past year. The report is typically issued to shareholders and other stakeholders who use it to evaluate the firm's financial performance. Typically, an annual report will contain the following sections[30]:

- General corporate information
- Operating and financial highlights
- Letter to the shareholders from the CEO
- Narrative text, graphics, and photos
- Management's discussion and analysis (MDA)
- Financial statements, including the balance sheet, income statement, and cash flow statement
- Notes to the financial statements
- Auditor's report
- Summary of financial data
- Accounting policies

Similar in many ways, these documents are designed to help inform potential investors or current shareholders about the company's performance.The annual report is sent to shareholders each year ahead of the annual shareholder meeting and voting for the board of directors. The deadline for filing a 10-K is between 60 and 90 days after the end of the company's fiscal year, depending on the size of the company. Generally, 10-Ks are found on the SEC website, while the annual report should be available on the company's website—usually under the investor relations section. Where an annual report may include company information, financials, and a letter from the CEO, the 10-K will include various risks and a detailed discussion of operations. Both documents are important when analyzing a company, although the 10-K should be preferred by analysts, given its more comprehensive nature.[31]

3.11.3 Polish Dataset

In the UCI machine learning repository there is a dataset about bankruptcy prediction of Polish companies. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. The data was collected from Emerging Markets Information Service (EMIS), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. The data contains 10503 instances, 64 financial ratios and 5 years of financial statements prior to bankruptcy. It has been used extensively in the literature as well as in a Kaggle competition.[9][32][10]

3.11.4 Stock Exchange

A stock exchange is a marketplace where securities, such as stocks and bonds, are bought and sold. Bonds are typically traded Over-the-Counter (OTC), but some corporate bonds can be traded on stock exchanges. Stock exchanges allow companies to raise capital and investors to make informed decisions using real-time price information. Exchanges can be a physical location or an electronic trading platform. Though people are typically familiar with the image of the trading floor, many exchanges now use electronic trading.

Companies that are publicly listed on a stock exchange must conform to reporting standards that are set by regulating bodies. This includes having to regularly and publicly report their financial statements and earnings to their shareholders. The actions of a company's management are constantly under public scrutiny and directly affect the value of the company. Public reporting helps ensure that management will make decisions that benefit the goals of the company and its shareholders, thereby acting efficiently.

As a result, there are various researches that have listed their source of data as the individual stock exchange such as the Athens Stock Exchange[33][34], the Taiwan Stock Exchange[35] and the Istanbul Stock Exchange[36].

Chapter 4

Sample Data and Implementation

In this chapter we describe the sample dataset and its covariates parameters. Firstly we analyse the ratios used to build our models in order to give a more comprehensive view of the sample. We then present our results and compare it to those of a previous conducted research, which performs as a benchmark.

4.1 Sample Data

The source of the data we use is the National Bank of Greece and the business database containing the financial data of the companies named ICAP. More specifically, the bankruptcy deposits used in the research are related to the years of 2003 and 2004. The financial statements were collected from the database provided from ICAP and are related to a period of up to three years prior to bankruptcy. Accordingly we divided our data in three subsets.

- "2 years before": The dataset includes data from financial statements worth of one year prior to bankruptcy.
- "1 years before": The dataset includes data from financial statements worth of two years prior to bankruptcy.
- "last year": It includes data from financial statements worth of three years prior to bankruptcy.

In order to build our sample we managed to collect financial information of 49 bankrupt firms and 96 firms that were still operating in the years aforementioned. Thus, our sample consists of 145 firms and 13 individual financial ratios for each of the three years prior to declaring bankruptcy.

The ratios included in the research are the following:

- Scaled Change of Assets
- Size
- Change of Net Income

- Gross Profit Margin
- Capital Employed Turnover
- Stockholders Equity Turnover
- Capital Employed/Net Fixed Assets
- Debt/Equity
- Equity/Capital Employed
- Working Capital/Total Assets
- Average Collection period for Receivables
- Average Payment Period
- Average Turnover Period for Inventories

So as to be able to understand the reasons behind corporate bankruptcy, we will take a look in each of these ratios, to explain their meaning and what they represent.

4.1.1 Scaled Change of Assets

Asset prizes are directly observable and are readily available from the various markets where trading is occured. However, instead of the prices themselves we are more interested in various derived data and statistical summaries of those. The most common types of derived data are a first order of measure in the change of the asset prices in time. The scaled change of the asset prices is called the rate of return, which in the simplest form is simply the difference in price between two time points divided by the price in the first. More often, though, it is the difference in the logarithm of the price at the two time points. Rates of return are scaled in such a way in order to represent annual rates.[37]

4.1.2 Size

The size of the company is determined via the logarithm of the division of the total assets with the GDP Price Index. The logarithm of the total assets is considered as a proxy measure of the size of the company. The Growth Domestic Product price index is a measure of inflation in the prices of goods and services produced in the country. The GDP price index includes the prices of goods and services exported to other countries. A primary benefit of measuring this index is that it can show the growth of the economy over time, or its lack thereof. Therefore, this ratio is related both with the growth of the company itself but also with the economy of the country within the company operates.[38]

4.1.3 Change of Net Income

This ratio symbolizes the annual percentage change in a company's net income. The calculation is a given year's net income minus the prior year's net income, divided by the prior year's net income. If this figure is positive, the company's net income is growing; if it's negative, net income is generally declining. However, it's best to look at the Net Income percent Change for each of the past several years. Sometimes a one-time charge can depress earnings for a single year, even though the company's earnings are generally rising; in other cases, earnings may have only started slipping in the most recent year. Annual Net Income percent Change figures give a more long-term perspective than quarterly figures.[39]

Net Income is calculated by subtracting the total expenses from the total revenue. Changes in net income are endlessly scrutinized. In general, when a company's net income is low or negative, a myriad of problems could be to blame, ranging from decreasing sales to poor customer experience to inadequate expense management. It's important to know that net income is not a measure of how much cash a company earned during a given period. This is because the income statement includes a lot of non-cash expenses such as depreciation and amortization that aren't the same as cash expenses.[39]

4.1.4 Gross Profit Margin

Gross profit margin is a metric analysts use to assess a company's financial health by calculating the amount of money left over from product sales after subtracting the cost of goods sold (COGS). Sometimes referred to as the gross margin ratio, gross profit margin is frequently expressed as a percentage of sales. The gross profit margin shows the amount of profit made before deducting selling, general, and administrative costs, which is the firm's net profit margin.[40]

If a company's gross profit margin wildly fluctuates, this may signal poor management practices and/or inferior products. On the other hand, such fluctuations may be justified in cases where a company makes fundamental operational changes to its business model, in which case temporary volatility should be no cause for alarm. In other words, if a company decides to automate certain supply chain functions, the initial investment may be high, but the cost of goods ultimately decreases due to the lower labor costs resulting from the introduction of the automation.[40]

Furthermore, product pricing adjustments may also influence gross margins. If a company sells its products at a premium, with all other things equal, it has a higher gross margin. But this can be a delicate balancing act, because if a company sets its prices overly high, fewer customers may buy the product, and the company may consequently hemorrhage market share.[40]

4.1.5 Capital Employed Turnover

Turnover ratios measure how efficiently the facilities including the assets and liabilities of the organization are utilized. Turnover ratios formula include inventory turnover ratio, receivables turnover ratio, capital employed turnover ratio, working capital turnover ratio, asset turnover ratio and accounts payable turnover ratio.[41]

The Capital Employed Turnover is calculated by dividing the sales by the average capital employed and it indicates the efficiency with which a company utilizes its capital employed with reference to sales. This ratio helps the investors or the creditors to determine the ability of a firm to generate revenues from the capital employed and act as a key decision factor for lending more money to the asking firm.[42]

4.1.6 Stockholders Equity Turnover

Stockholders Equity Turnover is provided by a company's annual sales divided by its average stockholders' equity. Equity turnover is used to calculate the rate of return on common equity, and is a measure of how well a company uses its stockholders' equity to generate revenue. The higher the ratio is, the more efficiently a company is using its capital. It is also commonly referred to as Capital Turnover.[43]

It is important to note that the ratio varies substantially, depending on how capital-intensive an industry may be. In addition, company management can skew the ratio in their favor by using more debt instead of equity. Doing so will increase the ratio, but can place a business in serious jeopardy if margins fall, since the organization will no longer be able to pay its debts. The ratio assumes that the key company improvement standard is revenues, when in fact it is usually more important to generate cash flow or profits. Thus, the ratio may place an emphasis on the wrong target.[43]

4.1.7 Capital Employed/Net Fixed Assets

Capital-employed provides a snapshot of how a company invests its money. However, it can be problematic to define capital-employed because there are so many contexts in which it can exist. However, most definitions generally refer to the capital investment necessary for a business to function.[42]

Capital investments include stocks and long-term liabilities, but it can also refer to the value of assets used in the operation of a business. Put simply, it is a measure of the value of assets minus current liabilities. Both of these measures can be found on the balance sheet. A current liability is the portion of debt that must be paid back within one year. In this way, capital employed is a more accurate estimate of total assets.[42]

Net Fixed Assets calculates the residual value of the assets. In other words, it theoretically calculates how much life or use these assets have left in them by comparing the total purchase price with the total amount of depreciation that has been taken since the assets were purchased.[44]

A low ratio can often mean that the assets are outdated because the company has not replaced them in a long time. In other words, the assets have high amounts of accumulated depreciation indicating their age.[44]

4.1.8 Debt/Equity

The debt-to-equity (D/E) ratio is calculated by dividing a company's total liabilities by its shareholder equity. The ratio is used to evaluate a company's financial leverage. The D/E ratio is an important metric used in corporate finance. It is a measure of the degree to which a company is financing its operations through debt versus wholly-owned funds. More specifically, it reflects the ability of shareholder equity to cover all outstanding debts in the event of a business downturn.[45]

However the ratio can be distorted by retained earnings/losses, intangible assets, and pension plan adjustments, further research is usually needed to understand a company's true leverage. To reflect more on that, the D/E ratio is difficult to compare across industry groups where ideal amounts of debt will vary, thus investors will often modify the D/E ratio to focus on long-term debt rather than for short-term debt and payables.[45]

4.1.9 Equity/Capital Employed

Equity is typically referred to as shareholder equity which represents the amount of money that would be returned to a company's shareholders if all of the assets were liquidated and all of the company's debt was paid off.[43]

Equity is found on a company's balance sheet and is one of the most common financial metrics employed by analysts to assess the financial health of a company. Shareholder equity can also represent the book value of a company. Equity can sometimes be offered as payment-in-kind.[43]

Paired with capital employed, this ratio attempts to give a better understanding of a company's financial health, as it gives a better perspective of the liabilities a company on average has.

4.1.10 Working Capital/Total Assets

The working capital over total assets ratio, sometimes referred to as the net working capital ratio, measures the net liquid assets of a business as a percentage of it's total assets. The ratio is an indicator of the short term liquidity and financial strength of the business and indicates it's ability to finance short term obligations.[46]

The working capital over total assets ratio formula calculates the ratio by dividing the current assets less the current liabilities by the total assets of the business. It should be noted that the term working capital is broader than the usual working capital definition and, for the purposes of this ratio, refers to the difference between current assets (including cash) and current liabilities.[46]

It should be noted that there is no correct value for the working capital over total assets ratio, generally a high ratio is a good thing. Its level will vary from industry to industry, and therefore it is important when making comparisons, to determine an industry ratio benchmark based on financial statements of similar businesses.[46]

The working capital total assets ratio can be negative, indicating that current assets are greater than current liabilities. In general a negative ratio is viewed as a sign that the business is in financial distress and does not have the necessary liquid assets to pay its current liabilities as they fall due. A continually declining working capital to assets ratio indicates that the current assets of the business are declining, a sign of a loss making business facing financial difficulties.[46]

4.1.11 Average Collection period for Receivables

The average collection period is the amount of time it takes for a business to receive payments owed by its clients in terms of accounts receivable. Companies calculate the average collection period to make sure they have enough cash on hand to meet their financial obligations. The average collection period is calculated by dividing the average balance of accounts receivable by total net credit sales for the period and multiplying the quotient by the number of days in the period.[47]

A company's average collection period is indicative of the effectiveness of its accounts receivable management practices. Businesses must be able to manage their average collection period in order to ensure they operate smoothly. A lower average collection period is generally more favorable than a higher average collection period. A low average collection period indicates the organization collects payments faster. There is a downside to this, though, as it may indicate its credit terms are too strict. Customers may seek suppliers or service providers with more lenient payment terms. The average balance of accounts receivable is calculated by adding the opening balance in accounts receivable (AR) and ending balance in accounts receivable and dividing that total by two. When calculating the average collection period for an entire year, 365 may be used as the number of days in one year for simplicity.[47]

4.1.12 Average Payment Period

Average payment period is a solvency ratio that measures the average number of days it takes a business to pay its vendors for purchases made on credit. Average payment period is the average amount of time it takes a company to pay off credit accounts payable. Many times, when a business makes a purchase at wholesale or for basic materials, credit arrangements are used for payment. These are simple payment arrangements that give the buyer a certain number of days to pay for the purchase.The average payment period calculation can reveal insight about a company's cash flow and creditworthiness, exposing potential concerns.[48]

In short, payment period is a sensor for how efficiently a company utilizes credit options available to cover short-term needs. As long as it is in line with the average payment period for similar companies, this measurement should not be expected to change much over time. Any changes to this number should be evaluated further to see what effects it has on cash flows.[48]

4.1.13 Average Turnover Period for Inventories

Inventory turnover is a ratio showing how many times a company has sold and replaced inventory during a given period. A company can then divide the days in the period by the inventory turnover formula to calculate the days it takes to sell the inventory on hand. Calculating inventory turnover can help businesses make better decisions on pricing, manufacturing, marketing and purchasing new inventory.[49]

Inventory turnover measures how fast a company sells inventory and how analysts compare it to industry averages. A low turnover implies weak sales and possibly excess inventory, also known as overstocking. It may indicate a problem with the goods being offered for sale or be a result of too little marketing. A high ratio implies either strong sales or insufficient inventory. The former is desirable while the latter could lead to lost business. Sometimes a low inventory turnover rate is a good thing, such as when prices are expected to rise (inventory pre-positioned to meet fast-rising demand) or when shortages are anticipated.[49]

The speed at which a company can sell inventory is a critical measure of business performance. Retailers that move inventory out faster tend to outperform. The longer an item is held, the higher its holding cost will be, and the fewer reasons consumers will have to return to the shop for new items.[49]

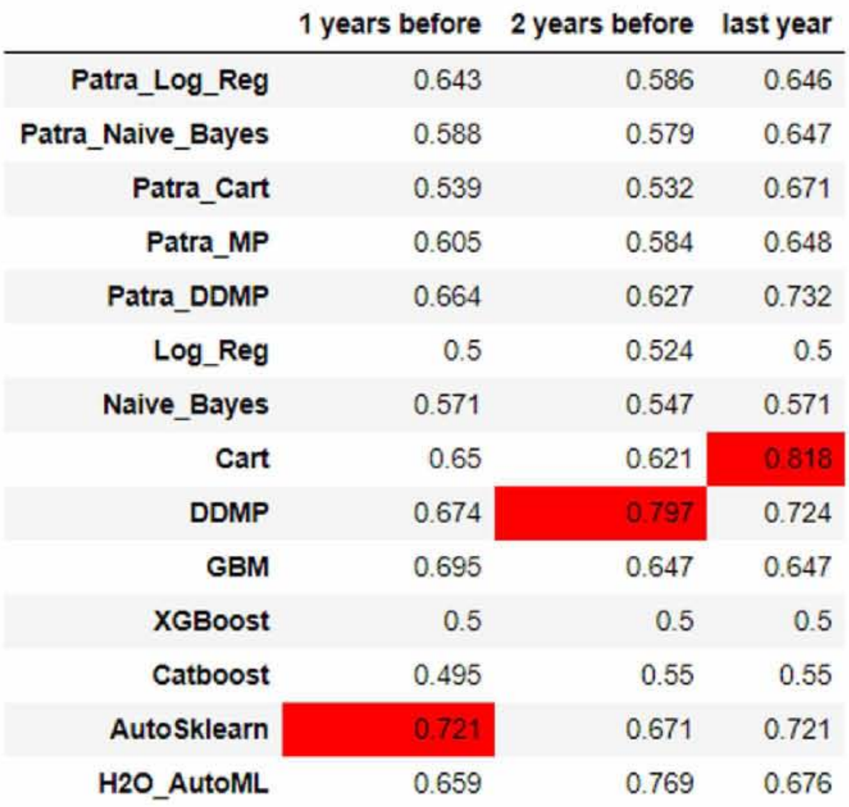
4.2 Empirical Results

During research, as mentioned in the previous chapters, a plethora of algorithms was implemented. The algorithms that were developed include many of the most common machine learning methods that have been used to solve the problem of predicting corporate bankruptcy. These include logistic regression, naive bayes classifier, decision trees, neural networks and various boosting methods such as extreme gradient boosting, cat boost and the vanilla gradient boosting classifier. Moreover, the libraries H2O and Autosklearn were tested so as to use the technology of automated machine learning to possibly find better models as these libraries, since they are quite new in the field, have not been used extensively in the literature.

In order to find the best models during the research, extensive use of the library GridSearchCV was made, so as to make sure that the parameters of each algorithm will be tuned in the best way possible. What is more, since the dataset has been obtained by previous researchers and similar methods have been used, we compare our results with them.[24]

During the research some very interesting results were unearthed. First of all, logistic regression and naive bayes presented some serious underperformance, with the models being random or vely close to it. Furthermore, extreme gradient boosting and cat boost, while being praised as algorithms with very good results, performed very poorly as well, probably because the dataset was too small for these algorithms to perform the desired way.

Below we will present the totality of our results before going further, as well as the tabs that start with the prefix "Patra" are the algorithms that were implemented in previous research, as aforementioned[24]. The metric of measure is the area under the roc curve.



	1 years before	2 years before	last year
Patra_Log_Reg	0.643	0.586	0.646
Patra_Naive_Bayes	0.588	0.579	0.647
Patra_Cart	0.539	0.532	0.671
Patra_MP	0.605	0.584	0.648
Patra_DDMP	0.664	0.627	0.732
Log_Reg	0.5	0.524	0.5
Naive_Bayes	0.571	0.547	0.571
Cart	0.65	0.621	0.818
DDMP	0.674	0.797	0.724
GBM	0.695	0.647	0.647
XGBoost	0.5	0.5	0.5
Catboost	0.495	0.55	0.55
AutoSklearn	0.721	0.671	0.721
H2O_AutoML	0.659	0.769	0.676

As presented above, AutoSklearn, deep dense multilayer perceptron(Neural networks) and Cart Decision Trees had the best performances out of all the algorithms for the datasets "1 years before", "2 years before" and "last year" respectively. Moreover, these algorithms have outshined previous obtained results with an increase of 8.6%, 27.1% and 11.7 % respectively. What is interesting however, is the variety of algorithms that have managed to obtain the best results for each dataset. Moreover, the rest of the algorithms have performed significantly well.

In the "1 years before" dataset, the best result, as can be seen is obtained from AutoSklearn at 0.721, while the lowest(of the algorithms that have not been excluded because of randomness) is 0.65. In the "2 years before" the results fluctuate more, as the range is from 0.621 to 0.797 and in the "last year" dataset, that includes the wholeness of our sample, we range from 0.647 to 0.818. What is important to note is that the algorithms that have performed well, are ones that are generally considered good when the data are unstructured. What can be taken away from that point, is that our dataset possibly includes some notes of randomness rather than linearity, which is to be expected considering the nature of the problem we are trying to solve. To continue we will look at the parameters and the models that have yielded the best results so far.

For the dataset last year, Classification and Regression Trees have performed the best out of algorithms, with 0.818 AUC. In order to achieve this, the impurity measure used was Gini. The number of leafs constructed were 6 and the maximum features considered to build a node were found by transforming logarithmically the number of features in our sample.

For the dataset "2 years before" the best performance was yielded by Neural networks. The model was constructed by using 66 input nodes and 4 hidden layers. The layers had 44, 77, 11 and 11 nodes respectively. For each layer the activation function used was ReLU. The way the number of nodes was found was by using the keras-tuner package, using the same logic as GridSearchCV, so as to find the best model possible. A wide range of parameters were used to make this search as even the slightest change can matter due to their stochastic nature. However, if the model is considered computationally expensive, H2O has yielded very similar results, while Cart,GBM and autosklearn rank a bit lower on this dataset.

For the dataset "1 years before", AutoSklearn has performed the best yielding a result of 0.721. AutoSklearn is not parameterizing as the other algorithms, as it is an automated pipeline. The things you can experiment with is the time needed to run the models, how many models will be built and how many times each of them will be tested. Again, if this model can be considered computationally expensive, Gradient Boosting Classifier has very good results (0.695) and can be built by following these rules: learning rate is to be set at 0.4 , loss is set to "deviance", the maximum depth is 6 while the number of estimators used to build the model is 70.

Chapter 5

Summary of the thesis research

In this chapter, we attempt to present a short overview of the previous chapters as well as some conclusions that were derived through our research. Furthermore, we make an effort to explain the way this thesis was constructed by outlining our research.

5.1 Accounting Ratios

Ratios	Size	Efficiency	Profitability	Liquidity-Leverage
Size	✓			
Scaled Change of Assets			✓	
Change of Net Income			✓	
Gross Profit Margin			✓	
Capital Employed Turnover		✓		
Stockholders Equity Turnover		✓		
Capital Employed/Net Fixed Asset			✓	
Average Collection Period for Receivables		✓		
Average Payment Period		✓		
Average Turnover Period for Inventories		✓		
Debt/Equity				✓
Equity/Capital Employed				✓
Working Capital/Total Assets				✓

TABLE 5.1: Accounting Ratios

The accounting ratios mentioned in the table 5.1 are the ones that were used in our research. Apart from the size of the company, which is in a category by itself, the rest of the ratios can be easily constructed by the financial statements of the company. Based on the way the ratios are constructed, they are divided into three categories, efficiency ratios, profitability and liquidity ratios.

Efficiency ratios, also known as activity ratios, are used by analysts to measure the performance of a company's short-term or current performance. All these ratios use numbers in a company's current assets or current liabilities, quantifying the operations of the business. An efficiency ratio measures a company's ability to use its assets to generate income. An efficiency ratio can calculate the turnover of receivables, the repayment of liabilities, the quantity and usage of equity, and the general use of inventory and machinery.

Profitability ratios are a class of financial metrics that are used to assess a business's ability to generate earnings relative to its revenue, operating costs, balance sheet assets, and shareholders' equity over time, using data from a specific point in time. For most profitability ratios, having a higher value relative to a competitor's ratio or relative to the same ratio from a previous period indicates that the company is doing well. Ratios are most informative and useful when used to compare a subject company to other, similar companies, the company's own history, or average ratios for the company's industry as a whole.

Liquidity is the ability to convert assets into cash quickly and cheaply. Liquidity ratios are most useful when they are used in comparative form. This analysis may be internal or external. Liquidity ratios are an important class of financial metrics used to determine a debtor's ability to pay off current debt obligations without raising external capital. Liquidity ratios measure a company's ability to pay debt obligations and its margin of safety through the calculation of metrics including the current ratio, quick ratio, and operating cash flow ratio.

A combination of ratios, should provide with an entire outlook of a company's financial health. Most successful studies include a combination of 2 up to 21 ratios. The ratios are chosen usually based on the intuition of the researcher and their correlation, as well as their popularity in the literature.

5.2 Algorithms

Algorithms	Undestandability	Computational Expensive	Effectiveness on small datasets	Effectiveness on large datasets
Logistic Regression	✓		✓	✓
Naïve Bayes	✓		✓	
Cart	✓		✓	✓
Neural Networks		✓	✓	✓
Gradient Boosting Machine	✓	✓		✓
XGBoost	✓	✓		✓
CatBoost	✓	✓		✓
AutoSklearn		✓	✓	✓
H2O		✓	✓	✓

TABLE 5.2: Algorithms

The algorithms used in the research are listed in the table 5.2. Algorithms such as Logistic Regression, Naive Bayes, Cart decision trees and the boosting algorithms are more understandable for the user, because of the mathematical background of the algorithm, as well as, the ease of visualizing the results. On the contrary, algorithms such as Neural Networks and the libraries H2O and Autosklearn, due to their complexity, are often considered as a black-box. In addition, Neural networks, the boosting algorithms, H2O and Autosklearn, are computationally expensive and demand a lot of resources in comparison to the rest of the algorithms used in the research. Furthermore, according to our research and the literature, it can be inferred that different algorithms are effective in different situations. As it can be seen in the table 5.2, Logistic Regression can be effective in both large and small datasets, which is not the case for the statistical algorithm, Naive Bayes, which works best in small samples. Decision trees, Neural networks and the libraries H2O and Autosklearn are effective and efficient in both cases, while the family of boosting algorithms work best in large samples. Moreover, while seemingly

Logistic Regression has only advantages, the algorithm can not handle effectively non linear relationships. More particularly, no algorithm presents only advantages, which can be inferred as well from our results.

5.3 Data Availability

Datasets	Easily accessible	Extensive use in the research	Easily Replicable	Free	Large Dataset
Polish Dataset	✓		✓	✓	✓
Stock Exchanges		✓			
Financial Databases		✓			✓
Financial Reports (e.g. 10-K)	✓	✓		✓	✓

TABLE 5.3: Available Datasets

In table 5.3, the ways that one can retrieve samples for the problem of bankruptcy are mentioned. One of the most common datasets, is the Polish dataset, which contains information about Polish companies throughout several years. The dataset was used in a Kaggle competition, therefore it has been analysed extensively.In addition, many researchers, in order to find data about the problem of corporate failure, have gathered information from numerous stock exchanges. The most common ones are the Athens Stock Exchange, the Istanbul stock exchange and the Taiwan Stock Exchange. Another common approach is gathering data from financial databases. Standard and Poor Capital IQ service, Wind Financial Terminal Database & CCER Economic Financial database and Korea Credit Guarantee Fund, are some of the financial databases that have been used extensively in the literature. Another approach that researchers have taken is gathering financial information for the companies from the financial statements and building their sample by themselves. The 10-K report and the Annual Report are public reports that contain such information about US companies, thus it is common to find samples from these sources in the literature.

5.4 References

Title	Model Development	Research	Related to Greek Companies
ZETATM analysis:A new model to identify bankruptcy risk of corporations	✓		
Financial Ratios and the Probabilistic Prediction of Bankruptcy	✓		
Financial Ratios As Predictors of Failure	✓		
FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY	✓		
A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies	✓		
Predicting corporate bankruptcy: Where we stand?		✓	
Business Bankruptcy Prediction Based on Survival Analysis Approach	✓		
Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach	✓		
Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction	✓		
Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks: A case study in Polish companies	✓		
The limitations of bankruptcy prediction models: Some cautions for the researcher		✓	
A Multicriteria Discrimination Method for the Prediction of Financial Distress: The Case of Greece	✓	✓	✓
The pre-bankruptcy procedure for consolidation - evaluation of the effects on greek entreprises		✓	✓
A Review of Bankruptcy Prediction Studies: 1930-Present		✓	

PITFALLS in the APPLICATION OF DISCRIMINANT ANALYSIS IN BUSINESS, FINANCE, AND ECONOMICS		✓	
Forecasting with neural networks. An application using bankruptcy data	✓		
An application of support vector machines in bankruptcy prediction model	✓		
PREDICTING BANKRUPTCY OF INDUSTRIAL FIRMS IN GREECE	✓		✓
Methodological Issues Related to the Estimation of Financial Distress Prediction Models	✓	✓	
Deep Dense Neural Networks for Bankruptcy Prediction	✓		✓
Efficiency of Machine Learning Techniques in Bankruptcy Prediction	✓	✓	✓
Effectiveness of semi-supervised learning in bankruptcy prediction	✓	✓	✓
Corporate Bankruptcy Prediction: a high dimensional approach	✓		
Machine Learning Methods of Bankruptcy Prediction Using Accounting Ratios	✓		
Corporate Bankruptcy Prediction Using Machine Learning Techniques	✓		

TABLE 5.4: References

In table 5.4, we have divided our references in two categories, based on what is the major role of each paper that was mentioned. Therefore, the references fall under the category of model development or research. Moreover, since, our research is related to Greek companies, we also included papers and publications that are related to Greek enterprises as well. Apart from the references mentioned, we have used various articles and sources online such as investopedia, wikipedia, h2o official site, machine learning mastery etc., from which we gathered information about the mathematical background of the algorithms and economical features as well. These are not mentioned in the table since their main role was informative and not critical to our research, but they can be found in the reference section.

Chapter 6

Conclusions and Future Work

Machine Learning and Data Mining techniques can produce a number of rules that may improve the decisions or the predictions that have been made by human experts on several fields. As it concerns bankruptcy forecasting or other similar financial tasks, the consequences of wrong estimations may induce tremendous costs to the corresponding organizations.

In this research, the ability of various machine learning algorithms to predict financial distress and corporate bankruptcy is compared. More specifically, we are interested in performing this research in for Greek companies. The aim of the research is to find a way to predict as accurately as possible, financial distress and to find these indications that will alert the corporate management to take measures to prevent that from happening. The main target, is to help middle sized companies as these type of corporations are the core of a nation's industry. Thus, being able to help these companies to take preemptive measures to ensure their viability, will strengthen in whole the core of the country itself.

To do that we collected a sample of 145 corporations, of which 49 were bankrupt and 96 were still operating at the years we have their financial data. The years of bankruptcy were 2003 and 2004 and in the database, 3 years worth of financial data prior to bankruptcy were included. The database was divided in to three subsamples in a time series manner. Therefore we have one database with data for one year prior to bankruptcy, one for a total of two years and so on.

From the models built, the best performance was achieved by neural networks, decision trees as well as AutoSklearn, which is a new library for automated machine learning. The findings are in accordance with literature, as various researchers have found that neural networks and decision trees have been performing this task successfully.

Furthermore, we have found that the best overall performance achieved was from decision trees, predicting corporate bankruptcy three years up to bankruptcy. What is interesting though, is that this success doesn't translate with an ease of prediction of financial distress two years prior to bankruptcy, while on the last year of concern, the algorithms perform better. That said, the earlier that red flags are raised, the better for the company. Being warned only one year prior to filing bankruptcy, for the most part it might be obvious for the management to have financial concerns. However, three years ahead of time, can be proven helpful to a number of companies to change their course and head into a period of consolidation. The continuing years, can serve a more consultatory role, for a corporation to see if their performance has improved.

Particularly, we have deemed it appropriate to consider the performance of predictive instruments (effectiveness) but also, and mainly, the economic and organisational sustainability (efficiency) of the same. The efficiency analysis represents a fundamental view of the possible uses of the diagnostic instruments within a company/profession. Indeed, models that are particularly effective but excessively costly and cumbersome for the organisational apparatus of the user would still be impossible to use. The effectiveness and efficiency parameters, therefore, must be examined within a systematic viewpoint. Naturally, the assessment of these criteria is characterised by a certain degree of subjectiveness as it depends on the specific characteristics and needs of the user.[50]

Concerning this latter aspect, we have seen that instruments prepared with more up to date techniques (neural networks and automated machine learning) are difficult to sustain by small-medium Greek users. The complexity that accompanies the use of said instruments is particularly demanding on the user's organization as well as having considerable direct and indirect costs. What emerges – at least as far as a full application is concerned – is a type of trade-off between what has been researched in literature on matters of prediction models and what, instead, is required in a day-to-day operation. Even though in recent years researchers have moved towards the elaboration of instruments using increasingly sophisticated techniques, what operators are asking for is the availability of reliable models that are sustainable at the same time. Only by respecting these conditions will models be used profitably on the field. What is optimistic, however, is that Decision Trees represent a comprehensive and computationally effective way to predict bankruptcy. Moreover, this technique has achieved very good results at the three year benchmark making it ideal for this kind of use.[50]

In the future, we aim to include more financial ratios to the existing ones. In the literature a great variety of indicators have been used with success for this task. For that reason, we believe that a more extensive research on this part can yield better results. Furthermore, our dataset, as described beforehand, contains information about the years 2003-2004 and up to three years prior to bankruptcy. Our target is to manage to collect more up to date financial data, as through the years, the economic state of a country changes, which has a direct impact on this research. For example, as the Greek economy is in more ways than one, different than the economy of the US, models that are being used to predict corporate bankruptcy in US, can not be used effectively in Greece. Following the same logic, outdated data can not necessarily be an efficient indicator of the financial health of a modern company. Furthermore, a more in depth analysis can be made per industry of operation. In chapter 4, we discussed about the financial ratios used in this research. Many of which, the ideal range of values varies per industry, therefore making it harder for the model to learn and discern failing companies from healthy ones. Thus, a per industry research, should be more valuable and yield more satisfactory results in this type of problem.

Bibliography

[1] E. I. Altman, R. G. Haldeman, and P. Narayanan, "ZETATM analysis A new model to identify bankruptcy risk of corporations," *Journal of Banking and Finance*, vol. 1, no. 1, pp. 29–54, 1977.

[2] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research*, vol. 18, p. 109, 21 1980.

[3] W. H. Beaver, "Financial Ratios As Predictors of Failure," *Journal of Accounting Research*, vol. 4, p. 71, 1966.

[4] E. I. Altman, "FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[5] P. J. FitzPatrick, "A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies," *Certified Public Accountant*, 1932.

[6] M. A. Aziz and H. A. Dar, "Predicting corporate bankruptcy: Where we stand?," *Corporate Governance*, vol. 6, no. 1, pp. 18–33, 2006.

[7] M.-C. Lee, "Business Bankruptcy Prediction Based on Survival Analysis Approach," *International Journal of Computer Science and Information Technology*, 2014.

[8] H. Ahn and K. j. Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach," *Applied Soft Computing Journal*, 2009.

[9] M. Ziba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, 2016.

[10] L. Hardinata, B. Warsito, and Suparti, "Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks: A case study in Polish companies," in *Journal of Physics: Conference Series*, 2018.

[11] J. S. Grice and M. T. Dugan, "The limitations of bankruptcy prediction models: Some cautions for the researcher," *Review of Quantitative Finance and Accounting*, vol. 17, no. 2, pp. 151–166, 2001.

[12] M. Doumpos and C. Zopounidis, "A Multicriteria Discrimination Method for the Prediction of Financial Distress: The Case of Greece*," Tech. Rep. 2.

[13] I. Tsenes, *The pre-bankruptcy procedure for consolidation - evaluation of the effects on greek entreprises*. PhD thesis, 2018.

[14] A. Panousiadis, *Risk Analysis and Models of Bankruptcy*. PhD thesis, 2016.

[15] wikipedia, "Edward Altman."

[16] J. Gissel, D. Giacomino, and M. Akers, "A Review of Bankruptcy Prediction Studies: 1930-Present," *Journal of Financial Education*, 2007.

[17] R. A. Eisenbeis, "PITFALLS in the APPLICATION OF DISCRIMINANT ANALYSIS IN BUSINESS, FINANCE, AND ECONOMICS," *The Journal of Finance*, vol. 32, no. 3, pp. 875–900, 1977.

[18] D. Fletcher and E. Goss, "Forecasting with neural networks. An application using bankruptcy data," *Information and Management*, vol. 24, no. 3, pp. 159–167, 1993.

[19] K. S. Shin, T. S. Lee, and H. J. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications*, vol. 28, pp. 127–135, 1 2005.

[20] T. Grammatikos and G. Gloubos, "PREDICTING BANKRUPTCY OF INDUSTRIAL FIRMS IN GREECE," tech. rep.

[21] M. E. Zmijewski, "Methodological Issues Related to the Estimation of Financial Distress Prediction Models," *Journal of Accounting Research*, vol. 22, p. 59, 1984.

[22] "h2o AutoML Tutorial," 2017.

[23] M. Feurer, A. Klein, K. Eggensperger, J. Springberg, M. Blum, and F. Hutter, "Auto-sklearn: Efficient and Robust Automated Machine Learning," *Automated Machine Learning*, pp. 113–134, 2019.

[24] S. A. Alexandropoulos, C. K. Aridas, S. Kotsiantis, and M. Vrahatis, "A Deep Dense Neural Network for Bankruptcy Prediction," *Communications in Computer and Information Science*, vol. 1000, 2019.

[25] S. Kotsiantis, D. Tzelepis, and V. Tampaks, "Efficiency of Machine Learning Techniques in Bankruptcy Prediction," *2nd International Conference on Enterprise Systems and Accounting*, 2005.

[26] S. Karlos, S. Kotsiantis, N. Fazakis, and K. Sgarbas, "Effectiveness of semi-supervised learning in bankruptcy prediction," *2016 7th International Conference on Information, Intelligence, Systems and Applications*, 2016.

[27] S. Jones, "Corporate Bankruptcy Prediction: a high dimensional approach," 2017.

[28] L. Yachao and W. Yufa, "Machine Learning Methods of Bankruptcy Prediction Using Accounting Ratios," 2018.

[29] "10-K."

[30] "Annual Report."

[31] "Annual Report vs. 10-K: What's the Difference?."

[32] B. Mattsson and O. Steinert, "Corporate Bankruptcy Prediction Using Machine Learning Techniques," 2017.

[33] S. Z. Mantziaris, *Bankruptcy Prediction Models: An Empirical Analysis of Altman's Z Score Model in Forty Greek Companies in the Period of Economic Recession*. PhD thesis, 2015.

[34] Charalambakis Georgios, "*Πρόβλεψη εταιρικής αποτυχίας ελληνικών εισηγμένων εταιρειών με χρήση διακριτής ανάλυσης*," 2018.

[35] M.-C. Lee, "Business Bankruptcy Prediction Based on Survival Analysis Approach," *International Journal of Computer Science and Information Technology*, 2014.

[36] B. Yildiz and S. Agdeniz, "a Comparative Study of Machine Learning Algorithms As an Audit Tool in Financial," *TIDE Academia Research*, 2019.

[37] J. C. Duan, W. K. Härdle, and J. E. Gentle, *Handbook of computational finance*. 2012.

[38] J. Sindhuja, "The Size of a Firm: Definition, Measures and Concepts."

[39] "Net Income."

[40] "Gross Profit Definition."

[41] "Turnover Ratios Formula."

[42] "How to Calculate Capital Employed From a Company's Balance Sheet."

[43] "Equity Turnover Ratio."

[44] "Net Fixed Assets."

[45] "Debt-To-Equity Ratio - D/E."

[46] "Working Capital to Total Assets Definition and Explanation."

[47] "Average Collection Period."

[48] "Average Payment Period."

[49] "Inventory Turnover Ratio,"

[50] G. Cestari, G. Risaliti, and M. Pierotti, "Bankruptcy Prediction Models: Preliminary Thoughts on the Determination of Parameters for the Evaluation of effectiveness and efficiency," *European Scientific Journal*, vol. 9, 2013.