Vaasan yliopisto
UNIVERSITY OF VAASA

Aleksi Tanskanen

Predicting Corporate Bankruptcy with Financial Ratios and Macroeconomic Predictors

Evidence from Finnish data

School of Accounting and Finance
Master’s thesis in Finance
Master’s Degree Programme in Finance

Vaasa 2020

UNIVERSITY OF VAASA
School of Accounting and Finance
Author: Aleksi Tanskanen
Title of the Thesis: Predicting Corporate Bankruptcy with Financial Ratios and Macroeconomic Predictors : Evidence from Finnish data
Degree: Master of Science in Economics and Business Administration
Programme: Finance
Supervisor: Vanja Piljak
Year: 2020 **Pages:** 82

ABSTRACT:

Bankruptcy is a severe and permanent state of a firm where all stakeholders are facing the consequences, not just the investors. The literature of bankruptcy prediction is an extensive area where new statistical methods have been applied recently.

The purpose of this thesis is to study benefits of using machine learning methods in bankruptcy prediction instead traditional methods such as logistic regression and Z-score by using Finnish data. Furthermore, this thesis tests the use of macroeconomic variables together with firm specific predictors. Lastly, machine learning algorithm called random forest is tested against logistic regression. The adaptation of random forest in bankruptcy prediction is not studied comprehensively.

This thesis employs dataset of 96 995 Finnish firms between the years 1999 and 2019. 2595 firms of this dataset are stated as bankrupt, representing 2.7% of all observations. The financial ratios are derived from Altman's Z-score's variables which reflect the financial state of a firm. The effect of macroeconomic events on predictability of bankruptcy, is tested by employing different macroeconomic predictors such as change in gross domestic product. The robustness checks include careful data cleaning and validating models by splitting data into training and test data.

The results from Finnish data encourage the use of machine learning methods in bankruptcy, especially random forest algorithm. Predictability by using random forest outperformed all other methods introduced in this thesis. Furthermore, the utilisation of macroeconomic predictor in bankruptcy prediction is justified together with firm specific predictors. Particularly, household debt as a proportion of available income shows a significant predictive power on bankruptcy. Lastly, the random forest performed better than logistic regression. This thesis provides encouraging results on bankruptcy prediction in practical purposes against traditional methods such as Z-score that are still used today.

KEYWORDS: bankruptcy prediction, machine learning, random forest, Z-score

Table of Contents

Figures

Tables

Abbreviations

LR	Logistic Regression
RF	Random Forest
GDP	Gross Domestic Product
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
MDA	Multivariate Discriminant Analysis
AUC	Area Under Curve
ANN	Artificial Neural Network
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic

1 INTRODUCTION

Bankruptcy is situation when a business or a person becomes bankrupt (Cambridge Online 2020). In general, bankruptcy is a legal statement that debtholder is unable to repay the debts. In the event of a corporate default, severe consequences are in form of discontinuation of the business. Therefore, bankruptcy is not only a matter of debtholders. Stakeholders such as shareholders, employees, management, and government have direct consequences due to financial distresses. The multiplicative effects of financial distresses in the economy are evident by looking at recessions from the history. Recently, the global economic activity has declined due to the COVID-19 pandemic. There have been discussions, whether this decline of economic activity will cause recession and several bankruptcies in the future, especially if the pandemic is prolonged. Thus, bankruptcy prediction today is even more current topic in the field of finance. Due to the differences in the literature of distress and bankruptcy prediction, bankruptcy is used as synonym for financial distress. A state of a firm is explained in more detail in *Status of failed and healthy firms* section (see 5.1.2).

Despite that firms have a risk of default; lending has been an accelerator of economic growth in the past centuries. Majority of businesses have expenses before the actual income, which is why lending (i.e. investing) plays a key role in an economy. Recently, global debt to GDP ratio has reached all-time high of 322% (Institute of International Finance 2020). Even without the latest purchase plans of debt instruments by Federal Reserve System (FED) and European Central Bank (ECB), the level of debt was still historically high in the end of 2019. Increasing the rate of corporate debt and obligations will affect organization’s financial stability. The less financial leeway a firm has, the more vulnerable it is for financial distresses. However, ever cheapening credit in the future can pay out old debts. This can lead to unnatural balances between firms and changes in financial ratios.

Investors are seeking for firms that are solvent until the maturity of the debt i.e. when the liability is settled. Banks and other investors are striving to maximise the profit of their credit portfolio. Profit is created by the positive correlation between yield of a debt instrument and probability of default, in other words risk-return trade-off. The presence of a corporate bankruptcy has created several credit scoring models in the history trying to predict this likelihood (Altman 2018). Bankruptcy prediction models use variety approaches and methods, but their main source of predictors comes the financial statements.

In most research papers regarding bankruptcy prediction, the focus has been on corporates' internal factors such as financial ratios. Financial ratios have been used as predictors of bankruptcy. The external macroeconomic factors have received less attention (Hol 2007). Controversially, bankruptcies are clustered around economic cycles, and larger companies are less vulnerable to macroeconomic factors (Filipe et al. 2016). For example, in Finland almost 90% of employees are employed by firms that have personnel less than five persons (Tilastokeskus 2020). Small and medium sized enterprises (SME) are more sensitive to macroeconomic risks due to harder access to financing (Filipe et al. 2016). High employment rate of SME's in the economy combined with a higher probability default, gives a strong motive to research more about bankruptcy prediction in Finland.

The relationship between financial ratios and bankruptcy was identified already back in 1930s. The prediction of bankruptcy became popular area to study after the Great Recession (Fitzpatrick, 1932). Rating agencies and financial entities introduced advanced techniques that could predict solvency by using quantitative data analysis in the beginning of 1900s. Univariate ratio analysis and peer-group comparisons were applied in corporate rating purposes. The advantage of these metrics was based on databases which allowed to distinguish the effect of time and industry factors. However, the scope of databases was limited for a long period of time. (Altman 2018)

A multivariate discriminant analysis (MDA) tool Z- score was introduced by Edward Altman in 1968 in the Journal of Finance. Prediction rate of over 94% percent, gave the Z- score attention and it is still used by some professionals. The number of credit scoring models in the last 30 years has increased vastly but the methods and data differ. Growth of databases has enabled machine learning to become more popular in the field of bankrupt prediction. These machine learning methods provide even more accurate models compared to MDA. However, some of the algorithm processes are not always understood by the user due to complexity. The causality and relationships in machine learning (especially neural network) techniques may be unshown. Thus, the use of some of ML techniques amongst practitioners and researchers remain uncertain. (Altman 1968, Altman 2018)

1.1 Previous studies & Hypotheses

In this section significant previous bankruptcy models are discussed. Thereafter, three hypotheses are conducted around the bankruptcy prediction. The three hypotheses are structured based on encouraging results from previous literature.

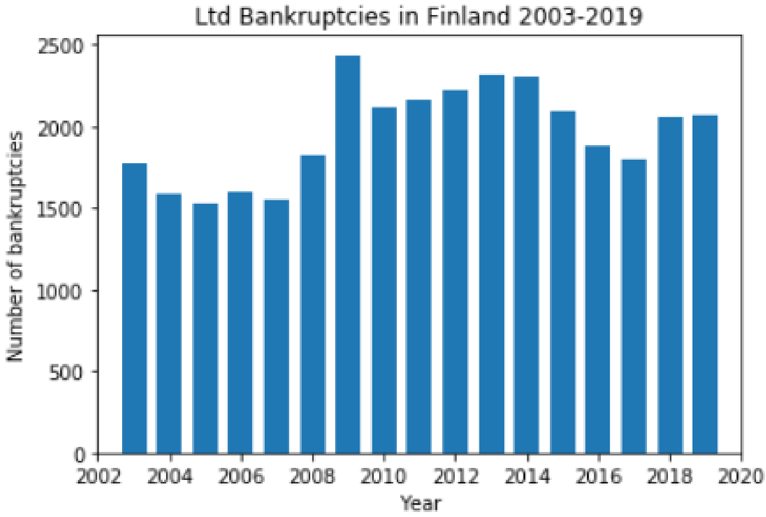
The earliest bankruptcy prediction models were discovered in the 1800s but the contribution for today’s quantitative analysis remains low. Interestingly, some models from 50 years ago still get attention in today’s financial literature. Furthermore, some of these techniques are still applied by practitioners. Univariate discriminant analysis of the financial ratios was carried out by Beaver in 1966. Individual financial ratios were found to be robust predictors, some even 5-year prior to bankruptcy. The selection process of financial ratios by Beaver (1966) was influenced by three aspects: popularity of ratios in literature, performance in previous studies and use of cash flow ratios. Development of new ratios was not intended by Beaver, but he rather tested the prediction power of existing ones. Two years later, the discovery of the multivariate discriminant analysis (MDA) in bankruptcy prediction called Z-score by Altman (1968) revolutionised prediction model. The use of several predictors at the same time, put the Beaver’s findings into a practical form. A high model predictability of bankruptcy over 90% and ease of usage led to popularity of Z- score amongst researchers and practitioners in the field of finance. Z-score model is still studied and used by financial professionals as a benchmark for their own models. (Altman et al., 2017, Altman 1968) Nonetheless, the original Z-score predicts the default probability based on the sample of American firms over 50 years ago. Nowadays corporates use derivatives to hedge their businesses which has made the ratio analysis more complex.

The dynamics of the corporate world have changed over time which has caused the financial ratios to change as well. (Altman 2018)

Logit bankruptcy model was introduced by James Ohlson in 1980. The model overcame statistical assumptions regarding the popular MDA as normal distribution of the predictors was not required. Additionally, the result of the model was stated as a probability unlike in Z-score, where the value itself is ambiguous number. (Ohlson 1980)

The empirical part of this study classifies companies by using different machine learning methods. Logistic Regression (LR) and Random Forest (RF) are selected as methods of building an optimal model. Additionally, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are made for re-estimation of Z''-score variables (LDA) and QDA to challenge the test the statistical assumptions in LDA. LR is widely used in the previous literature and has proven its effectiveness compared to Z''-score in classification problems such as Altman et al. (2017). The use of other machine learning methods such as decision trees are not as popular in the literature of bankruptcy prediction. Still, decision tree models have shown their ability to perform well in other sciences see (Muchlinski et al. 2015) as well as in finance (Rudd et al. 2017). Additionally, Muchlinski et al. (2015) justify the use of RF in rare event binary classification problems. Utilizing RF might be handy when the data is imbalanced. The proportion of bankruptcies per year is relatively low compared to total number of firms. The number of limited liability companies' bankruptcies between 2003 and 2019 is shown in the Figure 1.

Figure 1. Ltd. bankruptcies in Finland (Statistics Finland 2020)



According to Statistics Finland (2020), the number of bankruptcies in Ltd. firms has been below 2000 on average for almost 20 years. The total number of Ltd. Companies in January 2019 was 272 084 (Finnish Patent and Registration Office, 2020). Thus, bankruptcies represent for only about 0.7% of all Ltd.'s in the economy. The proportion of bankruptcies to total Ltd.'s is expected to remain approximately the same through different time periods. Therefore, consideration of the emerging new companies is not in the interest in this thesis. Minor changes in the ratio of bankruptcies to total firms is not expected distort the results.

The first hypothesis suggests that the classification performance of area under curve (see 3.1.2) is improved by new statistical methods. Altman et al. (2017) found only a little improvement by re-estimating the Z''-score with new data but found greater Area Under Curve (AUC) by using LR. The use of additional variables in the study generally improved the model's performance, but the results were different across countries. By having different predictors than in Z''-score (X1, X2, X3 and X4) makes no fair comparison between Z''-score and other methods of classification while testing hypothesis H_1.

H_1: *Z''- score is outperformed by Logistic Regression and Random Forest.*

For the sake of fairness, same predictors are used in all three methods. The Z’’-score will be used as a benchmark for logistic regression and decision tree approach. This procedure is following the same criterion as in previous studies such as Altman et al. (2017).

Strong support of utilising macroeconomic variables is evident from the literature of bankruptcy prediction (Filipe et al. 2016, Laamanen 2015, Hol 2007 & Altman 1983). Altman (1983) found that the failure rate of businesses increases with a lower real economic growth, stock market performance, money supply growth and increased business formation. Laamanen (2015) used in her thesis accommodation and restaurant industry data from Finland to predict the failure of firms. Significant improvement of the model was found by using gross domestic product (GDP) as an additional predictor. Hol (2007) found prediction power of GDP, production index and money supply (M1) by using data from Norwegian firms from 1990s. Altogether, there is a strong evidence about the correlation between economic cycles and occurrence of bankruptcy. Therefore, this thesis studies this correlation effect by analysing the performance of new predictors.

> H_2*: The performance of the bankruptcy models can be improved by including a macroeconomic predictor*

The use of RF is not popular in the literature of finance. However, there is evidence of benefits of applying RF on different statistical problems proves its superiority (Muchlisnki et.al. 2015). Rare binary events were not distinguished by LR as good as with RF in the study. This supports the use of RF. As previously mentioned, bankruptcy is a rare event in the economy compared to total number of healthy firms. A binary class-imbalance might deteriorate the performance of LR, but no balancing of data should not be done from sample bias point-of-view.

> H_3: *Random forest outperforms the use of Logistic Regression*

Joshi et. al. (2018) used RF to predict the bankruptcy from carefully selected variables. In this study, the use of RF outperformed traditional decision tree method by reducing variance and diminished overfitting. Causality is a key interest in science. For the most part in the literature of finance, study of causality carried out by regression which is a great tool for analysing this. However, the practical perspective and non-linearities in bankruptcy prediction need more attention. Therefore, new methods should be studied without preconceptions.

1.2 Structure of the thesis

The second chapter of this thesis reviews the theoretical aspect of bankruptcy and discusses about the popular bankruptcy prediction models from the literature of finance. The third chapter discusses the choice of firm specific and macroeconomic variables used in the thesis. In the fourth chapter, the methodology and basics of statistics used in this thesis are explained. The fifth chapter discusses about the data and predictors that are derived financial statements and macroeconomic data. Sixth chapter describes the univariate properties of financial ratios and is continued by constructing the different prediction models. The seventh chapter summarizes the results obtained from chapter six and reflects them on to hypotheses. Lastly, suggestions for future research are discussed.

2 THEORY OF BANKRUPTCY

This chapter discusses the reasons behind the business operations that lead to a bankruptcy. Thereafter, relevant prediction models and their statistical methods are presented.

2.1 Bankruptcy

The purpose of this section is to define what is a bankruptcy and what causes firms to fail. Yet, the purpose of this study is not to investigate operational level errors that could lead to a bankruptcy, but the predictability of bankruptcy using financial data. Thus, operational discuss remains limited. Quantitative data allows stakeholders to exploit a bankruptcy model that utilises income statements and other public sources for macroeconomic data.

2.1.1 Definition of a bankruptcy

A corporate firm’s balance sheet consists of assets and liabilities & owner’s equity. Assets are the items that company owns and liabilities & owner’s equity are those items that firm owes to other participants. Owner’s equity is not becoming due, but it is still considered as a liability. Owner’s equity is paid out as dividends to shareholders if sufficient funds are found. In order the business to continue, assets should be greater than liabilities in the long run. If the capital required to pay back the liabilities is not sufficient, a company may become insolvent. Insolvency can be temporal until new capital is accumulated. However, prolonged insolvency could lead to default. In the event of default, legal reorganization might be beneficial to stakeholders. Corporate reorganization inhibits the use of capital which is used as a last resort to make company solvent again. Failing of firm’s reorganization leads to a legal statement of bankruptcy. Bankruptcy is the most severe form of financial distress and usually has serious consequences to third parties. (Laitinen & Laitinen 2004)

2.1.2 Path to a bankruptcy

Financial ratios reflect the financial state of a firm. Furthermore, the ratios are a consequence of events from the operational level. These events take place before they are visible in numbers. Therefore, predictability of bankruptcy one- or two-years prior is essential. External factors such as change in GDP, will often be reflected on financial ratios as well. In other words, financial ratios and macroeconomic predictors might be correlated and incomplete correlation could lead to a higher predictability (Altman et al. 1984). Thus, predictors are divided into internal and external factors. This categorization will be applied later in empirical part of the thesis.

Laitinen (1990) divided the reasons leading to a bankruptcy into nine different categories concerning for example experience of management, strategy, marketing, poor adaption in new situations, risk diversification (vulnerable key roles, old equipment), systematic risks (country specific business cycle, devaluation of a currency) and increased competition in the industry. These paths to a bankruptcy are visible in all industries and finally they are reflected in the financial ratios of the firm. Still some macroeconomic factors such as current interest rate is instantly observable. This allows a model with macroeconomic predictors to react faster than a model with firm specific ratio. Firm-specific factors reflect the historical performance and are derived from the financial statement. Thus, macroeconomic factors might give early signals of the bankruptcy and improve the bankruptcy model's prediction.

Lussier (2005) investigated the effect of 15 firm specific variables in the bankruptcy prediction in the real estate industry by using logistic regression. Controversially, this study used non-financial predictors with the real estate industry data. Lussier found that relevant industry experience by management, higher age, use of professional advisors, specific business plan and appropriate capital structure will lead to a higher probability of success. The data was limited only for real estate, yet consensus of factors leading to bankruptcy is coherent in the literature.

2.2 Bankruptcy models

This section discusses the popular recent bankruptcy models in the literature. First the Beaver model is discussed and then continuing to the three Altman's Z-score models. Lastly, the most recent models in bankruptcy prediction are briefly discussed.

The history of the credit scoring models goes all the way back to 1800s when money lenders needed to have information about the lender's credibility. Information about the creditability was mostly subjective and in a qualitative form. In the early 1900s the scoring system took steps towards a quantitative analysis. Data was collected from peer firms and use of a timespan enabled a robust analysis of the credibility of a firm. Last 50 years, the big data has shown its superiority of analysing the creditworthiness of a firm. (Altman 2018)

2.2.1 Beaver's model

A remarkable ratio-based study called *Financial Ratios as Predictors of Failure* was written by Beaver in 1966. This study researched the relationship between financial ratios and failure of a firm by using a univariate analysis. The purpose of the study was not to create a perfect failure model but rather investigate the prediction power of the financial ratios. By using the cashflow to total-debt ratio, Beaver could classify firms reliably into bankrupt and solvent even five years before the event. However, another important finding was made. If a financial ratio can predict the bankruptcy before it happens, the ratio analysis may provide useful information to management for changes.

All the financial ratios did not predict the failure, but for example cashflow-to-total debt ratio performed extremely well in prediction. This Beaver's pioneering study served as a starting point on further studies of multivariate analysis of failure prediction.

2.2.2 Z- score model

The original Z-score-model was created by Altman in 1968. This study stood as a continuum to a Beaver's (1966) study. Altman shifted from univariate approach to a multivariate discriminant approach to predict the bankruptcy which enabled to use several financial ratios at the same time. The ratios were given certain weights by their significance of prediction ability. The data of bankrupt firms was gathered from National Bankruptcy Act from 1946 to 1965. The total asset size of the data ranged from 0,7 and 25,9 million USD while mean size being 6,4 million USD. Altman found out that this group was not homogenous from size and industry point-of-view. The data of non-bankrupt firms were constructed randomly but stratified by size and industry. Additionally, the year of observation in the bankrupt sample was matched with the non-bankrupt to counteract the possible bias of time effect. (Altman 1968)

The purpose of the MDA method is to divide the data into different groups of interest, bankrupt and non-bankrupt. After grouping, the MDA finds the linear combination of the variables that separates the groups the best. Set of coefficients (weights) for the variables are derived which indicates the importance of the specific financial ratio. The variables were categorized liquidity, profitability, leverage, solvency, and activity ratios. Altman chose 5 variables out of 22 based on popularity in previous literature and relevancy to the study. The Z-score model had five financial ratios: X_1 = Working Capital/Total Assets, X_2 = Retained Earnings/Total Assets, X_3 = Earnings Before Interest and Taxes/Total Assets, X_4 = Market Value of Equity/Book Value of Total Debt and X_5 = Sales/Total Assets. (Altman 1968)

Working Capital/Total Assets ratio (X_1) measures the net liquid assets to total assets of the firm. Working capital is defined by the subtraction of current liabilities from current assets. Commonly, consistent negative profit will result in decreasing current assets to total assets. Two popular optional liquidity ratios of current and quick ratios showed lower significance on univariate and multivariate basis compared to X_1. (Altman 1968)

Retained Earnings/Total Assets (X_2) measures the accumulated profit related to total assets during the lifetime of the firm. Retained earnings is correlated with the age of the firm well. Thus, older firms tend to have bigger X_2 than younger ones due to accumulation of profits over the years. The discrimination against young firms is raised but over 50% of manufacturing firms failed in the first five years and over 31% in the first three years (The Failure Record 1965; Altman 1968)

Earnings Before Interest and Taxes (EBIT)/Total Assets (X_3) measures the true productivity of the firm without considering taxes and capital structure in the form of interest. The continuum of the firm is based on the earning power of its assets which EBIT measures. The liabilities can be paid out in the future with strong EBIT and taxes are based on positive profit. Furthermore, this ratio is popular amongst corporate prediction literature. (Altman 1968)

Market Value of Equity/Book Value of Total Debt (X_4) is measured by dividing the common and preferred stock values by all debt. This ratio indicates how much the firm can lose its asset value before it becomes insolvent (liabilities are greater than the assets). This ratio was introduced as a new measure to the literature and suggested to outperform a more common related measure of net worth/total debt. (Altman 1968)

Sales/Total Assets (X_5) indicates the ability of the firm to generate sales to its assets. Also, this ratio has been used as a metric for the management's performance in competitive markets. However, on a univariate basis X_5 contributes the least but combined with other variables it is the second important of all variables. (Altman 1968)

Arbitrary number of Z is calculated by the equation (1) and then compared to the three range zones.

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \qquad (1)$$

$Z' > 2,9$	Safe Zone
$1,81 < Z' < 2,99$	Gray Zone
$Z' < 1,81$	Distress Zone

The non-bankrupt sector is considered healthy which means that the probability of bankruptcy in two years is unlikely. The grey area sector has a high probability of misclassification and thus it is not reasonable to make conclusions of such a firm. Altman also describes it as “zone of ignorance”. The classification performance by using the training data was high (94%) for initial sample (33 observations) and 95% for all data (66 observations). However, the sample size is considerably small which can lead to generalization problems for new unseen data. (Altman 1968)

2.2.3 Z’- and Z’’- score- model

In 1983, Altman updated the original 1968 model by using in X_4 the book value of equity instead of market value. This transformation made Z’-score model applicable for private manufacturing firms as well. The coefficients of the model were re-estimated, shown in equation 2. (Batchelor 2018)

$$Z' = 0.717X_1 + 0.847X_2 + 0.3107X_3 + 0.420X_4 + 0.998X_5 \qquad (2)$$

$Z' > 2,9$	Safe Zone
$1,23 < Z' < 2,9$	Gray Zone
$Z' < 1,23$	Distress Zone

After ten years, two new models called Z’’-score were introduced by Altman in 1993. The scope of this model was expanded to non-manufacturing firms (see Equation 3) and companies from emerging markets (equation 4) as well. The variable of X_5 was removed due to minimization of the industry effect of asset turnover. By doing so, Z’’-score model was less sensitive for different industries. However, the X_4 variable was substituted back with the market value instead of book value. Naturally, after modifications the model coefficients were re-estimated. (Batchelor 2018)

$$Z'' = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4 \tag{3}$$

$$Z'' = 3.25 + \ 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4 \tag{4}$$

Z'' > 2.6	Safe Zone
1.1 < Z'' < 2.6	Gray Zone
Z'' < 1.1	Distress Zone

2.2.4 Ohlson model

In the 1980s James Ohlson's study was published in the Journal of Accounting Research about bankruptcy prediction. In this study, he attempted to find better model for bankruptcy by using 1970-1976 data. Ohlson used conditional logit model to predict the probability of failure due to strict statistical assumptions of MDA. MDA requires the sample of failed and non-failed firms' variance-covariance matrices to be equal and their predictors to be normally distributed. A major contribution to the previous literature was achieved by using data that was released before the bankruptcy declaration. By using this approach, Ohlson achieved realistic forecasts about the probability of failure as the model was trained on data obtained before the bankruptcy. (Ohlson 1980)

Table 1. Ohlson's predictors (Ohlson 1980)

	Variable	**Explanation**
1	**SIZE**	*log(total assets/CNP prioce-level index)*
2	**TLTA**	*Total Liabilities / Total Assets*
3	**WCTA**	*Working Capital / Total Assets*
4	**CLCA**	*Current Liabilities / Current Assets*
5	**OENEG**	*1: Total Liablities > Total Assets, 0: otherwise*
6	**NITA**	*Net Income / Total Assets*
7	**FUTL**	*Funds provided by operations / Total Liabilities*
8	**INTWO**	*1: Net income < 0 for last two yearss, 0: otherwise*
9	**CHIN**	*(NIt - NIt-1) / (|NIt| + |NIt-1|), where NI = Net Income*

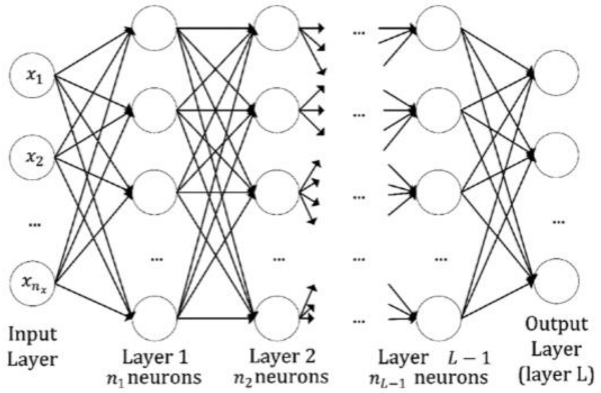
Nine different predictors were used in the study and they are shown in the Table 1. Ohlson did not attempt to create any “new or exotic” ratios but rather chose predictors purely based on previous literature. He found four statistically significant factors affecting the probability of bankruptcy. These factors were size, financial structure, measure of performance and measure of liquidity. Three latter factors are identical Laitinen study (1992, 190), where profitability is inevitable part for continuum of business. However, without a stable liquidity and financial structure, profitability becomes meaningless. (Ohlson 1980)

2.2.5 New models

Until 1990s, the credit risk models used in the literature have been dominated by MDA and logit models. Previously, univariate models have made contribution to the literature by analysing the ratios. However, i.e. Beaver (1966) only studied the effect of specific ratios individually but applicable real-life model was not put into practice. After 1990’s the machine learning methods have increased popularity in the literature. Big data and an increase in the computational power have made this transition possible towards more complex models. Decision trees, K-Nearest Neighbour, Support Vector Machine (SVM) and Naive Bayes classifier are just few examples of machine learning classification methods which have received more attention recently.

Excellent performances of Artificial Neural Networks (ANN) has gained popularity in the literature in past decades. The benefit of an ANN is that it can detect highly non-linear and complex patterns from the data independently by using neural network. A structure of an ANN is presented in Figure 2.

Figure 2. Artificial neural network structure (Michelucci 2018)



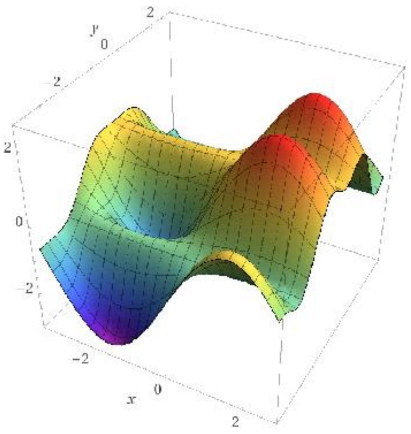
The ANN is constructed from three types of layers called input layer, hidden layer(s) and output layer. The hidden layers have so-called neurons which take in values with certain weights *w* from previous layer. This weighted sum (commonly referred to *z*) is then passed to an activation function (non-linear) which calculates the value *f*(*z*) for the next neuron. The procedure can be repeated with several layers until the output layer is reached, giving the final prediction. This makes the ANN very complicated to understand by human, but the exceptional performance and increased computational power have made ANNs popular lately. (Michelucci 2018)

Using SVM however, has shown encouraging results in the field of bankruptcy prediction. Briefly, SVM utilises non-linear boundaries to find categories in a multidimensional feature space. A hyperplane refers to plane that has -1 dimensions than the original space. This non-linear hyperplane has been able to separate classes effectively. Min & Lee (2005) studied the use of SVM in prediction, and found it to be preferable compared to ANN, LR, and MDA. The use of ANN is more likely to overfit and the success of ANN is heavily influenced by the user. (Min & Lee 2005)

Min et. al. (2006) integrated SVM with genetic algorithm which improved the original SVM model. Additionally, the use of structural risk minimization principle used by SVM outperforms the popular empirical risk minimization used e.g. in ANN (Min & Lee 2005). The benefit of structural risk minimization comes from finding the global minimum risk instead of a local one (Min & Lee 2005).

ANN uses gradient descent to find the minimum empirical risk which can be difficult with non-convex problems. Stochastic gradient descent used also in regression problems to fit the model. The Figure 3. shows an imaginary 3-dimensional plot of challenge that ANNs face. The original image (Lagandula 2019) does not represent ANN loss, but it illustrates the problem of finding global minimum well in ANN. X and y axes present weights *w* of features whereas *z* indicates the loss respect to *x* and *y*. The ANNs algorithm moves towards the smaller error (risk) by using gradient descent. However, the learning rate of ANN determines the size of steps to be used. As we can see, there are several "valleys" in the Figure 3. The minimization process can get trapped into one of these "valleys" (local minimum), leading to a false interpretation of global minimum. (MIT Introduction to Deep Learning 2020)

Figure 3. Empirical Risk Minimization with two features (Lagandula 2019)



The controversy about the superiority of methods remain uncertain. Both of SVM and ANN have shown great performance results but techniques such as of logistic regression and MDA are still popular in the literature.

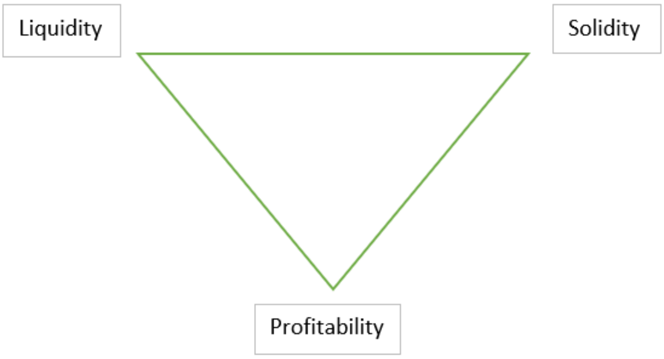
3 CHOICE OF FIRM SPECIFIC AND SYSTEMATIC VARIABLES

This chapter validates the use of predictors used in this thesis. First, the choice firm specific predictors are discussed based on previous studies. Thereafter, the macroeconomic risks are introduced and the use of three different macroeconomic predictors are justified.

3.1 Firm specific variables

The firm-specific variables should explain the probability of bankruptcy as much as possible. Laitinen (1990, 170) used liquidity, solidity, and profitability to describe the continuation of business. The triangle (Figure 4) describes the relationship between three main components of business continuum. The base of the triangle consists from profitability which the most crucial part of these three aspects. In the long run, company needs to be profitable to exist. Profitability holds the two parts of solidity and liquidity and is responsible for stability of this triangle.

Figure 4. Prerequisite for business continuum (Laitinen 1990, 171)

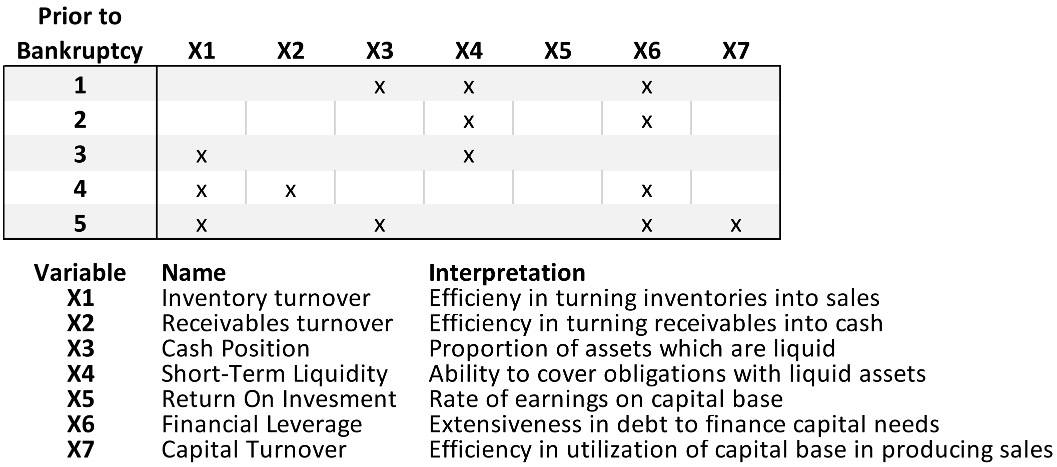


Each of these three categories have different financial ratios that can predict the bankruptcy. If one field is removed, triangle will not be stationary anymore (Laitinen 1990, 170-171). A firm is as strong as the weakest link of these three categories (Laitinen 1990, 172). Therefore, at least one variable should reflect at least one category in models constructed in this thesis.

Four firm specific variables used in this study are identical with the Z’’-score’s X1, X2, X3 and X4. This allows to test the hypothesis (H_1) about the performance of statistical methods. Additionally, the variables are suitable for private companies which can extend the scope of data used from Orbis database. First variable of X1 reflects the liquidity of a firm by dividing the working capital by total assets. Working capital defines the short-term operational flexibility and it is turned into a ratio for comparison purposes with other firm sizes. The total assets represent the total ownership of long-term and short-term assets and is the total size of the firm. In variable X2, retained earnings are divided with total assets for the same purposes as previously. The retained earnings itself reflects the age of a firm but also the profitability. Two identical firms with same age and total assets can differ by the profit margin. Therefore, values of X2’s are different. In other words, X2 is linked to the most important feature of profitability in the triangle (Figure 6). X3 value is derived from EBIT divided by total assets. The EBIT stands for the base of triangle as X3 indicates the current profitability. The capital structure and tax load are not considered (see 2.2.2). The variable of X4 represented the book value of equity to book value of total debt. This insolvency measure indicates when liabilities are greater than assets. However, the original Z-score used market value of equity which reacts faster to changes in equity than book value. However, both variations of the X4 reflect solidity (stability) of a firm. Altman (1968) used sales/total assets (X5) as fifth ratio that contributed to the industry specific properties and competition conditions. However, X5 was found non-significant on an individual level but enhanced the performance of the bankruptcy model. (Altman 1968)

The time between the features (data) obtained and event of bankruptcy ($t = 0$) should be considered while constructing a bankruptcy model. Zavgren & Friedman (1988) studied the significance of ratios depending on the time to bankruptcy. Five different ratios (not the same X predictors as in Z-score) and their significance were studied five years prior to the bankruptcy from 1979 to 1983. The results are shown in the Figure 5.

Figure 5. Significance (α=95%) of variables in Logit model (Zavgren &Friedman 1988)



Prior to Bankruptcy	X1	X2	X3	X4	X5	X6	X7
1			x	x		x	
2				x		x	
3	x			x			
4	x	x				x	
5	x		x			x	x

Variable	Name	Interpretation
X1	Inventory turnover	Efficieny in turning inventories into sales
X2	Receivables turnover	Efficiency in turning receivables into cash
X3	Cash Position	Proportion of assets which are liquid
X4	Short-Term Liquidity	Ability to cover obligations with liquid assets
X5	Return On Invesment	Rate of earnings on capital base
X6	Financial Leverage	Extensiveness in debt to finance capital needs
X7	Capital Turnover	Efficiency in utilization of capital base in producing sales

Zavgren & Friedman (1988) concluded that the variables have either long- or short-term prediction abilities. For example, equity turnover X7 indicates the ability to accumulate sales on capital and was significant only 5 years prior to bankruptcy. Due to the uncertainty about the costs (misclassification of a firm) of type 1 and type 2 error, total classification error was used to evaluate the performance with different timespans. From years 1 to 5, classification errors were 18%, 17%, 28%, 27% and 20% respectively. It is noteworthy that statistically non-significant variable can enhance the performance of the model. These findings encourage to use predictors t ≤ 3 years in this thesis concerning liquidity (X3 and X4) and solidity (X6).

The exact year of bankruptcy is not indicated by Orbis database. However, bankruptcy is expected to occur one or two years after the indication of last available year (Altman et.al. 2017).

3.2 Macroeconomic predictors

Macroeconomic events are important events predicting the bankruptcy (see Laitinen (1990), Laitinen & Laitinen (2004), Filipe et. al. 2014, Hol (2006)). Utilising the macroeconomic data could benefit from frequent predictor update, unlike in pure microeconomic models (financial ratios) where data is received annually or quarterly in the form of financial statements. Nevertheless, all data in this thesis is annual for simplicity and availability.

Laitinen (1990: 27) found that 61% of increase in bankruptcies in Finland during were explained by business cycle, inflation, ease of financing and trade balance. Business cycle and inflation from these categories contributed most to the probability of bankruptcy. As for Filipe et. al. (2014) used three categories of country-specific systematic variables of business cycle, credit conditions and insolvency codes.

Laitinen & Laitinen (2004) divide macroeconomic factors into four categories: *business cycle, inflation, ease of financing* and *trade balance*. The analysis is made from the perspective of Finnish economy, but the findings generalize well with other studies and countries. *Business cycle* is linked to bankruptcies through demand. In an economic downtrend, less demand for the products and services are needed. Therefore, income financing decreases and results in a deteriorated liquidity. A firm may need to invoke for liabilities which results in a higher gearing. However, in an economic uptrend, the demand for goods and services is high. Surviving from liabilities in this kind of environment is easier, but firms can still expand too fast. Expanding too fast might result in poor management and financing which increases the risk of a bankruptcy. (Laitinen & Laitinen 2004)

Second important macroeconomic factor affecting bankruptcies, *inflation* can have positive and negative effects. Negative effects can result from higher prices of purchases from production if the firm is not able to pass the inflated prices to the customers. On the other hand, an indebted firm can benefit from increased inflation when the value of liabilities decreases. The nominal interest rate being lower than inflation rate, firm is profiting by having liabilities. (Laitinen & Laitinen 2004)

The third macroeconomic factor, *ease of financing* might affect to the probability of bankruptcy both ways. By having strict rules for financing, a firm with a poor liquidity can experience default and/or the cost of debt increases. Strict financing, however, can reduce riskier projects and might decrease the number of new firms in the economy due to stricter rules or higher expected rate of return for capital. Newer firms tend to fail in the early years and thus scarcity of them can reduce overall bankruptcies in the economy. The net effect of financing is dependent on whether the money is being used to give aid for liquidity or investing in new riskier firms. (Laitinen & Laitinen 2004)

The fourth factor is the *trade balance* of an economy. Increase in exports can expand the markets and leads to a higher demand for domestic products and services. This generally helps the business environment and reduces risk of bankruptcy. Exporting goods is still considered riskier than selling domestically which results in a higher proportion of riskier firms. Changes in import can have both good and bad effects. Increase of imports can mean tightened competition of domestic firms. Eventually, the lower demand can lead to bankruptcies. Increased imports can also mean cheaper factors of production. This allows firms to produce goods and services with a lower production costs, ultimately easing the competition. (Laitinen & Laitinen 2004)

The firm specific and systematic risks of firm distress was studied by Filipe et.al. (2016) on European small and medium size enterprises (SMEs) during 2000 – 2009. SMEs were found to be sensitive on same firm specific predictors. However, the effect of macroeconomic predictors varied within the data by groups of countries. Another major finding was that the smaller SMEs were more sensitive to systematic risks than larger ones. 15 different macroeconomic variables were studied, and they were categorized by business cycle, credit conditions, financial market and insolvency codes. The of the significance of the macroeconomic variables was following; fit models by using firm specific ratios, include one systematic variable at a time, calculate the AUCs and keep the predictors with the highest AUC values. To validate the causality and coefficient estimates by LR, a correlation between systematic variables were measured. A correlation coefficient of over 0.6

between two features resulted in exclusion with the lower AUC. The firm-specific predictors were found significant in the generic model 2, even when FX rate, unemployment, economic sentiment indicator and change in bank lending were included in the model. Additionally, all these macroeconomic factors were statistically significant on the 0.1% level. A shift from generic model to a regional model, resulted in major changes in the magnitudes of systematic coefficients. The significance of GDP change and bank lending contributed well in the prediction of distress and were inversely related to probability of distress. (Filipe et.al. 2017)

In this thesis, *GDP change (%)* and *household debt level & interest expense on available income* are used as a macroeconomic predictors. The GDP change describes the overall state of an economy. A negative GDP means decreasing amount of goods and services produced in the economy and the natural expected results is a higher rate of bankruptcies. Also, Filipe et.al. (2017) found GDP to be a very crucial part of predicting distress in a regional model. This thesis uses even more restricted data from Finland which could result in a good prediction power of a nation's GDP change. Contradictorily, Hol (2007) found no significance of *GDP change* but only with the *GDP gap* in Norwegian unlisted firms. The two predictors were used together in the model and could result in wrong conclusions about the significance. Therefore, Hol (2007) highlighted the contrast between her finding about the GDP with Altman (1971). In the early study by Altman (1971), an inverse relationship between nationwide failure rate of railway companies with overall economic activity (real Gross National Product, GNP), stock market performance (S&P 500 index) and money supply conditions were found. Another study by Altman (1983) studies the effect of macroeconomic events on businesses. In short, the business failure rate on American firms during 1951-1978 was increased by cumulative effects of reduced real GNP, stock market performance, money supply and enhanced new business formation (Altman 1983). Based on the literature review, the use of GDP change as a macroeconomic predictor is strongly suggested.

The *household debt & interest expenses of available income* provide unique point-of-view to bank lending on households. These two predictors reflect the state of an economy by different views. The amount of debt is usually high when people are confident about their future. This could result in longer contracts and higher gearing of households. In a normal monetary policy, the interest rates tend to be higher when economy is growing in an uptrend and vice versa. This phenomenon is reflected in the interest expenses that households pay. However, the salaries and available income higher during an uptrend. At the same time, the available income might be lower due to layoffs. Ultimately, these macroeconomic predictors reflect, how well households are doing at certain time. Many firms are directly influenced by the private spending, some with a lag. Thus, bankruptcy prediction models could benefit by utilising *GDP change (%)* and *household debt level & interest expense on available income* predictors.

4 STATISTICAL METHODS

The focus of this thesis is to find a well-performing bankruptcy model and compare them by ROC AUC (see 4.1.2). Bankruptcy is always a caused by real-world events such as choices of management and changes business environment. Bankruptcy usually does not happen overnight but with incremental negative changes in business. Thus, these events can be observable from data even five years prior (see Beaver 2.2.1). Consequently, this thesis relies heavily on statistical methods trying to identify these early signs. The quantitative feature of this thesis motivates to introduce the statistical methods in more detail. In this thesis, a binary classification models are utilised, meaning that a single firm can only be either bankrupt or non-bankrupt (healthy). First necessary concepts of machine learning with validation criterion of the models. Lastly, four different statistical methods are presented.

4.1.1 Basic concepts of machine learning

Three basic principles of machine learning are *data, hypothesis space* and *loss function*. These principles cover all the choices that are made to predict the dependent variable from the independent variables. The first component, *data* consist of *features* and *labels*. *Features* can be derived from the data points which are fundamental measured values. These features are usually referred as independent *x* values. For example, in this paper data points could be income statement values but then computed as a specific ratio like in Z-score. In other words, a f*eature* can be any predictor value that can be computed from data points. *Label* is something that *features* are trying to predict. A label is usually referred as *y* value. In this paper, the *label* indicates whether a firm will be bankrupt or not by indication of 0 (healthy) and 1 (bankrupt). (Jung 2018)

Hypothesis space considers all the possible ways to describe the relationship between *features* and *labels*. In other words, a single hypothesis can be anything that gives an outcome based on *x* values. A *hypothesis map* is a function that approximates the true label of *y* from the *features*. Another way to describe the hypothesis map is a map that describes the relationship between x and y. It is a design choice what *hypothesis map* is used. However, computationally efficient *hypothesis map* that can approximate the *label* well from *features* is a desirable choice. (Jung 2018)

The third element in machine learning is the *loss function*. A *Loss function* determines, which predictor map out of hypothesis space should be used. To find a good predictor, penalty should be given from an error. The error is calculated from the difference between true label *y* and predicted label of *ŷ*. The loss can be expressed as a function (Equation 5) *features*, *labels*, and *predictor map* (*h*). A popular *loss function* (*L*) called *squared error loss* is commonly used for example in linear regression.

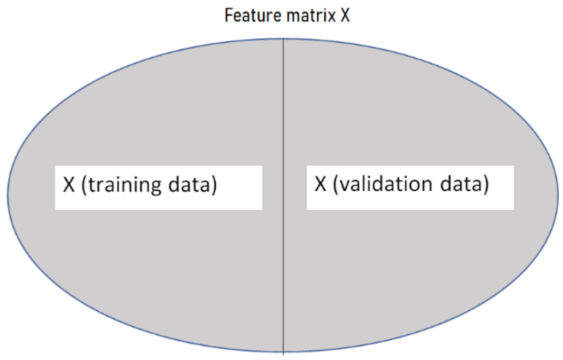
$$L\left((x,y),h\right) = \left(y - h(x)\right)^2 \tag{5}$$

For instance, predicting true label value (*y*) of 10 with predicted value ($\hat{y}$ = h(x)) results in 4 units of penalty. The choice of loss function should be analysed carefully when constructing the machine learning model e.g., *squared error loss* works well in coherent data but is sensitive for outliers which may result in poor performance of the model. (Jung 2018)

4.1.2 Validation of the model

Two important concepts of machine learning called training error and test error (validation error). A training error is the error from the sample. In other words, the model is trained and tested by the same sample data. A small training error might lead to wrong conclusions about the model’s performance since new data can perform differently. By feeding unseen data to the model, the error might increase dramatically. This usually means that the model is overfit, and the model is biased towards the training sample. The complexity of the model is often positively correlated with the probability of overfitting. There are different techniques to overcome overfitting and bias. A popular technique of splitting data into two sets, training data and validation data is utilised in this thesis. (Jung 2018)

Figure 6. Training and Validation data split



The validation data set is used to calculate the validation error (empirical risk) once the model is trained by the training set. This metric is more reliable than training error as it gives indication, how the model performs with new unseen data. Another popular technique to validate a model is to repeat this procedure of random splitting *k* times. This method is called k-fold cross-validation but due to a large dataset, k-fold cross-validation is left out. (Jung 2018)

There are two AUCs calculated in this thesis. First one calculates the AUC of Receiving Operating Characteristic (ROC). The ROC is used a diagnostic for binary classification problems to compare the overall performances of statistical models. The x-axis indicates the *false positive rate* and y-axis the *true positive rate* (*recall*). *False positive rate* indicates how many healthy firms are predicted as bankrupt out of all true negatives. The *true positive rate* states the ratio of bankrupt firms predicted correctly from all true positives (bankrupt). The threshold of a model is changed so that the points are received with varying values of *false positive* and *true positive rates*.

The second graph is called Precision – Recall AUC which uses the *true positive rate* in the x-axis and y-axis shows the precision. Precision describes the ratio of true positives out of *true positives* and *false positives*. This graph is more suitable for imbalanced dataset as the *true negatives* (i.e. true healthy) do not affect the results. Therefore, a careful analysis of the minor class can be made.

4.1.3 Logistic regression

Logistic regression model is simple to use and popular method classifying observations into two categories. Let us assume feature space of X matrix with label space of Y = {-1, 1} and predictor *h* of hypothesis space. Let us say that in this thesis, y = -1 would mean bankrupt and y = 1 non-bankrupt firm. A linear map of h(x) = w^TX gives any number that might be non-equal with the labels {-1, 1}. Logistic regression model can determine the level of confidence of observation belonging into one of the two categories. If a value greater than 0 is given, there is over 50% probability is that the company is healthy. A negative value of h(x) means bankrupt. A value of zero would mean equal probability of between the two classes. The absolute value of |h(x)| indicates the level of confidence when threshold is at 0. The greater |h(x)| is, the greater confidence the model has about the observation. On the other hand, a high confidence of observation being misclassification, gives a lot of penalty for the model. The equation of logistic regression is expressed as follows. (Jung 2018)

$$h^{(w)}(x) = w^T x = \left(\frac{1}{m}\right) \sum_{i=1}^{m} \log\left(1 + e^{\left(-y^{(i)} w^T x^{(i)}\right)}\right) \tag{6}$$

$$\hat{y} = \begin{cases} 1 \; if \; h^{(w)}(x) \; \geq 0 \\ -1 \; otherwise \end{cases} \tag{7}$$

The equation minimizes the empirical risk (the error based on the sample) giving out the optimal weights *w* for the features *x* in X. This is generally done by stochastic gradient descent, which is out of the scope of this thesis. However, the true error (real error of the population) is not obtained as the empirical risk is only based on the sample. Once the model is complete, observations from the sample can be classified by the value of ŷ (see equation 7). (Jung 2018)

4.1.4 Decision trees

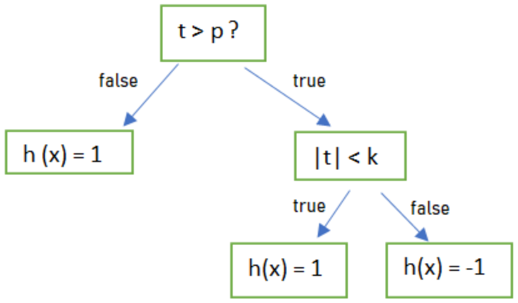
Decision trees have gained popularity by their ability to solve financial problems with scattered data (see Rudd et. al. 2017) but they also benefit from visualization properties (see

Scikit-Learn: Decision Trees 2020, James et. al. 2017). Furthermore, a decision tree is capable of handling categorical and numerical (regression) data, which makes them resilient to use. That said, the prediction accuracy of decision tree methods cannot often compete with some traditional linear and non-linear models (James et. al. 2017). Fortunately, decision trees can be improved by variations of trees. One variation of decision trees is the random forest (RF) which will be used in this thesis. The generalization properties of RF (less variance) comes with a cost of reduced interpretation of the model (James et. al. 2017).

A decision tree consists *decision nodes* and *leaf nodes* (end nodes) which are connected to each other. The starting point of a decision tree is called a *root node* and the end of the tree *leaf node,* where the predicted label ŷ is given. At *decision nodes*, hypotheses of features are tested. A hypothesis with least amount of impurity gets chosen and a path to next node is determined by the true or false outcome. In other words, decision tree is a stepwise algorithm for solving the predicted label ŷ from features *x*. (Jung 2018)

A loss function measures the impurity (separation ability) in each node and is determined by the task we are solving (classification or regression). A common loss function for categorical data is called Gini. Other methods such as entropy and misclassification are options for the loss function in the Python Scikit-Learn package for classification purposes. For numerical data, mean squared error or mean absolute error can be used to calculate minimum impurity at the node. A structure of a simple decision tree is shown in the Figure 7. For the sake of simplicity, binary labels {1, -1} are identical with the example in logistic regression.

Figure 7. A simple decision tree



The top node (t > p) is the root node of the tree and it is the least impure of all hypotheses. The Boolean decision (true/false) determines the path of an observation to a next *decision node* or a *leaf*. Finding an optimal decision tree (hypothesis *h*) requires iteration over certain number of trees as finding optimal decision tree is not a convex problem like in linear regression. A convex problem refers to a global minimum that can be approximated by derivation and stochastic gradient decent methods for example. The decision tree algorithm replaces the leaf nodes by decision nodes as much as possible, trying to achieve the least amount of empirical risk (training error). By doing so, the tree might grow too deep and computational resources are limited for large amount of data. Another problem with expanding the decision tree have to do with overfitting. Large and complicated decision trees can achieve zero training error but generalize to new data poorly. (Jung 2018)

RF is special case of decision tree methods. This technique tackles high variance of the model by bagging the trees and decorrelates them with random selection of predictors. The variance refers to the variation of models created if the data is split and used separately. Bagging (bootstrap aggregation) method is a powerful technique to minimise the variance. When bagging, a model is trained by a subset of the data with replacement (can have one several times same observation) from the original sample. This allows shrinkage of the variance of deep trees (usually high validation error) by averaging them together. By having a one strong predictor in the data, the several trees even with bagging technique, will probably have the same *root node* in all the trees. Homogenous trees are not desirable due to correlation. Since bagging might not be enough to achieve a low variance, RF chooses a subset of predictors *p* randomly in each node. Decorrelated trees are more likely to achieve less variance and important predictor is still not lost in the procedure. (James et. al. 2017)

4.1.5 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is usually referred as multiple discriminant analysis (MDA) in bankruptcy prediction literature due to a use of several predictors in the analysis. The use of LDA and QDA (see 4.1.6) in the analysis is used as an insight of the Z-score. These methods are not related to hypotheses. There are many statistical assumptions on LDA but less with QDA which makes the analysis interesting.

LDA exploits linear combinations of variables. Purpose of LDA is to minimize the variance within groups and maximize the variance between group means. By doing so, the best separation of groups is achieved.

The *Bayes' theorem* is used to form LDA model and is in the following form:

$$\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{8}$$

where,

K = total number of classes
π = overall probability of random observation belonging to a class
$f_k(x)$ = density function of *X* from class *k*

The notation states the probability that observation belongs to class *k* given *x*. The probability of π_k is easily determined by the proportion of class represented in training data. The function of $f_k(x)$ is not easily obtained but estimate of *Gaussian distribution* can be used. The Gaussian distribution (normal) is in the one-dimensional form following:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}^{(-\frac{1}{2\sigma_k^2}(x-\mu_k))^2} \tag{9}$$

where,

μ_k = class mean of *k*
σ_k^2 = class variance of *k*

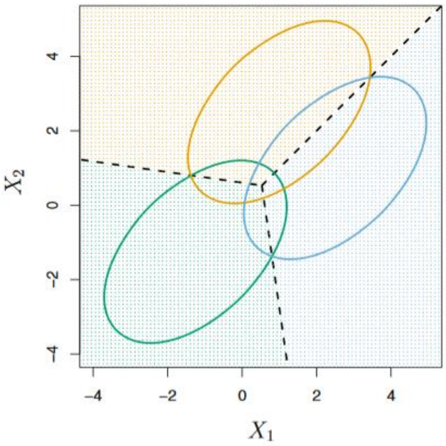
For K > 1, equal variances are assumed. By replacing the $f_k(x)$ in equation 8 by Gaussian distribution (equation 9), taking the log of replaced equation and simplifying, we get:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad (10)$$

The equation is maximized when the distances between the group means are maximised and the variations within the groups are minimised. (James et.al. 2017)

LDA is a useful classification method for classification when the classes are separated well. A stable feature of LDA advocates the utilization instead of logistic regression. Especially, by having a small sample size and normality in all predictors, the LDA can be beneficial. Two important assumptions about the variables X = (X_1, X_2, X_3,…,X_p, where p = amount of predictors) are made. First, the predictors are expected to have a Gaussian normal distribution. Second, the covariance matrix is expected to be the same in all classes. The Figure 8 illustrates the linearity of LDA with three classes and two predictors. (James et.al. 2017)

Figure 8. Example of LDA with p=2 and k=3 (James et.al. 2017)



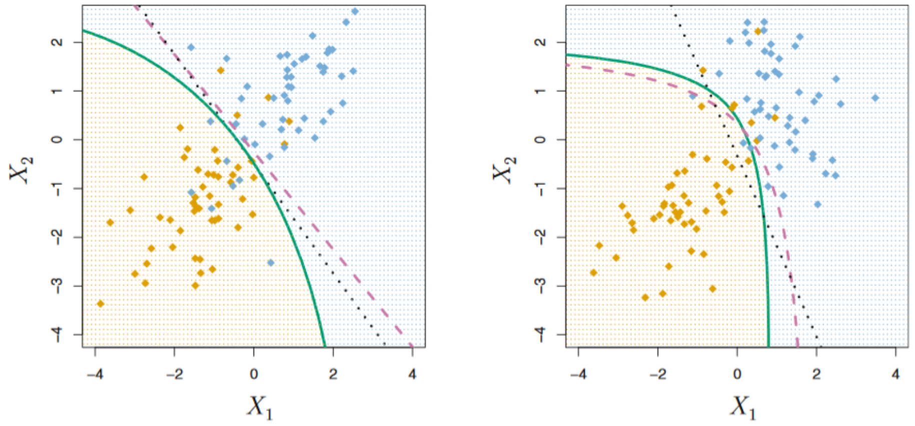
The circles represent 95% probabilities of three different classes relation to two predictors of X_1 and X_2. The dashed lines show the linear decision boundaries of classification. (James et.al. 2017)

For example, in a binary classification problem, LDA penalises the Type 1 and Type 2 error the same amount. In this thesis this would mean same amount of penalty for misclassified in bankrupt and healthy category. Thus, the overall error rate is used to define the coefficients of LDA. If a large class-imbalance between two categories is present, LDA can perform poorly on such dataset. To overcome this issue, the threshold in classifying can be adjusted. In this thesis, it is desired to observe more instances from the bankrupt class. However, this comes with the cost of misclassifying healthy firms. Therefore, it is a matter of research problem which threshold to use. (James et.al. 2017)

4.1.6 Quadratic Discriminant Analysis

The Quadratic Discriminant Analysis (QDA) is considered as an alternative to LDA. QDA assumes that in each class the observations are following the Gaussian distribution. Yet, QDA differs from LDA by assuming that covariance matrices differ between classes. This changes the *x* in equation 10 to be quadratic. While LDA estimates covariance matrix by estimating p(p+1)/2 number of parameters, the QDA estimates significantly more (Kp(p+1)/2) parameters for each covariance matrix. This results in more required computational resources. The decision boundary for QDA is more flexible. However, small datasets may perform better i.e. validation error with LDA due to lower model variance. The flexibility between LDA and QDA is presented by the Bayes decision boundary in the Figure 9. (James et.al. 2017)

Figure 9. The Bayes, LDA & QDA models with equal and non-equal covariance matrices (James et.al. 2017)



A binary classification is presented by two pictures and predictors of X1 and X2. The red and blue dots represent the real classes of observations. The shaded area represents QDA, purple dotted Bayes decision boundary and black dotted LDA model. Left picture shows a situation where the covariance matrices of predictors are equal in all classes. Thus, the Bayes decision boundary is approximated better with LDA since it is linear. However, in right picture, the non-linear Bayes curve indicates of better approximation by using QDA. (James et.al. 2017)

The choice between LDA and QDA depends on the data and use. In general, LDA performs well for relatively small dataset due to low variance. However, large dataset with an explicit violation of equal covariance matrix assumptions, QDA is recommended to use. It is worth mentioning that the interpretability of QDA is hard with non-linearities. Thus, the use of LDA is recommended from causality and interpretability point-of-view.

5 EMPIRICAL DATA

In this chapter, the process of importing data, data cleaning, data manipulation, and creation of variables are introduced.

5.1 Firm specific data

5.1.1 Sample of firms

The firm specific sample data was imported from Orbis database by Bureau Van Dijk. The proportion of private companies in this database is over 99% and therefore requires the use of Z''-score in the analysis (Altman et. al. 2017). This thesis utilises only Finnish data to simplify the impact from macroeconomic events. The proportion of N.A. values in the Finnish data is small and advocates the use of this sample. The Z''-score was developed by using USA data which can affect the performance on Finnish firms, but the re-estimation of LDA of the variables should reflect the new data. Over-sampling can be created from class-imbalance (Balcaen & Ooghe 2006). However, due to small amount of difference in sample (2.7% bankrupt) and population (0.7% bankrupt), the results are expected to vary only a little or at all. By reducing the number of bankrupt firms, vital information about bankruptcies could be lost. The financial ratios of X1, X2 and X3 vary more in bankruptcy group than in healthy (see 6.1, Altman et.al. 2017). Therefore, all observations that are not considered as outliers, should be included.

5.1.2 Status of failed and healthy firms

Orbis classifies the active firms into five categories (*active, rescue plan, default of payment, insolvency proceedings, reorganization,* and *dormant*). Out of these, only *active* indicates a healthy firm. *Default of payment* is not the same as insolvency and bankruptcy. Insolvency refers to a situation where the debtor is unable to pay the debt whereas bankrupt is legal proceeding by court supervision. Default of payment might arise from several reasons which are not related to financial distresses. Thus, this category is not included in the study. In a *rescue plan*, the company is active, has paid the credits, but is in protection by initiative of the debtor. This is a precautionary step against financial difficulties, and usually there is a third-party supervisor to govern negotiations. *Rescue plan* is a serious situation for the continuation of business and is included in the bankruptcy category. *Insolvency proceedings* cover firms that were unable to pay credits but remain active. In this severe condition, firm attempts to regain the normal operations by paying debts under protection of law. Firms with i*nsolvency proceeding* status are included in the bankruptcy category. *Dormant* status refers to a registered but non-operating firm. One reason for *dormant* status is holding a name for the future purposes (excluded). *Reorganization* of business indicates reorganization, restructuring etc. However, this category does not include any financial distresses and is excluded from the study. (Orbis 2020)

Bankruptcy category has eight categories: *in liquidation, bankruptcy, dissolved (merger or take-over), dissolved (demerger), dissolved (liquidation), dissolved (bankruptcy), dissolved* and, *inactive (no precision)*. *Bankruptcy* is a legal proceeding to repay liabilities to creditors and is included in the bankruptcy category. Company is unable to continue and will be non-existent. *In liquidation* refers to a situation where the assets of the company are being sold. Selling assets can be voluntary which makes the causality from financial distress unclear. Thus, *in liquidation* category is excluded from the study. *Dissolved* refers to non-existence of a company for unknown reasons. Naturally, dissolved is not included in the study. *Dissolved (merger or take-over)* means merger or take-over which does not indicate explicitly financial distress. Thus, this category is excluded from the study. *Dissolved (demerger)* indicates that firm is no longer legal entity due to a division (excluded). *Dissolved (bankruptcy)* could mean that a firm is dissolved the end of bankruptcy process or company is stated bankrupt in insolvency or liquidation proceeding. This status indicates financial distress and therefore is included in the bankruptcy category. *Dissolved (liquidation)* category refers to “friendly” liquidation of assets and shows no indisputable financial distress (excluded). *Inactive (no precision)* means non-active and the reason for inactivity remains unclear (excluded). *Unknown* status is literally unknown and cannot be included in the study. (Orbis 2020)

5.1.3 Deriving firm specific variables

Different accounting standards across countries can make the data non-robust between countries. Fortunately, Orbis provides comparison table (Table 2) between variables and Finnish accounting terms.

Table 2. Correspondence Table of Variables (Orbis 2020)

English	**Finnish**
Shareholders’ funds	Oma pääoma + Tilinpäätössiirtojen kertymä
Other shareholders’ funds	Ylikurssirahasto + Arvonkorotusrahasto + Käyvän arvon rahasto + Muut rahastot + Tilinpäätössiirtojen kertymä
Capital	Osake-,osuus tai muu vastaava pääoma
Working Capital	*Calculated (derived)*
Operating P/L (EBIT)	Liiketoiminnan tulos
P/L for period (net income)	Tilikauden Tulos - Vähemmistöosuus
Total Shareholder's Funds and Liabilities	Taseen loppusumma
Currrent Liabilities	Lyhyt vieras pääoma
Non-Current Liabilities	Pitkäaikainen vieras pääoma + Pakolliset varaukset
Intangible Fixed Assets	Aineettomat hyödykkeet

As a reminder, the financial ratios were calculated as follows:

X1: Working Capital / Total Assets
X2: Retained Earning / Total Assets
X3: EBIT / Total Assets
X4: Book Value / Total Liabilities

Working Capital is calculated by Orbis database itself and needs no modification. *Total Assets* represents the total of balance sheet value and the corresponding variable in Orbis database is named as *Total Shareholders’ Funds and Liabilities*. However, *Retained Earnings* need to be derived by subtracting *Capital, Other Shareholders’ Funds* and *P/L for the Period (net income)* from *Total Assets* (*Total Shareholders’ Funds and Liabilities*). The *Total*

Liabilities is calculated by adding *Current Liabilities* and *Non-Current Liabilities* together. However, it is noteworthy that *Non-Current Liabilities* includes provisions. The *Book Value* is calculated by subtracting *Intangible Assets* and *Total Liabilities* from *Total Assets*.

5.1.4 Data cleaning (firm specific)

Altman et. al. (2017) removed firms (total assets less than $100 000 during observation period) from the sample due to unstable financial ratios in small firms. However, the size of the firm does not define the importance of detecting bankruptcy to stakeholders in this thesis. In other words, predictive models should identify probability of bankruptcy in all firms. Therefore, small firms are included in this study even if they are not coherent with bigger firms. Additionally, by restricting the scope of data, the model might get biased if applied with all sized firms. Firms with duplicate names were excluded. Revenue is not used as a predictor but observations having negative revenue were removed. Total assets with zero value are removed, indicating an error or "nonexistence" of a firm. Total liabilities with negative values are removed. Zero values (no liabilities at all) are changed to 1 EUR to make division possible in calculation of X4. Similar technique is also used by Filipe et. al. (2014). Clearing outliers by utilising *winsorizing* technique was not applied as in Altman et. al. (2017). Without careful analysis of the extreme values of predictors, no general assumption about outliers should not be made. The outliers should be analysed individually (or by some function). Thus, data is not expected to have outliers in the analysis.

5.2 Macroeconomic data

Gross Domestic Product (GDP), *household debt % of available income* and *interest expenses % of available income* are selected as predictive macro-economic variables in this thesis. All firm specific variables from financial statements are expected to available at the latest 6 months after financial period. Usually, companies report financial period from January to December. The levels of GDP change, household debt and household interest expenses are also expected to be available before 6 months after. Thus, bankruptcy predictions can be made latest 6 months after financial period.

5.2.1 Gross Domestic Product (GDP)

The GDP is selected as first macroeconomic variable based on previous studies such as Filipe et.al. (2014) and Laamanen (2017). In the Laamanen (2017) study, GDP contributed the most between the macroeconomic prediction models. In this thesis, Finnish GDP data is imported from Eurostat. Observations are chain linked yearly volumes, meaning inflation adjustment. This allows to distinguish the real growth of GDP. Figures are indexed by year 2010, and the yearly change is calculated as a percentage from previous year. Positive growth of GDP is expected to lower the probability of bankruptcy and vice versa.

5.2.2 Household debt and interest expenses of available income

Household debt % and *interest expenses %* (of available income) data measures the liability risk that households face. This data is gathered by the Bank of Finland during 1999-2019. By having a high *household debt %* and *interest expense %* the financial stability of a household deteriorates. Disruptions and unexpected situations in the economy can lead to default of payments due to layoffs, when debt-to-income ratio increases by lower available income. Naturally, this would lead to more defaults of loans and less private spending (decrease in demand). This negative cycle has a big impact on economy. Variability in *household debt %* and *interest expenses %* could indicate the state of business cycle which is a major cause of bankruptcies in Finland (Laitinen & Laitinen 2004). Additionally, this indicator might reflect the ease of financing. On the other hand, positive views of the future might enhance lending. Low interest rates and low financial requirements for debt are expected to result in higher rates of lending. These two ratios resemble single firm’s solidity measures but from a household point-of-view. These two variables are calculated as percentage change from previous year. Finally, despite the contradictory views of the effects of ratios, higher values of these ratios are expected to increase the probability of bankruptcy.

6 EMPIRICAL RESULTS

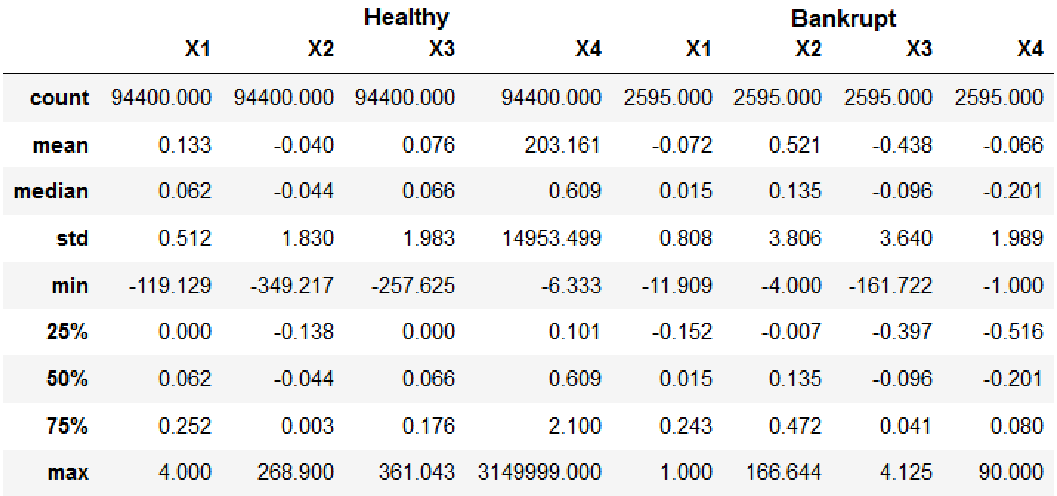
In this chapter, four different setups of predictors are constructed. During each setup, LR, RF, LDA and QDA models are utilised, except in the first setup the use of Z''-score is also used. The first setup is purely based on firm specific predictors and three last ones utilise one macroeconomic predictor along with firm-specific ones. The analysis of data begins with descriptive statistics of firm specific variables, continuing to macroeconomic variables. Thereafter, the building processes of predictive models and their performances are presented individually. Lastly, the results are summarized.

6.1 Descriptive statistics

The descriptive statistics for healthy and bankrupt firms are presented in the

Table 3. The number of observations (count), mean, median, standard deviation (std) and range with quantiles are specified groupwise. There are 94 400 (97.3%) healthy and 2595 (2.7%) bankrupt firms in the entire sample dataset. All financial ratios are positive in healthy firms except X2 is negative. For bankrupt firms this is opposite of X2 being positive. This finding is contradictory opposite from Altman et.al. (2017). X1 and X4 are negatively skewed in healthy group with due to higher means than medians due to higher values of the means.

Table 3. Descriptive statistics of firm specific variables



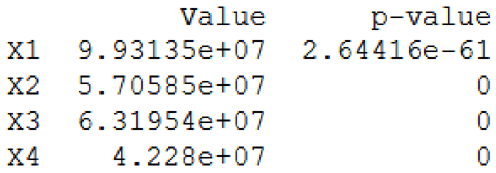
	Healthy				Bankrupt			
	X1	X2	X3	X4	X1	X2	X3	X4
count	94400.000	94400.000	94400.000	94400.000	2595.000	2595.000	2595.000	2595.000
mean	0.133	-0.040	0.076	203.161	-0.072	0.521	-0.438	-0.066
median	0.062	-0.044	0.066	0.609	0.015	0.135	-0.096	-0.201
std	0.512	1.830	1.983	14953.499	0.808	3.806	3.640	1.989
min	-119.129	-349.217	-257.625	-6.333	-11.909	-4.000	-161.722	-1.000
25%	0.000	-0.138	0.000	0.101	-0.152	-0.007	-0.397	-0.516
50%	0.062	-0.044	0.066	0.609	0.015	0.135	-0.096	-0.201
75%	0.252	0.003	0.176	2.100	0.243	0.472	0.041	0.080
max	4.000	268.900	361.043	3149999.000	1.000	166.644	4.125	90.000

X2 in the bankrupt category is also negatively skewed, whereas X3 indicates positive skewness. The range of minimum and maximum values in healthy firms is much larger than in bankrupt which could be an indication of outliers, especially in X4 of healthy group. The standard deviation in X1, X2 and X3 is higher for bankrupt firms except for X4. Balcaen & Ooghe (2006) stated that small firms have tendency to fail more frequently than larger ones. This could explain the variability in ratios.

6.1.1 Equality of medians (Mann Whitney U-test)

The four firm financial ratios of Z''-score are tested for equality of the medians between bankrupt and healthy (active) firms. The null hypothesis (H_0) of median test indicates identical medians and H_1 for unequal medians. The confidence level is chosen at 0.01.

Table 4. Mann Whitney U-test



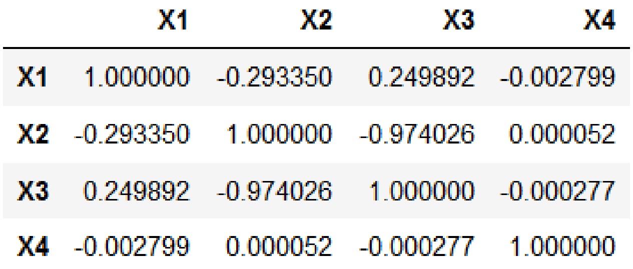
```
          Value      p-value
X1  9.93135e+07  2.64416e-61
X2  5.70585e+07            0
X3  6.31954e+07            0
X4    4.228e+07            0
```

The H_0 will be rejected and H_1 accepted with low p-values which can result from a large dataset. The medians between bankrupt and healthy firms are different for all four firm specific variables.

6.1.2 Correlation of firm specific variables

The correlation between the firm specific variables are shown in the Table 5. There is a strong negative correlation (-0.97) between variables X2 and X3 which can lead to multicollinearity. Between other variable pairs, no notable correlation exists. The second highest correlation between X1 and X2 is only -0.29. The "danger zone" for predictor correlation in regression problems is around 0.6 absolute value. The purpose of this thesis supports the use of all variables despite the possible multicollinearity. Multicollinearity can create problems regarding interpreting the coefficients but does not necessarily make the model perform poorly. Thus, the number of variables is not reduced. Especially, RF and QDA might perform well with correlated variables with non-linearities in the data.

Table 5. Correlation matrix: Firm specific variables



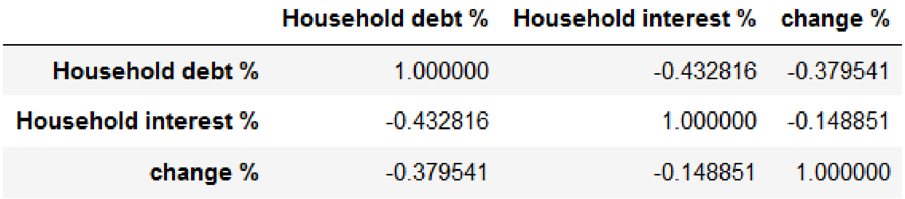
	X1	X2	X3	X4
X1	1.000000	-0.293350	0.249892	-0.002799
X2	-0.293350	1.000000	-0.974026	0.000052
X3	0.249892	-0.974026	1.000000	-0.000277
X4	-0.002799	0.000052	-0.000277	1.000000

Additionally, the correlation describes only linear dependence. Correlation does not explain exponential dependence or causality without further analysis.

6.1.3 Correlation of macroeconomic predictors

A similar correlation matrix (Table 6) is constructed based on macroeconomic predictors. There is no significant correlation found between any of the macroeconomic predictors. The highest absolute value is only 0.43 which is well below the limit of 0.6. This indicates that the macroeconomic variables explain different things and each predictor might give valuable information about the systematic risks that firms face. Due to the low level of correlation, predictors could be used in the same setup to get better results. However, for the simplicity, only single macroeconomic variable is used in each setup.

Table 6. Correlation of macroeconomic predictors



	Household debt %	Household interest %	change %
Household debt %	1.000000	-0.432816	-0.379541
Household interest %	-0.432816	1.000000	-0.148851
change %	-0.379541	-0.148851	1.000000

6.2 Prediction models

The four different setups are constructed by the following way (Table 7). The first setup uses firm specific predictors solely. Rest of the setups utilise only one additional macroeconomic variable at a time. Note that the analysis of Z’’-score is only related to the *setup 1*.

Table 7. Predictors of four setups

Setup	Firm-specific	Macroeconomic	Models
1	*X1, X2, X3, X4*	-	*LR, RF, Z''-score, LDA, QDA*
2	*X1, X2, X3, X4*	*GDP change %*	*LR, RF, LDA, QDA*
3	*X1, X2, X3, X4*	*Household debt / avail. income*	*LR, RF, LDA, QDA*
4	*X1, X2, X3, X4*	*Household interest exp. / avail. income*	*LR, RF, LDA, QDA*

The building process of the four models starts with testing the firm specific predictors. The first (H_1) hypothesis is tested by comparing the performances of logistic regression and random forest against Z’’-score. Thereafter, macroeconomic variables are added to the model to improve the model (H_2). After all models (2,3 and 4) have been created, the third hypothesis (H_3) of superiority of RF is evaluated. Each time predictors are changed; the evaluation is referred as a “setup” instead of “model” due to several models in one setup.

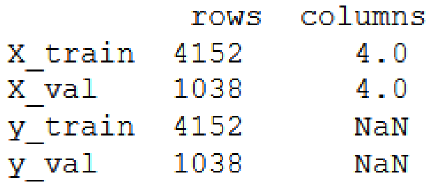
In each setup, the models are built by the same procedure (see appendix). To validate each model, the data is split into training and validation (test) sets. The size of the random validation set is 20% (n=19 399) of all data (n=96 995). The ratio of bankruptcy firms (2.68%) and healthy (97.32%) remains the same training and validation sets.

6.2.1 Micro: firm specific predictors (*setup 1*)

In this first setup, only four Z’’-score predictors of X1, X2, X3 and X4 are used. The sizes of training and validation sets (.shape attribute in Python on DataFrame) are shown in the

Table 8. Note, the NaN (Not a Number) values inherit from zero dimensionality (one column), not from empty nor NaN values in the data.

Table 8. Sizes of training and validation sets (*setup 1*)



```
          rows  columns
X_train   4152      4.0
X_val     1038      4.0
y_train   4152      NaN
y_val     1038      NaN
```

First, the logistic regression object is created by using Python scikit-learn package. Intercept is allowed for more flexible fit. The model is trained with training data `X_train` matrix and `y_train` vector. The prediction vector `logreg_pred` contains the predictions of the model by using the validation dataset `X_val`.

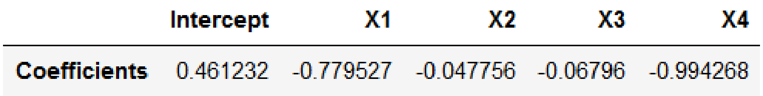
Second method, Random Forest object (`r_forest`) is created by using the same scikit-learn package and the hyperparameter `n_estimators` is set at 100 trees. Increasing the number of trees is better concerning overfitting and generalizing but comes with the cost of computational resources. In this thesis, an increase in the number of trees over 100 does not improve the validation error. Random forest model is trained and predicted with identical dataset as in logistic regression.

Third model of Z''-score is calculated by the formula (4). Z''-score has three different categories for firms. In this binary classification study, only healthy firms are labelled as healthy (0), leaving out gray zone and distress firms with a label (1). Thus, the threshold is set at 2.6. This must be done since other models do not have an opportunity to leave out (i.e. gray area in Z-score models) uncertain observations.

The re-estimation of Z''-score is done by linear discriminant analysis (`lda`) by scikit-learn package. The same procedure was followed for training and prediction of the model. Linear discriminant analysis requires same distributions in all predictors. However, by using quadratic discriminant analysis (qda) this problem could be tackled. So as a fifth method, quadratic discriminant analysis was carried out.

First, a quick look at the coefficients of logistic regression is done. Previous studies tend to look at the coefficients of logistic regression models. Coefficients describe the effect of predictors to the dependent variable. The coefficients of the logistic regression are shown in the Table 9. Increase in each of variables decrease the risk of getting bankrupt in the future as the label for bankruptcy is 1 and healthy 0. However, one should be careful with this statement because the correlation between X2 and X3 is very high. Multicollinearity can change the signs and the magnitude of the coefficients. The significance test of each coefficient was not tested. Calculating p-values remains unnecessary from predictive point-of-view.

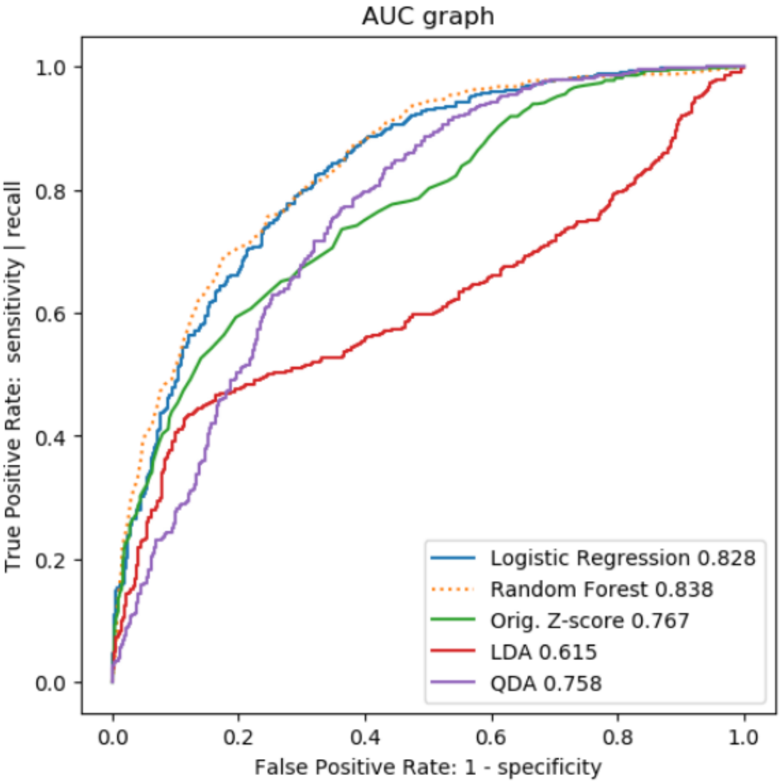
Table 9. Logistic regression coefficients



	Intercept	X1	X2	X3	X4
Coefficients	0.461232	-0.779527	-0.047756	-0.06796	-0.994268

The absolute performance of each model was analysed by using the ROC AUC which indicates the overall classification property of a predictive model. The ROC line of Z''-score was obtained by changing the Z-value threshold (see appendix) which is the value that classifies observations into bankrupt and healthy firms. The ROC AUCs of five models are shown in the Figure 10.

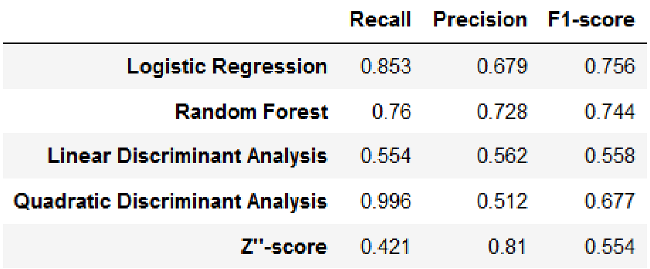
Figure 10. ROC AUC graph (*setup 1*)



The RF model performs the best (0.84) with four variables. The second-best model is the LR model (0.83). The Z''-score model performs poorly compared to previously mentioned models but outperforms the LDA. Thus, hypothesis H_1 is accepted. The additional models QDA model (0.76) outperforms the LDA Model (0.62), suggesting un-equal distributions and unequal covariances of predictors, which is a prerequisite for LDA. ROC AUC of 0.5 refers to model which is not capable of separating classes. However, the purpose of a prediction model determines the real-life usability. One could argue that, discovering almost all bankrupt firms is essential, even with a cost of misclassification of healthy firms. This changes the performance metrics, leaving Random Forest behind with four variables used. The upper right corner demonstrates this kind of preference where LR and QDA perform better than RF. In fact, the Z''-score outperforms Random Forest at the very end of the right upper corner (high sensitivity). Regardless of the minor benefit of Z''-score, the overall metric for model performance is used as ROC AUC.

To get a more in-depth analysis of the classification, a classification table is constructed and shown in Table 10. The thresholds of the models are kept at the default which may weaken the comparison. However, this results in the best overall performance regarding the loss function. The metrics are selected due to imbalanced dataset. ROC AUC graph is a good metric but lacks detailed information in the high sensitivity area. *Recall* defines the ability of the classifier to find positive (bankrupt) observations. The number of *true positives* is divided by the sum of *true positives* and *false negatives* to compute *recall*. Therefore, this metric indicates how well the bankruptcies are spotted from the data. The range of the ratio varies between 0 and 1, where 1 being the best value. QDA spots almost (99.6%) all the bankruptcies from the data whereas LR can spot 85%. These levels are considerably good but the cost of high performance in *recall* weakens the *precision*. RF performs poorly at this threshold level as only 76% of bankrupt firms are found. *Precision* metric is the ability of a model to not label healthy firms as bankrupt.

Table 10. Classification metrics (*setup 1*)



	Recall	Precision	F1-score
Logistic Regression	0.853	0.679	0.756
Random Forest	0.76	0.728	0.744
Linear Discriminant Analysis	0.554	0.562	0.558
Quadratic Discriminant Analysis	0.996	0.512	0.677
Z''-score	0.421	0.81	0.554

Best value of *precision* is at 1 and worst at 0. The ratio is calculated by the same way as in *recall*, but the *false negatives* are replaced by *false positives*. Surprisingly, the Z’’-score did best not label healthy firms incorrectly. However, a low recall shows no practical use of Z’’-score. The second-best precision comes from RF and performed relatively well in both recall and precision metrics. *F1-score* is a popular metric in classification which takes both, recall and precision into account (see equation 11).

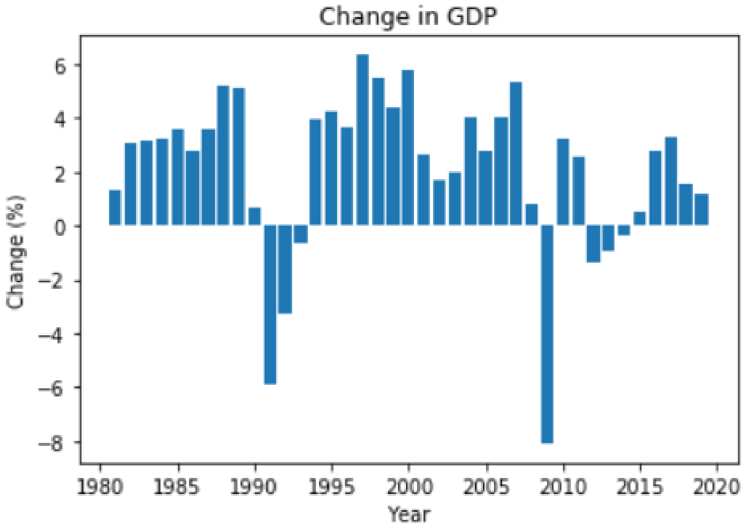
$$F1 = \frac{2*(precision*recall)}{(precision+recall)} \qquad (11)$$

Therefore, *F1-score* can spot the performance on class-imbalanced dataset. LR (0.76) and RF (0.74) have almost same F1-scores and perform the best out of the five models. QDA performs well compared to LDA. Again, this suggests violations against LDA statistical assumptions.

6.2.2 Macro: GDP change % (*setup 2*)

In the *setup 2*, the first macroeconomic variable included describes the yearly change in GDP (`GDP change %`) from previous year. The data of new predictor `GDP change %` is shown in the Figure 11.

Figure 11. Yearly GDP change in Finland (%)

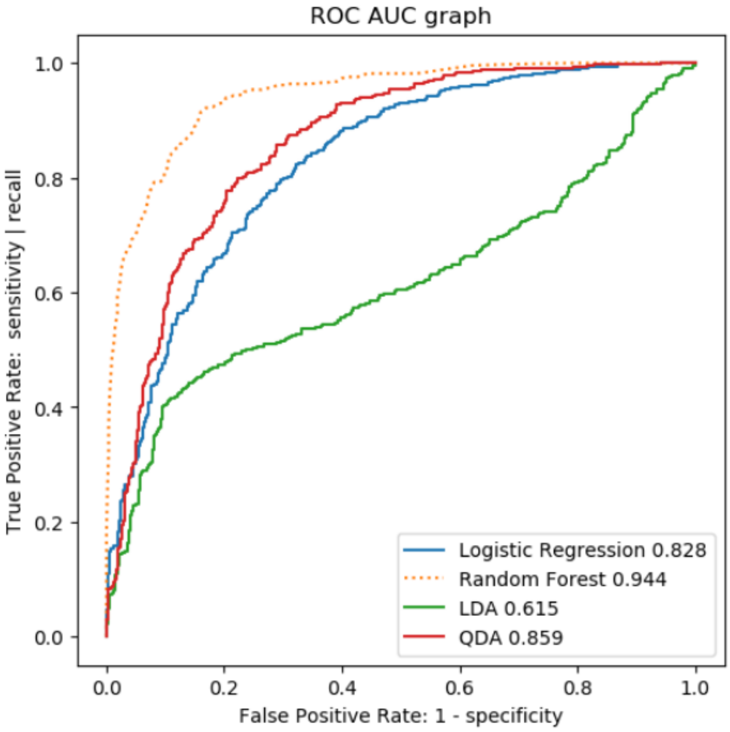


There have been two major events in the last 30 years, 1990s depression and 2008 financial crisis. During and after these events the yearly change in GDP declined dramatically, -8% being the worst reading in 2009. Additionally, the GDP growth has been negative between the years 2012 and 2014. If the change in GDP is compared with the absolute number of bankruptcies (Figure 1), a negative correlation can be observed. Negative correlation could have explanatory power to the probability of bankruptcy. Additionally, the "tough" years seem to last approximately the same period. This is not evident from Figure

11 as it measures the growth compared to last year. Thus, the growth 2010 and 2011 gives too optimistic indications from absolute value of GDP. The absolute value of GDP after year 2011 is in fact over 2% less than before the drop during 2009. Note that the timespan of firm dataset starts from the year 1999 and therefore 1990s depression cannot be tested.

All the models were constructed with the similar process as in firm specific model but Z''-score was left out due unfair comparison. Z''-score does not allow the utilisation of macroeconomic predictor. The results with `GDP change %` added are shown in the Figure 12.

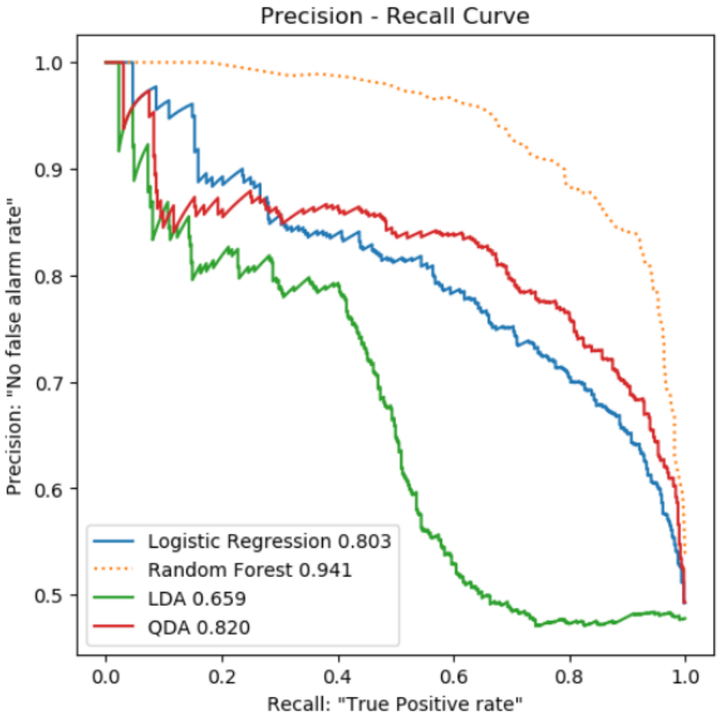
Figure 12. ROC AUC graph (*setup 2*)



The best performance (0.94) is achieved by RF model which is exceptionally good. The QDA performs the second best (0.86) and improves from the *setup 1*. LR underperforms compared to the Random Forest and QDA. As a matter of fact, the performance of LR is the same with a new predictor. This finding the opposite with the findings from Laamanen (2017). Improvement of ROC AUC was achieved by using Finnish restaurant and accommodation data with a change in GDP. Non-linear properties of RF and QDA suggest that the data cannot be explained linearly as well as by these two metrics. Despite the same ROC AUC in LR, the second hypothesis (H_2) can already be accepted on behalf Random Forest and QDA.

Another metric is used with different thresholds of four models. A *precision-recall curve* is an illustrative metric with imbalanced data where observing the positive class is in key interest. In bankruptcy prediction, this indeed is the case. By increasing the amount correctly classified negative class (healthy), does not affect the analysis. Now, the *recall* is illustrated by the x-axis and the *precision* by the y-axis. The precision describes the cost of false alarms that stakeholders face when a model predicts a healthy firm as a bankrupt. The interpretability of this graph is improved to ROC AUC as the precision considers only *true positive* and *false positives* in the equation. ROC AUC considered *true negatives* in the denominator which is a major class in this analysis. A perfect model in *precision-recall curve* would lie in the upper right corner. In this area, all bankrupt companies are spotted while “no false alarm” rate is 100% (healthy predicted as bankrupt). As in ROC AUC graphs, the area under Precision – Recall curve can be calculated as well. These values are shown in the legend of Figure 13.

Figure 13. Precision - Recall AUC graph (*setup 2*)

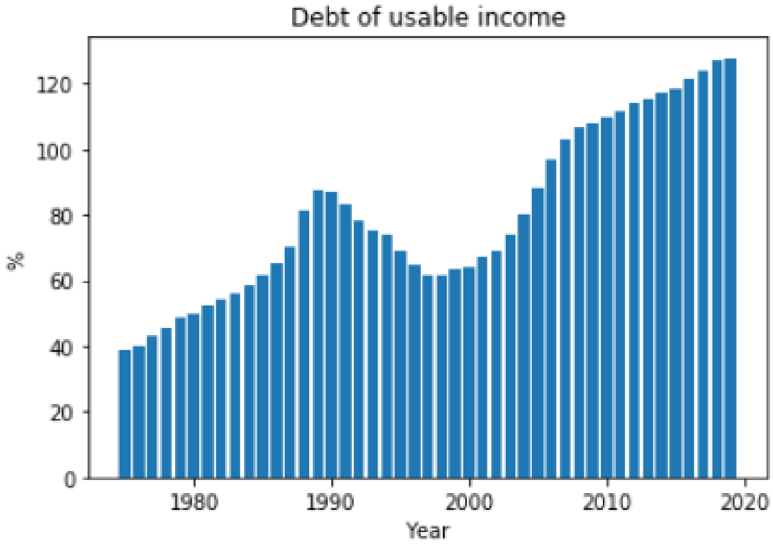


The best performance is achieved by RF (0.94) and shows a smooth approach towards ideal area of upper right corner. The false alarm rate remains relatively low at all thresholds. The second-best performance is achieved by QDA (0.82) and shows variability in the *precision* when *recall* is low. This means that the false alarms are rapidly increased when the threshold is increased at low *recall* levels. This unwanted behaviour applies in all four models except the RF. However, *recall* > 0.5 (over half bankrupt firms spotted) major oscillations are not observed by LR and QDA. LR and especially LDA perform poorly compared to the non-linear models of QDA and RF. This suggests that these LR and LDA cannot capture non-linear relationships between labels and predictors.

6.2.3 Macro: Household debt % (*setup 3*)

In the *setup 3*, the second macroeconomic predictor `Household debt %` is used in the models. `Household debt %` indicates, not only the amount of debt but also the changes in available income (equation 12). The level of debt is shown in the Figure 14.

Figure 14. Household debt as a proportion of usable income (Bank of Finland)

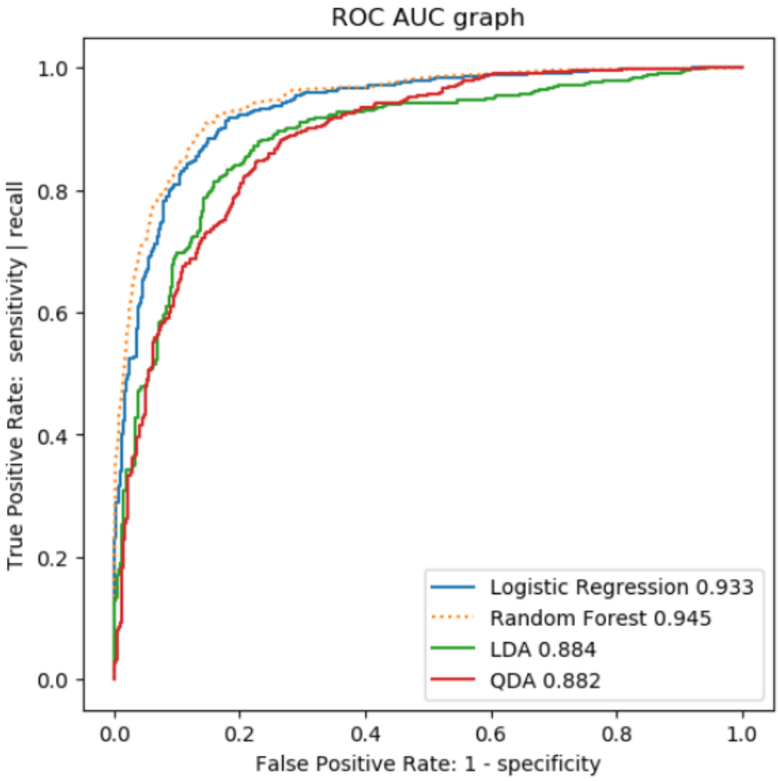


$$Household\ debt\ \% = \frac{Household\ debt}{Available\ income\ of\ household} \tag{12}$$

This illustration indicates similar negative effects in economy during 1990s and around 2008-2009 as in `GDP change %` predictor did. However, the level of household debt continued to increase after 2008 crisis what can be a result from lower interest rates for debt or lower income (layoffs etc.).

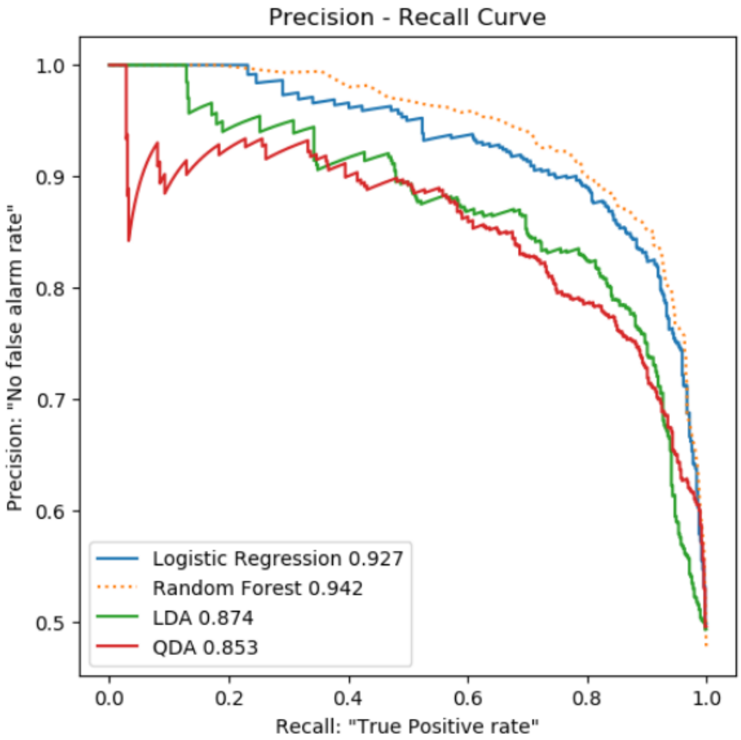
By using the `Household debt %` as a predictor along firm specific variables shows a significant improvement in all models. This suggests that this macroeconomic predictor affects the probability of bankruptcy. Households are not corporates, but their vulnerability and spending are related to companies' revenue. The performance of RF (0.95) is slightly improved, being still the best model in all three setups. The AUC of LR experienced an increase of 0.11 which is considered a major increase. The improvement of ROC AUC of LR is now along with the previous findings with Laamanen (2017) from the macro predictor point-of-view. As for LDA and QDA, the model performances were improved by 0.27 and 0.02 of AUC, respectively. Surprisingly, the LDA outperformed the QDA while the performance of QDA did not change remarkably. Good performance of LDA might indicate normality of macro predictor. These findings strongly support second hypothesis (H_2) of the benefit of utilising macroeconomic predictor in bankruptcy prediction. Even if the overall performance (ROC AUC) of QDA is worse than in LDA, QDA outperforms LDA in the higher sensitivity area (upper right corner).

Figure 15. ROC AUC graph (*setup 3*)



The Precision – Recall graph (Figure 16) illustrates the performances in the bankruptcy class. RF again performs the best (0.94) with a smooth line along the graph but LR does a lot better in prediction than in *setup 2* (0.80 vs. 0.93). The performance of QDA and LDA shows disturbance in the low *recall* levels. LR and RF seem to perform evenly throughout different threshold levels of the models. Therefore, the performances of these models are almost identical based on the Precision - Recall curve AUC and ROC AUC.

Figure 16. Precision - Recall AUC graph (*setup 3*)

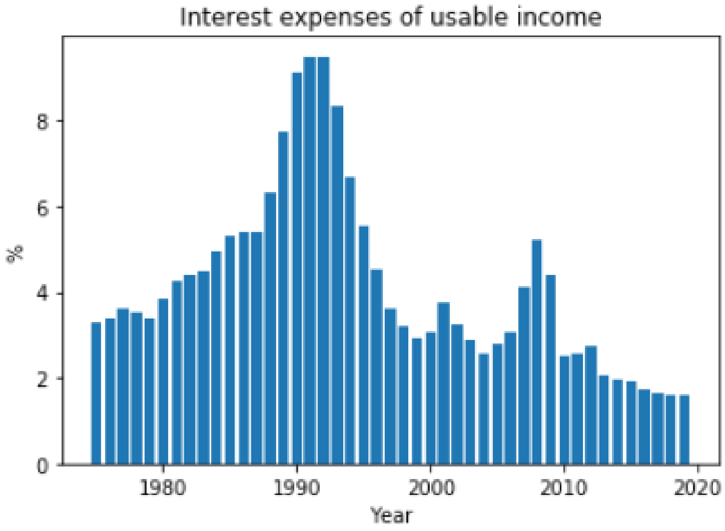


6.2.4 Macro: Household interest % (*setup 4*)

The `household interest %` of usable income indicates the interest expense load on households. This measure is essential as it reflects the interest rate levels in addition with the amount of debt households maintain. Otherwise, the denominator is identical with `household interest %`, reflecting the same available income. The correlation of these measures is relatively low (see Table 6) which was unexpected as the equations (12 and 13) are quite similar.

$$Household\ interest\ expense\ \% = \frac{Interest\ expenses}{Available\ income\ of\ household} \tag{13}$$

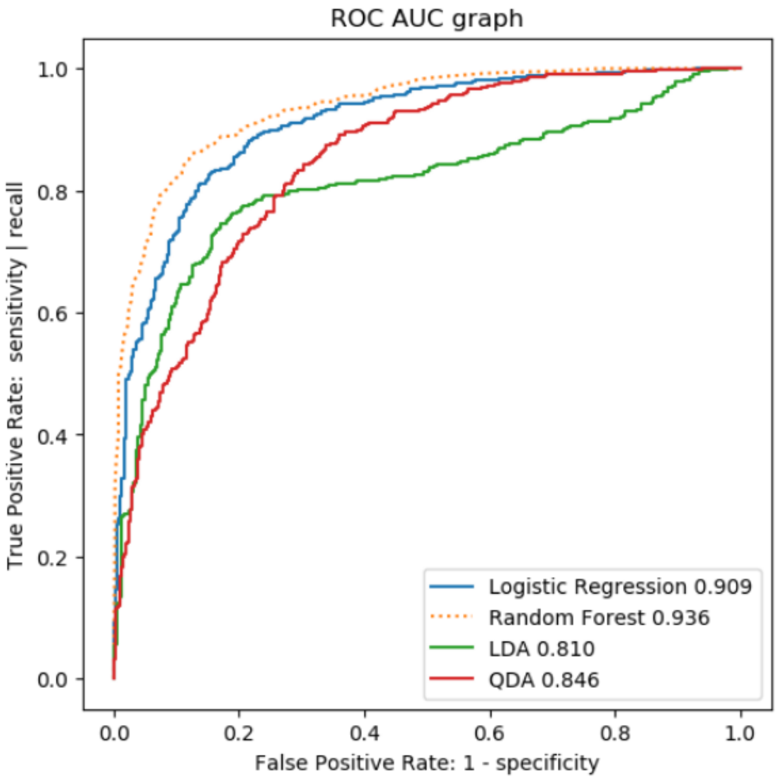
Figure 17. Household interest expenses of available income (Bank of Finland)



The correlation of depressions (1990s and 2008) and household interest expenses is evident from Figure 17. All macroeconomic predictors were able to detect these times of distresses in the economy, `household interest %` being no exception. However, the interest expenses have steadily declined after 2008 depression. This could result from correlation of lower interest rates since the amount of debt has grown during the same period (Figure 14).

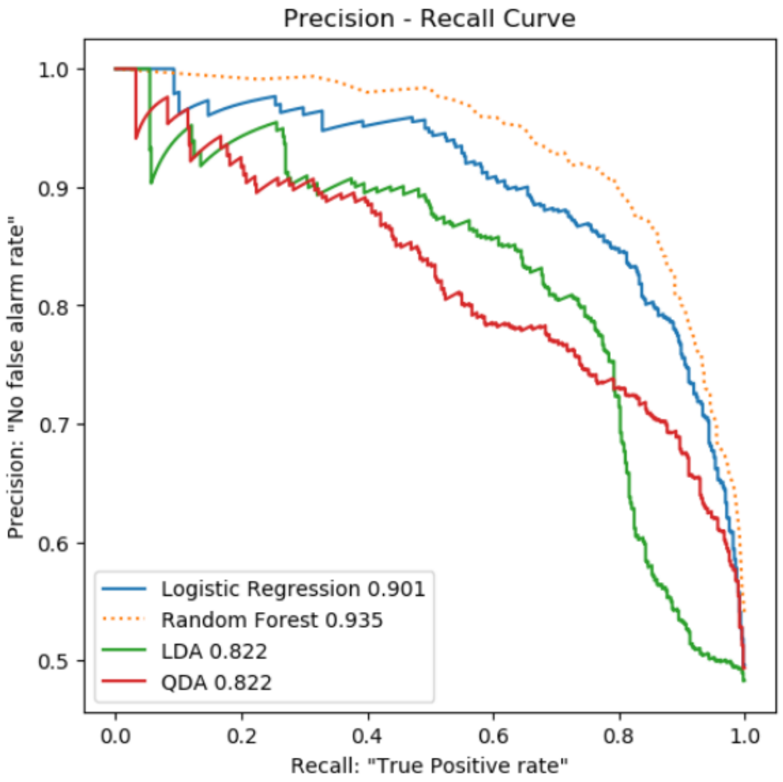
The overall performances of the models suffered slightly from *setup 3* but still remained at relatively high level compared to *setup 1* (Figure 18). Once again, RF outperformed all other models. LR performed well (0.94), especially compared to LDA and QDA. Still, in the higher sensitivity area, RF, LR and QDA perform almost equally well. For predictive real-life purposes using high sensitivity, only LDA would be left out based on this analysis. With a low *recall* LDA performs equally with all models but as the *recall* increases, the model starts to label healthy firms as bankrupt at a very high rate.

Figure 18. ROC AUC graph (*setup 4*)



The findings of precision – recall curve of the *setup 4* (Figure 19) show similar results regarding RF and LR in *setup 3* (Figure 16). The AUC of precision – recall curves are smaller in all four models, indicating that `Household debt %` is a better predictor compared to the interest expense level. However, the performances of LR and RF are still good and show a smooth curve unlike LDA and QDA. The variability of the curve (LDA and QDA) in low recall area is smaller and seem to perform better at that area. However, the key interest is the *recall* area > 0.8 where these models cannot compete with LR and RF.

Figure 19. Precision – Recall AUC graph (*setup 4*)



6.2.5 Summary of results

The Z’’-score did not perform well in this Finnish dataset with ROC AUC of 0.77. Therefore, the generalization of Z’’-score to new un-American data does not hold which lead to acceptance of H_1. These findings were different from Altman et.al. (2017) where the use of Z’’-score performed well on international data. The ROC AUC of Z’’-score in the study reached a value of 0.86 and LR reached the same results with international training data. Altman et.al. (2017) researched the predictive power of additional variables of year, size, age, country risk (SP country ranking) and industry. By adding all these variables into LR model, international ROC AUC was improved from 0.75 to 0.77. The best ROC AUC value of 0.89 from Finland was achieved by using all variables and country-specific training data. As a comparison, the LR model in this thesis exceeded AUC of 0.9 only by adding one macroeconomic variable of household debt (*setup 3*) or interest expense on available income (*setup 4*).

There were benefits by adding a macroeconomic predictor. All AUC metrics were improved when either GDP, household debt level or household interest expense data was fed into the model. Thus, the H_2 was accepted with confidence. Filipe et.al. (2016) showed that a regional model performed better compared to a generic model where the systematic risks are considered equal to all countries. This thesis does not answer the systematic differences between regions but benefit of utilising information outside from a firm is significant.

Muchlinski et.al. (2015) encouraged to investigate the performance of RF with LR as the dataset imbalance is similar with bankruptcy prediction. RF performed the best considering *ROC AUC* and *Precision – Recall AUC* in all setups which supports the hypothesis (H_3). LR did not perform poorly but the differences with RF in AUCs are significant, especially in *setup 2*. Rare event of a bankruptcy and non-linearities seem to support the use of RF instead of LR. However, to study the causality between predictors and labels, the LR is suggested to use. Coefficients and p-values of LR indicate better the causalities between predictors and labels. On the other hand, for practical use, RF is suggested over LR due to undisputable performance benefits.

Lastly, a feature standardising was used to test the performance of the models. The models performed worse than without the standardising which suggests that the variance of predictors is broad. The standardising makes the predictors to have unit variance and models could not capture the differences between bankrupt and healthy firms with the same manner. Additionally, a performance test by using all firm specific and macroeconomic predictors was done. The improvements in already well-performing models of RF and LR could only have a minor benefit. However, LDA and QDA did improve but never reached over 0.9 in these two AUC analyses. This was already achieved by LR and RF by single macro predictor. Therefore, the performances of LDA and QDA remained limited in this thesis.

7 CONCLUSIONS

7.1 Research results

This thesis studied questions of quality of Z’’-score in today’s global environment, use of macroeconomic predictor in bankruptcy prediction and performance of RF to LR in the same context.

First, hypothesis (H_1) was evaluated by comparing the original Z’’-score to LR and RF. The use of LR and RF was strongly supported, based on validation sample from Finnish firms. All models performed better when one macroeconomic predictor was added to a model. Therefore, the utilization of macroeconomic predictor in bankruptcy prediction agreed with previous literature. As a conclusion, the probability of bankruptcy can be detected with a greater reliability than merely from firm specific predictors. Thirdly, the use of RF method showed better performance in bankruptcy prediction than any other method. The use of LR is commonly seen in the literature by the causality purposes. In LR the coefficients and their p-values explain the effects of certain predictors. However, this was not intended by the thesis but rather find a well-performing and practical bankruptcy model.

7.2 Future research

This thesis focused on Finnish Ltd. companies between years 2000 and 2019. The scope of data is thus limited. In future research international aspect should be evaluated to see whether the international results are similar with the thesis. In this thesis, only four firm specific variables were used, but some additional predictors could lead to even better results (see Altman et.al. 2017). The proportion of bankrupt firms (2.7%) does not represent the whole population (0.7%) which can lead to biased results. This question should be carefully analysed with balanced data in the future research. To unbiasedly predict probability of bankruptcy, the training data should not include any observations from future. Bankruptcy models must be able to handle unexpected events of future. The covid-19 distress caused a lot of firms to go bankrupt in some industries (airlines, restaurants etc.).

All this happened in a relatively short period of time even compared to previous economic downtrends. Despite the satisfactory levels of financial ratios, firm might have gone bankrupt during this pandemic. Therefore, updating predictors more frequently (such as macroeconomic) is essential to predict bankruptcies well in the future.

8 REFERENCES

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

Altman, E. (1971). Railroad bankruptcy propensity. Journal of Finance, 26(2), 333-345. https://doi.org/10.1111/j.1540-6261.1971.tb00901.x

Altman, E. (1983). Why businesses fail. Journal of Business Strategy. 3(4). 15-21. https://doi.org/10.1108/eb038985

Altman, E., Bibeault, D., & Casey C. J. (1984). Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy. *Journal of Business Strategy*, 5, 102-108.

Altman, E., Iwanicz-Drozdowska M., Laitinen E. & Suvas A. (2017). Financial Distress Prediciton in an International Context: A Review and Empirical Analysis of Altman's Z-score Model. *Journal of International Financial Management & Accounting*, 28(2), 131-171. https://doi.org/10.1111/jifm.12053

Altman, E. (2018). A 50-Year Retrospective on Credit Risk Models, the Altman Z-Score Family of Models and Their Applications to Financial Markets and Managerial Strategies. *Journal of Credit Risk,* 14(4), 1-34. https://doi.org/10.1002/9781119541929.ch10

Amini, A. (2020, July 28). MIT Introduction to Deep Learning | 6.S191. Massachusetts Institute of Technology [Video]. YouTube. https://www.youtube.com/watch?v=njKP3FqW3Sk

Balcaen, S. & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38, 63-93. https://doi.org/10.1016/j.bar.2005.09.001

Bank of Finland (2020, September 17). Household gearing and interest expenses in Finland. https://www.suomenpankki.fi/fi/Tilastot/kuviopankki/rahoituksen-suhdannemittarit/yksityisen-sektorin-velkaantuneisuus-tai-velanhoitorasite/velkaantumisaste_ja_korkorasitus/

Batchelor, T. (2018). Corporate Bankruptcy: Testing the Efficacy of the Altman Z-Score. *International research Journal of Applied Finance*, 9(9), 404-414.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 4, 71-111. https://doi.org/10.2307/2490171

Filipe, S. F., Grammatikos, T. & Michala, D. (2016). Forecasting distress in European SME portfolios. *Journal of Banking and Finance*, 64, 112-135. https://doi.org/10.1016/j.jbankfin.2015.12.007

Finnish Patent and Registration Office. (2020, July). Yritysten lukumäärät kaupparekisterissä. https://www.prh.fi/fi/kaupparekisteri/yritystenlkm/lkm.html

Hol, S. (2007). The influence of the business cycle on bankruptcy probability. *International transactions in Operational Research*, 14, 75-90. https://doi.org/10.1111/j.1475-3995.2006.00576.x

Institute of International Finance. (2020, April 7). April 2020 Global debt Monitor: COVID-19 Lights a Fuse. https://www.iif.com/Publications/ID/3839/April-2020-Global-Debt-Monitor--COVID-19-Lights-a-Fuse

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Joshi, S., Ramesh, R. & Tahsildar, S. (2018). A Bankruptcy Prediction Model Using Random Forest. *Proceedings of the Second International Conference on Intelligent Computing and Control Systems*. https://doi.org/10.1109/ICCONS.2018.8663128

Jung, A. (2018). *Machine learning: Basic principles*. arXiv preprint arXiv:1805.05052

Laamanen, A. (2017). Taloudellisten tunnuslukujen ja makrotaloudellisten muuttujien kyky ennustaa konkurssi suomalaisissa majoitus- ja ravitsemusalan pk- yrityksissä. [Unpublished master's thesis]. Lappeenranta University of Technology.

Lagandula, A. C., (2020, July 29). *Learning Parameters, Part 0: Basic Stuff*. https://cdn-images-1.medium.com/freeze/max/1000/1*XrL0tp9rHSeN5SosuAvC_g.png?q=20

Laitinen, E. K. (1990) *Konkurssin ennustaminen*. Vaasan Yritysinformaatio Oy.

Laitinen, E. K. (1992) *Yrityksen talouden mittarit* (2nd ed.), Gummerus Kirjapaino Oy.

Laitinen, E. K. & Laitinen, T. (2004). *Yrityksen rahoituskriisin ennustaminen*. Talentum Media Oy.

Lussier, R. N. (2005). A Success Versus Failure Prediction Model for the Real Estate Industry. *Mid-American Journal of Business* 20(1). https://doi.org/47-53. 10.1108/19355181200500005

Michelucci, U. (2018). *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-3790-8

Min, J. H. & Lee, Y-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*. 28. 603-614. https://doi.org/10.1016/j.eswa.2004.12.008

Min, S-H. Lee, J. & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Systems with Applications*, 31, 652-660. https://doi.org/10.1016/j.eswa.2005.09.070

Muchlinski D., Siroky D., He J & Kocher M. (2015). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 1-17. https://doi.org/10.1093/pan/mpv024

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

Rudd, J. M. & Priestley, J. L. (2017). *A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit*. Kennesaw State University. [Gray Literature from PhD Candidates]

Scikit-Learn: Decision Trees. (2020) [Python Libarary] https://scikit-learn.org/stable/modules/tree.html

Statistics Finland (2020, August 15). Limited liability bankruptcies 2003-2019 [Data set]. http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__oik__konk__vv/statfin_konk_pxt_11f8.px/table/tableViewLayout1/

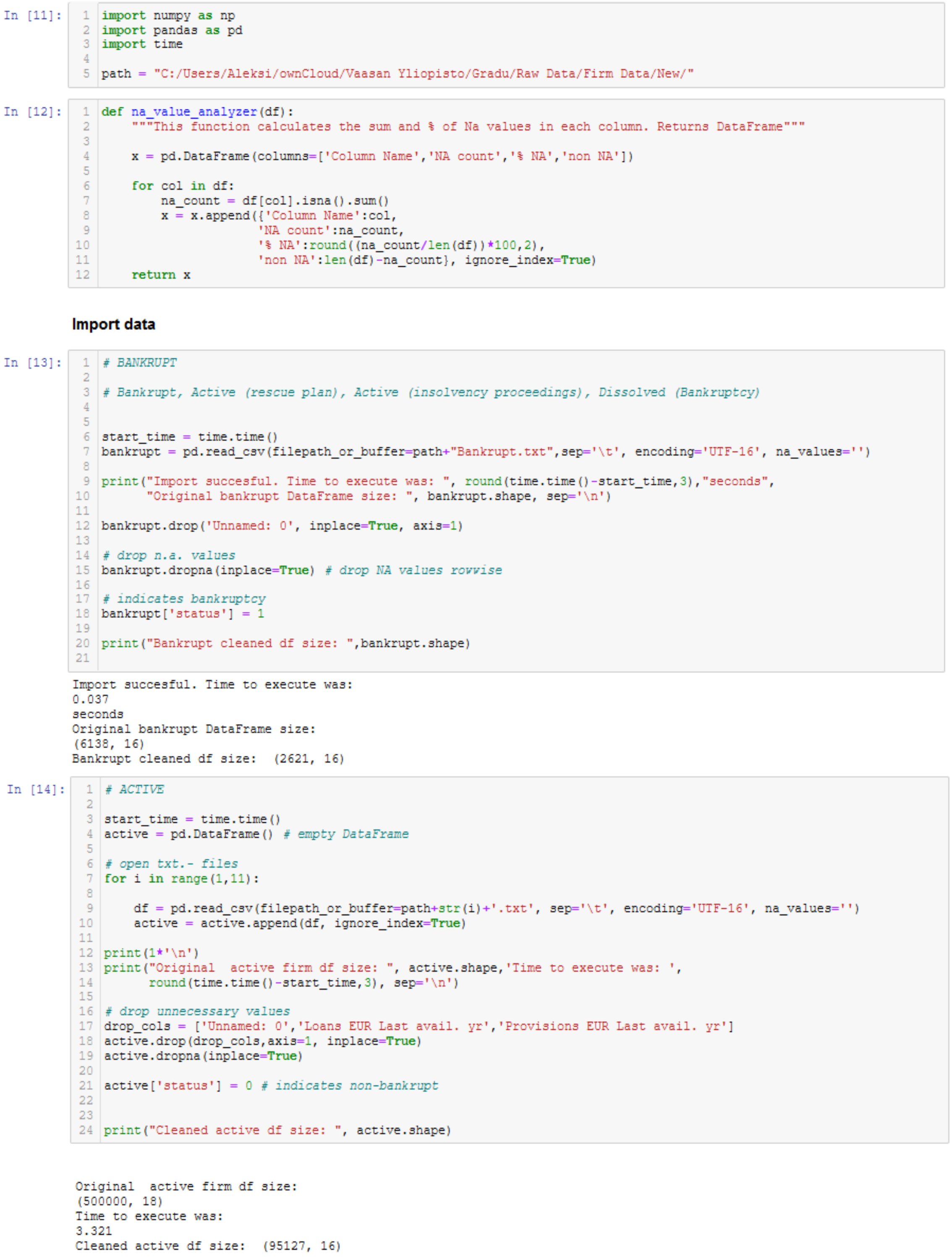
Tilastokeskus (2020, August). *Yritykset 2018*. http://www.tilastokeskus.fi/tup/suoluk/suoluk_yritykset.html

The Failure Record, Through (1965). New York: Dun & Bradstreet, Inc., 1966. p 10

Zavgren, C. V. & Friedman, G.E. (1988). Are Bankruptcy Prediction Models Worthwhile? An Application in Securities Analysis. *Management International Review*. 28(1). 34-44.

9 ATTACHMENTS

The code below covers all data-analysis made in this thesis and is divided into four parts: *Data Import, Creating Predictors, Descriptive Statistics (no editing) & Data-Analysis*



In [11]:
```
import numpy as np
import pandas as pd
import time

path = "C:/Users/Aleksi/ownCloud/Vaasan Yliopisto/Gradu/Raw Data/Firm Data/New/"
```

In [12]:
```
def na_value_analyzer(df):
    """This function calculates the sum and % of Na values in each column. Returns DataFrame"""

    x = pd.DataFrame(columns=['Column Name','NA count','% NA','non NA'])

    for col in df:
        na_count = df[col].isna().sum()
        x = x.append({'Column Name':col,
                      'NA count':na_count,
                      '% NA':round((na_count/len(df))*100,2),
                      'non NA':len(df)-na_count}, ignore_index=True)
    return x
```

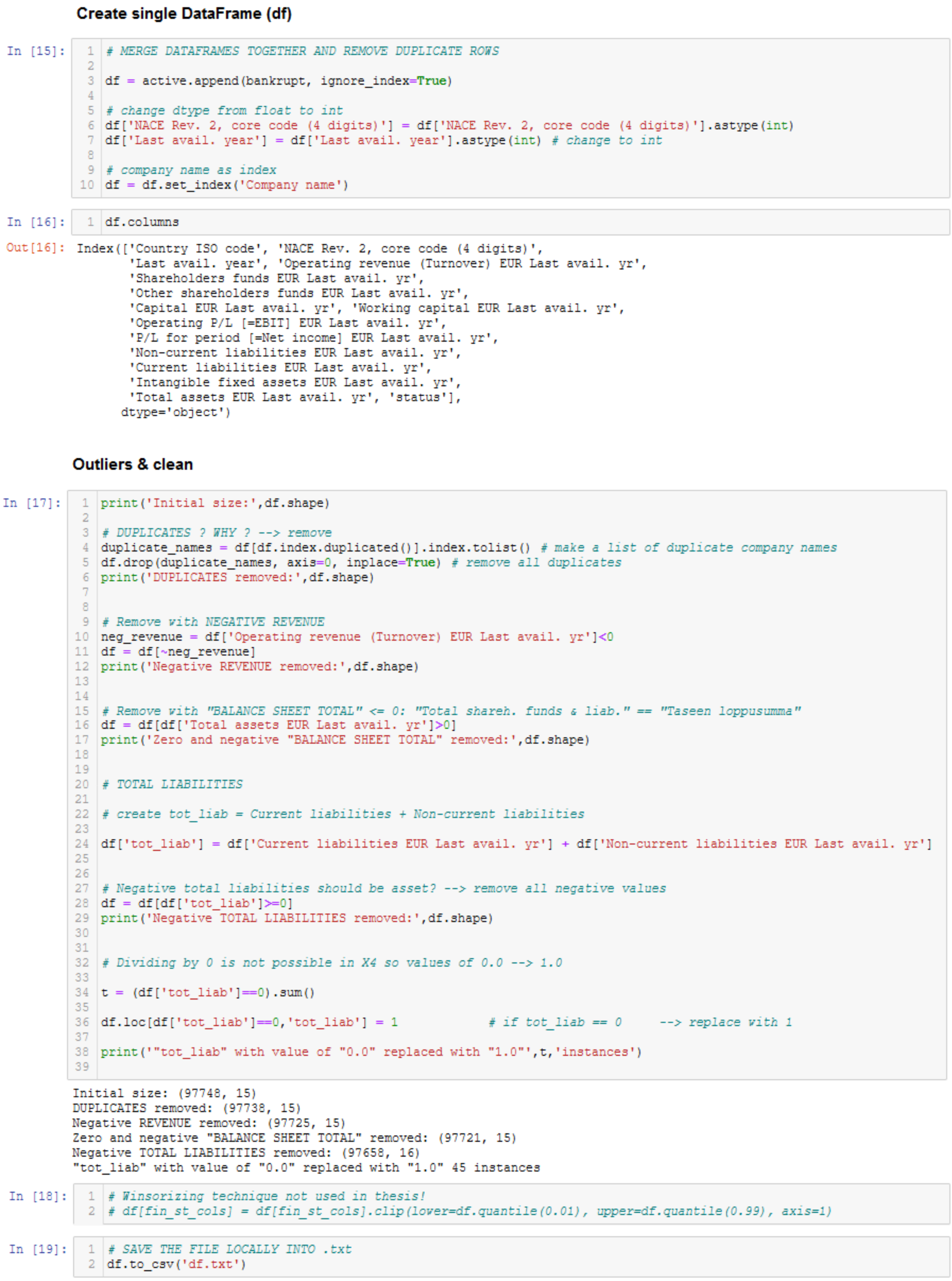
Import data

In [13]:
```
# BANKRUPT

# Bankrupt, Active (rescue plan), Active (insolvency proceedings), Dissolved (Bankruptcy)


start_time = time.time()
bankrupt = pd.read_csv(filepath_or_buffer=path+"Bankrupt.txt",sep='\t', encoding='UTF-16', na_values='')

print("Import succesful. Time to execute was: ", round(time.time()-start_time,3),"seconds",
      "Original bankrupt DataFrame size: ", bankrupt.shape, sep='\n')

bankrupt.drop('Unnamed: 0', inplace=True, axis=1)

# drop n.a. values
bankrupt.dropna(inplace=True) # drop NA values rowwise

# indicates bankruptcy
bankrupt['status'] = 1

print("Bankrupt cleaned df size: ",bankrupt.shape)

```

```
Import succesful. Time to execute was:
0.037
seconds
Original bankrupt DataFrame size:
(6138, 16)
Bankrupt cleaned df size:  (2621, 16)
```

In [14]:
```
# ACTIVE

start_time = time.time()
active = pd.DataFrame() # empty DataFrame

# open txt.- files
for i in range(1,11):

    df = pd.read_csv(filepath_or_buffer=path+str(i)+'.txt', sep='\t', encoding='UTF-16', na_values='')
    active = active.append(df, ignore_index=True)

print(1*'\n')
print("Original  active firm df size: ", active.shape,'Time to execute was: ',
      round(time.time()-start_time,3), sep='\n')

# drop unnecessary values
drop_cols = ['Unnamed: 0','Loans EUR Last avail. yr','Provisions EUR Last avail. yr']
active.drop(drop_cols,axis=1, inplace=True)
active.dropna(inplace=True)

active['status'] = 0 # indicates non-bankrupt


print("Cleaned active df size: ", active.shape)
```

```

Original  active firm df size:
(500000, 18)
Time to execute was:
3.321
Cleaned active df size:  (95127, 16)
```



Create single DataFrame (df)

In [15]:

```
# MERGE DATAFRAMES TOGETHER AND REMOVE DUPLICATE ROWS

df = active.append(bankrupt, ignore_index=True)

# change dtype from float to int
df['NACE Rev. 2, core code (4 digits)'] = df['NACE Rev. 2, core code (4 digits)'].astype(int)
df['Last avail. year'] = df['Last avail. year'].astype(int) # change to int

# company name as index
df = df.set_index('Company name')
```

In [16]:

```
df.columns
```

Out[16]:

```
Index(['Country ISO code', 'NACE Rev. 2, core code (4 digits)',
       'Last avail. year', 'Operating revenue (Turnover) EUR Last avail. yr',
       'Shareholders funds EUR Last avail. yr',
       'Other shareholders funds EUR Last avail. yr',
       'Capital EUR Last avail. yr', 'Working capital EUR Last avail. yr',
       'Operating P/L [=EBIT] EUR Last avail. yr',
       'P/L for period [=Net income] EUR Last avail. yr',
       'Non-current liabilities EUR Last avail. yr',
       'Current liabilities EUR Last avail. yr',
       'Intangible fixed assets EUR Last avail. yr',
       'Total assets EUR Last avail. yr', 'status'],
      dtype='object')
```

Outliers & clean

In [17]:

```
print('Initial size:',df.shape)

# DUPLICATES ? WHY ? --> remove
duplicate_names = df[df.index.duplicated()].index.tolist() # make a list of duplicate company names
df.drop(duplicate_names, axis=0, inplace=True) # remove all duplicates
print('DUPLICATES removed:',df.shape)


# Remove with NEGATIVE REVENUE
neg_revenue = df['Operating revenue (Turnover) EUR Last avail. yr']<0
df = df[~neg_revenue]
print('Negative REVENUE removed:',df.shape)


# Remove with "BALANCE SHEET TOTAL" <= 0: "Total shareh. funds & liab." == "Taseen loppusumma"
df = df[df['Total assets EUR Last avail. yr']>0]
print('Zero and negative "BALANCE SHEET TOTAL" removed:',df.shape)


# TOTAL LIABILITIES

# create tot_liab = Current liabilities + Non-current liabilities

df['tot_liab'] = df['Current liabilities EUR Last avail. yr'] + df['Non-current liabilities EUR Last avail. yr']


# Negative total liabilities should be asset? --> remove all negative values
df = df[df['tot_liab']>=0]
print('Negative TOTAL LIABILITIES removed:',df.shape)


# Dividing by 0 is not possible in X4 so values of 0.0 --> 1.0

t = (df['tot_liab']==0).sum()

df.loc[df['tot_liab']==0,'tot_liab'] = 1             # if tot_liab == 0     --> replace with 1

print('"tot_liab" with value of "0.0" replaced with "1.0"',t,'instances')

```

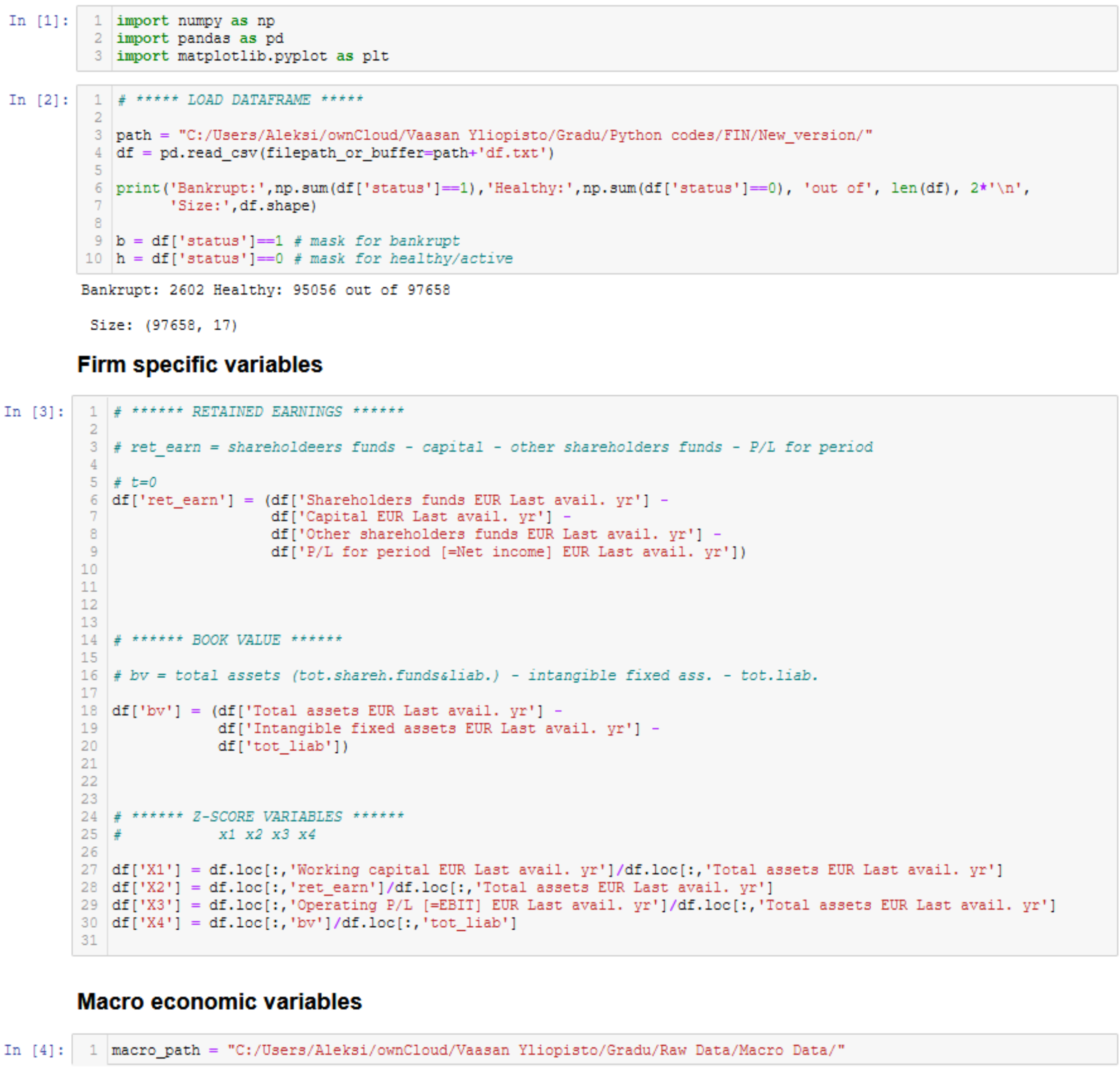
```
Initial size: (97748, 15)
DUPLICATES removed: (97738, 15)
Negative REVENUE removed: (97725, 15)
Zero and negative "BALANCE SHEET TOTAL" removed: (97721, 15)
Negative TOTAL LIABILITIES removed: (97658, 16)
"tot_liab" with value of "0.0" replaced with "1.0" 45 instances
```

In [18]:

```
# Winsorizing technique not used in thesis!
# df[fin_st_cols] = df[fin_st_cols].clip(lower=df.quantile(0.01), upper=df.quantile(0.99), axis=1)
```

In [19]:

```
# SAVE THE FILE LOCALLY INTO .txt
df.to_csv('df.txt')
```



In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:

```
# ***** LOAD DATAFRAME *****

path = "C:/Users/Aleksi/ownCloud/Vaasan Yliopisto/Gradu/Python codes/FIN/New_version/"
df = pd.read_csv(filepath_or_buffer=path+'df.txt')

print('Bankrupt:',np.sum(df['status']==1),'Healthy:',np.sum(df['status']==0), 'out of', len(df), 2*'\n',
      'Size:',df.shape)

b = df['status']==1 # mask for bankrupt
h = df['status']==0 # mask for healthy/active
```

```
Bankrupt: 2602 Healthy: 95056 out of 97658

 Size: (97658, 17)
```

Firm specific variables

In [3]:

```
# ****** RETAINED EARNINGS ******

# ret_earn = shareholdeers funds - capital - other shareholders funds - P/L for period

# t=0
df['ret_earn'] = (df['Shareholders funds EUR Last avail. yr'] -
                  df['Capital EUR Last avail. yr'] -
                  df['Other shareholders funds EUR Last avail. yr'] -
                  df['P/L for period [=Net income] EUR Last avail. yr'])




# ****** BOOK VALUE ******

# bv = total assets (tot.shareh.funds&liab.) - intangible fixed ass. - tot.liab.

df['bv'] = (df['Total assets EUR Last avail. yr'] -
            df['Intangible fixed assets EUR Last avail. yr'] -
            df['tot_liab'])



# ****** Z-SCORE VARIABLES ******
#            x1 x2 x3 x4

df['X1'] = df.loc[:,'Working capital EUR Last avail. yr']/df.loc[:,'Total assets EUR Last avail. yr']
df['X2'] = df.loc[:,'ret_earn']/df.loc[:,'Total assets EUR Last avail. yr']
df['X3'] = df.loc[:,'Operating P/L [=EBIT] EUR Last avail. yr']/df.loc[:,'Total assets EUR Last avail. yr']
df['X4'] = df.loc[:,'bv']/df.loc[:,'tot_liab']

```

Macro economic variables

In [4]:

```
macro_path = "C:/Users/Aleksi/ownCloud/Vaasan Yliopisto/Gradu/Raw Data/Macro Data/"
```

GDP (change %)

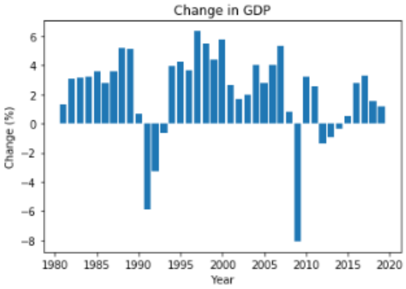
In [5]:

```
# import data
gdp = pd.read_csv(filepath_or_buffer=macro_path+'GDP (eurostat)/nama_10_gdp_1_Data.csv', na_values=[':',''] )

# keep only necessary data
gdp = gdp.loc[gdp['UNIT']=='Chain linked volumes (2010), million euro',['TIME','UNIT','Value']]
gdp.dropna(inplace=True)

gdp.reset_index(inplace=True, drop=True)

# modify blank space in string and change to numeric
gdp['Value'] = gdp['Value'].str.replace(" ","")
gdp['Value'] = pd.to_numeric(gdp['Value'])

# add two columns: lagged and % change
gdp['lagged -1'] = gdp['Value'].shift(periods=1)
gdp['change %'] = ((gdp['Value']/gdp['lagged -1'])-1)*100


print('Deleting year == 2020 leaves out',(df['Last avail. year']== 2020).sum(),
      'observations where bankrupt',
      ((df['Last avail. year']== 2020)&(df['status']== 1)).sum())

# 2020 GDP not known --> so need to delete corresponding rows from df
df = df.loc[(df['Last avail. year']!= 2020), :]

# make a dictionary for mapping and map
gdp_dict = (gdp.loc[:,['TIME','change %']]).set_index('TIME').to_dict()

# add new column 'GDP change %' to df
df['GDP change %'] = list(map(lambda x: gdp_dict['change %'][x], df['Last avail. year']))

```

```
Deleting year == 2020 leaves out 659 observations where bankrupt 7
```

In [6]:

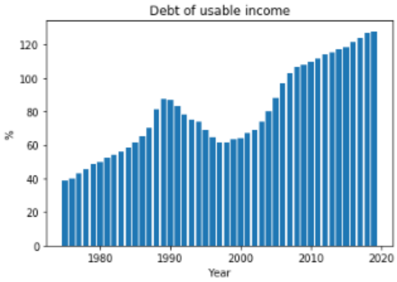
```
plt.bar(*zip(*gdp_dict['change %'].items()))
plt.xlabel('Year')
plt.ylabel('Change (%)')
plt.title('Change in GDP')
plt.show()
```

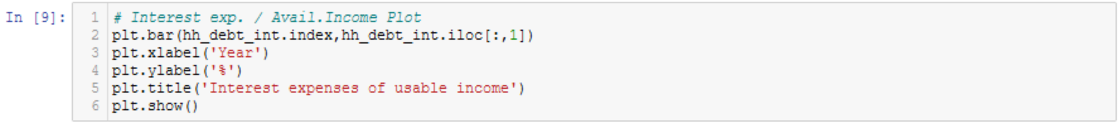


Household debt & Interest as proportion usable income (%)

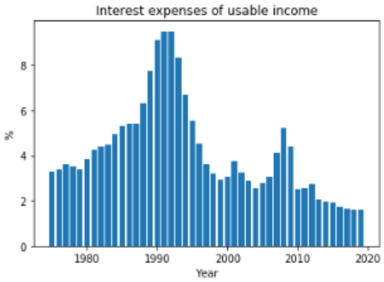
In [7]:

```
# household debt & interest are a percentage(%) of usable income (=Finnish definition) Bank of Finland??
hh_debt_int = pd.read_csv(filepath_or_buffer=macro_path+
                          'Household debt and interest/Household debt and interest.csv',
                         delimiter = ';')
hh_debt_int.dropna(inplace=True)

# change the column names
hh_debt_int.columns = ['Date','Household debt %','Household interest %']

# distinguish the years --> list
hh_debt_int['Date'] = hh_debt_int['Date'].str.split('.')

# remove dates, leave 'Date' as int
for i in range(len(hh_debt_int['Date'])):
    hh_debt_int['Date'][i] = int(hh_debt_int['Date'][i][2])

# remove whitespaces
hh_debt_int.loc[:,'Household debt %'] = hh_debt_int.loc[:,'Household debt %'].str.strip()
hh_debt_int.loc[:,'Household interest %'] = hh_debt_int.loc[:,'Household interest %'].str.strip()

# replace ',' with '.' and change to numeric
hh_debt_int['Household debt %'] = pd.to_numeric(hh_debt_int['Household debt %'].str.replace(',','.'))
hh_debt_int['Household interest %'] = pd.to_numeric(hh_debt_int['Household interest %'].str.replace(',','.'))

# take average of quarters
hh_debt_int = hh_debt_int.groupby(['Date']).mean()

# make dictionary
hh_dict = (hh_debt_int.loc[:,['Household debt %','Household interest %']]).to_dict()

# add new column 'Household debt %' and 'Household interest %'(map function) to df
df['Household debt %'] = list(map(lambda x: hh_dict['Household debt %'][x], df['Last avail. year']))
df['Household interest %'] = list(map(lambda x: hh_dict['Household interest %'][x], df['Last avail. year']))

```

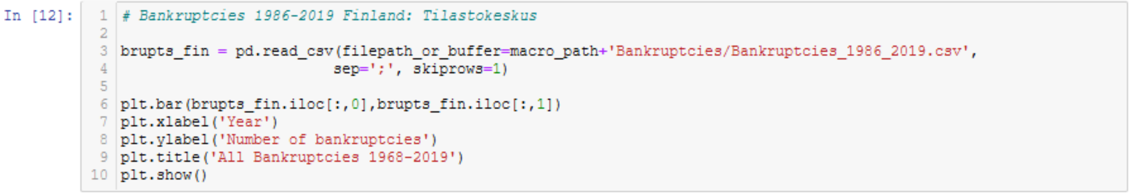




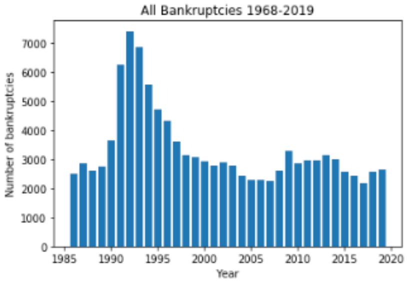
```
In [9]:
# Interest exp. / Avail.Income Plot
plt.bar(hh_debt_int.index,hh_debt_int.iloc[:,1])
plt.xlabel('Year')
plt.ylabel('%')
plt.title('Interest expenses of usable income')
plt.show()
```

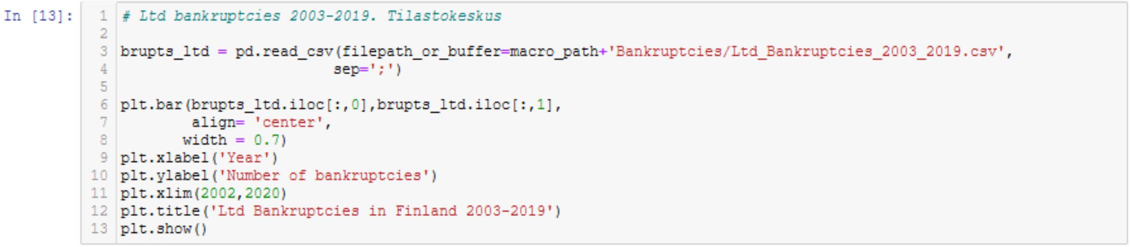


Bankruptcies in Finland 1986-2019

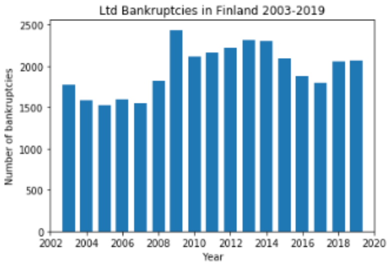


```
In [12]:
# Bankruptcies 1986-2019 Finland: Tilastokeskus

brupts_fin = pd.read_csv(filepath_or_buffer=macro_path+'Bankruptcies/Bankruptcies_1986_2019.csv',
                        sep=';', skiprows=1)

plt.bar(brupts_fin.iloc[:,0],brupts_fin.iloc[:,1])
plt.xlabel('Year')
plt.ylabel('Number of bankruptcies')
plt.title('All Bankruptcies 1968-2019')
plt.show()
```





```
In [13]:
# Ltd bankruptcies 2003-2019. Tilastokeskus

brupts_ltd = pd.read_csv(filepath_or_buffer=macro_path+'Bankruptcies/Ltd_Bankruptcies_2003_2019.csv',
                        sep=';')

plt.bar(brupts_ltd.iloc[:,0],brupts_ltd.iloc[:,1],
        align= 'center',
       width = 0.7)
plt.xlabel('Year')
plt.ylabel('Number of bankruptcies')
plt.xlim(2002,2020)
plt.title('Ltd Bankruptcies in Finland 2003-2019')
plt.show()
```



In [14]:

```python
# CORRELATION OF MACROPREDICTORS

hh_df = pd.DataFrame.from_dict(hh_dict)
gdp_df = pd.DataFrame.from_dict(gdp_dict)

macro_df = hh_df.join(gdp_df)

macro_df.loc[1999:,:].corr()
```

Out[14]:

	Household debt %	Household interest %	change %
Household debt %	1.000000	-0.432816	-0.379541
Household interest %	-0.432816	1.000000	-0.148851
change %	-0.379541	-0.148851	1.000000

In [15]:

```python
# save df as 'df2.txt' csv-file
df.to_csv('df2.txt')
```

Descriptive statistics and correlation of predictors (no data editing)

In [9]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from scipy.stats import mannwhitneyu, kstest

```

In [10]:

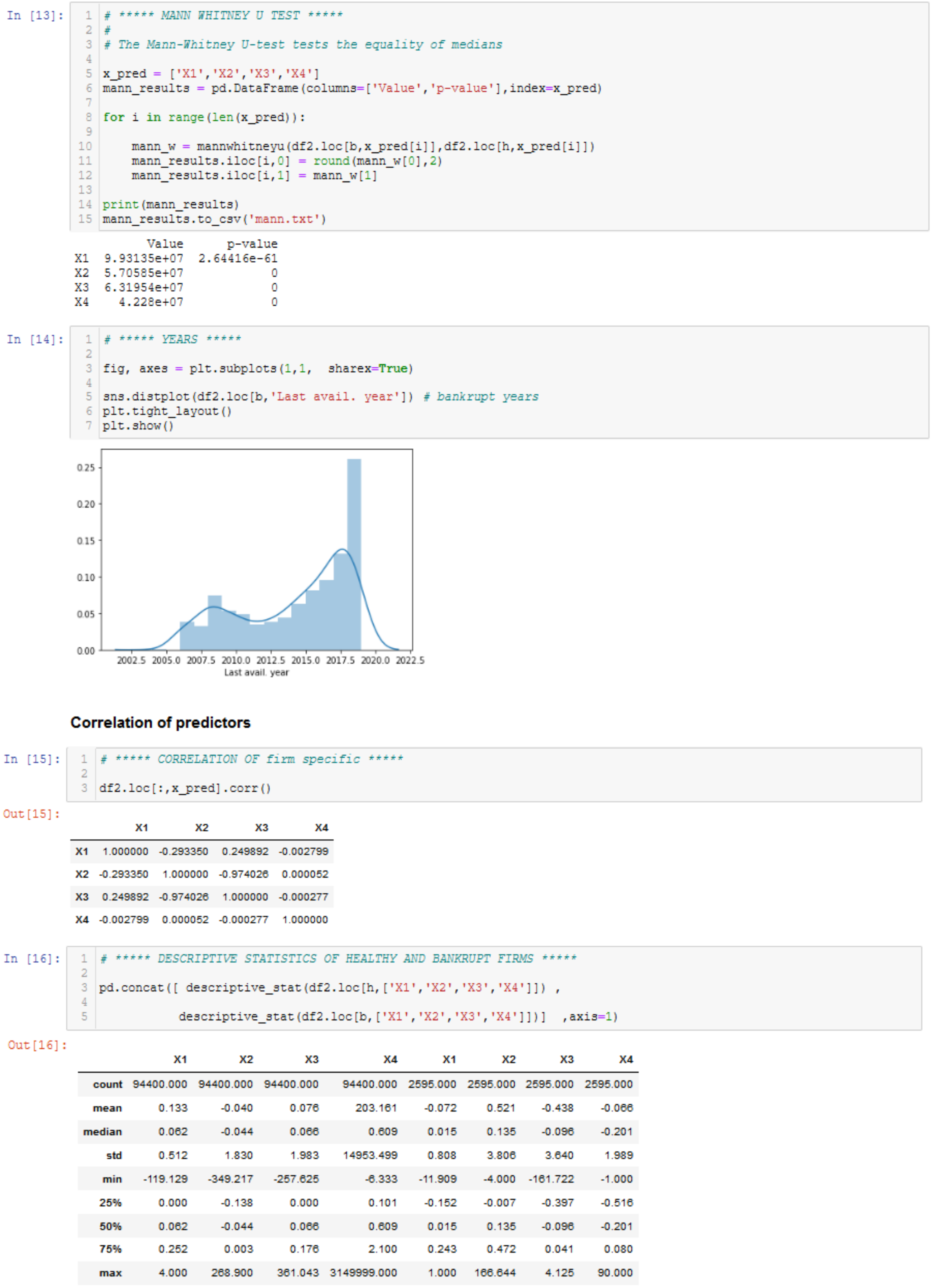
```python
# ***** LOAD DATAFRAME: df2 *****

path = "C:/Users/Aleksi/ownCloud/Vaasan Yliopisto/Gradu/Python codes/FIN/New_Version/"
df2 = pd.read_csv(filepath_or_buffer=path+'df2.txt')
df2.drop('Unnamed: 0',inplace=True,axis=1)

b = df2['status']==1 # mask for bankrupt
h = df2['status']==0 # mask for healthy/active
```

In [11]:

```python
def descriptive_stat(df):
    """This function returns a descriptive statistics from
    columns of DataFrame: returns df object"""

    desc = df.describe()
    desc.loc['median',:] = 0                   # create empty row for median
    ind = desc.index                           # get indexes
    order = [0,1,len(ind)-1,2,3,4,5,6,7]       # create order for indexes
    ind = [ind[i] for i in order]              # edit the original list
    desc = desc.reindex(ind)                   # reindex the df

    for col in df:
        desc.loc['median',col] = df.loc[:,col].median()

    return round(desc,3)

```

In [12]:

```python
# ****** KOLMOGOROV-SMIRNOV TEST *****
#
# KS tests the equality of the distributions
# kstest()
```



In [13]:

```
# ***** MANN WHITNEY U TEST *****
#
# The Mann-Whitney U-test tests the equality of medians

x_pred = ['X1','X2','X3','X4']
mann_results = pd.DataFrame(columns=['Value','p-value'],index=x_pred)

for i in range(len(x_pred)):

    mann_w = mannwhitneyu(df2.loc[b,x_pred[i]],df2.loc[h,x_pred[i]])
    mann_results.iloc[i,0] = round(mann_w[0],2)
    mann_results.iloc[i,1] = mann_w[1]

print(mann_results)
mann_results.to_csv('mann.txt')
```

```
          Value      p-value
X1  9.93135e+07  2.64416e-61
X2  5.70585e+07            0
X3  6.31954e+07            0
X4    4.228e+07            0
```

In [14]:

```
# ***** YEARS *****

fig, axes = plt.subplots(1,1,  sharex=True)

sns.distplot(df2.loc[b,'Last avail. year']) # bankrupt years
plt.tight_layout()
plt.show()
```

Correlation of predictors

In [15]:

```
# ***** CORRELATION OF firm specific *****

df2.loc[:,x_pred].corr()
```

Out[15]:

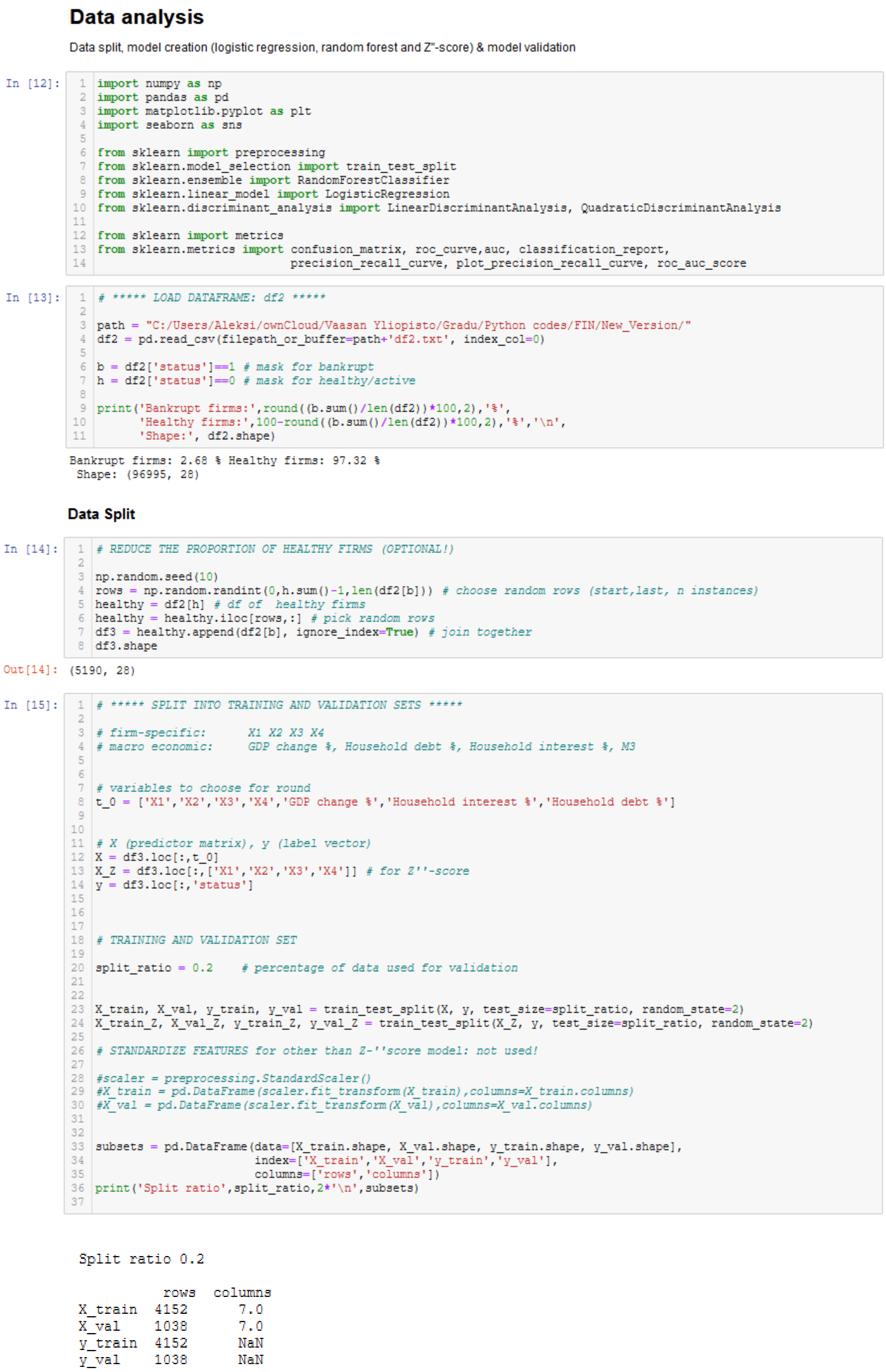
	X1	X2	X3	X4
X1	1.000000	-0.293350	0.249892	-0.002799
X2	-0.293350	1.000000	-0.974026	0.000052
X3	0.249892	-0.974026	1.000000	-0.000277
X4	-0.002799	0.000052	-0.000277	1.000000

In [16]:

```
# ***** DESCRIPTIVE STATISTICS OF HEALTHY AND BANKRUPT FIRMS *****

pd.concat([ descriptive_stat(df2.loc[h,['X1','X2','X3','X4']]) ,

            descriptive_stat(df2.loc[b,['X1','X2','X3','X4']])]  ,axis=1)
```

Out[16]:

	X1	X2	X3	X4	X1	X2	X3	X4
count	94400.000	94400.000	94400.000	94400.000	2595.000	2595.000	2595.000	2595.000
mean	0.133	-0.040	0.076	203.161	-0.072	0.521	-0.438	-0.066
median	0.062	-0.044	0.066	0.609	0.015	0.135	-0.096	-0.201
std	0.512	1.830	1.983	14953.499	0.808	3.806	3.640	1.989
min	-119.129	-349.217	-257.625	-6.333	-11.909	-4.000	-161.722	-1.000
25%	0.000	-0.138	0.000	0.101	-0.152	-0.007	-0.397	-0.516
50%	0.062	-0.044	0.066	0.609	0.015	0.135	-0.096	-0.201
75%	0.252	0.003	0.176	2.100	0.243	0.472	0.041	0.080
max	4.000	268.900	361.043	3149999.000	1.000	166.644	4.125	90.000



Data analysis

Data split, model creation (logistic regression, random forest and Z''-score) & model validation

In [12]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis

from sklearn import metrics
from sklearn.metrics import confusion_matrix, roc_curve,auc, classification_report,
                            precision_recall_curve, plot_precision_recall_curve, roc_auc_score
```

In [13]:

```
# ***** LOAD DATAFRAME: df2 *****

path = "C:/Users/Aleksi/ownCloud/Vaasan Yliopisto/Gradu/Python codes/FIN/New_Version/"
df2 = pd.read_csv(filepath_or_buffer=path+'df2.txt', index_col=0)

b = df2['status']==1 # mask for bankrupt
h = df2['status']==0 # mask for healthy/active

print('Bankrupt firms:',round((b.sum()/len(df2))*100,2),'%',
      'Healthy firms:',100-round((b.sum()/len(df2))*100,2),'%','\n',
      'Shape:', df2.shape)
```

```
Bankrupt firms: 2.68 % Healthy firms: 97.32 %
 Shape: (96995, 28)
```

Data Split

In [14]:

```
# REDUCE THE PROPORTION OF HEALTHY FIRMS (OPTIONAL!)

np.random.seed(10)
rows = np.random.randint(0,h.sum()-1,len(df2[b])) # choose random rows (start,last, n instances)
healthy = df2[h] # df of  healthy firms
healthy = healthy.iloc[rows,:] # pick random rows
df3 = healthy.append(df2[b], ignore_index=True) # join together
df3.shape
```

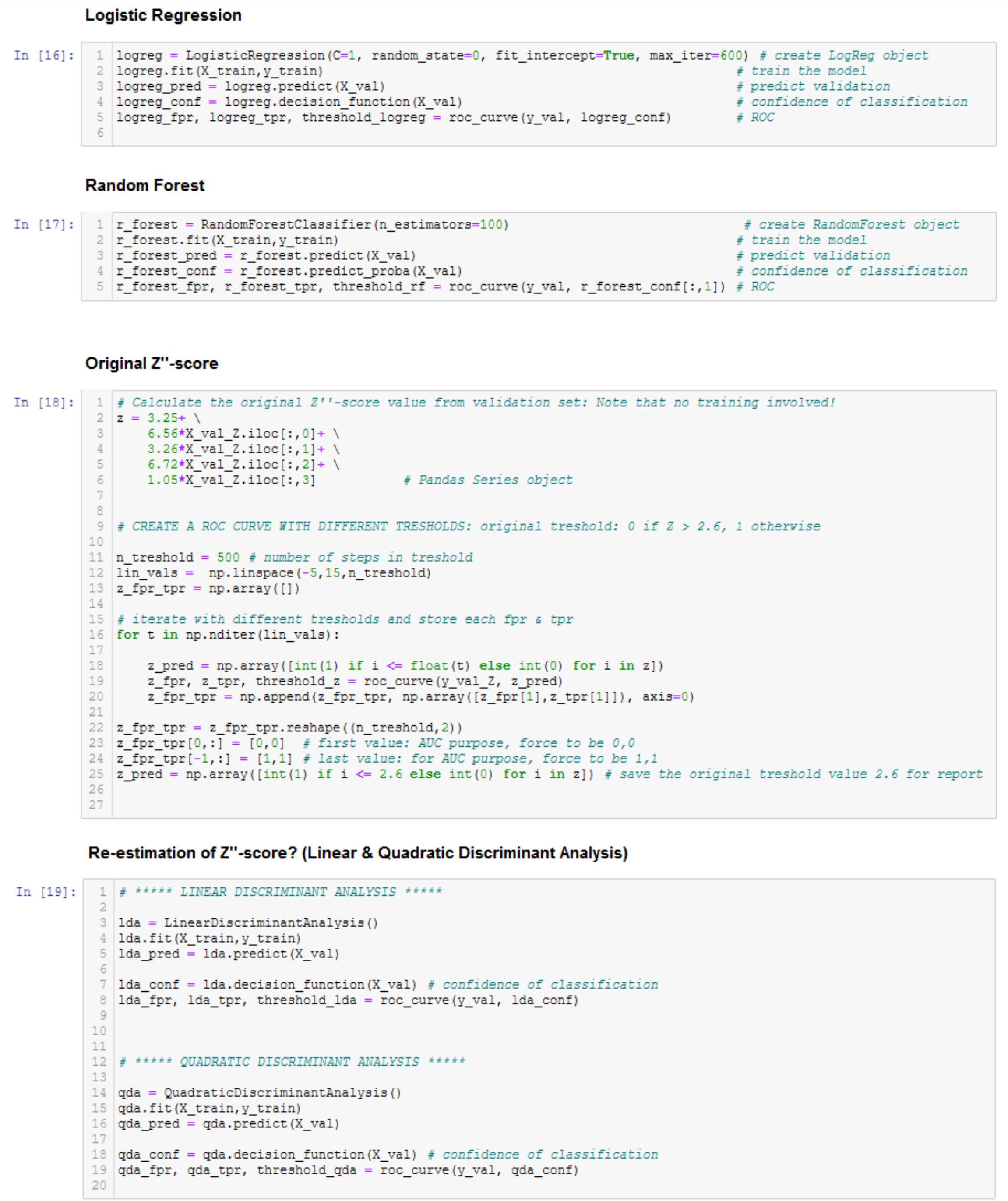
Out[14]:

```
(5190, 28)
```

In [15]:

```
# ***** SPLIT INTO TRAINING AND VALIDATION SETS *****

# firm-specific:      X1 X2 X3 X4
# macro economic:     GDP change %, Household debt %, Household interest %, M3


# variables to choose for round
t_0 = ['X1','X2','X3','X4','GDP change %','Household interest %','Household debt %']


# X (predictor matrix), y (label vector)
X = df3.loc[:,t_0]
X_Z = df3.loc[:,['X1','X2','X3','X4']] # for Z''-score
y = df3.loc[:,'status']



# TRAINING AND VALIDATION SET

split_ratio = 0.2    # percentage of data used for validation


X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=split_ratio, random_state=2)
X_train_Z, X_val_Z, y_train_Z, y_val_Z = train_test_split(X_Z, y, test_size=split_ratio, random_state=2)

# STANDARDIZE FEATURES for other than Z-''score model: not used!

#scaler = preprocessing.StandardScaler()
#X_train = pd.DataFrame(scaler.fit_transform(X_train),columns=X_train.columns)
#X_val = pd.DataFrame(scaler.fit_transform(X_val),columns=X_val.columns)


subsets = pd.DataFrame(data=[X_train.shape, X_val.shape, y_train.shape, y_val.shape],
                       index=['X_train','X_val','y_train','y_val'],
                       columns=['rows','columns'])
print('Split ratio',split_ratio,2*'\n',subsets)
```

```
Split ratio 0.2

          rows  columns
X_train   4152      7.0
X_val     1038      7.0
y_train   4152      NaN
y_val     1038      NaN
```



Logistic Regression

In [16]:

```
logreg = LogisticRegression(C=1, random_state=0, fit_intercept=True, max_iter=600) # create LogReg object
logreg.fit(X_train,y_train)                                                          # train the model
logreg_pred = logreg.predict(X_val)                                                  # predict validation
logreg_conf = logreg.decision_function(X_val)                                        # confidence of classification
logreg_fpr, logreg_tpr, threshold_logreg = roc_curve(y_val, logreg_conf)             # ROC

```

Random Forest

In [17]:

```
r_forest = RandomForestClassifier(n_estimators=100)                                   # create RandomForest object
r_forest.fit(X_train,y_train)                                                        # train the model
r_forest_pred = r_forest.predict(X_val)                                              # predict validation
r_forest_conf = r_forest.predict_proba(X_val)                                        # confidence of classification
r_forest_fpr, r_forest_tpr, threshold_rf = roc_curve(y_val, r_forest_conf[:,1])  # ROC
```

Original Z''-score

In [18]:

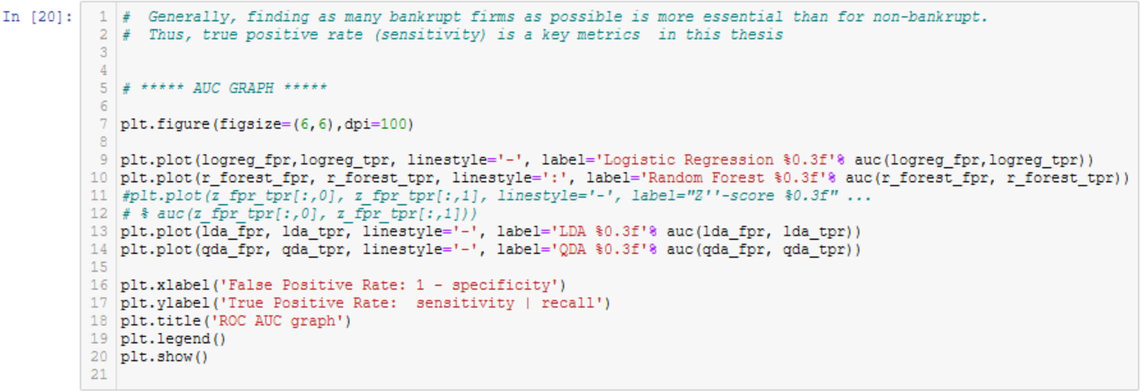
```
# Calculate the original Z''-score value from validation set: Note that no training involved!
z = 3.25+ \
    6.56*X_val_Z.iloc[:,0]+ \
    3.26*X_val_Z.iloc[:,1]+ \
    6.72*X_val_Z.iloc[:,2]+ \
    1.05*X_val_Z.iloc[:,3]           # Pandas Series object


# CREATE A ROC CURVE WITH DIFFERENT TRESHOLDS: original treshold: 0 if Z > 2.6, 1 otherwise

n_treshold = 500 # number of steps in treshold
lin_vals =  np.linspace(-5,15,n_treshold)
z_fpr_tpr = np.array([])

# iterate with different tresholds and store each fpr & tpr
for t in np.nditer(lin_vals):

    z_pred = np.array([int(1) if i <= float(t) else int(0) for i in z])
    z_fpr, z_tpr, threshold_z = roc_curve(y_val_Z, z_pred)
    z_fpr_tpr = np.append(z_fpr_tpr, np.array([z_fpr[1],z_tpr[1]]), axis=0)

z_fpr_tpr = z_fpr_tpr.reshape((n_treshold,2))
z_fpr_tpr[0,:] = [0,0]  # first value: AUC purpose, force to be 0,0
z_fpr_tpr[-1,:] = [1,1] # last value: for AUC purpose, force to be 1,1
z_pred = np.array([int(1) if i <= 2.6 else int(0) for i in z]) # save the original treshold value 2.6 for report

```

Re-estimation of Z''-score? (Linear & Quadratic Discriminant Analysis)

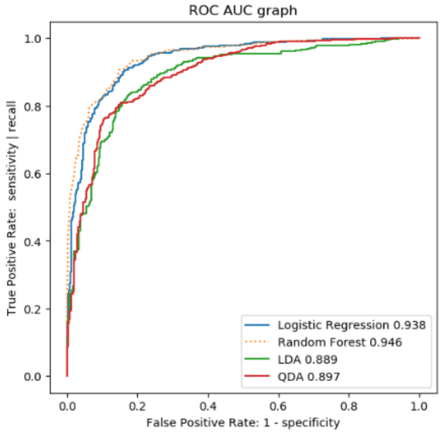
In [19]:

```
# ***** LINEAR DISCRIMINANT ANALYSIS *****

lda = LinearDiscriminantAnalysis()
lda.fit(X_train,y_train)
lda_pred = lda.predict(X_val)

lda_conf = lda.decision_function(X_val) # confidence of classification
lda_fpr, lda_tpr, threshold_lda = roc_curve(y_val, lda_conf)



# ***** QUADRATIC DISCRIMINANT ANALYSIS *****

qda = QuadraticDiscriminantAnalysis()
qda.fit(X_train,y_train)
qda_pred = qda.predict(X_val)

qda_conf = qda.decision_function(X_val) # confidence of classification
qda_fpr, qda_tpr, threshold_qda = roc_curve(y_val, qda_conf)

```

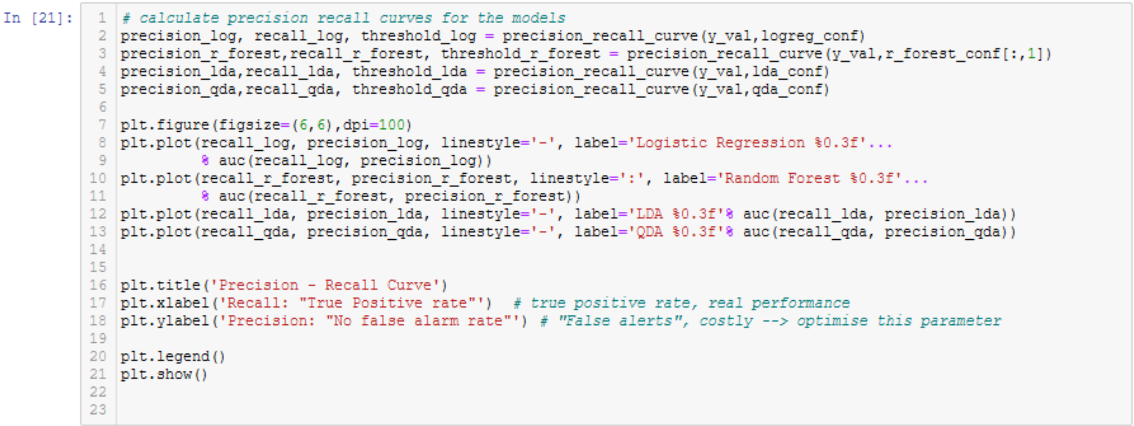
Plot AUC



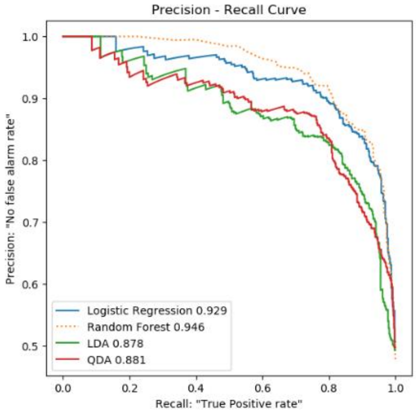
```
In [20]:
# Generally, finding as many bankrupt firms as possible is more essential than for non-bankrupt.
# Thus, true positive rate (sensitivity) is a key metrics  in this thesis


# ***** AUC GRAPH *****

plt.figure(figsize=(6,6),dpi=100)

plt.plot(logreg_fpr,logreg_tpr, linestyle='-', label='Logistic Regression %0.3f'% auc(logreg_fpr,logreg_tpr))
plt.plot(r_forest_fpr, r_forest_tpr, linestyle=':', label='Random Forest %0.3f'% auc(r_forest_fpr, r_forest_tpr))
#plt.plot(z_fpr_tpr[:,0], z_fpr_tpr[:,1], linestyle='-', label="Z''-score %0.3f" ...
# % auc(z_fpr_tpr[:,0], z_fpr_tpr[:,1]))
plt.plot(lda_fpr, lda_tpr, linestyle='-', label='LDA %0.3f'% auc(lda_fpr, lda_tpr))
plt.plot(qda_fpr, qda_tpr, linestyle='-', label='QDA %0.3f'% auc(qda_fpr, qda_tpr))

plt.xlabel('False Positive Rate: 1 - specificity')
plt.ylabel('True Positive Rate:  sensitivity | recall')
plt.title('ROC AUC graph')
plt.legend()
plt.show()
```

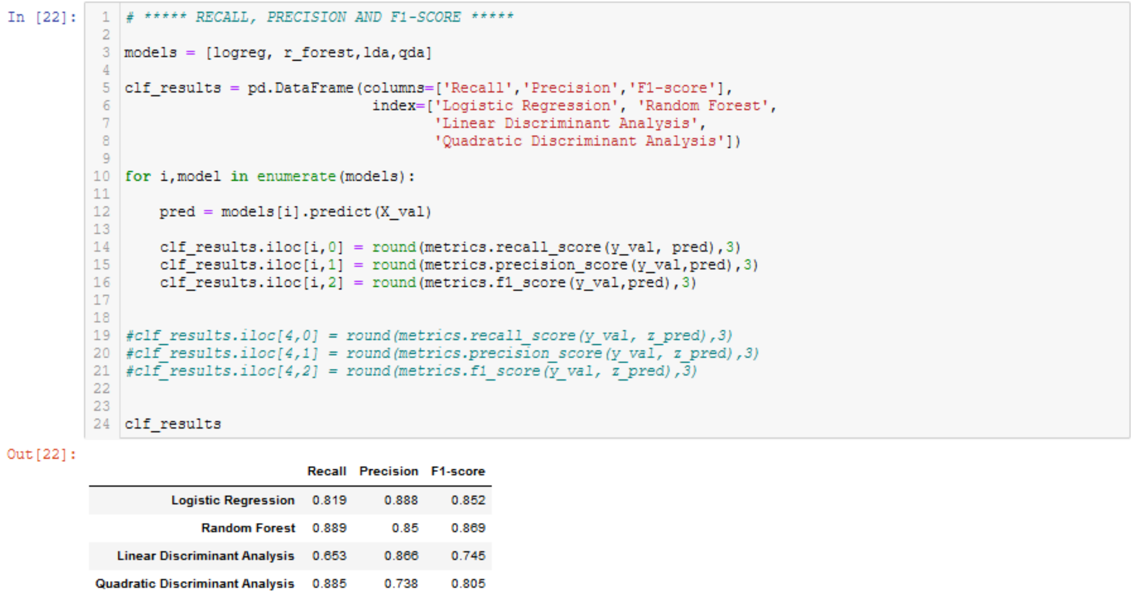


Plot Precision-Recall Curve



```
In [21]:
# calculate precision recall curves for the models
precision_log, recall_log, threshold_log = precision_recall_curve(y_val,logreg_conf)
precision_r_forest,recall_r_forest, threshold_r_forest = precision_recall_curve(y_val,r_forest_conf[:,1])
precision_lda,recall_lda, threshold_lda = precision_recall_curve(y_val,lda_conf)
precision_qda,recall_qda, threshold_qda = precision_recall_curve(y_val,qda_conf)

plt.figure(figsize=(6,6),dpi=100)
plt.plot(recall_log, precision_log, linestyle='-', label='Logistic Regression %0.3f'...
         % auc(recall_log, precision_log))
plt.plot(recall_r_forest, precision_r_forest, linestyle=':', label='Random Forest %0.3f'...
         % auc(recall_r_forest, precision_r_forest))
plt.plot(recall_lda, precision_lda, linestyle='-', label='LDA %0.3f'% auc(recall_lda, precision_lda))
plt.plot(recall_qda, precision_qda, linestyle='-', label='QDA %0.3f'% auc(recall_qda, precision_qda))


plt.title('Precision - Recall Curve')
plt.xlabel('Recall: "True Positive rate"')  # true positive rate, real performance
plt.ylabel('Precision: "No false alarm rate"') # "False alerts", costly --> optimise this parameter

plt.legend()
plt.show()
```





In [22]:

```
# ***** RECALL, PRECISION AND F1-SCORE *****

models = [logreg, r_forest,lda,qda]

clf_results = pd.DataFrame(columns=['Recall','Precision','F1-score'],
                            index=['Logistic Regression', 'Random Forest',
                                   'Linear Discriminant Analysis',
                                   'Quadratic Discriminant Analysis'])

for i,model in enumerate(models):

    pred = models[i].predict(X_val)

    clf_results.iloc[i,0] = round(metrics.recall_score(y_val, pred),3)
    clf_results.iloc[i,1] = round(metrics.precision_score(y_val,pred),3)
    clf_results.iloc[i,2] = round(metrics.f1_score(y_val,pred),3)


#clf_results.iloc[4,0] = round(metrics.recall_score(y_val, z_pred),3)
#clf_results.iloc[4,1] = round(metrics.precision_score(y_val, z_pred),3)
#clf_results.iloc[4,2] = round(metrics.f1_score(y_val, z_pred),3)


clf_results
```

Out[22]:

	Recall	Precision	F1-score
Logistic Regression	0.819	0.888	0.852
Random Forest	0.889	0.85	0.869
Linear Discriminant Analysis	0.653	0.866	0.745
Quadratic Discriminant Analysis	0.885	0.738	0.805