Master Thesis
Cand.Merc. Accounting, Strategy & Control
Copenhagen Business School

May 15th 2018

Can Publicly Available Non-Financial Data Significantly Improve Corporate Default Prediction Accuracy?

By:
Andreas Olczyk &
Malthe Dyvekær Nybjerg

Supervisor:
Benjamin Christoffersen

Number of standard pages: 114
Number of characters: 258.894

CBS
COPENHAGEN
BUSINESS SCHOOL
HANDELSHØJSKOLEN

Abstract

Huge amounts of financial and non-financial data has been made freely accessible to the general public by the Danish Business Register in an effort to digitize its operations and to share knowledge. It is the hope that the open data will create knowledge and opportunities for companies and ultimately create growth. The non-financial data is a record of all – recorded by the Central Business Register – non-financial attributes a company has had since its establishment.

This paper seeks to test whether this non-financial data can add any significant explanatory power to corporate default prediction models. Data used in the analyses consists of 93.000 observations for Danish ApS and A/S companies in the period 2013-2017, of which 7.909 were defaulted. The test is done, first, by analyzing the financial and non-financial data separately using logistic regression. Subsequently, the datasets are combined in order to create a full model and investigate whether this is more accurate than the two separately. Furthermore, the paper takes a critical stance to the financial ratios used and how these can, and are being, manipulated and how this may affect the models.

The principal findings in the paper clearly shows that the addition of the non-financial data adds significant accuracy. The combined model reaches an AUC score of 0,921 and log score of -0,1665 which is better than the strictly financial model with AUC of 0,876 and log score of -0,2164 and the strictly non-financial model with AUC of 0,698 and log score -0,2416. In addition, it is superior in any of the common measurements of model fit; log likelihood, log score, R^2's and classification ability.

Though the analysis showed that the addition of non-financial data technically improved a corporate default prediction model, the extraction, modelling and analysis of the data was so complex that the usefulness of the non-financial data in the Central Business Register is limited to those with very high technical abilities and computational capacities.

Acknowledgements

We want to dedicate this page to the people that helped with the creation of this thesis.

Benjamin Christoffersen, our supervisor and Ph.D. fellow at Copenhagen Business School, who tirelessly helped guide us in the right direction and navigate complex problems.

Kasper Jønsson, student and employee at Copenhagen Business School Department of Finance, who helped with extracting the data that is the basis of this thesis.

The Danish Business Register for putting us in contact with the right people and for giving us access to the central register.

Danish Center for Big Data Analytics Driven Innovation (DABAI) for helping us make sense of the data in the CVR register and for being the inspiration for this thesis.

Morten Fammé Hviid-Hansen, Head of Business Intelligence & Data Science at Deloitte, for granting us access to the ACE super computer at Deloitte.

Andreas Olczyk

Malthe Dyvekær Nybjerg

Copenhagen Business School, May 15th, 2018

Contents

1 Introduction

There are major benefits associated with the release of government data to the general public as open data. Everyone in society, including the government itself, benefits hereof. Several studies (e.g. (Hardy & Maurushat, 2017)) have found three major benefits: 1) it increases effectiveness and efficiency of government services, 2) because the data that is the basis of government decisions are made public, transparent and accountability is increased and 3) it makes the country more democratic by facilitating a broader basis for citizen participation.

The World Bank (World Bank Group, 2017) has also described the use and benefits of access to big data. It has identified three areas where big data and public access to government data can be transformational – service delivery, policymaking and citizen engagement (World Bank Group, 2017). It further describes, that not only is easy access to public data useful in the before-mentioned spheres, it also enhances the benefits of integration of data; public with non-public (business-, private- and/or organizational data) As such, it enables a more holistic analysis to be conducted and for the investigation of public/non-public interdependencies and relations.

1.1 Denmark as a Digital Front Runner

Denmark ranks number one on the Digital Economy and Society Index (DESI) according to the EU (European Commision, 2017). The DESI is an index that quantifies connectivity, human capital, use of internet, integration of digital technology and digital public services. Having ranked number one is no coincidence, but rather the result of the Danish government prioritizing digitalization and digital integration. Since 2001 the government has worked determinedly to update existing systems to give them a "digital lift" and to create new integrated systems (Danish Business Authority, 2018). The government's vision is that Denmark should be a "**Digital Front Runner**" in order to take full advantage of new digital possibilities and to create digital growth and development in Denmark (ibid).

One of the areas that has been digitalized, and has been a large focus area, is the Central Business Register (CBR). In 2013, the Danish Business Authority created a digital solution that allowed for system-to-system access to the data stored in the central register, as well as digital versions of financial statements. This entails that it is e.g. no longer necessary to manually look-up individual

businesses' financial/non-financial information one-by-one at various websites. This can now be done directly in internal systems that link to the central register. To implement this access to data, companies need only to code the functionality into their internal systems (e.g. internal invoice processing systems).

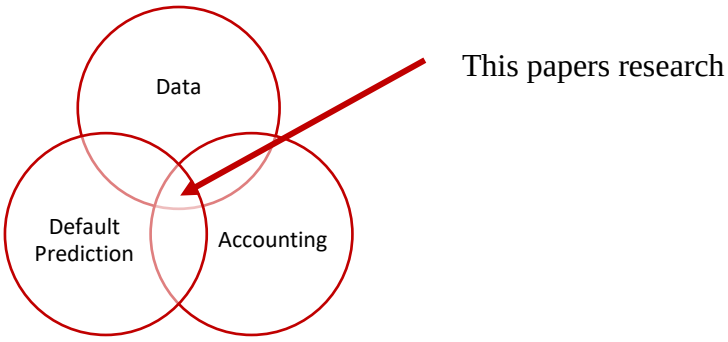
Access to massive amounts of register data also created the possibility for the government to analyze register data in a much more comprehensively and at an unprecedented scale. In an effort to make use of this new possibility the Danish Business Authority has established collaborations with the Danish Centre for Big Data Analytics Driven Innovation (DABAI), Copenhagen University, Aarhus University and Danish Technical University. One of the Danish Business Authority's main points of interest is to develop a tool that examines the data from the CBR and the financial statements for the companies, in order to create a model which is more accurate at predicting corporate default than traditional models. The aim is to identify companies in risk of default and intervene or prioritize industries that are more at risk of default; by investigating ***both*** financial and non-financial life patterns and events of companies.

A private application for this type of data and analysis could be within the banking sector. In order to determine if a company is eligible to receive a loan from a bank, the bank needs to assess how financially stable (risky) the lender is. That is, how big is the risk of default and therefore the risk of losing money on the loan. A model created using all publicly available data from the companies in a country would, all else equal, be more precise and more sensitive to country specific risk than traditional corporate default risk models that are created from foreign data.

1.2 The Area of Focus

In order to examine the abovementioned development and possibilities of open data, it is important to narrow the focus. Consequently, this paper focuses on the interactions between **data, accounting and default prediction**. Where data and accounting interacts are large scale analysis of firm characteristics, which can be used to generalize a population of companies. Between accounting and default prediction is the literature concerned with which accounting figures are important when predicting defaults. And between default prediction and data is purely statistical models that aims at building statistical models that can accurately, and statistically correct, predict corporate defaults. The interaction between data, accounting and default prediction is what is interesting; creating

statistically sound prediction models based on large scale analysis, but with a deeper understanding of the accounting figures.



Source: illustration by authors

1.3 Can CBR Data be Used to Predict Corporate Defaults?

While in theory an expansion of any data-set with relevant information leads to (at least some) enhanced strength of the corresponding prediction model, it is not necessarily the case for non-financial data. As such, this beg the questions; Can non-financial register data really be use to predict corporate defaults? Does the addition of non-financial register data implemented in standard default risk type models increase the precision of corporate default prediction? Is it possible to create a model that predicts corporate default with the integration of both non-financial register data and financial statement data?

There are several studies that have used large scale analysis of Danish financial data. But this data has been 'cleaned' beforehand by companies like Bisnode or Experian who charges a significant fee for delivering the data. But is it really that difficult to extract and process the publicly available free data?

This paper seeks to answer these questions by extracting the non-financial register - and financial statement data and analyze these using logistic regression. In order to analyze whether the addition of non-financial data has a significant positive effect on the precision of a prediction model, the Full Model (consisting of non-financial and financial data) is compared to the two individual models created to craft the Full Model (i.e. the Financial Model & the CBR Model).

1.4 Research Questions and Hypotheses

The aim of this paper is two folded. First, the paper investigates whether this newly accessible public data can be efficiently extracted and used to create a corporate default prediction model. Secondly, the paper analyze the effect of adding non-financial data to corporate default prediction models that uses financial data. This is approached by stating an overall research question. To answer this research question, a number of hypotheses are made, 4 of which are overall hypotheses aimed at answering the research question, and a number of sub-hypotheses stated for each variable used in the analyses.

The overall research question is therefore as follows:

"Does the addition of publicly available non-financial data from the Danish Central Business Register significantly improve corporate default prediction model accuracy for Danish Aps and A/S companies?"

Below are 4 main hypotheses that this paper seeks to investigate, stated in the order in which they are tested:

- **H1:** Financial data from the Danish Central Business Register can be used to predict corporate defaults.
- **H2:** Non-financial register data from the Danish Central Business Register can be used to predict corporate default.
- **H3:** The addition of publicly available non-financial data increases the accuracy of financially based default prediction models.
- **H4:** Financial data is superior to non-financial when predicting corporate defaults.

Hypothesis 1 states that the financial data that is in the Central Business Register can effectively be extracted, modeled, analyzed and used to build a corporate default prediction model. Proving that this hypothesis is true, hinges on whether the data can be effectively extracted, whether it contains the right data and whether the data is accurate.

Hypothesis 2 is the center of analyzing the effectiveness of the non-financial data. If the non-financial data in itself proves not to show distinction in the (non-financial) choices between default and non-default companies, then the data cannot be used to predict default. In such a case the data cannot improve a default prediction model, i.e. can the non-financial data too be extracted, modeled, analyzed and used to build a corporate default prediction model.

Hypothesis 3 is closely linked to the overall research question which this paper is seeking to investigate – whether the addition of this type of non-financial data can significantly improve default prediction.

Hypothesis 4 states that the non-financial data is not as effective at predicting corporate defaults as the financial data. Defaults are caused by a company not being able to repay its debts and the financial data is therefore hypothesized to be most effective.

1.5 Structure

This paper will first go through the literature surrounding corporate default prediction models. Then, in detail, explain the method of how we seek to build up our analysis and whether these model choices are sound. Variables used in the analysis are then selected, described and calculated and the raw datasets initially explored in order to explore the sample data's various trends and distributions. The analysis of the data consists of three parts; analysis of the financial data, analysis of the non-financial data and analysis of a model combining the two. Before the conclusion, a discussion is presented, aiming at incorporating an accounting-perspective of issues with utilizing financial accounting items (variables) in default prediction models. From the analyses a conclusion will be drawn, followed by a reflection of the usability of the findings, what this paper adds to the literature and the overall process of the papers analysis.

2 Literature Review

Over the past 100 years, massive amounts of research have been dedicated to analyzing and predicting corporate defaults. This research has moved its focus with the technological developments and has included both more data and more data types. Where early corporate default theories were centered on the attributes of the managers and workers, more recent literature is centered on financial statement-, ratio- and comparative analysis.

This section will review the literature and research about corporate default prediction, the various methods and variables used to build corporate default prediction models.

2.1 Statistical Techniques for Default Prediction

From a general perspective, the literature surrounding corporate default analysis and prediction can be grouped into the three types – **the classical** theories which is the models that has been the basis of most studies, **the modern** theories that include more modern statistical methods and techniques and take advantage of access to more data and **the alternative** studies that have tried to use alternative forms of data and methods, either by itself or together with classical or modern theories.

2.1.1 The Classical theories

2.1.1.1 Univariate Discriminate Analysis

Univariate discriminate analysis (UDA) is easily applied and interpreted. It offers a fast and simple way to analyze a single variable and has therefore been favored in studies examining the effect of a single independent variable.

Most noticeable of the UDA studies is **Beaver's** in **1966**. His study of paired matches of non-default and default firms, based on asset size and 3-diget SIC code, found several indicators that could be used to predict companies in risk of default up to five years in advance, with an accuracy of 78% (Beaver, 1966).

Other studies have used variables such as total liabilities over assets (Miller W. , 2009) and cash flow and return on assets (Bhargava, Dubelaar, & Scott, 1998). Even though UDA has proven its

effectiveness, much critique has been written about the method. Particularly the inconsistency and assumptions of linearity (see Keasey & Watson, 1991 and Amendola et al., 2006).

2.1.1.2 Multivariate Discriminant Analysis

The foundation for using multivariate discriminant analysis (MDA) was laid by **Altman in 1968**, **Deaking (1972), Edminister (1972) and Blum (1974)**. These studies used multiple accounting ratios to score companies' credit risk and determined default depending on a credit score threshold.

In **1968**, **Altman** published a paper in which he used multi factor (multivariate) discriminant analysis. Analyzing matched pairs of 33 default and 33 non-default US publicly traded firms. He developed a model, termed the "Z-score", which used 5 different accounting ratios to predict the probability of default (Altman, 1968). The paper demonstrated the advantage of having interacting variables in one analysis instead of analyzing variables one by one. The model has since then proven highly effective[1] in discriminating between default and non-default firms using financial ratios and has been used as the base-model for a large amount of studies since.

After Beaver and Altman, several studies tried to replicate and confirm or disprove their models. For instance, **Deakin** (**1972)** compared the works of Altman and Beaver using the same sample (Deakin 1972). He replicated Beavers study by using the same ratios and then searched for a linear combination of the 14 ratios which is used to "*devise a decision rule which will be validated over a cross-sectional sample of firms"* (ibid). The study showed that the discriminant analysis was the most effective of the two.

Shortly after **Edminister (1972)** published a paper, in which he analyzed smaller corporations and concluded that not all ratios and methods can be used to predict corporate defaults in these smaller corporations, but confirmed that some ratios did prove valuable in prediction models (Edminister, 1972). Furthermore, he recommended that corporate default models include a minimum 3 years of consecutive financial statements.

Blum (1974) analyzed the results and the sturdiness of the discriminant analysis. His sample set contained 115 failed firms paired with 115 operating firms based on asset size, industry, total sales

[1] The model proved that it was 95% accurate in predicting default within a year using the initial sample but only 79% accurate using the holdout sample.
(https://epublications.marquette.edu/cgi/viewcontent.cgi?article=1025&context=account_fac)

and number of employees. He was able to identify default firms with 94% accuracy within one year of default, 80% two years prior and 70% 3-5 years prior.

In **1977, Altman et al.** updated his z-score model (now termed the "Zeta model"). The paper compared linear and quadratic discriminant analysis and found these to have higher effectiveness of the original linear model using the original and holdout samples (Altman et al., 1977). They also introduced "prior probabilities of group memberships" and "cost of error estimation" into the classification model and then compared the performance of model with naïve corporate default classification strategies. They achieved great results with well over 90% accuracy for the hold out sample one year prior and 70 % accuracy up to five years prior.

The **1993 Altman's** paper updated the Z-score model again (now termed Z'-model). This time he adapted the model to include private firms. He did this by swapping market value of equity with book value of equity and recalculating all of the coefficients in the model.

The Z-score model was again modified by **Altman** in **1995** to non-manufacturing and emerging market firms (termed the Z"-model) (Altman, 1995). This version of the model minimized the potential industry effect by taking out the asset turnover ratio.

2.1.1.3 Conditional Probability Models

The conditional probability models (CPM) predicts default probability using the maximum likelihood estimator. The CPM includes the **linear probability** models that assume a linear probability of default, **probit models** that assumes normal distribution and **logistic regression models** (or logit models) that assumes a logistic distribution.

Studies using linear probability is somewhat limited, however the **1990** study by **Platt & Platt** in which they describe how the default probability varies by the same increment in response to equal change in the independent variable is an example (Platt & Platt 1990). However, as studies (e.g. **Aziz & Dar 2006)** have shown, the assumptions of linear probability is unrealistic and the dependent variable is arbitrarily distributed which decreases the predictive effectiveness.

Lennox examined the causes of corporate defaults by using a sample of 949 UK companies in the period from 1987-1994 (Lennox 1999). He identified the most important determinants for corporate default as "*profitability, leverage, cash flow, company size, industry sector and the economic cycle*" (ibid). He proved that cash flow and leverage have significant "non-linear effects" and adjusting for these effects can increase the predictive power. Moreover, he argues that logistic regression and probit

models can identify failing companies more accurately than models using discriminant analysis (UDA and MDA).

Zmijewski used **probit models** to analyze 40 default and 800 non-default companies listed on the American and New York stock exchanges and created a prediction model that had an estimation accuracy of 99% (Zmijevski 1984). The model used net income over total assets, total liabilities over total assets and current assets over current liabilities as variables in the model. The study also identified two biases – "choice-based sample bias" and "sample selection bias" (ibid, p. 59). The first is caused by the one-to-one match of default and non-default companies, creating an oversample of default companies, and the second is caused by the differences between default probabilities for companies with complete data and companies with incomplete data (ibid, p. 74).

The most used conditional probability model is the **logistic regression model (LR)**. The first generation of the logistic regression models was pioneered by **Joseph Berkson** in **1944**, but the most noticeable study was done by **Ohlson** in **1980**. This study, which included 105 default and 2.058 non-default US industrial companies, was created as a critical response to Altmans 1968 model (Ohlson, 1980). He identified three central critique points: 1) the model relies too much on assumptions, 2) because the output is an ordinal ranking device it offers very little intuitive interpretation and 3) default and non-default firms are matched according to size and industry which tend to be somewhat random. He claims the variables should be included to predict default risk, not for matching purposes. His model, termed the O-score, showed high predictive ability and consisted of 9 variables whereof two were dummy variables.

The second generation of logistic regression models, called multi-period or dynamic logistic regression models, was first created by **Shumway** in **2001**. He developed a discrete hazard model with logistic assumptions to forecast corporate default risk (Shumway, 2001). He proved that the efficiency of multi-period logistic regression was superior to single-period logistic regression models because it takes time varying variables into account, can incorporate more data and is able to distinguish the default risk period. His model included both basic accounting variables as well as equity market variables. He noticeably showed the usefulness of some market drivers, such as the company's market, past stock return and the idiosyncratic standard deviation of stock returns, that has previously been neglected. Numerous studies have since used his model and framework, such as **Chava & Jarrow (2004)** and **Beaver et al. (2005)**.

Among the second generation models is also **Lykke et al.** in **2004**. This study, made on behalf of the Danish National Bank, uses 300.000 annual statements from Danish companies, whereof 8.000 are default, from the period 1995-1999. The study creates three models: a basic model, a sector model and a model based on the number of employees. All three models includes both accounting variables as wells as three dummy variables describing non-accounting information. These include the "Remark"-variable that describes "*if there is a critical auditor comment in the account*" (ibid), a dummy variable describing the corporation type and firm age.

2.1.1.4 Market Based Structure Models

Another approach to corporate default prediction is the **market based structure models** (MBS). First created by **Merton** in **1974**, his seminal structure model of default classify company as default when total assets are lower than total liabilities (Merton, 1974). The model views equity as a call option on the assets of the firm and assumes the strike price of the option equal to the face value of the liability. This framework is useful even when accounting policies changes and is not sample dependent.

Market based models have been recommended by e.g. **Hillegest et al. (2004)** and **Miller (2009)** with the arguments that the superior performance is due to the fact that it includes more information, namely asset volatility.

2.1.2 Modern Default Prediction Studies

Modern default prediction studies have used artificial intelligence (AI) and advances in computational and data processing systems to come up with models that take advantage of the huge leaps that has been made within AI.

2.1.2.5 Neural Networks

Neural Networks (NN) techniques are "*inspired by the way biological nervous systems, such as the brain, process information*" (Stergiou & Siganos, 1996). NN "*can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques*" (ibid). Examples of NN include **Jaymeen & Murtaza (2000)** who developed a multi-layered neural network that uses unsupervised learning to predict corporate default within the computer, computer manufacturing and associating industries, using 65 companies whereof 6 were default. The model had a success rate of 73% using three years worth of data from financial statements.

Brédart (2014) used a sample of 3.728 Belgian SMEs whereof 1.864 have gone default in-between 2002-2012. The paper displayed a model that used three different financial ratios which describe solvency, liquidity and profitability. The precision of the model is "more or less 80%" (ibid).

NN theory has been criticized primarily on the bases that its input and processing happens in a 'black box' which makes it impossible to see the importance of each variable (Altman & Saunders (1998), Peursem & Pratt (2002), Kumar & Ravi (2007)).

2.1.2.6 Rough Set

Rough Set (RS) was introduced by **Pawlak & Sowinski** in **1991** and has proven to be a powerful tool for analyzing data and discovering patterns. This is done by lowering the degree of precision in order to make the data patterns more visible. Rough sets does not require a particular functional form and does not have restrictive assumptions of distribution (of both variables and errors).

Ahn et al. (2000) used rough set theory to predict corporate default. They used financial data from 2.400 companies, whereof 1.200 were defaulted, from the period 1994-1997 to create a model that was comprised of 8 financial ratios. Their study proved that rough sets together with neural networks outperform both NN and MDA.

RS theory has been criticized due to the fact that its effectiveness depends largely on the dataset. It also has a tendency to be affected too much by data noise and problems with multimodality. Studies have also show that RS combined with NN is much more effective than the two separately (e.g. Ahn, Cho & Kim 2000).

2.1.2.7 Decision Tree

Decision Tree (DT) models forecast default by partitioning data into sub-classes recursively and then until a final node of the tree consists of two risk outcomes – default or non-default. It is essentially a series of if-then-statements that has the purpose of dividing a heterogeneous dataset into multiple homogenous datasets.

Aoki & Hosonuma (2004) created a model that used such techniques to analyze 156 companies whereof half were default. The model analyzed 184 financial ratios and chose five of these that were the most significant, where interest coverage ratio was the most important. They were able to classify corporate default correctly 91,3% of the time.

2.1 Variables Used to Predict Corporate Default

For the purpose of creating a default prediction model, it is interesting to examine the different types of variables that has been used to predict default. Competing models and theories claim that a specific type of variables are superior to others, while other take the stance that default risk cannot be explained by one type of variable alone, but rather by a combination of variables types. In general, there are three types of data used in default literature; ratios, market data and non-financial data.

2.1.1 Ratios

Arguably the most used type of variable in corporate default prediction literature is ratios. Ratios have the advantage of making the variables comparable across different sizes of firms and industries as well as being a relative measure as opposed to an absolute measure and thereby limiting the range of the variable. Accounting ratios also often serve as the basis of auditors' assessment of the 'going concern' evaluation.

However several problems are also related to using ratios in corporate default prediction and financial analysis in general. **Kristóf (2008)** best describes two problems that are ever present in ratio analysis – instances when one of the figures in the ratio is zero and instances where both figures are negative and therefore yielding a positive number.

Another problem is the industry effects that might be present. These effects was describes in the **Platt & Platt (1990)** article mentioned above. They describe how ratios may differ considerably across industries as different industries results in different capital structure, profitability etc. In order to compensate for these industry effects, Platt & Platt recommends dividing the ratios with the industry mean times 100. This approach to dealing with industry effects is not exploited very much in the literature, only a few studies make up the entire research on this area (like **Dewaelheyns-Van Hulle (2004), Hillegeist et al. (2004), Berg (2007)** for example)

When analyzing the financial ratios through a statistical framework, in order to create a corporate default prediction model for example, most models used in the literature rely on the independence of variables. This is however not always possible, as some ratios are naturally correlated. An example of this could be the current ratio and the quick ratio. These rely, almost, on the same input and would therefore most likely also be correlated. In order to test whether the data is free of correlation principal component analysis have been proposed by studies like **Li-Sun (2011) and Xiaosi (2011)** or

Studenmund (2006) who recommends using variance inflation factor (VIF) tests**.** These recommendations have however been criticized by **Wang (2004)** and **Huang et al. (2012).**

2.1.2 Market Data

Aharony et al. (1980) takes a different approach to analyzing a company's default risk. They have the believe that ratios analysis "*have little or no definitive theoretical foundation*" and that "*financial ratios are simply utilized in various statistical procedures until they do, in fact, work*." (ibid). Their study, which used the firm- and industry specific variance (and risk) as model input, showed that there was a significant difference in the behavior of certain risk measures based on market data up to three years before default.

Similar to **Aharony, Jones & Swary (1980)** is the study by **Clark & Weinstein (1983)**. Their study used market based stock returns instead of variance measures but came to the same conclusion, that there was significant difference between default and non-default firms up to three years leading up defaults.

2.1.3 Non-Financial Data

Only a few of the abovementioned research articles have included non-financial measures or dummy variables in their research. Some studies have however successfully used and proven the effectiveness of this type of data. **Lykke et al. (2004)**, mentioned above, included three dummy variables ("*remarks*", corporation type and age).

Grunert et al. (2005) also pointed out that the literature around using non-financial measures for default prediction is ambiguous. They, like **Günther & Grüning's (2000)**, used two non-financial measures in their model. One for "*management quality*" and one for "*market position*" and proved their explanatory power by applying them to a model that predicted defaults of lenders in German banks.

In one of the largest studies, **Altman et al. (2010)** used a range of non-financial variables that they classified within 4 categories, "*type and sector*", size, age, "*reporting and compliance*" and "*operational risk*". Their study on 5,8 mil SMEs proved that when adding non-financial measures to default prediction models, the accuracy increased with up to 13%. They were also successful at creating a model that predicted corporate default for large amounts of companies with limited financial information.

Pervan & Kuvek (2013) also used "*management quality*" as a variable in their model along with "*quality of accounting information*", age, number of employees, "*dependence on key customers*" and firm owners personal credit rating. They successfully show an increase in precision from 82,8% to 88,1% by adding the non-financial variables to a model consisting solely of financial variables and applying it to clients of a Croatian commercial bank.

Stenbäk (2013) took another approach and used macroeconomic factors in his research. He used "*gross national income", "industry volume", "interest rate", "consumption",* and *"consumer confidence on economy*" (ibid, pg. 19). Most of the macroeconomic variables proved to be significant to his model and that the addition of the non-financial measures to a financial model increased its precision.

2.2 Summary

There has obviously been a significant development in the way corporate default is analyzed and predicted. Where the first literature was very subjective in nature and focused more on the performance and capabilities of the managers, the ground work for modern corporate default prediction was focused on a few very basic financial variables. As statistical frameworks developed, more and more complex financial variables were included and analyzed across a selection of very advanced techniques. However, there has been a "back to basics" trend in recent literature. From using very precise and specific accounting variables and ratios, to include non-financial variables also. Variables like age, number of employees and accounting quality has made their entrance into default literature and has been proven effective. This trend is very interesting, yet is still in its early stages of exploration due to the fact historical non-financial data is scares which makes it difficult to back-test models.

3 Method: Data Collection and Analytical Framework

The method by which the following results have been obtained can be divided into two sections; 1) data collection and analytical framework and 2) data modelling and handling. This chapter will address both processes chronologically, in order to provide an in-depth understanding of the work that has been done and the choices and decisions made. Then, in the following chapter determine which implications these decisions have on the analyses and the following conclusion.

3.1 Data Collection

3.1.1 General Source of Public Business Data: Central Business Register

The CBR is a Register with current and historical data for all current and previously existing businesses in Denmark. Since 2001, the CBR has been the legal entity responsible for registering and storing fundamental business data, used by the government, companies and the general public. The data in the register is submitted both by the business owners themselves and different government branches (e.g. the Danish Business Authority).

The reg register istry has undergone a drastic transition during the past 20 years. From data scattered across multiple government entities on microfilm, to a central digital register. But in 2013, as part of the government's "*2016-20 Digital Strategy*" (Agency for Digitalization, 2016), the CBR took the next step in the digitalization process and created a solution that allowed for system-to-system integration and access to all CBR data. This means that users of the data can now use the information in their own systems as well as export information. The information available is both financial statement data **(Financial Data)** as well as non-financial business data such as address, number of employees, industry code, type of corporation, etc. **(Register Data).**

3.1.2 Technical Export of CBR Data

Export of the CBR data is done via a so-called Application Programming Interface (API). An API is essentially a connection between two software programs, that allows them to 'talk' together. The CBR has opened up their systems in order to allow for the data to be accessed using an API. To access this data and be able to create an API it is necessary to register with, and be granted access by, the Danish Business Authority.

When an API is set-up and extraction commence, it is quickly discovered that the data provided is not all standardized and user-friendly, i.e. extracted easily in e.g. Excel, albeit this is probably already known, as that is the general set-up when one gains access to unmodified database data. The Register Data comes in a JSON file, a file format easily read by computers, and the Financial Data in a XBRL format; "*the open international standard for digital business reporting*" (XBRG.org, 2018). Upon extracting these files, it is necessary to modify them in order to use the data.

Due to the nature and size of the data, the data processing was done on a 24-core, 60 GB ram 'supercomputer' graciously made available by Deloitte Consulting for this project. This was necessary as ordinary personal computers does not have the computational or storage capacity to deal with data of this magnitude[2].

3.1.3 Technical Transformation of Extract to Useful Datasets

3.1.3.1 Transformation of Register Data from CBR

The output from the CBR for the Register Data come in one JSON file pr. CVR number (the company identification number). Each JSON file has the same layout with the same variables, even if the company has no data for the variable.

However, while it may be possible when extracting Register Data to filter for A/S and ApS (this research's scope) the only way to get the data is to write a "loop"-code, which simply entails that, even though you "filter" for specific data, the code will go through each and every CVR number to check the chosen variable. Upon finalizing the Register Data extract, we could conclude that the total amount of CVR numbers checked in the CBR was approx. 5 million, of which approx. 500.000 were either A/S or ApS companies, thus producing an equally large number of individual JSON files. In order to make those files useful for any analytical purpose, all files would have to be transformed into

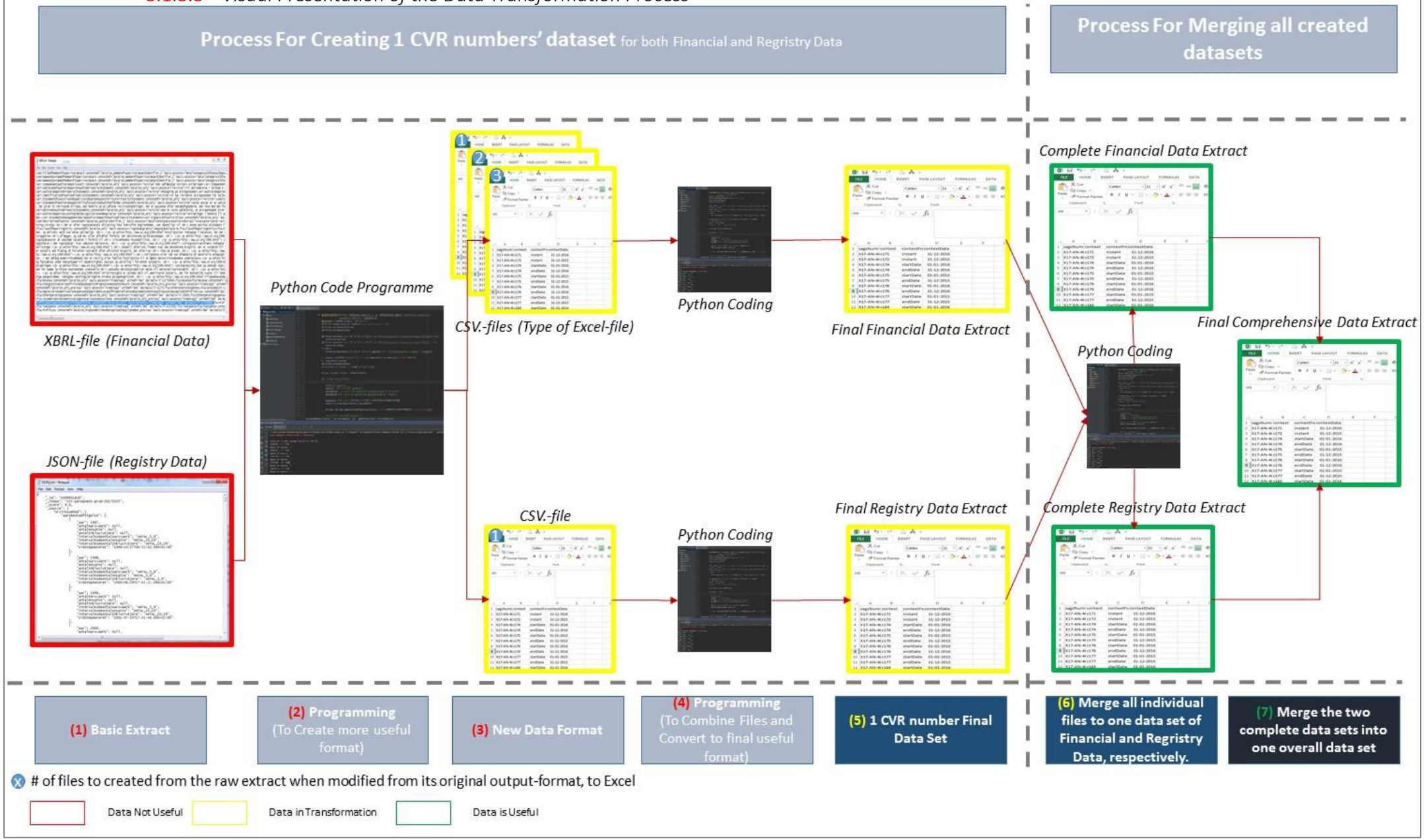
[2] Upwards of 1,5m files / 250 GB of raw data

one complete Excel-file. This was done through programming, utilizing the possibilities of Python (a general purpose coding program) and its `pandas` and `glob` packages.

3.1.3.2 Transformation of Financial Data from CBR

The same issue is present when extracting financial output. However, in this instance there can be only one filtering variable; the CVR number. Consequently it is necessary to create a list of CVR numbers for those companies which one wants to investigate prior to extraction. Also, differentiating from the Register Data extraction and transformation process, is that the output cannot be converted via Python into a single data-file for each CVR number, but must be converted into three different, relating data-files for each CVR numbers, using Python. Of course this complicates the process further, as those files must be then be combined before combining all individual "master"-files into a final dataset. The process for data transformation for both datasets are visualized below, using real-life images of the layout. Notice, the first part of the process exhibit the process of creating one dataset for one CVR-number, as a result, this is the process which must be replicated approx. 500.000 times.

3.1.3.3 *Visual Presentation of the Data Transformation Process*



3.2 Analytical Framework

As it is this paper's aim to test whether or not non-financial publicly available data enhances the strength of prediction when examining probability of default, it is necessary to create a base, in the form of a statistical model, upon which the non-financial data can be added. If the right model is selected, it would enable the creation of a base-model; in this case a model which is based on Financial Data only. This model is then to be extended to incorporate non-financial data. The three models' results would then be somewhat comparable and you might also compare either model to other models or standard within the sphere of the topic.

3.2.4 The Statistical Framework

In order to analyze whether the non-financial data contributes positively to the accuracy of corporate default prediction models, a statistical framework needs to be chosen. This framework has to 1) be able analyze multiple variables in the same model, 2) allow for analysis of different types of variables (continuous, categorical, binary etc.) and 3) be in line with what statistical frameworks that has previously been used in corporate default literature.

Looking at what statistical models that has previously been used, as described in the literature review chapter, multiple different statistical methods fulfill all three points. For example: **MDA** has been widely used in corporate default literature (for instance in the famous "z-score" by Altman). This model is easily applicable and can analyze not only the significance of each of the variables used in the analysis, but also how these variables interact with each other. However, there are several problems with MDA. One major flaw is that the outcome cannot be forced to be in the range between 0 and 1 which is needed in order to evaluate probability. Instead, MDA outcome can take any real number (also negative). For this reason, Altman has had to define three "*zones*" for the output of the model, in order to classify whether a company is expected to go default, maybe go default or probably survive (Altman, 1968). Furthermore, there are also problems relating to heteroscedasticity and normality.

A model that is able to accommodate all three points as well as deliver an outcome within the 0 to 1 range and does not rely on the assumptions of normality and homoscedasticity is **LR.** LR estimates the probability of an event occurring given a set of explanatory variables that can be several different types of variables.

As an alternative to LR, neural networks, rough sets and decision trees can be used. However, these do not allow for the interpretation of the independent variables or deliver an outcome in the 0 to 1 range.

Therefore, LR was chose as the statistical framework to be used in this study. The section below will describe LR in brief and outline the assumptions as well as how we intend to use LR in our analyses.

3.2.5 Logistic Regression

Both logistic regression and discriminant analysis can be used to calculate the categorical probability of an event given a number of categorical or continuous variables. One significant difference between the two statistical frameworks is, however, that discriminant analysis assumes all variables in the model to be normally distributed. Because some of the variables in the analyses are categorical, this assumption cannot be fulfilled. This is the reason that LR is recommended when working with these types of variables (see e.g. Sharma, 1996).

As described above, LR estimates the probability, π_I, of a binary response variable, Y, taking the value of 0 or 1 given the set of explanatory variables x_1, x_2, …, x_k.

The basic model for LR with multiple independent variables can be formulated as:

$$P(Y = 0|x) = P = \frac{e^{\beta' x}}{1+e^{\beta' x}} \quad 1)$$

$$P(Y = 1|x) = 1 - P = \frac{1}{1+e^{\beta' x}} \quad 2)$$

β' being a vector of the coefficient $(\beta_0, \beta_1, \ldots \beta_n)$ of the explanatory variables $(x_{0,} x_{1,} \ldots x_n)$. The two equations above is the equivalent of:

$$ln\left(\frac{P}{1-P}\right) = \beta' X =: l \quad 3)$$

l being the logistic function of the probability p.

The expression $\frac{1}{1+e^{\beta' x}}$ is what is called the "Sigmoid" (or logit) function. This function "forces" the value of $\theta^T x$ (or more generally, any real value) into the 0 to 1 range such that $h_\theta(x)$ can be interpreted as probability. The goal of LR is to find a value for θ that is large when observation *x* is in the group with the event (default) occurring and small when the event is not occurring (non-default).

Equations 1 and 2 can also be written as:

$$P(Y = y_i) = P_i^{1-yi}(1 - P_i)^{yi} \quad 4)$$

P_i being the probability of default of the i^{th} observation and y_i the random variable Y (that assumes either 1 or 0).

LR relies on the maximum likelihood method that maximizes the function *L*:

$$L = \prod_{i=1}^{n} P_i^{1-yi}(1 - P_i)^{yi} \quad 5)$$

LR also relies on some basic assumptions[3]:

I. The dependent variable should be measured on a dichotomous scale (0 or 1)
II. The observations should be independent and the dependent variable should have "*mutually exclusive and exhaustive categories"* (Lærd Statistics, 2018).
III. The relationship between any continuous independent variables and the logistic transformation of the dependent variable should be linear.
IV. There should be no high intercorrelations (multicollinearity) in the independent variables.
V. There should be no (or very few) strongly influential outliers.

3.2.5.1 Evaluating and Comparing Model Performance

The different LR models will be evaluated and compared on a number of parameters. **Firstly**, the models are evaluated on how well they model the data, which is notoriously difficult to do with LR models without understanding the datasets in depth. Most model fit estimates that are known from other linear models, like linear regression (R^2 etc.), cannot be directly used for LR. There are however several ways to compare different LR models. The way this papers evaluates model *performance* is by looking at the **chi-square distribution, the classification table, the -2 Log Likelihood (-2LL), the average LogScore (LS) and the Receiver Operating Characteristics (ROC) curve and the related Area Under the Curve level (AUC level)**.

The **chi-square distribution** in the 'Omnibus Test of Model Coefficients' will tell if the variables in the model adds explanatory power compared to a model consisting of only a constant (the intercept). If there is no, or very little, significant explanatory power in the variables, the addition of these will be deemed not significant.

[3] In reality there are several other assumptions for LR however these are not easily tested and not as important

The **LS** is a 'proper scoring rule'. *"Scoring rules provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes*" (Gneiting & Raftery, 2007). In essence, the LS rewards predictions that are close to the actual outcome and punish those far from the actual outcome. It can be written as follows:

$$\text{Individual LS} = log(\hat{p}_{i,t} * a + (1-a) * (1-\hat{p}_{i,t}) \quad 6)$$

$$\text{Average LS} = \bar{x} = \frac{\sum_{i=1}^{n} log(\hat{p}_{i,t} * a + (1-a) * (1-\hat{p}_{i,t})}{n-1} \quad 7)$$

Where $\hat{p}_{i,t}$ is the predicted probability $\hat{p}$ for company *i* at time *t* and *a* is the actual event outcome – 1 if default and 0 if non-default. This means that if the model correctly classifies a default then the score is the log of the predicted probability $\hat{p}$, whereas if the model classifies a default incorrectly then the score is the log of $1-\hat{p}$.

The motivation for using the LS instead of a simple deviance-from-the-actual measure is that the LS takes into account both whether the classification was correct and thereby 'punished' both wrong classifications and estimated probabilities.

The **ROC** curve plots the true positive rate (sensitivity) against the false positivity rate (1-specificity) for different cutoff points. The closer the curve is to the upper left corner, and to AUC level of 1,0, the better the model. This measure can also easily be compared across models of the same dataset but also different datasets.

The **Classification Table** is a confusion matrix showing the number of correct and incorrect classifications by the model. This tells something about the precision of the model, based on the number of correct guesses given the threshold (cutoff point). This threshold is by default set to 0,50, meaning that any probability above 0,50 will be classified as a default (1) and any below as a non-default (0). This threshold is also what will be used in the analyses.

Comments will also be made for the **pseudo R^2 estimates**; the Cox & Snell R^2 and Nagelkerke R^2. These are 'artificial' R^2 estimates that tries to do the same as R^2 in linear models and tell something of how well the model explains the data. Both pseudo R^2's has the approach of trying to calculate the improvement from null model to fitted model. The **Cox & Snell R^2** is essentially a transformed likelihood ratio test and can be written as:

$$\textbf{Cox \& Snell } \mathbf{R^2} = 1 - \left\{\frac{L(M_{intercept})}{L(M_{full})}\right\}^{\frac{2}{N}} \quad 8)$$

Very similar to the Cox & Snell R^2 is the **Nagelkerke R^2**. This is basically the same estimate except it has been adjusted such that the range of possible outcome values for the estimate is between 0-1. This is not the case with Cox & Snell R^2 as a perfect fit would result in an estimate of $1 - L(M_{full})^{\frac{2}{N}}$ which is less than one. To extend this range, the Naglekerne R^2 divides the Cox & Snell model with $1 - L(M_{full})^{\frac{2}{N}}$ such that he full estimate is:

$$\textbf{Nagelkerke } \mathbf{R^2} = \frac{1-\left\{\frac{L(M_{intercept})}{L(M_{full})}\right\}^{\frac{2}{N}}}{1-L(M_{intercept})^{\frac{2}{N}}} \quad 9)$$

Secondly, the significance, based on the Wald test, and coefficient estimates are evaluated against the sub-hypotheses made in chapter 5. This evaluation determines whether each variable behaves in the way that was hypothesized. That is, are the variables negatively or positively correlated with default risk, and whether they are significant to default prediction.

Thirdly, the models -2LL, average LS and AUC level estimates is used to compare the different models and how these compare to each other. This makes it possible to determine whether the addition of the non-financial data to the financial data adds any significant explanatory power.

3.2.5.2 Applying Logistic Regression

Before the LR can be performed, the raw dataset needs to be analyzed in order to determine if it lives up to the assumptions of the LR. Variables not living up to the assumptions are excluded from the further analyses.

The initialization dataset is then analyzed using IBM's statistical software 'SPSS'. The LR model will be made using the 'Enter' method. 'Forward' and 'Backwards' selection methods were also considered however as the aim is to test every variable and their significance the 'Enter' method was chosen.

To test the **first assumption**, of the data falling into a dichotomous scale, a simple histogram of the 'event' variable is made. If some variables does not fall within the 0 or 1 category they are excluded from the analyses.

To test the **second assumption,** of independence between the observations in the model, a simple rational applies. The observations are considered to be independent from each other as the different observations are from different companies.

The **third assumption,** which states that there must be a linear relationship between the independent variable and the log odds, is tested using a scatterplot for each variable. This scatterplot will have the standardized (Pearson) residuals on the Y-axis and the independent variable on the X-axis. Any obvious pattern in the residuals would indicate a non-linear relationship between the two.

The **fourth assumption** of no high inter-correlations is tested using a Variable Inflation Factor (VIF) statistics score. This score will tell if there are some variables that are any variables that are closely correlated. The threshold VIF score used to potentially exclude variables is set at 10. This level is both a common rule of thumb but also used in previous literature (e.g. Hair et al. 1995)

The **fifth assumption** deals with influential outliers; these should be removed. To deal with potential outliers the raw data is capped in such a way that the 10 % lowest values are set to the 10 % lowest value and the 10 % highest values are set to the 10 % highest value. More on this cap of data in Chapter 3.

3.3 Data Modelling and Handling

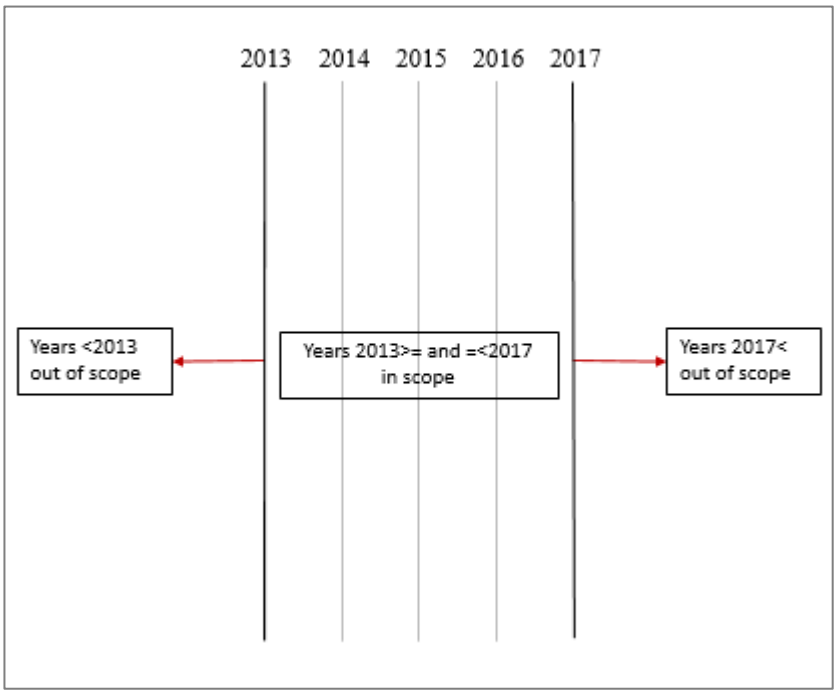
Having discussed in the above how to get and analyze data, this section will focus on *how to deal with data prior to and post analyses*. It is the aim to convey how data was handled in practice, what choices were made and how these have influenced the models' interpretation.

3.3.1 The Aim of Data Modelling

The aim of data modelling is very simplistic and easily understood, yet infinitely more difficult to execute. The aim of data modelling to produce a dataset – or several datasets – which is structured in such a way that allows for it to be used in LR, i.e. it must be structured to fit the input-format used in programs like SPSS, R and STATA. Essentially the format is straight forward; the data must be structured such that one row represents one observation, which holds one or multiple variables as well as one event-variable (in our case 1 = Default, 0 = Not Default). Remember, as described earlier, the reason for employing LR was partly due to the fact that our event-variable is binary. Further, a secondary, albeit just as important aim of data modelling, is to calculate (create) those variables which will constitute the collective set of variables and as such observations.

3.3.2 Understanding How the Extracts Were Made and Fitted to Scope

In order to understand what has been done to the raw data and how, the starting point and thus outlook of the raw data must be understood. Remember that it was necessary to create a list of CVR-numbers within the scope prior to extracting Financial Data – this was not the case for Register Data as it was possible to be filtered differently when extracting. Consequently Register Data allowed for the production of a coherent list of all A/S and ApS CVR-numbers, as well as to state their operating status for every year of their lifetime (as well as all else Register Data of course). Hence, it was necessary to extract Register Data first. Upon extracting, a list of in-scope CVR-numbers was created by applying various filters to filter out those companies which start/end-date rendered them out of scope. As a result, a list of 250.000 in-scope CVR-numbers was created, which could be used to extract Financial Data. Graphically the scope is illustrated as such:



Source: illustration by authors

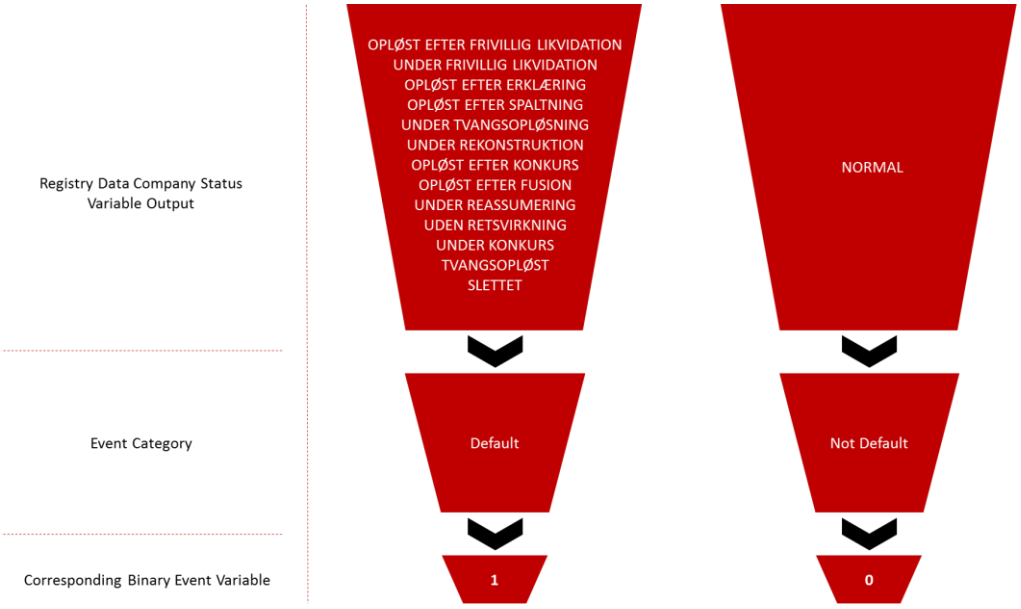
Once extracted and technically merged into two individual datasets, both would hold the raw data for every company within the scope of the research; remember, the scope was defined as all Danish companies with the legal status A/S and ApS. Further, we limited the scope to focus on a 5 year period (from 2013-2017) which was primarily due to the fact that XBRL-data was not accessible prior to 2013.

Thus, depending on the number of variables within the two datasets, one observation would (ideally) exist pr. year pr. variable; for Financial Data that equals approx. 26,25m variable-observations scattered across approx. 1 million rows and numerous columns (**21 variables** (accounting figures) x **5yrs** x **250.000** CVR-numbers). The same figure for Register Data was approx. 12,5m (**10 variables** x **5yrs** x **250.000** CVR-numbers). A total of 38,75m variable-observations.

3.3.3 Definition and Creation of Event Data

As mentioned in the above section, it is absolutely essential to create an event variable, which means it is a necessity to create a dataset which holds that information. Often, when working with binary events it is no problem to quickly generate such data, as the raw data may already contain it. Unfortunately, this is not the case for our data. Regardless, in any scenario the starting point is to define what an event is. Hence, as the general focus in this paper is predictability of default, it is no surprise that the actual default itself becomes the event. As such we should create our binary event variable to be 1 = default (event occurs) and 0 = no default occurs (absent of event, i.e. business is operating). Thus, creating the needed dataset should be unproblematic considering the Register Dataset already holds the information about company status (default or not); you would simply need to filter the data. However, filtering was made difficult by the large array of possible company statuses relating to default. Statuses were not, it turned out, binary in nature (default/not default). Rather, there were statuses which represented a degree of default, financial distress or some other transition phase such as "*UDEN RETSVIRKNING, UNDER FRIVILLIG LIKVIDATION, UNDER REKONSTRUKTION, UNDER KONKURS, UNDER TVANGSOPLØSNING, OPLØST EFTER KONKURS, TVANGSOPLØST, OPLØST EFTER FRIVILLIG LIKVIDATION, OPLØST EFTER ERKLÆRING, UNDER REASSUMERING, SLETTET, OPLØST EFTER FUSION, OPLØST EFTER SPALTNING*" (Danish Business Register, 2018). Hence, in order to continue it became critical to determine and define the 'event' variable in a manner which accommodated the plurality of possible statuses.

Therefore, an event is labelled 'Default' and encompass *any status other than "NORMAL" for which it can be argued that the company in question is default or in default proceedings*. In other words, for every year within the timeframe, it is analyzed what status a company has, and if it is not "NORMAL" but one of the other above, then an event occur. This enables the binary variable of 1 or 0, regardless of a non-binary set of event variables. This can be illustrated as:



Source: illustration by authors

When this has been defined it is possible to lookup the various statuses for each year for each CVR-number and return either 1 or 0. The final product is an overview of all selected company's 'lifecycle' reduced to a binary code. See below for an example:

CVR #	2013	2014	2015	2016	2017
Xxxxxxx1	0	0	0	1	1
Xxxxxxx2	0	0	0	0	0
Xxxxxxx250.000	0	1	1	1	1

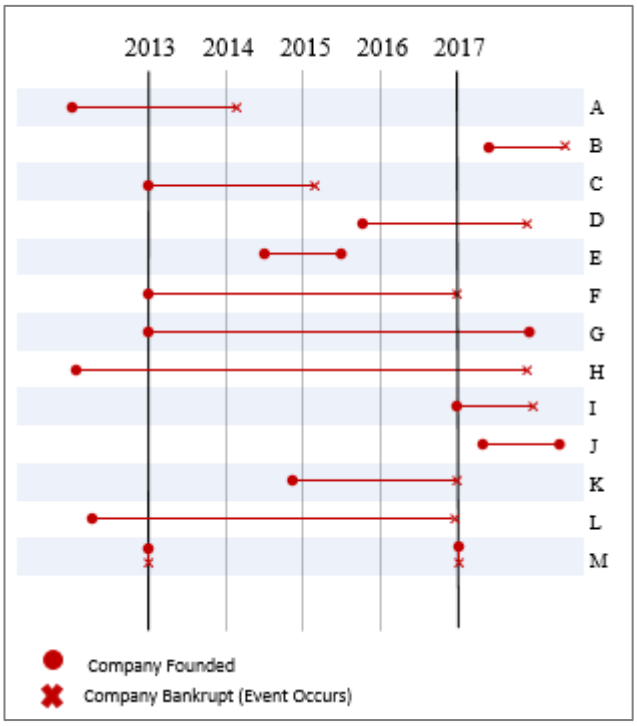
Source: table by authors

CVR1: in normal operation in 2013-2015, until an event (default) occurs in 2016. The subsequent years are given a 1 to indicate that no operation is taking place. This is also done to facilitate calculations in Excel.

CVR2: Fully operational within the period, no event.

CVR250.000: in normal operation for 1 year until event (default) occurs in 2014. The subsequent years are given a 1 to indicate that no operation is taking place. This is also done to facilitate calculations in Excel.

Based on the above, it is fair to state that there exists an almost endless amount of lifecycle scenarios. See a list below (non-comprehensive):



Source: illustration by authors

However, while it is of some interest, it is not the main focus to analyze the scenarios, but merely to create observations and data based on the usage of these scenarios' information about company lifecycle. In sum, a dataset exhibiting pr. CVR-number pr. year status has been produced enabling us to investigate whether or not a company is default or operating in a given year.

The theoretical need for this data has been explained above, yet it remains to be explicitly stated how it is used in practice. This will be discussed later in this chapter.

3.3.4 Creation of Observations and Final Output

Having successfully extracted data and created an event dataset, it becomes possible to start building the final output, which consists of x-number of rows, each amounting to one observation which is made up by several variables (columns) with different values.

3.3.4.1 Difficulties Prior to Calculation

Using historical data

Although this may seem fairly doable, there are some difficulties to be overcome. As a starting point, consider what it is we wish to accomplish; to predict an outcome. By utilizing LR, we use historical data in order to find the most fitting model, yet the usage of historical data must be tied to the defined event occurrence. Thus, when having located an event, we must produce variable-observations based

on historical data, but which historical data (year) to use? Does it matter? In short, "YES", it matters. Essentially, as long as the approach is consistent, the choice of how many years prior to event occurrence data is retrieved from, only affects the interpretation of the final model; e.g. "this model can predict default 1 year before it occurs based on data from x-years prior".

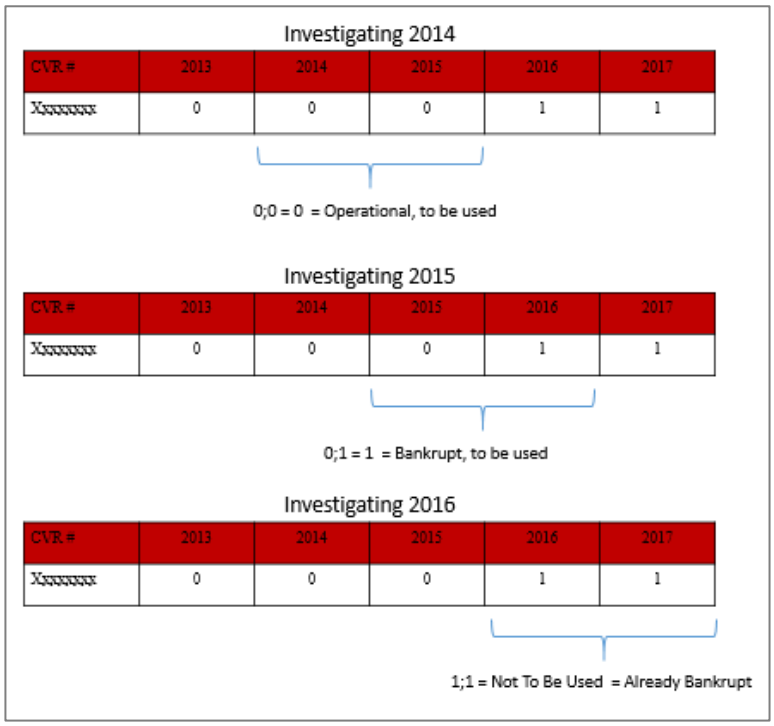
There can be several reasons as to why different researchers have chosen different time-lags. Some want to (dis-)prove some specific hypothesis, others simply due to data availability. At the outset, our research focused on a 1-year lag (if an event occurred in 2014, we would utilize 2013 figures for variable value calculations). However, by trial-and-error we settled on a 2-year lag. The main reason being the availability of data. It became apparent that companies often did not report financial statements in the year prior to default or in any case only a partial amount of the complete data. As a consequence we operate with a 2-year lag. Lykke et al. (2014) in a similar research states that *"[o]n average it takes 19 months from the accounting year of the last account until a failure is announced*". This is also the reason why this study has not extended its time-frame (scope) to include 2018, as that would increase the *"risk of accounts having a wrong response variable, as an apparently active company could in fact have failed. In addition, due to the time lag it is difficult to specify the exact timing of a failure"* (Ibid).

When calculating values, we produce calculations for all variables, for all companies, in those years were it is possible, considering the time-lag (2014-2016). Thus, it is important to create a system which will make sure that the observations (rows) are given the correct binary event value (1, 0) which correspond to the time-lag. Hence, we revisit the event dataset made earlier in the process. We use this dataset to lookup each CVR-number in each year to find its current status. However, in practice this is not quite enough. Recall that a company lifecycle could look like this:

CVR #	2013	2014	2015	2016	2017
Xxxxxxxx	0	0	0	1	1

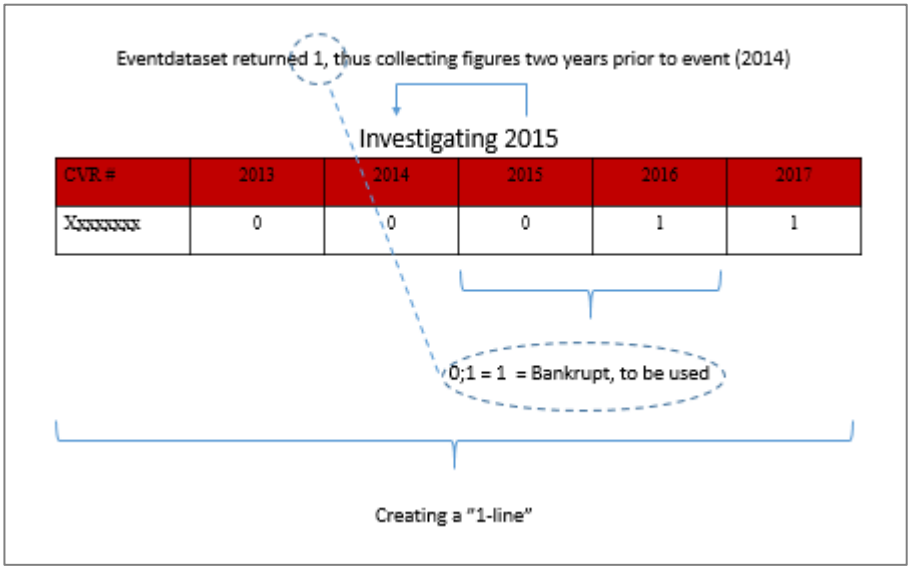
In such a case, you would get a 1 (event occurs) in 2016 and 2017. However, it is only true for 2016. Thus, a system had to be created that made sure that a 1-line in 2017 would not be used/calculated. Hence, by using Excel's many options and formulas, various IF-statements secured that this would not happen. More importantly to understand is, that what is wanted is to know the points in time at which a company is operational prior to the event, as this enable further analysis. Now, before calculation would take place, our models would check if the year in question returned a 1 or 0. If returning a 0 it would move on to check if the following year returns a 1 or 0. If it returning a 1, then

it can be concluded that the company will go default next year, but is currently in operation. The below tables should illustrate the various scenarios:



Source: illustration by authors

Thus, with such a set-up, it is ensured that we can utilize the same framework for each year and constantly ensure that we produce either '0 or 1-lines' (observations) which is based on information two years prior to the event. See the example below:



Source: illustration by authors

As a final note to this issue, it should be mentioned that this format excludes calculations for 2013 as the data (Register- and Financial Data) which is to be used lies outside the scope. Likewise, the event data which is to be used for 2017 lies outside the scope, and thus renders calculations out of scope. Also, the choice to be consistent in the use of a two-year lag has a direct impact on the interpretation of the model. This will be revisited later in the chapter.

Missing Values

Another issue which must be addressed is whether or not data is missing in the datasets. Of course, ideally no data would be missing, yet in most cases – especially with large datasets – that is not the case. It is important to deal with missing data, as the reason why it is missing and how it is handled can affect the result, usage (extrapolation) and interpretation of the model. Data can be missing in one of three ways (Allison, 2002):

Missing Completely At Random (MCAR)

If data is missing completely at random, then the reason that $X_{(i)}$ is missing does not depend on Y, nor on $X_{(-i)}$. i.e.:

$$X_{(i)} \perp\!\!\!\perp (X_{(-i)}, Y)$$

In other words, the data is missing completely at random and there are no explanatory/dependent variable to explain its absence.

Missing At Random (MAR)

If data is missing at random, then the reason that $X_{(i)}$ is missing does not depend on Y, but depends on $X_{(-i)}$. i.e.:

$$X_{(i)} \perp\!\!\!\perp (Y)$$

Missing Not At Random (MNAR)

If data is not missing at random, then the reason that $X_{(i)}$ is missing depends on Y and $X_{(-i)}$, i.e.:

$$X_{(i)} \not\perp\!\!\!\perp (X_{(-i)}, Y)$$

Considering the large datasets with which this research is engaged and the number of variables, it is possible to argue for the presence of all of the above types of missing data in some form or another. Although an initial investigation of missing data in the Financial Dataset, based on Little's MCAR test (Garson, 2015), revealed an insignificant p-value which entails that we fail to reject the Little's null-hypotheses, thus supporting the assumption that data is in fact MCAR. However, it would be

incorrect to assume the absence of either of the other types of missing data; e.g. MAR is very likely to be present as several of the accounts used for the Financial Data's variables' calculation is sub-totals – and thus in nature dependent on another value ($X_{(-i)}$). In terms of the Register Data, no values were found to be missing. This has to do with the fact that the database is very much up-to-date and that we extract almost only binary variables; has the company changed address within year x, Yes (1) /No (0)? Nevertheless, it could be argued that it is possible that the database do not exhibit the true reality completely; it only conveys what has been reported. That is, a company may have changed its address physically, but not yet within the Central Business Register. However, this is extremely difficult, if not impossible, to test and in any case the data is not per se missing – we do have a variable value.

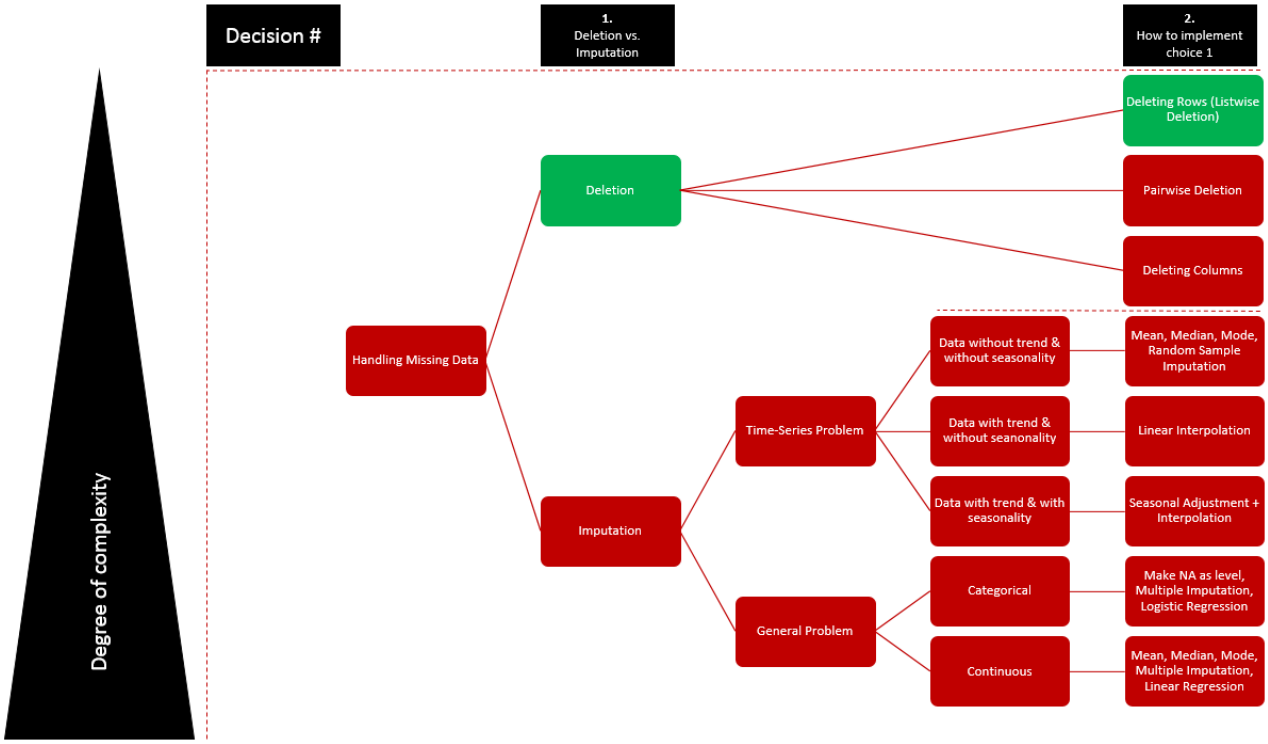
As for the Financial Data, the reason as to why data is MCAR/MAR, is most likely found in the way in which figures are reported and entered into the XBRL-database. The database consists of approx. 4.900 different accounts for reporting Profit/Loss and Balance-sheet items; a rough estimate would render 50-250 accounts sufficient depending on the level of aggregation. Further, there is no account-mapping (guide) declaring the level of aggregation or differentiating sub-total accounts from single item accounts. Nor are there any official, useful guide for how to use database output for analytical purposes in general. This of course is problematic; which accounts to choose when there exists approx. 10-20 different accounts all relating to e.g. depreciation? Moreover, a quick analysis of IFRS's policies for entering data revealed a surprisingly un-standardized process – company owners (or their accountants/auditors) may freely choose the account which they deem most appropriate. Only general rules for entering values exists – rather surprising for an organization which purpose is to standardize company financial data – granted, it may be rather difficult to standardize considering the many formats and rules of individual countries/regions (GAAP differences). This means, in effect, that the cost related to e.g. depreciation, which should go to the same account for all companies, may be entered on various account depending on who is performing the entry.

As a result we made an extract of 5.000 random in-scope companies prior to extracting the complete Financial Dataset, in which we had to choose which accounts one wants. Subsequently we tested which accounts of those available in IFRS, related to an overall accounting line-item, e.g. depreciation, had the most entries and chose that account, if also its description did not violate the accounting definition of such an account (i.e. for calculation purposes it is important that we do not use e.g. IFRS account "Depreciation Property Plant And Equipment" as simply Depreciation, simply

because that account had the most entries – in this specific case an analyst would be unable to confirm by the name of the account, that this is in fact the total depreciation amount).

As it, at this point, it is clear how the data is missing and to some degree why it is missing, focus turn to how to deal with it. There are countless possibilities of how to deal with missing data, yet most (or all) fall within the categories "Imputation" and "Deletion" (Garson, 2015). With imputation, the aim is to 'fix' the dataset by utilizing the known variables' values and create the most educated estimate to fill in where data is missing. With deletion, the aim is also to 'fix' the dataset, but instead of creating estimates to fill in where data is missing, data is 'simply' deleted all together, leaving only those observations which have data for all variables.

While there are disputing arguments, within the missing-data-literature, of best practice, consensus seems to be reached about the fact that 1) there is a need to actively deal with missing data 2) neither of the options (Imputation/Deletion) have been proven superior to the other on a general basis and 3) there is, in general, no choice which is not prone to create a bias. As a consequence there is no 'correct' choice. Thus, after investigation of the literature, the following decision-tree has been made, exhibiting our line of thought in terms of options towards dealing with missing data:



Source: illustration by authors

As the above decision-tree shows, Deletion is chosen as the main response to missing data. Predominantly due to its simplicity. In practice a Listwise Deletion (LD) was executed. LD is straight forward; delete all observations (rows) for which one or more variables' values are missing. Hence, upon calculating various ratios and values for different years in Excel, then those observations (rows) for which one or more ratio(s) or value(s) (variables) were missing, would be deleted all together. Essentially, if an observation has 20 of the 21 variables, then it would be rejected. In doing so, the datasets were 'fixed'. The specific calculations of ratios etc. and their relation to prediction of default, will be discussed later in Chapter 5. As a final note, the decision to use 'Deletion' accommodates the error-problems related to ratio calculations presented by Kristóf (2008), yet simultaneously creates a significant bias – this will be revisited in Chapter 4.

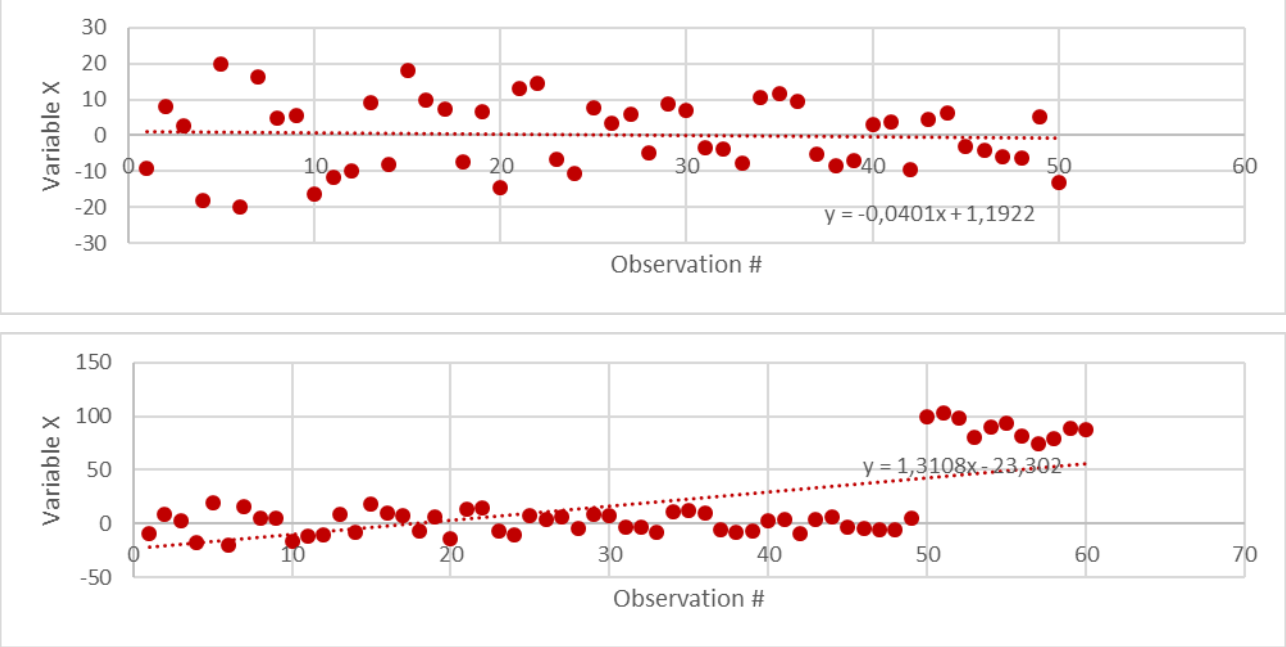
3.3.4.2 Difficulties After Variable Calculation

Outliers

In statistic, an outlier is an observation point which is distant from other observations. Yet "how distant" an observation should be to be categorized as an outlier remains a topic of discussion in the literature. It is more common to discuss methods for detection of outliers than a general rule of "distance" from the main data. Barnett and Lewis (1994) have greatly investigated what may be labeled as the "formal" techniques for detection outliers, among which Grubbs' test is one of them. Often said formal tests are engaged in some kind of mathematically proven method for detecting outliers such as hypotheses testing. However, such tests are often restricted in their use due to assumptions of e.g. the need for a univariate dataset or a normal distribution (Grubbs, 1950). In reality, data is not necessarily complying with such assumptions and thus renders those methods inadequate. Consequently, informal methods are more commonly applied in such cases. An informal method is characterized by any way in which an outlier is detected by means of some predetermined threshold set by the investigator. In other words, you may simply select some measure, based on e.g. mean, median, quantile or similar, and categorize any variable-observation above/below that threshold as an outlier.

The reason why detecting and dealing with outliers is important, in general, but especially when working with regression models, is that a single outlier may greatly affect the coefficient related to a variable in the model and thus alter the effect of that variable, resulting in a model which over/under-estimates. Consider the two scatterplots below; assume that we have built a regression model with one variable for which the trend-line (dotted line) is the regression line. Its function y = -0,04x + 1,19

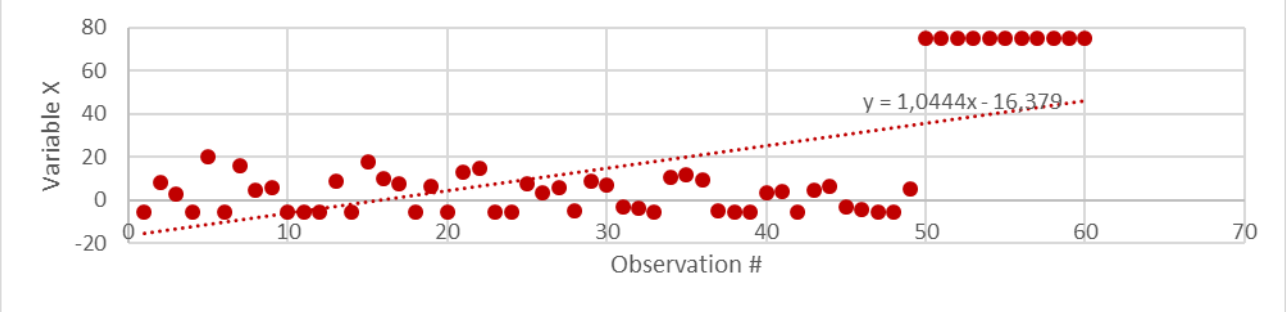
reveals that the effect of the variable is -0,04 (approx. 0). Now, examine the scatterplot below, it is based on exactly the same data, except 10 more observations are added - all are fairly 'distant' from all other observations. As a result, its function changes dramatically to y = 1,3x – 23,3. Thus, the effect of the variable just changed and would entail a different interpretation and result of the regression model.



Source: illustration by authors

Hence, utilizing an informal method for detecting outliers, our data revealed that outliers were present only in the Financial Dataset. Conducting a quantile-analysis, in which we divided observations into 10 equally large groups, enabled outliers to be detected by setting the threshold to the 1st and 9th quantile. Any value above/below is deemed an outlier.

Having detected the outliers it is important to ensure that they will not distort the general outcome of the model. To do so, one may employ one of a large variety of options. The most direct method is to delete outlier observations. However, this will create a large bias; the aim is to use all observations available, while ensuring that their effect is somewhat controlled. In doing so, observations are capped and their effect is somewhat controlled – in this example, any observation below the 1st quartile and above the 3rd quartile is capped at the quantile value, lowering the coefficient of the model by 0,3 to y = 1,0x – 16,3.



Source: illustration by authors

This procedure has been utilized consistently across all variables.

Shumway (2001) and Zmijewski (1984) both successfully utilized a similar detection/capping-method, albeit e.g. Shumwey (2001) used the 99th and 1st quantile as upper/lower limit capping 2% of his data, compared to this research's 20%. However, his analysis consisted of a considerable smaller dataset (observations) with smaller standard deviations and thus calls for the usage of a higher/lower quantile.

As a final note, financial institutions (banks etc.) have been excluded from this study all together, as they often tend to display significantly different financial ratios, due to asset size, and default characteristics, due to government bailouts, than other industries. This is 'normal procedure' for corporate default prediction studies, and seminal studies, like Altman (1968) and Beaver (1966), have also excluded these types of corporations.

Size of variables' values

Having capped all variables by the same method, next it was determined that all variables should – to some degree – operate within the same interval (sizes should be somewhat equal; all ratios range fairly within -/+10, yet those financial variables which were not calculated would simply be a given value – i.e. not a ratio – such as 4.884.993 for current assets. In order to properly 'down-size' these accounting figures, the natural logarithm was taken for each figure. However, as we operate with figures below and at 0, it was necessary to create a range of "IF"-formulas, to accommodate those calculation errors which would otherwise have occurred when taking the log of 0:

If variable > 0 then ln(variable)

If variable < 0 then ln(absolute value of variable)*-1

If variable = 0 then ln(variable+1)

The reason why the size of the non-ratio variables were changed were to accommodate SPSS's limit (settings) of only showing three decimals when calculating the value of the coefficients in the LR model. Thus, prior to 'down-sizing' the beta values for the non-ratio figures were presented as 0.00, yet in reality the true value were e.g. 0.00000000645. Hence, by 'down-sizing' the figures would yield coefficients within more or less the same interval, facilitating the use of SPSS and by extension interpretation of the models' outcome.

3.4 Summary

In short, data was extracted from publicly available databases, subsequently made useful (changing format to Excel) by programming in Python. Further, the analytical framework regarding LR were examined and explained. Hereinafter focus was turned to how data was handled after extraction with special focus given to dealing with missing data, outliers and variation in 'size'. At this point data is 'fixed'/clean and structured in a manner allowing for LR analysis to be conducted in SSPS. Both datasets consists of a variety of variables and a large number of observations. The specific variables and their relation to default prediction will be discussed in Chapter 5.

4 Model Validity Evaluation

This chapter aims at listing and investigating the potential implications, in terms of biases, that the applied method may cause to the interpretation of this paper's study and its conclusions drawn from it. Each bias will be investigated and explained how they each affect the analyses. Then it is outlined exactly what the resulting models and conclusions will actually describe and how they should be interpreted.

4.1 Biases

6 main biases have been identified upon examining the choices made throughout the process described in the chapter above:

1. Company Scope-limit Bias
2. Period Bias
3. Aggregation of Event Determining Variables Bias
4. Account-item Bias
5. Missing Data (Deletion) Bias
6. Time-lag of Model Bias

4.1.1 Company Scope-limit Bias

The decision to limit the scope in terms of company types create a direct bias. It is an unfair assumption that all other company types resembles that of an A/S or ApS in a manner which allows a direct extrapolation of the model. As such, the data extracted is biased towards A/S and ApS, ultimately depicting only a partial amount of the complete data (all data from all company types). In turn the model is only fairly extrapolated to other A/S and ApS companies and cannot be used for generalization for all corporation types.

4.1.2 Period Bias

This paper's research and by extension data collection is limited to a 5-year-scope due to the availability of financial data. A 5-year-scope decreases the likelihood of cyclical effects. However, to some degree, this limitation creates the assumption that all 5-year-scopes are somewhat alike in

terms of company data – development, growth etc. If this assumption is violated, it would cause a great bias when extrapolating the model (to other companies within the scope; A/S, ApS).

Consider the scenario; conducting the exact same model on data from 2004-2008 and 2013-2017. It is easy to imagine that the two models would not produce a similar output – of course no two models are ever similar in output, but if the assumption is to hold, then the result must to some degree be alike. Thus, the bias is somewhat unknown (untested) if one does not produce two different researches/models. However, whether or not a bias exists may be informally concluded by the analyst by assessing whether or not the scope-period is characterized by a large degree of extraordinary events (economic/political or the like).

It is assessed that the period 2013-2017 did not experience a significant number of extraordinary political/economic events – impacting the risk of default – compared to e.g. the period of 2004-2008. As such the period is deemed 'normal' and is less likely to hold any significant bias; yet it remains untested and it is important to state that a bias may exist.

4.1.3 Aggregation of Event Determining Variables Bias

Recall that companies could have different statuses over the course of its lifecycle. This created a problem when event data needed to be binary. Thus, to alleviate this issue, the statuses were aggregated to determine when a company was operational or default. However, the aggregation may have caused a bias. Essentially, the aggregation was quite strict, yet easily understood and implemented; if the company status was **anything else** than "NORMAL", then it was deemed default. Is it fair to assume that a company with the status "OPLØST EFTER FUSION" or "UNDER FRIVILLIG LIKVIDATION" is to be considered default? Perhaps not. Those statuses, and the related historical company data, may resemble the companies which have status "NORMAL". This effectively entails that it is likely that the event data is distorted. If so, a bias is likely to be created when creating (calculating) observations – it is likely that a '1-observation' (an observation which variables are indicators of how a company's financial health is prior to default) should, in fact, have been an '0-line'. The effect of the bias is to potentially 'improve'[4] the variables' values for some of those observations which are categorized as 'default', as it is fair to assume that a company which

[4] "Improve" is to be understood such that e.g. as a group, for those ratios which indicates a higher risk of default the higher the ratio, the average will be lowered as a company which is "wrongly" characterized as a 1-observation is likely to have a better (lower) ratio. Ultimately, these figures distort the trend in the data.

e.g. is "OPLØST EFTER FUSION" has not necessarily had the same financial development as those which are in fact "UNDER KONKURS".

The only way to alleviate this bias is to decrease sample size and only include those companies which have 'clean' statuses, i.e. 'default' and 'normal'. However, in order to increase the number of default-observation (1-line observations) the choice to aggregate statuses was made.

However, *this abovementioned bias is only an issue when thinking of this study as a strictly 'default' prediction model.* This model, in essence, predicts that a company enters any other state than normal and not only default.

4.1.4 Account-item Bias

Unlike other research papers within the sphere of default prediction, our data is based on extracts directly from the source and not on an intermediary's standardized and 'clean' data (e.g. Bisnode provides access to IFRS-data which have been formatted, standardized and cleaned). Both methods holds pros/cons, yet the former is often more difficult to use, but holds more possibilities to model data as wanted, as there is no interface. The difficulty became apparent in terms of using and mapping accounts. In the above method chapter it was explained that the consistency with which figures were reported on various accounts were surprisingly low. As a result, investigation of 'useful' accounts were explored. As a consequence it is fair to assume that those accounts which were chosen as the 'correct' accounts, do not necessarily show the entire reality. Thus, a bias is created based on the neglecting of those observations for which figures were 'incorrectly' reported, yet reported nevertheless with the correct value under a different account. The effect of the 'incorrect' reporting is an inability to calculate certain ratios. The effect of which is to create a model which is biased towards those companies which have reported their figures 'correctly', leaving out other companies which should have been included. An observant reader would notice that this relates somewhat to the above missing-data section in chapter 3. As the way in which companies reported their figures on different accounts is completely random, it supports the analysis that data is missing at random.

4.1.5 Missing Data (Deletion) Bias

The above account-reporting bias is, as mentioned, directly relatable to the missing-data issue. In response to missing-data, this research chose to delete all observations (rows) which did not have all variables – this was only the case for Financial Data. As a result, a bias is created. In effect, the model does not capture those companies which are 'overly distressed'. It might be that companies which do

not report some figures or none at all, are in fact those which are most interesting to study in relation to defaults. Yet these never enter the final dataset. It is the same issue as if one were to study the relationship between BMI and diabetes, to find what levels of BMI are likely to have diabetes. Then imagine that the researcher collects the data at the hospital. Everyone would show up for examination, except those which are inherently too overweight to go anywhere. In such a case the sample would be shewed and incomplete, lacking information about those which may hold the most information. This would effect that the final regression model's coefficients for predictability would be incorrect to some degree. In sum, the model is biased towards those companies which have reported every figure correctly and for which every ratio can be calculated. An observant reader would notice that this bias resembles what was previously referred to as "sample selection bias" by Zmijevski (1984), as noted in the literature review.

4.1.6 Time-lag of Model Bias

As it was exhibited in Chapter 3, a choice was made to consistently use data two years prior to event occurrence (default). This is perhaps somewhat unusual. Often analysts and researchers wish to work with the most recent data available. In the case of default prediction based on financial data, that would be the most recent annual report. One reason for this is that by using the latest available data, the model interprets as "predicting default one year from today, based on the most recent data", which strengthen the model in terms of extrapolation and trustworthiness. Further, it can be argued, that anything but the most recent data is uninteresting for anyone assessing the financial health of a company; no one cares how well/not well the company were doing 10 years ago, right?

Nevertheless, the choice was made due to two reasons 1) it simplified the Excel modelling of data. With the large datasets with which we worked, entering another layer of formulas/coding to find the latest/most recent annual report and select those figures, is very demanding and process heavy. Thus, choosing to use a consistent time-lag will ease the model and calculation time. 2) In choosing to use a consistent time-lag, it was first examined to use a 1-year lag. This, however, provided too few useful observations - especially in terms of default-observations - as many of those companies had not reported full annual statements the year prior to default. Eventually a 2-year time-lag was chosen.

4.2 What the Model Predicts

Given the information in this and the preceding section, it is important to establish exactly what this models predicts. The outcome of this model is strictly not a probability of default, but:

a probability of entering any other state than normal.

From this point onwards, for simplicity reasons, the term default will be the overall term for:

a company entering any other state than normal

Because the model uses a 2 year time lag between the default actually happening and the data used, the model predicts the default 2 years after the data, that is analyzed, is from. In essence, the predictive outcome can be written as follows:

Probability of default today = $\hat{P}\left(Y_{i,t} \mid X_{i,t-2}\right)$ 10)

or

Probability two years from today = $\hat{P}\left(Y_{i,t+2} \mid X_{i,t}\right)$ 11)

$\hat{P}$ being the predicted probability of default, $Y_{i,t}$ being

Equation 10 describes that the probability of a company going default today is given by the data from two years prior. Alternatively, the model can predict the probability of default two years from today given the data from today using equation 11.

Using conditional probability calculations it is possible to calculate the probability of going default within a 2 year period:

Probability of default *within* 2 years = $\frac{\hat{P}\left(Y_{i,t} \mid X_{i,t-2}\right)+\hat{P}\left(Y_{i,t+2} \mid X_{i,t}\right)}{2}$ 12)

Overall this means that the model predict:

the probability of Danish ApS and A/S companies of entering any other state than normal two years after the date of the analyzed data.

5 Feature Engineering and Dataset Exploration

This section will firstly describe how the raw data was transformed, as well as which variables were extracted to be used in the analysis in the next section. Then initially explore the two datasets in order to make sense of what is in the data, and whether there is any distinction between default and non-default companies in the raw data. This is meant as a first test of the data quality.

5.1 Financial Model Variables

5.1.1 Variables and their Usage

When assessing a corporation's probability of default, you are faced with an overwhelming amount of necessary choices among which are defining scope and limitations, choose a (prediction) model, identify assessment variables and choose how to handle data (e.g. structure, method, how to deal with missing/incomplete data). Individually, and combined, all these areas of choices have a direct impact on the result as well as the way in which the findings are interpreted.

Nevertheless, while no two studies are completely alike - neither in result or findings – all must concern themselves with identifying a single, or more commonly, a set of variables which outcome (observations) when studied enables the analyst to infer some conclusion of the relationship between variables' outcome and the event of interest (e.g. default) and thus by extension utilize such knowledge to predict a future outcome. However, the strength of any prediction model is not only to choose any variables, but the right (explanatory) variables – meaning those with the highest relative significance for predicting the event.

However, as it is virtually impossible to test all possible variables, a selection has to be made – such selection may occur naturally based on one's access to data and limitations, but may also occur by what may be referred to as 'educated guessing'. Essentially, one may possess some knowledge about cause and effect, theory and/or previous studies of the investigated event, all which enables a selection

of variables based on basic, pure rationality and information, consequently producing to the model most trustworthy and likely to succeed.

While this paper's aim is not (necessarily) to produce a prediction model which outperforms existing models, but rather to test the value of a certain type of data when predicting default, one may question the importance of the selection of variables which are not of same data type - after all we simply want to investigate if certain data (non-financial public data) have significant positive effect in a prediction model and as such simply need to test significance levels against other type of data, correct? No, it is of great importance that the financial variables are selected carefully as these essentially provides a basis for inter-variable comparison.

Hence, to enable the most trustworthy conclusion, the benchmark for comparison ought to be set as high as possible. Luckily, as it was presented in the literature review, the academic sphere has for quite some time been engaged in credit assessment or prediction/probability of default and thus has produced numerous variables which may be utilized in this relation. It is the rationale behind our chosen financial variables which this section aims to address.

Firstly, parties most interested in predicting default and conduct credit assessments are, naturally, those whom stand to incur a (financial) loss – i.e. stakeholders. As a financial loss can only be mitigated by an equal and opposite gain, it is natural that stakeholders are, primarily but not exclusively, interested in estimating 1) the probability of default and 2) in the event that a corporation goes default, how large is the 'gain'. Thus, in general, this effectuate an interest in a corporation's ability to drive/sustain ordinary operations, its liquidity, strength of cash flow, debt liabilities and value of assets. As such, we have identified several financial variables which contribute to assessing the economic health of a corporation. As explained above these have been chosen based on rational thinking rooted in the idea of causality.

Below is an overview of our financial variables and a brief explanation of the rationale behind using them. A sub-hypothesis is made for each variable and which will be confirmed or rejected in the analysis chapters.

5.1.1.1 Current Ratio

Formula:

$$\frac{Current\ Assets}{Current\ Liabilities}$$

Description: The Current Ratio explains how well a company is equipped to cover its most immediate liabilities. As such, the ratio can be used to make a rough estimation of a company's financial health and is of interest to stakeholders because it also gives an idea of how well a company can remain efficient / not incur any liquidity problems in the nearest future. The higher the ratio the better.

Hypothesis 5: Current Ratio has a significant *negative* correlation with default risk

5.1.1.2 Quick Ratio

Formula:

$$\frac{Current\ assets - inventories}{Current\ liabilities}$$

or

$$\frac{(cash\ and\ equivalents\ +\ marketable\ securities\ +\ accounts\ receivable)}{current\ liabilities}$$

Description: The Quick Ratio is, similarly to the Current Ratio, a liquidity ratio. It is also concerned with much the same balance sheet figures and as such can be interpreted in a relatively likewise manner. The main difference, however, between the two ratios are the fact that the Quick Ratio takes the "current"-aspect a step further. Meaning that it only concerns itself with those figures within current assets which quickly can be made/already are liquid. This entails that e.g. inventory is disregarded. As such, the Quick Ratio is often lower than the Current Ratio, as the total amount of assets to cover current liabilities are now lower in comparison. However, as an upside, as a stakeholder, using the Quick Ratio, you get a more conservative/trustworthy idea of what assets are actually almost 100% useful – here and now – to cover current liabilities; think, in the extreme event, what if you were not able to sell the inventory ever? Then the Current Ratio would be quite misleading if those assets' value made up a significant amount of total current assets.

Hypothesis 6: Quick Ratio has a significant *negative* correlation with default risk

5.1.1.3 Cash Ratio

Formula:

$$\frac{Cash\ and\ cash\ equivalents}{Current\ liabilities}$$

Description: If calling the Quick Ratio "conservative" compared to the Current Ratio, then the Cash Ratio is "extremely conservative". Essentially the general idea for the ratio is the same as the two other. However, as the name entails, this ratio is only concerned with how much cash a company has to cover its current liabilities. Thus, are the company able to pay off its current liabilities with the most liquid asset of all; cash. Of course, if this ratio is high (>1), a stakeholder may be "more at ease" about the company's financial health.

Hypothesis 7: Cash Ratio has a significant *negative* correlation with default risk.

5.1.1.4 Net Working Capital

Formula:

$$Current\ Assets - Current\ Liabilities$$

Description: Net Working Capital holds, in many ways, the same indication value as the Current Ratio. It is however not a ratio, but an absolute sum. It indicates the ability of a company to service it short term debt with its short-term assets. Thus, the sum that is left, if any, can be added to a Free Cash Flow. Indicating, that the sum of 'free cash', which can be utilized to increase business performance. This is also why a too high ratio (>2) is commonly not preferred, as it indicates an unused potential of investing in further development/growth.

Hypothesis 8: Net Working Capital has a significant *negative* relationship with default risk.

5.1.1.5 Current Assets / Total Assets

Formula:

$$\frac{Current\ Assets}{Total\ Assets}$$

Description: Current Assets are those assets which are to be utilized within one year. Total assets are of course the sum of all tangible and intangible assets, regardless of time-perspective. As Current Assets is a function of total assets, then the ratio is naturally limited to the interval of 0-1. Thus, a ratio close to 1, indicates that the company intend to use most of its assets during the next year. Thus, one may by simple rationality argue, that the higher the ratio, the more likely it is that the company

can quickly liquidate their assets if experiencing liquidity issues. This would be preferable to any stakeholder, unless the business otherwise can alleviate the liquidity problem. In some regard, this ratio is similar to the Current Ratio, only it is not concerned with liabilities, but with the level of total assets which can be easily be liquidated.

Hypothesis 9: Current Assets / Total Assets has a significant *negative* relationship with default risk.

5.1.1.6 Coefficient of Financial Stability

Formula:

$$\frac{Non\ current\ assets}{Equity + non\ current\ liabilities}$$

Description: This ratio is fairly similar to the Current Ratio, only now focus is turned to 'non-current', instead of the immediate future. Thus, we analyze if a company is suited to pay its non-current liabilities (that is any liability that extends beyond 1 year). The rational of incorporating this ratio is to make sure that all aspects (not only the present) financial health of the company is accessed. Also, it incorporates the value of equity as a liability. Equity, although it is often discussed, can have various explanations as to what it covers; e.g. the terms is used differently when one discuss the equity one may hold when owning a real estate property compared to the equity one holds when owning a stock ('owners' or stakeholders' equity'). However, let's denote equity as stakeholders' equity. In that case, when a firm goes default, if the owners of that company (equity owners), are to walk away without debt, then the company must be able to pay off all its debt on its own. However, if we view equity owners not as owners per se, but as creditors, then their stake in the company of course is viewed as debt and should be included. This is what is incorporated in the above equation. If the ratio is above 1, then the company is able to serve long-term debt as well as to cover owners' equity.

Hypothesis 10: Coefficient of Financial Stability has a significant *negative* relationship with default risk.

5.1.1.7 Return on Assets

Formula:

$$\frac{Net\ Profit}{Total\ Assets}$$

Description: The above ratios and figures are all in some way related to the liquidity of the company. Turning the focus, Return On Assets (ROA) looks at the profitability of a company. The ratio conveys

how much (free) capital the company generates when investing 1 unit of capital; in other words, how well it utilizes its assets. This is of course of great interest to any stakeholder, as it is an indication of whether or not a company is efficient in creating a profit or not, which in turn ought to be negatively correlated with the probability of default. While one may argue that the equation as such does not incorporate liabilities (service of liabilities and equity) this is not entirely true; Total assets = Total liabilities + equity. Also, within the calculation of Net profit one accounts for interest expenses (service of debt) and as such the measure do to some extend incorporate this, which of course is a further indication that if ROA > 0, then cost of debt is accounted for in some extend and thus liabilities are as well. However, as a measure, ROA may not be preferable across industries because the level of assets vary substantially. If one wanted the ROA measure to concern itself 'purely' with the operational profitability of the company, one may add back interest expense to Net Profit. In any case, it is fair to assume that the higher the ROA, the less likely the company is to go default.

Hypothesis 11: Return on Assets has a significant *negative* relationship with default risk.

5.1.1.8 Return on Equity

Formula:

$$\frac{Net\ Profit}{Equity}$$

Description: Return On Equity (ROE) is quite similar to ROA, although it is better at cross-industry comparison and now turns focus towards the amount of capital injected into the firm by stakeholders. The higher the ratio, the better and more efficient the company is at making use of those funds, which by extension reveals how efficient the company's operations are. Essentially a positive ROE conveys the cash (or level of cash) generated for a somewhat 'free' use and thus is a determinant of future growth and development. As such, the higher the ratio, the more a company is earning on its invested capital and by extension the more it is able to cover it financial future which makes it less likely to go default.

Hypothesis 12: Return on Equity has a significant *negative* relationship with default risk.

5.1.1.9 *Indebtedness Factor*

Formula:

$$\frac{Total\ Liabilities}{Retained\ earnings + depreciations}$$

Description: While this ratio is not necessarily a common one, it addresses a central point in securing continuous operation. Companies which makes a profit can utilize that profit in various ways; it can pay it out to stakeholders as dividends, retain it within the business to facilitate growth, or both. Thus, all else equal, a company which retains 100% of earnings are better suited to service future short- and long-term debt (should the need arise) and thus by extension, better suited to sustain its operation, ultimately making it less likely to go default. Let's assume that a company is 100% debt financed. In this case, all else equal, that company's probability to go default increases if it does not built a sizable 'buffer' (cash) to service interest expense or negative equity (cover a loss). This is what the above ratio is concerned with measuring in terms of the firms total debt/liabilities. The reason for adding depreciation to retained earnings is because it is an accounting figure and thus does not relate to the actual cash flow of the company. That is, depreciation is an accounting cost which decreases the Profit and thus the earnings amount which can be retained. Hence, to give a more realistic view the amount is added back.

Hypothesis 13: Indebtedness Factor has a significant *positive* relationship with default risk.

5.1.1.10 *EBITDA / Total Liabilities*

Formula:

$$\frac{EBITDA}{Total\ Liabilities}$$

Description: EBITDA (Earnings Before Interest Tax Depreciation and Amortization) is a common measure used to assess a company's operational profitability. It is the amount left to cover interest, tax, and amortization and ultimately produce a profit/loss. The figure is often used by analysts because it does not incorporate account-costs such as depreciation/impairments and as such is a better starting point for analyzing whether or not the operating activities (the core business) is profitable. It is also often used as a starting point to calculate the Free Cash Flow, again because it is quite cash-centric. The EBITDA/Total Liabilities ratio essentially conveys how much the operating profit can cover of the company's total liabilities at this point. If one took the inverse relation, one would be able to estimate how many years, assuming the same level of profitability (EBITDA) and no new

liabilities, it would take the company to pay off all its liabilities. It is of course interesting for a stakeholder to know these figures, as a company, which has a difficulties covering it total liabilities with its operating profit in a reasonable time, all else equal, is more likely to default.

Hypothesis 14: EBITDA / Total Liabilities has a significant *negative* relationship with default risk.

5.1.1.11 EBIT / Total Liabilities

Formula:

$$\frac{EBIT}{Total\ Liabilities}$$

Description: EBIT (Earnings Before Interest and Tax) and the workings of the ratio has all the same contributions as the above EBITDA/Total Liabilities. The only difference is that EBIT considers accounting-costs such as depreciation and as a consequence does not provide a completely fair/true estimate of the operating profit/loss which may be utilized to cover interest and tax and produce a profit. The difference between EBIT and EBITDA is, as the name suggest, not incredible radical. However, in some cases the difference lies within their respective usefulness for analysts etc. Often both are a fair starting point to calculate Free Cash Flow, which is what analysts is most often concerned with to be fair. However, choosing which one to use for that purpose may very well depend on the industry. As said, EBITDA is essentially cash-centric and thus a fair measure for Free Cash Flow, but only if the company in question is not capital intensive which one must subtract to get to Free Cash Flow, in such a case, it may be better to start from EBIT. Moreover, some analyst often find that EBITDA is harder to find as depreciations and amortizations may be included in various parts of the Profit/Loss statement.

Hypothesis 15: EBIT / Total Liabilities has a significant *negative* relationship with default risk.

5.1.1.12 Shareholder Equity Ratio

Formula:

$$\frac{Total\ Shareholders'Equity}{Total\ Assets}$$

Description: The shareholder equity ratio determines how much shareholders would receive in the event of a company-wide liquidation. The ratio, expressed as a percentage, is calculated by dividing total shareholders' equity by total assets of the firm, and it represents the amount of assets on which shareholders have a residual claim. While the ratio may only have little, if any, indication on a

company's financial health, it may hold great value to a stakeholder or potential investor etc. as the risk of entering into the business is somewhat mitigated in the event of a default. Also, if one knows that Total Assets = Total Liabilities + Total Equity, then the shareholder equity ratio -1 indicates how much of the company's assets are financed by debt/borrowings. As such the figure may indirectly have some predicting ability of default, if one submit to the understanding that the more leveraged a company is, the more likely it is to go default.

Hypothesis 16: Shareholders Equity Ratio has a significant *negative* relationship with default risk.

5.1.1.13 Liabilities / Total Asset

Formula:

$$\frac{Total\ Liabilites}{Total\ Assets}$$

Description: This ratio tells the inverse relation of the Shareholder Equity Ratio. It examines how much of a company's assets are financed by debt/borrowings. Thus, a high ratio would indicate a high leverage. To be leveraged is not specifically a problem for companies – almost all have debt – but if a company becomes over leveraged, then it may use a lot or all of its free cash simply to pay off interests and loans. Also, the company, all else equal, is doing business based on the mercy of the lender. In sum, to be highly leveraged it not preferable, yet debt is not an issue if the business can sustain it. In any case, it is fair to assume that a company with a high degree of debt/borrowings are more likely to go default than one that is debt free – without liabilities.

Hypothesis 17: Total Liabilities / Total Assets has a significant *positive* relationship with default risk.

5.1.1.14 Coverage Ratio

Formula:

$$\frac{Equity}{Non\ current\ assets}$$

Description: This ratio examines the equity's coverage of the company's Non-current assets. As for the Shareholders' Equity Ratio, it examines how much of the (non-current) assets the shareholder is entitled to, should the company default. Also, it tells a bit about a company's capital structure, as a high ratio would indicate a low level of leverage for its Non-current assets.

Hypothesis 18: The Coverage Ratio has a significant *negative* relationship with default risk.

5.1.1.15 *EBITDA / Total Assets & EBIT / Total Assets*

Formula:

$$\frac{EBITDA}{Total\ Assets}$$

and

$$\frac{EBIT}{Total\ Assets}$$

Description: Both of these ratios measure a company's ability to utilize its assets efficiently. This allows the organization to see the relationship between its resources and its income, and it can provide a point of comparison to determine if an organization is using its assets more or less effectively than it had previously. The difference between the two ratios is the difference between EBITDA and EBIT. The higher the ratios the more efficient the company is at utilizing its assets and the less likely it is that a company would incur a default.

Hypothesis 19: EBITDA / Total Assets and EBIT / Total Assets has a significant *negative* relationship with default.

5.1.1.16 *Level of selected figures*

In addition to the above ratios we have selected a range of different figures, primarily from the balance sheet, to investigate if the level of these in themselves have a predictive significance. The chosen figures are:

I. Current Assets
   - **Hypothesis 20**: Current Assets has a significant *negative* relationship with default risk

II. Current Liabilities
   - **Hypothesis 21**: Current Liabilities has a significant *positive* relationship with default risk

III. Cash and Cash Equivalents
   - **Hypothesis 22**: Cash and Cash Equivalents has a significant *negative* relationship with default risk

IV. Profit / Loss
   - **Hypothesis 23**: Profit / Loss has a significant *negative* relationship with default risk

V. Total Assets
   - **Hypothesis 24**: Total Assets has a significant *negative* relationship with default risk

VI. Total Equity
   - **Hypothesis 25**: Total Equity has a significant *negative* relationship with default risk

5.1.2 Overview Financial Model Variables

Below table shows the different financial variables, which category they are in (what they are measuring), their formula and our hypothesis of how these variables are expected to correlated with default risk.

Financial Model Variable Codes Description

Category of Measure	Variable Name	Variable Description	Expected Correlation With Default
Liquidity	**Current Ratio**	$\frac{\text{Current Assets}}{\text{Current Liabilities}}$	Negative
	Quick Ratio	$\frac{\text{Current Assets} - \text{inventories}}{\text{Current Liabilities}}$	Negative
	Cash Ratio	$\frac{\text{Cash}}{\text{Current Liabilities}}$	Negative
	Net Working Capital	$\text{Current Assets} - \text{Current Liabilities}$	Negative
	Current Assets to Total Assets	$\frac{\text{Short Term Assets}}{\text{Total Assets}}$	Negative
	Coefficient of Financial Stability	$\frac{\text{Non current Assets}}{\text{Equity} + \text{non current Liabilities}}$	Negative
	Indebtness Factor	$\frac{\text{Total Liabilities}}{\text{Retained Earnings} + \text{Depreciations}}$	Positive
	Shareholder Equity Ratio	$\frac{\text{Equity}}{\text{Total Assets}}$	Negative
	Liabilities to Total Assets	$\frac{\text{Total Liabilities}}{\text{Total Assets}}$	Positive
	Coverage Ratio	$\frac{\text{Equity}}{\text{Long Term Assets}}$	Negative
Profitability	**Profit / Loss**	Profit / Loss	Negative
	Return on Assets	$\frac{\text{Total Profit}}{\text{Total Assets}}$	Negative
	Return on Equity	$\frac{\text{Total Profit}}{\text{Total Equity}}$	Negative
	EBITDA to Assets	$\frac{\text{EBITDA}}{\text{Total Assets}}$	Negative
	EBITDA to Liabilities	$\frac{\text{EBITDA}}{\text{Total Liabilities}}$	Negative
	EBIT to Liabilities	$\frac{\text{EBIT}}{\text{Total Liabilities}}$	Negative
Size	**Total Assets**	Total Assets	Negative

	Equity	Total Equity	Negative
	Current Assets	Current Assets	Negative
	Current Liabilities	Current Liquidity	Positive
	Cash and Cash Equivalents	Cash and Cash Equivalents	Negative

Source: table by authors

5.2 Financial Model Data Exploration

Of the 154.838 total observations possible, the final sample size of the dataset that was used in the analysis consisted of 93.779 (≈ 61%) observations of which 7.909 were default and 85.870 were non-default - an approximate 8,5% / 91,5% split. These 93.779 observations include only the observations which had valid data in all of the variables as described above.

In the initial exploration of the Financial Data, significant differences between the default and non-default companies can be seen. Below is a chart showing all 22 variables, their averages, median as well as the 25%, 50% and 75% deciles and how much these differs (in pct.) between default and non-default companies.

	Average			Median			25%			50%			75%		
	Default	Non-Default	*Diff.*	Default	Non-Default	*Diff.*	Default	Non-Default	*Diff.*	Default	Non-Default	*Diff.*	Default	Non-Default	*Diff.*
Cash And Cash Equivalents	322.133	480.455	-33,0%	46.034	146.249	-68,5%	158	3.372	-95,3%	46.034	146.249	-68,5%	417.582	1.030.757	-59,5%
Cash Ratio	0,35	0,48	-26,8%	0,03	0,15	-77,9%	0,00	0,00	-100,0%	0,03	0,15	-77,9%	0,54	1,01	-46,6%
Coefficient of Financial Stability	0,43	0,47	-9,0%	0,37	0,40	-6,0%	0,02	0,12	-85,3%	0,37	0,40	-6,0%	0,93	0,87	6,3%
Coverage Ratio	1,72	2,54	-32,2%	1,05	1,70	-38,4%	0,03	0,52	-93,7%	1,05	1,70	-38,4%	3,36	4,44	-24,4%
Current Assets	2.287.191	2.809.630	-18,6%	852.852	1.407.543	-39,4%	161.016	247.307	-34,9%	852.852	1.407.543	-39,4%	3.716.936	5.653.733	-34,3%
Current Ratio	1,60	1,95	-17,8%	0,91	1,22	-25,4%	0,28	0,37	-23,6%	0,91	1,22	-25,4%	2,20	3,39	-34,9%
EBIT / Assets	0,00	0,05	-97,9%	0,00	0,03	-100,0%	-0,06	0,00	-	0,00	0,03	-100,0%	0,05	0,14	-64,6%
EBIT / Total Liabilities	0,00	0,06	-96,5%	0,00	0,03	-100,0%	-0,07	0,00	-	0,00	0,03	-100,0%	0,06	0,15	-60,8%
EBITDA / Assets	4.420	8.595	-48,6%	0	0	-100,0%	-89.753	0	0,0%	0	0	-100,0%	94.372	0	-
EBITDA / Total Liabilities	0,25	0,05	432,4%	0,52	-0,02	-2399,7%	-0,02	-0,02	0,0%	0,52	-0,02	-2399,7%	0,52	-0,02	-2399,7%
Indebtedness Factor	1,51	2,50	-39,9%	1,30	1,97	-34,0%	-2,10	1,03	-303,9%	1,30	1,97	-34,0%	4,58	5,01	-8,6%
Net Working Capital	62.643	678.860	-90,8%	-4.000	149.465	-102,7%	-810.973	-514.104	57,7%	-4.000	149.465	-102,7%	734.630	2.114.992	-65,3%
Profit / Loss	227.657	1.166.673	-80,5%	0	454.149	-100,0%	-370.653	-44.029	741,8%	0	454.149	-100,0%	637.926	2.695.281	-76,3%
Quick Ratio	1,48	1,80	-17,6%	0,78	1,02	-23,7%	0,20	0,27	-25,2%	0,78	1,02	-23,7%	2,05	3,15	-34,8%
Return on Assets	0,00	0,13	-101,5%	0,00	0,08	-100,0%	-0,20	-0,02	1056,2%	0,00	0,08	-100,0%	0,14	0,34	-60,0%
Return on Equity	0,10	0,20	-51,7%	0,00	0,12	-100,0%	-0,10	0,00	3878,8%	0,00	0,12	-100,0%	0,39	0,47	-18,7%
Shareholder Equity Ratio	0,58	0,82	-29,3%	0,50	0,75	-33,5%	0,00	0,26	-100,0%	0,50	0,75	-33,5%	1,00	1,40	-28,7%
Short Term Assets / Total Assets	0,45	0,41	9,2%	0,49	0,38	29,5%	0,10	0,07	38,4%	0,49	0,38	29,5%	0,85	0,79	7,6%
Short Term Liabilities	1.950.611	2.030.749	-3,9%	976.674	1.062.290	-8,1%	155.844	219.415	-29,0%	976.674	1.062.290	-8,1%	3.775.620	3.983.590	-5,2%
Total Assets	7.115.023	9.239.482	-23,0%	2.940.240	5.503.383	-46,6%	827.796	1.716.547	-51,8%	2.940.240	5.503.383	-46,6%	11.004.238	17.578.936	-37,4%
Total Equity	4.182.047	7.447.148	-43,8%	731.797	3.243.339	-77,4%	23.212	550.503	-95,8%	731.797	3.243.339	-77,4%	5.694.442	15.190.598	-62,5%

Source: table by authors

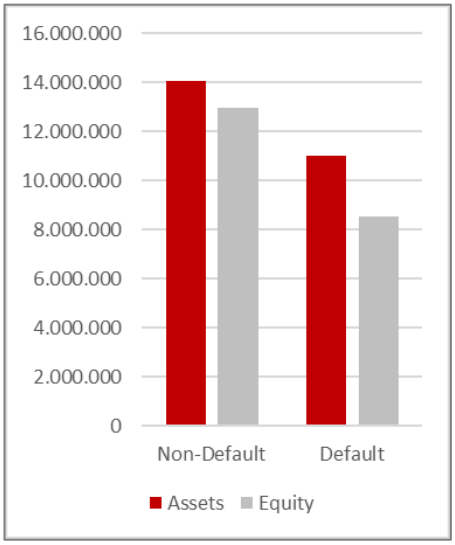
Marked with red are the differences where the default companies had a lower calculated value than the non-default companies. It is clear that most variables are lower for default companies compared to non-default companies. This also makes sense for a number of reasons. For one, as described

above, larger firms tends to have lower default rates and you would therefore expect most of the default companies in the analysis to be smaller companies and therefore have lower values for the asset / liabilities based variables. Also, and more importantly to this analysis, the default companies were simply performing worse on average than the non-default companies.

Below is a more detailed breakdown of some selected variables and how these differ between default and non-default companies.

5.2.3 Balance Sheet Based Variables

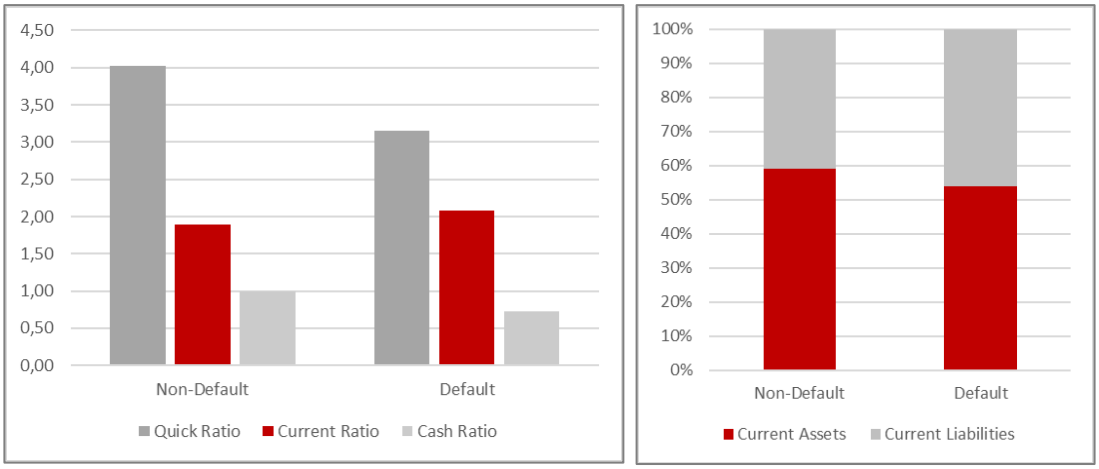
Looking at some of the more basic and aggregated balance sheet measures, here too can large differences be observed. The chart below shows the average asset and equity for default and non-default companies. This only strengthens the point made about firm size and default risk. The chart shows that non-default companies have an average asset size of about 14m whereas default companies have an average asset size of about 11m . The same difference can be seen on the equity where non-default companies have an average equity size of about 13m and default around 8,5m.



Source: illustration by authors

The balance sheet ratios tells much of the same story. The Quick Ratio is higher for non-default companies than for default companies indicating that non-default companies have more current assets compared to current liabilities. When looking at the Current Ratio, non-default companies have a lower average value than default companies, indicating that non-default companies have a significant amount of inventories (on average)/ability to cover its current liabilities.

The Cash Ratio is, like the Quick Ratio higher for non-default companies than default companies at around 1,0. This indicates that the non-default companies can repay on average 100% of their immediate expenses with the cash they have at hand compared to only about 70% for default-companies. The split between Current Assets and Current Liabilities can be seen on the chart below. Here it can be seen that default companies have a, marginally, higher proportion of current liabilities than non-default companies. Note, these ratios seem unusually high.



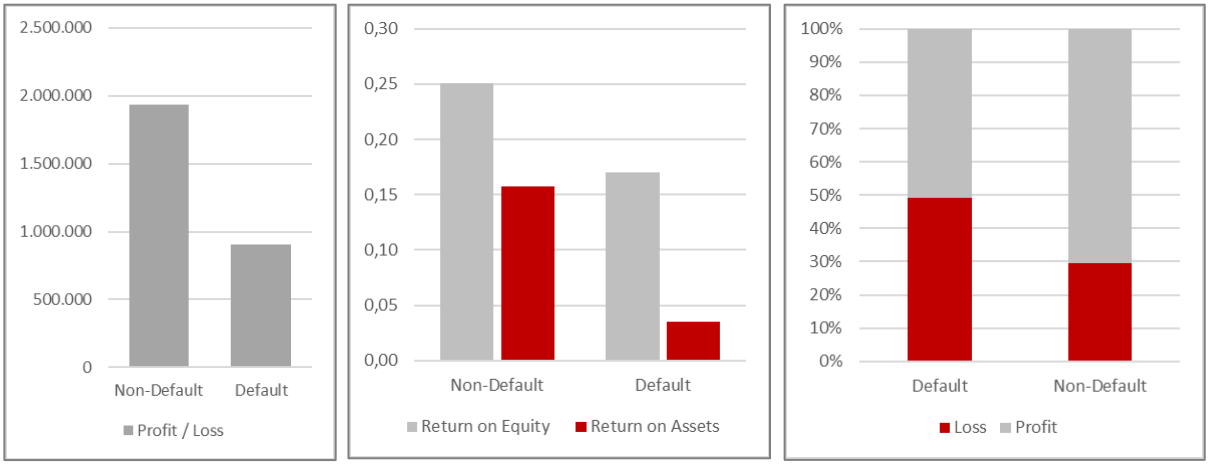
Source: illustration by authors

5.2.4 Profitability Based Measures

Where you would expect most differences between default and non-default companies would be on the profitability measures. The three charts below shows a selection of profitability measures. On the chart to the left is the average yearly profit / loss. Even though both categories of companies seems to be making a profit on average, the average profit just shy of 2m for non-default companies is twice as much as default companies.

On the chart in the middle, the balance sheet based return measures can be seen. Both the Return on Equity and Return on Assets are significantly higher for non-default companies compared to non-default companies. Return on Equity for non-default companies is about 0,25 compared to about 0,17 for default companies. Larger is the difference on the Return on Assets measure. Here non-default companies have a value of about 3,5 time the size of default companies – 0,16 compared to 0,03. This is makes sense as for default companies the total assets were about 15% lower and total profit about 50 % lower than non-default companies.

The chart on the right show the proportion of companies that are making a profit and a loss. For non-default companies, about half are making a profit and half are not. For default companies this is significantly different. Here only about 30 % are making a profit while the remaining 70% are not.



Source: illustration by authors

5.3 CBR Model Variables

From the CBR data 10 different main variables was extracted from the many hundreds of possibilities. These 10 variables were selected based on 3 criteria. **Firstly**, the variables have to be present for most of the companies. Variables with very few observations will most likely not prove significant in the analysis. **Secondly**, the variables have to have an assumed positive or negative correlation with default probability. Variables that are not intuitively correlated with default risk are not included (although some might prove to have a significant impact on default risk[5]). **Finally**, because the variables are observed over a short (five year) period, the events behind the variables have to occur on a somewhat frequent basis in order for them to be present during this period. A list of the 10 variables extracted can be seen below.

5.3.5 Variables and their Usage

Below is a list of the 10 variables. The list includes the variable name, type and a description. Below the list will be a description of the intuition behind each variable and their expected correlation with

[5] As an example, companies that are located on the ground floor might have a larger default risk than those located at the 4th or 5th floors but there is no intuitive explanation for this.

corporate default. Each variable will have a sub-hypothesis that is an initial statement of how the variable is expected to behave with regards to default risk. A *positive* correlation indicating that as the variable *increases*, so does default risk and vice versa. These sub-hypotheses will then be confirmed or rejected in the analysis.

CBR Data Variable Codes Description

Variable Name	Variable Type	Variable Description
Age	Continuous numerical	Age of company in years from creation year to 2018
Capital injection	Binary	1 if company has received one or more capital injection during the period
Corporation type changes	Delta	Number of changes in corporation type over period (e.g. from ApS to A/S)
Corporation type	Binary	1 if A/S, 0 if ApS
Employees interval	Interval	0 if 0, 1 if 0, 2 if 2-4, 3 if 5-9, 4 if 10-19, 5 if 20-49, 6 if 50-99, 7 if 100-199, 8 if 200-499, 9 if 499+
Main Industry Code changes	Delta	Number of main industry code changes over period
Main Industry Code	Categorical	Main industry code
Name changes	Delta	Number of name changes in period
No audit	Binary	1 if company has chosen not to be audited during the period, 0 otherwise
Telephone no. changes	Delta	Number of telephone number changes over period

Source: table by authors

5.3.5.1 Age

Description: The expected effect of company age can be split into two contradicting arguments.

Firstly, companies with long lifetimes (e.g. 50-100 years) are less likely to go default as they have displayed a more sustainable business model that has ‘survived’ for a long time. However, companies with long lifetimes also have higher risk of their products or services becoming obsolete because of trends, technological innovation etc., and therefore higher risk of going default.

Secondly, companies with very short lifetimes (e.g. 1-3 years) are expected to have a lower risk of going default, as they have not had enough time to go into financial distress.

Dunne & Hughes (1994) found that there might be a link between firm age and the growth patterns it goes through. Their analyses showed that younger firms tend to be more unstable in their growth

compared to older firms and are therefore in greater risk of defaulting. This finding has been both confirmed and rejected in several studies since (e.g. Ishikawa et al. 2014, Altman et al. 2011 and Law & Roache 2015).

This study will take point of departure in the first argument, stating a negative correlation between company age and risk of default, as the rationale is easily understood and not yet significantly disproven in the default risk literature.

Hypothesis 26: Age is significantly *negatively* correlated with default risk.

5.3.5.2 Capital Injection

Description: The expected effect of whether or not a company has received capital injection from its owners can also be split into two contradicting arguments.

Firstly, if a company receives capital injection it might be because the company needs money for matters such as operational- or interest expenses or in order to repay some debt. Thus, such injection could possibly be a sign of financial distress.

Secondly, a capital injection could also be an indication of facilitating expansion, taking advantage of the 'good times'. Moreover, one may argue that investors/owners, knowing the company and its default risk the best, would not inject capital in a 'lost cause'.

Lin, Chang & Lin (2013) describe how injection of capital by the government into banks could be beneficial in restoring firm stability. It is our hypothesis that this is also true for non-government capital injections and non-bank-firms.

Hypothesis 27: Capital injection is significantly *negatively* correlated with default risk

5.3.5.3 Corporation Type Changes

Description: We see it as unlikely that a company that changes corporation type within a year also goes default that same year. This is due to the fact that most company type changes are (presumably) related to the growth of a company when changing from a one-man company to an ApS type for example. This change is most likely due to the fact that the company is growing and the owners wants to limit personal liability, be able to reinvest potential profits etc.

Hypothesis 28: Corporation type change is significantly *negatively* correlated with default risk.

5.3.5.4 Corporation Type

Description: It is our initial hypothesis that there is a difference in the probability of default between the two types of corporation types. Numbers from Denmark's Statistics (Danmarks Statistik, 2018) show that A/S corporations have much higher revenue (6x on average) and number of employees (4x on average). In other words, they are larger corporations. This makes scenes as there are benefits of converting ApS to A/S with regards to creditworthiness, raising capital etc. Dunne & Hughes (1994) show that larger corporations displayed more stable growth patterns than smaller firms.

Hypothesis 29: Corporation type ApS is significantly *positively* correlated with default risk.

5.3.5.5 Employee Interval

Description: Making the same argument as the corporation type variable, that larger firms have lower default risk, we expect this to be true for this study as well. The number of employees can be seen as a proxy for firm size[6] and therefore a higher employee interval could mean a larger firm size. There is however a risk that higher employee numbers could mean overstaffing a company and therefore a higher default risk, but we find the first argument to be the most likely.

Hypothesis 30: Employee interval is significantly *negatively* correlated with default risk.

5.3.5.6 Change in Main Industry Code

Description: If a company changes main industry code it most likely means that the company has changed it focus of operation. A change in focus is not necessarily a bad thing as this could mean that the company has realized it can profit more in a different industry or has redefined what it is doing. However, if a company has to change industry code (because of a change in industry) this means that industry the company was in has proven not to be profitable and therefore forced the business to do something else. This scenario could mean that the company is underperforming and therefore presumably more likely to go default.

Hypothesis 31: Change in Main Industry Code is significantly *negatively* correlated with default risk.

5.3.5.7 Main Industry Code

Description: Many studies have proven that there are significant differences in default risk in different industries. Some industries, like energy / natural resource and transportation has historically proven to have higher default rates than insurance and banking industries (Vazza & Kraemer, 2016).

[6] Firm size here defined in terms of revenue

It is therefore out belief that this study will also show that there are significant differences in default risk between the different industries.

Hypothesis 32: There will be significant *differences* between the different industry codes.

5.3.5.8 Name Changes

Description: Name changes to a company is a rather drastic change to a company's brand. Changing a company name is therefore not something that happens often unless there is a reason. This reason could be that the company is simply rebranding, but it could also mean that the company has changed its owners, change what it is doing, its location etc. These last changes we see as negative and outweighing the positive motivations for changing the company name.

Hypothesis 33: Name Changes is significantly *positively* correlated with default risk.

5.3.5.9 Choice of no Auditing

Description: According to the CBR and the Danish Financial Statement act §135 a company classified as financial statement class "B" can choose to not be audited if it does *not* exceed two of these following three criteria two years in a row: 1) total assets of 4 million DKK, 2) turnover of 8 million DKK or 3) an average of number of full time employee (or equivalent) of 12.

Since it is optional for companies to choose not to be audited it is an active choice that the companies have to make. The reasons for not having its books audited can be that companies simply does not want to spend the money on audit but could also mean that the companies are in financial distress and does not want to have its books audited. We find the last argument most plausible and therefore see the indication of no audit as an indicator of greater default risk.

Hypothesis 34: Choice of no auditing is significantly *positively* correlated with default risk.

5.3.5.10 Number of Telephone Changes

Description: Making the same argument as with the number of name changes variable, the change of main telephone number is a significant change in the company's profile. The reasons for changing the main telephone number could be simply be due to a change in the telecommunications supplier, but could also be in order to be harder to get a hold of by authorities, creditors, customers etc. This last case would presumably be highly correlated with default risk as companies effectively hiding from its stakeholders cannot do this in the long run.

Hypothesis 35: Change in Telephone number is significantly *positively* correlated with default risk

Below is an overview of the above arguments and their expected correlation with default risk.

Variable Name	Expected correlation with default risk
Age	Negative
Capital injection	Positive
Corporation type changes	Positive
Corporation type	-
Employees	Negative
Main Industry Code changes	Positive
Main Industry Code	-
Name changes	Positive
No audit	Positive
Telephone no.	Positive

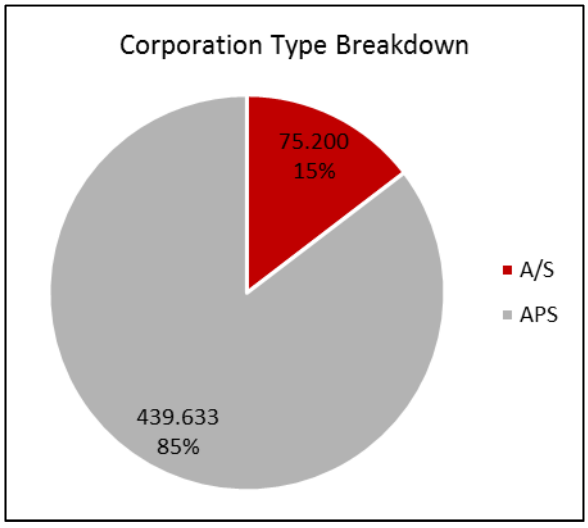
Source: table by authors

5.4 CBR Model Data Exploration

The raw data extract, extracted between February and April 2018, contains CBR records for **5.126.512** companies. A cross reference from the data available at the CVR register on the last day of extraction reveals that there are 5.139.038 companies available on the website – an extraction error of 0,24%. Below is a breakdown of the datasets.

5.4.6 Corporation Types

The dataset shows that of the 5.136.162 companies in the extract, 514.833 (≈10%) of these were either ApS or A/S. The pie chart shows the distribution between the two corporation types.



Source: illustration by authors

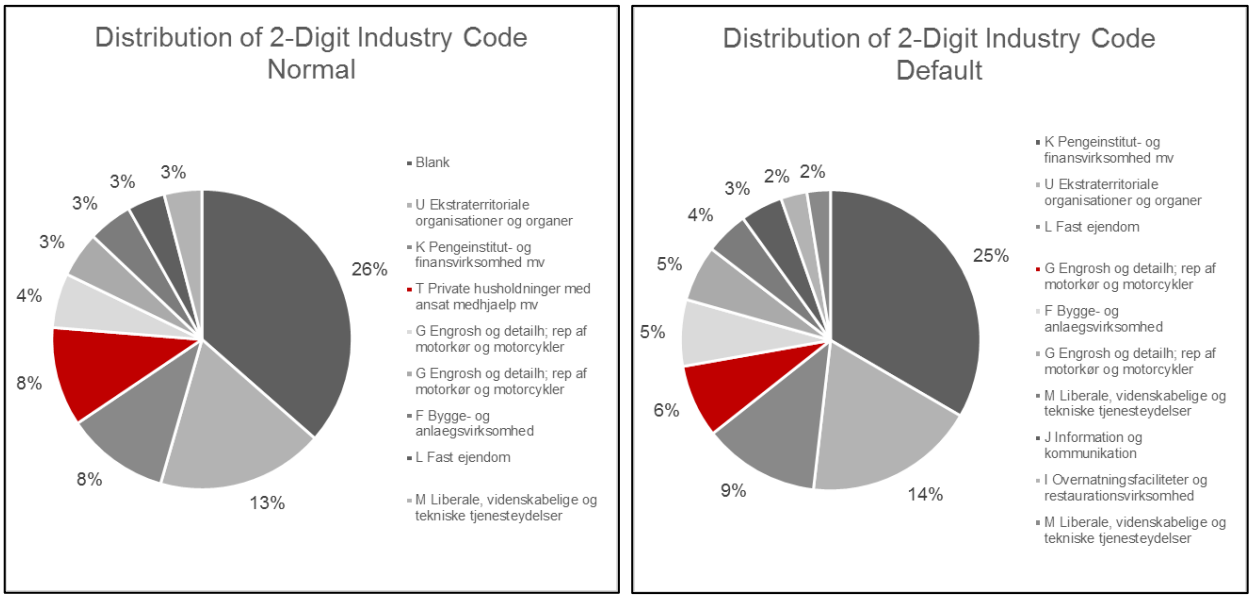
Further analysis reveals a breakdown of the different statuses for each corporation type.

	A/S	ApS
NORMAL	28.265	227.007
OPLØSTEFTERKONKURS	14.121	58.776
TVANGSOPLØST	4.984	55.717
OPLØSTEFTERERKLÆRING	3.088	42.390
OPLØSTEFTERFRIVILLIGLIKVIDATION	11.454	27.354
OPLØSTEFTERFUSION	10.352	12.523
OPLØSTEFTERSPALTNING	1.694	5.662
UNDERKONKURS	766	6.175
UNDERTVANGSOPLØSNING	74	2.100
UNDERFRIVILLIGLIKVIDATION	268	1.532
Ophørt	67	187
SLETTET	42	82
UNDERREASSUMERING	22	77
UDENRETSVIRKNING	0	28
UNDERREKONSTRUKTION	1	12
OPLØSTEFTERGRÆNSEOVERSKRIDENDEHJEMSTEDSFLYTNING	1	4
OPLØST	0	4
UNDERREASUMMERING	0	3
OPLØSTEFTERGRÆNSEOVERSKRIDENDEFUSION	1	0

Source: illustration by author

5.4.7 Main Business Area

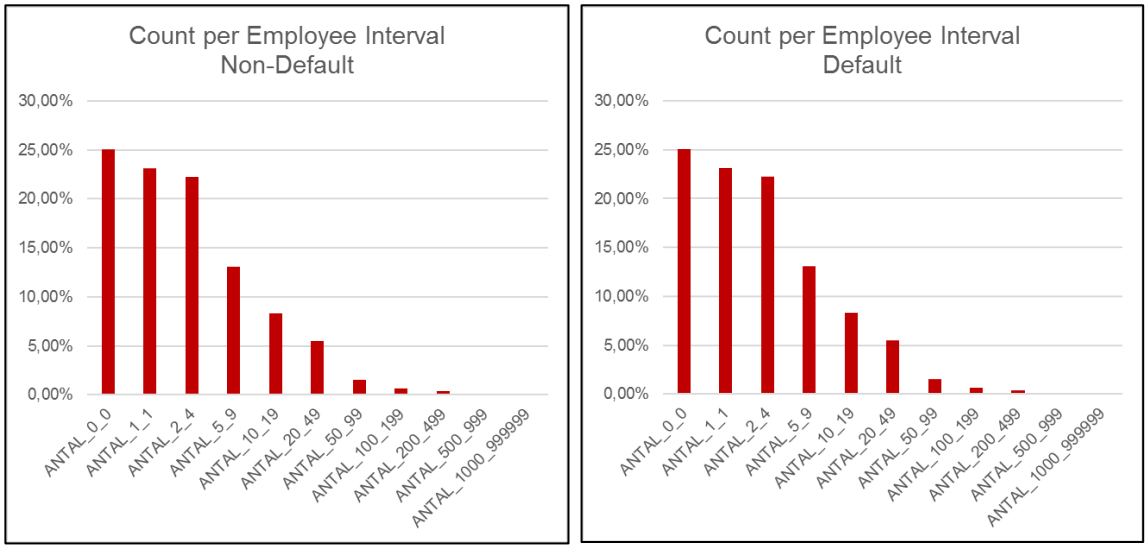
Looking at the main business areas for the companies, there is a clear difference between the companies that categorized as default and non-default. The most used industry code, with 25,2%, for the non-default companies is the industry code 64, indicating that it is operating within banking or finance – this support the choice to exclude such companies from the study. The most used industry code, with 25,3%, for non-default companies is ‘blank’. Since it is not mandatory to disclose a company’s industry code, many companies have this value as ‘blank’. It is however interesting that 25,3 % of the companies that have gone default have a blank industry code whereas this number is only 0,2 % of the operating companies.



Source: illustration by author

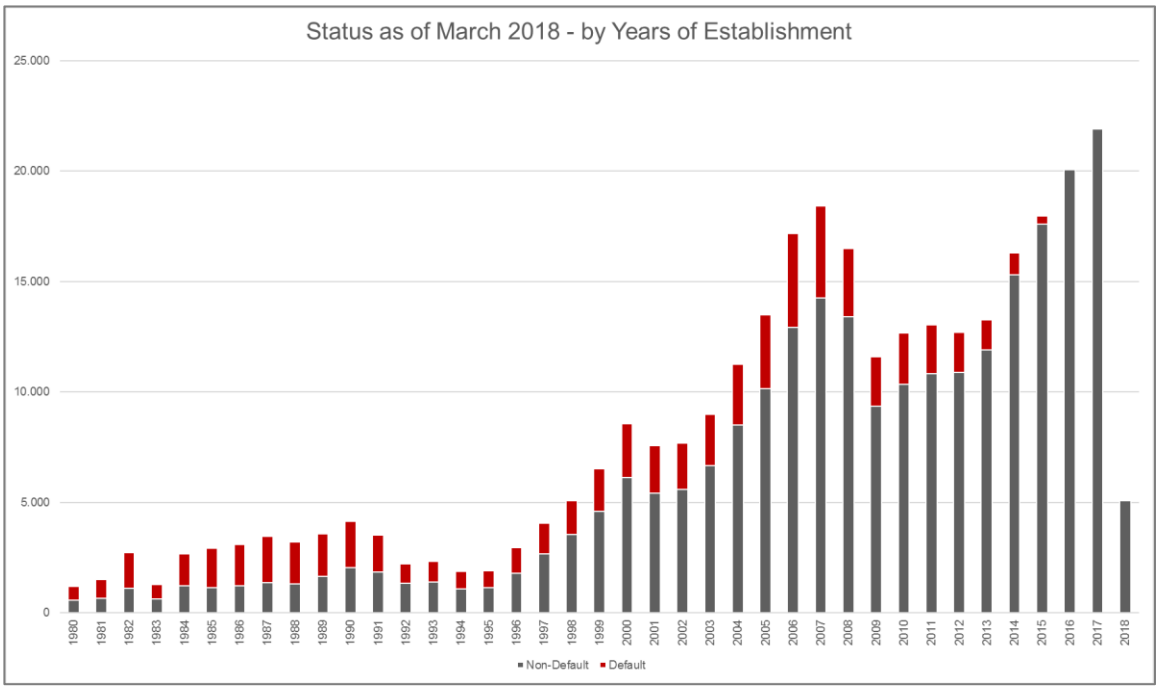
5.4.8 Employee Interval

In line with most literature and research within corporate default, there is a higher concentration of default for companies with fewer employees. Research have shown that smaller companies have a higher risk of default than larger companies and because the number of employees can be used as a proxy for company size, the distribution shown in the charts is no surprise. There is, however, a tendency for the interval with no employees to be used for both companies with no employees (where the only "employee" is the owner) and as a ‘bucket’ for companies that does not which to disclose their employee interval.



Source: illustration by author

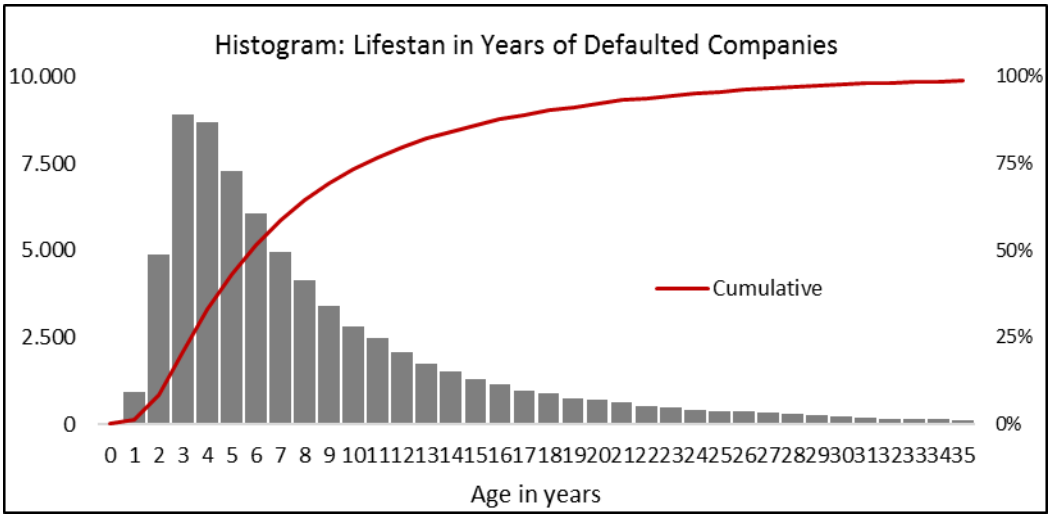
5.4.9 Additional Statistics



Source: illustration by authors

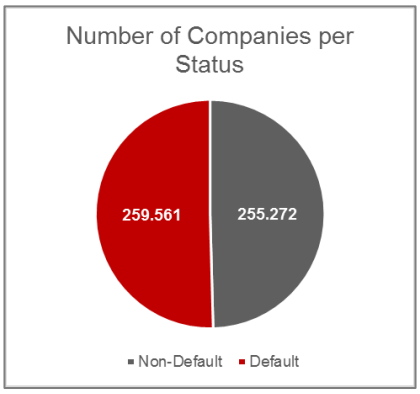
The above chart shows how many companies were established each year, and the proportion (red) which, as of March 2018, are default. Here we see that the number of companies that are created varies over time, most likely due to national or global economic environments. For example, there is a large increase in the number of companies started between 2003-2008. This was the period leading up to the financial crises – a time where the economy was booming. Then there was a drop in the number created, but then from 2014 a significant increase.

Below is a graph that shows the average lifespan of defaulted Danish companies. The graph shows that most companies are between 3-5 years at the time of default. It also shows that only very few older companies, with ages above 15 for instance, go default. In the other end of the scale is companies that are less than two years old. These companies also seems to be less likely to go default.



Source: illustration by authors

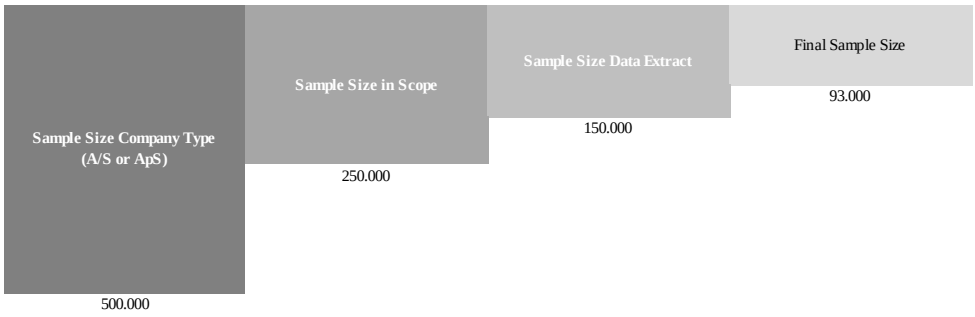
A breakdown of the statuses of the companies can be seen below. The split between default and non-default is almost 50/50. Of cause from these 500.000 companies, the ones that fall within the scope of this analysis have to be extracted.



Source: illustration by authors

The extraction process can be summarized by the graph below. Of the 5m total companies, approximately 500.000 were A/S or Aps company types. Of these 500.000 approximately 250.000 were active or default within the time period of the study (2013-2017). Of the 250.000 companies, it was possible to extract financial statements and CBR data for approximately 150.000 companies (primarily due to time and computational constraints). Of these 150.000 companies, a total of approximately 93.000 full observations could be used in the analyses after removing unwanted observations due to missing data.

This process of removing and carrying onwards data is graphically shown below:



Source: illustration by authors

5.5 Summary

This chapter was focused on selecting and calculating the variables that is to be used in the further analysis. The 10 non-financial variables were selected based on three criteria; their rational explanatory power, their presence for most for most of the companies and correlation with default. The 21 financial measures were selected based on which accounting figures were available and can be divided into three groups that tells something about a company's size, profitability and liquidity. These factors were deemed significant with regards to default prediction.

An short exploration of the two datasets based on the variables selected was also conducted. This was in order to make an initial analysis of whether the selected variables were differentiated from a superficial perspective. Luckily both dataset explorations showed significant differences between default and non-default companies, which indicates that the variables must have, at least some explanatory power.

6 Analysis of Financial Model

Before running the LR with the Financial Data it is necessary to first test whether the data lives up to the assumption of LR. From Chapter 3 we know that there are 5 basic assumptions of LR.

6.1 Testing Assumptions of Logistic Regression

The first is that the dependent variable is on a dichotomous scale. This assumption is easily met, as described in Chapter 3 above, the classification of whether a company is default or not, is very clear. The first assumption is therefore met.

The second assumption, which has to do with the independence of the observations, is also considered to be met. This is due to the fact that each observation is for a different company. It *can* be argued, that because what is considered an observation, is a given company's attributes for a given year and the same company therefore can have multiple observations if the company is active in more than one year within the time-scope, the observations are not fully independent. That said, a company's performance in one year, is not directly a result of how the company performed the year prior. The observations are, however, considered independent and therefore the second assumption is met.

The third assumption describe how there should be a linear relationship between the independent variables and the logistic transformation of the dependent variable. This is tested using scatterplots of the Pearson residuals against each variable (see appendix 2 for scatterplots). Based on these scatterplots, it can be concluded that the assumption is met.

The fourth assumption is concerned with the inter-correlations of the independent variables. To test for this assumption we test for the VIF statistics. Any VIF above 10 is considered non-linear and therefore considered not to live up to the assumption. Below is a test for VIF.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1,448	,031		46,115	,000		
	Current liquidity	-,036	,001	-,241	-54,781	,000	,486	2,056
	Quick Ratio	,000	,000	,005	,732	,464	,219	4,577
	Cash Ratio	-,001	,001	-,008	-1,583	,113	,370	2,702
	Net working capital	-,004	,002	-,012	-2,202	,028	,308	3,242
	Short term assets to total assets	,029	,005	,036	5,394	,000	,210	4,763
	Coefficient of financial stability	-,024	,003	-,039	-9,386	,000	,551	1,815
	Return on assets	,002	,005	,003	,405	,686	,236	4,239
	Return on equity	-,003	,002	-,005	-1,335	,182	,644	1,552
	Indebtedness factor	,000	,000	-,005	-1,452	,147	,898	1,114
	EBITDA/Total Liabilities	-1,235	,014	-,383	-86,777	,000	,486	2,059
	EBIT/Total Liabilities	,150	,045	,071	3,330	,001	,021	48,385
	Shareholder equity ratio	,007	,002	,018	3,114	,002	,269	3,719
	liabilities to total asset ratio	-,158	,004	-,109	-35,483	,000	,996	1,004
	Coverage ratio 1	-,001	,000	-,013	-2,808	,005	,410	2,436
	CurrentAssets	-,005	,002	-,016	-2,105	,035	,174	5,755
	ShorttermLiabilitiesOther ThanProvisions	,003	,003	,009	1,101	,271	,154	6,510
	CashAndCashEquivalents	-,004	,000	-,029	-7,333	,000	,594	1,683
	ProfitLoss	-,017	,001	-,093	-25,204	,000	,698	1,432
	Assets	,017	,003	,040	5,045	,000	,148	6,740
	Equity	-,027	,002	-,090	-11,778	,000	,162	6,183
	EBITDA / Assets	-,024	,004	-,028	-5,928	,000	,430	2,326
	EBIT / Assets	-,226	,047	-,105	-4,840	,000	,020	49,683

a. Dependent Variable: Event

Source: illustration by authors from SPSS

The test shows that 2 variables – EBIT / Total Liabilities and EBIT/ Total Assets – have a higher VIF than 10. Instead of excluding both variables with VIF values above 10, we look at which variables have high VIF scores. In the middle of the chart is the ratio EBIT/total liabilities and at the very bottom is EBIT/total assets. These both have too high VIF scores, which makes sense as there might be a correlation between current assets and current liabilities and the two therefore are similar in size which makes the two ratios correlated as well. EBIT/assets are therefore excluded from the analysis as that ratio is more concerned with profitability than liquidity/debt/liabilities, which most often can tell more about a company's financial health. Taking out the EBIT/assets ratio, and running the test again gives the results below. This shows that there are no independent variables with too high VIF scores.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1,446	,031		46,047	,000		
	Current liquidity	-,036	,001	-,241	-54,761	,000	,486	2,056
	Quick Ratio	,000	,000	,005	,792	,429	,219	4,576
	Cash Ratio	-,001	,001	-,008	-1,577	,115	,370	2,702
	Net working capital	-,005	,002	-,013	-2,368	,018	,309	3,239
	Short term assets to total assets	,025	,005	,032	4,757	,000	,214	4,668
	Coefficient of financial stability	-,024	,003	-,040	-9,637	,000	,552	1,811
	Return on assets	-,001	,005	-,001	-,194	,846	,240	4,175
	Return on equity	-,003	,002	-,005	-1,402	,161	,644	1,552
	Indebtedness factor	,000	,000	-,005	-1,543	,123	,898	1,114
	EBITDA/Total Liabilities	-1,234	,014	-,382	-86,739	,000	,486	2,059
	EBIT/Total Liabilities	-,057	,014	-,027	-4,076	,000	,213	4,696
	Shareholder equity ratio	,006	,002	,016	2,707	,007	,271	3,691
	liabilities to total asset ratio	-,158	,004	-,109	-35,466	,000	,996	1,004
	Coverage ratio 1	-,001	,000	-,011	-2,272	,023	,416	2,405
	CurrentAssets	-,004	,002	-,013	-1,721	,085	,175	5,718
	ShorttermLiabilitiesOther ThanProvisions	,003	,003	,009	1,198	,231	,154	6,508
	CashAndCashEquivalents	-,004	,000	-,029	-7,347	,000	,594	1,683
	ProfitLoss	-,017	,001	-,093	-25,211	,000	,698	1,432
	Assets	,017	,003	,039	4,929	,000	,148	6,736
	Equity	-,027	,002	-,089	-11,665	,000	,162	6,180
	EBITDA / Assets	-,027	,004	-,031	-6,578	,000	,437	2,291

a. Dependent Variable: Event

Source: illustration by authors from SPSS

The fifth and final assumption of minimizing outliners was also handled. In the dataset, there were a small part of the data that contained errors. These errors can be caused by some accounting figure being extremely high/low and can be traced back to either 1) errors in the XBRL data 2) errors when extracting data from the XBRL files or 3) errors when merging the XBRL files[7]. Regardless of the origination of these errors, they were handled by capping the ratios using quantile analysis and minimizing the "size" of certain variables to make the data more uniform – see Chapter 3.

6.2 Evaluating the Financial Model

6.2.1 Is Data At All Significant or Not?

After removing the two variables the remaining was inserted into a LR model. SPSS automatically creates an "Omnibus Tests of Model Coefficients". This test, essentially, produces the likelihood ratio, which in simple terms is a ratio used to compare models. It does so by calculating a pseudo chi-square, degrees of freedom and significance for what is called "Step", "Block" and "Model". "Step"

[7] Numerous studies, among those one by the creators of the format, have shown that XBRL has an error rate above 10%. See e.g. Charlie (2017).

shows the values for any additional data entered into the model. As there initially is no data entered, "Step" provides the results for the data first entered by assuming that the data is an addition to a constant. Hereinafter "Step" provides the results for any additional data entered in addition to the current data. "Block" is then the combined results for a defined number of "Steps". "Model" essentially refers to the results calculated based on the total dataset. As such, it is possible to enter a dataset, then take a "Step" and enter additional data, creating a second model, and review the significance of the "Step" to determine whether or not the 'new' data provided any value as well as to examine the model's total results. Thus, when only operating with one dataset – not adding additional data – "Step", "Block" and "Model" will hold the same results, essentially comparing the model to a model consisting of only a constant, i.e. a test that shows if, at all, the data has any significance and explanatory value. The test performed (see below table) shows a significance score of 0.000 for the "Model" which is below the threshold of 0.05. Hence, it can be concluded that the entered data is 100% significant and hold explanatory value.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	13670,904	21	,000
	Block	13670,904	21	,000
	Model	13670,904	21	,000

Source: illustration by authors from SPSS

6.2.2 How to Analyze a Logistic Regression Model

Having determined that the data is of some value, investigation continues into the strength of the model. It is very difficult, in general, to conclude a model's 'worth' or 'strength', simply by analyzing various scores, as many of these scores/calculations are dependent on dataset-specifications such as number of observations, number of variables or type of data (categorical, binary, values). Thus, to some degree the true assessment of a model's scores/outcomes must be conducted based on a profound understanding of the data; how it is created, how it is handled and the nature of observations/variables – see Chapters 1-5. Nevertheless, bearing in mind that the interpretation of scores/results are subject to such understanding, it is necessary to determine a point of departure for the analysis. Hence, this study has chosen to analyze and comment on 6 scores/levels of outcomes which seems to be fairly widely used within the LR sphere/literature. As a set, these results should create some coherent understanding of the model's strength. This same structure of analysis is utilized consistently throughout the paper.

6.2.3 Log Likelihood and Pseudo R^2 Analysis

The -2 Log Likelihood (-2LL) is a measure which is most commonly used to compare two models. It is calculated as:

$$-2LL = -2 * \ln\left(\frac{Likelihood\ for\ null\ model}{Likelihood for\ alternative\ model}\right) \quad 12)$$

The -2LL use the sum likelihood of two models – if there are not two, then null and alternative model is necessarily the same. The result is then checked against what may be labelled 'critical values'. A chi-square distribution table pairing degrees of freedom (df) with the level of significance, produce these 'critical values', which the -2LL must exceed in order to reject the scenario that there is a 5% chance, that the model's predicted likelihood is obtained chance. In doing so, the -2LL is able to take account for the difference in number of variables used in the two models which are being compared (1 variable = 1 df). This is perhaps the most important feature of the -2LL. As the -2LL is a measure of model fit, just like the R^2 which states that the best 'fit' is obtained when the sum of residuals is lowest, the -2LL too is preferred to be as low as possible, but still high enough to exceed the 'critical value'. As such, the -2LL does not enable much inference into how 'good' the model is without a reference point (another -2LL score), yet is preferred 'low'. Also calculated are two pseudo R^2s. Cox & Snell's and Nagelkerke's. To some extend these both assess a model's 'fit'. Both uses a comparative (ratio) framework somewhat equal to that of the -2LL; the lower the figure, the better the fit. The difference between the two R^2s is that Cox & Snell's may not reach 1 if the model has a perfect 'fit'. Nagelkerke's, however, is an adjusted version of Cox & Snell's which ensures that the result is between 0 and 1.

Looking at the Model Summary statistics below, the -2LL and pseudo R^2's can be found. The model shows a -2LL score of about 40.574,63 and Cox & Snell and Nagelkerke R^2s of 0,136 and 0,309, respectively. Upon finalizing analyses of all models, it was found that the Financial Model's -2LL was -4.742,33 compared to the CBR model (45.316,96). Thus, as the Financial Model's -2LL exceeds the 'critical value', it can be concluded that the Financial Model holds a better fit to the data than the CBR model, based solely on this score.

Model Summary

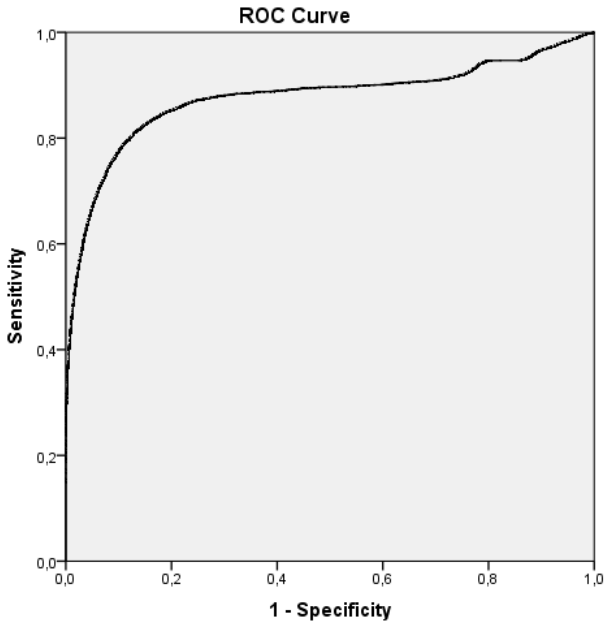
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	40574,632[a]	,136	,309

Source: illustration by authors from SPSS

6.2.4 ROC Curve (AUC score) Analysis

An ROC curve in simple terms is a graph which shows a model's ability to separate two categories. When the model has been made, it can be tested against the data used to create it. As such, the model calculates log odds and based on that score/odds determines whether or not an observation must be categorized – in our case – as default or not. Sometimes the model predicts correctly, this can be labelled a 'true positive'. Sometimes the model is incorrect, classifying an observation as default, when in fact it is not true; this can be labelled a 'false positive' (type 1 error). The ROC curve shows the relationship between the true positives and false negatives (type 2 error). Thus, if the model is useless, then it is no better at predicting default than flipping a coin (50/50) and the curve would be a straight, 45 degree line; the ROC curve would thus tell that the true positive rate is the same as the false positive at any point. However, if the ROC curve is not a straight line, then the model has some predictive value. The more towards the left and top border the curve is, the better the model is. AUC is then the calculated amount of the total area which is below the ROC line – thus if the model is perfect, there would be no area above the curve and AUC would be 1.

Calculating the ROC curve shows that the financial model has a very high AUC score of 0,876. Below is the ROC curve:



Source: illustration by authors from SPSS

This score can classify the model as 'good'. Comparing this score to other Danish default prediction studies, like Lykke et al. (2004) who had an average AUC of 0,825 shows that this model is as good or better at prediction correct outcomes.

6.2.5 Classification Table: Prediction vs. Actual

Next is the Classification Table. This shows that the model had predicted 92,9 % correct. The table shows that the model categorized at least some observations as default. More precisely it predicted correct on the non-default companies ≈ 100% of the time but only ≈ 16 % of the time for default companies. The model only predicted 1.264 of the observations as default compared to the real number of default observations of 7.909.

It may seem a bit odd, that the ROC curve and AUC score as well as the -2LL score are all presenting very reasonable levels, yet when it is tested, the model only predict that 16% (of the total amount of true defaults) will default within a year. However, notice the 'cut value' below the classification table. It is at 0,5 (or 50%), which entails that for the model to classify an observation as default, that observation's predicted log odds of default must be =>50%. Hence, the Classification Table is not necessarily an indication of how good a 'fit' the model has or how well it predicts the outcomes; e.g. if the cut value was set to 0%, then all observations would be classified as default. As such, the 'correctness' of the model can be somewhat manipulated if one only presented the classification table to an unaware audience. However, there is no correct level for the cut value. Essentially it depends

of the purpose of the model. The lower the cut value, the more likely it is that the model will predict false positives, yet this might be preferable – think of the cost of Type 1 and Type 2 errors. Imagine a study which uses LR to assess the probability of a person to have cancer. In such as case, the cost of a Type 2 error – in which the model predicts that there is no chance of the subject to have cancer, but the subject, in fact, does have cancer – is very, very high. In such a case, it may be preferable to lower the cut value, such that more subjects will be classified as "Does have cancer" (positives) and thus examined, only to find that many of them does not have cancer (false positives).

Classification Table[a]

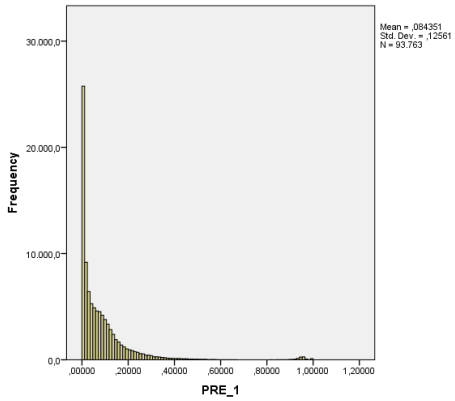
			Predicted		
			Event		
	Observed		0	1	Percentage Correct
Step 1	Event	0	85851	3	100,0
		1	6645	1264	16,0
	Overall Percentage				92,9

a. The cut value is ,500

Source: illustration by authors from SPSS

Hence, the cut value must be related to the purpose of the model. It was decided for this paper's study to use a cut value of 50%, which compared to e.g. Lykke et. al. (2004)'s 2.5%, seems extremely high. This thus entails that when an observations is classified as default, the likelihood of it actually happening is relative high – the log odds is above 50%. One could argue, that in practice it might be worth setting the threshold lower, yet this depends on e.g. the creditor's risk appetite. Further, it is thus possible to 'manipulate' the outcome to show a model which is very good at 'Classification', as the cutoff value yielding the best 'Classification' of the data can be calculated. See example in Appendix 6.

A quick gaze at the distribution of predicted probabilities show that the majority, by far, has a calculated probability of default below 0.20:



Source: illustration by authors from SPSS

Thus, if the cut value were set at 0.05 (5%) the model is almost certain to classify most of the observations as default. Contrary, it explains why the financial model only has a 16% 'hit-rate' on default observations; not many observations yield a log odds above 50%.

6.2.6 LogScore

Remember from Chapter 3 when the LS was explained that *"[…] if the model correctly classifies a default then the score is the log of the predicted probability $\hat{p}$, whereas if the model classifies a default incorrectly then the score is the log of* $1 - \hat{p}$ " As the log-function is inherently exponential, then an incorrectly classified default will be 'punished' relatively harder, than a correctly predicted default is 'rewarded'. Essentially, if the prediction likelihood for an observation is 0,2 (20%) and the true classification of the observation is default, then the prediction is not that great (LS = log of 0,2 = -1,6) However, if the prediction likelihood is 80% and the classification is still default, then it is a good prediction which yields a higher LS (log of 0,8 = -0,22). Notice, the relationship between the two LSs is not linear; 0,8/0,2 = 4 and 4*-0,22 ≠ -1,6. Thus, a low likelihood which creates an incorrect classification compared to the actual classification, is 'punished' hard. The sum of all observations' LSs divided by n-1 equal the mean LS. The aim is to obtain the highest possible (mean) LS. Moreover, the LS is not only used in isolation, but also frequently used to compare models.

The average of the LS for the Financial Model is calculated to be -0,2164. Compared to the CBR model's LS of -0,2416, then it is slightly better.

Report

LogScore

Mean	N	Std. Deviation
-,2164	93763	,79592

Source: illustration by authors from SPSS

6.2.7 Variables' Coefficients and Significance Analysis

The coefficient estimates as well as the significance level for each variable can be found in the table below. Overall, looking at the significance levels, 7 out of 21 variables were not significant.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CurrentRatio	-,798	,013	3529,820	1	,000	,450
	QuickRatio	-,005	,004	1,680	1	,195	,995
	CashRatio	-,034	,014	5,926	1	,015	,967
	NetWorkingCapital	-,005	,002	8,366	1	,004	,995
	CurrentAssetsToTotalAssets	,537	,080	44,645	1	,000	1,711
	CoefficientOfFinancialStability	,081	,045	3,290	1	,070	1,084
	ReturnOnAssets	,167	,093	3,231	1	,072	1,181
	ReturOnEquity	-,202	,030	44,137	1	,000	,817
	IndebtednessFactor	,002	,002	1,122	1	,290	1,002
	EBITDAToTotalLiabilities	-41,730	,630	4383,591	1	,000	,000
	EBITToTotalLiabilities	-1,843	,216	72,578	1	,000	,158
	ShareholderEquityRatio	-,065	,031	4,552	1	,033	,937
	TotalLiabilitiesToTotalAssetRatio	-4,254	604,017	,000	1	,994	,014
	CoverageRatio	,009	,005	3,634	1	,057	1,009
	CurrentAssets	-,015	,016	,926	1	,336	,985
	CurrentLiabilities	-,064	,016	16,558	1	,000	,938
	CashAndCashEquivalents	-,019	,003	37,083	1	,000	,981
	ProfitLoss	-,013	,002	46,714	1	,000	,987
	Assets	-,049	,019	6,534	1	,011	,952
	Equity	-,019	,002	69,979	1	,000	,981
	EBITDAToAssets	,600	,055	119,117	1	,000	1,822
	Constant	29,516	3998,169	,000	1	,994	6,583E+12

Source: illustration by authors from SPSS

This section will go through each variable and comment on their coefficients and their significance levels.

6.2.7.1 Current Ratio

With one of the highest (absolute) coefficient estimates, the Current Ratio can be said to have very high influence on default risk. Its negative coefficient of -0,798 shows that an increase of 1 in the current ratio results in a decrease in log odds of 0,798. The variable is also highly significant with significance level of 0,000. This negative relationship was what was also hypothesized and this make sense; as the ratio of current assets to current liabilities increases, the company becomes more liquid and therefore default is less likely.

Hypothesis 5 can therefore not be rejected.

6.2.7.2 Quick Ratio

Subtracting inventories from current assets and dividing with current liabilities gives you the Quick Ratio. In the LR model this ratio has a very low coefficient estimate of -0,005 and is considered not significant with a significance level of 0,195. It is interesting to see that the subtraction of inventories has such a large effect on the explanatory power of the variable. Not only is the variable less significant, the coefficient estimate has also turned from very negative to slightly negative. From this it can be concluded that inventories has significant effect on default risk in this model and that hypothesis 16 about the variables effect can be rejected. Most likely, inventory's effect rooted in the assumption that the inventory account constitutes the majority of companies' current assets and as such make up much of the company's potential liquidity, which in turn is negatively correlated with probability of default.

Hypothesis 6 can therefore be rejected.

6.2.7.3 Cash Ratio

Cash and cash equivalents over current liabilities yields the Cash Ratio. This variable has an estimated coefficient of -0,034 and a significance level of 0,025. Using only cash and cash equivalents instead of current assets or current assets minus inventories does have a significant negative impact on default risk. Though rather small in coefficient size, this measure does show that it is important to have readily available cash on hand in order to be able to serve its current obligations.

Hypothesis 7 can therefore not be rejected.

6.2.7.4 Net Working Capital

The net working capital variable has a coefficient estimate of -0,005 and a significance level of 0,004. Though rather small, there is a positive effect (in terms of a reduction in log odds) when there is an increase in net working capital. This relationship was also what was hypothesized as a higher net working capital means the company is more liquid. The measure is more significant than the three variables above, but its impact (coefficient) is lower.

Hypothesis 8 can therefore not be rejected.

6.2.7.5 Current Assets to Total Assets

With a semi large coefficient of 0,537 the Current Assets to Total Assets Ratio has a high positive correlation with default risk. An increase in short term assets to total asset would therefore mean an

increase in the default risk. It makes sense that if a company's assets, to a greater extend, becomes more "current", then its non-current (long-term) assets are decreased, deteriorating its ability to cover its non-current liabilities. This entails that while the company is relatively stable in the present, the future is significantly more unstable, which ought to result in an increase in risk of default. This behavior is also what was hypothesized and as the ratio is also highly significant with a significance level of 0,000.

Hypotheses 9 can therefore not be rejected.

6.2.7.6 Coefficient of Financial Stability

A company's ability to serve its long term debt is summarized by the Coefficient of Financial Stability. This ratio has an estimated coefficient of 0,081 but is considered non-significant with a significance level of 0,070. Because the measure is focused on the long run financial liquidity, we hypothesized that there would be a negative relationship between default risk and the coefficient of financial stability ration. However, it is neither the case that the ratios is negatively correlated nor that it is significant.

Hypothesis 10 can therefore be rejected.

6.2.7.7 Return on Assets

The Return on Assets shows a positive correlation with default risk contrary to what would be expected. The ratio has an estimated coefficient of 0,167 but is shown not to be significant with a significance level of 0,072. This is not a huge surprise as net profits is not necessarily correlated with the size of the assets. Large companies with small returns but with huge assets, for example, might skew the explanatory power of the ratio by having a very small ratio but without going default.

Hypothesis 11 can therefore be rejected.

6.2.7.8 Return on Equity

The net profit over equity shows how much profit a company is making on the equity. The model has estimated this ratio to have a negative relationship with default with an estimated coefficient of -0,202. This makes sense, as higher the ratio, the more profitable the business - all else equal. In turn, making it less likely that the company would default. The significance level indicates that this measure is significant to the model with a significance level of 0,000.

Hypothesis 12 can therefore not be rejected.

6.2.7.9 *Indebtedness Factor*

Showing both very little coefficient estimate and significance, this ratios is clearly not a factor that should be included when calculating default risk. With a coefficient estimate of 0,002 and a significance level 0,290 we can reject our hypothesis about the variable. This is somewhat surprising as we expected that companies that retained a larger proportion of their retained earnings would be more liquid and therefore less likely to go default. This is however not the case, according to this study.

Hypothesis 13 can therefore be rejected.

6.2.7.10 *EBITDA/Total Liabilities*

With the highest (absolute) coefficient estimate, the EBIDTA over Total Liabilities is estimated to be the most influential variable in the model. The coefficient of -41,730 shows that just a very slight increase in the ratio will result in a much lower default risk. This measure is not only highly influential but also highly significant with a significance level of 0,000. It is understandable that the ratio is so influential; essentially it exhibits a company's ability to cover its liabilities with its operational profit. That is, do the company make money by its core operations and how much? This is interesting to examine, as other profitability measures, such as Net Profit, incorporate non-cash-items and as such may present the company worse than what is actually true; consider the case of Maersk Drilling that took a $1,75bn write-down on its assets, which affected a Net Profit of -$0,6bn. However, if this non-cash-item were disregarded, the Net Profit would have been $0,5bn, revealing that the actual business is profitable. Nevertheless it shows that the hypothesis holds (Offshore Energy Today, 2017).

Hypothesis 14 can therefore not be rejected.

6.2.7.11 *EBIT / Total Liabilities*

Also with a high coefficient estimate, the EBIT / total liabilities variable has a coefficient of -1,843. The ratio is very closely linked to EBITDA / total liabilities (though not too inter-correlated as the VIF test shows). The variable is considered highly significant with a significance score of 0,000. The behavior of the variable is also what was hypothesized.

Hypothesis 15 can therefore not be rejected.

6.2.7.12 Shareholder Equity Ratio

The Shareholder Equity Ratio is displaying similar behavior to what is expected. The variable has a negative estimated coefficient of -0,065 and is deemed significant with a significance level of 0,033. This behavior would imply that as shareholders' equity increases, it can be said that the company's assets are less financed by debt and more by equity. This of course is a positive development for any company as it is less dependent on any third party service provider (bank etc.) to sustain operations. Also, in the event of default, shareholders are entitled to a larger amount of the assets.

Hypothesis 16 can therefore not be rejected.

6.2.7.13 Total Liabilities / Total Assets

The least significant variable is the Total Liabilities / Total Assets. Although the variable has a highly negative estimated coefficient of -4,254 the significance level of 0,937 is so insignificant that the coefficient does not matter. A possible explanation as to why the ratio is insignificant could be that almost all companies' ratio is close to 1.

Hypothesis 17 can therefore be rejected.

6.2.7.14 Coverage Ratio

Also very insignificant to default risk prediction is the coverage ratio. This variable has a very small (absolute) coefficient of 0,009 but a significance level of 0,057. The ratio is very similar to shareholders' equity ratio, only it is concerned with the non-current part of total assets. As such it is expected that it would have similar contributions, yet it has not.

Hypothesis 18 can therefore be rejected.

6.2.7.15 EBITDA / Total Assets

A company's ability to utilize its assets display the opposite behavior from what was hypothesized. The estimated coefficient is positive at 0,600 and is very significant to the model with a significance level of 0,000. From this it can be deduced that an increase in a company's ability to make an operational profit on its assets is positively correlated with default risk. This means an increase in operational efficiency increases the default risk, which does not make sense. At this point we can produce no other reasonable explanation as to why this relationship exists, other than our data for this measure is somehow incorrect (which does not seem to be the case, as the two measures are, on their own, significant and hold the anticipated relationship (coefficient)) or that companies take on more

risk when increasing their operational profit or by lowering (tuning) their amount of assets without lowering its level of operation.

Hypothesis 19 can therefore be rejected (EBIT/Assets was excluded from the analysis due to too high VIF level and as such this ratio has not been tested. Therefore the rejection of hypothesis 29 is based on EBITDA/Assets alone).

6.2.7.16 Current Assets

The level of Current Assets a company has is insignificant to the default prediction. As the Current Assets is not a ratio but a level value, it tells something about the size of the company. Therefore, you would expect that the higher the Current Assets the lower the default risk, assuming that larger firms are less likely to default. Although this is not what the model has estimated. It has estimated a coefficient of -0,015 but a significance level of 0,336. The hypothesis that as Current Assets go up, default risk go down can be rejected. Hence, it supports the argument that it does not matter how many assets one have got, but how they are financed and/or if they are profitable.

Hypothesis 20 can therefore be rejected.

6.2.7.17 Current Liabilities

Contrary to the estimates for Current Assets, is the Current Liabilities estimate. This variable is considered significant with a level of 0,000 and has a coefficient estimate of -0,064. This relationship implies that as the (ln-transformed) size of the Current Liabilities goes up, default risk goes down. This is in line with what we would expect; the larger the company's current liabilities, the less likely it is to default – again assuming that larger companies are less likely of default than smaller companies (which have been advocated for in the literature).

Hypothesis 21 can therefore not be rejected.

6.2.7.18 Cash and Cash Equivalents

Showing much of the same behavior as current assets, the cash and cash equivalents measure has an estimated coefficient of -0,019 and a significance level of 0,000. This implies that an increase in the cash at hand decreases the default risk, as would be expected. The measure is also estimated to be very highly significant. From these estimates it can be concluded that cash at hand has a negative correlation with default risk.

Hypothesis 22 can therefore not be rejected

6.2.7.19 *Profit / Loss*

The Profit or Loss a company makes is significantly correlated with the default risk of a company. The relationship is negative with a coefficient of -0,013 and a significance level of 0,000. This measure is perhaps the most obvious indicator of financial distress and the estimated relationship is therefore no surprise.

Hypothesis 23 can therefore not be rejected.

6.2.7.20 *Total Assets*

The size of the assets in a company proved to be a significant predictor of default risk, with a coefficient of -0,049 and a significance level of 0,011. This is no surprise as most of the default prediction literature shows, that larger companies are less likely to go default for a number of reasons.

Hypothesis 24 can therefore not be rejected.

6.2.7.21 *Equity*

Similar to total assets, the size of equity does also have a significant relationship with default risk. The estimated coefficient is -0,019 with a significance level of 0,000. As equity is another proxy for firm size, we expect the same relationship as with total asset size – that the larger the firm size the smaller the default risk.

Hypothesis 25 can therefore not be rejected.

6.3 Summary

The estimated Financial Model shows significant explanatory power when it comes to predicting probability of default. The analysis first showed that the data entered possessed superior explanatory power to that of a model created on the basis of the constant alone. The -2LL score was 40.574,632 and the pseudo R^2's showed levels of 0,136 and 0,309. By the Classification Table it could be seen that the model correctly classified 92,9% of observations correctly overall, but only 16% of the default companies correctly. This, seemingly low level of default prediction, was discussed in relation to the model's 'cutoff value', which at 50% is rather high, especially considering that the majority of the observations' log odds lies below 20%. Thus, in the light of this, the low level is accounted for. Calculating the ROC curve showed that the AUC was at a level of 0,876 – a score *as* high as or higher than other Danish studies (e.g. Lykke et al., 2004). Moreover, the model's LS was investigated and

it showed a score of -0,2164. While it is difficult to examine the strength of a prediction model without a reference point, it is concluded that the model exhibits an all-around medium/high strength. This was expected, as the literature on many occasions has validated the use of financial data for default prediction. Thus, this model adds to such validation within the literature.

Further, every variable were examined to determine their respective impact on the model, as well as to test the sub-hypotheses. It was found that 14 out of the 21 variables were significant with a significance level of 95%.

Going through each variable used in the model shows that the behavior of 8 of the variables was different from what was hypothesized. 7 of these (Quick Ratio, Coefficient of Financial Stability, Indebtedness Factor, Liabilities / Total Assets, Coverage Ratio, Return on Equity and Current Assets) because the significance levels were too high and therefore deemed the ratio insignificant and 1 (EBITDA/Assets) because it displayed different behavior from the hypothesized (positive/negative relationship to default).

7 Analysis of CBR Model

This analysis is similar to that of the Financial Model's analysis. However, as the various levels/scores have been explained earlier, this analysis will be less descriptive but simply convey the outcome.

7.1 Testing Assumptions of Logistic Regression

The first is that the dependent variable has to be on a dichotomous scale. Using the same argument as above, this assumptions is assumed to be met.

The second assumptions is also considered to be met using the same argument as above.

The third assumptions of a linear relationship between the independent variables and the logistically transformed default probability is difficult to meet when having so many categorical values. However, this assumption is assumed to be met.

The fourth assumption dealing with the inter-correlations of the independent variables. This is tested with the calculation of the VIF score. As below chart show, no variable has a VIF score above 10 which is why the variables are considers to not be inter-correlated and the fourth assumption met.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	,125	,004		32,328	,000		
	Corp. Type	-,009	,003	-,012	-3,654	,000	,834	1,199
	Corp. Type Changes	-,062	,004	-,081	-16,298	,000	,363	2,758
	Age	-,001	,000	-,050	-14,446	,000	,770	1,299
	Main Business Area Changes	-,021	,002	-,043	-10,310	,000	,530	1,886
	Employee Interval	-,003	,001	-,020	-6,519	,000	,923	1,083
	Name Changes	,038	,002	,073	16,163	,000	,445	2,248
	Audit	,659	,005	,443	120,695	,000	,672	1,488
	Capital Injection	-,243	,005	-,176	-47,925	,000	,673	1,486
	Telephone Changes	-,017	,002	-,024	-7,903	,000	,954	1,049
	Main Industry Code (short)	-,001	,000	-,024	-7,484	,000	,896	1,116

a. Dependent Variable: EVENT

Source: illustration by authors from SPSS

The fifth and final assumption is concerned with minimizing potentially strongly influential outliers. As most of the data is either binary or a count per year, this assumption is easily checked using a simple variable summary chart showing maximum, minimum and average values. The table below shows that there are no extreme values and the fifth assumptions is therefore assumed to be met.

	Min	Max	Average
Corp. Type	0	1	-
Corp. Type Changes	0	3	0,15
Age	0	115	11,32
Main Industry Code (short)	3	21	-
Main Industry Code Changes	0	4	0,30
Employee Interval	0	10	-
Name Changes	0	5	0,26
Audit	0	1	0,04
Capital Injection	0	1	0,04
Telephone Changes	0	4	0,15

Source: table by authors

Further, notice that the model consists of 10 main variables. Some of these variables have sub-variable categories (e.g. Main Industry Code). This means that the total amount of variables is not 10, but 36.

7.2 Evaluating the CBR Model

7.2.1 Is Data At All Significant or Not?

The above test showed that the data complied with all the assumptions and that no variables had to be excluded. Below Omnibus Test of Model Coefficients shows that the addition of the independent variables is a positive addition to the baseline model without any explanatory variables.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8931,220	36	,000
	Block	8931,220	36	,000
	Model	8931,220	36	,000

Source: illustration by authors from SPSS

7.2.2 Log Likelihood and Pseudo R^2 Analysis

The Model Summary table below shows that the model has a -2LL of 45.316,96 and a pseudo R^2 at 0,091, from the Cox & Snell R^2, to 0,207, from the Nagelkerke R^2. Compared to the Financial Model, these levels of lower, indicating a worse relative 'fit' of the model.

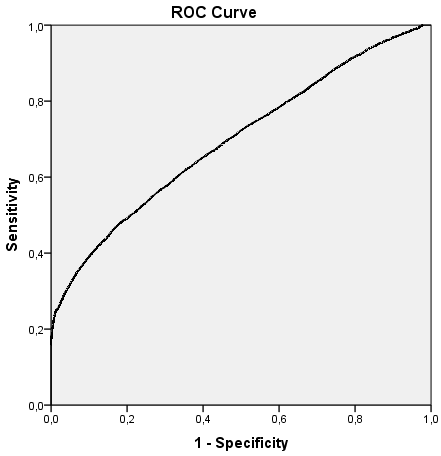
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	45316,960[a]	,091	,207

Source: illustration by authors from SPSS

7.2.3 ROC Curve (AUC score) Analysis

The ROC curve shows that the AUC for the model is calculated to be 0,698.



Area Under the Curve

Test Result Variable(s):

Area
,698

Source: illustration by authors from SPSS

As the CBR Model's AUC is less than the Financial Model's, its ability to distinguish between false positives and true positives is lower, meaning that it has moved towards a straight diagonal line, which would be the same a tossing a coin (50/50). However, in absolute terms the score is in the lower end of the scale, yet considering the simplicity of the data, the score is not bad altogether.

7.2.4 Classification Table: Prediction vs. Actual

The model's ability to correctly classify defaults is described in the table below. The table shows that the model classifies ≈ 100% correct with regards to non-default companies but only 16,7% correct for default companies. Overall the proportion of correct classifications was ≈ 92,9%

Classification Table[a]

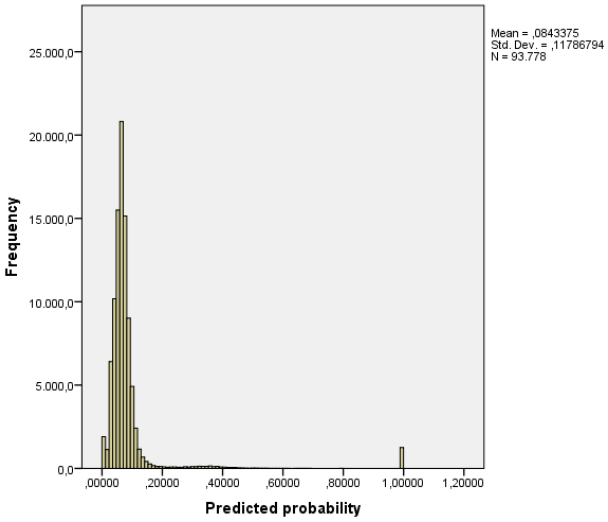
Observed		Predicted EVENT 0	Predicted EVENT 1	Percentage Correct
Step 1 EVENT	0	85840	29	100,0
	1	6588	1321	16,7
Overall Percentage				92,9

a. The cut value is ,500

Source: illustration by authors from SPSS

Hence, the CBR Model, employing the same cutoff value as the Financial Model, has actually correctly classified 16,7% of the defaults, which is 0,7% better than the Financial Model did. Considering the amount of variables used and the simplicity of the data, this is rather impressive.

A histogram of the predicted probabilities can be seen below. This shows that the vast majority of the predicted probabilities lies within the range from 0,08 and 0,12. There is however a small part of the observations that are centered around 1,0. Thus, once again the cut point could be discussed.



Source: illustration by authors from SPSS

7.2.5 LogScore

The calculated LS for the model is calculated below:

Report

LogScore

Mean	N	Std. Deviation
-,2416	93778	,64016

Source: illustration by authors from SPSS

At -0,2416 the CBR Model's log score is lower than the Financial Model by 0,0252. The difference is deemed appropriate/expected, yet the size of the difference is a bit surprising. The CBR Model holds a relatively decent LS.

7.2.6 Variable Coefficients and Significance Analysis

Next, the individual independent variable coefficients and significance levels are analyzed.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Corp.Type(1)	,124	,040	9,593	1	,002	1,132
	Corp.TypeChanges	-,942	,063	225,561	1	,000	,390
	Age	-,023	,002	162,940	1	,000	,978
	MainIndustryCodeChanges	-,300	,036	67,910	1	,000	,741
	EmployeeInterval			72,048	10	,000	
	EmployeeInterval(1)	2,098	1,015	4,277	1	,039	8,151
	EmployeeInterval(2)	1,977	1,015	3,796	1	,051	7,220
	EmployeeInterval(3)	1,897	1,015	3,495	1	,062	6,664
	EmployeeInterval(4)	1,889	1,015	3,465	1	,063	6,613
	EmployeeInterval(5)	1,757	1,015	2,995	1	,084	5,794
	EmployeeInterval(6)	1,786	1,016	3,092	1	,079	5,965
	EmployeeInterval(7)	1,758	1,020	2,970	1	,085	5,801
	EmployeeInterval(8)	1,689	1,029	2,693	1	,101	5,412
	EmployeeInterval(9)	1,681	1,043	2,599	1	,107	5,370
	EmployeeInterval(10)	2,366	1,075	4,847	1	,028	10,654
	NameChanges	,440	,030	211,081	1	,000	1,553
	Audit(1)	-22,602	634,479	,001	1	,972	,000
	CapitalInjection(1)	20,615	634,479	,001	1	,974	897749470,0
	TelephoneChanges	-,298	,038	60,733	1	,000	,742
	MainIndustryCodeShort			214,375	18	,000	
	MainIndustryCodeShort(1)	,201	,067	8,992	1	,003	1,222
	MainIndustryCodeShort(2)	,178	,240	,552	1	,457	1,195
	MainIndustryCodeShort(3)	-,189	,330	,328	1	,567	,828
	MainIndustryCodeShort(4)	,453	,057	63,830	1	,000	1,573
	MainIndustryCodeShort(5)	,391	,052	57,504	1	,000	1,478
	MainIndustryCodeShort(6)	,372	,098	14,482	1	,000	1,451
	MainIndustryCodeShort(7)	,766	,083	85,899	1	,000	2,152
	MainIndustryCodeShort(8)	,408	,067	37,463	1	,000	1,504
	MainIndustryCodeShort(9)	,091	,046	3,960	1	,047	1,096
	MainIndustryCodeShort(10)	-,001	,057	,000	1	,988	,999
	MainIndustryCodeShort(11)	,233	,058	16,001	1	,000	1,263
	MainIndustryCodeShort(12)	,418	,072	33,443	1	,000	1,519
	MainIndustryCodeShort(13)	,642	1,074	,357	1	,550	1,900
	MainIndustryCodeShort(14)	,232	,202	1,313	1	,252	1,261
	MainIndustryCodeShort(15)	,005	,115	,002	1	,966	1,005
	MainIndustryCodeShort(16)	,403	,160	6,362	1	,012	1,496
	MainIndustryCodeShort(17)	,446	,137	10,620	1	,001	1,562
	MainIndustryCodeShort(18)	-17,494	23204,931	,000	1	,999	,000
	Constant	-2,450	1,016	5,818	1	,016	,086

a. Variable(s) entered on step 1: Corp.Type, Corp.TypeChanges, Age, MainIndustryCodeChanges, EmployeeInterval, NameChanges, Audit, CapitalInjection, TelephoneChanges, MainIndustryCodeShort.

Source: illustration by authors from SPSS

7.2.6.1 Age

The measure of a company's Age also has significant explanatory power in the model. The variable has an estimated coefficient of -0,023 meaning an increase in age results in a decrease in default risk. This effect is significant with significance level of 0,000.

Hypothesis 26 can therefore not be rejected.

7.2.6.2 Capital Injection

Like the Audit variable it is a surprise that Capital Injection is not a factor when calculating default risk. The coefficient estimate is very high with an estimate of 20,615 but with a significance level of 0,974. Again, this can either be caused simply because there is no relationship between the independent and dependent variable, because of errors in the extraction of the data or, as may be the case with Audit, there are not enough observations for the variable.

Hypothesis 27 can therefore be rejected.

7.2.6.3 Corporation Type Changes

The dummy variable indicating whether a company has changed its corporation type in a given year has proven to be a very significant variable when calculating default risk. The estimated coefficient of -0,942 indicates a very strong negative relation between the corporation type change and default risk. This is most likely because the companies that change corporation type change from ApS to A/S indicates that they are expanding or growing. Only one changed from A/S to ApS. The significance further proves that the variable has large explanatory power by having a significance level of 0,000. This means that change in corporation type has a significant influence on default risk.

Hypothesis 28 can therefore not be rejected.

7.2.6.4 Corporation Type

The dummy variable for Corporation Type, with 1 indicating ApS and 0 indicating A/S corporation type, is a significant variable in predicting corporate default. The variable has a positive coefficient of 0,124, which can be interpreted as ApS type corporations being more likely to go default. The variable is significant with a significance level of 0,002. In general an ApS corporation is smaller than an A/S corporate and as such less likely to fail. Hence, the output is as expected.

Hypothesis 29 can therefore not be rejected.

7.2.6.5 Employee Interval

The overall categorical variable for the number of employees (the employee interval) is significant. However of the 10 different categorical classifications, only 2 are significant – employee interval 1 and 10.

Employee interval 1 are those companies with no employees. This variable has an estimated coefficient of 2,098 and a significance level of 0,039. This employee interval 1 is often used for companies that does not wish to disclose their number of employees and the positive relation is also what was hypothesized.

Employee interval 10 means the company has 499+ employees. This dummy has a positive correlation of 2,366 and a significance level of 0,028. This behavior does not make a lot of sense when having the relationship between firm size and default risk in mind. The coefficient means that a higher number of employees means a higher default risk. A double check of the number of defaults with employee interval 10 reveals that only 1 of them have gone default. The result is therefore somewhat questioned. Nevertheless, the hypothesis about employee interval is rejected.

Hypothesis 30 can therefore be rejected.

7.2.6.6 Main Industry Code Change

The dummy variable indicating whether a company has changed its main industry code in a given year has a significant effect on default risk. The estimated coefficient of -0,300 indicates that companies that change their corporation type has lower risk of going default. This variable is also significant with significance level 0,000. As with change in corporation type, a change in main industry code may very well indicate expansion and as such growth and financial stability. Therefore the coefficient and significance is not surprising. Thus, the behavior of the variable is what would be expected and what was hypothesized.

Hypothesis 31 can therefore not be rejected.

7.2.6.7 Main Industry Code

The overall main industry code variable is estimates to be significant with significance level of 0,000. Of the 18 different industry codes, 11 are significant.

Main industry code 1 is linked to industry code 3 which is ***manufacturing/production*** companies. According to the model estimates, having this industry code means that there is an increase in the default risk due to the positive coefficient of 0,201 and significance level 0,003.

Main industry code 4 is linked to industry code 6 which is ***construction*** companies. This industry code has a coefficient of 0,453 and a significance level of 0,000. This coefficient is the second highest of the significant variable coefficients and must therefore have a somewhat high impact on default risk.

Main industry code 5 is linked to industry code 7 which is ***wholesale and retail*** companies. This industry code has an estimated coefficient almost as high as main industry code 4 with a coefficient of 0,391 and significance level of 0,000

Main industry code 6 is linked to industry code 8 which is ***freight and transport*** companies. This category has an estimated coefficient of 0,372 and significance level 0,000.

Main industry code 7 is linked to industry code 9 which ***hotel/motel and restaurant*** companies. This main industry code has the highest coefficient with 0,766 and significance 0,000. This means that there is a very high correlation between the companies that have gone default and the main industry code 7.

Main industry code 8 is linked to industry code 10 which is ***information and communication*** companies. This main industry code has a significant explanatory power with a coefficient of 0,408 and significance level 0,000.

Main industry code 9 is linked to industry code 11 which is ***(non-banking) financial*** companies. This group has the lowest coefficient of the significant companies with 0,091 and is also the least significant (among the significant) industry codes with significant level of 0,047.

Main industry code 11 is linked to industry code 13 which is ***liberal, scientific and technical services*** companies. This variable has the second lowest coefficient of 0,233 and significance level of 0,000.

Main industry code 12 is linked to industry code 14 which is ***administrative*** companies. This industry code has the third highest coefficient with 0,418 and a significance level of 0,000

Main industry code 16 is linked to industry code 16 which is ***educational*** companies. The coefficient is estimated to be 0,403 and with a significance level of 0,012.

Main industry code 17 is linked to industry code ***other services*** companies. This industry code has an estimated coefficient of 0,446 and a significance level of 0,001.

The main takeaway from the analysis of the Main Industry Codes is, that there is significant differences between both coefficients and significance levels for the different industry codes. The highest coefficient for a significant observation is for hotel/motel and restaurant companies. This is no surprise as this industry has historically been plagued by defaults (as described in this resent article for example Nathan (2018)). The least significant industry code is the ***private house with hired help.*** This industry code is as insignificant as it can be, with significance level of 1,0, but also with a very low coefficient estimate of -17,49.

In appendix 3 is an overview of all the main industry codes and how they rank based on coefficient estimates and significance levels.

Hypothesis 32 can therefore not be rejected.

7.2.6.8 Name Change

If a company changes its name in a given year, it has a significant effect on default risk. The estimated coefficient of the Name Change variable is 0,440 and with a significance level of 0,000. This means that if a company changed its name it is more likely to go default. This is also what we hypothesized, however it is surprising the coefficient is so large and the p-value so low.

Hypothesis 33 can therefore not be rejected.

7.2.6.9 Audit

Very surprisingly is the estimation made about the dummy variable indicating whether a company has chosen not to be audited. This variable *does* have a very low coefficient estimate of -22,602 however the variable is also very insignificant with significance level of 0,972. It was hypothesized that the indication of not to be audited was highly correlated with default risk, however as the significance level indicates, this is not the case. Potential causes for this could be that there simply is no correlation between the independent and dependent variables or because the extraction of the data has not captured all of the no-audit-observations or that there are not enough observations for the variable.

Hypothesis 34 can therefore be rejected.

7.2.6.10 Telephone Change

If a company has changed its telephone number in a given year, it is less likely to go default. The model has estimated a coefficient of -0,298 with a significance level of 0,000. This is both different

from what was hypothesized but also surprising in terms of just how much influence it has. It implies that changes in telephone number is positive (in terms of a decrease in default risk).

Hypothesis 35 can therefore be rejected.

7.3 Summary

The CBR Model produced a -2LL score of 45.316,96 which is 4.742,33 higher than the Financial Model. As such, the Financial Model is 'better' than the CBR Model. In addition, the CBR Model's two calculated R^2s were below that of the financial model's, which simply support the above statement of relative 'fit'. Moreover, at 0,698, the CBR Model's AUC level was deemed to be in the lower end of the scale, with a substantial gap to the financial model's 0,876. While the CBR Model's LS at -0,2416 was not necessarily poor, it was slightly lower than the Financial Model's. The only factor for which the CBR Model is superior to the Financial Model is for the classification of defaults. The CBR Model predicted 16,7% correct compared to the Financial Model's 16%. Albeit only a very small difference, it is worth mentioning. Hence, based on the above comparison of scores/levels the conclusion may seem straight forward. However, consider the data used to produce the CBR Model. It consisted only of either binary or a count per year. In addition it had far fewer variables than the Financial Model. Bearing these factors in mind, a fair conclusion is that the CBR Model produce satisfying scores/levels on all parameters; scores/levels which are not far off from the more complex, larger (number of variables) Financial Model. As such, while the CBR Model 'loos' the battle, it still put up a fair effort. After all, the CBR Model does predict default better than a model based on a constant and thus validates its own usability for predicting default.

Finally, when analyzing the variables it was found that 6 of the 10 variables showed the expected behavior (2 were rejected on the bases of significance and 2 on the bases of correlation with default). Audit, Capital Injection, Employee interval and Telephone Changes did not behave as expected. Audit had, as expected a negative impact on the probability of default, yet were unexpectedly insignificant. The same is true for Capital Injection. It is believed that this might be due either some data error occurred when extracting or too few observations for which these variables were present (too few observations with the binary value 1; have had capital injection in year; have chosen not to be audited). Moreover, against expectation, Employee Interval (number of employees) were positively correlated with the probability of default. As number of employees often is an indication of the size of a company, it was expected to be negatively correlated. Finally, Telephones Changes

made in a year showed a negative correlation with the probability of default. This seems odd as it essentially states that the more a company changes its telephone number, the less likely it is to default. Nevertheless, all in all the coefficients' level and the significances assigned to each of the variables were overall satisfying.

8 Analysis of the Full model

8.1 Testing the Significance of Adding Variables Sequentially

8.1.1 Is Data At All Significant or Not?

The real test is whether the two models put together is more precise than any of the two separately. To test this we will sequentially add first the financial statement data (the Financial Model) and then the register data (CBR Model) to craft a Full Model. For each step there is a likelihood ratio test that determines whether the addition of the variables is significant to the model – as explained in section 6.2.1. This is done because some of the performance estimation models is dependent on the number of variables. The Full Model will then be compared to the two separate models on 6 parameters – -2LL, classification score, pseudo R^2's, AUC level and average LS.

The first addition of the Financial Model's variables to the model is, not surprisingly, significant as the model without the variables consists only of the constant (intercept). The model is now identical to the Financial Model. Below is the likelihood ratio test results in the Omnibus Tests of Model Coefficients showing that the "step" (entitled "Step 2") (the addition of the Financial Model variables) is significant – just as was shown earlier.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12694,161	21	,000
	Block	12694,161	21	,000
	Model	12694,161	21	,000

Source: illustration by authors from SPSS

Next is the addition of the CBR Model variables. This adds 36 more variables to the model such that it now holds 57 variables. The below is the post-step 2 Omnibus Tests of Model Coefficients which shows that the step of adding the new variables is significant.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8046,667	36	,000
	Block	8046,667	36	,000
	Model	20740,829	57	,000

Source: illustration by authors from SPSS

These two steps shows that the addition of the variables is significant. Next we test how the Full Model performs compared to the two models by themselves.

8.1.2 Log Likelihood and Pseudo R^2 Analysis

Below is the Model Summary for the Full Model, which shows the -2LL and pseudo R^2's. Full Model has an -2LL 31.232,071 and a pseudo R^2 range from 0,198 to 0,466.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	31232,071[a]	,198	,466

Source: illustration by authors from SPSS

If we compare these three figures to the two models, we see that the Full Model is better than the two models by themselves.

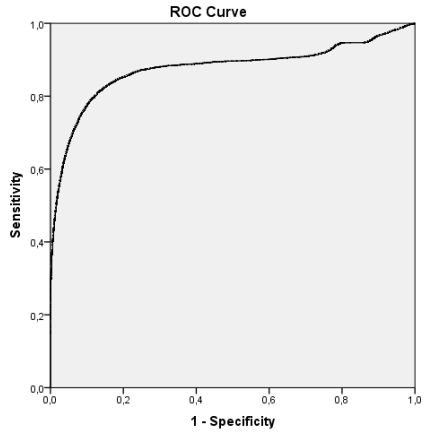
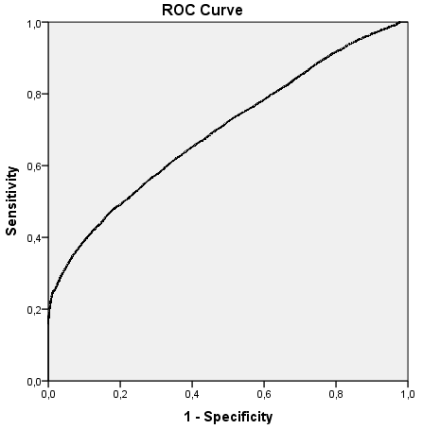
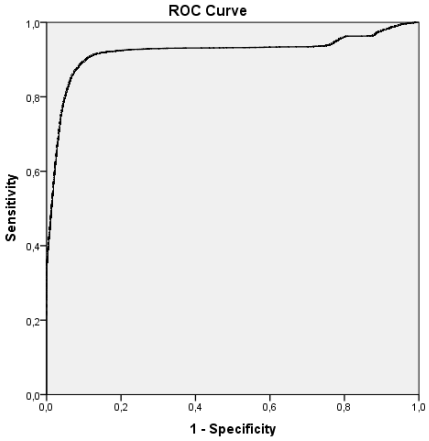
	-2 Log Likelihood	Cox & Snell R^2	Nagelkerke R^2
Financial Model	40.574,53	0,136	0,309
CBR Model	45.316,96	0,091	0,207
Full Model	**31.232,07**	**0,198**	**0,466**

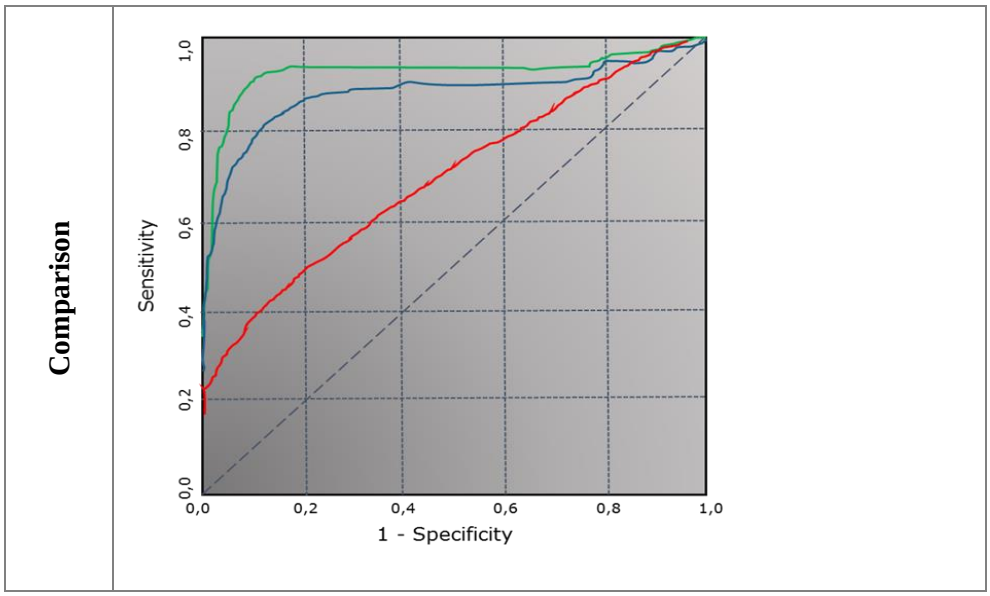
Source: table by authors

The comparison shows that the Financial Model is better than the CBR Model, but that the Full Model is the better of the three, as would be expected.

8.1.3 ROC Curve (AUC score) Analysis

Furthermore, the ROC curve also shows the superior performance of the Full Model. Below is the ROC curve and AUC score. The Full Model has an AUC score of 0,921. If we compare the ROC curves and AUC scores across the three models, the Full Model is by far the best.

Financial Model		**Area Under the Curve** Test Result Area ,876
CBR Model		**Area Under the Curve** Test Result Variable(s): Area ,698
Full Model		**Area Under the Curve** Test Result Area ,921



Source: illustration by authors from SPSS

8.1.4 Classification Table: Prediction vs. Actual

Looking at how well the Full Model correctly predicts classifications (defaults / non-defaults) the Full Model also outperforms the two models separately. Below is the Classification Table as well as a comparison table.

Classification Table[a]

Observed			Predicted EVENT 0	Predicted EVENT 1	Percentage Correct
Step 1	EVENT	0	86268	56	99,9
		1	4852	2587	34,8
	Overall Percentage				94,8

a. The cut value is ,500

Source: illustration by authors from SPSS

	% non-default	% default	% overall
Financial Model	**100,0**	16,0	92,9
CBR Model	**100,0**	16,7	92,9
Full Model	99,9	**34,8**	**94,8**

Source: table by authors

The two separate models are very close to each other, prediction wise, but the Full Model is by far better at classifying correct defaults.

8.1.5 LogScore

The calculated LS for all three models is shown below. The Full Model shows a significant improvement compared to the two apart, with a mean LS of -0,1665. Comparing this to the other models, we see that the Financial Model was marginally (11,64%) better than the CBR Model but the Full Model was substantially better (23,06%) than the superior of the two models, the Financial Model.

Financial Model

Report

LogScore

Mean	N	Std. Deviation
-,2164	93763	,79592

CBR Model

Report

LogScore

Mean	N	Std. Deviation
-,2416	93778	,64016

Full Model

Report

LogScore

Mean	N	Std. Deviation
-,1665	93763	,74184

Source: illustration by authors from SPSS

8.1.6 Variable Coefficients and Significance Analysis

Below is an overview of the Full Model's variables' coefficients and significance followed by a table comparing the three models' coefficients and significance levels. Highlighted in the comparison table are those significances which are above the threshold of 0,05 (in bold red) and those variables for which the combination (creation of the Full Model) meant a significant change in either coefficient or significance (row in light green).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Currentliquidity	-,702	,016	1859,150	1	,000	,496
	QuickRatio	,003	,005	,475	1	,491	1,003
	CashRatio	-,029	,016	3,266	1	,071	,972
	Networkingcapital	-,006	,002	10,504	1	,001	,994
	Shorttermassetstototalassets	,535	,096	30,766	1	,000	1,707
	Coefficientoffinancialstability	,067	,053	1,588	1	,208	1,070
	Returnonassets	,172	,107	2,586	1	,108	1,187
	Returnonequity	-,145	,036	16,704	1	,000	,865
	Indebtednessfactor	,003	,003	1,562	1	,211	1,003
	EBITDATotalLiabilities	-45,121	,757	3550,853	1	,000	,000
	EBITTotalLiabilities	-1,568	,249	39,539	1	,000	,209
	Shareholderequityratio	-,041	,036	1,246	1	,264	,960
	liabilitiestototalassetratio	-,842	,069	147,412	1	,000	,431
	Coverageratio1	,006	,005	1,044	1	,307	1,006
	CurrentAssets	-,042	,019	4,942	1	,026	,959
	ShorttermLiabilitiesOtherThanProvisions	-,042	,018	5,349	1	,021	,958
	CashAndCashEquivalents	-,017	,004	21,540	1	,000	,983
	ProfitLoss	-,015	,002	47,851	1	,000	,985
	Assets	-,051	,023	4,864	1	,027	,950
	Equity	-,015	,003	27,482	1	,000	,986
	EBITDAAssets	,610	,064	90,693	1	,000	1,840
	Corp.Type(1)	,229	,050	21,012	1	,000	1,258
	Corp.TypeChanges	-,282	,072	15,199	1	,000	,754
	Age	,013	,002	53,825	1	,000	1,013
	MainIndustryCodeChange	-,173	,041	17,679	1	,000	,841
	EmployeeInterval			20,132	10	,028	
	EmployeeInterval(1)	1,847	1,016	3,304	1	,069	6,342
	EmployeeInterval(2)	1,777	1,016	3,058	1	,080	5,913
	EmployeeInterval(3)	1,701	1,016	2,803	1	,094	5,481
	EmployeeInterval(4)	1,800	1,016	3,136	1	,077	6,051
	EmployeeInterval(5)	1,650	1,017	2,633	1	,105	5,208
	EmployeeInterval(6)	1,791	1,018	3,099	1	,078	5,998
	EmployeeInterval(7)	1,824	1,023	3,179	1	,075	6,197
	EmployeeInterval(8)	1,610	1,036	2,414	1	,120	5,005
	EmployeeInterval(9)	1,863	1,047	3,166	1	,075	6,443
	EmployeeInterval(10)	1,436	1,138	1,592	1	,207	4,204
	NameChanges	,324	,038	74,406	1	,000	1,383
	Audit(1)	-28,643	616,231	,002	1	,963	,000
	CapitalInjection(1)	19,515	616,231	,001	1	,975	298860781,5
	TelephoneChanges	-,288	,045	41,520	1	,000	,750
	MainIndustryCodeshort			40,509	18	,002	
	MainIndustryCodeshort(1)	-,184	,084	4,817	1	,028	,832
	MainIndustryCodeshort(2)	,267	,292	,839	1	,360	1,306
	MainIndustryCodeshort(3)	-,241	,375	,413	1	,520	,786
	MainIndustryCodeshort(4)	,076	,074	1,064	1	,302	1,079
	MainIndustryCodeshort(5)	-,103	,069	2,197	1	,138	,902
	MainIndustryCodeshort(6)	,049	,123	,159	1	,690	1,050
	MainIndustryCodeshort(7)	,311	,106	8,677	1	,003	1,365
	MainIndustryCodeshort(8)	,105	,085	1,537	1	,215	1,111
	MainIndustryCodeshort(9)	,062	,054	1,358	1	,244	1,064
	MainIndustryCodeshort(10)	-,003	,069	,002	1	,966	,997
	MainIndustryCodeshort(11)	-,032	,073	,189	1	,664	,969
	MainIndustryCodeshort(12)	,090	,090	1,001	1	,317	1,094
	MainIndustryCodeshort(13)	1,810	1,114	2,641	1	,104	6,110
	MainIndustryCodeshort(14)	,012	,233	,003	1	,958	1,012
	MainIndustryCodeshort(15)	,046	,133	,119	1	,730	1,047
	MainIndustryCodeshort(16)	-,418	,219	3,645	1	,056	,658
	MainIndustryCodeshort(17)	-,016	,167	,009	1	,925	,984
	MainIndustryCodeshort(18)	-17,609	21398,622	,000	1	,999	,000
	Constant	13,467	1,160	134,801	1	,000	705680,928

a. Variable(s) entered on step 1: Corp.Type, Corp.TypeChanges, Age, MainIndustryCodeChange, EmployeeInterval, NameChanges, Audit, CapitalInjection, TelephoneChanges, MainIndustryCodeshort.

Source: illustration by authors from SPSS

	Coefficient		Significance Level	
	Full	**Partial**	**Full**	**Partial**
Financial Variables				
Current liquidity	-0,702	-0,798	0	0
Quick ratio	0,003	-0,005	**0,491**	**0,195**
Cash ratio	-0,029	-0,034	**0,071**	0,015
Net working capital	-0,006	-0,005	0,001	0,004
Current assets / total assets	0,535	0,537	0	0
Coefficient / financial stability	0,067	0,081	**0,208**	**0,07**
Return on assets	0,172	0,167	**0,108**	**0,072**
Return on equity	-0,145	-0,202	0	0
Indebtedness factor	0,003	0,002	**0,211**	**0,29**
EBITDA / Total Liabilities	-45,121	-41,73	0	0
EBIT / Total Liabilities	-1,568	-1,843	0	0
Shareholder equity ratio	-0,041	-0,065	**0,264**	0,033
Liabilities / total asset ratio	-0,842	-4,254	0	**0,994**
Coverage ratio	0,006	0,009	**0,307**	**0,057**
Current Assets	-0,042	-0,015	0,026	**0,336**
Current Liabilities	-0,042	-0,064	0,021	0
Cash and cash equivalents	-0,017	-0,019	0	0
Profit / Loss	-0,015	-0,013	0	0
Assets	-0,051	-0,049	0,027	0,011
Equity	-0,015	-0,019	0	0
EBITDA / Assets	0,61	0,6	0	0
CBR Variables				
Corp. Type(1)	0,229	0,124	0	0,002
Corp. Type Changes	-0,282	-0,942	0	0
Age	0,013	-0,023	0	0
Main Business Area Changes	-0,173	-0,3	0	0
Employee Interval		0	0,028	0
Employee Interval(1)	1,847	2,098	**0,069**	0,039
Employee Interval(2)	1,777	1,977	**0,08**	**0,051**
Employee Interval(3)	1,701	1,897	**0,094**	**0,062**
Employee Interval(4)	1,8	1,889	**0,077**	**0,063**
Employee Interval(5)	1,65	1,757	**0,105**	**0,084**
Employee Interval(6)	1,791	1,786	**0,078**	**0,079**
Employee Interval(7)	1,824	1,758	**0,075**	**0,085**
Employee Interval(8)	1,61	1,689	**0,12**	**0,101**
Employee Interval(9)	1,863	1,681	**0,075**	**0,107**

Employee Interval(10)	1,436	2,366	**0,207**	0,028
Name Changes	0,324	0,44	0	0
Audit(1)	-28,643	-22,602	**0,963**	**0,972**
Capital Injection(1)	19,515	20,615	**0,975**	**0,974**
Telephone Changes	-0,288	-0,298	0	0
Main Industry Code (short)		0	0,002	0
Main Industry Code (short)(1)	-0,184	0,201	0,028	0,003
Main Industry Code (short)(2)	0,267	0,178	**0,36**	**0,457**
Main Industry Code (short)(3)	-0,241	-0,189	**0,52**	**0,567**
Main Industry Code (short)(4)	0,076	0,453	**0,302**	0
Main Industry Code (short)(5)	-0,103	0,391	**0,138**	0
Main Industry Code (short)(6)	0,049	0,372	**0,69**	0
Main Industry Code (short)(7)	0,311	0,766	0,003	0
Main Industry Code (short)(8)	0,105	0,408	**0,215**	0
Main Industry Code (short)(9)	0,062	0,091	**0,244**	0,047
Main Industry Code (short)(10)	-0,003	-0,001	**0,966**	**0,988**
Main Industry Code (short)(11)	-0,032	0,233	**0,664**	0
Main Industry Code (short)(12)	0,09	0,418	**0,317**	0
Main Industry Code (short)(13)	1,81	0,642	**0,104**	**0,55**
Main Industry Code (short)(14)	0,012	0,232	**0,958**	**0,252**
Main Industry Code (short)(15)	0,046	0,005	**0,73**	**0,966**

Source: table by authors

Looking first at the **financial variables**, the Financial Model contained 21 variables whereof 7 were deemed insignificant to the model. This has changed in the Full Model such that there are still 7 insignificant variables, however 2 of the variables that were insignificant in the Financial Model is now significant in the Full Model and 2 of the significant variables in the Financial Model is now insignificant in the Full Model. **Net Working Capital** has gone from significant in the Financial Model to insignificant in the Full Model, though retaining the same approximate coefficient. **Shareholders Equity Ratio** has gone from significant in the Financial Model to insignificant in the Full Model. The variable has changed its coefficient by a small amount but the significance level by a lot. The **Liabilities to Total Assets Ratio** was very insignificant, with significance level 0,994 in the Financial Model but is now very significant, with significance level 0,000, in the Full Model. The same is true for the **Current Assets** measure. This measure was very insignificant in the Financial Model but very significant in the Full Model.

The CBR Model variables shows much of the same development. 8 (identical) CBR Model variables were significant in both the CBR Model and the Full Model, however the categorical variables (Employee Interval and Main Industry Code) shows different coefficients and significance levels.

While it is not directly possible to compare the coefficients and significance levels of the separate models with the Full Model one-to-one, it is still interesting to look at them. Keeping in mind that the coefficients and significance levels are calculated based on three different models with three different datasets (all with more/less variables than the other as well as differing types of variables), the variables that changes coefficient and/or significance can be interpreted in two ways; 1) some CBR Model variables have Financial Model variables that act as 'proxies' for the CBR Model variables but are more significant, rendering the CBR Model variables insignificant or 2) the resulting model has a different composition of variables and some variables therefore are insignificant to that particular model.

Nevertheless, it can be useful examine to facilitate further discussion and also help grasp the general development of the model as well as to increase one's belief in the deduced conclusions of the models' overall strengths.

8.2 Summery

The full model, in short, proved superior in every aspect of which the models have been investigated:

Measure	(1) Full Model	(2) Financial Model	(3) CBR Model	(1-2) Difference	(1-3) Difference
-2LL	31.232,07	40.574,63	45.316,96	-9.342,56	-14.084,89
Cox & Snell R^2	0,198	0,136	0,091	0,062	0,107
Nagelkerke R^2	0,466	0,309	0,207	0,157	0,259
AUC Score	0,921	0,876	0,698	0,045	0,223
Classification (correctly classified default as %)	34,8%	16%	16,7%	19%	18%
LogScore	-0,1665	-0,2164	-0,2416	0,0499	0,0751

Source: table by authors

As per above table, the Full Model has a significantly lower -2LL score, a higher level for all 'fitting'-related scores (R^2s and AUC score), is better at classifying a correct amount of defaults and holds a higher LS.

At the outset of this study it was difficult to imagine that the Full Model would not be better than the two other models individually. Yet, had the Full Model only exceeded either of the two other models marginally, one could have questions its superiority. However, as it is clear from the summary table, the conclusion that the Full Model is in fact superior to the other models is rooted in a significant difference in both individual 'fitting'-measures as well as in comparison-measures. As such, the Full Model is validated.

In terms of the Full Model's variables, analysis showed that a range of variables – a total of 32 (25/36 of the non-financial variables and 7/21 of the financial variables) – were insignificant. While, in general, not many changes were observed in relation to variables' coefficients and significance when combining the two models, a few *did* stand out; Net working capital and Shareholder's Equity Ratio turned insignificant in the Full Model. Contrary, Total Liabilities/Total Assets as well as Current Asset measure turned significant. Moreover, Corporation Type Change saw a relatively sharp absolute decrease in its coefficient from -0,942 to -0,282. Age also experienced changes in its coefficients, yet in this case the sign switched from negative to positive, changing the interpretation of the effect of the variable Age altogether (from -0,023 to 0,013). However, the effect for this specific variable is very small and it is thus unlikely that this little change will affect the outcome significantly.

9 Discussion

This will be a discussion of the accounting manipulation techniques employed to boost business performance and its effect on default prediction models. This section is intended to be a critical look at the variables that goes into the model and how these can be manipulated and therefore potentially decrease the effectiveness of the model.

9.1 What is Accounting Manipulation

Roughly speaking, accounting manipulation is the act of manipulating accounting figures in such a way that it improves some (or all) aspect(s) of the reported figures (e.g. the annual report). Obviously there is a fine line between manipulating the accounting figures and committing a fraudulent action, which is one of the main reasons why companies are audited. However, some actions of accounting manipulation are not directly illegal and as such can be used to improve the outlook of a company's financial health. In effect, those ratios or figures, used in any model employing annual report accounting figures, may be incorrect, ultimately impacting the performance and 'correctness' of the models' output. So why are the accounts being manipulated?

9.2 Why Does Accounting Manipulation Occur?

Essentially it is to improve business performance or, more correctly, individual performance. Gibbs and Lazear (2015) advocates, in their seminal work "Personnel Economics in practice", for the belief that all individuals are self-centered and everyone will thus, at any given point, act accordingly. Essentially building a foundation for the well-established principal-agent problem. It is their conviction that everyone from board members to executives to managers and common employees, will act in a manner which maximizes their own utility, at the expense of the company's. This is a well-known belief which is the starting point of this discussion. Nevertheless, the literature revolving around such behavior is relatively extensive, as it is a central element in the performance measurement sphere.

9.3 How is Accounting Manipulation Performed?

There are many ways in which accounting manipulation (also referred to as 'technical accounting') can be used to manipulate accounts and financial statements. Below are 10 selected cases which we find especially interesting to this study:

Technique	Affects	
Name	**Income Statement**	**Balance Sheet**
1) Accelerated revenue	X	X
2) Delay expenses	X	X
3) 'Non-recurring' expenses	X	
4) Other income or expenses	X	
5) Pension plans smoothing	X	
6) Off-sheet items	X	X
7) Synthetic lease	X	X
8) Extend credit to boost sales	X	X
9) Recording bogus revenue	X	
10) Boasting operating income through improper classification	X	
11) Boasting operating income by shifting losses to the balance sheet	X	x

Source: table by authors

9.3.1 Accelerating Revenues

When accelerating revenue, the aim is to move future sales to the current period. This could be achieved by booking a lump-sum payment at time *t* when the service or product is to be provided in time *t*+1 or over the period of several years. E.g. a software provider receives upfront payment for a service which is to be delivered over the course of 5 years. The correct accounting method would be to amortize the revenue over the 5 years period, such that it is gradually distributed to depict the true relationship between revenue earned and 'amount' of service/product provided. Instead, it is possible to book the revenue in the current period and thus inflate revenue. Of course, the opposite is also a possibility; to postpone revenue. E.g. it is legal to distribute revenue received for a product in the current period, over the course of several periods. This would effect that a company could produce a revenue in times with no sales (Jeter & Chaney, Advanced Accounting, 2011).

Another way of accelerating revenue is through 'channel stuffing'/'registration of consignment as revenue' whereby a company makes a large shipment to a distributor at the end of a period and records the sale as revenue. However, as the goods are not sold yet the company would not have received any cash from the distributor. Further, as the distributor most likely have the right to return the unsold goods, the company should keep the products classified as a type of inventory until the distributor

has sold the product. This entails that the revenue can booked and affect the income statement, but the cost of those products (Cost of Goods Sold) will not go into the statement until a later period in which the distributor actually sells the products. Until then, the value of the goods will simply be placed on the balance sheet (Ibid).

Accelerating revenue can cause several accounting measures to be distorted. More specifically it can impact the profitability measures like EBITDA over Assets or Return on Equity. Both are significant in the Full Model.

9.3.2 Delay Expenses (Capitalizing vs. Expensing costs)

A cost is capitalized when it is booked to the balance sheet instead of the income statement. This is completely legal, yet it is an account technique reserved mainly for investment items, i.e. when a company invests in an asset, it does not (necessarily) view the cost as an expense, but rather an investment. Thus, upon taking the asset into use, the company will amortize its costs to the income statement and gradually the cost's expense is accounted for. Hence, to capitalize a cost is a way to move costs/expenses out of the income statement, which of course, results in an increased Profit (Coombs, Hobbs, & Jenkins, Management Accounting, 2005).

In the 1990s AOL (America Online) became the center of attention after it was found that they had capitalized the costs related to the making and distributing its CDs. AOL viewed the marketing campaign related to the distribution of the CDs as a long-term investment and thus capitalized the cost over an (self-determined) extended period. The more appropriate treatment would have been to expense the cost in the period the CDs were shipped (Financial Times, 2010).

Increasing the Profit via capitalization can affect especially the return measure variables used in the model. Artificially high returns can cause artificially high ratios which can affect the default probability as all profitability measures (except Return on Assets) are significant in the Full Model.

9.3.3 "Non-recurring" Expenses (One Time Items)

Non-recurring items or OTIs are essentially any revenue/cost which is extraordinary in nature. Companies have different ways of dealing with extraordinary costs. Some companies try to account for these extraordinary events in their books (by making accruals for example) to help in- and external analyst examine ongoing operating results. However, in some cases it is found that companies have an extraordinary cost each year. While this may be correct, some also ensure that they accrue for an extraordinary cost/income/less income each year, such that they are able later on to 'discover' that

they have reserved too much and thus are able to put something (an amount) back into the income statement (Jeter & Chaney, Advanced Accounting, 2011). This amount can be used to either increase revenue and profit or decrease expenses.

9.3.4 Other Income or Expenses

Perhaps one of the most misused accounts, the 'Other'-account is used by companies to book any 'excess' reserves (like those from the non-recurring expenses). Further, it is the perfect place to hide costs by netting them against other newfound income or simple burry it along with other more or less random costs (Coombs, Hobbs, & Jenkins, 2005). Thus unwanted cost or large credit notes can be 'hidden' or netted against each other to hide their true nature/origin of the costs/credit note. Further, the account may be used to book 'wrongly classified costs' (see below section 9.3.10).

9.3.5 Pension Plans Smoothing

Some companies have defined benefit plans (e.g. pension plans) which the company regularly pays into as an expense. Commonly the level of payments into the corporate pension fund is subject to two factors 1) governmental legislation and 2) whether or not the pension fund is under- or overfunded. However, these two factors create the possibility for the company to use the pension fund as 'cookie jar'. If the government suspend/postpone pension payments, it would give some leeway for companies to utilize those funds for other projects; as was the case during the Obama Registration in 2014 when a $10.8 billion transport bill was signed extending a 'pension-smoothing' provision for another 10 months (The Wall Street Journal, 2014). Further, whether or not a pension fund is over- or underfunded is subject to the company's internal calculations and policies – even though some official regulation exists in e.g. GAAP. Hence, the company is in complete control to determine the current state (over- or underfunded) of the fund, by employing accounting techniques (change demographic, payout plan, minimum interest rate etc.). This entails that the company is ultimately able to artificially create a situation in which its payments decrease, freeing up funds for other projects in a bull-market, simply by technically changing the fund's status/level (American Academy of Acturaies, Fundamentals of Current Pension Funding and Accounting, 2004)

While this is a technique still in use, the topic has been given some attension since the 1990s and as such amedments to both GAAP and IFRS have closed the ability to 'smooth' pension plans. Nevertheless, as the calculations of most of the standards/regulations set are determined internally in

the companies, it is still very much a possibility to take advantage of the situation (Investopedia, 2018).

Utilizing the 'cookie jar' could possibly create additional revenue and as such increase the revenue (profit) for a limited period. Ultimately affecting all profitability measures.

9.3.6 Off-sheet Items

The establishment of subsidiaries is a common company constellation when expanding into new markets, establishing production facilities or entering new partnerships. However, a subsidiary can also be used to house liabilities or costs which the parent company do not want recorded in its financial statements. As a subsidiary is a separate legal entity which is not necessarily owned wholly by the parent company, it is not required by the parent company to record the subsidiaries liabilities or expenses, in effect hiding them from the investors, analysts and shareholders (Jeter & Chaney, Advanced Accounting, 2011). This type of accounting manipulation could mean that liabilities are significantly lower than what would otherwise be the case, in effect skewing both liquidity and size measures of which most were significant in the Full Model.

9.3.7 Synthetic Leases

A synthetic lease is a technique used to keep costs off of the balance sheet. A lease is a long-term contract, during which the company pays a fixed amount per year. The cost of which is recorded in the income statement. However, at the termination of the lease agreement the company can have an obligation to buy the asset which was leased. Yet, because of the nature of the lease, such a liability is not recorded on the balance sheet until point of purchase. As a result, a synthetic lease in which the lease is more or less artificially created, the company can hide future liabilities from investors, analysts and shareholders (Ibid).

Removing such significant liabilities can affect the Coefficient of Financial Stability, Indebtedness Factor, Capital Structure (Total Liabilities to Total Assets) etc. The last measure is significant to the Full Model above and an increase in the ratio (which can be the case when deflating liabilities) will result in reduced default risk as the correlation for the measure is negative.

9.3.8 Extend Credit to Boost Sales

A simple but effective technique to boost sales is to extend the payment credit to customers. This, of course, is only possible if the company is relative liquid and will not foster problems for driving its

business (must be done without significantly affect working capital and reserves). If credit extension is possible, it is likely that customers will purchase a higher volume or the firm will acquire new customers, ultimately increasing revenue. This maneuver, however, is often likely only to be utilized over a limited time frame. While this is far from uncommon to do, it is difficult for analysts etc. to detect and as such the increase in revenue can be analyzed incorrectly. Finally, by extending the credit period, the firm would encounter some changes to its account receivables and thus its balance sheet (Investopedia, 2018).

9.3.9 Recording Bogus Revenue

Recording bogus revenue is a technique by which a company records a revenue which is not real – it is bogus. Global Crossing, a telecommunication company, were the center of a scandal in which it had recorded revenue from an alleged sale of services/products to another company, yet the sale was fiction; the counterparty company in turn replicated the technique and the companies essentially swapped ales, recording a revenue which never took place. Of course there is no chance for an analyst or other interested party to discover such fraudulent schemes (CNN, 2002). It is important to note that the recording of bogus revenue is less of a technique than it is direct fraud, as it is not legal to record bogus revenue.

9.3.10 Boasting Operating Income Through Improper Classification

Operating income, often referred to as EBITDA (Earnings Before Interest, Tax, Depreciation and Amortization) and EBIT (Earnings Before Interest and Tax) are measurements of the company's ability to create a profit based on its core operation. This means that the profit does not include costs related to e.g. financing, but simply focus on how well the asset(s), when in operation, produces a profit. Thus, naturally, as the notion of EBITDA/EBIT describes, it does not consider e.g. interest expenses or income. Thus, if a company classifies interest income on loans, to e.g. its franchisees, as sales, it is able to increase the operating profit (EBITDA/EBIT) artificially. Hence, an improper classification can lead to improved performance (Jeter & Chaney, Advanced Accounting, 2011).

EBITDA over total liabilitieis is the financial variable with the largest (absolute) coefficient, with a coefficient of -45,121. This means that default prediction is very sensitive to changes in the variable and therefore very sensitive to changes in EBITDA, such that may be caused by improper classification of interest.

9.3.11 Boasting Operating Income by Shifting Losses to the Balance Sheet

Shifting any losses from the income statement to the balance sheet inherently increases the bottom line. However, shifting the losses are not always easy (or legal). In the case of Enron, a quite complex scheme of Joint Ventures and subsidiaries (or so called Special Purpose Entities) were utilized to move around losses and realize gains on e.g. external as well as own stocks. The idea may partially resemble the above "Off-sheet items", yet is intrinsically more complex (CFA Institute, 2016).

9.3.12 Summary

In sum, there are many ways in which a company can manipulate its accounts. All of which has the same purpose; to improve the outlook of business performance. And it is a fact, that such accounting techniques/maneuvers exists (some of which is directly fraud, i.e. bogus revenue) as it has been seen conducted not only within SMEs, but within large companies as well:

Company Name	($bn)	By which accounting method
Enron	64	i.a. Boasting operating income by shifting losses to the balance sheet
Bernie Madoff	64	All of the above – a complete Ponzi scheme
Lehman Brothers	50	i.a. Boasting operating income through improper classification
WorldCom	11	Delay Expenses (Capitalizing vs. Expensing costs)
Freddie Mac	5	Non-recurring items
AIG	4	Boasting operating income through improper classification
HealthSouth	2	Accelerating Revenue
Waste Management	2	Recording Bogus Revenue
Satyam	1	Recording Bogus Revenue
Tyco International	0.5	Recording Bogus Revenue
Source	(Corporate Finance Institute (CFI), 2018)	

Source: table by authors

9.4 Effect on Default Prediction Models

The effect of accounting techniques on any model utilizing financial data (financial statements) is two-folded. Let's assume that every company utilizes accounting techniques. In such a case, the accounting techniques would not necessarily change the strength, reliability or interpretation of the model. However, it is unlikely that this is the case, and so this assumption probably does not hold. Thus, to the extent that the model is based on a significant amount of incorrect/tampered data arising from some companies' use of accounting techniques, the model itself will be incorrect.

Arguably, for larger datasets this is not a problem per se, as the majority of companies as expected not to utilize accounting techniques, to an excessive degree. Regardless, as long as some companies

are 'falsifying results (figures)', then any prediction model would be less accurate in assessing the probability of default than analysts (on average), as there are simply no way in which the model can anticipate or deduce these accounting techniques' influence on the figures and subsequently account for them. Contrary, one may argue that any model could analyze a much higher number of companies than a single analyst.

9.4.13 Detection of Accounting Techniques/Fraudulent Behavior

In order to make the above argument, the line of thought was a bit (over-)simplistic and plain. It is of course possible to create a model which – to some extent – could take into account some of the effects of accounting techniques. Most likely this would be done by calculating some test variables, testing for improper behavior – a so-called "Red Flags". Consider these 3 scenarios:

1. Accounts Receivables grow faster than sales
2. Inventory grows faster than sales
3. Profit/Loss grow faster than Operation Cash Flow (OCF)

9.4.13.1 1) Accounts Receivables grow faster than sales

Accounts Receivables vs. Sales Growth										
	Year									
	1	2	3	4	5	6	7	8	9	10
Assumptions										
Total Sales Growth Pr. Year	20%									
Credit Sales as % of total sales	10%	10%	10%	20%	30%	40%	30%	20%	10%	5%
Normal Sales as % of total sales	90%	90%	90%	80%	70%	60%	70%	80%	90%	95%
Sales										
Credit Sales	10	12	14	35	62	100	90	72	43	26
Normal Sales	90	108	130	138	145	149	209	287	387	490
Total Sales	**100**	**120**	**144**	**173**	**207**	**249**	**299**	**358**	**430**	**516**
Y-o-Y Growth										
Credit Sales	N/A	20%	20%	140%	80%	60%	-10%	-20%	-40%	-40%
Normal Sales	N/A	20%	20%	7%	5%	3%	40%	37%	35%	27%
Accounts Receivables (AR)	N/A	20%	20%	20%	20%	20%	20%	20%	20%	20%
Difference										
Normal Sales - AR	N/A	0%	0%	13%	15%	17%	-20%	-17%	-15%	-7%
Red Flag		OK	OK	Red Flag	Red Flag	Red Flag	OK	OK	OK	OK

Source: table by authors

To detect whether or not a company is inflating revenue by extending credit to its customers, one could analyze the relationship between accounts receivables (the amount of money the company is owed from customers) and its sales structure.

In the above, notice that "Total sales pr. year" is 20%. Consequently, the company's accounts receivable is expected to grow at 20% a year as well. Now, focus on the change in sales structure; from year 1-3 it stays the same, which entails that the 20% growth in Total sales is due to an equal growth in credit- and normal sales – which is apparent in the last table "Difference". However, in year 4 the sales structure has changed and the company have now credit sales which amount to 20% of total sales compared to 10% the year before. The company's Total sales still have increased by 20%, but it can be concluded that the increase was due to a larger increase in credit sales than in normal sales (year-on-year growth for credit sales is 140%, while normal sales only increased by 7%). This means that the difference between the increase in accounts receivables and normal sales is 13%, showing that while the company has increased sales, an increasing amount of those sales are not yet paid/turned into cash. This creates a "Red Flag" as it is a sign of revenue being accelerated.

9.4.13.2 *2) Inventory grows faster than sales*

Inventory vs. Sales Growth										
	Year									
	1	2	3	4	5	6	7	8	9	10
Assumptions										
Sales Growth	0%	0%	10%	20%	20%	20%	10%	15%	20%	30%
Investory Growth	0%	0%	10%	20%	20%	20%	20%	30%	40%	50%
Sales (Income Statement)										
Sales	100	100	110	132	158	190	209	240	289	375
Inventory (Balance Sheet)										
Inventory	100	100	110	132	158	190	228	297	415	623
Y-o-Y Growth										
Sales	N/A	0%	10%	20%	20%	20%	10%	15%	20%	30%
Inventory	N/A	0%	10%	20%	20%	20%	20%	30%	40%	50%
Difference										
Sales - Inventory		0%	0%	0%	0%	0%	10%	15%	20%	20%
Red Flag/OK	N/A	OK	OK	OK	OK	OK	Red Flag	Red Flag	Red Flag	Red Flag

Source: table by authors

In this example, sales- and inventory growth is growing at a similar rate until year 7. From year 7 and onwards, sales continue to increase, yet inventory accelerates its growth, growing more than sales.

In the case that inventory grows faster than sales, cost of sales, or accounts payables, the potential issue may be that inventory is obsolete, requiring a write-off or that the company have failed to charge the cost of sales on some sales. Of course such behavior should generate a "Red Flag" for any analyst.

9.4.13.3 3) Profit/Loss Grow Faster than Operating Cash Flow (OCF)

Operating Cash Flow vs. Net Income Growth										
	Year									
	1	2	3	4	5	6	7	8	9	10
Assumptions										
Net Income Growth	0%	0%	10%	20%	30%	40%	30%	20%	10%	0%
Growth in Accounts Receivable	0%	0%	5%	10%	15%	20%	25%	30%	40%	45%
Growth in Accounts Payable	0%	0%	5%	10%	15%	20%	25%	30%	40%	45%
Growth in Depreciation	0%	0%	5%	10%	15%	20%	25%	30%	40%	45%
Growth in Inventory	0%	0%	5%	10%	15%	20%	25%	30%	40%	45%
OCF										
Net Income	100	100	110	132	172	240	312	375	412	412
+ Depreciation	10	10	11	12	13	16	20	26	36	53
-Accounts Receivable	-50	-50	-53	-58	-66	-80	-100	-130	-181	-263
-Accounts Payable	-25	-25	-26	-29	-33	-40	-50	-65	-91	-131
+ decrease in inventory	20	20	21	23	27	32	40	52	73	105
OCF	**55**	**55**	**63**	**80**	**112**	**169**	**223**	**258**	**249**	**176**
Difference in Growth										
Net Income Growth		0%	10%	20%	30%	40%	30%	20%	10%	0%
OCF Growth		0%	14%	28%	40%	51%	32%	16%	-4%	-29%
Difference		0%	4%	8%	10%	11%	2%	-4%	-14%	-29%
Red Flag		OK	OK	OK	OK	OK	OK	Red Flag	Red Flag	Red Flag

Source: table by authors

In this case, focus is on cash generation compared to Profit/Loss. The cash generated from operating activities (OCF: Operating Cash Flow) is calculated by adding back depreciation and decrease in inventory to Profit/Loss, and then subtracting any increase in accounts receivables and any decrease in accounts payables. This leaves the amount of cash generated from operations.

Comparing the growth of OCF and Profit/Loss enable the analyst to estimate the quality of earnings. If Profit/Loss grows faster than OCF then some of the growth is due to a decrease in non-cash related items; e.g. accelerated depreciation or inflated accounts receivables. In sum, the company exhibits healthy growths rates in Profit/Loss, yet its ability to turn revenue into cash is decreasing due to some element of changed behavior in non-cash related items. This should produce a "Red Flag".

9.5 Summary

There can be little doubt that accounting techniques are utilized to influence financial statements' performance outlook. It was shown that these techniques are employed to affect either the income

statement or the balance sheet. As a result, any analysis conducted on financial statements could possibly be incorrect. By extension, this study's models may to some degree be affected. Insofar as that is the case, some margin of error is yet to be accounted for, which entails that the models inherently are sub-optimal. However, as not all accounting techniques are detectable via investigation of a single financial statement, it is unfair to assume the models to fully – or at all – account for it.

The detection of accounting techniques is to some degree possible by investigating the relationship between various financial statement figures. Also, it is important to state that the above examples are not necessarily signs of employment of accounting techniques, but merely a process by which one can filter out those companies to be further investigated. The above are just a few, simplistic "Red Flag"-checks which can be made of a large variety of checks. These can be found in appendix 4.

10 Conclusion

This papers main objective was to investigate whether the addition of non-financial data could add significant explanatory power to corporate default prediction models. In order to examine this, non-financial information for 5m Danish companies were extracted from the Danish Central Business Register. Of these, a total of 500.000 ApS and A/S companies were identified of which 250.000 were within the time scope of this analysis (2013-2017). From this number it was possible to extract financial data form 150.000 of which a total 93.000 companies had complete financial and non-financial information and were used in the analysis.

The extraction and modeling proved quite complex due to the nature of the raw data. Several steps had to be taken in order to transform the extracted data's formats, to a format that could be used in statistical analysis. This need for complex transformation limits the direct usability of the data, as it requires advanced technical abilities and computational capacities.

To analyze the technical usefulness of the data, analyses based on logistic regression was first made separately for financial an non-financial data. The Financial Model's data proved to be a valid data source and therefore significant when used in a logistic regression model. The output from the Financial Model shows an AUC score of 0,876, a classification precision of 92,9 % and a log score of -0,2164. The model showed that 14 out of 21 variables were estimated to be significant.

***H1 can therefore not be rejected**. Confirming that Danish publicly available financial data can be used to predict corporate defaults for Danish ApS and A/S companies.*

Non-financial data was also analyzed in the same manner. Not surprisingly, this data-type showed less explanatory power than the financial data. Of the 10 selected main variables, 7 were estimated to be significant resulting in an AUC score of 0,698, a classification precision of 92,2 % and a log score of -0,2416. The CBR Model thus proved somewhat able to distinguish between default and non-defaults, but less effective than the Financial Model. Overall the CBR Model had lower scores in all evaluating measures, other than classification of defaults, compared to the Financial Model.

***H4 can therefore not be rejected**. Confirming that while the non-financial model possesses some explanatory power it is less precise than the financial model.*

When combining the two datasets, of financial and non-financial data, the resulting Full Model showed significant improvement compared to the two separately. The Full Model had an AUC score of 0,921, a classification precision of 94,8% and a log score of -0,1665. This means that the AUC score was improved by 5,14 % compared to the Financial and 37,30 % compared to the CBR Model. Further the Full Model showed an improvement in the log score of 23,15 % compared to the Financial Model and 31,08 % compared to the CBR Model. Based on these findings we can answer our research question;

yes, the addition of publicly available non-financial data from the Danish Central Business Register does significantly improve corporate default prediction models for Danish ApS and A/S companies.
H2 and H3 can therefore not be rejected.

In terms of the sub-hypotheses that were made, 12 out of 31 were rejected, 8 because of the variable being insignificant and 4 due to the variable having a different correlation with default than hypothesized. See a list below for an overview of the sub-hypotheses:

Variable Name	H. No.	Test result	Showing following correlation to default
Current Ratio	5	Not rejected	Significant negative correlation
Quick Ratio	6	Rejected	Insignificant negative correlation
Cash Ratio	7	Not rejected	Significant negative correlation
Net Working Capital	8	Not rejected	Significant negative correlation
Current Assets to Total Assets	9	Not rejected	Significant positive correlation
Coefficient of Financial Stability	10	Rejected	Insignificant positive correlation
Return on Assets	11	Rejected	Insignificant positive correlation
Return on Equity	12	Not rejected	Significant negative correlation
Indebtedness Factor	13	Rejected	Insignificant positive correlation
EBITDA to Liabilities	14	Not rejected	Significant negative correlation
EBIT to Liabilities	15	Not rejected	Significant negative correlation
Shareholder Equity Ratio	16	Not rejected	Significant negative correlation
Liabilities to Total Assets	17	Rejected	Insignificant negative correlation
Coverage Ratio	18	Rejected	Significant positive correlation
EBITDA to Assets	19	Rejected	Significant positive correlation
EBIT to Assets	19	Rejected	Significant positive correlation
Current Assets	20	Rejected	Insignificant negative correlation
Current Liabilities	21	Not rejected	Significant negative correlation
Cash and Cash Equivalents	22	Not rejected	Significant negative correlation
Profit / Loss	23	Not rejected	Significant negative correlation
Total Assets	24	Not rejected	Significant negative correlation
Equity	25	Not rejected	Significant negative correlation
Age	26	Not rejected	Significant negative correlation
Capital injection	27	Rejected	Insignificant negative correlation
Corporation type changes	28	Not rejected	Significant negative correlation
Corporation type	29	Not rejected	Significant positive correlation

Employees	30	Rejected	Significant positive correlation
Main Industry Code changes	31	Not rejected	Significant negative correlation
Main Industry Code	32	Not rejected	Significantly different correlations
Name changes	33	Not rejected	Significant positive correlation
No audit	34	Rejected	Insignificant negative correlation
Telephone no.	35	Rejected	Significant negative correlation

Source: table by authors

11 Reflections

This chapter seek to reflect on this paper's outcome, findings and the process by which these were obtained. Further, it aims at specifying its contribution to the established literature as well as to propose areas of interest for future studies.

11.1 Usability of the Model

The overall purpose of this paper, was to examine the explanatory effect of non-financial public data when predicting default. Evidence produced by this study has shown, that such information does in fact provide some value. However, during the study several decisions were made which limited the models[8] ability to be extrapolated in general – mainly 1) the decision to limit the scope to A/S and ApS companies 2) the decision to aggregate various lifecycle statuses and 3) the decision to consistently use a two-year lag, resulting in a model which can only predict default in t based on data from t-2.

Nevertheless, if these decisions are kept in mind and the model is extrapolated according to the underlying assumptions, then **the usability of the model is fair/OK.** However, if the aim is to use the model for 'commercial' purposes then it lacks automation. Not only would it be necessary to establish a permanent link to the CBR database, it would require extensive code writing to produce calculations and to craft a user-friendly interface. However, it is not deemed impossible at all.

In sum, depending on how one intent to use the Full Model, it is doable provided you have a good/very good amount of knowledge about coding (API links and data processing) as well as how to handle large quantities of data in a data processing software (e.g. excel); as such, people uneducated in the necessary fields are unlike to gain any insight or results, even if experienced with corporate default prediction.

[8] "The model" refers to the combined, Full Model

11.2 Process Evaluation – a Helicopter Perspective

6 months. That is how long it has taken to collect, handle and analyze the publicly available data – financial and non-financial. Of these 6 months, 5 were dedicated solely to retrieving/extracting and handling the data; setting up API-link, formatting, re-formatting and combining data (coding). Thus, in relation to the above section "Usability of the model", it is concluded that crafting and continuous use of the model is relative difficult, yet doable.

The aim of this section is to reflect on the process of the study and how it could have otherwise been conducted. As outlined in Chapter 3 "Method", the process was lengthy which was mainly due to the collection and handling of data. The study demanded that all available data for 500.000 companies (financial and non-financial) were extracted and then subsequently narrowed down to 93.000 useful observations. Hence, another obvious possibility to conduct a similar study would be to limit the number of companies to be investigated/observations, as well as 'pairing' observations (e.g. one might have used 60 companies in total; 30 which had defaulted and 30 which had not. In such a case, it is almost obsolete to create a complex framework to extract and handle data; one could simply collect the annual reports manually. However, if conducted in this manner, all companies (observations) must be useful and as such, all figures must be available (no missing data). This would leave little room for random sampling.

The missing data was also an extensive hurdle to overcome in this study. Having decided to employ public available data directly from the source, instead of through an intermediary (such as Bisnode), issues arose in relation to locating the correct accounts. This, too, would have been much easier if the sample was much smaller and perhaps handled manually.

Moreover, the data extraction and handling required substantial computational processing power, and as such a standard computer is incapable of handling the amount of data analyzed. Hence, the procurement of a 'supercomputer' from Deloitte was essential for the study. A smaller sample size would have alleviated this issue.

In sum, without the use of a 'clean' database it is difficult to imagine that the study could have been conducted on the same scale more efficiently – if at all. Only very small process tweaks could, in hindsight, have been made. However, it may be possible to create the same sort of study and produce the same results and conclusion with a sample size 1/1550[9] of this study's. Further, this study employs

[9] 60 observations = 93.000 x 0,00064. 0,00064 = 1/1550

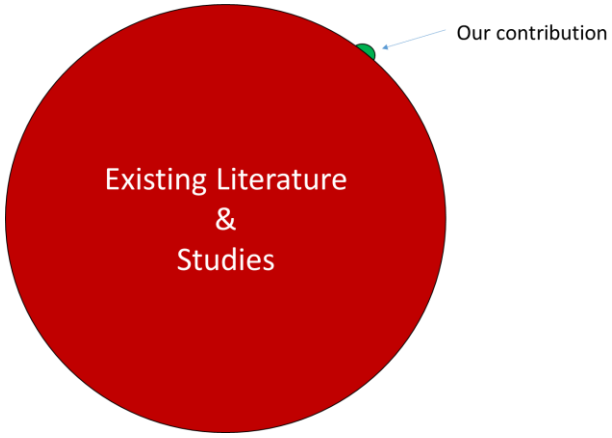
a relative large number of variables, which all must be present for an observation to be used. Therefore, lowering the number of variables would, all else equal, create more useful observations. The same is true if missing data was handled with imputation instead of deletion. This would also have allowed the model to be more widely extrapolated. It is difficult to conclude whether one method is superior to the other, but one relationship seems to be prevalent; a trade-off between sample size and data quality is – without the use of professional database providers' services – present when utilizing the Danish CBR database. Contrary, the smaller the sample size and use of e.g. 'pairing' observations, limit the conclusions and extrapolation possibilities of a model.

11.3 This Paper's Contribution to the Literature

Lykke et. al. (2004) correctly clarifies that *"[…] earlier studies such as Altman (1968), Altman et al. (1977), Dambolena and Khoury (1980), Hennawy and Morris (1983), Betts and Belhoul (1987), Platt and Platt (1991) failure-rate models are estimated on data not randomly selected and often consisting of less than 100 accounts. The companies in these studies are often "paired", so the active and failed companies have some of the same 9 characteristics (usually of the same size and from the same sector). This way of selecting data might cause selection bias. As stated in Shumway (2001), most of the existing literature on estimating failure-rate models is based on a single account from each company."*.

By such statement, this study arguably provides a substantial contribution to the literature; data is (to some degree) randomly selected; consists of substantially more accounts; studies are not 'paired'; model is based on more than a single account from each company.

Moreover, none of the above studies – apart perhaps from Lykke et. al. (2004) – incorporate the use of non-financial data into their prediction models on a large scale. As such, this study take a leap forward into a little investigated area of the literature.



Source: illustration by authors

However, the above statement is based on Lykke et. al. (2004)'s presentation of past and contemporary studies in 2004. Since then, studies seeking to investigate the value of non-financial data have (slowly) been published:

Author	Name of Study	Sample Size	Non-Financial Variable(s)	Year
M. Lykke, K. Pedersen, H. Vinther	A Failure-Rate Model for the Danish Corporate Sector	300.000	• Number of employees • Remarks from audits • Age	2004
J. Grunert, L. Norden, M. Weber	The role of non-financial factors in internal credit ratings	278	• Management Quality • Market Position	2005
E. Altman, G. Sabato, N. Wilson	The value of non-financial information in small and medium-sized enterprise risk management	2.5 mil	• AUDITED • Audit qualification: severe • Audit qualification: going concern • Late filing (log of days late) • No cash flow statement • CCJ number • CCJ real value • Log of age • Age 3–9 years • Subsidiary • Subsidiary negative net worth • Size (log) • Size squared (log) • Industry insolvency	2010
I. Pervan, T. Kuvek	The relative importance of financial ratios and non-financial variables in predicting of insolvency	825	• Quality of accounting information • Firm owners personal credit performance • Management quality	2013
T. Stenbäk	Corporate Default Prediction with Financial Ratios and Macroeconomic Variables	31.880	• GNI • Industry volume • Interest rate • Consumption • Consumer confidence on economy	2013

A. Olczyk, M. D. Nybjerg (not yet published)	Can Publicly Available Non-Financial Data Significantly Improve Corporate Default Prediction Accuracy?	93.000	• Age • Capital Injection • Corporation Type (Change in type) • Corporation Type (Status) • Number of employees • Main industry code (Change in code) • Main industry code (Status) • Name changes • Choice of no audit • Telephone number changes	2018

Source: table by authors

While the list above is not exhaustive, research in this area is relatively limited. This paper's study differentiate itself from the above in 1) the types variables employed 2) the number of financial and non-financial variables employed 3) its direct focus on default (not risk management, credit ratings, etc.) 4) its direct use of open public data (most of the above studies relies on databases which have been 'cleaned' prior to use and in general 'ready to use' data) 5) level of resources available and 6) its focus on Danish companies only.

In addition, this paper's data and findings are in line with similar studies such as Altman et al (2010) based on US corporations; that public available non-financial data is a significant contribution to default prediction models. Hence, with much fewer resources – compared to Altman et al. (2010) and Lykke et al. (2004) – the authors of this paper's study were able to reach *similar* conclusions and to some degree extent the knowledge within the field. This should thus provide a contribution to the literature, in that it in itself underlines a development in the field and usage of non-financial public data. It is therefore our sincere wish that this study is not only judged by its findings but also by the precedence it set in the *hands-on use* of public data.

11.4 Further Research

It is the belief of the authors, having been invested in literature review and prediction model creation, that the new topic of interest within prediction model creation is the use of AI (Artificial Intelligence) and ML (Machine Learning). We see some studies being published in present time which all praise the use of AI/ML, but at the same time conclude that their real benefit (power) is still somewhat to be revealed, i.e. it is a fairly young field. Preliminary studies have been conducted by Wang & Srinivasan (2015) on the use of AI in predicting energy needs; Goyal & Kaur (2016) on the use of ML in predicting loan risk (default); Bacham & Zhao, (2017) on the possibilities of utilizing ML for

credit risk modeling; Khandani, Adlar, & Lo (2010) on the use of ML in predicting consumer credit-risk models.

Insofar as the examples of studies given above is a relatively fair representation of the contemporary state of usage/knowledge creation within the AI/ML/prediction model sphere, then further research is encouraged. The initial papers published on the topic all conclude that the value of AI/ML is high. In relation to this paper's study, the use of AI/ML would be beneficial e.g. in detecting accounting manipulation/techniques. Also, it is possible that AI/ML can utilize multiple different trend analytics tools simultaneously (not only relying on LR) and thus increase the models' accuracy.

11.5 Summary

The models is evaluated to be fairly usable in practice if one is aware of the biases they may contain. However, due to the complex nature of extracting, modelling an analyzing the data, further development of the model, such as including more companies or newer information, is very difficult. This means that people not highly skilled in data- and computer science in reality, have very little usefulness of the open data in the CBR.

In hindsight very little could have been done differently with respects to the data handling and processing within the timeframe and resource limitations. The data is complex, unstructured and does require a lot of modelling and the methods and results are therefore thought of as very acceptable.

This paper have contributed to the literature within corporate default prediction by differentiating itself in multiple ways. The findings in the paper was in line with other studies within the field, proving both that this analysis is valid and that other studies results are replicable.

12 References

Agency for Digitalization. (2016). *A Stronger and More Secure Digital Denmark.* Agency for Digitalization.

Aharony, J., Jones, C., & Swary, I. (1980). An Analysis of Risk and Return Characteristics of Corporate Bankruptcy Using Capital Market Data. *The Journal of Finance, 35*(4), 1001-1016.

Ahn, B., Cho, S., & Kim, C. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications, 18*(2), 65-74.

Allison, P. D. (2002). *Missing Data.* London: Sage Publications.

Altman, E. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk*.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Default. *The Journal of Finance, 23*(4), 589-609.

Altman, E. I. (1977, June). Zeta analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance, 1*(1), 29-54.

Altman, E. I. (1995). *Emerging markets corporate bonds: A scoring system.* The New York University Salomon Center Series on Financial Markets and Institutions.

Altman, E. I. (2005). *Corporate financial distress and bankruptcy* (Vol. 1). Wiley.

Altman, E., & Rijken, H. (2011). Toward a Bottom-Up Approach to Assessing Sovereign Default Risk. *Journal of Applied Corporate Finance*.

Altman, E., & Saunders, A. (1997). Credit risk measurements: developments over the last 20 years. *Journnal of Banking and Finance, 21*(11-12), 1721-1742.

Amendola, A., Bisogno , M., Restaino, M., & Sensini, L. (2006). Forecasting corporate bankruptcy: empirical evidence on Italian data. *Euromed Journal of Business, 6*(3), 294-312.

American Academy of Acturaies. (2014). Fundamentals of Current Pension Funding and Accounting. *American Adacemy of Acturaies*.

Aoki, S., & Hosonuma, Y. (2004). *The Application of Econophysics: Proceedings of the Second Nikkei Econophysics Symposium.*

Aziz, A., Emanuel, D., & Lawson, G. (1988). Bankruptcy prediction - An investigation of cash flow based models. *Journal of Management Studies, 25*(5).

Aziz, M., & Dar, H. (2001). Predicting corporate bankruptcy where do we stand? *Corporate Governance: The international journal of business in society, 6*(1).

Bacham, D., & Zhao, J. (2017). Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. *Moody's Analytics Risk Perspective*.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons Ltd.

Beaver, W. H. (1966). Financial ratios as predictor of faliure. *Journal of Accounting Research, 4*, 71-111.

Berg, D. (2006). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry, 23*(2), 129-143.

Berton, J., Gorham, U., Jaeger, P., Sarin, L., & Choi, H. (2014). Big data, open government and e-government: Issues, policies and recommendations. *Information Polity, 19*(1), 5-16.

Bhargava, M., Dubelaar, C., & Scott, T. (1998). Predicting bankruptcy in the retail sector: an examination of the validity of key measures of performance. *Journal of Retailing and Consumer Services, 5*(2), 105-117.

Blum, M. (1974, Spring). Failing company discriminant analysis. *Journal of Accounting Research, 12*(1), 1-25.

Brédart, X. (2014). Bankruptcy Prediction Model Using Neural Networks. *Accounting and Finance Research, 3*(2).

CFA Institute. (2016). *Enron Corporation: Financial Scandals, Scoundrels & Crises*. Retrieved from CFA Institute: https://www.econcrises.org/2016/12/07/enron-corporation-2001/

CFA Institute. (2016, 12 7). *Enron Corporation: Financial Scandals, Scoundrels & Crisis*. Retrieved from CFA Institute: Bond/Debt Default / Equities / Fraud: https://www.econcrises.org/2016/12/07/enron-corporation-2001/

Charlie, A. (2017, 12). *Quarterly XBRL-based Public Company Financial Report Quality Measurement (Nov 2017)*. Retrieved from Digital Financial Reporting: http://xbrl.squarespace.com/journal/2017/12/1/quarterly-xbrl-based-public-company-financial-report-quality.html

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance, 8*(4), 537-569.

Chidi, G. (2002). *Global Crossing battles accounting controversy*. Retrieved from CNN.com: http://edition.cnn.com/2002/TECH/internet/02/12/global.crossing.probe.idg/index.html

Clark, T., & Weinstein, M. (1983). The Behavior of the Common Stock of Bankrupt Firms. *The Journal of Finance, 38*(2), 489-504.

CNN. (2002). Global Crossing battles accounting controversy. *CNN SCI-TECH*.

Coombs, H. M., Hobbs, D., & Jenkins, D. E. (2005). *Management accounting: principles and applications.* London: SAGE Publications .

Coombs, H., Hobbs, D., & Jenkins, D. (2005). *Management Accounting: Principles and Applications.* London: SAGE Publication.

Corporate Finance Institute (CFI). (2018, 05 07). *CFI*. Retrieved from CFI: Top Accounting Scandals: https://corporatefinanceinstitute.com/resources/knowledge/other/top-accounting-scandals/

Corporate Finance Institute (CFI). (2018). *Top Accounting Scandals*. Retrieved from Corporatefinanceinstitute.com: https://corporatefinanceinstitute.com/resources/knowledge/other/top-accounting-scandals/

Danish Business Authority. (2018). *Ny strategi skal gøre Danmark til digital frontløber.* Danish Business Authority.

Danish Business Registry. (2018). *CVR samler og udstiller data*. Retrieved from Erhvervsstyrelsen.dk: https://erhvervsstyrelsen.dk/hvad-er-cvr

Danmarks Statistik. (2018). *Statistikbanken.dk*. Retrieved from dst.dk: http://www.statistikbanken.dk/statbank5a/Graphics/MakeGraph.asp?menu=y&maintable=GF5&pxfile=20183201022542165406O7GF5.px&gr_type=0&PLanguage=0

Deakin, E. B. (1972). A Discriminant Analysis of Predictors of Business Faliure. *Journal of Accounting Research, 10*(1), 167-179.

Dewaelheyns, N., & Van Hulle, C. (2006). Corporate failure prediction modeling: Distorted by business groups' internal capital markets? *Journal of Business Finance & Accounting, 33*(5-6), 909-931.

Dunne, J., & Hughes, A. (1994). Age, Size, Growth and Survival: Uk Companies in the. *Journal of Industrial Economics, 42*(2), 114-140.

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

European Commision. (2017). *The Digital Economy and Society Index (DESI).* Retrieved from European Commision: https://ec.europa.eu/digital-single-market/en/desi

Financial Times. (2010, 05 13). *Financial Times*. Retrieved from Financial Times: Alphaville: Home: https://ftalphaville.ft.com/2010/05/13/229611/ten-easy-lessons-in-cooking-the-books/

Financial Times. (2010). *Retrieved from Financial Times: Alphaville*. Retrieved from ftalphaville.com: https://ftalphaville.ft.com/2010/05/13/229611/ten-easy-lessons-in-cooking-the-books/

Garson, D. G. (2015). *Missing Values Analysis & Data Imputation.* Asheboro, NC: Statisical Publishing Associates.

Gibbs, L., & Lazear, E. (2015). *Personal Economics in Practice.* John Wiley & Sons.

Gneiting, T., & Raftery, A. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association, 102*(477).

Grubbs, F. (1950). *The Annals of Mathematical Statistics.* The Institute of Mathematical Statistics.

Grunert, J., Norden, L., & Weber, M. (2005). The Role of Non-financial Factors in Internal Cretid Ratings. *Journal of Banking & Finance*, 509-531.

Günther, T., & Grüning, M. (2000). Einsatz von Insolvenzprognoseverfahren bei der Kreditwürdigkeitsprüfung im Firmenkundenbereich. *Die Betriebswirtschaft, 60*, 39-59.

Hair, J., R., A., R., T., & Black, W. (1995). *Multivariate Data Analysis* (3rd ed.). Macmillan.

Hardy, K., & Maurushat, A. (2017, Feb.). Opening up government data for Big Data analysis and public benefit. *Computer Law & Security Review, 33*(1), 30-37.

Hillgeist, S., Keating, E., Cram, D., & Lundsted, K. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies, 9*(1), 5-34.

Huang, S., Tang, Y., Lee, C., & Chang, M. (2012). Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress prediction. *Expert Systems with Applications, 39*(3), 3855-3861.

Hui Li, J. (2011). Principal component case-based reasoning ensemble for business failure prediction. *Information & Management, 48*(6), 220-227.

Investopedia. (2018). *Top 8 Ways Companies Cook the Books*. Retrieved from Investopedia.com: https://www.investopedia.com/articles/analyst/071502.asp

Investopedia. (2018, 04 02). *Top 8 Ways Companies Cook the Books* . Retrieved from Investopedia: https://www.investopedia.com/articles/analyst/071502.asp

Ishikawa , A., Fujimoto, S., Mizuno, T., & Watanabe, T. (2014). Firm Age Distributions and the Decay Rate of Firm Activities. *The International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014*, (pp. 187-194).

Jeter, D., & Chaney, P. (2011). *Advanced Accounting.* John Wiley.

Kaur, R., & Goyal, A. (2016). Accuracy Prediction for Loan risk Using Machine Learning Models. *International Journal of Computer Science Trends and Technology*.

Keasey, K., & Watson, R. (1991). Financial Distress Prediction Models: A Review of Their Usefulness. *British Journal of Management, 2*(2), 89-102.

Khandani, A., Adlar, K., & Lo, A. (2010). Consumer Credit-Risk Models Via Machine-Learning Algorithms. *Journal of Banking & Fianance*.

Kristóf, T. (2008). Gazdasági szervezetek fennmaradásának és fizetőképességének előrejelzése. Corvinus University of Budapest.

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

Law, D., & Roache, S. (2015). *Assessing Default Risks for Chinese Firms: A Lost Cause?* International Monetary Fund.

Lennox, C. (1999, Jul.-Aug.). Identifying Failing Companies: A Reevaluation of the Logit, Probit and DA Approaches. *Journal of Economics and Business, 51*(4), 347-364.

Lykke, M., Pedersen, K., & Vinther, H. (2004). Danmarks Nationalbank Working Paper 16. National Banken.

Lærd Statistics. (2018). *Binomial Logistic Regression using SPSS Statistics*. Retrieved from statistics.laerd.com: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

Massman, C., Bell, G., & Turtle, H. (1998). The comparison of enterprise bankruptcy forecasting method. *The Financial Review, 33*(2), 34-54.

Miller, W. (2009). *Comparing models of corporate bankruptcy prediction: distance to default vs. Z-score.* Morningstar Inc.

Miller, W. (2009). Comparing Models of Corporate Bankruptcy Prediction: Distance to Default vs. Z-Score. *SSRN Electronic Journal*.

Nathan, I. (2018, Feb.). *Rekordmange konkurser blandt hovedstadens restauranter*. Retrieved from DR.dk: https://www.dr.dk/nyheder/regionale/hovedstadsomraadet/rekordmange-konkurser-blandt-hovedstadens-restauranter

OECD. (2017). *Open Government Data*. Retrieved from OECD.org: http://www.oecd.org/gov/digital-government/open-government-data.htm

Offshore Energy Today. (2017, 11 7). *Offshore Energy Today: Business Guide*. Retrieved from Offshore Energy Today: Business Guide: https://www.offshoreenergytoday.com/impairments-drag-maersk-drilling-to-the-red/

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131.

Pawlak, Z., & Sowinski, R. (1994). Rough set approach to multi-attribute decision analysis. *European Journal of Operational Research, 72*(3), 443-459.

Pervan, I., & Kuvek, T. (2013). The Relative Importance of FInancial Ratios and Non-Financial Variables in Perdicting of Insolvency. *Croation Operational Research Review*.

Platt, & Platt. (1990, Mar.). Development of a Class of Stable Predictive Variables. *Journal of Business Finance & Accounting, 17*(1), 31-51.

R., M. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance, 29*(2), 449-470.

Schilit, H. (2002). *Financial Schenanigans: How to Detect Accounting Gimmicks & Fraud in Financial Reports.* McGraw-Hill.

Shah, J., & Murtaza, M. (2000). A neural network based clustering procedure for bankruptcy prediction. *Latin American Business Review, 18*(2).

Sharma, S. (1995). *Applied multivariate techniques.* John Wiley & Sons.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business, 74*(1), 101-124.

Stenbäk, T. (2013). Corporate Default Prediction with Financial Ratios and Macroeconomic Variables. *Working Paper*.

Stergiou, C., & Siganos, D. (1996). *Neural Networks*. Retrieved from Imperical College London Department of Computing: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

Studenmund, A. (2006). *Using Econometrics, a Practical guide.* (5th, Ed.) NYC: Addison-Wesley.

The Wall Street Journal. (2014). *Welcome to the World of Pension Smoothing.* The Wall Street Journal.

The Wall Street Journal. (2014). Welcome to the World of 'Pension Smoothing. *The Wall Street Journal*.

Van Peursem, K., & Pratt, M. (2002). A new Zealand failure prediction model: development and international implications. *Advances in International Accounting, 15*, 229-247.

Vazza, D., & Kraemer, N. (2016). *Default, Transition, and Recovery: 2016 Annual Global Corporate Default Study And Rating Transitions.* SPG Global Report.

W., B., & McNichols, M. (2005). Have financial statements become less informative? evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies, 10*(1), 93-122.

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications, 41*(5), 2353-2361.

Wang, Z., & Srinivasan, R. (2015). A Review of Artifical Intellegence Based Building Energy Prediction with a Focus on Ensemble Prediciton Model.

World Bank Group. (2017). *Big Data in Action for Governments.* World Bank Group.

XBRL.org. (2018). *An Introduction to XBRL*. Retrieved from XBRL.org: https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/

Xiaosi, X., Ying, C., & Haitao, Z. (2011). The comparison of enterprise bankruptcy forecasting method. *Journal of Applied Statistics, 38*(2), 301-308.

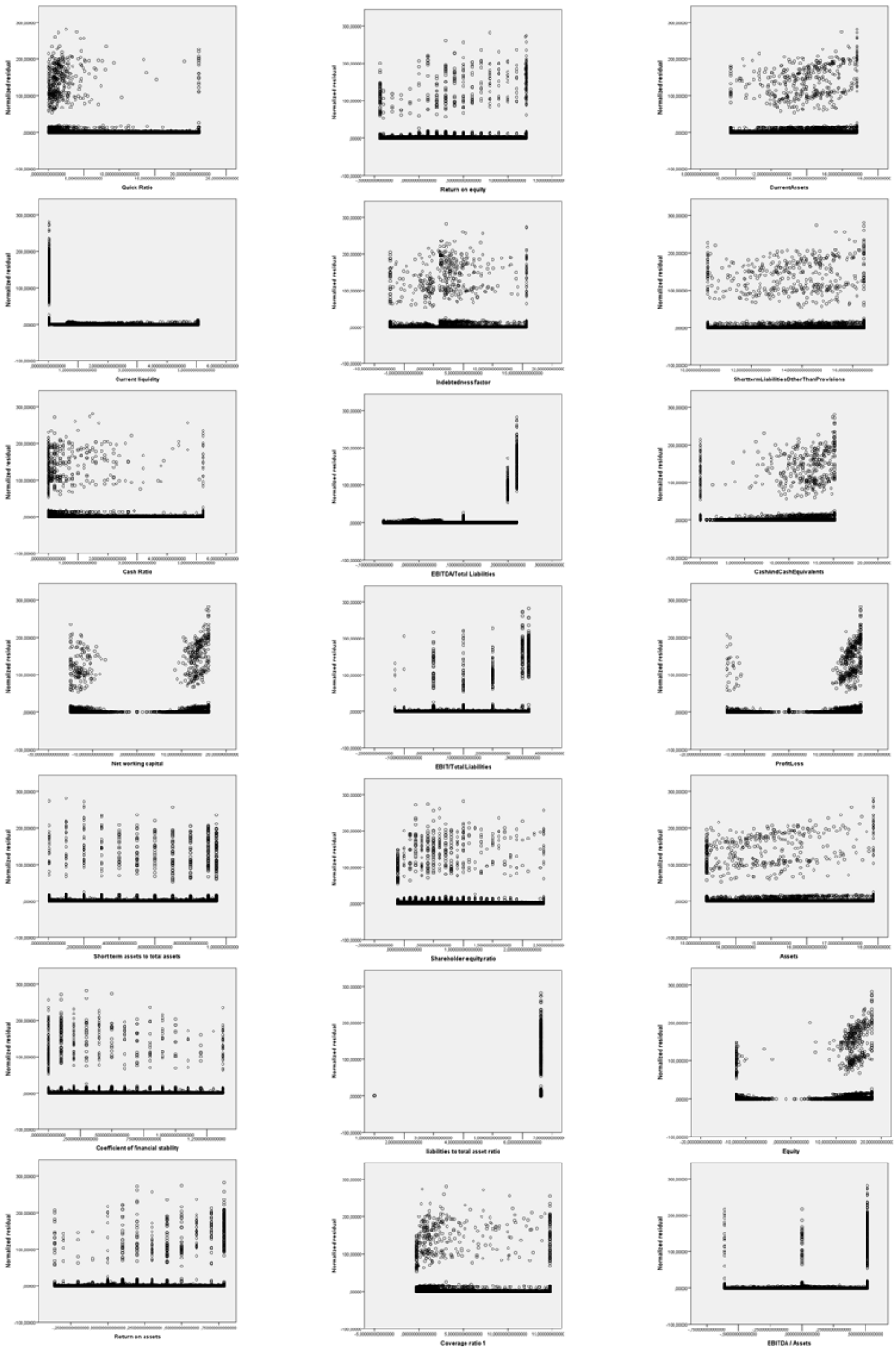
Zmijewsk, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research, 22*, 59-82.

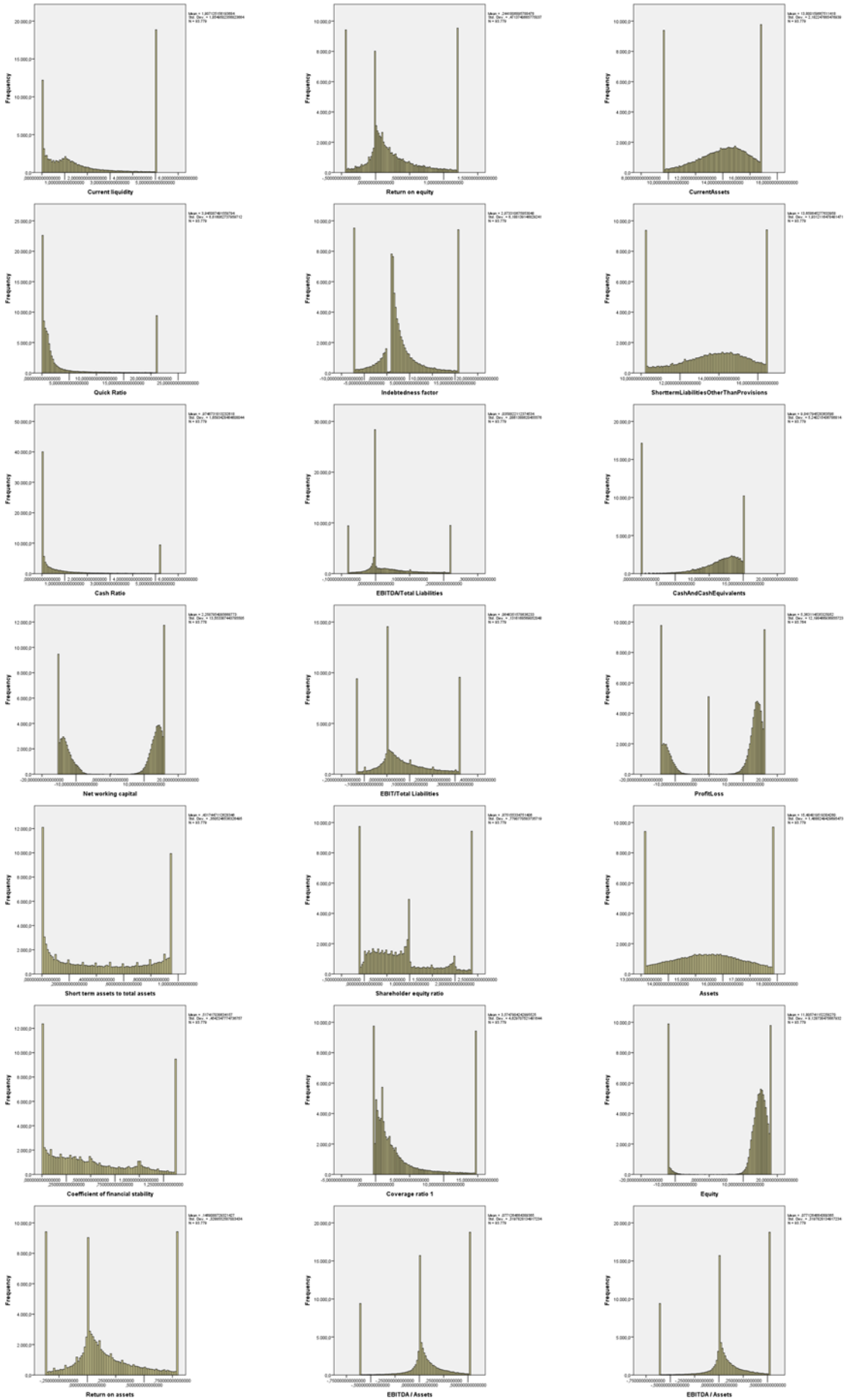
13 Appendix

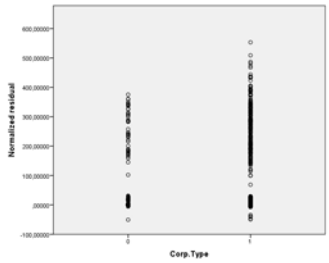
13.1 Appendix 1 – Industry Codes

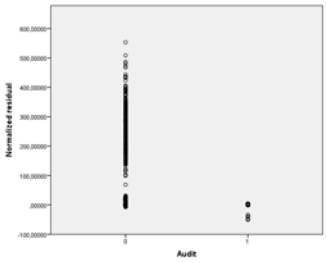
Number	Description
1	A Landbrug, jagt, skovbrug og fiskeri
2	B Råstofindvinding
3	C Fremstillingsvirksomhed
4	D El-, gas-, fjernvarmeforsyning
5	E Vandforsyning; kloakvaesen, affaldshåndtering mv
6	F Bygge- og anlaegsvirksomhed
7	G Engrosh og detailh; rep af motorkør og motorcykler
8	H Transport og godshåndtering
9	I Overnatningsfaciliteter og restaurationsvirksomhed
10	J Information og kommunikation
11	K Pengeinstitut- og finansvirksomhed mv
12	L Fast ejendom
13	M Liberale, videnskabelige og tekniske tjenesteydelser
14	N Administrative tjenesteydelser og hjaelpetjenester
15	O Offentlig forvaltning og forsvar; socialsikring
16	P Undervisning
17	Q Sundhedsvaesen og sociale foranstaltninger
18	R Kultur, forlystelser og sport
19	S Andre serviceydelser
20	T Private husholdninger med ansat medhjaelp mv
21	U Ekstraterritoriale organisationer og organer
22	(X) Uoplyst

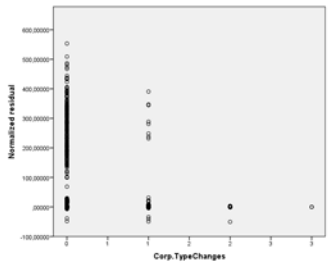
13.2 Appendix 2 – Scatterplots and Histograms

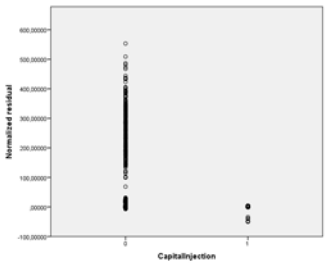


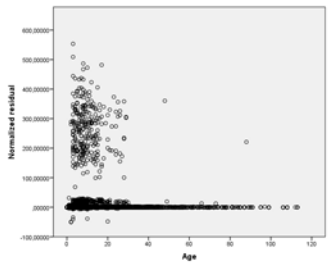

Frequency
Current liquidity
Return on equity
CurrentAssets
Quick Ratio
Indebtedness factor
ShorttermLiabilitiesOtherThanProvisions
Cash Ratio
EBITDA/Total Liabilities
CashAndCashEquivalents
Net working capital
EBIT/Total Liabilities
ProfitLoss
Short term assets to total assets
Shareholder equity ratio
Assets
Coefficient of financial stability
Coverage ratio 1
Equity
Return on assets
EBITDA / Assets
EBITDA / Assets

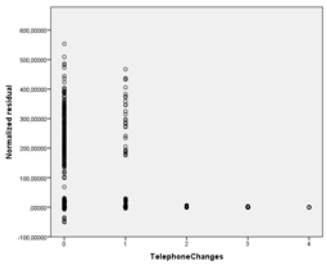

Normalized residual
Corp.Type

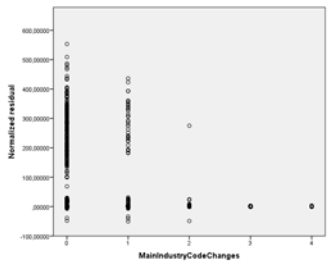

Normalized residual
Audit

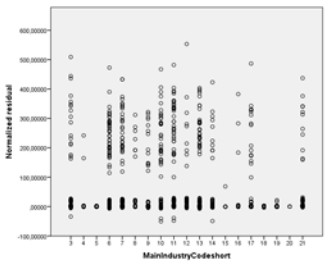

Normalized residual
Corp.TypeChanges

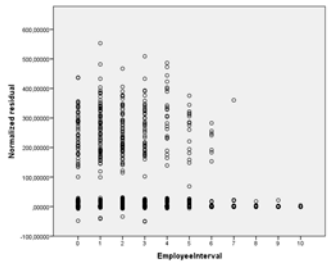

Normalized residual
CapitalInjection

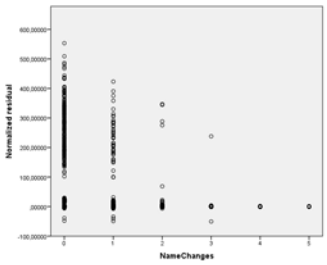

Normalized residual
Age


Normalized residual
TelephoneChanges

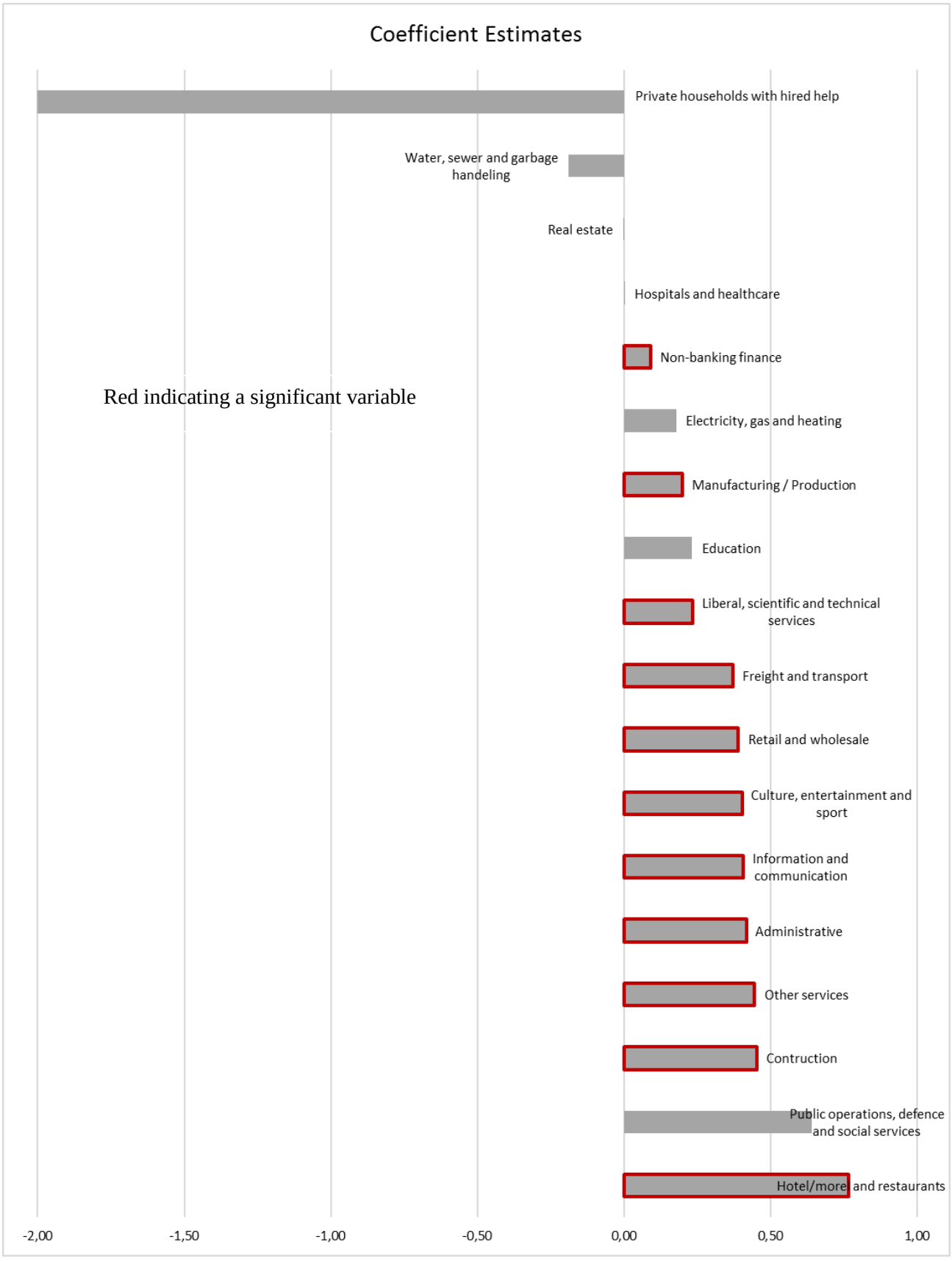

Normalized residual
MainIndustryCodeChanges

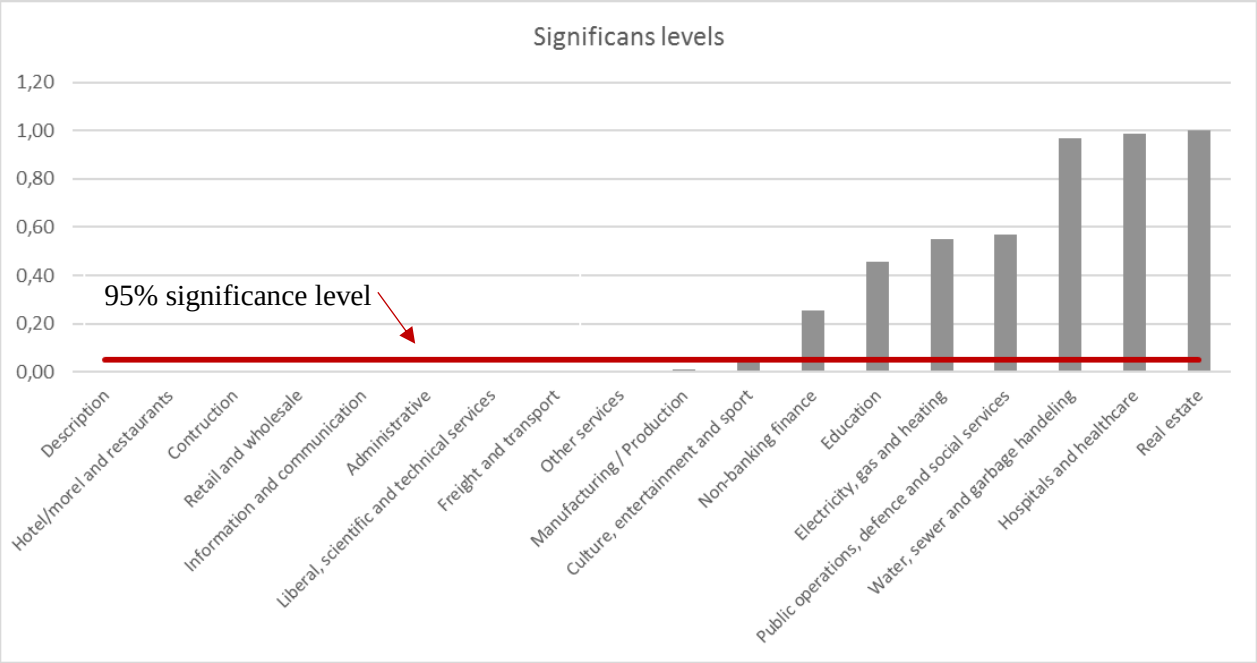

Normalized residual
MainIndustryCodeshort


Normalized residual
EmployeeInterval


Normalized residual
NameChanges

13.3 Appendix 3 – Coefficients' Estimates for Full Model




Significans levels
1,20
1,00
0,80
0,60
0,40
0,20
0,00
95% significance level
Description
Hotel/morel and restaurants
Contruction
Retail and wholesale
Information and communication
Administrative
Liberal, scientific and technical services
Freight and transport
Other services
Manufacturing / Production
Culture, entertainment and sport
Non-banking finance
Education
Electricity, gas and heating
Public operations, defence and social services
Water, sewer and garbage handeling
Hospitals and healthcare
Real estate

13.4 Appendix 4 – Red Flag Checks – ways to detect accounting manipulations

1. Cash and equivalents as percentage of total assets declines form prior period	Liquidity issues
2. Receivables grow substantially faster than sales	Perhaps aggressive revenue recognition-recording revenue too soon or granting extended credit terms to customers
3. Receivables grow substantially slower than sales	Receivables may have been reclassified as another asset category
4. Bad debt reserves decline relative to gross receivables	Under-reserving and inflating operating income
5. Unbilled receivables grow faster than sales or billed receivables	A greater portion of revenue may be coming from sales under the percentage-of-completion method
6. Inventory grows substantially faster than sales, cost of sales, or accounts payable	Inventory may be obsolete, requiring a write-off; company may have failed to charge the cost of sales on some sales
7. Inventory reserves decline relative to inventory	Under-reserving and inflating operating income
8. Prepaid expenses shoot up relative to total assets	Perhaps improperly capitalizing certain operating expenses
9. Other assets rise relative to total assets	Perhaps improperly capitalizing certain operating expenses
10. Gross plant and equipment increases sharply relative to total assets	Perhaps capitalizing maintenance and repair expenses
11. Gross plant and equipment declines sharply relative to total assets	Failing to invest in new plant and equipment
12. Accumulated depreciation declines as gross plant and equipment rises	Failing to take sufficient depreciation charge - inflating operating income
13. Intangible assets rise sharply relative to total assets	Perhaps tangible assets were reclassified into intangibles to avoid expensing them in future periods
14. Accumulated amortization declines as intangibles rise	Failing to take sufficient amortization charge - inflating operating income

15. Growth in accounts payable substantially exceeds revenue growth	Failed to pay off current debts for inventory and supplies- will require larger cash outflow in future period
16. Accrued expenses decline relative to total assets	Perhaps company released reserves – inflating operating income
17. Deferred revenue declines while revenue increases	Either new business is slowing or company released some reserves to inflate revenue
18. Cost of goods sold grows rapidly relative to sales	Pricing pressure results in lower gross margins
19. Cost of goods sold declines rapidly relative to sales	Company may have failed to transfer the entire cost of the product form inventory
20. Cost of goods sold fluctuates widely from quarter to quarter relative to sales	Unstable gross margin could indicate accounting irregularities
21. Operating expenses decline sharply relative to sales	Perhaps improperly capitalizing certain operating expenses
22. Operating expenses risk significantly relative to sales	Company may have become less efficient, sending more for each unit sold
23. Major portion of pretax income comes from one-time gains	Core business may be weakening
24. Interest expense rises materially relative to long-term debt	Higher cash outflow expected
25. Interest expense declines materially relative to long-term debt	Perhaps improperly capitalizing interest expense
26. Amortization of software costs grows more slowly than capitalized costs	Perhaps improperly capitalizing certain operating expenses
27. Cash flow from operations materially lags behind net income	Quality of earnings may be suspect or expenditure for working capital may have been too high
28. Company fails to disclose details of cash flow from operations	Company may be trying to hide the source of the operating cash problem

29. Cash inflows come primarily from asset sales, borrowing, or equity offerings	Signs of weakness, especially if cash comes exclusively from asset sales, borrowing, or equity offerings

Source: (Schilit, Financial Shenanigans: How To Detect Accounting Gimmicks & Fraud in Financial Reports, 2002)

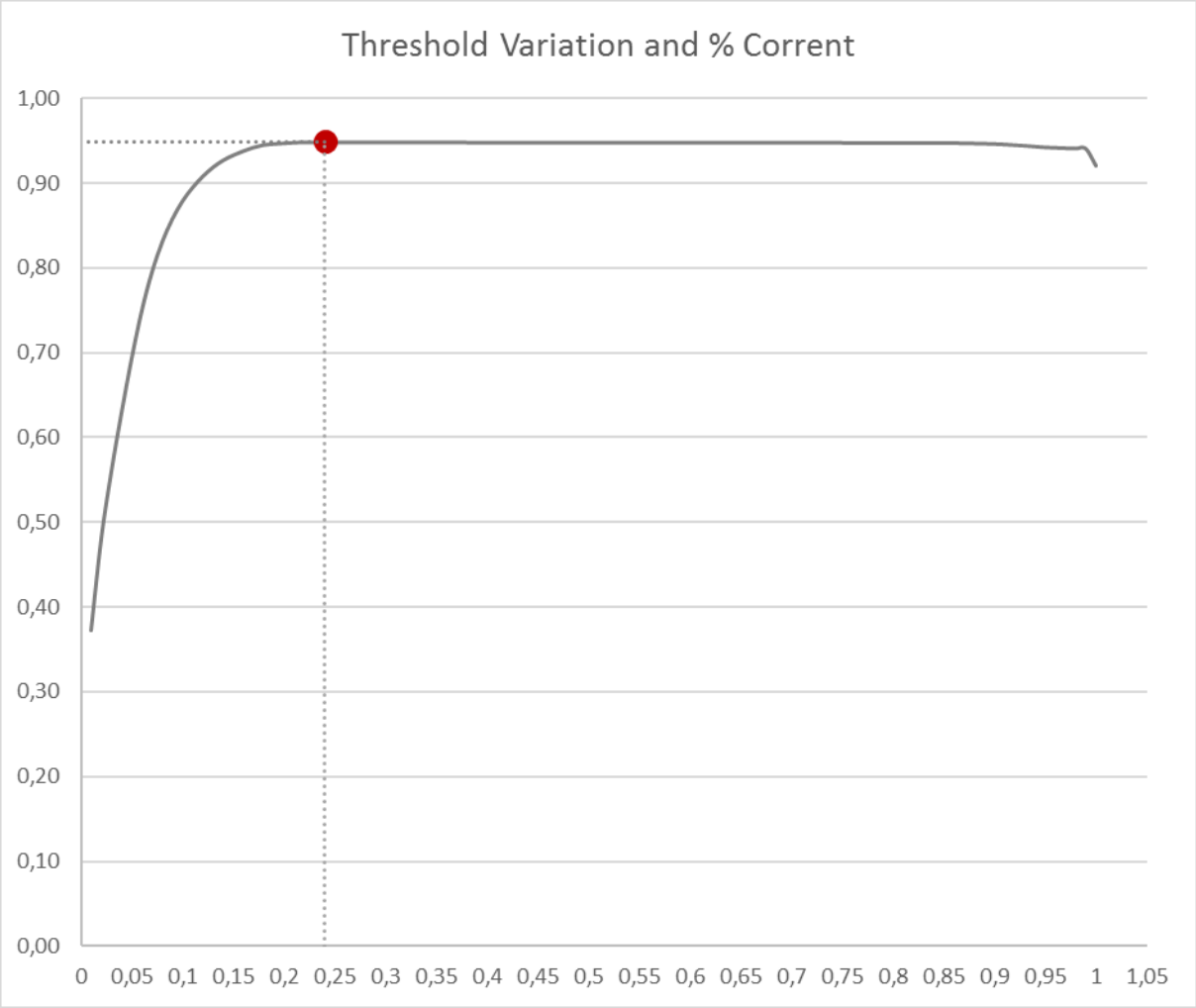
13.5 Appendix 5 – VIF Test Full Model

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1,218	,028		42,792	,000		
	Current liquidity	-,019	,001	-,129	-31,520	,000	,462	2,162
	Quick Ratio	-1,954E-5	,000	,000	-,089	,929	,269	3,715
	Cash Ratio	-,001	,001	-,006	-1,412	,158	,371	2,694
	Net working capital	,000	,000	-,020	-4,241	,000	,333	3,001
	Short term assets to total assets	,032	,005	,042	6,950	,000	,212	4,710
	Coefficient of financial stability	,002	,003	,003	,640	,522	,355	2,816
	Return on assets	-,001	,005	-,001	-,200	,841	,183	5,474
	Return on equity	-,006	,002	-,011	-3,150	,002	,662	1,511
	Indebtedness factor	,000	,000	,004	1,407	,159	,843	1,186
	EBITDA/Total Liabilities	-,881	,013	-,281	-68,422	,000	,458	2,183
	EBIT/Total Liabilities	-,051	,013	-,025	-4,064	,000	,207	4,834
	Shareholder equity ratio	,000	,002	,000	,100	,920	,321	3,115
	liabilities to total asset ratio	-,141	,004	-,100	-36,011	,000	,996	1,004
	Coverage ratio 1	-2,725E-5	,000	,000	-,105	,917	,388	2,575
	CurrentAssets	-,002	,001	-,016	-2,322	,020	,154	6,501
	ShorttermLiabilitiesOtherThanProvisions	-,003	,001	-,025	-3,673	,000	,171	5,836
	CashAndCashEquivalents	-,001	,000	-,018	-4,845	,000	,590	1,695
	ProfitLoss	-,001	,000	-,028	-6,029	,000	,358	2,796
	Assets	-,003	,001	-,016	-2,669	,008	,204	4,898
	Equity	-,001	,000	-,040	-8,161	,000	,317	3,156
	EBITDA / Assets	-,005	,004	-,006	-1,284	,199	,399	2,508
	Corp. Type	-,015	,002	-,020	-6,315	,000	,757	1,320
	Corp. Type Changes	-,034	,003	-,047	-9,990	,000	,355	2,817
	Age	,001	,000	,018	5,534	,000	,737	1,357
	Main Business Area Changes	-,012	,002	-,026	-6,658	,000	,525	1,906
	Employee Interval	-,001	,000	-,007	-2,523	,012	,885	1,131
	Name Changes	,024	,002	,047	11,304	,000	,439	2,280
	Audit	,768	,006	,452	127,290	,000	,613	1,632
	Capital Injection	-,110	,007	-,056	-15,932	,000	,625	1,600
	Telephone Changes	-,016	,002	-,024	-8,366	,000	,948	1,054
	Main Industry Code (short)	,000	,000	-,004	-1,367	,172	,755	1,325

a. Dependent Variable: EVENT

13.6 Appendix 6 – Threshold Variation (Cutoff Value)



Classification Table[a]

			Predicted		
			EVENT		Percentage Correct
	Observed		0	1	
Step 1	EVENT	0	84776	1548	98,2
		1	3260	4179	56,2
	Overall Percentage				94,9

a. The cut value is ,240

13.7 Appendix 7 – Abbreviations

ABBREVIATION	DESCRIPTION
AI	ARTIFICIAL INTELLIGENCE
API	APPLICATION PROGRAMMING INTERFACE
CBR	CENTRAL BUSINESS REGISTER
CPM	CONDITIONAL PROBABILITY MODEL
CVR-number	CENTRAL BUSINESS REGISTER NUMBER (CORPORATE ID NUMBER)
DT	DECISION TREE
EBIT	EARNINGS BEFORE INTEREST AND TAX
EBITDA	EARNINGS BEFORE INTEREST, TAX, DEPRECIATION AND AMORTIZATION
EVENT STATUSES (DANISH)	EVENT STATUSES (ENGLISH)
-SLETTET	-DELETED
-OPLØST EFTER FRIVILLIG LIKVIDATION	-DISOLVED BY VOLUNTARY LIQUIDATION
-UNDER FRIVILLIG LIKVIDATION	-VOLUNTARY LIQUIDATION
-OPLØST EFTER ERKLÆRING	-DISOLVED BY DECLARATION
-OPLØST EFTER SPALTNING	-DISOLVED BY DIVISIONS
-UNDER TVANGSOPLØSNING	-FORCED DISOLVEMENT
-UNDER REKONSTRUKTION	-RECONSTRUCTION
-OPLØST EFTER KONKURS	-DISOLVED BY DEFAULT
-OPLØST EFTER FUSION	-DISOLVED BY MERGER
-UNDER REASSUMERING	-REASSUMPTION
-UDEN RETSVIRKNING	-NOT A LEGAL ENTITY
-UNDER KONKURS	-DISOLVED BY DEFAULT
-TVANGSOPLØST	-FORCED DISOLVEMENT
GAAP	GENERAL ACCEPTED ACCOUNTING PRACTICE
IFRS	INTERNATIONAL FINANCIAL REPORTING STANDARDS
JSON	JAVASCRIPT OBJECT NOTATION
LD	LISTWISE DELETION
LR	LOGISTIC REGRESSION
LS	LOG SCORE
MBS	MARKET BASED STRUCTURE MODELS
MDA	MULTIVARIATE DISCRIMINANT ANALYSIS
ML	MACHINE LEARNING
NN	NEURAL NETWORK
OCF	OPERATING CASH FLOW
ROA	RETURN ON ASSETS
ROC	RECEIVER OPERATING CHARACTERISTICS
ROE	RETURN ON EQUITY
RS	ROUGH SET
UDA	UNIVARIATE DISCRIMINANT ANALYSIS
VIF	VARIANCE INFLATION FACTOR
XBRL	EXTENSIBLE BUSINESS REPORTING LANGUAGE