Constructing a predictive model for assessing the bankruptcy risk of Finnish SMEs

Iiro Sokka

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Zeist, the Netherlands 07.04.2020

Supervisor

Prof. Markku Maula

Advisor

Esa Mäkeläinen, M.Sc. (Econ.)

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Iiro Sokka

Title Constructing a predictive model for assessing the bankruptcy risk of Finnish SMEs

Degree programme Industrial Engineering and Management

Major Strategy and Venturing **Code of major** SCI3050

Supervisor Prof. Markku Maula

Advisor Esa Mäkeläinen, M.Sc. (Econ.)

Date 07.04.2020 **Number of pages** 99+7 **Language** English

Abstract

Bankruptcy prediction is a subject of significant interest to both academics and practitioners because of its vast economic and societal impact. Academic research in the field is extensive and diverse; no consensus has formed regarding the superiority of different prediction methods or predictor variables. Most studies focus on large companies; small and medium-sized enterprises (SMEs) have received less attention, mainly due to data unavailability. Despite recent academic advances, simple statistical models are still favored in practical use, largely due to their understandability and interpretability.

This study aims to construct a high-performing but user-friendly and interpretable bankruptcy prediction model for Finnish SMEs using financial statement data from 2008–2010. A literature review is conducted to explore the key aspects of bankruptcy prediction; the findings are used for designing an empirical study. Five prediction models are trained on different predictor subsets and training samples, and two models are chosen for detailed examination based on the findings.

A prediction model using the random forest method, utilizing all available predictors and the unadjusted training data containing an imbalance of bankrupt and non-bankrupt firms, is found to perform best. Superior performance compared to a benchmark model is observed in terms of both key metrics, and the random forest model is deemed easy to use and interpretable; it is therefore recommended for practical application. Equity ratio and financial expenses to total assets consistently rank as the best two predictors for different models; otherwise the findings on predictor importance are mixed, but mainly in line with the prevalent views in the related literature.

This study shows that constructing an accurate but practical bankruptcy prediction model is feasible, and serves as a guideline for future scholars and practitioners seeking to achieve the same. Some further research avenues to follow are recognized based on empirical findings and the extant literature. In particular, this study raises an important question regarding the appropriateness of the most commonly used performance metrics in bankruptcy prediction. Area under the precision-recall curve (PR AUC), which is widely used in other fields of study, is deemed a suitable alternative and is recommended for measuring model performance in future bankruptcy prediction studies.

Keywords bankruptcy prediction, credit risk, machine learning, SMEs

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Iiro Sokka	
Työn nimi Ennustemallin kehittäminen suomalaisten PK-yritysten konkurssiriskin määritykseen	
Koulutusohjelma Industrial Engineering and Management	
Pääaine Strategy and Venturing	**Pääaineen koodi** SCI3050
Työn valvoja Prof. Markku Maula	
Työn ohjaaja Esa Mäkeläinen, KTM	
Päivämäärä 07.04.2020	**Sivumäärä** 99+7 **Kieli** Englanti

Tiivistelmä

Konkurssien ennustaminen on taloudellisten ja yhteiskunnallisten vaikutustensa vuoksi merkittävä aihe akateemisesta ja käytännöllisestä näkökulmasta. Alan tutkimus on laajaa ja monipuolista, eikä konsensusta parhaiden ennustemallien ja -muuttujien suhteen ole saavutettu. Valtaosa tutkimuksista keskittyy suuryrityksiin; pienten ja keskisuurten (PK)-yritysten konkurssimallinnus on jäänyt vähemmälle huomiolle. Akateemisen tutkimuksen viimeaikaisesta kehityksestä huolimatta käytännön sovellukset perustuvat usein yksinkertaisille tilastollisille malleille johtuen niiden paremmasta ymmärrettävyydestä.

Tässä diplomityössä rakennetaan ennustemalli suomalaisten PK-yritysten konkurssiriskin määritykseen käyttäen tilinpäätösdataa vuosilta 2008–2010. Tavoitteena on tarkka, mutta käyttäjäystävällinen ja helposti tulkittava malli. Konkurssimallinnuksen keskeisiin osa-alueisiin perehdytään kirjallisuuskatsauksessa, jonka pohjalta suunnitellaan empiirinen tutkimus. Viiden mallinnusmenetelmän suoriutumista vertaillaan erilaisia opetusaineiston ja ennustemuuttujien osajoukkoja käyttäen, ja löydösten perusteella kaksi parasta menetelmää otetaan lähempään tarkasteluun.

Satunnaismetsä (random forest) -koneoppimismenetelmää käyttävä, kaikkia saatavilla olevia ennustemuuttujia ja muokkaamatonta, epäsuhtaisesti konkurssi- ja ei-konkurssitapauksia sisältävää opetusaineistoa hyödyntävä malli toimii parhaiten. Keskeisten suorituskykymittarien valossa satunnaismetsämalli suoriutuu käytettyä verrokkia paremmin, ja todetaan helppokäyttöiseksi ja hyvin tulkittavaksi; sitä suositellaan sovellettavaksi käytäntöön. Omavaraisuusaste ja rahoituskulujen suhde taseen loppusummaan osoittautuvat johdonmukaisesti parhaiksi ennustemuuttujiksi eri mallinnusmetodeilla, mutta muilta osin havainnot muuttujien keskinäisestä paremmuudesta ovat vaihtelevia.

Tämä diplomityö osoittaa, että konkurssiennustemalli voi olla sekä tarkka että käytännöllinen, ja tarjoaa suuntaviivoja tuleville tutkimuksille. Empiiristen havaintojen ja kirjallisuuslöydösten pohjalta esitetään jatkotutkimusehdotuksia. Erityisen tärkeä huomio on se, että konkurssiennustamisessa tyypillisesti käytettyjen suorituskykymittarien soveltuvuus on kyseenalaista konkurssitapausten harvinaisuudesta johtuen. Muilla tutkimusaloilla laajasti käytetty tarkkuus-saantikäyrän alle jäävä pinta-ala (PR AUC) todetaan soveliaaksi vaihtoehdoksi, ja sitä suositellaan käytettäväksi konkurssimallien suorituskyvyn mittaukseen.

Avainsanat konkurssien ennustaminen, luottoriski, koneoppiminen, PK-yritykset

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and research motivation

Corporate bankruptcy and other types of financial failure incur significant costs not only on the failing company and its stakeholders, but also on the wider business ecosystem and economy at large (Alaka et al., 2018; Bauweraerts, 2016), and predicting the occurrence of bankruptcies has long been a prominent topic in business literature and an integral part of credit risk modeling. The subject particularly garnered attention in the wake of the 2007–08 financial crisis that exposed vulnerabilities in financial systems and highlighted the importance of credit risk management (Florez-Lopez & Ramon-Jeronimo, 2015; Gupta & Chaudhry, 2019).

Despite increased awareness of the costs of bankruptcy and the promising results achieved with novel techniques in recent literature, most financial institutions still rely on traditional statistical models (Zhang & Thomas, 2015). As technology is becoming a more and more integral part of the finance industry, the adoption of modernized bankruptcy prediction models into practice is both easier and more necessary than ever. New means of financing, such as crowdfunding and peer-to-peer lending, have emerged in recent years as viable alternatives to traditional bank loans, particularly among SMEs (Kgoroeadira et al., 2019; Xiang et al., 2018). As competition among credit institutions increases, efficient credit risk management, especially in times of economic instability and crises, is a vital factor in remaining relevant in the industry.

Small and medium-sized enterprises (SMEs) comprise the vast majority of companies globally; in the EU, they accounted for 99% of all enterprises in 2015 (Papadopoulos et al., 2015). SMEs are crucial to the development of national economies as well as the global economy as a whole (Gupta et al., 2018) and play a particularly vital role in job creation (de Wit & de Kok, 2014), especially during periods of economic downturn and overall high unemployment (Moscarini & Postel-Vinay, 2012). The importance of SMEs has been recognized in the EU, for example through the Small Business Act (European Commission, 2008). Nonetheless, smaller companies often have difficulty obtaining credit (Beck et al., 2006), partially because small business loans are subject to disproportionately high capital requirements under the Basel III framework (Bams et al., 2019).

Corporate failure prediction studies are mostly focused on large companies (Gordini, 2014), mainly due to the better availability of financial data (Ciampi, 2015) as well as the possibility of using market-based information where listed companies are concerned (Filipe et al., 2016). This has impeded the development of SME bankruptcy prediction literature, even though it is widely recognized that they should be treated separately from larger firms in credit risk modeling (Altman et al., 2010; Figini et al., 2017). Improving SME bankruptcy prediction models would be mutually beneficial: it could help financial institutions reduce their risk exposure, and thereby improve SMEs' financing options through lower risk premia (Tobback et al., 2017).

Data scarcity is a common issue in bankruptcy prediction: most recent studies employ data samples of only 400 firms or fewer (Veganzones & Séverin, 2020). Barboza et al. (2017) note that in many cases the available sample is limited in other ways in addition to small size, for example to the clients of a specific credit institution; this may induce bias and distort the results. Additionally, the majority of literature is focused on a handful of large economies such as the US, China, South Korea, France, and the UK (see e.g. Alaka et al., 2018).

Despite extensive research, findings on the key determinants of bankruptcy are largely inconclusive and contradictory; there is strong evidence that the best predictor variables differ significantly between data samples (Balcaen & Ooghe, 2006), and research also shows notable differences between countries in terms of the effectiveness of predictor variables and their interactions (Filipe et al., 2016; Laitinen & Lukason, 2014). Findings concerning a specific population of firms cannot therefore be assumed to apply to others, and prediction models may not be accurate outside their original context. To obtain reliable information regarding the population of interest, targeted study is needed.

The choice of prediction technique is subject to much debate, and numerous alternatives have been proposed. While machine learning methods typically outperform classic statistical models (Ravi Kumar & Ravi, 2007; Veganzones & Séverin, 2020), there is considered to be a trade-off between predictive performance and model interpretability (Alaka et al., 2018; Virág & Nyitrai, 2014). The explanatory factors of bankruptcy are a subject of interest, and therefore "black box" models may be of limited usefulness. Business practitioners usually prefer easily understandable and interpretable tools, even at the expense of a slightly higher predictive performance (Jones et al., 2017; Sun et al., 2014). Despite this, the current trend in academic literature seems to be towards increasingly complex and technically sophisticated methods (see e.g. Song & Peng, 2019; Sun et al., 2020; Zhang et al., 2019), while the practical usability of the prediction model is disregarded.

The main motivation for the research topic of this thesis stems from the fact that the practice of bankruptcy prediction seems to be lagging far behind the academic progress of recent years. Few scholars address practical considerations directly, and therefore it may be difficult to implement a bankruptcy prediction model based on contemporary literature. By addressing the key aspects of bankruptcy prediction and forming a cohesive picture of the modeling process from a practical perspective, this thesis contributes to bridging the gap between business practitioners and academia. A second notable contribution is related to the issues of small sample size and focus on certain geographic markets in the extant literature, as well as the lack of SME-specific research: this thesis addresses the aforementioned concerns by employing a sample of over 125 000 Finnish SMEs.

1.2 Objectives and research questions

The primary aim of this thesis is to investigate potential means of improving the bankruptcy prediction model used by Valuatum Oy (Valuatum), a Finnish provider of financial analysis software. On a more general level, the goal is to identify the key aspects in the design and implementation of bankruptcy prediction models, and to apply them in practice to an empirical study in order to establish an academically and practically viable model. The prediction model should not only perform well, but also be interpretable and easy to implement in practice. The first research question relates to the choice of prediction methods:

Q1: Which bankruptcy prediction techniques provide the best balance of performance and usability in the context of Finnish SMEs?

A second key area of interest are the data and predictor variables used. This thesis attempts to find the most relevant accounting-based predictor variables for Finnish SMEs, and additionally explores how the variables should be chosen and whether their number should be limited. The second research question is formulated as follows:

Q2: Which accounting-based predictor variables are the most important for Finnish SMEs, and how should the variable set be composed?

1.3 Research design and scope

The structure of this thesis is twofold. In the first, theoretical part, a literature review is conducted to gain insights into the state of the art in bankruptcy prediction and to explore the design and characteristics of bankruptcy prediction models. This is followed by an empirical study comparing the performance of different prediction models. This thesis is exploratory by nature and does not attempt to explain the causal mechanisms of bankruptcy; no hypotheses are therefore formed.

The literature review provides a brief overview of corporate failure prediction as a field of academic research and describes the process and key aspects of constructing a failure prediction model. The use of accounting-based predictor variables is discussed in depth to form a solid basis for the choice of variables in the empirical study. Different prediction methods and data preprocessing techniques applied in previous studies are explored in order to align the design of the empirical study with the objectives of this thesis.

The empirical part of this thesis consists of a quantitative study comparing the performance of bankruptcy prediction models, which are designed based on the findings of the literature review. Different combinations of prediction methods, data preprocessing techniques and predictor variables are analyzed to find the best-performing alternatives.

The scope of this thesis in terms of the firm population of interest is limited to

Finnish SMEs. Existing bankruptcy prediction research on Finnish data is scarce, and additional findings can be valuable. SMEs are targeted because they are vital to the economy, but underrepresented in the extant literature; new evidence is certainly welcomed. Furthermore, access to an extensive financial statement database provides a unique opportunity for studying SME bankruptcies, as data unavailability is a common hurdle to conducting new studies; the opportunity must be capitalized on.

Due to considerations related to the practical applicability of the results of this thesis, input variables for the empirical study are limited to quantitative variables calculated using financial statement data. The time scope of the data used for predictor variables is limited to 2008–2010, and the period for which bankruptcies are observed is 2011–2012. Because of practical limitations, no attempts are made to generalize the findings of the empirical study to other populations or time periods.

1.4 Structure of the thesis

The remainder of this thesis is structured as follows. Section 2 reviews academic literature on bankruptcy prediction and establishes the theoretical background for the empirical study. The data and variables used in the empirical study are described in Section 3, and methodological choices are detailed in Section 4. Empirical results are presented in Section 5, followed by a discussion of the results, recommendations for Valuatum, and proposals for further research in Section 6.

2 Literature review

2.1 Overview of corporate failure prediction

Corporate bankruptcies and other types of failure, and particularly predicting their occurrence, have long been a popular research topic; one of the first known works was published nearly a century ago by FitzPatrick (1932). The first major steps in the field were taken in the 1960s, with seminal studies by Beaver (1966) and Altman (1968) describing statistical methods for predicting bankruptcies based on financial ratios; most papers in the 60s and 70s built on these methods. New approaches were later developed, Ohlson's (1980) proposal of using logistic regression becoming particularly influential. The statistical methods used by the aforementioned authors remained the most popular choices in failure prediction until well into the 21st century (Balcaen & Ooghe, 2006; Dimitras et al., 1996; Veganzones & Séverin, 2020). Progress in information technology and data processing capabilities in the 1990s and 2000s provided new opportunities and led to the introduction of machine learning methods to failure prediction. Alternatives to accounting-based models also emerged, most notably market-based models using option pricing theory.

Interest towards bankruptcy prediction has increased in the 21st century. The changes established to capital requirements by the Basel II framework (BCBS, 2004) encouraged the development of improved risk models (Agarwal & Taffler, 2008; Angelini et al., 2008; Kirkos, 2015). The 2007–08 financial crisis is mentioned by many authors (Gupta & Chaudhry, 2019; Huang & Yen, 2019; Succurro et al., 2019) as a wake up call that alerted both academics and practitioners to the potentially catastrophic consequences of bankruptcy. It also revealed shortcomings in bank regulation and led to the revision of the Basel framework, Basel III (BCBS, 2011), which further promotes risk management and the development of credit risk models.

In addition to developing novel techniques and applying them in different ways, failure prediction research has branched into various other directions in search for more accurate models. While company financial statements and market-based data remain the most commonly used sources of input variables, various other factors such as macroeconomic variables and corporate governance indicators have been introduced.

A separate branch of bankruptcy research focuses on identifying and explaining the root causes of bankruptcy from a theoretical perspective (see Amankwah-Amoah, 2016, for a summary of relevant research). Corporate failure prediction and the causal theory of bankruptcy remain mostly disjointed. Some studies concentrate on the process of firm failure and use it for prediction purposes (du Jardin, 2015, 2017; Laitinen, 1993), but the failure process is mainly presented in terms of financial characteristics, and the root causes are not examined. Laitinen & Lukason (2014) link causes of failure to financial indicators, but do not establish a prediction model; Ooghe & De Prijcker (2008) present a typology of failure processes and the corresponding developments in financial ratios, but provide no quantitative evidence. In general, the connection between bankruptcy theory and empirical failure prediction research seems tenuous at best.

The definition and boundaries of what corporate failure prediction is are somewhat indistinct. While some studies focus strictly on predicting bankruptcies, others use the same methodologies for assessing the default risk of corporate loans. The related subjects of credit risk and credit scoring further add to the confusion, as they are used in both corporate and consumer contexts. The explicit meaning of ’corporate failure’ is also challenging due to varying legal definitions, as well as the use of non-legally based ones; this subject is discussed further in the following section. The issue of nomenclature and defining the field of study are seldom addressed in the literature, and many authors discuss corporate and consumer credit risk jointly without making a clear distinction between the two (see e.g. Li et al., 2016; Nanni & Lumini, 2009); however, due to the obvious similarities and empirical evidence of the same methodologies performing well in both contexts, this approach may be considered justified. In addition, due to its lack of connection to empirical research, bankruptcy theory provides no compelling argument in favor of making sharp distinctions between corporate and consumer credit risk in empirical studies.

2.2 Definition of corporate failure

Literature on corporate failure prediction has approached the subject from various perspectives, and no universally accepted definition exists for what constitutes corporate failure (Veganzones & Séverin, 2020). Juridical definitions such as bankruptcy are commonly used as a measure of failure, while some studies employ varying interpretations of "financial distress" or other similar concepts (Balcaen & Ooghe, 2006). Due to the lack of an explicit definition, corporate failure prediction largely overlaps with default prediction literature, where the main focus is on forecasting loan defaults rather than only outright business failures.

Many authors use the different terms almost interchangeably and directly draw on bankruptcy prediction literature for the purposes of credit risk modeling (e.g. Petropoulos et al., 2016; Sigrist & Hirnschall, 2019); as Platt & Platt (2002) note, most prediction models rely on bankruptcy data, even if the object of prediction is defined as something else. Additionally, consumer credit risk is in some studies addressed in conjunction with corporate failure and default modeling (Lin et al., 2012b; Nanni & Lumini, 2009), although the main focus in these studies is on the technical aspects of prediction models. Regardless of the specific definition used, failure is predominantly modelled as a binary variable that divides firms into failed and non-failed populations (Veganzones & Séverin, 2020; Yu et al., 2014).

In the modern literature, bankruptcy is the prevalent measure of corporate failure. Compared to other measures, it offers an unambiguous indication of failure and provides a clear dichotomy between failed and non-failed companies (Veganzones & Séverin, 2020). Due to its nature as an explicitly defined legal process, the occurrence of bankruptcy can be determined precisely and is an objective measure for the time of failure (Balcaen & Ooghe, 2006). Additionally, bankruptcy data is usually easily obtainable (Platt & Platt, 2002).

There are also particular problems related to bankruptcy as a measure of failure. Hill et al. (1996) note the some bankruptcies occur suddenly, with no prior signs of financial trouble; research on the causes of bankruptcy supports this notion by showing that unexpected environmental jolts such as natural disasters may be a contributing factor (Amankwah-Amoah, 2016). Such outliers to the typical failure process complicate modeling and can deteriorate predictive performance. Moreover, classifying firms by bankruptcy disregards other types of legal proceedings through which a failed firm can be terminated (Balcaen & Ooghe, 2006). The legal process of bankruptcy can be lengthy, and companies may encounter severe, irremediable financial problems long before bankruptcy is declared (Pompe & Bilderbeek, 2000; Theodossiou, 1993). Bankruptcy proceedings are specific to the prevailing legal framework, and therefore the resulting models may not be generalizable across countries. Finally, it should be noted that bankruptcy does not necessarily mean failure: Gupta & Chaudhry (2019) argue that filing for bankruptcy can, in some situations, be an attempt to gain strategic advantage.

As opposed to bankruptcy, definitions of financial distress vary greatly between studies and application contexts, from temporary cash flow problems to liquidation (Sun et al., 2014). Some studies make use of indicators that combine several distress factors such as loan defaults, filed bankruptcies and other types of insolvency proceedings (Filipe et al., 2016; Li et al., 2016). Financial ratios are also used, typically with predefined rules or thresholds that establish the failed/non-failed status of the studied companies (Lin et al., 2012a; Molina & Preve, 2012). Defining failure through financial distress rather than bankruptcy can thus allow constructing a model that is better suited to its context and requirements. Lin et al. (2012a) also suggest that extending the definition of failure to cover more than officially declared bankruptcies can help avoid data scarcity issues related to the small number of bankruptcies in real-world data. However, Veganzones & Séverin (2020) assert that using arbitrary subjective definitions easily leads to biased results, and that bankruptcy is therefore the preferred definition for corporate failure prediction models.

2.3 Accounting-based predictor variables

2.3.1 Main advantages and drawbacks

Throughout the history of corporate failure prediction, financial statement data have been the most common source of explanatory variables (Acosta-González et al., 2019; Succurro et al., 2019). Many of the pioneering works such as Beaver (1966), Altman (1968) and Ohlson (1980) rely entirely on financial ratios, and in modern research they are still the most widely used category of predictors (Veganzones &

Séverin, 2020). Even when alternative explanatory variables are used, they are typically applied in conjunction with accounting-based variables (Calabrese et al., 2019; Ciampi, 2015; Tobback et al., 2017).

Accounting-based variables are popular in failure prediction, because they offer a relatively objective, quantitative measure of a company's performance and financial status (Balcaen & Ooghe, 2006). The concept of bankruptcy is based on the inability to pay outstanding debts, and thus directly connected to financial statement variables (du Jardin, 2015), whereas links between bankruptcy and alternative predictors such as corporate governance indicators are more ambiguous and difficult to identify. Additionally, accounting-based ratios typically serve as the basis of debt covenant conditions, and can therefore contain information that reflects a company's credit risk (Agarwal & Taffler, 2008; Hillegeist et al., 2004).

Accessibility and reliability are also important arguments in favor of accounting-based variables: companies' financial statements are usually the most readily available type of information, especially for smaller, unlisted companies (Zoričák et al., 2020). Global standardization efforts further promote the availability, transparency and reliability of accounting data. Adoption of the eXtensible Business Reporting Language (XBRL) standard for digital financial reporting has been found to reduce account manipulation through earnings management practices (Kim et al., 2019) and decrease information asymmetry in the stock market (Yoon et al., 2011); additionally, financial statement data from several countries have been made publicly available in the XBRL format (XBRL International, 2020). The International Financial Reporting Standards (IFRS) promote the comparability of financial statements internationally, thus aiding the development of more widely generalizable failure prediction models.

The predictive capacity of accounting-based explanatory variables is indisputable and has been demonstrated in a multitude of studies throughout the history of corporate failure prediction. Many of the financial ratios proposed as determinants of failure by early scholars are still used today, and their predictive power remains significant; this robustness is also evidenced by Beaver et al. (2005), who find that an accounting-based model retains its predictive ability with only a minor performance decline over the period 1962–2002.

Despite its many benefits, the use of accounting data in corporate failure prediction is not without its complications. Zavgren (1985) and Zmijewski (1984) argue that financial information alone is not sufficient for predicting failure; Hillegeist et al. (2004) point out that financial statements are made on a going-concern basis and are therefore inadequate for failure prediction by design, and demonstrate that a model using market information is superior to accounting-based models. The notion of accounting data being insufficient is supported by empirical evidence: for example, Jones (2017) finds that in a model combining several different categories of predictors, the relative importance of financial ratios is lower than that of various others, such as ownership structure and market-based information. Lukason & Laitinen (2019) find that the duration of the firm failure process is often quite short, particularly for SMEs: even the latest financial statement may not sufficiently reflect the impending failure, which makes bankruptcy prediction extremely difficult.

Accounting data are often assumed to constitute a reliable and accurate indicator of a company's overall financial status. However, financial statements can be manipulated in diverse ways to distort this perception (du Jardin et al., 2019; Serrano-Cinca et al., 2019). While there are varying motivations for misrepresenting accounting figures, firms in financial distress may be particularly inclined to do so, for example in order to avoid covenant violations (DeFond & Jiambalvo, 1994). On a related note, Lukason & Camacho-Miñano (2019) find that firms encountering financial distress are more likely to delay the publication of their financial statements, which further indicates that the problems companies face may not be observable from their accounting figures. Furthermore, in many jurisdictions, only large companies are obligated to publish their financial statements (Balcaen & Ooghe, 2006). Data for smaller firms can be difficult to obtain, more likely to contain errors and misrepresentation, and their correctness is difficult to verify.

Due to the lack of connection with bankruptcy theory and the causes of failure, most empirical studies simply pick predictor variables based on the results of previous research, largely ignoring the theoretical underpinnings (Ooghe & De Prijcker, 2008). On the other hand, Kirkos (2015) notes that failure prediction studies rarely try to extract meaningful findings from the observed effects of specific predictor variables. The possible internal and external factors leading a company to bankruptcy are numerous (Amankwah-Amoah, 2016), and financial statement variables seem unable to capture them efficiently.

Various alternative predictors have been introduced to counter the shortcomings of financial data. Some commonly used examples include market-based variables, macroeconomic data and firm characteristics such as ownership and management structures. There is ample empirical evidence of the predictive power of non-financial variables (see e.g. Andrikopoulos & Khorasgani, 2018; Jones, 2017; Liang et al., 2016). Recent studies have explored such factors as relationships between management-level personnel (Tobback et al., 2017), spatial dependence (Calabrese et al., 2019), and textual information extracted from annual reports (Lohmann & Ohliger, 2020; Wang et al., 2018); research on alternative predictors is diverse and growing in popularity. However, the subject is outside the scope of this thesis and will therefore not be discussed in detail; the following section focuses solely on accounting-based predictors.

2.3.2 Key predictor categories

Despite significant variation in the specific variables and their combinations, failure prediction studies using financial statement data typically employ predictors measuring similar key aspects of a company's financial situation, such as profitability, solvency and liquidity, capital structure, and activity (Balcaen & Ooghe, 2006; Dimitras et al., 1996; Ravi Kumar & Ravi, 2007). The vast majority of accounting-based predictors are financial ratios; unscaled variables are dependent on company size, whereas ratios can be used to compare different-sized firms. However, Beaver (1968) notes that ratios can mask the effects of individual financial characteristics due to offsetting effects of their components, and Balcaen & Ooghe (2006) further assert that a similar effect can occur with detailed ratios within a more general one. For example, a seemingly normal level of working capital could hide a massive surplus inventory and cash shortage. Therefore, a sufficient level of specificity must be used in defining ratios for failure prediction.

The specific types of variables found to be most important vary between studies. For example, Liang et al. (2016) propose solvency and profitability as the key categories while Lin et al. (2012a) mention profitability, growth and employee efficiency as the best predictors; Gupta et al. (2014) find that ratios based on operating cash flow do not add to the performance of their model, while Jones et al. (2017) list multiple cash flow ratios among the most effective predictors. These types of findings corroborate that, as has long been suggested, the impact of different financial ratios is highly sample-specific (Bauweraerts, 2016; Edmister, 1972). It should also be noted that the categories of predictors are not precisely defined: there is significant overlap between them, and interpretations and terminology vary between studies. In general, no consensus has formed over the decades of research, and knowing which variables to use in a prediction model can be difficult (Yu et al., 2014).

Recent studies have shed some light on the possible reasons behind the conflicting findings. Laitinen & Lukason (2014) find that the explanatory factors of failure vary between Finnish and Estonian companies, and Filipe et al. (2016) similarly note regional differences in a study covering eight European countries. Firm size is also a significant factor: (Gupta et al., 2015) show that the process and predictors of failure vary between medium-sized, small, and micro companies. The numerous non-financial determinants of bankruptcy cannot be captured by accounting data, and therefore complicate the interpretation of the effects of specific financial ratios.

Profitability

The significance of profitability in failure prediction is perhaps self-evident; a company that does not generate positive returns is bound to fail sooner or later. Lukason & Laitinen (2019) demonstrate that capturing the different types of firm failure processes requires that both annual and cumulative profit measures are used, and additionally suggest that changes in profitability should be incorporated in prediction models. One advantage of cumulative profitability indicators such as retained earnings is that they implicitly incorporate the company's age (Altman, 1968), a factor which is known to affect risk of insolvency.

How profitability impacts the failure probability of a company depends on numerous external and internal factors. According to Andreeva et al. (2016), profitability is among the key predictors in most studies concerning SMEs, and Lukason & Laitinen (2018) find that profitability plays a larger role in the failure probability of exporting firms than that of non-exporting firms. Lohmann & Ohliger (2019b) discover that extremely low or high levels of profitability are more likely to indicate risk of failure for young companies than for mature ones. A potential explanation could be that newly established firms are less stable than mature ones and have no accumulated excess funds, and may therefore be more vulnerable to major short-term losses; the link between high profitability and probability of failure could be explained by high-growth startups that intentionally pursue a risky strategy.

Solvency and liquidity
Solvency, i.e. a company's ability to reimburse its outstanding debts, is a commonly used measure in failure prediction. A distinction is commonly made between overall solvency and short-term solvency, which is typically referred to as liquidity. From a conceptual standpoint, bankruptcy and most related definitions of financial distress essentially mean a company's inability to meet its financial obligations, and therefore financial indicators measuring solvency can be considered naturally suitable for failure prediction (du Jardin, 2015). Solvency and liquidity are among the most common predictors used in failure prediction (Serrano-Cinca et al., 2019), and their usefulness is substantiated by empirical evidence (Bauweraerts, 2016; Dimitras et al., 1996; Liang et al., 2016).

Traditional measures of liquidity include the current ratio (current assets to current liabilities) and quick ratio (current assets less inventories to current liabilities). These measures have been criticized for being static in nature; they measure a company's ability to repay its existing debts at a given moment using its existing assets, but do not take the company's ongoing operations into account (Yli-Olli & Virtanen, 1989). Shulman & Cox (1985) note that the traditional static ratios are primarily interesting from a liquidation perspective, and propose an integrative approach that incorporates operating liquidity to provide more relevant information from a going concern viewpoint; they argue that increases in inventory or receivables (both of which improve current and quick ratios) can signal a shortage of cash rather than improved liquidity. In failure prediction, the essential question is whether a company is able to continue its operations, and therefore the going concern perspective can be considered more suitable. Altman (1968) suggests that the ratio of working capital to total assets is a better predictor of bankruptcy than current or quick ratio; this seems to validate Shulman's view that operational requirements should be taken into consideration, but could also be because popular ratios are more likely to be subject to manipulation in an attempt to convey a falsely positive picture of the company's health (Beaver, 1968). Altman et al. (2010) reiterate the importance of working capital in failure prediction, and point out that it is especially effective with SMEs, which typically rely more on trade credit than bank loans.

Despite taking operations into account, the aforementioned liquidity indicators are solely based on balance sheet figures; alternative measures have been suggested to address the operational aspect more directly. The defensive interval (Davidson et al., 1964), which measures quick assets in relation to (projected) operating expenses is used for failure prediction by Beaver (1966), while Laitinen (1993) employs dynamic measures of liquidity and solvency based on cash flow. Contemporary studies seldom address the issue explicitly, but most nonetheless include dynamic liquidity and solvency indicators alongside static ones; it seems that their importance in failure prediction is widely accepted.

Capital structure
The capital structure of companies is a major area of interest in the field of corporate finance and has given rise to an extensive body of research. One of the key subjects of interest is the relative proportion of debt financing (leverage). The seminal work of Modigliani & Miller (1958, 1963) indicates that using debt financing instead of equity provides an advantageous tax shield, thereby implying that debt must entail adverse effects that offset the benefits; the (expected) costs of financial distress and bankruptcy are cited as one such factor. To assess this notion, Molina (2005) studies the effects of leverage and finds that higher leverage has a significant negative impact on credit ratings, indicating a higher risk of failure.

While the impact of leverage on failure risk is substantial, the structure of a firm's assets may also have some bearing on its probability of bankruptcy. Tangible assets can be used as collateral for financing, and are informationally transparent; therefore, they impact financially distressed companies' ability to obtain additional funds and thus increase their chance of survival (Keasey et al., 2015). On the other hand, Jones (2011) finds that higher capitalization of intangible assets is connected to a higher risk of bankruptcy.

Ample empirical evidence can be found in bankruptcy prediction literature to confirm the predictive ability of capital structure ratios (Beaver et al., 2005; Charalambakis & Garrett, 2019; Jones et al., 2017; Lukason & Laitinen, 2018; Succurro et al., 2019). Firms typically have target leverage ratios towards which they adjust (Mari & Marra, 2019); Löffler & Maurer (2011) use this phenomenon to forecast companies' future leverage ratios and find that these forecasts are useful bankruptcy predictors.

Activity
Activity ratios measure how efficiently a company utilizes its assets, thus providing a view of the soundness of its operations. Activity has been recognized as a key category in ratio analysis for a long time, and included in the early failure prediction models (Altman, 1968; Beaver, 1966). Typical examples of activity metrics include turnover ratios (De Bock, 2017; Liang et al., 2016; Wang & Ma, 2011) and measures of employee efficiency (Lin et al., 2012a; Volkov et al., 2017). Empirically, the usefulness of activity ratios as bankruptcy predictors has been demonstrated on many occasions (e.g. Charitou et al., 2004; Liang et al., 2016; Zoričák et al., 2020); Pompe & Bilderbeek (2005) show that activity ratios can have predictive power as long as five years before company failure.

Company characteristics
While not actually financial ratios, various background data about company characteristics are typically available together with financial statements, and are commonly used in accounting-based failure prediction models. Company age is one factor that has been found significant: in general, young firms tend to have a higher risk of

failure than older ones (Altman et al., 2017; Bauweraerts, 2016). However, Henderson (1999) contrarily suggests that older firms may in some cases be more likely to fail due to becoming obsolete. Additionally, Lohmann & Ohliger (2019b) find that the relationships between accounting-based predictor variables and probability of failure are different for young and mature firms.

Company size has also been shown to affect corporate failure risk (El Kalak & Hudson, 2016; Gordini, 2014; Ohlson, 1980), and is often used as a predictor in bankruptcy prediction. The most common proxies used for company size are total sales and total assets; both are directly available in the financial statement and therefore easy to incorporate into a prediction model. The specific characteristics of a firm's industry also have a major impact on the probability and determinants of failure (Charitou et al., 2004; Hu & Ansell, 2007; Lohmann & Ohliger, 2019a; Platt & Platt, 1990). Some studies therefore limit the examined firms to a specific industry, for example manufacturing (Altman, 1968; Kuběnka & Myšková, 2019; Shin & Lee, 2002), or exclude some, such as financial companies (Agarwal & Taffler, 2008; Jackson & Wood, 2013; Liang et al., 2016).

Growth/change variables
The nature of business failure as a process occurring over time, instead of a sudden event, has been understood for a long time. However, according to Argenti (1976) the duration of the bankruptcy process is commonly underestimated. Dimitras et al. (1996) discover in their literature review that most studies from the 1960s to 1990s disregard the time dimension of bankruptcy, although some exceptions (e.g. Laitinen, 1991, 1993; Theodossiou, 1993) can be found. In modern literature, the temporal aspect of failure has been discussed somewhat more frequently, with focus on both predictor variables and modeling techniques.

A common approach to incorporating the time dimension through predictor variables is to use data from multiple periods preceding the failure, either by including variables measuring the change in various financial characteristics, or by creating a multi-period model (du Jardin, 2015). The effectiveness of annual growth variables as failure predictors has been demonstrated empirically (Jones et al., 2017; Lin et al., 2012a). Nyitrai (2019) proposes an indicator variable that assesses whether a financial ratio's value is smaller than, greater than, or within the range of its values in previous years, and finds that it has significant predictive power.

2.3.3 Predictor variable selection

In corporate failure prediction, the choice of predictor variables is one of the key choices that affect the performance of the model. The typical approach in the literature has been to begin with a large initial set of variables assembled more or less arbitrarily based on prior research, from which the predictors are chosen based on different statistical or empirical criteria (Balcaen & Ooghe, 2006). This practice is still common in contemporary research, but many authors also disregard the latter step and simply use a selection of predictors picked from earlier literature (Barboza et al., 2017; Charalambakis & Garrett, 2019) or apply arbitrary selection criteria (Lohmann & Ohliger, 2019b; Zoričák et al., 2020). Some studies use pre-existing data sets and are simply constrained to whichever features the data happen to contain (Le et al., 2019; Tsai & Cheng, 2012). Regardless of the chosen approach, the final set of predictors in most studies contains at least variables corresponding to the previously discussed key categories of profitability, solvency/liquidity, capital structure, and activity.

Apart from variable selection based on prior research alone, the two main approaches are filter and wrapper methods; additionally, some prediction methods have built-in mechanisms to perform feature selection in the process of training the model (Li & Sun, 2011). Filter methods evaluate the general characteristics of the data to find the best subset of variables, typically using statistical measures such as mutual information or the χ^2 statistic (Zhang et al., 2019). Wrapper methods train a specific model using different subsets of features and compare predictive performance to find the optimal set of predictors (Kohavi & John, 1997). Filter methods are efficient and scalable, and can help avoid overfitting; moreover, a generic selection of "best" variables can be more useful than one optimized for a particular algorithm (Guyon & Elisseeff, 2003). Wrapper methods typically produce high performance, but the selection of best features is not generalizable to other classification methods (Peng et al., 2005). They are also computationally very demanding, especially if the initial feature set is large, and can easily cause overfitting (Lin et al., 2014). The performance of different methods varies; du Jardin (2017) argues that filters are useful with statistical methods, but machine learning techniques work better with wrapper methods.

An alternative to selecting a subset of features is feature extraction, in which the predictor variables are mapped to a lower-dimensional space, creating new features as linear or nonlinear combinations of the original ones (Bennasar et al., 2015). A variety of both linear and nonlinear feature extraction techniques have been successfully applied in bankruptcy prediction (Verikas et al., 2010), but they are generally used less often than feature selection (Alaka et al., 2018). One potential explanation for this is interpretability: the effects of the original predictor variables are difficult to decipher from the extracted features (Zoričák et al., 2020), whereas feature selection keeps the original financial ratios intact.

As with many other aspects of corporate failure prediction, the findings regarding different feature selection methods are inconclusive. In two studies that both use neural networks, Tsai (2009) compares different filter and feature extraction methods and finds that filtering by t-test gives the best results, while du Jardin (2010) shows that wrapper selection outperforms filter methods; the superiority of wrapper methods is supported by the findings of Lin & Lu (2019). Lin et al. (2014) design a new wrapper algorithm and find that its selected subset of features yields better performance than literature-based and filter selection alternatives. Wrapper methods seem to perform well in most contexts, but are not categorically superior to filter methods: for example, Liang et al. (2015) compare two wrapper and three filter methods with six different classifier algorithms and find that filter methods perform better with some classifiers and datasets, while losing out to wrapper methods with others.

As the impact of individual variables to failure prediction models is likely to be sample-specific (Edmister, 1972; Zavgren, 1985), the exclusively literature-based approach to feature selection can be considered somewhat problematic. Selecting a large, diversified set of predictors could help ensure that the most important variables are included. On the other hand, Veganzones & Séverin (2020) note that this results in the inclusion of irrelevant and redundant variables, which may decrease model performance, and conclude that multiple variable selection techniques should be applied to discover the most essential predictors. An alternative view is presented by Jones (2017), who shows that bankruptcy can be predicted accurately in a high-dimensional setting, i.e. using a large number of predictors, and that feature selection is not necessary.

2.4 Sample selection and data preprocessing

In the context of corporate failure prediction, real-world populations of firms typically consist overwhelmingly of non-failed companies, with failed companies representing only a small fraction of the total sample (Fan et al., 2018; Sueyoshi & Goto, 2009). Classification methods are typically designed to maximize the overall prediction accuracy (number of correct predictions divided by total number of predicted instances) without regard to class distribution; if the number of observations in one class is considerably larger than in the other, the model will focus on accurately predicting the majority class while disregarding the minority class (Kim et al., 2015; López et al., 2013). The presence of class imbalance can therefore impair the classifier's ability to correctly detect failed companies (Sun et al., 2020); Veganzones & Séverin (2018) find that an imbalance greater than 1:4 significantly weakens failure prediction performance. The small number of bankruptcies in the data is a major challenge in predicting business failure, since it leads to frequent misclassification of failing companies as healthy, which is much more costly than falsely predicting that a healthy company will become insolvent (Modina & Pietrovito, 2014).

Class imbalance issues are typically tackled either by using a classifier algorithm designed to take the imbalance into account, or by preprocessing the data to adjust the class distribution prior to training the classifier (López et al., 2013). Both of these approaches are frequently seen in failure prediction literature. Chen et al. (2011), Kim et al. (2015), and Xiao et al. (2012) use algorithmic modifications to mitigate the effects of imbalance; Faris et al. (2020) and Zhou (2013) employ sampling techniques to balance the data prior to training, and Sun et al. (2018) find that a learning ensemble using varied sampling rates is efficient in combatting the imbalance problem. Due to the rarity of bankruptcies, removing non-bankrupt firm observations (undersampling) can deplete the data severely, and therefore oversampling methods such as synthetic minority oversampling technique (SMOTE)

(Chawla et al., 2002), which create additional instances of the minority class to balance the data, are more popular. In some studies, the small number of bankruptcies has been utilized by framing the task as anomaly detection instead of the typical binary classification (Fan et al., 2018; Zoričák et al., 2020). These studies use unsupervised learning, in which the algorithm searches for outliers based on feature values only, without knowing the class (failed or non-failed) of the observations.

Despite the naturally occurring class imbalance in real-life bankruptcy data, most studies use a sample that has an equal number of failed and non-failed firms at the outset (Alaka et al., 2018; Gruszczyński, 2019). While this approach may seemingly alleviate the class imbalance problem, Zmijewski (1984) shows that it produces distorted prediction results: the model's ability to correctly classify bankrupt firms is overestimated. When the test sample distribution of failed and non-failed firms changes from 1:1 to a more realistic level (1:20 in Zmijewski's paper), the misclassification rate of failed firms grows drastically. Du Jardin (2015) observes that balancing the data leads to better detection of bankrupt firms and may not be an issue in classification tasks, but also notes that it produces unreliable probability estimates. Veganzones & Séverin (2020) conclude that balanced samples are preferable due to yielding better performance. Regardless of the balance or imbalance in the training data, Zmijewski's (1984) findings show that reporting the results truthfully requires a test sample that reflects the proportions of failed and non-failed firms in real-world populations.

Depending on the variables used and the companies studied, the range of values in the data can vary widely. Some prediction methods, for example neural networks, require that the feature values are in the same range (Angelini et al., 2008). A typical approach to this is min-max scaling, i.e. applying a monotonic transformation that limits the values to a given range, $[0, 1]$ being the most common choice (Bao et al., 2019; Liang et al., 2018).

The distributions of many financial variables used in bankruptcy prediction also tend to be highly skewed, i.e. not normally distributed (Jones et al., 2015; Laitinen & Lukason, 2014; Nyitrai & Virág, 2019). Most statistical prediction methods assume normality (Jones et al., 2017), but many machine learning methods are also known to achieve better accuracy if predictors are normally distributed (Son et al., 2019). This issue can be tackled by transforming the data; the Box-Cox family of power transformations (Box & Cox, 1964) are a popular choice. Jones et al. (2015, 2017) demonstrate that variable transformation improves the predictive performance of many failure prediction models.

Financial data typically contain outlier values that significantly deviate from the rest of the observations. Outlier values can considerably influence prediction models (Hu & Ansell, 2007; Shumway, 2001), and therefore most studies aim to eliminate their effects. A common technique for processing outliers is winsorization (Boritz & Kennedy, 1995; Lukason & Laitinen, 2018; Serrano-Cinca et al., 2019), which sets all outlier values to a limit value, which is usually defined as a specific percentile of the data. Various alternative methods exist, some of which are more or less arbitrary and subjective (see e.g. Molina, 2005). To take the high dimensionality of financial data into account, techniques such as the local outlier factor (Breuniq et al., 2000) have been used (Figini et al., 2017).

Most failure prediction studies appear to consider outlier processing a part of the standard modeling procedure, and do not specifically examine its effects on predictive performance. The general assumption is that it improves the model; however, (Yu et al., 2014) find empirically that outlier values of certain predictor variables are mostly observed with bankrupt companies, implying that they carry a notable amount of information and should not be removed or modified. Zoričák et al. (2020) also note that outliers occur naturally in financial data, and therefore a bankruptcy prediction model must be able to handle them similarly to non-outliers. It does seem reasonable to assume that a connection exists between outliers and firm failure: such characteristics as abnormally low profitability can certainly be seen as indicators of poor performance and failure risk. Still, outlier processing does not necessarily impede the ability to detect failing firms: for example, Pawełek (2019) winsorizes outliers from the non-bankrupt class only and finds that it improves predictive performance.

Missing predictor values are a common issue in all fields of business research, and especially prevalent in corporate failure prediction (Jones et al., 2017). The information in financial statements is often incomplete; this is particularly true for SMEs, since they are usually subject to less strict accounting-related regulations than larger companies (Andreeva et al., 2016; Ciampi & Gordini, 2013; Zhou & Lai, 2017). Typically, methods for handling missing data require that whether a datum is missing does not depend on its (unobserved) value: the data should be missing either independently of any observed or missing data (missing completely at random, MCAR) or potentially depending on some observed values but not on the missing values themselves (missing at random, MAR) (Hastie et al., 2009). However, in failure prediction this is often not the case; for example, companies that perform poorly and have a high risk of bankruptcy are more likely to delay the reporting of annual accounts than healthy companies (Lukason & Camacho-Miñano, 2019).

The most common approach to missing values in failure prediction and other fields of research is casewise deletion (Jones et al., 2017), which consists of simply discarding observations with missing feature values. A major issue with casewise deletion, especially in failure prediction, is that it may lead to significant depletion of the training data and introduce bias if data are not MCAR, which is seldom the case with accounting data (Acosta-González & Fernández-Rodríguez, 2014). An alternative to casewise deletion is missing value imputation, which replaces the missing values according to some predefined rule. Techniques range from simply using the mean or median of the feature, to sophisticated algorithmic approaches (Xia et al., 2017). Despite the prevalence and implications of missing data in failure prediction, most studies do not address the issue at all or only note it in passing. For example, García et al. (2019, p.90) merely mention that "an imputation method integrated into the ensemble model was used to handle missing data" and do not discuss the subject otherwise.

The effectiveness of different methods of handling missing data in failure prediction remains somewhat unclear. Florez-Lopez (2010) finds that casewise deletion and simple substitution perform worse than more sophisticated techniques, and Zhou & Lai (2017) show that imputation improves performance compared to non-imputed data. On the other hand, Jones et al. (2017) indicate that imputation using singular value decomposition does not improve the performance of most prediction methods when compared to casewise deletion. It can be reasonably assumed that the amount and patterns of missing data vary considerably between data sets due to a variety of reasons, from legal and accounting factors to different data handling and storage technologies used by data providers. As with many other aspects of failure prediction, there are no definitive answers as to how missing data should be addressed. Observations with missing values are unreliable, and may distort prediction results; on the other hand, removing or imputing the missing values can have the same effect, because the data are often not missing at random.

2.5 Prediction methods

The prediction method used plays a key role in the design of a failure prediction model and has a major impact on its performance. The context and objectives of a study are central to the choice of modeling approach; for example, assessing the effects of specific predictor variables may require tools that are not optimal when the aim is simply to achieve maximal predictive performance. The commonly used methods can be roughly categorized into traditional statistical methods and machine learning or artificial intelligence-based methods.

2.5.1 Statistical methods

Failure prediction literature up until the 1990s was dominated by statistical methods, the foremost of which in early literature was multiple discriminant analysis (MDA). Although a quadratic variant is applied by some authors, most studies use a linear MDA model. Terminology in the literature varies; some papers refer to these as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). The basic principle of MDA is to find the linear (or non-linear in the case of quadratic MDA) combination of a given set of variables that best discriminates between distinct, mutually exclusive groups (failed and non-failed firms) (Altman, 1968; Deakin, 1972). MDA first rose to prominence with Altman's (1968) Z-score model, which engendered a large number of further studies using a similar approach (Balcaen & Ooghe, 2006). MDA is subject to restrictive assumptions regarding the characteristics of the data used, namely multivariate normality of predictor variables, equal covariance matrices across the bankrupt and non-bankrupt firms, and specified misclassification costs and prior probabilities (Karels & Prakash, 1987); however, many studies fail to ascertain that the assumptions are not violated, and as a result the MDA method is often applied incorrectly and produces non-generalizable results (Joy & Tollefson, 1975; Zavgren, 1985).

The use of MDA in corporate failure prediction decreased in the 1980s (Dimitras et al., 1996), in large part due to the increasing popularity of the logistic regression (logit) method. Logit was first applied to failure prediction by Ohlson (1980), who argued that it avoids both the restrictive data assumptions and the arbitrariness of the procedure of matching failed and non-failed firms that is typical to studies using MDA. Despite these advantages, logit also suffers from certain limitations, such as sensitivity to the input variables' multicollinearity and extreme non-normality (Chou et al., 2017; Nyitrai & Virág, 2019).

Various modifications and improvements to basic logistic regression have been proposed in failure prediction literature, such as multinomial logit (Johnsen & Melicher, 1994), mixed logit (Jones & Hensher, 2004) and quadratic interval logit (Tseng & Lin, 2005). The logit method remains prominent in contemporary literature and is used in various contexts. Lukason & Laitinen (2018) assess the importance of different financial predictors for exporting and non-exporting firms using logit models. Li et al. (2016) propose a hybrid approach using a combination of logistic regression and machine learning. A stepwise logit model has also been applied for selecting the most useful predictor variables (Bauweraerts, 2016; Modina & Pietrovito, 2014).

In addition to reliance on assumptions about the data, statistical methods have a major drawback in that they typically assume a specific relationship between predictors and dependent variable (firm failure), and are therefore unable to model the complex nonlinear dependencies occurring in the data (Balcaen & Ooghe, 2006). Although evidence is not completely unanimous (Laitinen & Kankaanpää, 1999; Pompe & Bilderbeek, 2000), statistical methods are, due to the aforementioned deficiencies, typically outperformed by the alternatives presented in modern studies (Veganzones & Séverin, 2020).

Despite the emergence of novel prediction techniques in the last decades, traditional statistical methods such as logit and MDA maintain a strong position in academic research to this day (Giriūniene et al., 2019; Jones et al., 2015). They also remain widely used by practitioners: for example, Zhang & Thomas (2015) claim that as much as 95% of credit scorecards used by banks are based on logistic regression. Bemš et al. (2015) suggest that the popularity of the logit model among practitioners is to some extent due to being recommended in the Basel II framework (BCBS, 2004). Moreover, logit and other statistical methods are generally easy to interpret and can help in understanding which factors have the greatest impact on bankruptcy risk (Jones et al., 2015; Veganzones & Séverin, 2020). Furthermore, factors such as low computational cost (Alaka et al., 2018) and extensive prior literature make statistical methods an attractive choice.

2.5.2 Machine learning methods

Machine learning (ML) techniques first appeared as an alternative to statistical methods in corporate failure prediction in the 1990s. Most of the first studies (Boritz & Kennedy, 1995; Fletcher & Goss, 1993; Odom & Sharda, 1990; Wilson & Sharda, 1994) applied neural networks (NN), which are based on interconnected nodes that mimic the functioning of the human brain. In their extensive review of failure prediction studies, Ravi Kumar & Ravi (2007) find that NNs generally perform well and achieve better results than statistical methods (for some examples see Angelini et al., 2008; Ciampi & Gordini, 2013; López Iturriaga & Sanz, 2015). Another popular machine learning technique in failure prediction are support vector machines (SVM) (Min & Lee, 2005; Shin et al., 2005; Van Gestel et al., 2003). SVMs function by mapping inputs to a high-dimensional feature space, in which an optimal hyperplane is constructed to divide the sample to two distinct classes (Cortes & Vapnik, 1995).

Although based on different principles and functionality, NNs and SVMs share many characteristics. They are able to model complex nonlinear relationships, and do not rely on restrictive assumptions regarding the input data (Ravi Kumar & Ravi, 2007). NNs (Angelini et al., 2008; Ciampi & Gordini, 2013) and SVMs (Huang et al., 2004; Kim & Sohn, 2010) are reported to achieve high predictive performance, but both are prone to overfitting, i.e. they adapt too closely to the training data, thus reducing out-of-sample performance (Alaka et al., 2018; Jackson & Wood, 2013). They are also somewhat challenging to implement, as the selection of appropriate hyperparameters is difficult (Ravi Kumar & Ravi, 2007). Additionally, both NNs (Figini et al., 2017; López Iturriaga & Sanz, 2015; Sun et al., 2011) and SVMs (Verikas et al., 2010; Yao & Chen, 2019) are criticized for being "black box" methods: the models' internal structure and the impact of different input variables is difficult to decipher. This may limit the usefulness of NNs and SVMs in a corporate failure prediction context, because interpretability is important to business practitioners (Jones et al., 2017; Nyitrai, 2019). Despite their shortcomings, NNs and SVMs remain the most popular individual ML classifiers in bankruptcy prediction (Alaka et al., 2018), but a wide variety of alternative techniques have also been used.

Genetic algorithms imitate Darwinian evolutionary principles for complex, nonlinear problem solving (Ravi Kumar & Ravi, 2007). In failure prediction they have been applied both to feature selection (Back et al., 1996) and to the actual classification task (Gordini, 2014). Zhou et al. (2014) point out that the value of genetic algorithms to business practitioners may be limited due to the stochastic nature of the method; running the model twice on the same data may produce different results. Case-based reasoning significantly differs from typical machine learning methods: instead of modeling generalized relationships, it compares each observation to previously known data and finds the closest match (Shin & Lee, 2002). Examples of case-based reasoning in failure prediction include Bryant (1997) and Sartori et al. (2016). Numerous other techniques have been applied, for example rough sets theory (McKee, 2000, 2003), data envelopment analysis (Horváthová & Mokrišová, 2018; Sueyoshi & Goto, 2009) and Kohonen maps (du Jardin & Séverin, 2011).

Decision trees (DT), which are constructed by recursively partitioning data using some predefined splitting criteria, are a relatively common machine learning method in failure prediction (Alaka et al., 2018), and are increasing in popularity due to certain advantageous characteristics such as interpretability, capacity to handle mixed types of data, and ability to model nonlinear relationships (Delen et al., 2013; Hastie et al., 2009; Serrano-Cinca et al., 2019). The main weakness of DTs is predictive performance: although conflicting findings exist (Olson et al., 2012), they are usually reported as being inferior to NN and SVM (Chen, 2012; Ravi Kumar & Ravi, 2007). Nonetheless, decision trees have been prominent in recent literature due to their widespread use in ensemble learning.

2.5.3 Ensemble learning

The use of ensemble methods, which combine the predictions of several individual classifiers ("base learners") to improve model performance, has greatly increased in recent bankruptcy prediction literature (Veganzones & Séverin, 2020). The basic principle of ensemble learning is that the base learners misclassify different observations and thus compensate for each other's errors, which almost invariably leads to increased performance compared to the individual classifiers (Lin et al., 2012b). The two main approaches to building an ensemble are averaging and boosting methods.

Classification ensembles can, depending on the chosen method, use various kinds of base learners. Decision trees are considered one of the most suitable options due to their instability, which promotes diversity in the ensemble, thus supporting the underlying principle of combining weak learners to achieve high predictive performance (Breiman, 1996); they are also the most widely used base learner in failure prediction ensembles. Alternatives such as NN and SVM base learners have also been used in failure prediction and credit risk literature (Nanni & Lumini, 2009; Sun et al., 2017; Tsai et al., 2014), but to a smaller extent than decision trees.

Averaging or committee methods are based on training multiple base learners in parallel and combining their predictions using some suitable method, for example majority voting for classification and mean prediction for regression (Hastie et al., 2009). One of the most common averaging ensemble methods is bagging (Breiman, 1996), in which the base learners are trained using a bootstrap sample of the training data, that is, a subset of the sample drawn randomly with replacement. The random subspace method (Ho, 1998) uses bootstrap sampling similarly to bagging, with the difference that a subset of the features is drawn, instead of a subset of firm observations. Another commonly used averaging method is the random forest algorithm (Breiman, 2001), which is essentially decision tree bagging, with random subsampling of features during the process of growing each tree.

As opposed to averaging methods that train classifiers separately from each other, in boosting (Schapire, 1990) the ensemble is created by training classifiers sequentially. After the training of each base learner, incorrectly classified instances are identified, and the next base learner will prioritize correctly classifying those observations that were misclassified in the previous iteration (Hastie et al., 2009). The outputs of the individual classifiers are then combined to form the final prediction. Many variants of the boosting algorithm have been proposed; the most popular ones are AdaBoost (Freund & Schapire, 1997) and gradient boosting (Friedman, 2001), which both feature prominently in bankruptcy prediction literature.

The findings regarding the performance of ensemble methods in bankruptcy and credit risk literature vary. In general, ensembles tend to outperform standalone classifiers (Veganzones & Séverin, 2020). On the other hand, du Jardin (2018) remarks that, while an ensemble almost certainly outperforms any of its own base classifiers, it is not guaranteed to perform better than other types of standalone models. However, empirical evidence strongly suggests that ensembles are superior to methods such as NN and SVM (Alfaro et al., 2008; Barboza et al., 2017; Sun et al., 2011; Wang et al., 2014), which are commonly found to be the highest performing standalone classifiers (Alaka et al., 2018).

Many failure prediction studies implement multiple ensemble classifiers with mixed results. Tsai et al. (2014) compare bagging and boosting using NN, SVM and DT base learners, and find that decision tree boosting performs best, and has the additional advantage of lower computational cost compared to the next best options, NN-bagging and SVM-bagging. This supports Schapire's (1990) original notion of combining weak, unstable base learners, but contradicts the suggestion of Abellán & Mantas (2014) that bagging also requires weak classifiers. Barboza et al. (2017) find no notable difference in the performance of bagging, boosting and random forest models, while the results of Jones et al. (2017) indicate the same for generalized boosting, AdaBoost and random forest; in both studies, the distinction between ensembles and standalone models is much larger than that between the specific ensemble methods.

In addition to predictive performance, ensemble methods can be, depending on the choice of base learner, much more interpretable than NNs or SVMs, for example. Jones et al. (ibid.) demonstrate that the impact of individual predictors in a tree-based ensemble using the relative variable importance (RVI) measure. De Bock (2017) proposes a spline-rule ensemble and, in addition to RVIs, presents partial dependence plots that show each predictor's contribution to the outcome, as well as the interactions between different predictors. The ability to quantify and visualize a model's functioning can significantly increase its attractiveness, especially to business practitioners.

In search of improved predictive performance, some recent studies present elaborate, complex models that combine multiple ensembling approaches and other techniques. For example, Wang & Ma (2011) and Zhu et al. (2017) find that RS-boosting, which combines the random subsample approach to a boosting algorithm, outperforms separate random subspace and boosting models. Zhu et al. (2019) further develop this approach into the RS-Multiboosting method, which integrates the Multiboosting algorithm (Webb, 2000) with RS. Another example of the complex nature of contemporary failure prediction methods is the E-SMOTE-ADASVM-TW algorithm proposed by Sun et al. (2020), which embeds the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) into ADASVM-TW (Sun et al., 2017), a time weighting-integrated AdaBoost ensemble of SVM classifiers. While many novel methods seem promising, evidence of their performance is scarce, and no comparison has been made between the various proposed techniques. Due to the sample-specificity of failure prediction models, the results of such methods should be interpreted with some caution, at least until supporting evidence emerges.

2.5.4 Alternative methods

Different modeling methods have been used to address the process nature and temporal dimension of failure; these mainly fall under the categories of statistical and machine learning methods, but are presented here separately due to the different approach. Shumway (2001) proposes a hazard model for failure prediction that demonstrates good predictive performance; hazard models have been applied in later studies, for example by Löffler & Maurer (2011) and El Kalak & Hudson (2016). Other suggested alternatives include Markov-type models (Petropoulos et al., 2016; Volkov et al., 2017) and self-organizing maps to quantify failure process types (du Jardin, 2018; du Jardin & Séverin, 2011). However, in general these techniques have remained in the minority, with most studies either using different means, such as dynamic variables measuring change in accounting ratios over time (Nyitrai, 2019), or disregarding the time dimension altogether.

As an alternative to the accounting-based approach, market information and the work of Black & Scholes (1973) and Merton (1974) on option pricing theory and contingent claims have been used to develop failure prediction models (Hillegeist et al., 2004; Vassalou & Xing, 2004). Findings regarding the superiority or inferiority of such models are contradictory; for example, Hillegeist et al. (2004) and Jackson & Wood (2013) report contingent claims models outperforming accounting-based alternatives, while Reisz & Perlich (2007) and Agarwal & Taffler (2008) conversely find that accounting-based models are equal to or better than methods built on contingent claims literature. Regardless of predictive performance, market-based models have a significant disadvantage in that they cannot be used with unlisted companies and are thus inapplicable to most SMEs (Pompe & Bilderbeek, 2005). This is also the case in this thesis; therefore, literature on market-based models is not explored further.

2.6 Model evaluation

For binary classification tasks, the most common evaluation metrics are based on correct and incorrect classification of the two classes, commonly labeled the positive (1) and negative (0) class. The formulation of the classification task in bankruptcy prediction, and therefore also the terminology used, varies between studies. Some authors denote failed firms as the positive and non-failed as the negative class, while in other studies the class definitions are reversed. In both the literature review and empirical part of this thesis, the positive class (1) refers to failed and the negative class (0) to non-failed companies. Performance indicators are derived from the total number of each of the four possible outcomes: correctly classified failed firms (true positives), failed firms misclassified as non-failed (false negatives), correctly classified non-failed firms (true negatives) and non-failed firms misclassified as failed (false positives).

One of the most common classification performance metrics in failure prediction is accuracy, i.e. the number of correctly classified firms relative to the total number of observations. It is also widely used in bankruptcy prediction; however, accuracy may not be an optimal measure where imbalanced sets are concerned, as it does not differentiate between false positives and false negatives (Bao et al., 2019; García et al., 2015). For example, if the data set has a 95% majority of the negative class, any prediction model can achieve a seemingly high 95% accuracy by classifying all observations as negative, while the model's recall (proportion of positives predicted correctly) in this case is 0, and its predictions are of little practical value. Even with balanced data, the lack of differentiation between the two types of errors is problematic, since classifying a failing firm as healthy is significantly more costly than predicting that a healthy company will fail (Lohmann & Ohliger, 2019a; Succurro et al., 2019). Many authors therefore report the true positive and true negative rates, or the false negative and false positive rates, to specify the frequency of different types of errors (Delen et al., 2013; Liang et al., 2016; López Iturriaga & Sanz, 2015). Although the error rates show how well the model recognizes failed and non-failed firms, imbalanced data may still cause misleading results, for example a low error rate in the majority class obscuring the large number of misclassified instances. Alternative measures have been used to present a more truthful picture of the results, such as the Matthews correlation coefficient (Bao et al., 2019), which has been shown to be highly robust against class imbalance (Boughorbel et al., 2017), or the F_β score (Serrano-Cinca et al., 2019), which takes into account different preferences regarding the precision-recall trade-off (Van Rijsbergen, 1974).

One issue related to the previously discussed metrics is that they only contain information related to the performance of the model at a fixed classification threshold, and do not consider the trade-offs that have to be made when using the model. A widely used tool to rectify this is the receiver operating characteristic (ROC) curve; it is particularly useful when class imbalance is present in the data or the misclassification costs are unequal (Fawcett, 2006); both of these occur commonly in failure prediction. The ROC curve depicts the true positive and false positive rates at different classification thresholds (i.e. the cutoff value determining whether an observation is predicted as failed or non-failed). The resulting plot illustrates the trade-off between higher recall (correctly predicted failures) and increase in false positives (non-failed firms classified as failed) (Zhou & Lai, 2017). Jones et al. (2017) note that the visualization power of the ROC curve can be particularly useful to practitioners, for example when determining a cutoff threshold that balances recall and specificity in proportions suitable to a bank's risk tolerance and credit policy.

The ROC curve can also be quantified using the area under the curve (AUC), which is frequently used in failure prediction studies (Fan et al., 2018; Le et al., 2018; Nyitrai & Virág, 2019). The AUC score is commonly seen as immune to class imbalance (Veganzones & Séverin, 2020); however, it can be overly optimistic when imbalance is high (Davis & Goadrich, 2006), and some authors (Xia et al., 2017; Zhang et al., 2019) address this by using alternatives such as the H-measure (Hand, 2009).

The various evaluation measures all have their benefits and shortcomings. Veganzones & Séverin (2020) state that no individual measure can convey all the relevant information about a model's predictive performance, and that multiple measures should therefore be used. This does seem to be the standard in failure prediction literature: although exceptions do exist that only rely on one approach (Le et al., 2018; Li et al., 2016), most studies use accuracy and related metrics in conjunction with the AUC score.

2.7 Summary and implications for empirical study

Corporate failure is a complex and unpredictable process, and previous findings can never be considered truly generalizable. Factors such as varying accounting practices, industry characteristics and other traits specific to certain populations of firms make it very difficult to identify specific variables that are consistently useful predictors across different samples. However, the extant literature can be used as a general guideline for the composition of the predictor set, because the broad categories of relevant predictors have been identified.

Despite extensive research, there are also no conclusive results regarding the superiority of classification techniques for failure prediction (Barboza et al., 2017). All methods have characteristics that make them relevant, and the choice depends on the objectives of the researcher or user (Veganzones & Séverin, 2020). However, recent studies indicate that ensemble machine learning methods could provide a much sought-after combination of performance and interpretability (Jones et al., 2017). The choice of prediction methods plays a large role in the need for data preprocessing; this must be taken into account when designing the study.

Overall, the theoretical basis for corporate failure prediction is somewhat thin, and the field is empirically oriented. Every new study can add something to the existing body of knowledge, but as new information is hard to come by and is mainly based on empirical findings, it is important to design studies with generalizability and replicability in mind.

3 Data and variables

3.1 Sample

The data used in the empirical study comprise the summarized financial statements and assorted additional information of 126 545 Finnish companies in the years 2008–2010 and data on declared bankruptcies in Finland in the years 2011–2012. The companies in the sample only include public and private limited companies; other types of business entities such as general partnerships, limited partnerships and sole proprietorships are excluded. Both bankruptcy and financial statement data are originally obtained from the database of Bisnode Finland Oy (Bisnode) and accessed through a financial analysis platform provided by Valuatum. Companies in the sample are not limited based on financial or other characteristics: the sole criterion for including a company is the existence of its financial statements in the Valuatum database. Firms with missing values are included, because it cannot be assumed that values are missing at random. All companies in the sample published their financial statements and were non-bankrupt for the full duration of the years 2008–2010. Firms younger than three years are left out, because their bankruptcy processes are different from older firms (Lohmann & Ohliger, 2019b), and therefore the same prediction model might not be efficient. Furthermore, using three years' data for modeling would be complicated, if there was a need to accommodate firms that only have data available for one or two years.

Non-SMEs are removed from the data following the European Commission's recommendation, under which SMEs comprise enterprises that employ fewer than 250 persons, and either have an annual turnover not exceeding EUR 50 million or an annual balance sheet total not exceeding EUR 43 million, or both (Commission Recommendation of 6 May 2003, 2003). However, adhering fully to these limitations is not possible: the number of employees is missing for approximately 70% of the companies. Therefore, the employee headcount criterion is dropped, and SMEs are defined in this study as companies with annual turnover of no more than EUR 50 million or annual balance sheet total of no more than EUR 43 million. Despite the limitations imposed by the lack of data, it is likely that the staff headcount of most of the companies defined here as SMEs does not exceed the maximum of 249, since larger companies are subject to stricter regulation and scrutiny and can therefore be considered more likely to accurately report their number of employees. Additionally, for most companies that fall significantly short of the turnover and balance sheet total limits, a headcount of 250 or more is operationally inadvisable and financially unsustainable. The sample inevitably contains firms that do not qualify as SMEs under the Commission's recommendation; however, it seems reasonable to assume that such companies are a small minority.

The turnover and balance sheet total values used for establishing the SME status of the companies in the sample are taken from the most recent available fiscal year (2010). The filtering out of non-SMEs results in the exclusion of 900 companies, leaving a total of 125 645 companies and a total of 376 935 firm-year observations, three for each company. The descriptive statistics for the data are presented in Table B1. Due to the large number of predictor variables, only the values from the latest fiscal year (Y-1, 2010) are shown for practical reasons. The preceding years' values are largely similar and offer no significant additional information. The data contain a total of 124 252 companies that remained operational at least until the end of 2011 and 1393 companies that were declared bankrupt before the end of 2011, amounting to 98.9% and 1.1% of the sample, respectively.

Table 1: Summary of data sample

Number of companies	Non-bankrupt	Bankrupt
125645	124252	1393
	98.9%	1.1%

3.2 Training, validation, and test set

Bankruptcy prediction studies commonly separate the sample into training and test sets; model optimization such as feature selection and hyperparameter tuning are performed and tested on the training set using cross-validation (e.g. Liang et al., 2015; Son et al., 2019). However, this is often due to data scarcity; when possible, the preferred approach is to include a separate validation set for feature selection and hyperparameter tuning (Hastie et al., 2009). A further aspect to consider in this study is the use of sampling methods to balance the training set. If cross-validation were used on the balanced set, features and hyperparameters would be optimized for maximum predictive performance on balanced data, while a significant class imbalance is present in the actual data the model is created to predict. By using a separate unadjusted validation set, this problem can be avoided.

There are no clear guidelines for the appropriate proportions in which to split the sample (ibid.); in this study, the training, validation and test sets contain 60%, 20% and 20% of the sample, respectively. The split proportions are in line with previous studies (see e.g. Alfaro et al., 2008; Son et al., 2019; Veganzones & Séverin, 2018; West et al., 2005). The train-validation-test split is performed separately and using different random seeds for the first and second rounds of modeling in order to reduce possible sample-specific effects.

3.3 Class imbalance

A notable class imbalance is present in the sample used in this study: the minority class, i.e. bankruptcies, comprise only 1.1% of the total number of companies.

Veganzones & Séverin (2020) suggest that balanced data yield more accurate predictions, but also note that large samples are needed to increase the robustness and reliability of the models, and that these two objectives are often mutually exclusive due to the small overall number of failed firms.

To avoid depleting the data too much, tools like the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) are often used. However, here the available data set is large and contains a total of 1393 bankrupt firms; assuming the proportions of bankrupt companies in the training, validation and test sets remain close to the original sample, the training data would contain approximately 835 bankrupt firms. This enables the use of a simpler method, random undersampling (RUS), which simply removes observations belonging to the majority class at random until the desired balance is achieved (Kim et al., 2015). A balanced set with equal numbers of bankrupt and non-bankrupt firms using this technique will still contain upwards of 1600 companies, whereas most recent studies use samples of 400 or fewer companies (Veganzones & Séverin, 2020). Zhou (2013) shows that RUS performs as well as more sophisticated methods when the original sample is large, and has the advantage of computational efficiency; this makes it a suitable choice for this study.

To assess the effects of class imbalance, prediction models are trained separately on the full, imbalanced training set and on sets balanced by random undersampling. In the first round of modeling, only a fully balanced (1:1) set is used in addition to the original, imbalanced one. In the second round, the best models are trained on sets with 1:1, 1:3, and 1:10 class distributions to further examine the effect of different levels of imbalance. Balancing is only applied to training sets; validation and test sets are kept intact to ensure that the model's performance is assessed using the true class distribution.

3.4 Output variable

The output variable in this study is a simple binary variable that indicates whether a company is expected to go bankrupt during the relevant period of observation; forecasted bankruptcy is indicated by "1" and non-bankruptcy by "0". The observation period for the output variable is two years: given the financial statements from 2008–2010, the models predict whether companies will go bankrupt during the years 2011–2012.

Each of the prediction models produces a numeric value $p \in [0, 1]$, corresponding to the probability of bankruptcy during the observation period. The binary classification output is obtained from the probabilities by determining a cutoff threshold value: predictions smaller than or equal to the threshold value are classified as 0 (non-bankrupt) and predictions exceeding the threshold as 1 (bankrupt). This study uses varying thresholds when assessing binary classification performance: instead of a fixed threshold, the cutoff values are adjusted so that each model achieves the same level of correctness with regard to non-bankrupt companies. The choice of cutoff thresholds is described in Section 4.5. In practical use, the cutoff value can be adjusted to suit the specific use case: for example, a bank could use a threshold of 0.1 to classify loan applicants with over 10% probability of bankruptcy as too risky.

3.5 Predictor variables

3.5.1 Selection of initial variable set

The input variables in this study are chosen from three different sources that are explained in more detail in the following sections. First, a set of predictors is assembled from prior empirical studies. This selection is augmented with the predictors from the prediction model previously used by Valuatum. Lastly, due to the contingent nature of bankruptcy, and to address any perceived deficiencies in the predictor set assembled from the first two sources, some additional variables are picked from outside the established failure prediction literature.

All predictor values are based on the available financial statement data from 2008–2010. All variables that only use financial statement items from a single fiscal year are calculated for each firm-year observation. However, variables measuring change over time cannot be calculated for the earliest firm-year observations, as data from years preceding 2008 are not available for this study; these exceptions are specified in the following sections that present the predictor variables. The values of the same financial ratio in different years are treated as separate variables; for example, feature selection procedures may remove the 2008 and 2009 values of some ratio, while selecting the 2010 value as a relevant feature. The full list of abbreviations used in the predictor variable definitions is presented in Appendix A.

3.5.2 Variables from three prior studies

Existing empirical research on failure prediction is extensive, and a vast number of predictor variables have been used. Although the predictive ability of specific variables is not universal, the broad categories and types of useful predictors are mostly agreed upon. Earlier studies therefore provide a suitable starting point for assembling the set of predictors for this study. However, it must be noted that the choice of best predictors is usually sample-specific, and empirical findings are therefore not directly generalizable (Balcaen & Ooghe, 2006). To counter this issue, this study uses three different literary sources with diversified predictors, and the observed performance of the predictors is not used as a criterion for selection. After the predictor set is compiled, it is qualitatively assessed to ensure that the main categories are represented by a sufficient number of diverse predictors. The final selection of predictors from the three studies is presented in Table 2. The table is organized by category (profitability, capital structure, liquidity, solvency, activity, growth, size) for readability. However, the categories are not mutually exclusive, and many variables could justifiably be placed in a different category.

The first source is a review study by Dimitras et al. (1996), which provides a listing of all the financial variables used in a total of 59 reviewed failure prediction models, including such influential studies as Altman (1968) and Ohlson (1980). Due to the robustness of financial ratio predictors (Beaver et al., 2005) and the fact that the general view regarding the determinants of failure has not changed in the last decades, these variables form a good basis for the predictor set in this study. All of the predictors that appear in the papers reviewed by Dimitras et al. (1996, tables 4a & 4b), that can be calculated using available data, are used in this study.

The second article, by du Jardin (2015), is an empirical study that uses a set of 50 financial predictor variables selected based on prior literature and representing the main categories of liquidity, solvency, profitability, capital structure, activity, and turnover; in this study, turnover ratios are grouped under activity instead of being a distinct category. The selection process or exact sources of the variables are not specified. The same selection of predictors has been used in later studies (e.g. Veganzones & Séverin, 2018), indicating that it is considered useful and sufficient by itself. The variables overlap somewhat with those listed by Dimitras et al., but also include numerous new predictors, thus augmenting the variable set of this study.

The third study used for choosing the variables, conducted by Jones et al. (2017), also employs an approach that utilizes earlier research. All the categories used by du Jardin are represented, but with different emphases. For example, fewer liquidity and profitability ratios are used, but capital structure is well represented. Jones et al. use a somewhat more experimental approach and include predictors that are less common in the literature, such as intangible assets and asset write-downs. Additionally, the variables used by Jones et al. incorporate two aspects that are entirely absent from the other two papers: company size (proxied by total assets and sales) and growth variables. Some new variables are also included in the traditional predictor categories, although some overlap with the other two studies is observed.

Some variables from the three studies are omitted due to data unavailability. For example, market-based predictors are not included, because the information is not available in financial statements. Certain financial variables, such as financial debt and no credit interval, must also be excluded, because the available data are not sufficiently specific for calculating them. Dimitras et al. (1996) and du Jardin (2015) do not provide detailed variable definitions, and Jones et al. (2017) use a sample of US companies that differ notably from Finnish companies in terms of accounting standards and practices. Therefore, the exact formulas of the predictors used in this thesis may not be the same as in the original studies, and require a degree of interpretation. However, basic accounting terminology is quite universal, and therefore it can be assumed that the underlying principles justifying the chosen variables also apply in the context of this study.

Table 2: Predictor variables selected from prior studies

	(1)	(2)	(3)		(1)	(2)	(3)
Activity				Profitability			
AP/COGS*	x			EBIT/S*			x
AP/S	x	x		EBIT/TA	x	x	x
AR/S*	x	x	x	EBIT/TE		x	
CA/S		x		EBITDA/S		x	
CAPEX/TA			x	EBITDA/TA		x	
EBIT/VA		x		GP/S*	x	x	
EE/VA	x			GP/TA	x		
I/COGS			x	NI/S*	x	x	x
I/S*	x	x		NI/TA	x	x	
NI/VA		x		NI/TE		x	
NWC/S				OCF/S		x	
OCF/VA		x		OCF/TA		x	x
S/TA	x	x	x	OCF/TE		x	
VA/FA		x		RE/TA	x		
WC/S		x		VA/S		x	
WD/TA			x	VA/TA	x	x	
Capital structure				Solvency			
CA/TA	x	x		CA/TD	x		
CL/TD	x			EBIT/IE	x		
FA/TA	x			FE/EBITDA		x	
I/NWC	x			FE/NI		x	x
I/TA		x		FE/TA		x	
IA/TA			x	FE/VA		x	
NWC/TA		x		IE/GP	x		
QA/TA	x	x		IE/S	x		
SC/TC	x			NI/IE		x	x
TD/TE		x	x	OCF/TD			x
TE/TA	x	x		TD/TA	x	x	x
TE/TL*			x				
WC/TA	x	x	x	Liquidity			
				C/CA		x	
Annual growth				C/CL	x		
growth CAPEX			x	C/S	x		
growth NI			x	C/TA	x	x	
growth OCF			x	C+MS/CL		x	
growth TD			x	C+MS/S		x	
growth WC			x	CA/CL*	x	x	x
				CL/S		x	
Company size				CL/TA	x	x	
S*			x	NI/CL	x		
TA*			x	QA/CL*	x	x	

The appearance of the variables in the studied papers is indicated as follows:
(1): in at least one of the studies reviewed by Dimitras et al. (1996, tables 4a & 4b)
(2): in du Jardin (2015)
(3): in Jones et al. (2017)
Variables marked with (*) were also used in the previous prediction model used by Valuatum.

3.5.3 Variables from previous prediction model used by Valuatum

The variables utilized in the previous prediction model used by Valuatum have been found effective through empirical testing. Due to the sample-specificity of predictors' impact to the performance of the model, these variables can incorporate some useful properties that are not captured by scientific literature, and are therefore added to those chosen from prior studies. Many of the predictors in the Valuatum model are also found in the studies discussed in the previous section, and are listed in Table 2. The chosen Valuatum variables that do not appear in any of the three studies examined in Section 3.5.2 are listed in Table 3.

Many scholars note that bankruptcy risk is affected by the industry a company operates in, and incorporate its effect in risk modeling using a dummy variable. This study takes a different approach by utilizing data on past bankruptcies to calculate the average bankruptcy rate of different industries to measure the relative riskiness of the industry.

The previous Valuatum model includes an industry-specific bankruptcy risk indicator (ind_risk) that is also utilized in this study. Unlike the common approach of industry sector dummy variables, this variable incorporates concrete information about the riskiness of each industry in the two preceding years, as shown in Equation 1. Industries are classified using NACE Revision 2, as established in Regulation (EC) No 1893/2006 of the European Parliament and of the Council (2006). By default, the 4-digit NACE code is used; if there are fewer than 100 companies in the industry and/or no occurred bankruptcies in the previous two years, the parent industry (3-digit NACE code) is used. If the requirements are still not fulfilled, the wider industry sector corresponding to the 2-digit NACE code is used, and if this also fails, the industry risk defaults to 1.5%. The industry risk variable is calculated as

$$\text{ind\_risk}(Y) = \frac{b_{ind}(Y-1) + b_{ind}(Y)}{n_{ind}(Y-2)} \tag{1}$$

where Y is the year for which industry risk is calculated, n_{ind} is the total number of companies in the industry ind, and b_{ind} is the number of declared bankruptcies in industry ind during the specified year.

The industry risk variable can be seen as somewhat problematic, as it incorporates information from years preceding the period from which financial statements are available. However, it is not dependent on financial statement values and can be calculated even for firms that have no financial statements before 2008. In the Valuatum database, industry risk is available for each firm-year observation as a precalculated variable and does not require additional information; in the practical context of this study, it does not differ from typical financial ratios and is therefore included.

Certain variables are excluded because they are deemed unsuitable for the purposes of this study. The categorical variables S_growth_count and EBIT_count use financial statement data from outside the time scope defined in this study to describe the development of sales growth and EBIT, respectively, and are therefore omitted. The unscaled variables EBITDA, net income, net working capital, profit before depreciation, amortization and extraordinaries, and working capital, are also excluded: they combine information about the company's size and the various aspects of its financial status, and are therefore difficult to interpret. Firm size can be an impactful factor in failure risk, and to better observe its effects, no unscaled variables apart from the designated size proxies (see Table 2 section "Size") are included in the model.

Table 3: Predictor variables selected from previous Valuatum model

(EBIT+FI)/TA	TD/PBD	TL/S
(EBIT+FI)/TC	(TD-C)/EBITDA	(TL-C)/S
ind_risk	(TD-C)/TE	(TL-TD)/S
PBD/S		

3.5.4 Additional variables from miscellanous sources

Due to the variability of the observed effectiveness of specific predictors between studies and the unpredictable nature of bankruptcy prediction (Balcaen & Ooghe, 2006), this study uses a similar approach to Bauweraerts (2016) and Jones et al. (2017), and introduces some less commonly used predictor variables based on various academic sources and the author's own judgement. Some observed deficiencies of the predictor set constructed from the other two sources are also addressed. The additional variables are presented in Table 4.

While high leverage is linked to bankruptcy risk, Sanfilippo-Azofra et al. (2016) find that firms facing financial distress, in part due to already holding large amounts of debt, tend to seek additional financing by raising new equity. Naturally, companies may issue new shares for other reasons, many of which indicate positive developments rather than impending bankruptcy. However, in conjunction with other variables the annual growth of share capital (*growth SC*) could have predictive power. A precedent for the use of change in share capital as a predictor of failure is found in Kim & Sohn (2010). On the other hand, according to some authors distressed firms typically rely more on trade credit than healthy firms (Altman et al., 2010; Molina & Preve, 2012); change in accounts payable (*growth AP/S*) is therefore also included.

Change in total debt is already included from the study of Jones et al. (2017); to add a further dimension and perhaps eliminate the effect of changes in company size, growth in debt relative to assets (*growth TD/TA*) is added. For similar reasons, change in profitability ratios is also used for prediction (*growth GP/S; EBIT/S; PBD/S; NI/S*).

Employee efficiency is a key factor of business performance and can serve as a measure of failure risk (Lin et al., 2012a). Mature firms in particular can suffer from rigidity that contributes to weak performance and eventually bankruptcy; high employee expenses are one potential symptom (Kücher et al., 2018). The variables picked from extant literature already include one employee efficiency measure (*EE/VA*), but some alternatives (*EE/S, EE/PBD, EE/NI*) are added to better observe the effects. Unfortunately, the number of employees is not available for the majority of companies, and therefore is not used.

The ratio of receivables to sales typically depends on a company's credit policy, but can usually be assumed to remain quite stable over time. Altman et al. (2010) assert that small companies, when financially distressed, typically extend more credit to customers; additionally, significant changes in the relative amount of receivables can be a sign of earnings management, which is more common in troubled than healthy companies (Serrano-Cinca et al., 2019; Wells, 2001). However, the relationship between receivables and financial trouble is not entirely straightforward: Box et al. (2018) suggest that a permissive credit policy can be a means of gaining a competitive advantage. Due to the possible earnings management implications, total receivables are used for added robustness against attempts to obscure increases in accounts receivable; the variable is included both as a static ratio (*TR/S*) and its annual change (*growth TR/S*).

Using data from a three-year period and including annual growth variables can be considered sufficient for taking the process nature of bankruptcy into account. However, to capture developments occurring over time more efficiently, all of the growth variables are also included as two-year compound annual growth rates, CAGR (*cagr AP/S; CAPEX; EBIT/S; EE/NI; EE/PBD; EE/S; EE/VA; GP/S; NI; NI/S; OCF; PBD/S; SC; TD; TD/TA; TR/S; WC*).

Table 4: Additional predictor variables selected for modeling

EE/NI
EE/PBD
EE/S
TR/S
growth AP/S
growth EBIT/S
growth EE/NI
growth EE/PBD
growth EE/S
growth EE/VA
growth GP/S
growth NI/S
growth PBD/S
growth SC
growth TD/TA
growth TR/S
cagr AP/S
cagr CAPEX
cagr EBIT/S
cagr EE/NI
cagr EE/PBD
cagr EE/S
cagr EE/VA
cagr GP/S
cagr NI
cagr NI/S
cagr OCF
cagr PBD/S
cagr SC
cagr TD
cagr TD/TA
cagr TR/S
cagr WC

3.6 Missing values

Analyzing missing values is challenging, if not impossible, with the data used for this study: missing values in the database are coded as zeroes, and are therefore indistinguishable from actual zero-valued variables. Furthermore, variables with calculation errors such as division by zero may be omitted and appear as missing values. As there is no reliable means of ascertaining whether a specific data point is a missing value or not, no actions are taken to either remove or impute missing values.

Even if it were possible to detect missing values, keeping them unaltered in the data is a valid approach in a bankruptcy prediction context. Many failing firms do not report their annual accounts in the last years before failure (Balcaen & Ooghe, 2006), and Lukason & Camacho-Miñano (2019) find that low profitability and liquidity, which are also known determinants of failure, are linked to delays in financial reporting. This indicates that companies facing financial distress are more likely to have missing values in their financial statements; imputing missing values or removing firms with missing data could induce bias and distort the prediction results (Zmijewski, 1984). Apart from the logit model, the prediction models used in this study are based on decision trees, which are by nature robust against missing values (Hastie et al., 2009). However, the inability to differentiate between zeroes and missing values is somewhat problematic, particularly because of the heteroskedasticity of the variables. A value of zero may indicate different things for each predictor; for example, a gross margin of zero is quite concerning, while zero annual change in share capital does not appear to indicate anything specific. Therefore, when missing values for predictor variables are interpreted as zeroes, the implications are different in each case.

Nonetheless, the issue of missing values as zeroes is present in the data and cannot be rectified within the scope of this thesis; it must simply be accepted as one aspect of the unreliability of data that is common in financial research. Furthermore, the prediction model developed in this study will eventually be used in practice with similarly processed data. Therefore, any adjustments to combat missing data could deteriorate its predictive capacity and give misleading results.

Although the existence of true missing values cannot be investigated, some predictor values in the latest fiscal year (2010) are examined to assess the quality of the data. It should be noted that sum variables, such as fixed assets, current liabilities or EBIT, are calculated from individual financial statement items; if they are non-zero, it indicates that at least one of their constituent financial statement items is also non-zero. Out of 125 645 companies, the sample contains 72 firms with zero total assets, and 951 firms with zero total equity; these numbers do seem reasonable and indicate that at least some balance sheet items are available for almost all companies. Total liabilities are zero for 7 639 companies, which also seems plausible and gives no reason to suspect major issues with missing data. Sales are zero for 13 090 and EBITDA for 7 162 companies; for 2 536 firms, sales, EBITDA and net earnings are all zero. This indicates that the sample may contain some inactive companies, but the number of zero values does not seem unreasonably high, and income statement figures can also be considered quite reliable.

3.7 Outliers

The values of the data obtained from the Valuatum database are limited to $\pm 10^8$; any values outside these limits have been automatically adjusted to the corresponding limit value. Despite the modification, these values remain outliers: as the descriptive statistics (Table B1) show, the observed minimum and maximum values for all but three variables are far from these limits. One explanation for the absence of outliers is that division by zero is treated as an error in the Valuatum system, and results in a zero value instead of $\pm \infty$, which would appear as $\pm 10^8$ in the sample.

Due to the treatment of zero division, applying additional outlier handling methods would result in increased inconsistency. If the data were winsorised, for example, the sample would contain two different types of outliers that have been adjusted using different methods. Additionally, outliers are a common occurrence in real-life financial data, and therefore prediction models must be able to process them (Zoričák et al., 2020). For these reasons, and to maintain the modeling process as simple as possible, outliers are not removed or modified in any way, apart from the trimming of extreme values that has been applied prior to extracting the data.

3.8 Variable transformations

Financial ratios tend to be non-normally distributed, and various transformations have been used in the literature to make them more suitable for failure prediction. However, decision trees, and therefore also ensembles built from them, are immune to monotonic transformations (Hastie et al., 2009). Empirical evidence from prior credit risk literature corroborates ensemble models' indifference to variable transformation (Jones et al., 2015, 2017). To keep the modeling process simple and reduce the number of extra steps required, no variable transformations are applied in this study.

4 Methodology

4.1 Empirical design and implementation

The empirical study is conducted in two phases. In the first phase, the performance of five different classification methods is compared. Each classifier is separately trained using balanced and imbalanced training data, and three different feature selection methods are applied, resulting in a total of 30 distinct models. The aim of the first phase is to assess the impact of some of the key modeling choices on bankruptcy prediction performance and to find the most promising combinations.

Due to the large number of different classifiers and the considerable size of the data sample, it is not feasible to perform thorough hyperparameter tuning and repeated testing for the full selection of prediction models; therefore, only the best candidates are picked for further examination and development. The exact number of combinations of classification method, training data balance and feature selection approach is determined by assessing the results of the first phase. During the second phase, the prediction models are trained using a more detailed set of hyperparameters, and test scores are averaged over multiple runs for added robustness. The classification methods chosen for further examination in the second phase are random forest and gradient boosting. Class imbalance is also explored further; feature selection is found to be of no significant benefit and is therefore not used in the second phase.

To ensure the reliability and usability of the prediction models both during this study and in continued commercial use, a robust and well documented software platform is required. The project is implemented using the Python programming language, including various third-party libraries; most notably, the scikit-learn package (Pedregosa et al., 2011) is used for developing, training and testing the classification models. Scikit-learn is fully open-source, features a large selection of off-the-shelf classification methods and other necessary tools, and is extensively documented. Moreover, it ranks high among the most used machine learning libraries on the popular version control platform GitHub (Elliott, 2019) and has a proven track record of commercial application (Scikit-learn, 2020), making it an attractive choice for the purposes of this thesis. The documentation and source code of scikit-learn can be consulted for technical details on the learning algorithms presented in the following section.

4.2 Classification methods

4.2.1 Method selection

The classification methods used in this thesis are chosen based on three main criteria. First, there must be a sound theoretical basis and sufficient empirical evidence demonstrating the method's predictive ability. Second, the resulting model must be understandable and the effect of different predictor variables must be clearly interpretable. Third, the method should be easy to implement and modify when necessary, even by relatively inexperienced users.

Traditional statistical methods are burdened by extensive evidence of insufficient performance compared to more recent techniques, and the various requirements placed on input data complicate the modeling process. They are therefore not considered viable options for this study. However, logistic regression remains widely used in bankruptcy prediction literature, and is a common benchmark in modern studies (Serrano-Cinca et al., 2019; Tsai & Hsu, 2013). Comparing logit to alternative methods may be of particular interest to business practitioners due to its widespread use in current credit scoring models (Zhang & Thomas, 2015). For these reasons, logistic regression is included as one of the modeling techniques.

Based on the requirement of model interpretability, two of the most commonly used machine learning techniques are ruled out: support vector machines (Verikas et al., 2010; Yao & Chen, 2019) and neural networks (Figini et al., 2017; López Iturriaga & Sanz, 2015; Sun et al., 2011) are both commonly described as "black box" models that, while usually performing well, are difficult to interpret. Furthermore, neural networks are computationally demanding (Chung et al., 2008), and support vector machines may also require a substantial amount of computational resources (Huang et al., 2004). Decision trees have many attractive qualities, but suffer from instability and weak performance. The less commonly used machine learning methods are typically not available as off-the-shelf methods and literature on them is somewhat scarce, which limits their usability.

Following recent trends in bankruptcy prediction literature, ensemble techniques are identified as an appealing alternative. Three methods emerge as the most viable options based on frequency of use and observed performance in the literature, as well as practical considerations: random forests (RF), AdaBoost, and gradient boosting (GB). Random forests (García et al., 2019; Jones et al., 2017), AdaBoost (Barboza et al., 2017; Zhou & Lai, 2017) and gradient boosting (Brown & Mues, 2012; Pawełek, 2019) have all performed well in previous bankruptcy prediction and credit risk studies. They are widely studied and used, and have ready implementations in many popular machine learning libraries and platforms, including the scikit-learn package used in this study. This selection also ensures that both of the major ensemble techniques, averaging (RF) and boosting (AdaBoost, GB) are represented.

In this study, the random forest ensemble is implemented with decision trees as base learners, because this is required by design; the use of alternatives is not possible. Both boosting algorithms are also implemented with decision trees, which provide weak, unstable base learners in accordance with the underlying principles of boosting (Schapire, 1990). Furthermore, the lack of interpretability associated with the two major alternatives, NN and SVM, would still be present in an ensemble model, which rules out their use in this study. Empirical evidence from failure prediction also indicates that decision trees are the preferred choice for boosting in terms of performance (Marqués et al., 2012; Tsai et al., 2014). To examine the effects of ensembling versus individual classifiers, a standalone decision tree model is also trained.

Based on the results of the first modeling phase, two methods are chosen for further analysis in the second phase: random forest and gradient boosting. This choice of models is discussed in more detail in Section 5 together with the results.

4.2.2 Decision trees

Decision trees are a widely used machine learning technique that can be used for both classification and regression; the general term "classification and regression trees" (CART) was introduced by Breiman et al. (1984). This section focuses solely on decision trees for binary classification. There are many different algorithms for constructing tree classifiers; the technicalities are not discussed here in detail. In this thesis, decision trees are used both individually and as base learners for the random forest, AdaBoost and gradient boosting classifier ensembles.

Decision tree classifiers are constructed by recursively splitting the training set $\mathcal{X}$ into two smaller subsets ("nodes"), with the objective that the subsets formed in each split discriminate better between the two classes than their parent node. There are many different metrics for the goodness of a split, such as the Gini impurity and information gain. The split is performed by using the best predictor variable and a cutoff value, according to which the observations are divided into two smaller nodes. The splitting continues until a predetermined stopping criterion is reached; for example, the maximum depth of the tree is reached, or all instances in the node belong to the same class and no further splits are needed. After the tree has been trained, it can be used for classification by traversing the tree according to the split rules, until an unsplit terminal node ("leaf") is reached; the leaf indicates the class probabilities and final classification output for the instance under consideration. Algorithm 1 describes a basic algorithm for growing a decision tree classifier.

Algorithm 1 Decision tree for classification

```
Initialize tree $T$ with root node containing the full training sample
for all non-leaf terminal nodes $t$ in $T$ do
    Choose the best split point in the feature space
    if a stopping criterion is reached then
        $t$ is a leaf
    else
        Split $t$ into two child nodes
    end if
end for
```

Decision tree classifiers have many attractive qualities such as the capability to process mixed data types, outliers and missing values, and the ability to model complex nonlinear relationships (Hastie et al., 2009). They are also computationally efficient with large, high-dimensional data sets (Florez-Lopez & Ramon-Jeronimo, 2015), which is often useful in a failure prediction context. One notable advantage of decision trees is their interpretability: the classifier can be presented in the form of simple if-else rules, and is also easy to visualize.

Figure 1 shows an example of a (weak) decision tree classifier. The nodes display the splitting criteria, the Gini impurity of the node (lower impurity indicating better discriminative ability), and the number of firms belonging to each class in the node. The tree assigns a bankruptcy probability and the corresponding binary class to new observations based on the number of bankrupt and non-bankrupt samples in the leaf it belongs to. For example, a company with TE/TL Y-1 -0.05 and S Y-1 0.1 belongs to the leftmost leaf, and thus its probability of bankruptcy is $163/778 \approx 0.21$. The only combination that results in classification as bankrupt is (TE/TL Y-1 $\leq$ -0.032 && S Y-1 $\geq$ 0.128), with a failure probability of $456/764 \approx 0.60$. As can be seen, this tree produces somewhat inaccurate classification rules; three of the four leaves do not have a significant majority of one class; in other words, they cannot discriminate between the classes very efficiently. A leaf created mostly of instances in one class has good discriminative ability on the training set; of course, this does not necessarily indicate good out-of-sample performance.

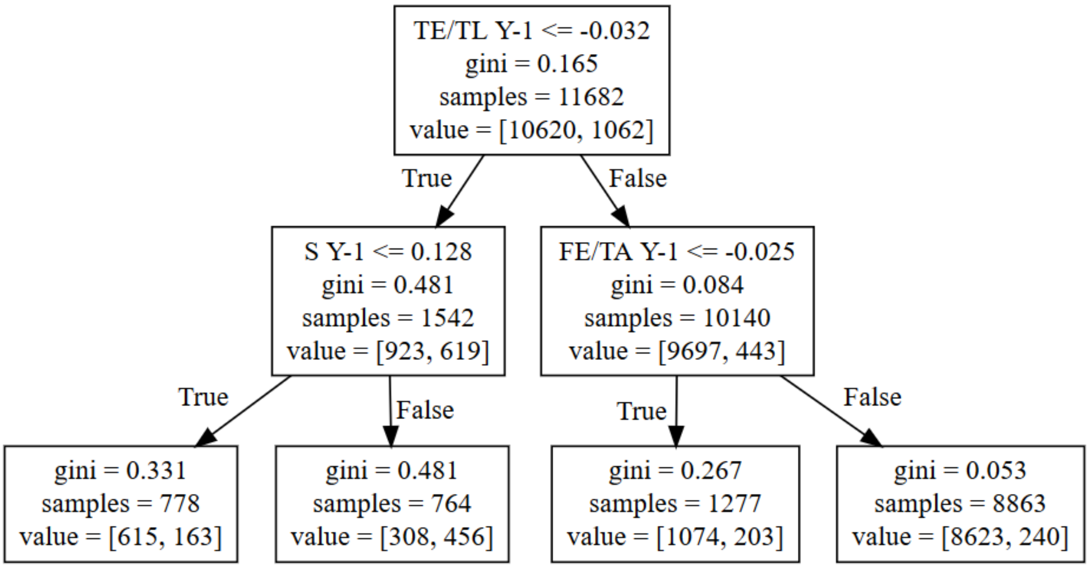

Figure 1: Simple decision tree classifier

4.2.3 Random forests

Random forests (Breiman, 2001) are an ensemble learning technique based on constructing multiple decision trees in parallel and joining their predictions to form the final classification or regression output. The random forest algorithm combines the principles of two averaging ensemble methods, bootstrap aggregating (bagging)

(Breiman, 1996) and random subspace (Ho, 1998). In traditional bagging, the trees are usually highly correlated and more likely to misclassify the same instances; random forest avoids this by using random feature subsets for each node split in the trees. Introducing randomness into the growing of individual decision trees reduces their correlation and produces more diverse trees, thereby improving the ensemble's predictive performance.

A random forest for classification (Algorithm 2), in its simplest form, is built by constructing an ensemble of decision trees using the following procedure, as outlined by Breiman (2001). First, a bootstrap sample (i.e. sample with replacement) of observations is drawn randomly from the training set. Using the bootstrap sample, the tree is built as described in Algorithm 1, by recursively splitting nodes using the best available feature. Instead of the full feature space, a random subset (with replacement) of the predictors is drawn at each node when considering the split. The output of the model is determined by a majority vote of the individual trees. The implementation of the scikit-learn package used in this study differs from the original in that the output of individual trees is combined by averaging the predicted probabilities rather than majority voting on the class predictions.

Algorithm 2 Random forest for classification (adapted from Hastie et al., 2009)

```
for $b = 1$ to $B$ do
    Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
    Grow a decision tree $T$ as follows:
    for all non-leaf terminal nodes $t$ in $T$ do
        Select $m$ of the total $p$ features randomly from the data
        Choose the best split point among the $m$ features
        if a stopping criterion is reached then
            $t$ is a leaf
        else
            Split $t$ into two child nodes
        end if
    end for
end for
```

4.2.4 Boosting methods

Boosting ensemble methods, first proposed by Schapire (1990), are based on the concept of combining a number of weak learners to form a strong classifier. Unlike random forest and other averaging techniques, boosting methods train base learners sequentially. After training a base learner, its observed performance is used for adjusting the priorities of the model: the next iteration focuses more on instances that were previously misclassified. The most popular and widely used boosting methods are AdaBoost (Freund & Schapire, 1997) and gradient boosting (Friedman, 2001). Although the original motivations for these two algorithms were quite different, their functioning is very similar: AdaBoost was, years after its inception, shown to be equivalent to a forward stagewise additive model, which in turn forms the basis for the gradient boosting approach (Hastie et al., 2009).

The basic principle of AdaBoost (Algorithm 3) is that, with each new learner, the algorithm adapts training data based on the errors made by previous learners. Before training a new base learner, the results of the preceding one are assessed, and sample weights are adjusted. Misclassified instances are assigned a higher weight, while the weight of correctly classified observations is reduced. This leads the algorithm to focus on the instances that are difficult to classify and improves its ability to predict them correctly. After the predetermined number of iterations have been completed, the predictions of the individual base learners are combined with a weighted vote based on their performance to produce the output of the AdaBoost model.

Algorithm 3 AdaBoost (Schapire, 2012)

```
Training data $(x_1, y_1), \dots, (x_m, y_m)$, where $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$
Initialize $D_1(i) = 1/m$ for $i = 1, \dots, m$
for t=1,...,T do
    Train weak decision tree using distribution $D_t$
    Get weak hypothesis $h_t : \mathcal{X} \to \{-1, 1\}$
    Aim: select $h_t$ to minimalise weighted error $\epsilon_t = \mathbf{Pr}_{i \sim D_t}[h_t(x_i) \neq y_i]$
    Choose $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
    for $i = 1, \dots, m$ do
        $$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$$
        Where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ is a distribution
    end for
end for
Output final hypothesis:
    $$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$
```

In contrast with AdaBoost, gradient boosting (Algorithm 4) approaches prediction from a numerical optimization perspective: it is essentially a gradient descent algorithm that minimizes the classifier's loss function on the training set. The model is initialized with a weak learner; in every subsequent iteration, a learner is fitted to the gradient of the loss function and added to the existing model. Thus, the error of the ensemble is reduced with each iteration, resulting in a high-performing final model.

Algorithm 4 Gradient boosting for binary classification (Hastie et al., 2009)

```
Initialize $f_{k0}(x) = 0,\ k = 0, 1$
for $m = 1$ to $M$ do
    Set
        $$p_k(x) = \frac{e^{f_k(x)}}{\sum_{\ell=1}^{K} e^{f_\ell(x)}},\ k = 0, 1$$
    for $k \in \{0, 1\}$ do
        Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1, 2, \ldots, N$
        Fit a regression tree to the targets $r_{ikm},\ i = 1, 2, \ldots, N$,
        giving terminal regions $R_{jkm}, j = 1, 2, \ldots, J_m$
        Compute
            $$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_t \in R_{jkm}} r_{ikm}}{\sum_{x_t \in R_{jkm}} |r_{ikm}|(1 - |r_{ikm}|)},\ j = 1, 2, \ldots, J_m$$
        Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm})$
    end for
end for
Output $\hat{f}_k(x) = f_{kM}(x),\ k = 1, 2, \ldots, K$
```

4.2.5 Logistic regression

Although the name might suggest otherwise, logistic regression (logit) is a linear method for classification. The logit model is similar to linear regression, but assumes a linear relationship between the predictor variables and the log-odds of the (binary) response variable, rather than directly between the predictors and the output. The logit model for binary classification can be expressed as follows:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2}$$

where p is the probability of the observation belonging to the positive class (1), $\beta_{0\ldots n}$ are the model coefficients, and $x_{1\ldots n}$ are the predictor variables. The model is first fitted to the training data using one of various solver algorithms in order to find the coefficients β. The predicted log-odds for an observation can be calculated by simply plugging in its predictor values to Equation 2.

4.3 Feature selection

Feature selection is considered an important part of failure prediction, as the presence of irrelevant or redundant variables can weaken predictive performance (Veganzones & Séverin, 2020). This study makes extensive use of decision trees, which essentially perform feature selection within the training algorithm when splitting nodes; therefore, the need for selecting a subset of predictors may not be as significant as with other classification methods. Nonetheless, there is evidence of feature selection improving the performance of standalone decision trees and ensemble models (Lin & Lu, 2019). For this study, one technique is chosen from both of the main feature selection categories, filter and wrapper methods. In addition to the feature selection methods, prediction models are trained without selection using the full set of predictors.

There are no reliable guidelines for selecting an appropriate number of features; quantitative measures can be used to assess the increase/decrease in performance from removing a feature, or the number of features can be chosen arbitrarily. For this study, 25 is chosen as a suitable number of predictors that keeps the model interpretable and can be assumed to contain enough information; many studies use fewer variables and obtain good results.

Mutual information (Battiti, 1994) (MI), a filter method, is used in this study to assess the relevance of the predictors statistically, without considering the different classification methods. The features are ranked by measuring their mutual information with the class output, i.e. how much relevant information about the bankruptcy status of the firms each feature holds. For binary classes, the mutual information of a feature f_i with the target class c is defined as

$$\mathrm{MI}(c; f_i) = \int P(0, f_i) \log \frac{P(0, f_i)}{P(0)P(f_i)} df_i + \int P(1, f_i) \log \frac{P(1, f_i)}{P(1)P(f_i)} df_i \qquad (3)$$

MI is advantageous compared to many alternative filter methods in that it can take nonlinear relationships into account (Bennasar et al., 2015). One downside is that MI is computationally complex and in practice has to be approximated (Battiti, 1994); especially with continuous variables, the approximation may be difficult if sufficient data are not available (Peng et al., 2005). However, data scarcity is not an issue for this study, and MI has been applied successfully in earlier bankruptcy prediction literature (Chan et al., 2006); it is therefore deemed an appropriate choice of filter method.

A wrapper method, recursive feature elimination (RFE), is also used for an alternative selection of the most important predictors. The basic principle of the method is to recursively consider smaller and smaller subsets until the best subset of some predetermined size is found. RFE trains a specific classifier using the considered variable subsets to assess the importance of the predictors. The technique is therefore applied separately to each of the prediction methods used in this study.

The process, as described by Guyon et al. (2002), is simple. The model is trained normally on the training data; the predictors are then ranked by relevance according to some criterion, and the lowest-ranking predictor or predictors are removed. This process is repeated recursively until a feature subset of the desired size is found. In this study, the ranking criterion for the predictors is relative variable importance (see Section 4.5.4) for standalone and ensembled decision trees, and model coefficients

β for the logit model (see Equation 2). Due to the large feature space and limited computational resources, 40% of predictors are discarded during each iteration of the elimination process.

4.4 Hyperparameter tuning

Each of the prediction methods involves several hyperparameters that can be adjusted to change the functioning of the algorithm. Literature on hyperparameter tuning is somewhat scarce; the optimal choice of parameters depends on a variety of factors from the technical implementation of the algorithms to the characteristics of the data, and definitive guidelines cannot be established. Therefore, the choice of which parameters to tune and which values to test relies to some extent on the documentation and default hyperparameter values of the scikit-learn package. This section describes how the key parameters for each classifier are chosen; the full list of tested hyperparameter values is presented in Appendix C.

The hyperparameter value combinations are tested using grid search, which takes a set of discrete values for each hyperparameter and exhaustively tests every possible combination. It is the most commonly used strategy for hyperparameter tuning in machine learning tasks in general (Bergstra & Bengio, 2012), and is also used commonly in corporate failure prediction (e.g. Sigrist & Hirnschall, 2019; Son et al., 2019; Volkov et al., 2017; Zoričák et al., 2020). Alternative, often more efficient methods exist, but these usually involve a degree of randomness, and grid search is therefore preferable from an academic perspective due to its transparency and reproducability.

Mantovani et al. (2018) find that for decision trees built using the CART algorithm, which is used in the scikit-learn implementation, the minimum number of observations in a node required to consider splitting (`min_samples_split`), and minimum leaf size (`min_samples_leaf`), which prevents splits if the resulting leaves would contain too few observations, are important. They further note that small, shallow trees perform poorly; the depth of the tree (`max_depth`) should not be limited too much. In this study, `min_samples_leaf` and `max_depth` are included in hyperparameter tuning; `min_samples_split` is omitted, because it functions similarly to `min_samples_leaf` in controlling tree complexity. Due to the imbalance of the training data, a hyperparameter used for reweighting the observations (`class_weight`) is also tested.

According to Breiman (2001), random forests are resistant to overfitting, and the number of base learners (`n_estimators`) can therefore be quite large. Van Rijn & Hutter (2017, 2018) find that restricting the growth of the trees using the `min_samples_leaf` parameter is more efficient than using `min_samples_split`; they also report that the maximum number of features to consider when splitting a node (`max_features`) has a major effect on performance. The parameters `min_samples_leaf` and `max_features` are chosen as the tuning parameters for the RF model. As with the individual decision tree, `class_weight` is also used.

For AdaBoost, the maximum depth of the individual trees (max_depth) and learning rate (learning_rate), which reduces the contribution of individual trees to avoid overfitting, are the key parameters; the number of iterations appears less important, and 50 iterations were sufficient in the empirical study by Van Rijn & Hutter (2017, 2018). Gradient boosting is very similar to AdaBoost in terms of the basic mechanism, and the same two aforementioned hyperparameters are therefore chosen for tuning for both models. Additionally, Friedman's (2002) suggested improvement to his original GB algorithm, stochastic gradient boosting, is taken into account. Stochastic gradient boosting trains each base learner on a bootstrap sample similarly to bagging; the size of the bootstrap sample as a fraction of total training set size (subsample) is therefore included. Class weights cannot be adjusted for the boosting methods, because it would interfere with the boosting algorithm.

For the logistic regression model, the main adjustable factor is regularization, which penalises model complexity. Adjustments are made by changing the value of the regularization parameter C. Additionally, two different penalty functions (penalty), ℓ_1 (lasso) and ℓ_2 (ridge regression), are tested. Similarly to decision tree and random forest models, class_weight is included.

In the first modeling phase, hyperparameter tuning is performed on a relatively narrow selection of different values (see Table C1) for practical reasons. The task of finding the best hyperparameters for each combination of classifier, feature selection method and training set composition is computationally demanding; moreover, it can be expected that a rough adjustment of the key parameters is enough to show the performance differences between the trained models.

For the finer tuning of hyperparameters in the second phase, observations from the first round are used as guidelines. For gradient boosting, a low learning rate (learning_rate$= 0.1$) appears to be a key performance factor. Although the interactions between hyperparameters can be complex, it seems safe to assume that low learning rates work better, and therefore the search in the second phase is concentrated on small values. The best combinations also use so-called "decision stumps", i.e. decision trees with only one split from the root node, corresponding to max_depth$= 1$; in the second phase, the search is focused on small values of max_depth. For random forests, min_leaf_size appears to be the most critical factor, with both 4 and 10 emerging as viable options. Minimum leaf size 1 is not feasible, most likely due to overfitting, because the trees can become arbitrarily complex, and the tested values are therefore adjusted upwards for the second round. Unadjusted class weights are used in all of the best parameter combinations for RF models trained with the full predictor set, and also for most RF models using feature selection. Therefore, the class_weight parameter is dropped altogether, and unadjusted weights are used in the second phase of modeling.

Additional hyperparameters are also considered for both methods in the second phase of modeling. The number of base learners n_estimators is added for both methods to ensure that relying on default values does not unnecessarily weaken model performance. For the same reason, the splitting criterion (criterion), which measures the goodness of potential node splits, is added for the RF method. To provide an alternative to minimum leaf size, the `max_depth` parameter is also used for RF; conversely, `min_leaf_size` is included for GB to complement the maximum tree depth parameter.

Tuning results from the second modeling phase, particularly those for the random forest method, may be of some interest for future studies. Entropy is consistently superior to the Gini coefficient, which is the default option for split criterion in the scikit-learn implementation: it is used in every one of the three best combinations for each RF model. Minimum leaf size 6 is chosen as the best value for each RF model; this somewhat contradicts the findings of Van Rijn & Hutter (2017, 2018), who report that minimum leaf size is the most important parameter and that small values (starting from 1) are preferable. Additionally, three of the four RF models (full data, 1:10, 1:3) benefit from the largest available number (250) of trees. Tuning results for GB show that a learning rate of 0.05 is preferred by all models; it seems possible that even lower values could have been beneficial. The GB classifiers trained on balanced data function best without applying stochastic gradient boosting (i.e. `subsample`= 1.0) and 100 base learners, while the model using the full training data set performs optimally with `subsample`= 0.75 and 250 base learners.

4.5 Model evaluation metrics

4.5.1 Threshold-dependent metrics

Despite the shortcomings discussed in the literature review, performance metrics derived from the confusion matrix as described in Table 5 can be useful. They are intuitive and easy to understand, and can be used to illustrate the differences between the behavior of the trained models.

Table 5: Confusion matrix (Kohavi & Provost, 1998)

	Predicted class	
Actual class	Negative (non-bankrupt)	Positive (bankrupt)
Negative (non-bankrupt)	True negative (TN)	False positive (FP)
Positive (bankrupt)	False negative (FN)	True positive (TP)

The true positive rate (recall), true negative rate (specificity) and positive predictive value (precision) are calculated for each model to provide a simple overview of binary classification performance. Additionally, the F_β score and Matthews correlation coefficient are used. The metrics and their formulas are presented in Table 6.

The F_β score, derived from Van Rijsbergen's (1974) effectiveness measure, is an indicator of classification performance calculated using precision and recall; more specifically, it is their weighted harmonic mean. The parameter β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" (ibid., p.371): $\beta > 1$ emphasizes recall, while $\beta < 1$ prioritizes precision.

The usefulness of the F_β score lies in the ability to easily measure classification performance according to the user's priorities and assumptions about misclassification costs. The typical β values are 1 for a neutral approach, 0.5 for emphasizing precision and 2 for emphasizing recall (Saito & Rehmsmeier, 2015), although the appropriate value depends strongly on the context. In the case of bankruptcy prediction, undetected bankruptcies are much more costly than healthy companies falsely classified as bankrupt; a value of $\beta = 3$ is therefore chosen. This is still a very conservative estimate: for example, Serrano-Cinca et al. (2019) estimate the different cost of false negatives and false positives based on prior literature and use $\beta = 35$. For comparison, the F_1 score (i.e. precision and recall considered equally important) is also calculated.

Table 6: Threshold-dependent performance measures

Recall (true positive rate)	$\frac{\text{TP}}{\text{TP+FN}}$
Specificity (true negative rate)	$\frac{\text{TN}}{\text{TN+FP}}$
Precision	$\frac{\text{TP}}{\text{TP+FP}}$
F_β	$(1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}$
Matthews correlation coefficient	$\frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}$

Boughorbel et al. (2017) note that F_β, although adjustable to specific needs, is sensitive to class imbalance, and suggest that the Matthews correlation coefficient (Matthews, 1975) (MCC) is a better alternative. Unlike measures such as accuracy and F_β, MCC takes into account all the different prediction outcomes (true positive, true negative, false positive, false negative) and thus provides a better summary of the performance of the classifier. MCC values range from -1 to 1, with 1 indicating a perfect classifier and -1 a classifier that predicts wrong on every instance; a MCC of 0.6 or higher is usually considered good performance (Bao et al., 2019).

The class imbalance in the data is likely to impact the prediction models significantly; they cannot be expected to perform optimally using the default classification threshold 0.5. Results between models trained on balanced and imbalanced data would not be comparable, and therefore different classification thresholds are used for each model. This can be done by using the predicted probabilities of the model and finding a cutoff threshold that produces the desired result. In this study, the main subject of interest is the ability to predict actual bankruptcies correctly, i.e. recall. Therefore, specificity is fixed at 95% for the classifiers in this study; this helps compare the performance of the models in terms of correct classification of bankrupt companies, when each model predicts non-bankrupt companies equally well. In other words, the thresholds are set so that each of the models produces a false alert for exactly 5% of non-bankrupt companies.

4.5.2 Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve (see e.g. Fawcett, 2006) is a widely used method for evaluating classifiers. The curve visualizes the trade-off between recall, or true positive rate (TPR), and false positive rate (FPR). Classification results and the corresponding (TPR, FPR) values are calculated at different classification thresholds. The values are plotted, typically with FPR in the x axis and TPR in the y axis. An example ROC curve is shown in Figure 2.

The ROC curve can be quantified by approximating the area under the curve (AUC), i.e. the area between the ROC curve and the x axis in $[0, 1]$. A line from (0,0) to (1,1), with an AUC of 0.5, represents a random guess; AUC$=$ 1 corresponds to a perfect classifier. The AUC of a prediction model is equivalent to the probability that it ranks a random instance from the positive class higher than a random instance from the negative class. It is thus equivalent to the Mann-Whitney U statistic (Mann & Whitney, 1947), and also relates closely to a number of other statistical measures.

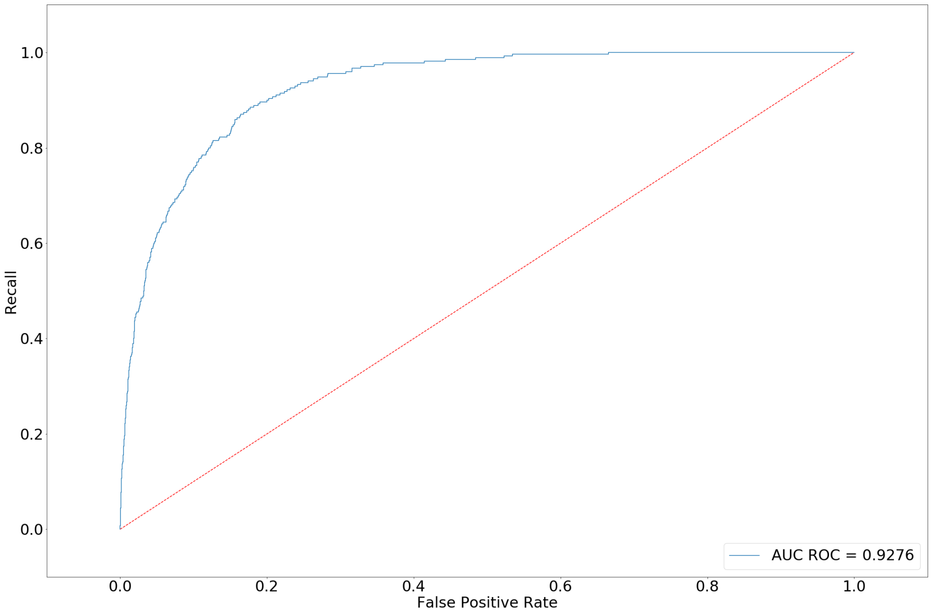

Figure 2: Receiver operating characteristic (ROC) curve

4.5.3 Precision-recall curve

Despite its widespread use in evaluating bankruptcy prediction models, the ROC curve may not be an optimal tool when significant class imbalance is present in the data; the precision-recall (PR) curve (Figure 3) may be a preferable alternative (Davis & Goadrich, 2006). The process of plotting the PR curve is similar to the ROC curve, with the exception that precision is used instead of FPR. An additional difference is that the PR curve usually shows recall (TPR) in the x axis rather than the y axis. Similarly to ROC, the PR curve can be quantified by approximating the area under the curve. Possible values range from 0 to 1, with PR AUC 1 corresponding to a perfect classifier.

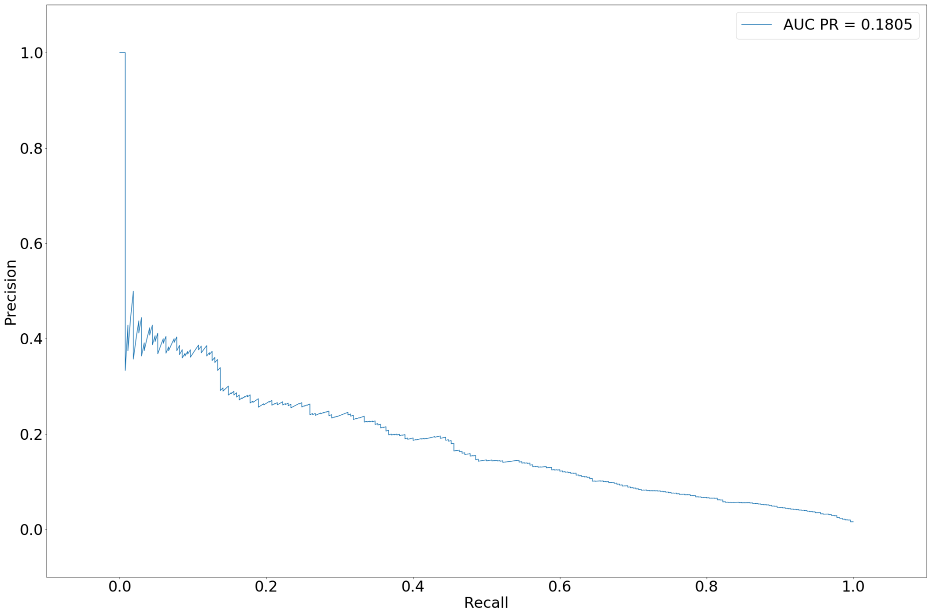

Figure 3: Precision-recall (PR) curve

The usefulness of the PR curve in imbalanced classification stems from the difference between FPR (and specificity, i.e. $1-\text{FPR}$) and precision. For example, in a scenario where class imbalance is high, a prediction model might produce output $\text{TN} = 1000$, $\text{FP} = 100$, $\text{FN} = 5$, $\text{TP} = 20$, where recall $= 20/25 = 0.8$, specificity $= 1000/1100 \approx 0.91$ and precision $= 20/140 \approx 0.14$. In terms of recall and specificity, performance seems good, but low precision reveals that a predicted positive is likelier to be a false alert than actual positive. Precision is much more sensitive to changes in the number of false positives, and therefore the PR curve is more informative than the ROC curve for imbalanced data (Saito & Rehmsmeier, 2015).

Because linear interpolation can produce overly optimistic results (Davis & Goadrich, 2006), the area under the PR curve (PR AUC) is approximated in this study using average precision (AP, `average_precision_score` in the scikit-learn package). It is calculated as

$$\text{PR AUC} \approx \text{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{4}$$

where P is precision and R recall at classification threshold n. Due to its better suitability to imbalanced learning, PR AUC is used as the performance criterion when optimizing hyperparameter values.

4.5.4 Predictor variable importance

To better understand the key determinants of bankruptcy, the importance of individual predictor variables is analyzed. For the logit model, the impact of predictor variables can be observed using the coefficients of the model. The magnitude of the coefficient indicates how strongly a single predictor affects the output of the model; additionally, the sign of the coefficient shows whether the predictor is directly or inversely related to bankruptcy risk.

For the other, decision tree-based models, the relative variable importance (RVI) measure (Breiman et al., 1984) is used. RVI evaluates individual nodes using the Gini importance, which measures how much the node reduces impurity compared to its child nodes, weighted by the number of observations n_t in the relevant nodes. The weighted Gini importance (I) is defined in terms of Gini impurity G:

$$G = p_0(1-p_0) + p_1(1-p_1) = 2p_0p_1, \quad I = \frac{N_p}{N}(G_p - \frac{N_{c1}}{N_p}G_{c1} - \frac{N_{c2}}{N_p}G_{c2}) \tag{5}$$

where p_0, p_1 are the proportions of instances belonging to classes 0 (non-bankrupt) and 1 (bankrupt) in a node, and N, N_p, N_{c1}, N_{c2} are the total sample size, number of instances in the parent node, and number of instances in each child node, respectively. To calculate the relative variable importance for predictor X in decision tree T, the Gini importances at each node t using X for splitting are summed:

$$\text{RVI}_X(T) = \sum_{t \in T} I_t \, \mathbf{1}(v(t) = X) \tag{6}$$

where $v(t)$ is the variable that is used for splitting node t. RVI scores are typically scaled for better interpretability; the scikit-learn implementation used in this study scales the values to sum to 1. The RVI scores from each individual tree are averaged to calculate the final scores for ensemble methods.

4.5.5 Benchmarking against original Valuboost model

To assess the usefulness of the results from a practical perspective, the results of the boosting-based bankruptcy prediction model previously used by Valuatum ("Valuboost") are used as a benchmark. PR AUC (0.19) is used as the primary measure for comparison, supplemented by ROC AUC (0.91). The results of the Valuboost model are obtained using some predictors that are omitted in this study (see Section 3.5.3) and therefore not entirely comparable; nonetheless, measuring the difference to previous results can give some indication of the performance of the models. The results of the Valuboost model are averaged over several runs with different training and test sets; the splits to training and test data are different from those used in this study.

5 Results

5.1 Predictive performance

5.1.1 First modeling phase

The results of the first phase of modeling presented in Table 7 and Table 8 show clear differences in the performance of the classifiers. The individual decision tree model (DT) is superior to logit: its PR AUC score is higher for every combination of balancing and feature selection, and ROC AUC score higher for four of the six cases. The performance difference is more pronounced in models trained on balanced data. Ensemble-based machine learning methods consistently outperform both of the standalone classifiers. The highest performing model in terms of ROC AUC is gradient boosting (GB) trained on balanced data using all available features (0.912). PR AUC scores, on the other hand, indicate that the random forest (RF) using all features and trained on the full, imbalanced training set performs best (0.190). The results are in line with recent findings showing the general effectiveness of ensemble methods (Veganzones & Séverin, 2020); more specifically, the predictive power of boosting methods and random forest have been demonstrated e.g. by Barboza et al. (2017), Jones et al. (2017), and Son et al. (2019). The performance of individual decision tree models is consistently weaker than that of the ensembles, but seems to behave in a similar manner in relation to the class distribution of the training data and the variable selection method. This result is quite expected and supports the notion that ensembling improves the predictive capacity of decision tree models (Hastie et al., 2009).

None of the ensemble models can outright be deemed superior to the others; comparing the results for all six combinations of class distribution and predictor set reveals no obvious patterns. In most cases the differences between ensembles are not particularly large; the most significant variation is observed in the PR AUC scores of the models trained on balanced data using either mutual information (MI) or recursive feature elimination (RFE). It could be speculated that this is mainly attributable to randomness, as these models have both a small training set and few predictors. ROC AUC scores are notably even with all combinations; the best performance is achieved with balanced data and full feature set.

The threshold-dependent metrics (recall, precision, F_1, F_3, MCC) indicate that, when adjusted to 95% specificity, the differences in classification performance are relatively small. Imbalanced data do not seem to produce inherently inferior models; in fact, only the models trained on the MI-selected predictors perform better on balanced data, while full-feature models and those with RFE selection appear slightly better with the original, imbalanced data.

In terms of PR AUC scores, all ensembles with the single exception of RF with MI features perform better on the imbalanced training set. Balanced data produce better ROC AUC scores for the ensembles when all predictors are used; with feature

Table 7: Model performance - full data

	ROC AUC	PR AUC	Recall	Precision	F_1	F_3	MCC
Logit	0.723	0.034	0.240	0.048	0.080	0.171	0.087
DT	0.691	0.072	0.415	0.077	0.130	0.288	0.160
RF	0.896	0.190	0.589	0.109	0.183	0.408	0.237
AdaBoost	0.901	0.173	0.605	0.111	0.187	0.418	0.243
GB	0.902	0.184	0.597	0.110	0.186	0.414	0.241
Logit	0.746	0.042	0.295	0.058	0.096	0.209	0.110
DT	0.735	0.081	0.422	0.085	0.141	0.302	0.172
RF	0.893	0.144	0.550	0.102	0.173	0.383	0.221
AdaBoost	0.900	0.139	0.558	0.104	0.175	0.388	0.224
GB	0.900	0.146	0.562	0.104	0.176	0.390	0.225
Logit	0.726	0.036	0.248	0.048	0.081	0.176	0.089
DT	0.771	0.087	0.450	0.080	0.135	0.307	0.171
RF	0.900	0.180	0.597	0.110	0.185	0.413	0.240
AdaBoost	0.896	0.161	0.581	0.105	0.179	0.401	0.231
GB	0.901	0.170	0.601	0.110	0.187	0.416	0.242

All performance metrics range from 0 to 1, except MCC, which ranges from -1 to 1; a score of 1 indicates a perfect classifier for each metric. For further explanations of the performance measures, see Section 4.5.

selection methods, the effects of class imbalance appear mixed. Logit shows slight overall improvement when balanced data are used; standalone DT achieves higher ROC AUC scores but lower PR AUC scores with balanced data. It must be noted that, due to substantial class imbalance in the original data, the balanced training set is significantly smaller; this may have an impact on predictive performance, even though sample size remains larger than in most recent studies.

The results of the first phase give no conclusive evidence to support the superiority of either balanced or imbalanced training data. If the aim were solely to create a binary classifier that performs well with the default classification threshold 0.5, balanced data would be required. However, in most contexts, including this study, it is more important to obtain reliable estimates of bankruptcy probability; the classification threshold can be adjusted as needed.

Of the two feature selection methods used, recursive feature elimination (RFE) mostly performs better than mutual information (MI). The predictors selected with RFE give equal or better results than predictors selected with MI; the sole exception is the AdaBoost model trained on imbalanced data, and even in this case the difference is negligibly small. This finding is in line with the literature, which mostly agrees that wrapper methods tend to yield better classification performance than filter methods (Liang et al., 2015; Peng et al., 2005). Since RFE selects the most relevant features for each classifier, it seems reasonable that it should

Table 8: Model performance - 1:1 balanced data

	ROC AUC	PR AUC	Recall	Precision	F_1	F_3	MCC
Logit	0.728	0.039	0.248	0.049	0.082	0.176	0.090
DT	0.809	0.043	0.539	0.054	0.098	0.283	0.146
RF	0.907	0.161	0.570	0.107	0.180	0.398	0.231
AdaBoost	0.904	0.152	0.562	0.104	0.176	0.390	0.226
GB	0.912	0.145	0.585	0.109	0.184	0.408	0.237
Logit	0.735	0.044	0.306	0.059	0.099	0.216	0.115
DT	0.843	0.057	0.504	0.074	0.129	0.319	0.174
RF	0.901	0.150	0.570	0.107	0.180	0.398	0.231
AdaBoost	0.901	0.133	0.574	0.106	0.179	0.398	0.231
GB	0.899	0.115	0.531	0.099	0.167	0.370	0.212
Logit	0.726	0.039	0.252	0.050	0.083	0.179	0.092
DT	0.831	0.049	0.457	0.063	0.111	0.281	0.149
RF	0.902	0.163	0.578	0.107	0.181	0.401	0.232
AdaBoost	0.904	0.152	0.562	0.104	0.176	0.390	0.226
GB	0.899	0.126	0.562	0.103	0.174	0.389	0.224

All performance metrics range from 0 to 1, except MCC, which ranges from -1 to 1; a score of 1 indicates a perfect classifier for each metric. For further explanations of the performance measures, see Section 4.5.

outperform MI, which does not tailor its selection to suit a specific model. However, due to the complexity of the interactions between predictors, it is not guaranteed that the most relevant features give optimal predictive performance out of sample.

The selections made by the MI and RFE feature selection methods are examined by comparing them to the 25 features that are observed as the most important in the models trained on all predictors. All of the feature selection results are obtained from the models using the full, imbalanced training sample. Since the ensemble models provide the best predictive performance, they are used for evaluating feature selection; standalone DT and logit models are notably weaker, and therefore their use of the different predictors can be considered less relevant. The findings are summarized in Table 9.

Out of the predictors ranked in each full-feature model's top 25, RFE selects 14 for RF, 12 for AdaBoost and 15 for GB. Compared to the variable rankings of the full-feature models, the average rank of the RFE-selected variables is 28 for RF, 44 for AdaBoost and 35 for GB. The MI method selects 10 for RF, 7 for AdaBoost and 11 for GB out of the top 25 of the full-feature models, and the average rank of the MI-selected variables in the variable importance ranking of the full-feature models is 72 for RF, 73 for AdaBoost, and 65 for GB. These results show that the RFE method fares notably better than MI in selecting those features that are found the most relevant in the trained models that use all predictors. It seems that the MI method drops more useful features and therefore leads to inferior predictive performance.

Table 9: Feature selection - no. of shared features between methods

	RF	AdaBoost	GB
All top 25 & MI	10	7	11
All top 25 & RFE	14	12	15
MI & RFE	11	8	9
All top 25 & MI & RFE	8	3	5

Furthermore, it is found that 8 predictors for RF, 3 for AdaBoost and 5 for GB are shared by all of the respective models' three subsets of 25 predictors (MI-selected, RFE-selected, best predictors from model using full feature space). The top predictors of the three ensembles trained using the full feature space are additionally compared to each other. RF and AdaBoost share 9, RF and GB 15, and AdaBoost and GB 14 of the 25 most important predictors; 6 predictors are shared by all three classifiers, with the latest year's (Y-1) equity to liabilities (TE/TL) and financial expenses to total assets (FE/TA) occupying the first and second places respectively for each ensemble method. Only a single predictor, TE/TL Y-1, is shared by all 18 ensemble-feature selection-training data combinations.

The most notable finding with regard to feature selection is that no predictors, perhaps apart from TE/TL Y-1, can be deemed categorically more important than others. There is significant variation in the predictors selected by the feature selection methods and those that emerge as the most important from the models trained on all predictors. Additionally, different predictors are relevant for the three ensemble classifiers. Therefore, it is very difficult to choose a subset of predictors that can be assumed to perform consistently well out of sample. Jones (2017) suggests that studying bankruptcy in a high-dimensional context is likely to capture meaningful predictor interactions that are overlooked if the feature space is too narrow; because of the contradictory outcomes of the feature selection methods, it can be assumed that is also the case in this study. The predictive power of specific variables varies between contexts (Balcaen & Ooghe, 2006; du Jardin, 2015), and therefore the features selected by MI and RFE in this study may not be relevant elsewhere; using the full feature space produces a more robust model. For these reasons, feature selection methods are not applied in the second phase of modeling.

As feature selection methods are not applied further, predictive performance using all predictors is the main criterion for choosing the classifiers to use in the second phase. Results on imbalanced data (Table 7) indicate that RF and GB outperform AdaBoost. GB is the superior model by a narrow margin on balanced data (Table 8), while AdaBoost and RF perform more or less equally well. AdaBoost is excluded, because it is the weakest ensemble overall based on predictive performance, and does not appear superior to the other two ensembles with any combination of feature selection and training data balance. Logit and standalone DT are also dropped due to clearly inferior performance. Thus, RF and GB are used in the second phase of modeling.

As the results on balanced and imbalanced data show no notable differences, class imbalance is explored further in the second modeling phase. As Veganzones & Séverin (2018) show, it is not necessary to balance the training data fully to 1:1 proportions; an imbalance may not impede performance if it is not too severe. To further assess the effects of class imbalance, the full imbalanced and 1:1 balanced training sets are again used in the second modeling phase. Additionally, the models are trained using 1:3 and 1:10 balanced sets; the proportions are chosen from both sides of the 1:4 limit suggested by Veganzones & Séverin (ibid.).

5.1.2 Second modeling phase

For the second phase, the two best prediction methods (RF and GB) are chosen for further analysis. Models are trained on the original, imbalanced training data, as well as balanced sets with proportions 1:1, 1:3, and 1:10. All models are trained using the full feature space; no feature selection methods are applied. Hyperparameters are tuned with a larger number of options, but otherwise similarly to the first phase: the data are split into training, validation and test sets (60%, 20%, and 20% of the total sample), the models are trained on the training set, and the validation set is used to assess performance with different parameter combinations. The test set remains unused during parameter tuning; this ensures that the sample size for parameter tuning and classification are similar, i.e. the training set, before possible random undersampling, comprises 60% of the total sample. The final classification results and variable importances are averaged over 10 runs; models are trained on the full sample, excluding the validation set used for hyperparameter tuning, in order to avoid data leakage (Kaufman et al., 2011). For each run, the data are divided into training and test sets by random splits of 75% - 25% (corresponding to 60% and 20% of the entire sample), with a different random number seed used for each run.

The results of the second modeling phase (Table 10) show a marked improvement in overall predictive performance. Both methods perform better than in the first phase on both imbalanced and fully balanced data: detailed hyperparameter tuning has a visibly positive effect, although it is difficult to assess which parameters have the greatest effect on performance. Similarly to the first phase, it can be observed that balancing the training data may slightly improve ROC AUC, while lowering PR AUC at the same time. The threshold-dependent metrics do not provide much additional information; the results are very similar for both models at all class imbalance levels, with the exception of the fully balanced training set.

A clear deterioration in all performance metrics is observed between the models trained on 1:3 and 1:1 training sets. Balancing to 1:10 or 1:3 yields a higher ROC

Table 10: Model performance - best models with fine-tuned parameters

	ROC AUC	PR AUC	Recall	Precision	F1	F3	MCC
RF full data	0.915	0.195	0.601	0.120	0.200	0.429	0.252
GB full data	0.914	0.180	0.602	0.121	0.201	0.430	0.253
RF 1:10	0.919	0.192	0.605	0.121	0.202	0.432	0.254
GB 1:10	0.915	0.179	0.593	0.119	0.199	0.424	0.249
RF 1:3	0.918	0.177	0.593	0.119	0.198	0.424	0.249
GB 1:3	0.917	0.173	0.597	0.120	0.200	0.427	0.251
RF 1:1	0.909	0.161	0.560	0.113	0.188	0.401	0.235
GB 1:1	0.913	0.157	0.576	0.116	0.193	0.413	0.242

All performance metrics range from 0 to 1, except MCC, which ranges from -1 to 1; a score of 1 indicates a perfect classifier for each metric. For further explanations of the performance measures, see Section 4.5.

AUC and similar MCC and F_β scores compared to the full imbalanced training sample; therefore it seems likely that the performance drop for the 1:1 data is due to training set depletion rather than different class distribution. The number of observations remaining in the fully balanced set is too small to achieve the performance of the models trained on larger samples. Unlike Zhou (2013) suggests, random undersampling may not be the best option despite the large initial sample size; alternatives that do not overly reduce training set size could perform better.

In terms of both PR AUC and ROC AUC, the RF classifier slightly outperforms GB, with the exception of ROC AUC on the 1:1 balanced training set. The performance difference is larger on the full and 1:10 balanced training sets; on the fully balanced 1:1 training set, GB outperforms RF in terms of all metrics except PR AUC. In general, GB appears less sensitive to different class distributions. It cannot be told with certainty whether this observation is truly due to class imbalance or not. As Breiman (2001) argues that RF is practically immune to overfitting, it could be assumed that the smaller training set size that accompanies closer class balance is the key factor: if the performance of the RF model improves or at least does not deteriorate when arbitrarily large numbers of training observations are added, removing part of the observations should result in weaker performance. GB, on the other hand, may overfit to the training data; while balancing the data removes some relevant information, it also reduces overfitting and therefore the change in performance is less prominent for GB. Hyperparameter tuning results support this assumption: the GB models consistently picked the smallest possible learning rate (0.05); it is possible that an even lower value would be needed to avoid overfitting.

Given the improved ROC AUC on 1:10 and 1:3 balanced data, as well as the weaker performance of models on 1:1 balanced data that is presumably due to insufficient data, the results corroborate the assertion of Veganzones & Séverin (2020) that a combination of balanced data and large sample size gives the best results. However, this view is based on the assumption that ROC AUC and threshold-dependent metrics are sufficient for measuring performance in corporate failure prediction. In highly imbalanced datasets, PR AUC is a more relevant measure, as it captures the performance impact of the imbalance better than ROC AUC (Davis & Goadrich, 2006). Examining PR AUC scores shows that the full training set gives the best results, with the 1:10 balanced set producing only slightly weaker results; the 1:3 and 1:1 training sets perform notably worse. The full training set has proportions of approximately 1:99; it is possible that optimal performance could be achieved somewhere between 1:99 and 1:10 imbalance. However, this cannot be reliably assessed without further studies, and therefore the conclusion based on PR AUC is that no resampling should be used: the original training set with imbalance corresponding to the real-world situation gives the best results.

Compared to the results of prior studies, it can be said that the ROC AUC scores achieved in this thesis indicate good performance. Although studies on different data are not truly comparable, most recent studies seem to reach ROC AUC scores from approximately 0.8 to somewhere upwards of 0.9 (see e.g. Jones et al., 2017; Son et al., 2019; Veganzones & Séverin, 2018; Zhou & Lai, 2017). The PR AUC scores of the models in this study do seem quite low; no useful comparisons to literature can be made, but it seems that the rarity of bankruptcies makes the prediction task very difficult in general.

Most of the models do not reach the performance level of the previous model used by Valuatum. In terms of ROC AUC, the results of all models are comparable or slightly superior to the benchmark score (0.91), but PR AUC of 0.19 is only reached by the RF model trained on the full or 1:10 balanced training set. Compared to using the unadjusted training set, 1:10 rebalancing yields an acceptable trade-off of higher ROC AUC with only a minor PR AUC deterioration. However, without further study there is not enough evidence to support the use of resampling. Therefore, the random forest model trained on the full training set is proposed as the best choice for practical application.

5.2 Predictor variable importances

The importance of specific variables and predictor categories is assessed using the best model, namely RF with fine-tuned hyperparameters trained on the full training set. To add an alternative perspective, the GB model, also with optimized hyperparameters and trained on the unadjusted data set, is used. The variable importances of the models using balanced data were briefly examined, but no notable differences to the models using the full training set were observed.

Similarly to the first phase, the 25 best predictors are examined as the main subject of interest; some observations from outside the top 25 rankings are also presented. The best predictors for the RF and GB models are presented graphically in Figure 4; Table 11 lists the RVI scores of the best predictors for each model individually, as well as the top 25 of the two models based on the predictors' average ranking. Average RVI is not used, because the scores of the RF and GB models are dissimilarly distributed, and the average RVI can therefore be uninformative and misleading.

9 out of 25 and 3 out of 10 of the most important predictors are shared by the models. The considerable variation supports the observations made in the first phase: there are no definitive answers as to which predictors are the most relevant. However, the similarities of the models' variable rankings can give some insights into the effect of different predictors. The common top 10 predictors are equity to liabilities (TE/TL), financial expenses to total assets (FE/TA) and financial expenses to EBITDA (FE/EBITDA), all from the latest financial year (Y-1). TE/TL Y-1 is ranked first and FE/TA Y-1 second in both models' relative variable importance ranking; this supports the similar findings for the three ensemble models in the first phase.

Despite sharing the best two predictors, significant differences are observed between the variable importances of the RF and GB models. For RF, 10 out of 25 best predictors are capital structure ratios, while for GB the number is only 2 out of 25. The top 25 ranking of RF includes fewer activity ratios than that of GB, but on the other hand more liquidity-related predictors. No growth variables, size proxies or industry variables are in RF's top 25, while for GB the list includes sales for all three years, as well as two growth variables and the industry risk variable for the latest year (Y-1). Both models have some profitability ratios ranked among the best 25, but none of these are shared by both. In addition to the different lists of best predictors, Figure 4 illustrates a major contrast between the models: the relative variable importances of GB are more unevenly distributed than those of RF. The variable importance scores are therefore not directly comparable between the two models; the fundamental difference in the design of the classification algorithms causes them to utilize the predictors in different manners.

For the RF model, both TE/TL and TE/TA (equity to assets) in each year (Y-3, Y-2, Y-1) are ranked in the top 25, the lowest being TE/TL Y-3 at rank 19. In addition to these, there are 4 other equity-related capital structure ratios among the best 25 predictors. On the other hand, the GB model's only capital structure predictors in the top 25 are TE/TL Y-1 and net debt to equity ((TD-C)/TE) Y-1. Total equity, and particularly TE/TL, undeniably has significant predictive power. Due to the differences between the variable importances of RF and GB, it is difficult to make any further conclusions regarding equity or other capital structure ratios.

Profitability ratios perform quite well for both models. They are slightly more prominent in the top predictors of GB (4 in total) and occupy places 6, 10, 17, and 21. For RF, the top 25 includes 3 profitability ratios, ranked 13th, 14th, and 18th. None of the predictors is shared by the two models. Additionally, it should be noted that both models have a large number of profitability ratios in relatively high rankings, approximately from 25 to 60. Altogether, it appears that profitability is, as suggested by the literature, a key determinant of bankruptcy. However, no single variable can be identified as particularly useful. Given the wide variety of business

Table 11: Relative variable importances

RF		GB		avg. rank RF & GB	
TE/TL Y-1	0.0189	TE/TL Y-1	0.1206	TE/TL Y-1	1
FE/TA Y-1	0.0154	FE/TA Y-1	0.0494	FE/TA Y-1	2
TE/TA Y-1	0.0136	AP/COGS Y-1	0.0395	FE/EBITDA Y-1	5
CL/TA Y-1	0.0131	**EE/PBD Y-1	0.0369	**EE/PBD Y-1	7.5
FE/EBITDA Y-1	0.0130	FE/EBITDA Y-1	0.0245	AP/COGS Y-1	9.5
TE/TL Y-2	0.0126	*(EBIT+FI)/TC Y-1	0.0203	FE/TA Y-2	9.5
FE/TA Y-2	0.0117	*(TL-TD)/S Y-1	0.0201	*(TD-C)/TE Y-1	10.5
FE/NI Y-1	0.0114	EE/VA Y-1	0.0200	CL/TA Y-1	11
TE/TA Y-2	0.0105	S Y-1	0.0196	FE/NI Y-1	13.5
*(TD-C)/TE Y-1	0.0093	GP/S Y-3	0.0188	*(TL-TD)/S Y-1	19
**EE/PBD Y-1	0.0089	*(TD-C)/TE Y-1	0.0184	*(EBIT+FI)/TC Y-1	20
TD/TE Y-1	0.0088	FE/TA Y-2	0.0183	*PBD/S Y-1	20.5
NI/TA Y-1	0.0083	*ind_risk Y-1	0.0168	EE/VA Y-1	22
*PBD/S Y-1	0.0076	**cagr AP/S	0.0129	NI/TE Y-1	23.5
FE/NI Y-2	0.0075	S Y-3	0.0128	TD/TE Y-1	25
AP/COGS Y-1	0.0074	AP/S Y-1	0.0121	FE/NI Y-2	25.5
TE/TA Y-3	0.0071	NI/TE Y-1	0.0117	*(TD-C)/TE Y-2	26.5
NI/S Y-1	0.0071	CL/TA Y-1	0.0102	AP/S Y-1	28.5
TE/TL Y-3	0.0071	FE/NI Y-1	0.0099	S Y-1	29.5
*(TD-C)/TE Y-2	0.0071	S Y-2	0.0094	SC/TC Y-1	31
(C+MS)/CL Y-1	0.0070	GP/S Y-1	0.0089	*ind_risk Y-1	32
SC/TC Y-1	0.0070	FE/EBITDA Y-2	0.0079	AP/COGS Y-2	32.5
*TD/PBD Y-1	0.0069	AP/COGS Y-2	0.0072	NI/TA Y-1	33
C/CL Y-1	0.0069	growth OCF Y-2	0.0070	FE/EBITDA Y-2	34
**EE/NI Y-2	0.0067	IA/TA Y-1	0.0070	EBIT/TA Y-1	35

*: variables from previous Valuatum model (Table 3)
**: additional predictor variables (Table 4)
Variables with no additional markings are from previous studies (Table 2).

models and cost structures, it is perhaps to be expected that diverse profitability variables are needed in order to create a model that works well across industries.

Activity variables are found effective, particularly for the GB model, for which they constitute 3 of the top 10 predictors. These are accounts payable to cost of goods sold (AP/COGS) at 3rd, employee expenses to profit before depreciation, amortization and extraordinaries (EE/PBD) at 4th, and employee expenses to value added (EE/VA) at 8th, each measured for the latest year Y-1. Also in the top 25 are accounts payable to sales (AP/S) Y-1 at 16th, and AP/COGS Y-2 at 23rd. For RF, activity ratios are somewhat less relevant, the best being EE/PBD Y-1 at 11th, followed by AP/COGS Y-1 (16th) and employee expenses to net income (EE/NI) Y-2 (25th). These rankings indicate that two operational aspects, accounts

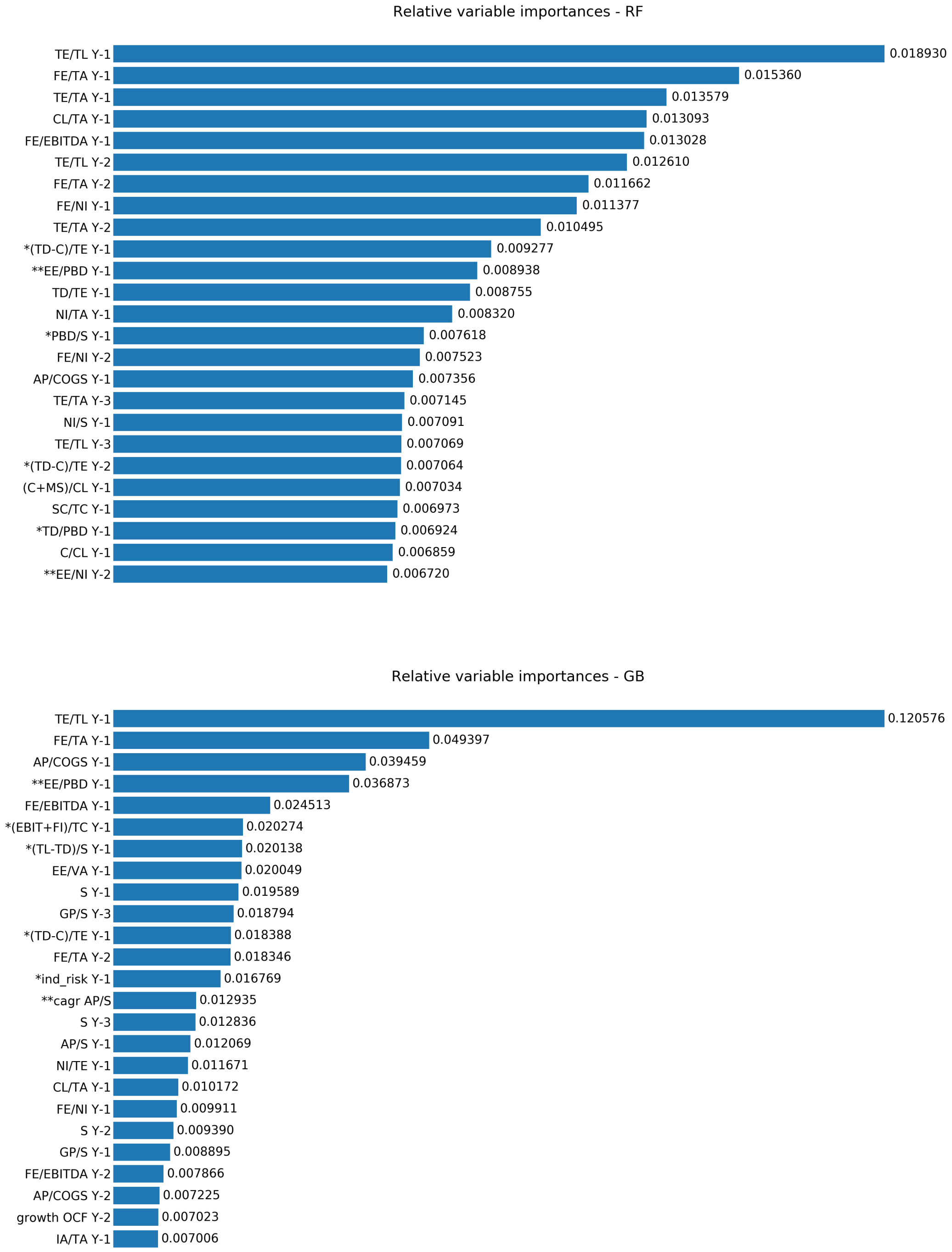

Figure 4: Relative variable importances - RF and GB

payable turnover and employee efficiency, are important bankruptcy determinants. Receivables and inventory turnover ratios are mostly insignificant for either model, and coarser measurements such as sales to total assets (S/TA) and (net) working capital to sales (WC/S, NWC/S) also perform quite weakly. Some predictors measuring cost structure, such as EBIT to value added (EBIT/VA) and net income to value added (NI/VA) are moderately relevant to both models. Contrary to Jones et al. (2017), capital expenditure does not appear particularly useful for either model.

Solvency, and in particular financial expenses, are important for both prediction models. In the RF model, the ratio to total assets (FE/TA) and to net income (FE/NI) for years Y-1 and Y-2, as well as the aforementioned FE/EBITDA Y-1, are in the top 25, with FE/NI Y-2 ranking lowest at 15. In addition to these solvency indicators, total debt to profit before depreciation, amortization and extraordinaries (TD/PBD Y-1) is included. For the GB model, FE/TA and FE/EBITDA for Y-1 and Y-2, as well as FE/NI Y-1, are among the best 25 predictors. An interesting observation regarding financial expenses is that similar predictors using only interest expenses (e.g. EBIT/IE, NI/IE) are of very little use; RF ranks all of them among the 20 worst predictors, while for GB the highest-ranking interest expense predictor is NI/IE Y-1 at 94th. Although total financial expenses (FE) includes interest expenses (IE) and therefore holds some of the same information, it does not seem reasonable that this would cause IE to be irrelevant. It seems probable that interest expenses are saved in the database under another variable, such as other financial expenses. Interest expenses are unlikely to be missing altogether: in this case, there would be a discrepancy between profit in the income statement and reported profit on the balance sheet, and such errors are checked against in the Valuatum system. Moreover, descriptive statistics (Table B1) show that IE ratios do have some non-zero values.

Liquidity measures are markedly rare in the top predictor ranking of both models. Current liabilities to total assets (CL/TA) Y-1 is the highest ranked liquidity variable for both RF (4th) and GB (18th); for GB it is the only one in the top 25. Although it is listed here under liquidity, CL/TA does not measure short-term payment ability and its status as a liquidity ratio is questionable. The only other liquidity variables are found in the RF model's ranking: cash and marketable securities to current liabilities ((C+MS)/CL) Y-1 at 21st, and cash to current liabilities (C/CL) Y-1 at 24th. In this aspect, the findings somewhat contradict the prevalent view in the literature that liquidity is one of the most important predictor categories; however, factors such as local accounting practices and bankruptcy legislation can certainly affect the usefulness of liquidity indicators. Another potential reason is the time period from which the data are collected: the financial crisis and subsequent recession may have caused temporary liquidity issues even to fundamentally healthy companies, making them less distinguishable from those close to failure.

A further point of interest regarding liquidity is that quick ratio (QA/CL) and current ratio (CA/CL), the most commonly used measures, are much less effective predictors than C/CL and (C+MS)/CL. It does seem plausible that the most liquid assets should hold the most predictive power, as inventories and receivables can be somewhat illiquid. In such a case a failing, indebted company is forced to exhaust its cash reserves to cover short-term obligations, while capital remains tied to less liquid assets; this view is shared by Beaver (1968), who notes that failing companies tend to have high inventory balances.

The models behave very differently where growth variables are concerned. RF is not able to use them effectively: the best growth variable is EE/PBD growth Y-1 at rank 91. On the other hand, for GB the compound annual growth rate from Y-3 to Y-1 (CAGR) of AP/S ranks 14th, and growth in operating cash flow (OCF) in Y-2 ranks 24th. Additionally, many growth indicators occupy ranks 25-65. Both profitability and activity are represented among these; growth measures of other variables such as share capital, total debt, or debt to assets, perform poorly.

On the whole, the observed relative variable importances are mostly in line with extant literature; the importance of solvency, activity, profitability, and capital structure is affirmed. The particular significance of equity ratio TE/TL should not be taken for a universal truth, but similar findings in the literature (Tian & Yu, 2017) support the observation. The considerable importance of financial expenses cannot be readily explained and may be particular to this sample; it is nonetheless an interesting finding that could prove useful for future studies. Employee efficiency has been found valuable before (Lin et al., 2012a), but many studies measure it with the number of employees; the results here show that employee expenses are a valid option for constructing the ratios. Employee expenses are also useful due to being a standard item in financial statements, whereas the number of employees is often not included by default; it might be useful to compare the two approaches in a context where both variables are available. One somewhat surprising result is the relatively weak performance of liquidity ratios. However, in most studies some of the typical predictor categories perform worse than others for no apparent reason; this study adds to the evidence showing that the usefulness of different predictor categories can vary significantly depending on the context.

6 Discussion and conclusions

6.1 Discussion of results

The aim of this study was to examine possible means of improving the performance of the bankruptcy prediction model used by Valuatum Ltd. This was carried out by means of a review of relevant literature and an empirical study designed in accordance with the findings of the literature review.

The literature review examined the key aspects of corporate failure prediction, including prediction techniques and the various (accounting-based) predictor variables and their importance. The field of failure prediction is very empirically oriented: theory on the causes of bankruptcy is mostly disconnected from prediction. Definitions and terminology in the literature are somewhat inconsistent, and failure prediction is often addressed as part of a wider conglomeration of literature including subjects such as financial distress prediction and both corporate and consumer credit risk.

The first research question concerned bankruptcy prediction methods: *"Which bankruptcy prediction techniques provide the best balance of performance and usability in the context of Finnish SMEs?"* The literature review showed that there is no consensus regarding the superiority of prediction methods; empirical results vary, and the preferred approach largely depends on the objectives of the study. In line with the goal of finding a practically applicable yet high-performing model, certain methods were excluded from consideration due to their lack of transparency and interpretability. For the empirical study, three decision tree-based ensemble machine learning methods were chosen: random forest, AdaBoost and gradient boosting. An individual decision tree classifier and a logistic regression model were trained as benchmarks. Empirical results showed that all the ensemble models perform well, while the standalone decision tree and logistic regression were noticeably inferior. Random forest performed consistently well; gradient boosting outperformed AdaBoost on imbalanced data, but was slightly inferior on balanced data. Random forest and gradient boosting were chosen for more detailed analysis and testing.

Further analysis showed that random forest outperforms gradient boosting by a small margin; however, gradient boosting appears more robust against changes in class distribution and sample size. Balancing the training set was observed to deteriorate model performance; however, a 1:10 balanced training set produces results closely comparable to the full training set. Nonetheless, the random forest model trained on unadjusted training data gave the highest PR AUC score, and is therefore the recommended choice for Valuatum.

The second area of interest were predictor variables: *"Which accounting-based predictor variables are the most important for Finnish SMEs, and how should the variable set be composed?"* The literature review indicated that certain key determinants of bankruptcy should be represented: at least profitability, solvency, liquidity, capital structure, and activity. A further observation from the extant literature was that the time dimension should be taken into account: in this thesis, the issue was addressed by using data from a three-year period and by including variables that calculate change in financial ratios over time. Based on three prior studies, a large set of predictors was assembled that represents the aforementioned categories. The variable set was augmented with predictors from Valuatum's previous model that had been empirically found to be effective, and some further variables were added based on various literary sources and observed potential deficiencies in the variable set.

Two feature selection techniques were applied to assess the impact of the composition and size of the predictor set. The findings indicated that feature selection mainly induces slightly inferior performance, although improvement was also observed in some cases. Considerable variability was observed between the variable importances of different models: predictors that played a key role in some models could be insignificant for others. Although some individual variables performed consistently well, no evidence was found to indicate that some subset of variables is clearly superior, even for the specific sample used in this study. The best classifiers used in this study are resistant against redundant features, and therefore there is no need to reduce the number of predictors. Thus, the suggestion for practical application is that no feature selection method should be applied.

Two predictor variables were found to be consistently good predictors for different classifiers. Total shareholders' equity to total liabilities in the latest year (TE/TL Y-1) was the most important, and financial expenses to total assets in the latest year (FE/TA Y-1) the second-most important feature for each of the ensemble models used in this study. However, the empirical study also demonstrated that relevant information can be extracted from a large variety of predictors. Other capital structure and solvency measures than the two aforementioned were also found useful. Activity ratios, accounts payable turnover and employee efficiency in particular, showed significant predictive power. Profitability measures were useful, but there was notable variability between the specific ratios used by different models. Contrary to prior literature, liquidity ratios were relatively weak predictors. The overall conclusion is that, as the literature suggests, different predictor categories should be represented and the selection of predictors should be sufficiently large and diverse.

From a practical perspective, the key objective was to achieve improved performance compared to the bankruptcy prediction model previously used by Valuatum. The best model in this regard was found to be the random forest classifier using no feature selection or data resampling methods, which slightly outperformed the previous model in terms of both key metrics, PR AUC (0.195 vs. 0.19) and ROC AUC (0.915 vs. 0.91). Although the improvement is small, the main goal of the thesis in terms of concrete, practicable results was reached. The proposal for Valuatum is to implement the random forest classifier, trained on an unadjusted training set and using the full selection of predictor variables as presented in

Section 3.5. However, since the performance differences between random forest and gradient boosting are relatively small, both models should be included when further analysis and development is conducted.

6.2 Academic and practical implications

From an academic perspective, the findings regarding the superiority of ensemble models to standalone decision tree and logit is not particularly surprising; similar findings are common in the recent literature. In this thesis, random forest slightly outperforms boosting methods, and gradient boosting is superior to AdaBoost. Very similar findings are reported by Brown & Mues (2012), who show that both random forest and gradient boosting perform well and can cope with class imbalance; random forest is found superior in the presence of imbalance similar to this study (1:99). García et al. (2019) also find that random forest performs slightly better than AdaBoost and stochastic gradient boosting; however, their results are too mixed to deem either boosting model better than the other. Contrasting results are also found: Jones et al. (2017) find that boosting models mostly perform better than random forest, although by a small margin, and Barboza et al. (2017) find no notable performance difference between AdaBoost and random forest. This thesis contributes to academic literature by adding evidence of the generally good bankruptcy prediction performance of ensemble models, and particularly random forest; however, no definitive conclusions can be drawn from the findings.

Feature selection methods are found to deteriorate rather than improve performance; this is somewhat contrary to the prevalent view that models should aim for simplicity to maximize predictive capacity (Veganzones & Séverin, 2020). However, decision tree-based classifiers are known to be resistant to irrelevant features (Hastie et al., 2009), and therefore the findings cannot be considered particularly surprising. Additionally, good performance in a high-dimensional context has been observed previously (Jones, 2017).

Perhaps the most interesting result in terms of predictive performance is found by comparing models trained on balanced and imbalanced training sets. As Veganzones & Séverin (2018) suggest, reducing imbalance appears to improve model performance in terms of area under the receiver operating characteristic curve (ROC AUC), the most commonly used metric for failure prediction models. However, the deficiencies of the ROC curve in a highly imbalanced setting have been noted in other fields of research (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015); area under the precision-recall curve (PR AUC) is proposed as a more suitable measure. PR AUC scores reveal that the original imbalanced training set yields superior performance compared to balanced training sets, which contradicts the ROC AUC results and challenges the consensus in the extant literature. Future studies should include PR AUC alongside the commonly used metrics, or otherwise pay attention to finding measures that are appropriate in the presence of significant class imbalance.

Although the performance of individual features is likely to be sample-specific, the findings strongly indicate that shareholders' equity and financial expenses are key determinants of bankruptcy in the Finnish SME context, and that ratios built on either of these two components are useful predictors. Another finding that may be of particular interest is the relatively high importance of employee efficiency ratios. In itself, this is not unexpected, but many studies use the number of employees, whereas this thesis utilizes employee expenses to construct the variables. Although no comparison can be made here between the two approaches, employee expenses seem to be a viable option, at least if the number of employees is not available.

In general, the variable importances are mainly in line with prior literature: capital structure, solvency, activity, and profitability are all found to have predictive power. The relative unimportance of liquidity indicators is somewhat surprising and conflicts with literary consensus. A further observation on liquidity is that traditional indicators such as current and quick ratio are particularly inefficient; predictors involving only the most liquid items perform better. The findings are not enough to refute the usefulness of liquidity variables entirely; however, specific attention should be paid to the design of individual ratios in order to capture the effects of liquidity as accurately as possible.

As noted before, bankruptcy prediction models tend to be sample-specific, and the models used in this thesis are unlikely to perform optimally in different contexts. A variety of factors such as the characteristics of local economy and legislation can greatly impact the relationships between predictor variables and firm failure. However, some general guidelines can be drawn from this thesis for practitioners seeking to construct a high-performing yet practical failure prediction model.

Ensemble classifiers are shown to be accurate; additionally, they are robust against data-related issues, such as missing values and outliers, that are common in financial data. By using decision trees as base learners, the models are also interpretable: the impact of different predictors can be easily quantified to assess their importance. The particular predictors that are most efficient are likely to be dissimilar for different classifiers and samples; therefore it is important that a sufficiently large number of variables is included, at least initially. Decision tree ensembles are also helpful in this respect, since they are not impeded by irrelevant or redundant predictors: the user can include a diverse collection of predictors without deteriorating model performance, and thus is more likely to discover the predictors that work best in the specific context.

6.3 Reliability and validity

6.3.1 Reliability

The reliability of this thesis is affected to some extent by the origin of the empirical data. The financial statements were originally obtained from Bisnode, a reputable commercial data provider; the data source can be considered reliable. The data are used as a key element in the Valuatum analysis platform and monitored for errors or anomalies, and can therefore be assumed to maintain high quality after being received from the data provider. Nonetheless, transferring data between dissimilar systems always involves the possibility of errors. In this study, the data are moved numerous times: from a firm's accounting system to its financial statement, which is then stored in the Bisnode database, whence it is conveyed to the Valuatum database, and finally extracted for use in this thesis. The finding that interest expenses are practically irrelevant, while total financial expenses are a strong predictor, indicates that some type of distortion may have occurred; other similar issues may be present in the data, even if they are not readily observable.

In addition to errors related to data storage and processing, a potential threat to the reliability of this study are accidentally or deliberately misleading financial statement figures. A variety of factors, from purposeful distortion through earnings management (du Jardin, 2019; Serrano-Cinca et al., 2019) to accountant errors, may cause financial statements to misrepresent the true financial status of a company. The value of a financial statement item could be different for two identical companies, and conversely two different companies could have identical values, purely due to choices made by the accountant.

Despite possible errors and inconsistencies, this thesis can be considered reliable. The data are from a reputable, widely used commercial source, and subject to both automatic and manual monitoring and observation. While single financial statement items are not entirely reliable, the empirical study mostly utilizes ratios constructed of aggregate values such as current assets, total liabilities, or EBIT, which can be assumed relatively robust in the presence of minor inconsistencies in individual financial statement items. Missing values or errors may be present, but they are consistently treated as zeroes. Literary source material is chiefly from peer-reviewed scientific journals of good repute, and thus reliable.

6.3.2 Internal validity

The main reason to question the internal validity of this thesis is the lack of underlying theory in corporate failure prediction (Balcaen & Ooghe, 2006). As noted in the literature review, bankruptcy theory remains disconnected from empirical prediction studies, and research design is mainly based on previous empirical findings. Addressing the issue is outside the scope of this thesis, and therefore prior empirical studies are used as the main guideline. Despite the lack of theoretical basis, the design of the study and methodological choices are backed by an extensive body of research spanning decades, and therefore causes no notable concerns regarding internal validity.

It is well established in the literature that accounting-based predictors alone are not sufficient for predicting firm failure. The prediction models constructed in this thesis cannot capture all of the numerous complex factors that affect bankruptcy, but neither are they supposed to do so. The main goal is predicting bankruptcy; omitting some explanatory factors is done as a methodological choice. Thus, the set of predictors used is not an issue in terms of internal validity.

The research process and technical implementation are a potential source of internal validity issues. To avoid possible concerns arising from methodological errors or inconsistencies, the research structure and process are designed in accordance with the relevant literature. Using off-the-shelf procedures for technical implementation helps avoid errors caused by algorithm design flaws. The Python programming language and the scikit-learn package are widely used tools in machine learning research, and can be considered appropriate for this study.

Many prediction methods are known to be unreliable if the available training data are scarce (Alaka et al., 2018). In the empirical part of this thesis, the sample contains the financial statements of over 125 000 companies, which eliminates potential issues caused by data scarcity. The large sample size also allows splitting the data into separate training, validation and test sets. This ensures that the measured out-of-sample performance is not distorted by data leakage (Kaufman et al., 2011). To further improve the robustness and validity of the results, they are averaged over ten runs of the classification procedure.

Any conscious or unconscious biases of the author can impact the research process; in this case, none are acknowledged. Particular attention must also be paid to the fact that this thesis is both an academic and a corporate project. The corporate employer did not impose any explicit or implied restrictions or objectives that could interfere with the research, and the objective from both perspectives is to construct a high-performing prediction model. Thus, there are no conflicts of interest or biases that threaten the internal validity of this thesis.

6.3.3 External validity

As has been discussed previously, this thesis involves some issues that raise concerns regarding external validity, i.e. generalizability of the results. The sample contains financial statement data from over 125 000 Finnish SMEs, and is undoubtedly large enough to be representative of the Finnish SME population. The companies are selected randomly, and therefore population-related bias should not be an issue. This thesis did not attempt to produce results that can be generalized to other populations (e.g. different countries), and the empirical results should not be assumed applicable outside the Finnish SME context.

The generalizability of the results across different time periods is questionable. The prediction models used in this study are only trained and tested on historical data from the years 2008–2010. As Balcaen & Ooghe (2006) point out, this may lead to problems related to non-stationarity and data instability: the relationships between predictor variables and bankruptcy risk change over time due to factors external to the model. For example, it could be assumed that the economic crisis and subsequent recession in 2008–2010 impact the relationships between financial ratios and bankruptcy risk. If input data from a period of economic upturn (or even a different recession) were used, the previously observed variable importances would not apply to the data, and the predictive performance of the models might be different.

Beaver et al. (2005) suggest that financial ratios are robust predictors with regard to time, but nonetheless the results cannot be assumed to generalize well to other periods without further testing. As Serrano-Cinca et al. (2019) suggest, intertemporal validation would increase the generalizability of the results; unfortunately, no suitable data for doing so were available in this study. Compared to prior bankruptcy prediction research, the external validity of this thesis can nonetheless be considered relatively high, because many studies suffer from the aforementioned issues, and additionally use much smaller samples. One important aspect to note is that the ecological validity of this study is certainly high, as it only uses real-world data without any kind of separate experimental setup.

6.4 Limitations

As already discussed in the previous sections, this thesis suffers from some limitations. Relying exclusively on accounting-based variables severely limits the possibilities of discovering the most important predictors overall; only the best financial predictors can be studied. The lack of alternative variables is also likely to impact predictive performance negatively. Although the focus on financial predictors is defined in the scope of this thesis, it is a major limitation that is worth mentioning.

Another limitation is the temporal scope of the empirical data. The results cannot be validated on data from different time periods, and their generalizability is therefore limited. A further issue is that the data are from 2008–2010, a period of financial crisis and recession; the usefulness of the prediction models may be limited to similar economic conditions only. Sole focus on Finnish companies also limits the applicability of the findings to other contexts.

This study predicts bankruptcies on a two-year forecasting horizon; the eventual declaration of bankruptcy may occur long after the proceedings have commenced, and therefore predicting a bankruptcy that takes place in two years or earlier may be of limited use (Balcaen & Ooghe, 2006). Although bankruptcy is the preferred choice of output variable due to having an unambiguous definition (Veganzones & Séverin, 2020), and therefore also used in this study, an alternative measure could be more useful if the prediction models are to be applied as an early warning system.

Despite the extensive number of different accounting-based predictors, all variables from the examined three studies cannot be applied due to lack of sufficiently granular information. Furthermore, there are countless financial predictors in other studies that have been found efficient, but are not used in this thesis; research on bankruptcy prediction is so extensive that some potentially relevant studies are certainly left unexplored, regardless of attempts to conduct a thorough literature review.

Some limitations are imposed by the methodological choice of using the scikit-learn package. The selection of classification methods, although extensive, does not include some of the (less frequently used) alternatives found in the literature. Other aspects such as feature selection and hyperparameter tuning are also limited to those options provided in scikit-learn. Naturally, there are no actual obstacles to using supplementary solutions, but it would conflict with the aim of simplicity and usability. An additional technical limitation is the availability of computational power: some aspects such as the number of hyperparameter combinations for tuning have to be limited, as the prediction modeling is conducted on an ordinary laptop computer with no additional resources.

The usefulness of the results for practical application may be somewhat limited due to the performance metrics used. The ROC and PR curves present an overall view of the models' performance, but are not directly suitable for finding the optimal cutoff threshold for a particular context with specific misclassification costs. On the other hand, F_β allows adjusting the weights based on the user's preference, but is limited to a single cutoff threshold at a time. Misclassification costs are not constant: they depend on the use case and on the firm under observation. Therefore, a more dynamic solution for evaluating model performance would be preferable.

6.5 Future research

The findings of this thesis point to some research topics that should be explored further. The use of the area under the precision-recall curve (PR AUC) for model evaluation is common in some fields, and it is considered more suitable for highly imbalanced data, but failure prediction studies still mostly rely on the area under the receiver operating characteristic curve (ROC AUC). Future studies should include PR AUC or other similar measures to ensure the results are examined from different angles. The widespread reliance on ROC AUC may hide the effects of class imbalance to some extent: as this study shows, balancing the data may improve ROC AUC while simultaneously decreasing PR AUC. Additionally, more research on the effects of different levels of imbalance is needed: this study is unable to provide much concrete evidence, as the depletion of the training data seems to play a part in predictive performance, thus masking the true impact of class imbalance.

Although a large number of predictors are tested in this thesis, research on additional financial variables is warranted. Since annual growth variables do not appear particularly efficient in the empirical study, additional efforts could be made to integrate the time dimension of the bankruptcy process into accounting-based models. For example, Nyitrai (2019) proposes a dynamic indicator variable with promising results; further research on similar concepts could be useful. This approach would also provide an interesting alternative to recent studies that build prediction models based on different failure processes and patterns (du Jardin, 2018; Lukason & Laitinen, 2019). Another potentially interesting subject would be to include comparisons to peer companies, for example by calculating financial ratios in relation to industry median. Such predictors could in many cases provide more relevant information than the financial ratio by itself, because they incorporate industry differences in business models and profit structures.

The literature on non-accounting predictor variables is already extensive, and this thesis provides no additional indication regarding future directions that should be pursued in academic research. However, from a practical perspective the use of non-accounting variables is certainly something worth exploring. The possibilities for Valuatum are somewhat limited due to the technical restrictions imposed by the current software platform, as well as the availability and integrability of new data sources. Macroeconomic indicators could be a suitable starting point, as they are readily available and do not require firm-specific data; experimentation could be carried out manually, without need to commit to costly integration work. The possibility of using market-based information with unlisted companies (see Andrikopoulos & Khorasgani, 2018) could also be considered.

As a final note on future research directions, this study shows that practicality and ease of use do not preclude high predictive performance. Many recent studies develop complex new prediction techniques and achieve excellent results, but in the end, bankruptcy prediction models are of very limited value if they are never applied in practice. Business practitioners could certainly benefit from scientifically developed state-of-the-art prediction models; on the other hand, wide-ranging adoption could advance scientific research further through additional empirical evidence, improved data accessibility, and increased general interest in corporate failure prediction. The ability to accurately predict bankruptcies is of immense value to individuals, companies, and society at large; it is in everyone's best interest to promote deeper cooperation between the scientific and business communities.

References

Abellán, Joaquín & Mantas, Carlos J (2014). “Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring”. *Expert Systems with Applications* 41, pp. 3825–3830.

Acosta-González, Eduardo & Fernández-Rodríguez, Fernando (2014). “Forecasting Financial Failure of Firms via Genetic Algorithms”. *Computational Economics* 43 (2), pp. 133–157.

Acosta-González, Eduardo; Fernández-Rodríguez, Fernando & Ganga, Hicham (2019). “Predicting Corporate Financial Failure Using Macroeconomic Variables and Accounting Data”. *Computational Economics* 53 (1), pp. 227–257.

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

Alaka, Hafiz A.; Oyedele, Lukumon O.; Owolabi, Hakeem A.; Kumar, Vikas; Ajayi, Saheed O.; Akinade, Olugbenga O. & Bilal, Muhammad (2018). “Systematic review of bankruptcy prediction models: Towards a framework for tool selection”. *Expert Systems with Applications* 94, pp. 164–184.

Alfaro, Esteban; García, Noelia; Gámez, Matías & Elizondo, David (2008). “Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks”. *Decision Support Systems* 45 (1), pp. 110–122.

Altman, Edward I. (1968). “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy”. *The Journal of Finance* 23 (4), pp. 589–609.

Altman, Edward I.; Iwanicz-Drozdowska, Małgorzata; Laitinen, Erkki K. & Suvas, Arto (2017). “Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman’s Z-Score Model”. *Journal of International Financial Management & Accounting* 28 (2), pp. 131–171.

Altman, Edward I.; Sabato, Gabriele & Wilson, Nicholas (2010). “The value of non-financial information in small and medium-sized enterprise risk management”. *The Journal of Credit Risk* 6 (2), pp. 95–127.

Amankwah-Amoah, Joseph (2016). “An integrative process model of organisational failure”. *Journal of Business Research* 69 (9), pp. 3388–3397.

Andreeva, Galina; Calabrese, Raffaella & Osmetti, Silvia Angela (2016). “A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models”. *European Journal of Operational Research* 249 (2), pp. 506–516.

Andrikopoulos, Panagiotis & Khorasgani, Amir (2018). “Predicting unlisted SMEs ’ default : Incorporating market information on accounting-based models for improved accuracy”. *The British Accounting Review* 50 (5), pp. 559–573.

Angelini, Eliana; di Tollo, Giacomo & Roli, Andrea (2008). “A neural network approach for credit risk evaluation”. *Quarterly Review of Economics and Finance* 48 (4), pp. 733–755.

Argenti, John (1976). “Corporate planning and Corporate Collapse”. *Long Range Planning* 9 (6), pp. 12–17.

Back, Barbro; Laitinen, Teija & Sere, Kaisa (1996). “Neural networks and genetic algorithms for bankruptcy predictions”. *Expert Systems with Applications* 11 (4), pp. 407–413.

Balcaen, Sofie & Ooghe, Hubert (2006). “35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems”. *The British Accounting Review* 38 (1), pp. 63–93.

Bams, Dennis; Pisa, Magdalena & Wolff, Christian C.P. (2019). “Are capital requirements on small business loans flawed?” *Journal of Empirical Finance* 52, pp. 255–274.

Bao, Wang; Lianju, Ning & Yue, Kong (2019). “Integration of unsupervised and supervised machine learning algorithms for credit risk assessment”. *Expert Systems with Applications* 128, pp. 301–315.

Barboza, Flavio; Kimura, Herbert & Altman, Edward (2017). “Machine learning models and bankruptcy prediction”. *Expert Systems with Applications* 83, pp. 405–417.

Basel Committee on Banking Supervision (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework.* Bank for International Settlements, Basel, Switzerland.

Basel Committee on Banking Supervision (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems.* Bank for International Settlements, Basel, Switzerland.

Battiti, Roberto (1994). “Using Mutual Information for Selecting Features in Supervised Neural Net Learning”. *IEEE Transactions on Neural Networks* 5 (4), pp. 537–550.

Bauweraerts, Jonathan (2016). “Predicting Bankruptcy in Private Firms: Towards a Stepwise Regression Procedure”. *International Journal of Financial Research* 7 (2), pp. 147–153.

Beaver, William H. (1966). “Financial Ratios As Predictors of Failure”. *Journal of Accounting Research* 4, pp. 71–111.

Beaver, William H. (1968). “Alternative Accounting Measures As Predictors of Failure”. *The Accounting Review* 43 (1), pp. 113–122.

Beaver, William H.; McNichols, Maureen F. & Rhie, Jung-Wu (2005). “Have Financial Statements Become Less Informative? Evidence from the Ability of Financial Ratios to Predict Bankruptcy”. *Review of Accounting Studies* 10 (1), pp. 93–122.

Beck, Thorsten; Demirgüç-Kunt, Aslı; Laeven, Luc & Maksimovic, Vojislav (2006). "The determinants of financing obstacles". *Journal of International Money and Finance* 25 (6), pp. 932–952.

Bemš, Július; Starý, Oldřich; Macaš, Martin; Žegklitz, Jan & Pošík, Petr (2015). "Innovative default prediction approach". *Expert Systems with Applications* 42 (17-18), pp. 6277–6285.

Bennasar, Mohamed; Hicks, Yulia & Setchi, Rossitza (2015). "Feature selection using Joint Mutual Information Maximisation". *Expert Systems with Applications* 42 (22), pp. 8520–8532.

Bergstra, James & Bengio, Yoshua (2012). "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research* 13, pp. 281–305.

Black, Fischer & Scholes, Myron (1973). "The pricing of options and corporate liabilities". *Journal of Political Economy* 81 (3), pp. 637–657.

Boritz, J. Efrim & Kennedy, Duane B. (1995). "Effectiveness of Neural Network Types for Prediction of Business Failure". *Expert Systems with Applications* 9 (4), pp. 503–512.

Boughorbel, Sabri; Jarray, Fethi & El-Anbari, Mohammed (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PLoS ONE* 12 (6).

Box, George E.P. & Cox, David R. (1964). "An Analysis of Transformations". *Journal of the Royal Statistical Society, Series B (Methodological)* 26 (2), pp. 211–252.

Box, Travis; Davis, Ryan; Hill, Matthew & Lawrey, Chris (2018). "Operating performance and aggressive trade credit policies". *Journal of Banking and Finance* 89, pp. 192–208.

Breiman, Leo (1996). "Bagging predictors". *Machine Learning* 24 (2), pp. 123–140.

Breiman, Leo (2001). "Random forests". *Machine Learning* 45 (1), pp. 5–32.

Breiman, Leo; Friedman, Jerome H.; Olshen, Richard A. & Stone, Charles J. (1984). *Classification and Regression Trees.* Boca Raton, FL, USA: Chapman & Hall/CRC Press.

Breuniq, Markus M.; Kriegel, Hans Peter; Ng, Raymond T. & Sander, Jörg (2000). "LOF: Identifying density-based local outliers". *SIGMOD Record (ACM Special Interest Group on Management of Data)* 29 (2), pp. 93–104.

Brown, Iain & Mues, Christophe (2012). "An experimental comparison of classification algorithms for imbalanced credit scoring data sets". *Expert Systems with Applications* 39 (3), pp. 3446–3453.

Bryant, Stephanie Mattox (1997). "A Case-Based Reasoning Approach to Bankruptcy Prediction Modeling". *Intelligent Systems in Accounting, Finance & Management* 6 (3), pp. 195–214.

Calabrese, Raffaella; Andreeva, Galina & Ansell, Jake (2019). ""Birds of a Feather" Fail Together : Exploring the Nature of Dependency in SME Defaults". *Risk Analysis* 39 (1), pp. 71–84.

Chan, Aki P.F.; Ng, Wing W.Y.; Yeung, Daniel S.; Tsang, Eric C.C. & Firth, Michael (2006). "Bankruptcy Prediction Using Multiple Classifier System with Mutual Information Feature Grouping". *IEEE International Conference on Systems, Man and Cybernetics.* Vol. 1, pp. 845–850.

Charalambakis, Evangelos C. & Garrett, Ian (2019). "On corporate financial distress prediction: What can we learn from private firms in a developing economy? Evidence from Greece". *Review of Quantitative Finance and Accounting* 52 (2), pp. 467–491.

Charitou, Andreas; Neophytou, Evi & Charalambous, Chris (2004). "Predicting Corporate Failure: Empirical Evidence for the UK". *European Accounting Review* 13 (3), pp. 465–497.

Chawla, Nitesh V.; Bowyer, Kevin W.; Hall, Lawrence O. & Kegelmeyer, W. Philip (2002). "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Chen, Mu Yen (2012). "Comparing traditional statistics, decision tree classification and support vector machine techniques for financial bankruptcy prediction". *Intelligent Automation and Soft Computing* 18 (1), pp. 65–73.

Chen, Ning; Ribeiro, Bernardete; Vieira, Armando S.; Duarte, João & Neves, João C. (2011). "A genetic algorithm-based approach to cost-sensitive bankruptcy prediction". *Expert Systems with Applications* 38 (10), pp. 12939–12945.

Chou, Chih-Hsun; Hsieh, Su-Chen & Qiu, Chui-Jie (2017). "Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction". *Applied Soft Computing Journal* 56, pp. 298–316.

Chung, Kim Choy; Tan, Shin Shin & Holdsworth, David K. (2008). "Insolvency Prediction Model Using Multivariate Discriminant Analysis and Artificial Neural Network for the Finance Industry in New Zealand". *International Journal of Business and Management* 39 (1), pp. 19–28.

Ciampi, Francesco (2015). "Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms". *Journal of Business Research* 68 (5), pp. 1012–1025.

Ciampi, Francesco & Gordini, Niccolò (2013). "Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises". *Journal of Small Business Management* 51 (1), pp. 23–45.

Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises (2003). *Official Journal of the European Union* L 124, pp. 36–41.

Cortes, Corinna & Vapnik, Vladimir (1995). "Support-Vector Networks". *Machine Learning* 20, pp. 273–297.

Davidson, Sidney; Sorter, George H. & Kalle, Hemu (1964). “Measuring the Defensive Position of a Firm”. *Financial Analysts Journal* 20 (1), pp. 23–29.

Davis, Jesse & Goadrich, Mark (2006). “The relationship between Precision-Recall and ROC curves”. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240.

De Bock, Koen W. (2017). “The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles”. *Expert Systems With Applications* 90, pp. 23–39.

De Wit, Gerrit & de Kok, Jan (2014). “Do small businesses create more jobs? New evidence for Europe”. *Small Business Economics* 42 (2), pp. 283–295.

Deakin, Edward B. (1972). “A Discriminant Analysis of Predictors of Business Failure”. *Journal of Accounting Research* 10 (1), pp. 167–179.

DeFond, Mark L. & Jiambalvo, James (1994). “Debt covenant violation and manipulation of accruals”. *Journal of Accounting and Economics* 17 (1-2), pp. 145–176.

Delen, Dursun; Kuzey, Cemil & Uyar, Ali (2013). “Measuring firm performance using financial ratios: A decision tree approach”. *Expert Systems with Applications* 40 (10), pp. 3970–3983.

Dimitras, A. I.; Zanakis, S. H. & Zopounidis, C. (1996). “A survey of business failures with an emphasis on prediction methods and industrial applications”. *European Journal of Operational Research* 90 (3), pp. 487–513.

Du Jardin, Philippe (2010). “Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy”. *Neurocomputing* 73 (10-12), pp. 2047–2060.

Du Jardin, Philippe (2015). “Bankruptcy prediction using terminal failure processes”. *European Journal of Operational Research* 242 (1), pp. 286–303.

Du Jardin, Philippe (2017). “Dynamics of firm financial evolution and bankruptcy prediction”. *Expert Systems with Applications* 75, pp. 25–43.

Du Jardin, Philippe (2018). “Failure pattern-based ensembles applied to bankruptcy forecasting”. *Decision Support Systems* 107, pp. 64–77.

Du Jardin, Philippe (2019). “Forecasting bankruptcy using biclustering and neural network-based ensembles”. *Annals of Operations Research.*

Du Jardin, Philippe & Séverin, Eric (2011). “Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model”. *Decision Support Systems* 51 (3), pp. 701–711.

Du Jardin, Philippe; Veganzones, David & Séverin, Eric (2019). “Forecasting Corporate Bankruptcy Using Accrual-Based Models”. *Computational Economics* 54 (1), pp. 7–43.

Edmister, Robert O. (1972). “An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction”. *The Journal of Financial and Quantitative Analysis* 7 (2), pp. 1477–1493.

El Kalak, Izidin & Hudson, Robert (2016). “The effect of size on the failure probabilities of SMEs : An empirical study on the US market using discrete hazard model”. *International Review of Financial Analysis* 43, pp. 135–145.

Elliott, Thomas (2019). *The State of the Octoverse: machine learning.* [Online] Available at: https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/, [accessed 2020-03-02].

European Commission (2008). *Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions - “Think Small First” - A “Small Business Act” for Europe.*

Fan, Shuoshuo; Liu, Guohua & Chen, Zhao (2018). “Anomaly detection methods for bankruptcy prediction”. *4th International Conference on Systems and Informatics, ICSAI 2017.* Vol. 2018-Janua, pp. 1456–1460.

Faris, Hossam; Abukhurma, Ruba; Almanaseer, Waref; Saadeh, Mohammed; Mora, Antonio M.; Castillo, Pedro A. & Aljarah, Ibrahim (2020). “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market”. *Progress in Artificial Intelligence* 9 (1), pp. 31–52.

Fawcett, Tom (2006). “An introduction to ROC analysis”. *Pattern Recognition Letters* 27 (8), pp. 861–874.

Figini, Silvia; Bonelli, Federico & Giovannini, Emanuele (2017). “Solvency prediction for small and medium enterprises in banking”. *Decision Support Systems* 102, pp. 91–97.

Filipe, Sara Ferreira; Grammatikos, Theoharry & Michala, Dimitra (2016). “Forecasting distress in European SME portfolios”. *Journal of Banking & Finance* 64, pp. 112–135.

FitzPatrick, Paul J (1932). “A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies”. *Certified Public Accountant.*

Fletcher, Desmond & Goss, Ernie (1993). “Forecasting with neural networks. An application using bankruptcy data”. *Information and Management* 24 (3), pp. 159–167.

Florez-Lopez, Raquel (2010). “Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data”. *Journal of the Operational Research Society* 61 (3), pp. 486–501.

Florez-Lopez, Raquel & Ramon-Jeronimo, Juan Manuel (2015). “Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment . A correlated-adjusted decision forest proposal”. *Expert Systems with Applications* 42 (13), pp. 5737–5753.

Freund, Yoav & Schapire, Robert E. (1997). “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. *Journal of Computer and System Sciences* 55 (1), pp. 119–139.

Friedman, Jerome H. (2001). “Greedy Function Approximation: A Gradient Boosting Machine”. *The Annals of Statistics* 29 (5), pp. 1189–1232.

Friedman, Jerome H. (2002). “Stochastic gradient boosting”. *Computational Statistics and Data Analysis* 38 (4), pp. 367–378.

García, Vicente; Marqués, Ana I. & Sánchez, J. Salvador (2015). “An insight into the experimental design for credit risk and corporate bankruptcy prediction systems”. *Journal of Intelligent Information Systems* 44 (1), pp. 159–189.

García, Vicente; Marqués, Ana I. & Sánchez, J. Salvador (2019). “Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction”. *Information Fusion* 47 (July 2018), pp. 88–101.

Giriūniene, Gintare; Giriūnas, Lukas; Morkunas, Mangirdas & Brucaite, Laura (2019). “A Comparison on Leading Methodologies for Bankruptcy Prediction: The Case of the Construction Sector in Lithuania”. *Economies* 7 (3).

Gordini, Niccolò (2014). “A genetic algorithm approach for SMEs bankruptcy prediction : Empirical evidence from Italy”. *Expert Systems with Applications* 41, pp. 6433–6445.

Gruszczyński, Marek (2019). “On unbalanced sampling in bankruptcy prediction”. *International Journal of Financial Studies* 7 (2).

Gupta, Jairaj; Barzotto, Mariachiara & Khorasgani, Amir (2018). “Does size matter in predicting SMEs failure?” *International Journal of Finance and Economics* 23 (4), pp. 571–605.

Gupta, Jairaj & Chaudhry, Sajid (2019). “Mind the tail, or risk to fail”. *Journal of Business Research* 99, pp. 167–185.

Gupta, Jairaj; Gregoriou, Andros & Healy, Jerome (2015). “Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter?” *Review of Quantitative Finance and Accounting* 45 (4), pp. 845–869.

Gupta, Jairaj; Wilson, Nicholas; Gregoriou, Andros & Healy, Jerome (2014). “The value of operating cash flow in modelling credit risk for SMEs”. *Applied Financial Economics* 24 (9), pp. 649–660.

Guyon, Isabelle & Elisseeff, André (2003). “An introduction to variable and feature selection”. *Journal of Machine Learning Research* 3, pp. 1157–1182.

Guyon, Isabelle; Weston, Jason; Barnhill, Stephen & Vapnik, Vladimir (2002). “Gene selection for cancer classification using Support Vector Machines”. *Machine Learning* 46, pp. 389–422.

Hand, David J. (2009). “Measuring classifier performance: A coherent alternative to the area under the ROC curve”. *Machine Learning* 77 (1), pp. 103–123.

Hastie, Trevor; Tibshirani, Robert & Friedman, Jerome (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* II. New York, NY, USA: Springer.

Henderson, Andrew D (1999). “Firm Strategy and Age Dependence : A Contingent View of the Liabilities of Newness, Adolescence, and Obsolescence”. *Administrative Science Quarterly* 44 (2), pp. 281–314.

Hill, Nancy Thorley; Perry, Susan E. & Andes, Steven (1996). “Evaluating Firms In Financial Distress: An Event History Analysis”. *Journal of Applied Business Research* 12 (3), pp. 60–71.

Hillegeist, Stephen A.; Keating, Elizabeth K.; Cram, Donald P. & Lundstedt, Kyle G. (2004). “Assessing the probability of bankruptcy”. *Review of Accounting Studies* 9, pp. 5–34.

Ho, Tin Kam (1998). “The Random Subspace Method for Constructing Decision Forests”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8), pp. 832–844.

Horváthová, Jarmila & Mokrišová, Martina (2018). “Risk of bankruptcy, its determinants and models”. *Risks* 6 (4).

Hu, Yu-Chiang & Ansell, Jake (2007). “Measuring retail company performance using credit scoring techniques”. *European Journal of Operational Research* 183 (3), pp. 1595–1606.

Huang, Yu-Pei & Yen, Meng-Feng (2019). “A new perspective of performance comparison among machine learning algorithms for financial distress prediction”. *Applied Soft Computing* 83.

Huang, Zan; Chen, Hsinchun; Hsu, Chia-Jung; Chen, Wun-Hwa & Wu, Soushan (2004). “Credit rating analysis with support vector machines and neural networks: A market comparative study”. *Decision Support Systems* 37 (4), pp. 543–558.

Jackson, Richard H.G. & Wood, Anthony (2013). “The performance of insolvency prediction and credit risk models in the UK: A comparative study”. *British Accounting Review* 45 (3), pp. 183–202.

Johnsen, Thomajean & Melicher, Ronald W. (1994). “Predicting corporate bankruptcy and financial distress: Information value added by multinomial logit models”. *Journal of Economics and Business* 46 (4), pp. 269–286.

Jones, Stewart (2011). “Does the capitalization of intangible assets increase the predictability of corporate failure?” *Accounting Horizons* 25 (1), pp. 41–70.

Jones, Stewart (2017). “Corporate bankruptcy prediction: a high dimensional analysis”. *Review of Accounting Studies* 22 (3), pp. 1366–1422.

Jones, Stewart & Hensher, David A. (2004). “Predicting firm financial distress: A mixed logit model”. *The Accounting Review* 79 (4), pp. 1011–1038.

Jones, Stewart; Johnstone, David & Wilson, Roy (2015). “An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes”. *Journal of Banking and Finance* 56, pp. 72–85.

Jones, Stewart; Johnstone, David & Wilson, Roy (2017). “Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks”. *Journal of Business Finance & Accounting* 44 (1-2), pp. 3–34.

Joy, O. Maurice & Tollefson, John O. (1975). “On the Financial Applications of Discriminant Analysis”. *The Journal of Financial and Quantitative Analysis* 10 (5), pp. 723–739.

Karels, Gordon V. & Prakash, Arun J. (1987). “Multivariate Normality and Forecasting of Business Bankruptcy”. *Journal of Business Finance & Accounting* 14 (4), pp. 573–593.

Kaufman, Shachar; Rosset, Saharon & Perlich, Claudia (2011). “Leakage in data mining: Formulation, detection, and avoidance”. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 556–563.

Keasey, Kevin; Pindado, Julio & Rodrigues, Luis (2015). “The determinants of the costs of financial distress in SMEs”. *International Small Business Journal* 33 (8), pp. 862–881.

Kgoroeadira, Reabetswe; Burke, Andrew & van Stel, André (2019). “Small business online loan crowdfunding: who gets funded and what determines the rate of interest?” *Small Business Economics* 52 (1), pp. 67–87.

Kim, Hong Sik & Sohn, So Young (2010). “Support vector machines for default prediction of SMEs based on technology credit”. *European Journal of Operational Research* 201, pp. 838–846.

Kim, Jeong-Bon; Kim, Joung W. & Lim, Jee-Hae (2019). “Does XBRL Adoption Constrain Earnings Management? Early Evidence from Mandated U.S. Filers”. *Contemporary Accounting Research* 36 (4), pp. 2610–2634.

Kim, Myoung-Jong; Kang, Dae-Ki & Kim, Hong Bae (2015). “Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction”. *Expert Systems with Applications* 42 (3), pp. 1074–1082.

Kirkos, Efstathios (2015). “Assessing methodologies for intelligent bankruptcy prediction”. *Artificial Intelligence Review* 43 (1), pp. 83–123.

Kohavi, Ron & John, George H. (1997). “Wrappers for feature subset selection”. *Artificial Intelligence* 97, pp. 273–324.

Kohavi, Ron & Provost, Foster (1998). “Glossary of terms: Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process”. *Journal of Machine Learning* 30, pp. 271–274.

Kuběnka, Michal & Myšková, Renáta (2019). “Obvious and hidden features of corporate default in bankruptcy models”. *Journal of Business Economics and Management* 20 (2), pp. 368–383.

Kücher, Alexander; Mayr, Stefan; Mitter, Christine; Duller, Christine & Feldbauer-Durstmüller, Birgit (2018). “Firm age dynamics and causes of corporate bankruptcy: age dependent explanations for business failure”. *Review of Managerial Science.*

Laitinen, Erkki K. (1991). “Financial ratios and different failure processes”. *Journal of Business Finance & Accounting* 18 (5), pp. 649–673.

Laitinen, Erkki K. (1993). “Financial predictors for different phases of the failure process”. *Omega* 21 (2), pp. 215–228.

Laitinen, Erkki K. & Lukason, Oliver (2014). “Do firm failure processes differ across countries: evidence from Finland and Estonia”. *Journal of Business Economics and Management* 15 (5), pp. 810–832.

Laitinen, Teija & Kankaanpää, Maria (1999). “Comparative analysis of failure prediction methods: The Finnish case”. *European Accounting Review* 8 (1), pp. 67–92.

Le, Tuong; Son, Le Hoang; Vo, Minh Thanh; Lee, Mi Young & Baik, Sung Wook (2018). “A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset”. *Symmetry* 10 (7).

Le, Tuong; Vo, Bay; Fujita, Hamido; Nguyen, Ngoc-Thanh & Baik, Sung Wook (2019). “A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting”. *Information Sciences* 494, pp. 294–310.

Li, Hui & Sun, Jie (2011). “Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study”. *Expert Systems with Applications* 38 (10), pp. 12747–12756.

Li, Kang; Niskanen, Jyrki; Kolehmainen, Mikko & Niskanen, Mervi (2016). “Financial innovation : Credit default hybrid model for SME lending”. *Expert Systems With Applications* 61, pp. 343–355.

Liang, Deron; Lu, ChiaChi; Tsai, Chih-Fong & Shih, Guan-An (2016). “Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study”. *European Journal of Operational Research* 252 (2), pp. 561–572.

Liang, Deron; Tsai, Chih-Fong; Dai, An-Jie & Eberle, William (2018). “A novel classifier ensemble approach for financial distress prediction”. *Knowledge and Information Systems* 54 (2), pp. 437–462.

Liang, Deron; Tsai, Chih-Fong & Wu, Hsin-Ting (2015). “The effect of feature selection on financial distress prediction”. *Knowledge-Based Systems* 73 (1), pp. 289–297.

Lin, Fengyi; Liang, Deron; Yeh, Ching-Chiang & Huang, Jui-Chieh (2014). “Novel feature selection methods to financial distress prediction”. *Expert Systems with Applications* 41 (5), pp. 2472–2483.

Lin, S. M.; Ansell, Jake & Andreeva, Galina (2012a). “Predicting default of a small business using different definitions of financial distress”. *Journal of the Operational Research Society* 63 (4), pp. 539–548.

Lin, Wei-Chao & Lu, Yu-Hsin (2019). “Feature selection in single and ensemble learning-based bankruptcy prediction models”. *Expert Systems* 36 (1).

Lin, Wei-Yang; Hu, Ya-Han & Tsai, Chih-Fong (2012b). “Machine Learning in Financial Crisis Prediction : A Survey”. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42 (4), pp. 421–436.

Löffler, Gunter & Maurer, Alina (2011). “Incorporating the dynamics of leverage into default prediction”. *Journal of Banking and Finance* 35 (12), pp. 3351–3361.

Lohmann, Christian & Ohliger, Thorsten (2019a). “The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models”. *Journal of Forecasting* 38 (5), pp. 375–389.

Lohmann, Christian & Ohliger, Thorsten (2019b). “Using accounting-based information on young firms to predict bankruptcy”. *Journal of Forecasting* 38 (8), pp. 803–819.

Lohmann, Christian & Ohliger, Thorsten (2020). “Bankruptcy prediction and the discriminatory power of annual reports: empirical evidence from financially distressed German companies”. *Journal of Business Economics* 90 (1), pp. 137–172.

López Iturriaga, Félix J. & Sanz, Iván Pastor (2015). “Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks”. *Expert Systems with Applications* 42 (6), pp. 2857–2869.

López, Victoria; Fernández, Alberto; García, Salvador; Palade, Vasile & Herrera, Francisco (2013). “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. *Information Sciences* 250, pp. 113–141.

Lukason, Oliver & Camacho-Miñano, María-del-Mar (2019). “Bankruptcy Risk, Its Financial Determinants and Reporting Delays: Do Managers Have Anything to Hide?” *Risks* 7 (3).

Lukason, Oliver & Laitinen, Erkki K. (2018). “Failure of exporting and non-exporting firms: do the financial predictors vary?” *Review of International Business and Strategy* 28 (3-4), pp. 317–330.

Lukason, Oliver & Laitinen, Erkki K. (2019). “Firm failure processes and components of failure risk : An analysis of European bankrupt firms”. *Journal of Business Research* 98, pp. 380–390.

Mann, H. B. & Whitney, D. R. (1947). “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. *The Annals of Mathematical Statistics* 18 (1), pp. 50–60.

Mantovani, Rafael Gomes; Horváth, Tomáš; Cerri, Ricardo; Junior, Sylvio Barbon; Vanschoren, Joaquin & de Carvalho, André Carlos Ponce de Leon Ferreira (2018). “An empirical study on hyperparameter tuning of decision trees”. *arXiv preprint 1812.02207*.

Mari, Carlo & Marra, Marcella (2019). “Valuing firm’s financial flexibility under default risk and bankruptcy costs: A WACC-based approach”. *International Journal of Managerial Finance* 15 (5), pp. 688–699.

Marqués, A. I.; García, V. & Sánchez, J. S. (2012). “Exploring the behaviour of base classifiers in credit scoring ensembles”. *Expert Systems with Applications* 39 (11), pp. 10244–10250.

Matthews, B. W. (1975). “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. *BBA - Protein Structure* 405 (2), pp. 442–451.

McKee, Thomas E. (2000). “Developing a bankruptcy prediction model via rough sets theory”. *Intelligent Systems in Accounting, Finance & Management* 9 (3), pp. 159–173.

McKee, Thomas E. (2003). “Rough sets bankruptcy prediction models versus auditor signalling rates”. *Journal of Forecasting* 22 (8), pp. 569–586.

Merton, Robert C. (1974). “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates”. *The Journal of Finance* 29 (2), pp. 449–470.

Min, Jae H. & Lee, Young-Chan (2005). “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters”. *Expert Systems with Applications* 28 (4), pp. 603–614.

Modigliani, Franco & Miller, Merton H. (1958). “The cost of capital, corporation finance and theory of investment”. *The American Economic Review* 48 (3), pp. 261–297.

Modigliani, Franco & Miller, Merton H. (1963). “Corporate Income Taxes and the Cost of Capital: A Correction”. *The American Economic Review* 53, pp. 433–443.

Modina, M. & Pietrovito, F. (2014). “A default prediction model for Italian SMEs: the relevance of the capital structure”. *Applied Financial Economics* 24 (23), pp. 1537–1554.

Molina, Carlos A. (2005). “Are firms underleveraged? An examination of the effect of leverage on default probabilities”. *Journal of Finance* 60 (3), pp. 1427–1459.

Molina, Carlos A. & Preve, Lorenzo A. (2012). “An Empirical Analysis of the Effect of Financial Distress on Trade Credit”. *Financial Management* 41 (1), pp. 187–205.

Moscarini, Giuseppe & Postel-Vinay, Fabien (2012). "The Contribution of Large and Small Employers to Job Creation in Times of High and Low Unemployment". *American Economic Review* 102 (6), pp. 2509–2539.

Nanni, Loris & Lumini, Alessandra (2009). "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring". *Expert Systems with Applications* 36, pp. 3028–3033.

Nyitrai, Tamás (2019). "Dynamization of bankruptcy models via indicator variables". *Benchmarking: An International Journal* 26 (1), pp. 317–332.

Nyitrai, Tamás & Virág, Miklós (2019). "The effects of handling outliers on the performance of bankruptcy prediction models". *Socio-Economic Planning Sciences* 67, pp. 34–42.

Odom, Marcus D. & Sharda, Ramesh (1990). "A neural network model for bankruptcy prediction". *IJCNN. International Joint Conference on Neural Networks*, pp. 163–168.

Ohlson, James A. (1980). "Financial Ratios and the Probabilistic Prediction of Bankruptcy". *Journal of Accounting Research* 18 (1), pp. 109–131.

Olson, David L.; Delen, Dursun & Meng, Yanyan (2012). "Comparative analysis of data mining methods for bankruptcy prediction". *Decision Support Systems* 52 (2), pp. 464–473.

Ooghe, Hubert & De Prijcker, Sofie (2008). "Failure processes and causes of company bankruptcy: A typology". *Management Decision* 46 (2), pp. 223–242.

Papadopoulos, George; Rikama, Samuli; Alajääskö, Pekka; Salah-Eddine, Ziade; Airaksinen, Aarno & Luomaranta, Henri (2015). *Statistics on small and medium-sized enterprises.* Statistical report. Eurostat, Luxembourg.

Pawełek, Barbara (2019). "Extreme Gradient Boosting Method in the Prediction of Company Bankruptcy". *Statistics in Transition* 20 (2), pp. 155–171.

Pedregosa, Fabian; Varoquaux, Gael; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent; Vanderplas, Jake; Passos, Alexandre; Cournapeau, David; Brucher, Matthieu; Perrot, Matthieu & Duchesnay, Édouard (2011). "Scikit-learn: Machine learning in Python". *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peng, Hanchuan; Long, Fuhui & Ding, Chris (2005). "Feature selection based on mutual information". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8), pp. 1226–1238.

Petropoulos, Anastasios; Chatzis, Sotirios P. & Xanthopoulos, Stylianos (2016). "A novel corporate credit rating system based on Student's-t hidden Markov models". *Expert Systems with Applications* 53, pp. 87–105.

Platt, Harlan D. & Platt, Marjorie B. (1990). “Development of a class of stable predictive variables: the case of bankruptcy prediction”. *Journal of Business Finance & Accounting* 17 (1), pp. 31–51.

Platt, Harlan D. & Platt, Marjorie B. (2002). “Predicting corporate financial distress: Reflections on choice-based sample bias”. *Journal of Economics and Finance* 26 (2), pp. 184–199.

Pompe, Paul P.M. & Bilderbeek, Jan (2000). “Faillissementspredictie: Een vergelijking tussen lineaire discriminantanalyse en neurale netwerken”. *Economisch en Sociaal Tijdschrift (Economic and Social Journal)* 54 (2), pp. 215–242.

Pompe, Paul P.M. & Bilderbeek, Jan (2005). “The prediction of bankruptcy of small- and medium-sized industrial firms”. *Journal of Business Venturing* 20 (6), pp. 847–868.

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

Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regulations on specific statistical domains (2006). *Official Journal of the European Union* L 393, pp. 1–39.

Reisz, Alexander S. & Perlich, Claudia (2007). “A market-based framework for bankruptcy prediction”. *Journal of Financial Stability* 3 (2), pp. 85–131.

Saito, Takaya & Rehmsmeier, Marc (2015). “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. *PLoS ONE* 10 (3), pp. 1–39.

Sanfilippo-Azofra, Sergio; López-Gutiérrez, Carlos & Torre-Olmo, Begoña (2016). “Coverage of financing deficit in firms in financial distress under the pecking order theory”. *E a M: Ekonomie a Management* 19 (4), pp. 104–116.

Sartori, Fabio; Mazzucchelli, Alice & Gregorio, Angelo Di (2016). “Bankruptcy forecasting using case-based reasoning: The CRePERIE approach”. *Expert Systems with Applications* 64, pp. 400–411.

Schapire, Robert E. (1990). “The Strength of Weak Learnability”. *Machine Learning* 5 (2), pp. 197–227.

Schapire, Robert E. (2012). *Boosting: Foundations and Algorithms.* Vol. L 393. MIT Press, pp. 1–39.

Scikit-learn (2020). *Who is using scikit-learn?* [Online] Available at: `https://scikit-learn.org/stable/testimonials/testimonials.html`, [accessed 2020-03-02].

Serrano-Cinca, Carlos; Gutiérrez-Nieto, Begoña & Bernate-Valbuena, Martha (2019). “The use of accounting anomalies indicators to predict business failure”. *European Management Journal* 37 (3), pp. 353–375.

Shin, Kyung-Shik; Lee, Taik Soo & Kim, Hyun-jung (2005). “An application of support vector machines in bankruptcy prediction model”. *Expert Systems with Applications* 28 (1), pp. 127–135.

Shin, Kyung-Shik & Lee, Yong-Joo (2002). “A genetic algorithm application in bankruptcy prediction modeling”. *Expert Systems with Applications* 23 (3), pp. 321–328.

Shulman, Joel S & Cox, Raymond A K (1985). “An Integrative Approach to Working Capital Management”. *Journal of Cash Management* 5 (6), pp. 64–67.

Shumway, Tyler (2001). “Forecasting bankruptcy more accurately: A simple hazard model”. *Journal of Business* 74 (1), pp. 101–124.

Sigrist, Fabio & Hirnschall, Christoph (2019). “Grabit : Gradient tree-boosted Tobit models for default prediction”. *Journal of Banking and Finance* 102, pp. 177–192.

Son, H.; Hyun, C.; Phan, D. & Hwang, H.J. (2019). “Data analytic approach for bankruptcy prediction”. *Expert Systems with Applications* 138, pp. 1–39.

Song, Yongming & Peng, Yi (2019). “MCDM-Based Evaluation Approach for Imbalanced Classification Methods in Financial Risk Prediction”. *IEEE Access* 7, pp. 84897–84906.

Succurro, Marianna; Arcuri, Giuseppe & Costanzo, Giuseppina Damiana (2019). “A combined approach based on robust PCA to improve bankruptcy forecasting”. *Review of Accounting and Finance* 18 (2), pp. 296–320.

Sueyoshi, Toshiyuki & Goto, Mika (2009). “Methodological comparison between DEA (data envelopment analysis) and DEA – DA (discriminant analysis) from the perspective of bankruptcy assessment”. *European Journal of Operational Research* 199 (2), pp. 561–575.

Sun, Jie; Fujita, Hamido; Chen, Peng & Li, Hui (2017). “Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble”. *Knowledge-Based Systems* 120, pp. 4–14.

Sun, Jie; Jia, Ming-yue & Li, Hui (2011). “AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies”. *Expert Systems with Applications* 38 (8), pp. 9305–9312.

Sun, Jie; Lang, Jie; Fujita, Hamido & Li, Hui (2018). “Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates”. *Information Sciences* 425, pp. 76–91.

Sun, Jie; Li, Hui; Fujita, Hamido; Fu, Binbin & Ai, Wenguo (2020). “Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting". *Information Fusion* 54, pp. 128–144.

Sun, Jie; Li, Hui; Huang, Qing-Hua & He, Kai-Yu (2014). "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches". *Knowledge-Based Systems* 57, pp. 41–56.

Theodossiou, Panayiotis T. (1993). "Predicting Shifts in the Mean of a Multivariate Time Series Process: An Application in Predicting Business Failures". *Journal of the American Statistical Association* 88 (422), pp. 441–449.

Tian, Shaonan & Yu, Yan (2017). "Financial ratios and bankruptcy predictions: An international evidence". *International Review of Economics and Finance* 51, pp. 510–526.

Tobback, Ellen; Bellotti, Tony; Moeyersoms, Julie; Stankova, Marija & Martens, David (2017). "Bankruptcy prediction for SMEs using relational data". *Decision Support Systems* 102, pp. 69–81.

Tsai, Chih-Fong (2009). "Feature selection in bankruptcy prediction". *Knowledge-Based Systems* 22 (2), pp. 120–127.

Tsai, Chih-Fong & Cheng, Kai-Chun (2012). "Simple instance selection for bankruptcy prediction". *Knowledge-Based Systems* 27, pp. 333–342.

Tsai, Chih-Fong & Hsu, Yu-Feng (2013). "A meta-learning framework for bankruptcy prediction". *Journal of Forecasting* 32 (2), pp. 167–179.

Tsai, Chih-Fong; Hsu, Yu-Feng & Yen, David C. (2014). "A comparative study of classifier ensembles for bankruptcy prediction". *Applied Soft Computing Journal* 24, pp. 977–984.

Tseng, Fang-Mei & Lin, Lin (2005). "A quadratic interval logit model for forecasting bankruptcy". *Omega* 33 (1), pp. 85–91.

Van Gestel, Tony; Baesens, Bart; Suykens, Johan; Espinoza, Marcelo; Baestaens, Dirk-Emma; Vanthienen, Jan & De Moor, Bart (2003). "Bankruptcy prediction with least squares support vector machine classifiers". *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr)* L 393, pp. 1–8.

Van Rijn, Jan N. & Hutter, Frank (2017). "An empirical study of hyperparameter importance across datasets". *CEUR Workshop Proceedings* 1998, pp. 1–39.

Van Rijn, Jan N. & Hutter, Frank (2018). "Hyperparameter importance across datasets". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* L 393, pp. 2367–2376.

Van Rijsbergen, C. J. (1974). "Foundation of evaluation". *Journal of Documentation* 30 (4), pp. 365–373.

Vassalou, Maria & Xing, Yuhang (2004). "Default risk in equity returns". *The Journal of Finance* 59 (2), pp. 831–868.

Veganzones, David & Séverin, Eric (2018). “An investigation of bankruptcy prediction in imbalanced datasets”. *Decision Support Systems* 112, pp. 111–124.

Veganzones, David & Séverin, Eric (2020). “Corporate failure prediction models in the twenty-first century: a review”. *European Business Review* (article in press),

Verikas, Antanas; Kalsyte, Zivile; Bacauskiene, Marija & Gelzinis, Adas (2010). “Hybrid and ensemble-based soft computing techniques in bankruptcy prediction : a survey”. *Soft Computing* 14 (9), pp. 995–1010.

Virág, Miklós & Nyitrai, Tamás (2014). “Is there a trade-off between the predictive power and the interpretability of bankruptcy models? The case of the first Hungarian bankruptcy prediction model”. *Acta Oeconomica* 64 (4), pp. 419–440.

Volkov, Andrey; Benoit, Dries F & Van den Poel, Dirk (2017). “Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination”. *Decision Support Systems* 98, pp. 59–68.

Wang, Gang; Chen, Gang & Chu, Yan (2018). “A new random subspace method incorporating sentiment and textual information for financial distress prediction”. *Electronic Commerce Research and Applications* 29, pp. 30–49.

Wang, Gang & Ma, Jian (2011). “Study of corporate credit risk prediction based on integrating boosting and random subspace”. *Expert Systems with Applications* 38 (11), pp. 13871–13878.

Wang, Gang; Ma, Jian & Yang, Shanlin (2014). “An improved boosting based on feature selection for corporate bankruptcy prediction”. *Expert Systems with Applications* 41 (5), pp. 2353–2361.

Webb, Geoffrey I. (2000). “MultiBoosting: a technique for combining boosting and wagging”. *Machine Learning* 40 (2), pp. 159–196.

Wells, Joseph T. (2001). “Irrational Ratios”. *Journal of Accountancy* 192 (2), pp. 80–83.

West, David; Dellana, Scott & Qian, Jingxia (2005). “Neural network ensemble strategies for financial decision applications”. *Computers & Operations Research* 32, pp. 2543–2559.

Wilson, Rick L. & Sharda, Ramesh (1994). “Bankruptcy prediction using neural networks”. *Decision Support Systems* 11 (5), pp. 545–557.

XBRL International (2020). *10 Countries with Open Data.* [Online] Available at: https://www.xbrl.org/the-standard/why/ten-countries-with-open-data/, [accessed 2020-03-14].

Xia, Yufei; Liu, Chuanzhe; Li, YuYing & Liu, Nana (2017). “A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring”. *Expert Systems with Applications* 78, pp. 225–241.

Xiang, Dong; Zhang, Yuming & Worthington, Andrew C. (2018). “Determinants of the Use of Fintech Finance among Chinese Small and Medium-Sized Enterprises”. *TEMS-ISIE 2018 - 1st Annual International Symposium on Innovation and*

Entrepreneurship of the IEEE Technology and Engineering Management Society L 393, pp. 1–39.

Xiao, Jin; Xie, Ling; He, Changzheng & Jiang, Xiaoyi (2012). "Dynamic classifier ensemble model for customer classification with imbalanced class distribution". *Expert Systems with Applications* 39 (3), pp. 3668–3675.

Yao, Jian-Rong & Chen, Jia-Rui (2019). "A New Hybrid Support Vector Machine Ensemble Classification Model for Credit Scoring". *Journal of Information Technology Research* 12 (1), pp. 77–88.

Yli-Olli, Paavo & Virtanen, Ilkka (1989). "On the long-term stability and cross-country invariance of financial ratio patterns". *European Journal of Operational Research* 39 (1), pp. 40–53.

Yoon, Hyungwook; Zo, Hangjung & Ciganek, Andrew P. (2011). "Does XBRL adoption reduce information asymmetry?" *Journal of Business Research* 64 (2), pp. 157–163.

Yu, Qi; Miche, Yoan; Séverin, Eric & Lendasse, Amaury (2014). "Bankruptcy prediction using Extreme Learning Machine and financial expertise". *Neurocomputing* 128, pp. 296–302.

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

Zhang, Jie & Thomas, Lyn C. (2015). "The effect of introducing economic variables into credit scorecards: An example from invoice discounting". *Journal of Risk Model Validation* 9 (1), pp. 57–78.

Zhang, Wenyu; He, Hongliang & Zhang, Shuai (2019). "A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm : An application in credit scoring". *Expert Systems With Applications* 121, pp. 221–232.

Zhou, Ligang (2013). "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods". *Knowledge-Based Systems* 41, pp. 16–25.

Zhou, Ligang & Lai, Kin Keung (2017). "AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data". *Computational Economics* 50 (1), pp. 69–94.

Zhou, Ligang; Lai, Kin Keung & Yen, Jerome (2014). "Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation". *International Journal of Systems Science* 45 (3), pp. 241–253.

Zhu, You; Xie, Chi; Wang, Gang-Jin & Yan, Xin-Guo (2017). "Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance". *Neural Computing and Applications* 28, pp. 41–50.

Zhu, You; Zhou, Li; Xie, Chi; Wang, Gang-Jin & Nguyen, Truong V. (2019). "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach". *International Journal of Production Economics* 211, pp. 22–33.

Zmijewski, Mark E. (1984). "Methodological Issues Related to the Estimation of Financial Distress Prediction". *Journal of Accounting Research* 22 (1984), pp. 59–82.

Zoričák, Martin; Gnip, Peter; Drotár, Peter & Gazda, Vladimír (2020). "Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets". *Economic Modelling* 84, pp. 165–176.

A Financial variable abbreviations

A variety of financial ratios are used as predictors in this study, as described in Section 3.5. The abbreviations used for different financial statement items and aggregate values throughout this thesis are presented in Table A1.

Table A1: Financial variable abbreviations

AP	Account payable
AR	Accounts receivable
C	Cash
CA	Current assets
CAPEX	Capital expenditure
CL	Current liabilities
COGS	Cost of goods sold
EBIT	Earnings before interest and taxes
EBITDA	Earnings before interest, taxes, depreciation and amortization
EE	Employee expenses
FA	Fixed assets
FE	Total financial expenses
FI	Total financial income
GP	Gross profit
I	Inventories
IA	Intangible assets (excl. goodwill)
IE	Interest expenses
MS	Marketable securities
NI	Net income
NWC	Net working capital
OCF	Operating cash flow
PBD	Profit before depreciation, amortization and extraordinaries
QA	Quick assets
RE	Retained earnings
S	Sales
SC	Share capital
TA	Total assets
TC	Total capital
TD	Total (interest-bearing) debt
TE	Total shareholders' equity
TR	Total receivables
VA	Value added
WC	Working capital
WD	Asset write-downs

B Sample descriptive statistics

Table B1 presents the descriptive statistics for the empirical data used in this thesis. The table only shows the statistics for the latest available fiscal year (2010, Y-1) and the compound annual growth rate variables. The two earlier years are omitted for practical reasons; the values of the features behave similarly and offer no additional insights. The values of the unscaled variables S and TA are presented in millions of euros, i.e. a value of 1.00 corresponds to EUR 1 000 000.

Table B1: Sample descriptive statistics

	mean	std	min	Q_1	Q_2	Q_3	max
AP/COGS Y-1	0.45	16.24	-2857.00	0.00	0.01	0.14	2850.00
AP/S Y-1	-0.40	203.41	-48722.49	0.00	0.00	0.07	39695.16
AR/S Y-1	43.29	632.10	-73000.00	0.00	8.74	36.27	84680.00
C/CA Y-1	0.40	0.45	-27.00	0.07	0.33	0.71	85.00
C/CL Y-1	2.89	26.31	-581.00	0.03	0.33	1.33	5009.00
C/S Y-1	0.74	11.01	-508.00	0.01	0.07	0.26	1320.71
C/TA Y-1	0.24	0.38	-15.00	0.02	0.12	0.37	85.00
(C+MS)/CL Y-1	3.59	36.67	-581.00	0.03	0.34	1.46	7185.00
(C+MS)/S Y-1	0.97	17.58	-508.00	0.01	0.08	0.28	3182.00
CA/CL Y-1	7.17	76.00	-1035.00	0.62	1.50	3.42	12545.20
CA/S Y-1	2.21	44.24	-642.00	0.14	0.32	0.69	9261.58
CA/S Y-2	2.31	66.14	-1042.00	0.14	0.31	0.69	19899.00
CA/S Y-3	1.85	28.78	-1308.00	0.14	0.30	0.62	3221.00
CA/TA Y-1	0.60	0.34	-4.00	0.29	0.67	0.92	6.00
CA/TD Y-1	5.15	112.30	-2092.12	0.00	0.03	1.26	14301.00
CAPEX/TA Y-1	-0.05	5.44	-1141.00	0.00	0.00	0.05	6.98
CL/S Y-1	1.26	29.03	-460.33	0.07	0.16	0.34	6546.85
CL/TA Y-1	0.51	4.34	-45.50	0.09	0.27	0.54	845.00
CL/TD Y-1	2.56	62.79	-387.00	0.00	0.03	1.00	11504.00
EBIT/IE Y-1	-1.40	50.52	-6958.00	0.00	0.00	0.00	1563.00
EBIT/S Y-1	-0.26	65.67	-22599.67	0.00	0.03	0.14	2675.00
EBIT/TA Y-1	0.03	2.54	-834.50	0.00	0.04	0.16	92.00
EBIT/TE Y-1	0.23	7.39	-1055.62	0.00	0.11	0.39	612.14
EBIT/VA Y-1	0.35	55.34	-5999.00	0.00	0.14	0.50	18079.73
EBITDA/S Y-1	0.03	15.49	-4071.00	0.00	0.07	0.21	2683.00
EBITDA/TA Y-1	0.07	1.32	-211.00	0.00	0.09	0.23	92.00
EE/VA Y-1	-0.49	3.56	-358.33	-0.87	-0.57	0.00	858.00
FA/TA Y-1	0.27	0.31	-4.60	0.01	0.12	0.46	5.00
FE/EBITDA Y-1	-0.35	153.10	-52905.33	-0.06	0.00	0.00	5737.00

	mean	std	min	Q_1	Q_2	Q_3	max
FE/NI Y-1	-0.05	72.69	-15504.00	-0.05	0.00	0.00	8841.00
FE/TA Y-1	-0.02	0.42	-106.00	-0.02	0.00	0.00	1.40
FE/VA Y-1	0.17	24.64	-2003.00	-0.03	0.00	0.00	5737.00
GP/S Y-1	0.63	2.91	-49.70	0.02	0.10	0.34	199.17
GP/TA Y-1	1.45	16.58	-58.26	0.01	0.13	0.66	4969.67
I/COGS Y-1	0.37	6.44	-227.42	0.00	0.00	0.05	759.76
I/NWC Y-1	0.25	8.07	-1559.00	0.00	0.00	0.22	432.67
I/S Y-1	0.36	6.13	-80.67	0.00	0.00	0.07	759.00
I/TA Y-1	0.12	0.34	-84.00	0.00	0.00	0.14	16.00
IA/TA Y-1	0.02	0.08	-0.35	0.00	0.00	0.00	2.50
IE/GP Y-1	-1.42	117.26	-27000.00	0.00	0.00	0.00	8833.33
IE/S Y-1	-0.02	0.92	-192.20	0.00	0.00	0.00	0.62
NI/CL Y-1	-0.08	284.09	-97748.50	-0.03	0.08	0.55	19295.00
NI/IE Y-1	-0.85	29.91	-4782.00	0.00	0.00	0.00	1402.00
NI/S Y-1	0.02	18.73	-4042.00	-0.00	0.02	0.10	2730.28
NI/TA Y-1	-0.02	3.73	-699.00	-0.02	0.03	0.13	548.67
NI/TE Y-1	0.08	5.47	-1074.38	0.00	0.10	0.31	652.50
NI/VA Y-1	0.06	111.02	-23154.00	0.00	0.12	0.49	27928.14
NWC/S Y-1	0.73	39.27	-3715.00	0.00	0.09	0.35	7937.92
NWC/TA Y-1	0.06	4.35	-844.00	-0.02	0.20	0.52	47.50
NWC/TA Y-2	0.13	3.07	-890.00	-0.02	0.20	0.52	20.80
NWC/TA Y-3	0.17	1.22	-152.00	-0.01	0.21	0.51	8.50
OCF/S Y-1	-0.13	32.53	-9296.00	0.00	0.05	0.19	3083.00
OCF/S Y-2	0.10	20.49	-1420.62	0.00	0.05	0.19	3949.50
OCF/S Y-3	0.23	45.48	-4397.90	0.00	0.07	0.24	11627.00
OCF/TA Y-1	0.07	3.42	-308.00	-0.01	0.08	0.22	764.00
OCF/TD Y-1	0.95	47.24	-6139.19	0.00	0.00	0.24	12230.00
OCF/TE Y-1	0.42	22.46	-3610.00	-0.05	0.14	0.49	3186.26
OCF/TE Y-2	0.42	54.30	-17027.50	-0.06	0.14	0.50	3031.00
OCF/TE Y-3	1.16	39.77	-2711.38	-0.02	0.22	0.72	7160.67
OCF/VA Y-1	0.11	112.07	-25133.33	0.00	0.20	0.64	13896.74
OCF/VA Y-2	0.64	137.29	-17012.00	0.00	0.19	0.60	34055.00
OCF/VA Y-3	0.85	135.73	-12880.00	0.00	0.25	0.70	27373.00
QA/CL Y-1	5.62	69.41	-585.00	0.37	1.12	2.72	12545.20
QA/TA Y-1	0.47	0.42	-15.00	0.17	0.44	0.78	85.00
RE/TA Y-1	-0.30	12.03	-2173.00	-0.01	0.20	0.53	56.00
S Y-1	1.19	9.09	-10.54	0.02	0.14	0.54	2377.70
S/TA Y-1	1.92	8.32	-42.00	0.21	1.15	2.42	1403.70
SC/TC Y-1	0.24	1.79	-200.00	0.02	0.07	0.24	200.00
TA Y-1	1.65	16.75	-0.03	0.04	0.15	0.52	2188.24
TD/TA Y-1	0.49	7.00	-74.00	0.00	0.05	0.44	1912.00

	mean	std	min	Q_1	Q_2	Q_3	max
TD/TE Y-1	7.54	1095.24	-16541.00	0.00	0.00	0.58	384871.00
TE/TA Y-1	0.07	8.59	-1912.00	0.14	0.49	0.81	59.00
TE/TL Y-1	0.26	4.91	-910.00	0.19	0.52	0.82	3.14
VA/FA Y-1	11.12	75.22	-9983.00	0.00	1.16	6.69	9810.00
VA/S Y-1	0.33	15.24	-4071.00	0.10	0.37	0.60	2698.00
VA/TA Y-1	0.67	1.71	-211.00	0.06	0.40	0.93	92.00
WC/S Y-1	0.25	38.20	-3715.00	-0.07	0.00	0.12	7929.67
WC/S Y-2	0.12	36.45	-5049.71	-0.07	0.00	0.12	8138.00
WC/S Y-3	0.26	22.54	-3049.00	-0.07	0.00	0.11	2714.50
WC/TA Y-1	-0.10	3.96	-833.00	-0.12	0.00	0.20	42.50
WD/TA Y-1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
growth CAPEX Y-1	1.04	321.84	-14268.80	-1.00	0.00	0.00	107897.33
growth NI Y-1	0.85	487.54	-19475.00	-1.05	-0.32	0.20	171170.00
growth OCF Y-1	-0.50	49.77	-7393.53	-1.37	-0.52	0.20	9277.87
growth TD Y-1	0.31	17.48	-135.00	-0.11	0.00	0.00	4497.86
growth WC Y-1	-0.85	236.71	-83673.00	-0.71	-0.05	0.27	1602.00
*(EBIT+FI)/TA Y-1	0.08	0.28	-8.67	-0.01	0.05	0.18	15.87
*(EBIT+FI)/TC Y-1	636.85	264648.36	-10000000.00	-0.01	0.07	0.26	10000000.00
*ind_risk Y-1	0.01	0.01	0.00	0.00	0.01	0.01	0.03
*PBD/S Y-1	-0.08	16.23	-4397.00	0.00	0.05	0.16	2686.00
*TD/PBD Y-1	-397.69	101718.01	-10000000.00	0.00	0.00	1.45	10000000.00
*(TD-C)/EBITDA Y-1	-79.10	28223.54	-10000000.00	-1.68	0.00	2.00	181726.00
*(TD-C)/TE Y-1	103.24	3332.08	-3819.00	-0.57	-0.04	1.42	690494.00
*TL/S Y-1	3.39	69.53	-1597.67	0.09	0.24	0.70	13799.50
*(TL-C)/S Y-1	2.42	70.10	-3091.00	-0.02	0.11	0.53	13770.00
*(TL-TD)/S Y-1	1.03	27.65	-419.00	0.06	0.14	0.27	6546.85
**EE/NI Y-1	-4.27	151.39	-20359.00	-3.81	-0.00	0.00	21097.00
**EE/PBD Y-1	-2.38	32.32	-1599.00	-3.31	-0.25	-0.00	3126.00
**EE/S Y-1	-0.30	2.38	-396.50	-0.41	-0.19	0.00	96.00
**TR/S Y-1	1.08	39.47	-1008.00	0.02	0.09	0.20	9253.33
**growth AP/S Y-1	0.80	133.38	-11957.29	-0.15	0.00	0.00	40749.98
**growth EBIT/S Y-1	-1.28	338.10	-114902.91	-1.00	-0.18	0.08	13374.31
**growth EE/NI Y-1	-0.12	30.10	-2199.95	-0.99	0.00	0.00	8108.00
**growth EE/PBD Y-1	-0.18	24.26	-1551.00	-0.75	0.00	0.00	7640.00
**growth EE/S Y-1	0.19	7.32	-243.79	-0.10	0.00	0.05	1520.17
**growth EE/VA Y-1	-0.08	24.61	-7519.00	-0.14	0.00	0.06	3004.10
**growth GP/S Y-1	0.10	30.87	-7410.00	-0.16	0.00	0.18	1774.00
**growth NI/S Y-1	-4.04	1674.20	-559009.44	-1.00	-0.21	0.05	132770.99
**growth PBD/S Y-1	-0.78	340.39	-99279.37	-1.00	-0.12	0.11	67566.24
**growth SC Y-1	0.14	31.95	-13.67	0.00	0.00	0.00	11212.00
**growth TD/TA Y-1	0.34	50.29	-102.32	-0.10	0.00	0.00	17500.00

	mean	std	min	Q_1	Q_2	Q_3	max
**growth TR/S Y-1	1.06	29.96	-713.50	-0.30	0.00	0.28	5962.00
**cagr AP/S	0.08	1.22	-1.00	-0.01	0.00	0.00	179.50
**cagr CAPEX	-0.01	2.18	-1.00	-0.61	0.00	0.00	567.94
**cagr EBIT/S	0.06	2.01	-1.00	-0.19	0.00	0.01	491.21
**cagr EE/NI	0.09	1.17	-1.00	-0.06	0.00	0.00	143.63
**cagr EE/PBD	0.07	0.97	-1.00	-0.07	0.00	0.02	180.32
**cagr EE/S	0.02	0.59	-1.00	-0.04	0.00	0.06	38.31
**cagr EE/VA	0.01	0.51	-1.00	-0.04	0.00	0.05	26.13
**cagr GP/S	0.02	0.75	-1.00	-0.14	0.00	0.09	50.97
**cagr NI	0.10	1.09	-1.00	-0.20	0.00	0.05	58.47
**cagr NI/S	0.10	2.42	-1.00	-0.18	0.00	0.00	501.74
**cagr OCF	0.07	1.05	-1.00	-0.25	0.00	0.01	133.13
**cagr PBD/S	0.05	1.31	-1.00	-0.16	0.00	0.04	210.13
**cagr SC	0.02	0.51	-1.00	0.00	0.00	0.00	104.89
**cagr TD	-0.02	0.93	-1.00	-0.10	0.00	0.00	139.09
**cagr TD/TA	-0.03	0.85	-1.00	-0.09	0.00	0.00	146.71
**cagr TR/S	0.13	1.66	-1.00	-0.19	0.00	0.20	350.10
**cagr WC	0.13	1.03	-1.00	-0.14	0.00	0.15	117.09

*: variables from previous Valuatum model (Table 3)
**: additional predictor variables (Table 4)
Variables with no additional markings are from previous studies (Table 2).

C Hyperparameter tuning options

Table C1 presents the hyperparameters that are tuned in the first and second modeling phases, as well as the different values tried for each parameter. The functions of the different hyperparameters are explained in Section 4.4. For further information, the scikit-learn package documentation can be consulted.

Table C1: Hyperparameter tuning options

First modeling phase		
Logistic regression	C	0.1, 1.0, 10.0
	class_weight	’balanced’, None
	penalty	’l1’, ’l2’
Decision tree	class_weight	None, ’balanced’
	max_depth	10, None
	min_samples_leaf	1, 4, 10
Random forest	class_weight	None, ’balanced’
	max_features	’log2’, ’sqrt’, 0.5
	min_samples_leaf	1, 4, 10
AdaBoost	learning_rate	0.1, 0.5
	max_depth*	1, 3, 6
Gradient boosting	learning_rate	0.1, 0.5
	max_depth	1, 3, 6
	subsample	0.5, 1.0
Second modeling phase		
Random forest	n_estimators	10, 100, 250
	criterion	’gini’, ’entropy’
	max_depth	5, None
	min_samples_leaf	3, 6, 15
	max_features	’log2’, ’sqrt’, 0.5
Gradient boosting	n_estimators	10, 100, 250
	learning_rate	0.05, 0.1, 0.25
	subsample	0.5, 0.75, 1.0
	max_features	0.5, None
	max_depth	1, 2, 3
	min_samples_leaf	1, 4, 10

*: AdaBoost accepts non-DT base learners and therefore does not have a `max_depth` parameter; the same effect is achieved by giving decision trees with the listed `max_depth` values to the `base_estimator` parameter.