Machine Learning in Credit Risk Management

Advanced Default Risk and Credit Ratings Modeling

A dissertation in partial fulfillment of the requirements for the degree of
Doktor der Wirtschaftswissenschaft (Dr. rer. pol.)

submitted to the

Faculty of Business, Economics,
and Management Information Systems

University of Regensburg

submitted by

Johannes Raab, M.Sc. (TUM)

Advisors:

Prof. Dr. Daniel Rösch (University of Regensburg)

Prof. Dr. Ralf Kellner (University of Passau)

Date of Disputation:

August 11th, 2022

Acknowledgments

First and foremost, I would like to thank Prof. Dr. Daniel Rösch for his highly valuable input, continuous support and guidance. It has always been an honor and pleasure working with him on this research project. I want to express my gratitude to Prof. Dr. Ralf Kellner for being my second advisor. I would like to thank Prof. Dr. Harald Scheule for being one of my co-authors and taking the time for numerous fruitful discussions on innovative ideas, pushing the manuscript to the best possible version.

I am grateful to all my colleagues at the University of Regensburg and Dr. Nagler & Company for their support and more than anything, for an amazing time. In particular, I wish to express my deepest gratitude to Dr. Martin Nagler for his ongoing support, guidance, and trust. I would also like to thank my co-author Dr. Maximilian Nagl for a great time of collaborative research and never-tiring discussions, i.e., *Coffee & Contributions*, as well as Felix Kircher for some sessions of *Meal & Motivations*.

I would like to dedicate this thesis to my wonderful family for always being my home base and grounding anchor, especially to Tina. Thank you for your love and for giving a rather perfect blend of patient listening, forgiveness of late-night dissertation work, and plenty of balance.

Keep on Smiling and *Be the Best Version of Yourself Every Day.*

Contents

List of Figures

List of Tables

Introduction

Motivation and area of research

The global survey on artificial intelligence (AI) by McKinsey (2021) confirms that businesses continue to adopt AI capabilities, e.g., machine learning (ML) or natural language processing (NLP). While 27% of respondents attribute at least 5% of their earnings to AI, a 23% increase from 2020, there are also ”AI high performers” (McKinsey (2021)) reporting an AI share of earnings of 20% or more. Weighing the costs and benefits leads to the question of how to manage and mitigate the risks arising from the use of AI. Explainability is among the AI risks considered most important, along with cyber security and regulatory compliance.

These developments may also have a significant impact on the financial industry, as shown by the increasing interest of regulatory authorities in different countries in the use of AI. The U.S. Federal Government (2021) surveyed financial institutions on the extent to which AI and ML have already been adopted and what AI use cases they see in particular. In doing so, the agencies put special focus on governance and risk control topics, while identifying the need for transparency as one of the key factors for successful and progressive utilization of ML. In addition, the European Banking Authority (2021a) invited the financial sector to discuss the use of ML in rating modeling, citing conditions for ML deployment such as interpretability by modelers, understandability by management, or justifiability to supervisors. According to the Bank of England (2020), half of the banks surveyed saw an increased importance of ML and data science due to the global Covid-19 pandemic. Hence, the institutions planned to increase investments in AI in many of their business areas.

In summary, this demonstrates the growing interest in the usage of ML methods and allows a glimpse of the importance these approaches may gain in the future methodology landscape of the financial industry. Particular attention should be paid to the explainability of AI, i.e., the need to understand and explain the overall functionality of the models and methods, which tend to be more ”black-box” in nature than traditional approaches (e.g., logistic regressions). However, this requires better knowledge of how to apply techniques that enable measuring the importance of features and their effects on model results, such as non-linear relations or interaction effects. As a consequence, financial institutions may get further support needed to justify the use of ML to both management and regulatory authorities.

Among the risks that credit institutions incur, credit risk, i.e., the risk arising from an obligor's failure to meet contractual obligations, assumes the central role. According to European Banking Authority (2021b), the share of credit risk in regulatory capital requirements, measured in terms of risk-weighted assets (RWA), for 131 major European banks as of September 2021 was 83.2%, corresponding to an RWA volume of 7,380 billion Euros. This amount demonstrates the materiality of credit risks and underlines the importance of comprehensible and reliable methods for measuring them. The probability of default (PD) of a debtor is one of the most important parameters in credit risk management (see, e.g., Ding et al. (2012)), along with the exposure at default (EAD) and the loss given default (LGD). In particular, PD models are used for loan pricing, loan loss provisioning, or capital allocations.

As an external assessment of an entity's creditworthiness, major rating agencies such as Moody's provide credit ratings at a single and commonly known scale (see, e.g., Moody's (2021b)). Thus, all participants in the financial markets are enabled to easily assess the risk characteristic of individual securities or corporates. As a way to quantify and manage risk, ratings are used extensively in private contracting and regulation (see, e.g., Becker and Milbourn (2011)), including their relevance for company valuations, determination of interest payments, investment decision processes, or allocation of regulatory capital (see, e.g., Hilscher and Wilson (2017)). This underpins the importance of external creditworthiness assessments, since diverse stakeholders in the economy, such as financial institutions, regulators, governments, employees, or the rated entity itself rely on credit ratings (see, e.g., Bonsall and Miller (2017)).

Quantitative information such as accounting and market-based variables have been found to be important factors in both modeling PD (see, e.g., Campbell et al. (2008) or Aretz et al. (2018)) and rating (see, e.g., Alp (2013) or Baghai et al. (2014)). Beyond financial ratios, also qualitative data is a valuable source of information that may be interesting in the credit risk management process. There is evidence that qualitative information such as business strategy or market position may be involved in the rating procedure of rating agencies (see, e.g., Moody's (2021b)) and text data may be exploited for additional risk differentiation of PD models (see European Banking Authority (2021a)). The great advantage of having public information freely available for retrieval and direct assessment comes along with major challenges such as handling huge amounts of data and processing text in a way that makes it usable and modelable. The application of ML and NLP techniques can be particularly useful in overcoming these challenges and may be key to incorporating the valuable information of textual data into the modeling task.

This thesis aims to address key challenges in the area of credit risk management that arise from the desire or need to apply machine learning and / or incorporate public information in modeling default probabilities and corporate creditworthiness. Overall, these challenges can be encountered in particular by the skills and tools of credit risk managers, so this thesis aims to provide use cases of innovative methods and approaches, thus expanding the methodological toolbox available to credit risk managers. The contributions are handled in three independent research papers presented in the Chapters 1, 2, and 3 of this thesis. The subsequent paragraphs provide an initial introduction to the motivation, background, and focus of each research paper.

Research paper I | *Revisiting the Relation between Corporate Default and Financial Frictions with Machine Learning*

Estimating the default risk of debtors is among the essential tasks in credit risk management (see, e.g., Ding et al. (2012)). Literature such as Campbell et al. (2008) has found that corporate default can be explained by financial frictions. These frictions can be proxied by fundamental features, i.e., cash liquidity, leverage, profitability, and size as well as market features, i.e., absolute share prices, market-to-book ratio, share returns in excess of an index return, and share return volatility. A commonly applied method for PD modeling is logistic regression (LR), but in recent times, literature and industry (e.g., supervisors) have acknowledged the better ability to predict default risk by ML models relative to LR (see, e.g., Petropoulos et al. (2020)). Due to the high value of interpretability and explainability, the application of explainable AI (XAI) techniques in credit risk is a key requirement for practitioners seeking to use ML in modeling risk parameters. In the first research paper (see Chapter 1), the relation between financial frictions and default risk is analyzed using machine learning methods. Default probabilities are predicted over different forecast horizons (1 to 24 months) and the degree of non-linearity is assessed.

Research paper II | *Default Risk and Public Information: A Machine Learning Approach*

Public information represents a comprehensive and valuable source of data that can be retrieved and further processed into a learning base for ML applications using NLP techniques. In different areas of application, researchers have recognized this potential and seized the opportunity to achieve better results, such as improved predictability or explanatory power (see, e.g., Donovan et al. (2021) or Frankel et al. (2022)). The methods for extracting modelable features from texts are diverse and include word frequency approaches, sentiment measures, advanced topic models, or language representation algorithms. Advanced NLP algorithms such as BERT (bidirectional encoder representations from transformers) by Devlin et al. (2019) enable to extract information from texts (e.g., corporate disclosures) and generate multi-dimensional text representations that can be further used in models, for example in artificial neural networks. However, the dimension of highly complex text representations is too high for use in standard models, i.e., logistic and linear regressions. Since these are still widely used in Finance (see, e.g., European Banking Authority (2021a)), the potential of advanced NLP methods cannot yet be fully exploited in important tasks such as risk management. The second research paper (see Chapter 2) targets the relation between default risk and public information. A methodological framework is proposed for extracting textual features from forward-looking Management's Discussion & Analysis (MD&A) sections of corporate reports and reducing dimensionality to generate features that can be modeled using both traditional methods (e.g., logistic regressions) and ML models.

Research paper III | *The Impact of Qualitative Information on Corporate Creditworthiness*

Determining the creditworthiness of a company, which is commonly measured by credit ratings of major rating agencies, is of great concern for various stakeholders in the economy, such as regulators, governments, financial institutions, employees, and the company itself (see, e.g., Bonsall and Miller (2017)). According to the outline of Moody's (2021b), rating agencies use quantitative variables to establish credit ratings, but also qualitative information such as management quality, competitive advantages, business strategy, market position, or salient characteristics of the entity being rated. These qualitative data may be contained in textual financial disclosures such as the MD&A sections of companies' filings to the Securities and Exchange Commission (SEC). However, when targeting the corporate creditworthiness, the impact of financial disclosures is considerably less investigated compared to quantitative ratios. In particular, the question of whether texts from corporate reports can provide independent additional information beyond the common quantitative drivers is of interest. Additionally, there is an ongoing discussion in the literature that these disclosures are becoming less readable, hampering investors to process the information when making investment decisions, see, e.g., Bonsall et al. (2017), Miller (2010), or Li (2008). Hence, the processing of textual information is becoming increasingly relevant and requires advanced methods. The third research paper (see Chapter 3) examines whether qualitative information in the form of the forward-looking MD&A sections of quarterly and annual corporate filings provides valuable information for the determination of a company's creditworthiness.

In particular, this thesis aims to complement the methodological toolbox available for credit risk management by providing use cases of innovative methods and approaches. First, machine learning models such as tree-based methods, i.e., gradient boosting and random forests, and artificial neural networks are applied to model and predict default probabilities (see Chapters 1 and 2). Second, corporate creditworthiness, expressed by credit ratings, is modeled and predicted by using Ordinal Artificial Neural Networks (OANN) following Cao et al. (2020) (see Chapter 3). Third, qualitative information in the form of text from the MD&A sections of companies' quarterly and annual SEC filings is assessed and processed into modelable text features. Methods used for this purpose include word clouds as graphical analysis tool (Chapter 2), t-distributed stochastic neighbor embedding (t-SNE) to reduce dimensionality following van der Maaten and Hinton (2008) (Chapter 2), dictionary-based sentiment analysis following Loughran and Mcdonald (2011) (Chapter 2), the transformer-based natural language processing model BERT by Devlin et al. (2019) (Chapter 2), and a Structural Topic Model (STM) following Roberts et al. (2016) (Chapter 3). Fourth, techniques of explainable AI are applied to evaluate the importance of input features to ML models and uncover drivers of performance improvements, including Accumulated Local Effects (ALE) plots following Apley and Zhu (2020) (see Chapter 1), and measures of first-order and higher-order feature importance following Kellner et al. (2022) (Chapters 2 and 3).

By using these innovative ML methods and explainable AI techniques, this thesis contributes to making the deployment of machine learning in credit risk management more reasonable and secure, as interpretability and explainability become part of the toolkit. The following paragraphs provide an overview on existing literature focusing on machine learning in credit risk management.

Literature

The use of machine learning methods in Finance has become increasingly popular over the last years. Recent studies include ML applications on big data analysis (e.g., Goldstein et al. (2021) or Mullainathan and Spiess (2017)), corporate culture (e.g., Li et al. (2021)), corporate governance (e.g., Erel et al. (2021)), risk premiums (e.g., Gu et al. (2020) or Bianchi et al. (2021)), or mortgage lending (e.g., Fuster et al. (2022)).

A relevant strand of research targets the modeling of important parameters in credit risk management (see, e.g., Donovan et al. (2021)). Besides covering Loss Given Default (LGD, see, e.g., Bellotti et al. (2021) or Kellner et al. (2022)), studies such as Mai et al. (2019), Sariev and Germano (2020), and Stevenson et al. (2021) focus on default prediction. From a regulatory point of view, Petropoulos et al. (2020) compare ML techniques with classical approaches, such as linear discriminant analysis and LR for forecasting bank insolvencies. They find that random forests provide superior performance on their out-of-time data set.[1]

Credit ratings provide a uniform and commonly known scale that allows participants in the financial markets to easily assess the risk characteristics of individual securities or companies. In addition, ratings serve as a way to quantify and manage risks and are used extensively in supervision and private contracting, see, e.g., Becker and Milbourn (2011). In general, they play an important role in the allocation of regulatory capital, in company valuations, in investment decision processes, or in the determination of interest payments, see, e.g., Hilscher and Wilson (2017). Due to this importance of ratings as a measurement of creditworthiness, studies such as Blume et al. (1998), Alp (2013), or Baghai et al. (2014) examine the evolution of corporate credit ratings and their relation to changes in rating standards.[2]

Computational power has increased greatly in recent years, making qualitative information treasures such as public information increasingly accessible, as evidenced by advances in text extraction methodologies and applications in Finance. Literature on the use of text data includes applications on returns (e.g., Frankel et al. (2022), Cohen et al. (2020), Durnev and Mangen (2020), or Muslu et al. (2015)), uncertainty risk (e.g., Friberg and Seiler (2017)), or the assessment of disclosure readability (e.g., Bonsall et al. (2017), Bonsall and Miller (2017), Lehavy et al. (2011), Miller (2010), or Li (2008)). Studies using word list-based text features and sentiment

[1] A comprehensive literature review regarding the modeling of default probabilities can be found in Chapter 1 (Section 1.2, Literature Review) and Chapter 2 (Section 2.2, Literature Review).

[2] A comprehensive literature review regarding credit ratings, its information content, and rating composition can be found in Chapter 3 (Section 3.2, Literature Review).

scores, e.g., based on the dictionary by Loughran and Mcdonald (2011), to predict bankruptcy risk include Nguyen and Huynh (2022), Tang et al. (2020), and Ahmadi et al. (2018).[3]

Contributions

This thesis contributes to the literature on modeling essential parameters of credit risk management and measurement, such as probability of default and credit quality, using machine learning techniques. In particular, it contributes to the literature on generating modelable features from text, using them in models of credit risk parameters, and understanding ML-based PD and rating predictions through the use of explainable AI. The main contributions of this thesis can be organized according to the independent research papers, which are presented in the Chapters 1, 2, and 3 of this thesis.

Contribution I | *Revisiting the Relation between Corporate Default and Financial Frictions with Machine Learning*

In the first research paper (see Chapter 1), machine learning approaches such as neural networks and tree-based methods are applied to investigate the relation between corporate default and financial frictions. As with bank credit risk management, the focus is on predicting default probabilities, so the models are calibrated on a training data set (”in-sample”), validated on a test data set (”out-of-sample”), and applied to a forward-looking data set (”out-of-time”). ML models are found to outperform logistic regression as benchmark model. This paper is among the first to scrutinize the conditions of this dominance based on information. For a comprehensive data set covering public U.S. corporate defaults over a 46-year period, various machine learning models are applied to predict default probabilities over different forecast horizons (1 to 24 months). The use of accumulated local effects (see Apley and Zhu (2020)) plots reveals that financial frictions have non-linear impacts on default risk and that this non-linearity decreases as information fades.

The dominance of ML methods compared to logistic regression is more pronounced for shorter default horizons as the non-linearity decreases as information fades over time. Since market information can absorb new information about a company’s future prospects in a more timely fashion, the ML dominance is more pronounced for market features than for fundamental features. Therefore, resources should be focused on measuring such information and optimizing processing times. This may include digitization of data collection and processing, decreasing

[3] The use of textual information in the literature is summarized in an extensive review and can be found in Chapter 2 (Section 2.2, Literature Review) and Chapter 3 (Section 3.2, Literature Review).

reporting and auditing times, publication periods, and reduction of human overlap. The findings enable financial analysts to generate more accurate predictions and financial institutions to make better resource allocations in terms of evaluating costs and benefits when deploying machine learning methods.

Contribution II | *Default Risk and Public Information: A Machine Learning Approach*

The second research paper (see Chapter 2) shows how text representations generated by the transformer-based NLP model BERT can be transformed into features usable in any machine learning application, but also in standard approaches of PD modeling such as logistic regressions. This can be achieved by applying t-SNE (see van der Maaten and Hinton (2008)) to further process the numerical BERT text representations and reduce their dimensionality. In this way, modelable numerical text features are obtained, which have the further advantage of being objective, i.e., they do not depend on expert knowledge such as word lists and generated training samples that would be required for fine-tuning BERT. Since standard PD models are still prevalent, e.g., in financial institutions, this approach is of great importance for risk professionals who want to integrate text into the models. The drivers of performance improvement in incorporating public information into PD models are deciphered by using XAI techniques according to Kellner et al. (2022).

Non-linear impacts of text, in particular higher-order effects of qualitative textual BERT features (about 43% of the total non-linearity) and interactions with quantitative firm characteristics are uncovered and quantified. The extracted text features provide additional, relevant information, as text accounts for about 36% of feature importance beyond common financial ratios such as a company's leverage. The use of artificial neural networks based on both financial ratios and text as input delivers an increase of AUC by 3.6% (in-sample) and 2.4% (out-of-sample) compared to the logistic regression based on financial ratios only. Model calibration as measured by the Brier Score can be increased by more than 50%, allowing more accurate PD predictions and thus more robust model-based decision making.

Contribution III | *The Impact of Qualitative Information on Corporate Creditworthiness*

The third research paper (see Chapter 3) shows the non-linear impact of (qualitative) public information and (quantitative) firm variables on corporate creditworthiness. In general, how to process text information is becoming more important and requires the application of advanced methods. Using an STM following Roberts et al. (2016), text data is transformed into topic probabilities that quantify how likely it is that an MD&A section contains a specific topic, e.g., real estate, financial risk, or sustainability. In addition to the discussion of *what* information to use, the question of *how* the information is connected to creditworthiness is also relevant. This paper is among the first to consider non-linear relations and interactions between common and novel factors of firms' creditworthiness using an Ordinal Artificial Neural Network (OANN) based on Cao et al. (2020). In addition, non-linear and joint effects of qualitative data on ratings are evaluated for the first time using XAI techniques.

The results of the paper show the importance of features extracted from the texts, i.e., text contributes significantly to explaining and predicting ratings. Structural differences in topic (probability) distributions across rating classes suggest differential effects of text on creditworthiness. A stress test analysis of the topic distribution shows that companies at the change-point between investment grade and non-investment grade tend to be more susceptible to changes in MD&A content. Hence, the MD&A section can make the difference to receive an investment grade rating or not. In this way, stakeholders and investors, especially in firms at the change-point, gain valuable insights into components of corporate creditworthiness.

Structure

The thesis is organized along the three independent research papers with varying co-authors.[4] Chapter 1 introduces the first research paper on the relation between corporate default and financial frictions using machine learning models such as neural networks and tree-based approaches. In Chapter 2, feature extraction from public information in the form of MD&A texts is discussed and the impacts of text on modeling the default probability of corporate bonds are deciphered using techniques of explainable AI. Chapter 3 is dedicated to the assessment of corporate creditworthiness and how it is influenced by qualitative information, i.e., topics extracted from the Management's Discussion & Analysis section of corporate reports. The Conclusion summarizes the thesis by discussing the main findings and providing an outlook for future research in the field of credit risk.

[4] At the beginning of each chapter, information is provided on the current status of the paper and the respective co-author(s). Since the studies were submitted to different journals with varying formal requirements, there may be minor formal differences across the chapters of this thesis.

Chapter 1

Revisiting the Relation between Corporate Default and Financial Frictions with Machine Learning

This chapter is a joint work with Daniel Rösch[*] and Harald Scheule[†], and corresponds to a working paper with the same name (it has been under review by the *Journal of Financial and Quantitative Analysis*).

Abstract

This paper shows that financial frictions have non-linear impacts on corporate defaults as they become more binding and that this non-linearity decreases as information fades. We show that machine learning techniques for default prediction dominate popular logistic regressions for the inclusion of non-linearity using accumulated local effects. In consequence, financial analysts should focus on information collection and timely processing to further enhance the prediction performance of machine learning techniques.

Keywords: Corporate Default; Financial Frictions; Fundamental Information; Machine Learning; Market Information; Non-linearity

JEL classification: C53; G21; G33

[*] University of Regensburg, Chair of Statistics and Risk Management, Universitätsstraße 31, 93040 Regensburg, Germany, email: daniel.roesch@ur.de

[†] University of Technology Sydney, Finance Discipline, Sydney, Australia, email: harald.scheule@uts.edu.au

1.1 Introduction

McKinsey (2020) confirms that firms attribute 20 percent or more of their earnings to artificial intelligence and machine learning (ML). The financial industry is one of the most affected industries. It may be debatable whether machine learning techniques costs outweigh prediction quality and the U.S. Federal Government (2021) has identified the need for transparency as the key factor to success. This paper explains the dominance of machine learning techniques by the non-linearity inherent in market prices and timely processing of information with respect to credit risk. The paper proposes to focus on information collection and processing in conjunction with the application of machine learning techniques.

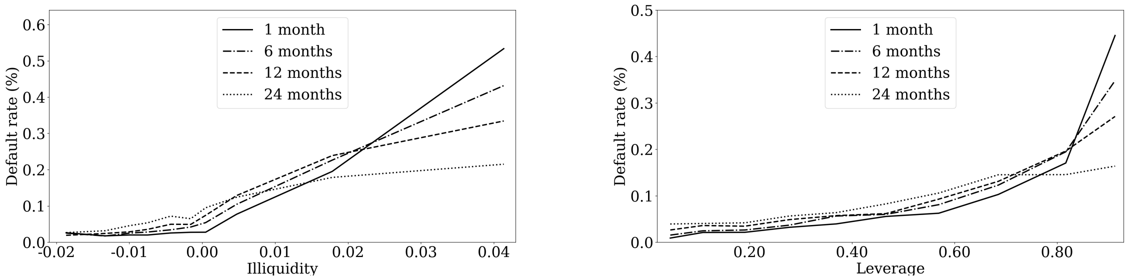
Probabilities of default (PD) of obligors are an essential parameter in credit risk predictions of financial institutions. Methods commonly applied in the industry include dynamic logistic regressions (LR) similar to Campbell et al. (2008), which are linear in terms of how variables are included in the model. Accounting-based fundamental measures such as cash liquidity, profitability, size and leverage or market-based information such as share returns in excess of an index return, market-to-book ratio, absolute share prices and share return volatility contribute significantly to the explanatory power of models.

We analyze a comprehensive data set consisting of over 2.8 million observations and 2,694 default events of public U.S. firms from 1975 to 2020. We document that the link between proxies for financial frictions and default is non-linear. The default risk increases at an accelerating rate as financial frictions become more binding. This non-linearity decreases with the time gap between feature measurement and default period and from market to fundamental features.

Figure 1.1 shows the default rate for different time distances from the measurement of negative profitability as a proxy for net cash shortfalls and hence illiquidity, and leverage. Illiquidity and leverage are important financial constraints underpinning the corporate finance literature (e.g., Tirole (2010)). Firms can default for a number of reasons, but default is often linked to a Chapter 11 reorganization if the firm faces temporary liquidity constraints and to Chapter 7 liquidation if a firm is deemed unviable in the long-term (for an overview, see White (1989)).

Machine learning techniques are strong in including non-linearity, and this paper provides the following three contributions. First, we model non-linear relations between both fundamental and market features and the default risk using the ML techniques neural networks and tree-based methods. ML models dominate LR due to the ability to model non-linear relations. We use accumulated local effects (ALE, see Apley and Zhu (2020)) plots to visualize non-linear effects of features on predicted PD. Second, we find that market features used in Campbell et al. (2008) are more non-linear. ML models dominate LR more when market features are included. Third, we consider multiple default time horizons and find that the non-linearity depreciates over time. We find that the out-performance of ML methods of the benchmark LR decreases as the default time-lag becomes longer and information fades.

Figure 1.1: Non-linear relation between default rate and financial frictions by time-lag



The remainder of the paper is structured as follows. Section 1.2 reviews the relevant literature. Section 1.3 explains default prediction methodologies, procedures for data splitting, cross-validation, hyper-parameter tuning, and the metrics of model validation. Section 1.4 presents the data used in this study. Section 1.5 provides detailed results of the implemented PD models, validation figures, and non-linearity plots. We perform a large number of robustness checks including back-testing using data splits by time, cross-validation of parameters, and hyper-parameter tuning. Section 1.6 discusses the findings and concludes.

1.2 Literature Review

Financial frictions have been found to explain corporate default and can be proxied by fundamental features, i.e., cash liquidity, profitability, size, and leverage as well as market features, i.e., share returns in excess of an index return, market-to-book ratio, absolute share prices, and share return volatility. There is a debate about which category provides better predictive performance. Agarwal and Taffler (2008) find that a market-based contingent-claims valuation approach (Hillegeist et al. (2004) and Bharath and Shumway (2008)) does not dominate the accounting-based Z-score model, and argue that default is more aligned with accounting information. Ohlson (1980) and Zmijewski (1984) focus on accounting-based models, and Reisz and Perlich (2007) find that accounting-based measures are best suited for short-term default forecasts, while market-based models are most relevant for medium- and long-term bankruptcy forecasts. Aretz et al. (2018) observe a positive default risk premium for portfolios constructed on the basis of financial frictions for an international sample of bankruptcy filings. We contribute to this debate by analyzing in detail the relation between financial frictions and the probability of default, as well as the evolution of this relation over time.

Recent studies on the application of ML methods in finance include Gu et al. (2020), who find increased predictive power from applying machine learning methods such as neural networks and trees to the measurement of asset risk premiums relative to regression-based models. Bianchi et al. (2021) measure bond risk premiums with neural networks and extreme trees based on macroeconomic and yield information, documenting the importance of inflation and labor markets. Further, using word embedding based on artificial neural networks, Li et al. (2021) document that a strong corporate culture correlates with corporate operational efficiency and firm value. Erel et al. (2021) show that machine learning methods support the selection process of board members and predict their performance using factors such as business network size or number of current and previous directorships. Goldstein et al. (2021) and Mullainathan and Spiess (2017) propose machine learning as powerful tools for analyzing big data. Fuster et al. (2022) analyze the impact of machine learning on mortgage lending and impact on credit access using logistic regressions and random forests. We contribute to this literature by applying machine learning approaches on corporate default prediction and financial frictions.

Common methods for predicting probability of default include logistic regression. Shumway (2001) analyzes an LR to multi-period, annual observations of both fundamental and market variables, and finds increased predictive accuracy compared to static models, as well as strong relations between bankruptcies and market variables such as stock return, market size and stock return volatility. This study is extended by Chava and Jarrow (2004) by including industry effects and reducing the observation interval to monthly values, which leads to improved predictions, while the inclusion of accounting variables in the model based solely on market variables contributes little to the overall predictive power. Beaver et al. (2005) observe that financial ratios are robust over time with only slight declines in predictive ability, while the inclusion of market-related variables can increase model stability. Generally, financial ratios are used as proxies for financial frictions that result in corporate default. These frictions often have a non-linear impact on default, e.g., a decline in profits from 0 to -1% has a greater impact on default than a decline from 10% to 9%. Non-linear relations between financial ratios and defaults have been included through non-linear feature transformations (see, e.g., Giordani et al. (2014)). For a detailed overview on default prediction studies, we refer to Traczynski (2017).

Literature and industry (e.g., regulators) have also acknowledged the better ability to predict probabilities of default by ML models relative to logistic regression. Among recent studies on the application of ML methods for enhanced credit risk prediction are Petropoulos et al. (2020), who compare ML techniques with classical approaches, such as logistic regression and linear discriminant analysis for bank insolvency forecasting and find that random forests provide superior performance on their out-of-time data set. By amending artificial neural networks with Bayesian regularization to reduce over-fitting the data, Sariev and Germano (2020) achieve increased accuracy in their corporate and retail default predictions.

The combination of standalone methods, such as combining several decision trees (DTs) to ensemble methods like random forests, can lead to further accuracy improvements of prediction models. Sigrist and Hirnschall (2019) apply gradient boosting, i.e., combining multiple DTs into a single strong classifier and a Tobit model and reach an advancement in default prediction accuracy. Tsai et al. (2014) study classifier ensembles of neural networks, support vector machines, and DTs, and find that DT ensembles using boosting have enhanced prediction performance.

Based on this review, we acknowledge that previous literature has found that ML methods are capable of predictions. We are first to scrutinize the conditions of this dominance based on information. Specifically, we analyze the non-linearity of the relation between financial frictions and the probability of default over different forecast horizons, i.e., the depreciation of information over time. This knowledge will enable financial analysts to generate more accurate predictions and financial institutions to make better resource allocations in relation to the cost-benefit trade-off underpinning the deployment of machine learning methods. As a result, the constitution and the banking system are expected to become more resilient.

Following the above literature, we aim to explain the dominance of ML methods based on financial frictions. We devise the following three research hypotheses:

- ML methods dominate logistic regressions in corporate default predictions due to non-linear relations between (fundamental and market) features and default outcomes;
- The dominance of ML methods is more pronounced for market features than for fundamental features since market information absorbs new information about a company's future prospects in a more timely fashion;
- The dominance of ML methods is more pronounced for shorter default horizons as the non-linearity decreases as information fades over time.

1.3 Research Framework

1.3.1 Default Prediction Models

In this study, we apply machine learning approaches to the prediction of probabilities of default (PDs) with a value between zero and one.[1] The various models applied for PD prediction differ in how the features enter the model. Logistic regression and neural networks estimate the outcome parametrically, whereas tree-based methods are non-parametric approaches. All approaches are applied for different settings, defined by hyper-parameters, and optimized.

Logistic Regressions

In a logistic regression (LR), the features linearly enter a non-linear transformation. The probability of default of firm i occurring in period $t = 1, \dots, T$ ($i \in N_t$ with N_t as the set of firms at the start of period t) is defined as

$$PD_{i,t} = P(D_{i,t} = 1 | \boldsymbol{x}_{i,t-1}) = \frac{\exp(\boldsymbol{\beta}' \, \boldsymbol{x}_{i,t-1})}{1 + \exp(\boldsymbol{\beta}' \, \boldsymbol{x}_{i,t-1})}, \tag{1.1}$$

with feature vector $\boldsymbol{x}$ (including a constant) observed at the end of period $t-1$ and parameter vector $\boldsymbol{\beta}$. In our study, we predict PDs for different prediction horizons, i.e., t includes periods such as 1 month, 6 months, 12 months, and 24 months. We use the LR as a benchmark model (see, e.g., Campbell et al. (2008), Giordani et al. (2014), or Butaru et al. (2016)).

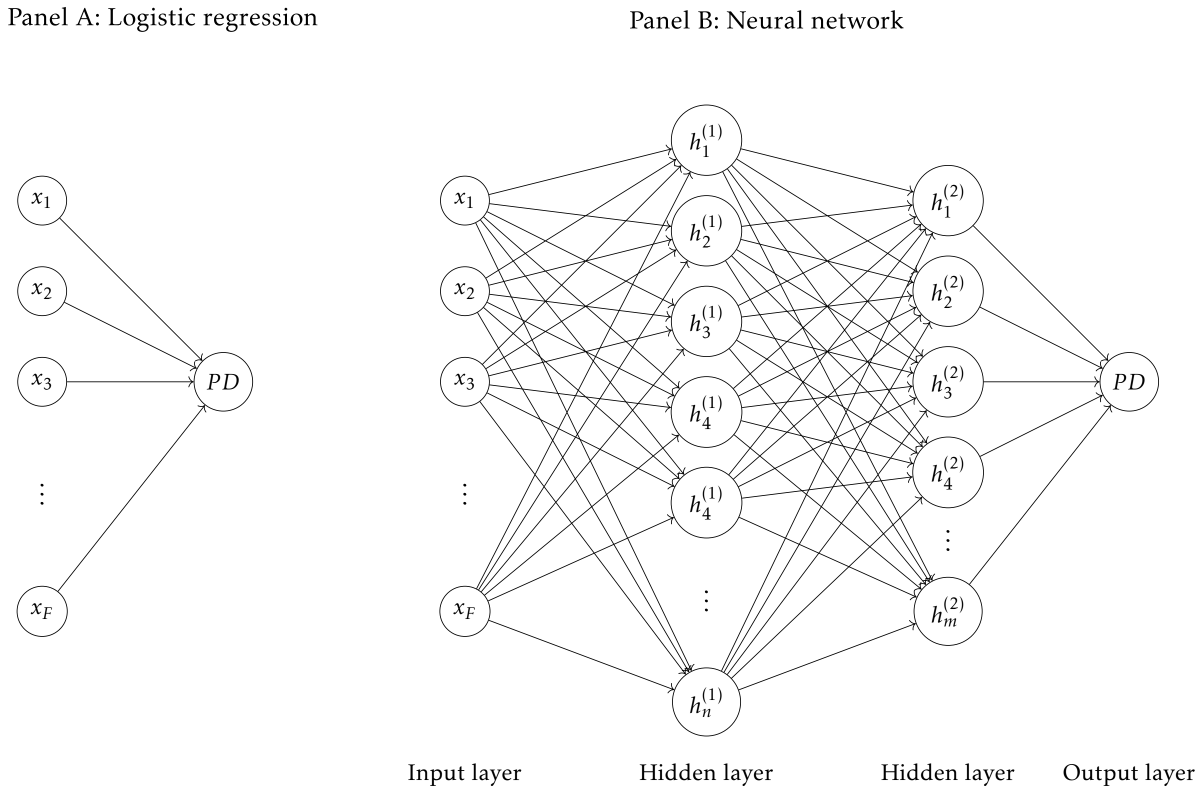
Neural Networks

A neural network (NN) allows non-linear modeling by setting up a flexible structure of connected layers of neurons. The time-lagged features (input layer) are passed to one (or more) so-called hidden layer(s) that non-linearly transform the input and feed the derived features on to an output layer. The schematic representation of an NN architecture compared to LR is displayed in Figure 1.2 (see, e.g., Hastie et al. (2009)). Panel A shows the relation between the PD and the input feature vector $\boldsymbol{x}$ of length F for LR, where $\boldsymbol{x}$ enters linearly. The same input is used in the NN (Panel B), but now the input layer is connected to the output layer via two hidden

[1] Model fit is measured by the binary cross entropy loss function, which is minimized as a target of the model fitting.

layers $h^{(1)}$ and $h^{(2)}$ with n and m neurons, respectively, allowing non-linear transformations in the hidden layers. For default prediction, typically a logistic output function is used in analogy to logistic regression.

Figure 1.2: Graphical overview, parametric approaches

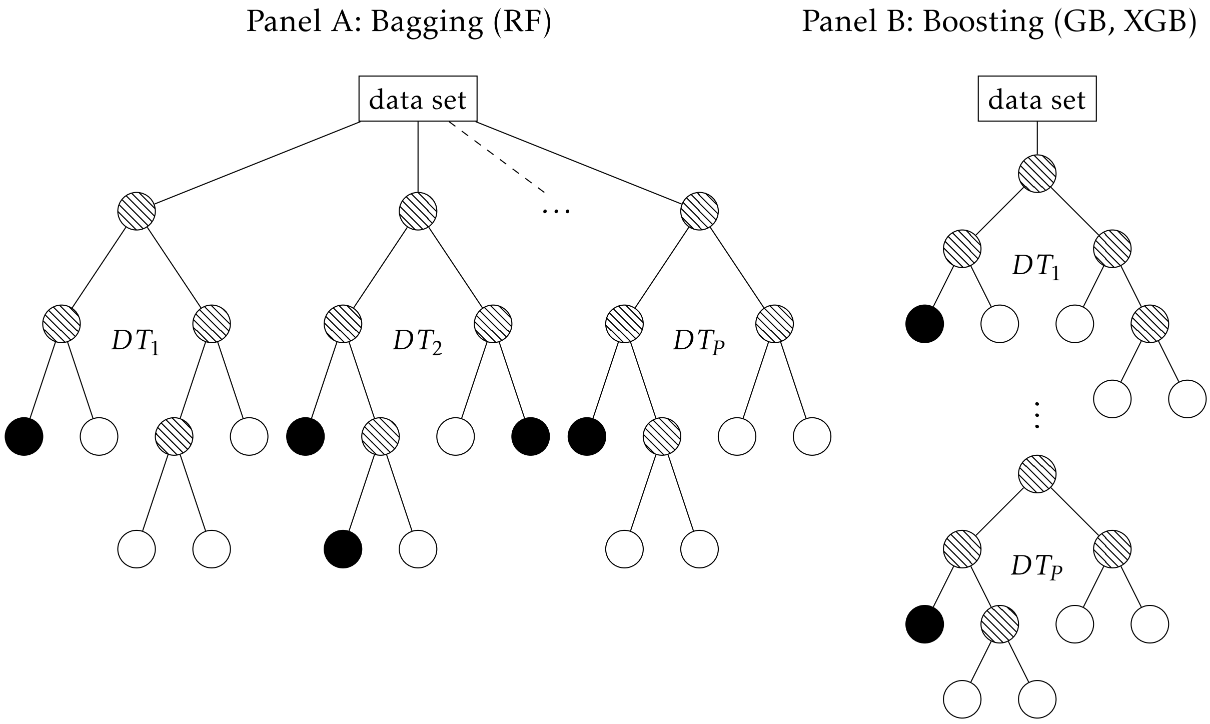


Notes: This figure shows the structure of the standard LR (Panel A) and the NN (Panel B), see, e.g., Hastie et al. (2009). Target variable is the probability of default $PD_{i,t}$ of firm i occurring in period t. LR and NN are parametric approaches. The LR equation in its basic form does not consider interactions and non-linearity. The shown NN example includes two hidden layers $h^{(1)}$ and $h^{(2)}$ with n and m neurons, respectively. The main improvement over the LR is the consideration of interactions and non-linearity. We omit indices for firm and time for simplicity.

Tree-based Methods

Methods based on decision trees divide the input feature space into different sub samples based on cutoff values for the features. This is executed consecutively until a stopping criterion applies and the PD is predicted. Tree-based methods are non-parametric approaches, so that the probability of default predicted by a single DT is equal to the default rate for each class. A major advantage of the application of decision trees is the simple decision tree-based interpretation of the relations between features and outcomes.

Figure 1.3: Graphical overview, non-parametric approaches



Notes: This figure shows the schematic illustration of tree-based ensemble methods (see, e.g., Hastie et al. (2009)). Panel A visualizes a random forest (RF), belonging to the family of bagging algorithms, and Panel B shows the schema of boosting algorithms, which include gradient boosting (GB) and extreme gradient boosting (XGB). In both panels, the shaded circles represent the (root) decision nodes of the trees, and the black and white filled circles represent the leaf nodes of the trees, i.e., default or no default. Tree-based methods take into account interactions and non-linearities and are non-parametric approaches, so the probability of default is represented by multiple leaf nodes. The PD for a single DT is predicted by the default rate for each class. While bagging algorithms such as RF build models in parallel, i.e., combining P randomly created decision trees $DT_1, DT_2, ..., DT_P$ applied to a data set, tree-based boosting techniques fit P decision trees sequentially (see Freund and Schapire, 1997). For ensemble methods, the default probability is predicted by averaging the PDs of the individual decision trees.

We explore three versions of tree-based methods: random forests, boosted and extreme boosted trees. Figure 1.3 shows that these methods consist of ensembles of P decision trees ($DT_1, DT_2, ..., DT_P$).

Random forests (RF, see Breiman (2001)) are based on a random selection of data and variables which are split into nodes. The PDs are predicted by DTs as the default rate for each leaf node. Multiple trees are formed and the default probability is predicted by averaging the PDs of the individual decision trees.

Gradient Boosting (GB, see Friedman (2001)) is another ensemble technique that fits the base classifiers sequentially and optimizes the model accuracy in every step.

Additionally, we apply extreme gradient boosting (XGB, see Chen and Guestrin (2016)), which may improve the model accuracy and provides improved execution speed by parallel computing and advanced regularization, i.e., Ridge (L1) and Lasso (L2) regularization.

1.3.2 Data Splitting, Cross-validation, and Hyper-parameter Tuning

Machine learning typically applies data splitting, cross-validation and hyper-parameter tuning techniques to achieve good predictive power. This allows the methods to identify dependencies or patterns within a portion of the data as well as possible. The fitted models can then be used to make predictions based on data that was not used to train the model.

The concept of splitting the data set into subsets is widely used in the literature (see, e.g., Blöchlinger and Leippold (2011)). We randomly split based on a specific ratio, e.g., 70:30. This means that 70 percent of the data is assigned to the training set (”in-sample”, IS) and the machine learning algorithm learns the structure and context of this data. To validate the fitted model, 30 percent of the data remains in the test set (”out-of-sample”, OS). In contrast to classical ML application areas such as image recognition or sentiment analysis, the prediction of the future is particularly important in credit risk. Capital adequacy, for example, is intended to cover future losses. To represent this, the data set in this study is split in time at the beginning of 2008 so that the models are fit to 70 percent of the observations from 1975 to 2007 (i.e., IS) and further validated on the remaining 30 percent of the data through 2007 (i.e., OS). Hence, we can make predictions for unknown data, i.e., data not used for model training (”out-of-time”, OT) for all observations from 2008 onwards[2].

Many machine learning methods require the determination of hyper-parameters, such as learning rate, regularization strength, or structure of a neural network architecture, which are set by the analyst. In order to find the optimal choice of hyper-parameters for PD prediction, cross-validation and parameter tuning are performed. In the present study, we conduct a five-fold cross-validation on the training data set for every machine learning technique applied. The cross-validation procedure splits the training data randomly into five parts of approximately equal size. The respective model is fitted on four (i.e., one part fewer) parts and the fifth part is predicted measuring the resulting prediction error. The permutation of the fifth prediction part leads to five prediction errors, which are averaged over all folds to assess the total performance of the selected hyper-parameter combination (see Rösch and Scheule (2020)).

[2] We tested different cut-off dates with consistent results.

1.3.3 Model Validation and Relative Improvement

We assess and compare the models' performances as the alignment of the model-predicted PDs and the observed defaults using two validation metrics. The first metric is discrimination, i.e., the ability of the PD model to distinguish between defaults and non-defaults. We calculate the area under the Receiver Operating Characteristic Curve (AUC, see Fawcett (2006) or Kang (2020)), which takes values between 0.5 and 1 for non-random forecasts. A high AUC value means that the model provides high accuracy for classification into default and non-default, depending on a cutoff value C as classification criterion. The number of correctly predicted defaults with the cutoff value C, $H(C)$, relative to the number of defaults N_D is defined as hit rate

$$HR(C) = \frac{H(C)}{N_D}. \tag{1.2}$$

Accordingly, the false alarm rate $FAR(C)$ is defined as number of non-defaults that were classified incorrectly as defaults, $F(C)$, relative to the number of non-defaults (N_{ND}):

$$FAR(C) = \frac{F(C)}{N_{ND}}. \tag{1.3}$$

The AUC then can be calculated as the area under the curve between HR and FAR for cutoff C (see Engelmann et al. (2003)):

$$AUC = \int_0^1 HR(FAR)\, d(FAR). \tag{1.4}$$

The relative improvement in AUC of an ML model compared to the LR is calculated as

$$\Delta AUC = \frac{AUC_{ML}}{AUC_{LR}} - 1. \tag{1.5}$$

To determine whether the AUC for a machine learning model under consideration is higher than the AUC of *LR*, we also perform a hypothesis test using bootstrapping (see Chava and Jarrow (2004)). Observations are randomly drawn with replacement from the list of default observations and predicted default probabilities and the AUC values are calculated. The mean AUC value for the LR model is subtracted from the mean AUC value for the ML model and these steps are repeated 2,000 times as a trade-off between accuracy and computational efficiency. The probability for the hypothesis that the AUC for the ML model predictions is higher than for LR model predictions is then calculated as a percentile rank and reported for different significance levels.

The second metric is calibration, i.e., calibration of PD model predictions and to observed default rates, using the Brier Score (BS, see Brier (1950) or Kruppa et al. (2013)) as the average of squared differences between PD predictions ($P\hat{D}_{i,t}$) and default observations ($D_{i,t}$):

$$BS = \sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot (P\hat{D}_{i,t} - D_{i,t})^2, \tag{1.6}$$

The lower the BS, the lower the deviation between the PD estimates and the default observations, and the better the model calibration. As with the AUC, we perform a hypothesis test to determine whether the BS for a given machine learning model is lower than the BS of *LR*, using bootstrapping.

The relative improvement in BS of an ML model compared to the LR is calculated as

$$\Delta BS = 1 - \frac{BS_{ML}}{BS_{LR}}. \tag{1.7}$$

In the following empirical analysis, we report the measures for discrimination of Equation (1.4) and calibration of Equation (1.6) as well as their relative improvements of Equation (1.5) and of Equation (1.7).

1.4 Data

The data set in this study includes over 2.8 million monthly observations and 2,694 default events of U.S. corporate bonds over a period from 1975 to 2020, covering multiple episodes of economic upturns and downturns. Several fundamental and market variables are collected and combined with the information on the default or non-default of the firms.

1.4.1 Dependent Variables

We measure firm defaults over different time horizons using a default indicator constructed from default information of the US bankruptcy register and Moody's credit rating agency.

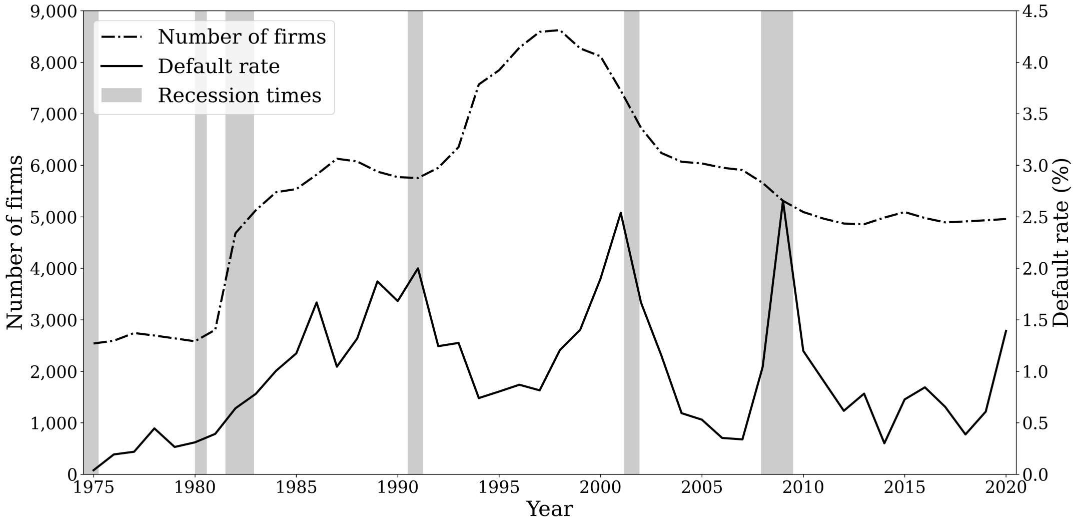
Table 1.1: Number of defaults and firms

Year	Firms	Defaults	Firm months
1975	2,539	1	19,305
1976	2,593	5	29,189
1977	2,743	6	31,247
1978	2,694	12	31,091
1979	2,639	7	30,455
1980	2,582	8	29,757
1981	2,805	11	30,734
1982	4,681	30	47,805
1983	5,125	40	55,790
1984	5,476	55	60,761
1985	5,537	65	61,044
1986	5,817	97	62,326
1987	6,126	64	66,290
1988	6,073	80	66,596
1989	5,876	110	64,337
1990	5,769	97	63,304
1991	5,752	115	62,959
1992	5,950	74	65,300
1993	6,352	81	69,319
1994	7,569	56	83,423
1995	7,844	63	86,046
1996	8,282	72	89,719
1997	8,589	70	94,853
1998	8,623	104	94,047
1999	8,267	116	88,942
2000	8,115	154	88,533
2001	7,449	189	82,292
2002	6,724	112	75,793
2003	6,238	72	70,583
2004	6,067	36	68,268
2005	6,035	32	67,919
2006	5,952	21	67,296
2007	5,906	20	66,070
2008	5,656	59	64,579
2009	5,309	141	60,341
2010	5,092	61	58,088
2011	4,963	45	56,928
2012	4,868	30	55,737
2013	4,854	38	55,126
2014	4,983	15	56,587
2015	5,091	37	58,133
2016	4,975	42	56,634
2017	4,890	32	55,512
2018	4,909	19	55,577
2019	4,931	30	55,896
2020	4,956	70	55,764

Notes: This table presents the number of firms, defaults, and firm months per year over time. The data set covers over 2.8 million firm months and 2,694 default events. The number of defaults is cyclical and higher in times of recession.

Table 1.1 presents the number of firms, defaults, and firm months. In total, the data basis comprises over 2.8 million firm months and 2,694 defaults from 1975 to 2020. Figure 1.4 shows the distribution of firms in the whole data set over time and the default rate per year. Default rates peak in economic downturns as indicated by the National Bureau of Economic Research, i.e., in 1991, 2001 and 2009.

Figure 1.4: Number of firms in the data set and default rate over time



Notes: The figure shows the distribution of the number of firms in the data set from 1975 to 2020 (dot-dashed line) and the default rate over time (solid line). The yearly default rate peaks in 2009 at 2.66%. The shaded areas (gray bars) indicate times of recessions as given by the National Bureau of Economic Research.

1.4.2 Proxies for Financial Frictions

All variables used in this study indicate financial frictions. An increase in financial friction/feature value, e.g., higher illiquidity or leverage, means an increase in default risk. We follow the proxies identified by Campbell et al. (2008).

We reverse the causality of some of these ratios by multiplying them by minus one. The ratios whose signs are changed include $CASHMTA$, $NIMTAAVG$, $EXRETAVG$, and $PRICE$, as marked in Tables 1.2, 1.3, and 1.4, and Figures 1.1, 1.6, 1.7, 1.9, and 1.10. Note that $RSIZE$ changes from a negative to a positive relation from a uni-variate to a multi-variable setting.

The ratios used include fundamental information such as cash liquidity, profitability, size, and leverage, and market-based information such as share returns in excess of an index return, market-to-book ratio, absolute share prices, and share return volatility. Quarterly accounting data on firm level from Compustat as well as daily equity and S&P index market prices from CRSP are the basis of our information features.

First, fundamental features are generated. A company's cash and short-term assets to market value of total assets multiplied by minus one, $-(CASHMTA)$ is an indicator of short-term illiquidity. The market-valued total assets are calculated as the sum of market value of a company's equity and the book value of its liabilities. Market value is preferred over the book value of total assets due to including new information about the prospects of the firm more quickly (see Campbell et al. (2011)). $-(NIMTAAVG)$ is measured by the net income to market-valued total assets, averaged over the previous four quarters and multiplied by minus one. The firm's relative size $RSIZE$ equals the logarithm of the ratio of the company's market capitalization to the S&P500 index market cap. It is an important variable in order to control for potential size effects, such as differences in exploiting scalability. Both numerator and denominator have market values and provide a co-move, i.e., the ratio is market- and time-invariant and no longer driven by market movements. The firm's leverage $TLMTA$ is quantified by total liabilities to market-valued total assets, capturing the capital structure of a firm.

Second, these variables are amended by adding market features. We define market variables as ratios where the numerator is based on market prices. The excess return of each firm's stock averaged over the last twelve months and multiplied by minus one $-(EXRETAVG)$ indicates underperformance relative to the S&P 500. The market-to-book ratio MB measures the efficiency and growth prospects. The logarithmic stock price per share truncated at a cap of \$15, $-(PRICE)$ is a measure of proximity to default. When stock prices fall, a company is moving toward default, and very low stock prices indicate that the distance to default is small (in absolute terms). The standard deviation $SIGMA$ measures the company's daily stock return over the previous three months, thus captures the recent volatility. Low volatility may indicate a limited trading and hence ability to refinance. A moderate volatility may indicate the ability of the enterprise to outperform the reference index on the stock market. A high volatility may indicate a greater likelihood to experience financial distress.

Table 1.2 depicts summary statistics for all features of non-defaulted firms (Panel A) and defaulted firms only (Panel B). On average, the defaulted firms are more illiquid, less profitable, smaller, and more leveraged than the group of non-defaulted companies. The non-default group has a lower negative excess return, market-to-book ratio, negative stock price, and stock return volatility. All means are significantly different from those of the other group, as indicated by the test statistic of the t-test (see Welch (1947)) in Panel C. The descriptive results are consistent with the literature such as Campbell et al. (2008) and show that defaulted firms differ from non-defaulted firms in both fundamental and market information.

Table 1.2: Summary statistics of features

	Fundamental features				Market features			
Feature	$-(CASHMTA)$	$-(NIMTAAVG)$	$RSIZE$	$TLMTA$	$-(EXRETAVG)$	MB	$-(PRICE)$	$SIGMA$
				Panel A: Non-default group				
Count	2 813 602							
Mean	−0.094	0.000	−10.441	0.443	−0.001	1.978	−0.874	0.531
Std. dev.	0.105	0.017	1.933	0.284	0.015	1.460	0.439	0.325
Min.	−0.385	−0.020	−13.700	0.037	−0.030	0.380	−1.180	0.164
25%-qu.	−0.130	−0.011	−11.930	0.190	−0.010	0.950	−1.180	0.284
50%-qu.	−0.052	−0.003	−10.520	0.416	−0.001	1.520	−1.110	0.437
75%-qu.	−0.018	0.004	−9.030	0.683	0.008	2.530	−0.690	0.686
Max.	−0.002	0.046	−6.810	0.926	0.029	5.940	2.110	1.350
				Panel B: Default group				
Count	2,694							
Mean	−0.054	0.027	−12.319	0.747	0.010	2.947	−0.075	1.028
Std. dev.	0.075	0.021	1.617	0.235	0.023	2.422	0.624	0.380
Min.	−0.385	−0.020	−13.700	0.037	−0.030	0.380	−1.180	0.164
25%-qu.	−0.066	0.009	−13.700	0.630	−0.011	0.570	−0.490	0.735
50%-qu.	−0.026	0.035	−12.910	0.864	0.023	1.890	−0.050	1.227
75%-qu.	−0.009	0.046	−11.470	0.926	0.029	5.940	0.360	1.350
Max.	−0.002	0.046	−6.810	0.926	0.029	5.940	1.810	1.350
				Panel C: T-test statistics for the difference in means of the default and non-default group				
T-statistic	19.58***	82.33***	−50.40***	55.34***	35.93***	34.39***	94.28***	79.31***

Notes: This table shows the summary statistics for the features used in the PD models. The count, mean, standard deviation (std. dev.), minimum (min.), 25%-quantile (25%-qu.), median (50%-qu.), 75%-quantile (75%-qu.), and maximum value (max.) are shown for the non-default group (Panel *A*) and for the default group (Panel *B*). Panel *C* shows the statistics of the t-test for the difference in means of the default and non-default group (see Welch (1947)). The significance is indicated for the 1% (***), 5% (**) and 10% (*) level. Potential outliers are controlled by winsorizing each variable at the 5th and 95th percentile of its distribution.

1.5 Empirical Results

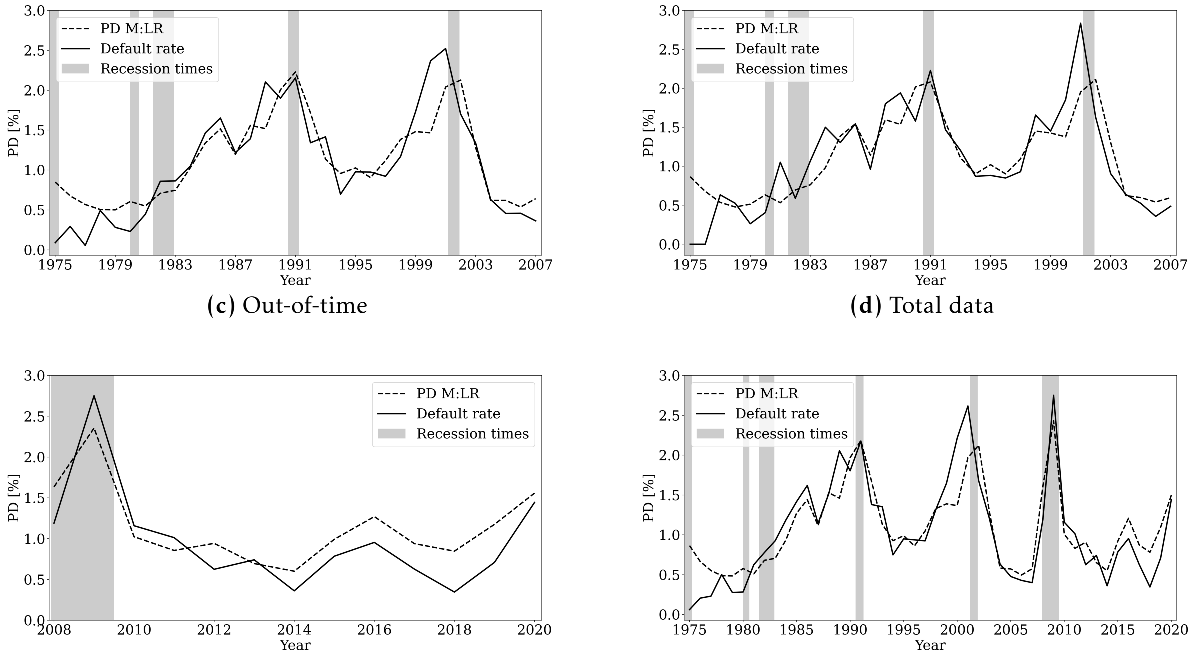
1.5.1 Logistic Regression

Logistic regression is a very common method for predicting default probabilities. Therefore we use a logit model as benchmark model. Figure 1.5 shows the time series of the mean default rate (solid line) and the mean probability of default predicted by the logit model (dashed line) for the different data samples and a one-month default time-lag. The number of corporate defaults increases during recessions (gray-shaded areas) and was particularly noticeable during the Global Financial Crisis.

Figure 1.5: Logistic regression fit for default horizon of 1 month

(**a**) In-sample

(**b**) Out-of-sample



(**c**) Out-of-time

(**d**) Total data

Notes: This figure summarizes the fit of the PD model *LR* for the default time-lag of 1 month: (a) in-sample data set, (b) out-of-sample data set, (c) out-of-time data set, (d) total data. The dashed lines correspond to the logistic regression models (fitted to each data set) and the solid lines represent the time series of the default rate. The shaded areas (gray bars) indicate recessions as defined by the National Bureau of Economic Research.

Table 1.3 shows the coefficients when applying the base model *LR* on the different data samples of this study: IS, OS, OT, and the total data set. All coefficients of the sub-models are significant at the 1%-level and comparable to Campbell et al. (2008) in sign and size[3]. The signs of the

[3] There is an ongoing debate about whether p-values are appropriate because they shrink towards zero as the number of observations increases (see Demidenko (2016)). We analyze the importance of features by drawing ALE plots in Section 1.5.3.

coefficients of all variables (including the intercept) are constant for the different samples considered, showing a consistent influence of the features. All financial ratios are defined as financial frictions and have a positive sign, so that probability of default increases with increasing feature values. Lower cash holdings ($CASHMTA$), net incomes ($NIMTAAVG$), excess stock returns ($EXRETAVG$), and stock prices ($PRICE$) increase the default risk. Higher market capitalizations ($RSIZE$), leverages ($TLMTA$), market-to-book equity values (MB), and volatility of stock returns ($SIGMA$) increase the default.

Table 1.3: Coefficient table of the logistic regression model for the one-month time-lag: in-sample, out-of-sample, out-of-time, and entire data set

		LR			
	Sample name	In-sample	Out-of-sample	Out-of-time	Total data
	Sample period	1975-2007	1975-2007	2008-2020	1975-2020
Fundamental features	$-(CASHMTA)$	4.846	3.144	3.304	4.259
		(0.420)***	(0.557)***	(0.520)***	(0.279)***
	$-(NIMTAAVG)$	35.313	38.189	10.485	30.443
		(1.831)***	(2.828)***	(2.773)***	(1.331)***
	$RSIZE$	0.111	0.123	0.147	0.123
		(0.024)***	(0.038)***	(0.034)***	(0.017)***
	$TLMTA$	3.433	3.563	4.005	3.570
		(0.119)***	(0.182)***	(0.222)***	(0.090)***
Market features	$-(EXRETAVG)$	18.985	14.911	8.315	16.200
		(1.497)***	(2.202)***	(2.349)***	(1.093)***
	MB	0.141	0.163	0.167	0.152
		(0.012)***	(0.019)***	(0.018)***	(0.009)***
	$-(PRICE)$	0.682	0.742	1.304	0.821
		(0.068)***	(0.104)***	(0.107)***	(0.050)***
	$SIGMA$	1.325	1.455	1.710	1.474
		(0.098)***	(0.153)***	(0.149)***	(0.072)***
	Intercept	−8.734	−8.989	−8.633	−8.764
		(0.306)***	(0.483)***	(0.443)***	(0.221)***
	# observations	1,448,522	620,796	744,284	2,813,602
	# defaults	1,458	622	614	2,694
	Pseudo-R^2	0.2334	0.2512	0.2454	0.2376
	AUC	0.9066	0.9215	0.8866	0.9045
	Brier Score (‰)	0.9893	0.9831	0.8071	0.9403

Notes: This table shows the coefficients of the logistic regression model for the one-month default time-lag based on different samples: in-sample, out-of-sample, out-of-time, and the entire data set. The standard errors are given in parentheses. The significance is indicated for the 1% (***), 5% (**) and 10% (*) level.

Our study also focuses on the analysis of default risk over different default time horizons. Table 1.4 summarizes the results for the base model *LR* applied to the in-sample data set and different time-lags for the default indicator: 1, 6, 12, and 24 months[4].

[4] The coefficients of the logistic regression for longer time-lags of 36, 48, or 60 months are consistent with 24 months in sign, size, and significance level. The number of observations drops with higher order of lags.

Table 1.4: Coefficient table of the logistic regression model for the in-sample data set: 1, 6, 12, and 24 months time-lag

	LR	Time-lag in months			
		1	6	12	24
Fundamental features	$-(CASHMTA)$	4.846	4.244	3.695	3.067
		(0.420)***	(0.384)***	(0.361)***	(0.353)***
	$-(NIMTAAVG)$	35.313	30.699	27.590	18.298
		(1.831)***	(1.755)***	(1.766)***	(1.807)***
	$RSIZE$	0.111	0.106	0.076	0.026
		(0.024)***	(0.023)***	(0.022)***	(0.020)***
	$TLMTA$	3.433	2.961	2.315	1.798
		(0.119)***	(0.109)***	(0.104)***	(0.102)***
Market features	$-(EXRETAVG)$	18.985	18.210	22.485	13.156
		(1.497)***	(1.514)***	(1.629)***	(1.653)***
	MB	0.141	0.129	0.117	0.148
		(0.012)***	(0.013)***	(0.015)***	(0.016)***
	$-(PRICE)$	0.682	0.503	0.276	0.053
		(0.068)***	(0.070)***	(0.074)***	(0.080)***
	$SIGMA$	1.325	1.384	1.219	0.936
		(0.098)***	(0.096)***	(0.097)***	(0.100)***
	Intercept	−8.734	−8.339	−8.165	−8.218
		(0.306)***	(0.286)***	(0.280)***	(0.269)***
	# observations	1,448,522	1,441,224	1,432,473	1,418,851
	# defaults	1,458	1,470	1,426	1,434
	Pseudo-R^2	0.2334	0.1723	0.1188	0.0551
	AUC	0.9066	0.8831	0.8324	0.7459
	Brier Score (‰)	0.9893	1.0116	0.9902	1.0082

Notes: This table shows the coefficients of the logistic regression model for the in-sample data set based on different time-lags (months): 1, 6, 12, and 24 months. The standard errors are given in parentheses. The significance is indicated for the 1% (***), 5% (**) and 10% (*) level.

The coefficients of all variables (including the intercept) show constant signs over different time-lags, documenting a consistent influence of financial frictions over the time-lags. Validation metrics including the AUC and Brier Score may not be compared for different time-lags as they depend on the distribution of the independent variable (see, e.g., Blochwitz et al. (2005)). Since the default indicator is lagged for different horizons, the number of observations and defaults can vary between different lags due to the data structure.

1.5.2 Machine Learning Models

Capturing Uni-variate Non-linear Relations

Performance measures for the various PD models depend strongly on the impact of financial frictions on default risk. A uni-variate approach is taken by plotting the mean default rate against the binned values of the individual characteristics. We find strong non-linearity in these relations (see circle markers in Figures 1.6 and 1.7 indicating mean default risk). The results are important for the analysis of models based on all features, as they show how ML methods can model non-linearity, which enhances their predictive power. To visualize this, the following figures include the fit of uni-variate random forest models (solid lines) and uni-variate logistic regressions (dashed lines).

Figure 1.6 focuses on fundamental features. The upper left figure shows that the one-month average default rate curve for lower cash holdings ($-(CASHMTA)$) has an increasing convex shape (circle markers). Companies with a low level of cash holdings have liquidity constraints and high default risk. Acharya et al. (2012) confirm a negative relation for short prediction horizons and that it may turn to positive as prediction horizons become longer, arguing that higher liquidity may also be sign of higher credit risk due to cash hoarding. The non-linearity decreases with longer default horizons. Likewise, the relation between features and default rates is convex and the default rates peak for low profitability ($-(NIMTAAVG)$), small relative market capitalization ($RSIZE$), and high leverage ($TLMTA$). These relations have been well documented in prior literature. Note that $RSIZE$ has a negative relation to default in the uni-variate setting but changes to the positive in a multi-variable setting (see Subsection 1.5.3). Using datasets which are smaller in the time series, Campbell et al. (2008) find a positive relation of $RSIZE$ to default risk for short horizons and a negative one for time horizons greater than 12 months. Also, Aretz et al. (2018) find mixed results for different countries.

Figure 1.7 focuses on market features. The upper left figure shows that the one-month average default rate for negative excess returns ($-(EXRETAVG)$) follows a smirk shape (circle markers). Companies with lower average excess returns have higher default rates. Interestingly, default rates also increase for high excess returns as the risk increases. The same can be observed for market-to-book ratio MB, which measures efficiency and growth prospects. Low market-to-book firms are low efficiency firms with high default risk. High market-to-book firms are often start-ups with a high potential to grow but also to fail. Griffin and Lemmon (2002), in conjunction with stock returns, find increased distress risk for firms both with low and high MB values. Likewise, default rates increase with lower share prices ($-(PRICE)$). The logarithmic stock price per share truncated at a cap of \$15 and multiplied by minus one is a measure of proximity to default. When stock prices fall, companies are closer to default, and very low stock prices indicate that the distance to default is short. The standard deviation $SIGMA$ measures the daily stock return over the previous three months and captures the recent volatility.

A low volatility may indicate a limited trading and hence, ability to refinance. A moderate volatility may indicate the ability of a firm to outperform the reference index on the stock market. A high volatility may indicate a greater likelihood of experiencing financial distress. As the default time-lag increases, the non-linearity decreases as the circle markers lie more and more on a straight line. While ML techniques dominate in most cases, their relative dominance decreases over time due to the shrinkage of linearity. In other words, ML techniques excel when information is collected and processed in a timely fashion.

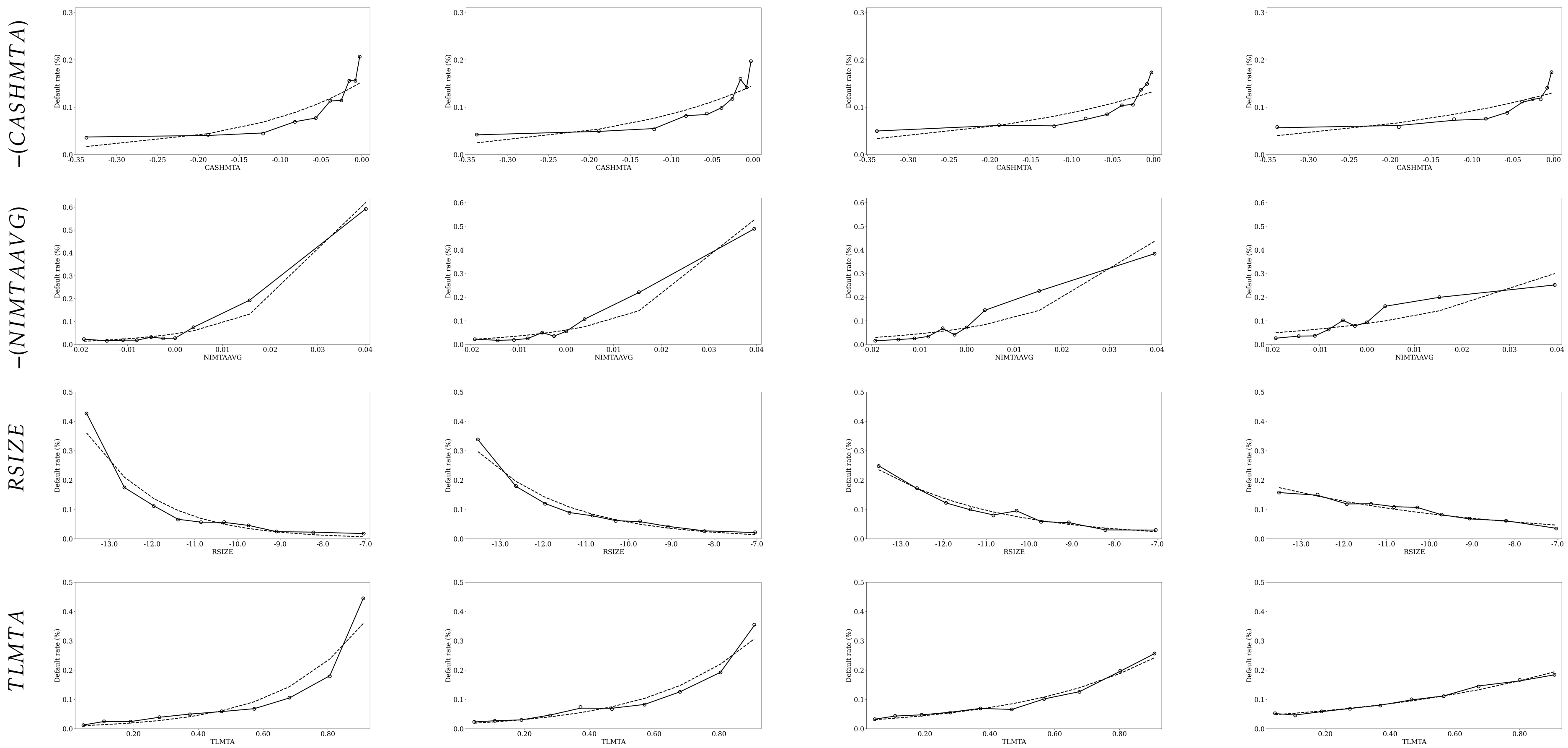
The logistic regression model (dashed line) fails to capture this non-linear pattern and the random forest model (solid line) provides a better fit. Results and interpretations remain the same for the uni-variate neural network model and the uni-variate (extreme) gradient boosting model and are available upon request.

Hyper-parameter Selection for Multi-variable Models

We now provide the multi-variable analyses using ML approaches. The first step in initializing the ML models is to cross-validate in-sample data to find the optimal combination of hyper-parameters. We then predict out-of-time PDs. Panel A of Table 1.5 lists the parameter grid for common hyper-parameters that are validated for the in-sample data set of every feature set on each of the applied models. For example, the structure of a neural network is mainly determined by the number and size of hidden layers. Thus, (16, 8) refers to an NN with two hidden layers of 16 and 8 neurons, respectively (see also Figure 1.2). In addition, the degree of regularization, which is determined by parameters such as (initial) learning rate, maximum depth of decision trees, or L1 and L2 regularization, is cross-validated to counteract possible over-fitting of the methods.

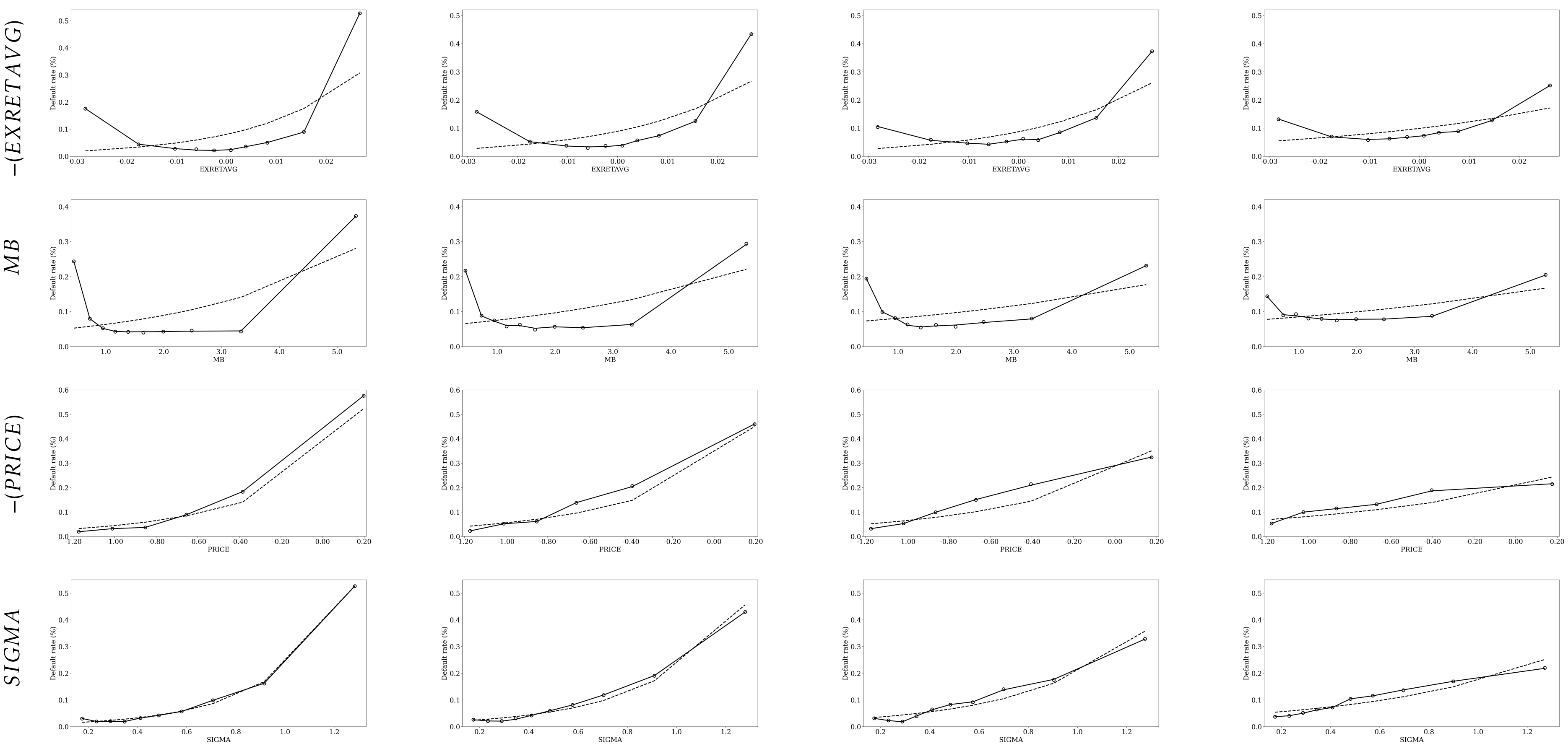
The hyper-parameter combinations resulting from the optimization based on the total feature set, the fundamental feature set, and market feature set are illustrated in Panel B of Table 1.5. The size of the NN architecture (number of layers and neurons) increases with longer default lags for the total and fundamental features, while the network structure remains comparatively constant for the market features. The number of decision trees combined into ensembles shows no obvious trend for different methods (*GB*, *XGB*, *RF*), horizons, or feature sets. This suggests that the structure of tree-based methods depends on the type of ensemble, i.e., bagging or boosting, and the overall hyper-parameter interaction of the methods in question.

Figure 1.6: Non-linearity analysis for uni-variate in-sample fit: fundamental features by time-lag



Notes: This figure shows the realized default rate (circle markers) and the predicted probabilities of default from uni-variate logistic regression (dashed lines) and from the uni-variate random forest model (solid lines), for the fundamental features of the in-sample data set from 1975 to 2007. The columns contain the subplots for a default time-lag of 1 month, 6 months, 12 months, and 24 months. The subplots show that the strength of the non-linear relation between the variables and the default rate decreases with increasing default time-lag.

Figure 1.7: Non-linearity analysis for uni-variate in-sample fit: market features by time-lag



Notes: This figure shows the realized default rate (circle markers) and the predicted probabilities of default from uni-variate logistic regression (dashed lines) and from the uni-variate random forest model (solid lines), for the market features of the in-sample data set from 1975 to 2007. The columns contain the subplots for a default time-lag of 1 month, 6 months, 12 months, and 24 months. The subplots show that the strength of the non-linear relation between the variables and the default rate decreases with increasing default time-lag.

Table 1.5: Hyper-parameter optimization

Panel A: Parameter grid for selected hyper-parameters

Model	Hyper-parameter	Parameter grid	Model	Hyper-parameter	Parameter grid
NN	size hidden layers	[(8), (16), (32), (64), (128), (16,8), (32,16), (64,32), (128,64), (32,16,8), (64,32,16), (128,64,32)]	*XGB*	# estimators	[50, 100, 150, 200, 250, 300, 350, 400, 450, 500]
	initial learning rate	[0.0025, 0.005, 0.0075, 0.01, 0.1]		learning rate	[0.01, 0.1, 0.2, 0.5, 1]
	alpha	[0, 0.000001, 0.00001, 0.0001, 0.001]		max. depth	[3, 4, 5, 10, 20, 30]
	activation function	[logistic, hyperbolic tan: tanh, rectified linear unit: relu]		gamma	[0.1, 0.01, 1/8]
	solver	[stochastic gradient descent: sgd, optimized sgd: Adam]		regularization L2	[0, 0.1, 0.5, 1.0, 5.0, 10.0]
				regularization L1	[0, 0.1, 0.5, 1.0, 5.0, 10.0]
GB	# estimators	[50, 100, 150, 200, 250, 300, 350, 400, 450, 500]			
	learning rate	[0.01, 0.1, 0.2, 0.5, 1]	*RF*	# estimators	[50, 100, 150, 200, 250, 300, 350, 400, 450, 500]
	max. depth	[3, 4, 5, 10, 20, 30]		max. depth	[2, 3, 5, 7, 10, None]
	max. # leaf nodes	[2, 3, 5, 10, 20]		min. samples split	[2, 3, 5, 7, 10, 11, 12, 15, 20]

Panel B: Selected hyper-parameter values from grid search

Model	Hyper-parameter	Total features				Fundamental features				Market features			
		1	6	12	24	1	6	12	24	1	6	12	24
NN	size hidden layers	(64,32)	(64,32)	(128,64,32)	(128,64)	(32)	(32,16,8)	(64,32,16)	(128,64,32)	(64,32)	(16,8)	(32,16)	(16,8)
	initial learning rate	0.1	0.0025	0.1	0.1	0.1	0.1	0.1	0.1	0.0025	0.1	0.1	0.1
	alpha	0.001	0.00001	0.001	0.000001	0.00001	0.000001	0.001	0.001	0.0001	0	0	0
	activation function	tanh	tanh	relu	relu	tanh	tanh	tanh	relu	tanh	tanh	tanh	tanh
	solver	sgd	adam	sgd	sgd	sgd	sgd	sgd	sgd	adam	sgd	sgd	sgd
GB	# estimators	400	350	350	350	400	100	300	150	450	300	400	400
	learning rate	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.01	0.01
	max. depth	3	10	10	10	3	3	10	4	3	10	3	3
	max. # leaf nodes	3	3	3	3	5	20	3	3	5	3	20	20
XGB	# estimators	400	350	150	200	350	200	100	350	400	100	100	100
	learning rate	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1
	max. depth	3	5	5	4	5	3	3	5	20	3	3	5
	gamma	0.01	0.125	0.125	0.1	0.125	0.1	0.01	0.125	0.125	0.01	0.125	0.01
	regularization L2	1	0	10	0.1	0	5	10	0	5	10	0.5	10
	regularization L1	5	10	0.5	10	10	0	1	10	10	1	0.5	10
RF	# estimators	200	450	450	450	300	100	250	450	400	150	100	250
	max. depth	10	10	7	7	7	7	7	7	7	5	5	5
	min. samples split	3	11	7	2	12	15	3	5	7	12	3	11

Notes: This table shows a selection of common parameters that are subject of grid search and tuning of ML model hyper-parameters, while a broader set of parameters was tested. For every model applied, the best parameter combination for the in-sample data set per feature set (fundamental features: $-(CASHMTA)$, $-(NIMTAAVG)$, $RSIZE$, $TLMTA$; market features: $-(EXRETAVG)$, MB, $-(PRICE)$, $SIGMA$) and time-lag (1 month, 6 months, 12 months, 24 months) is presented. The grid search was randomized with 300 iterations and 5-fold cross-validation. The AUC measure was used to evaluate the best fit.

Evaluation of Predictive Performance

After tuning the hyper-parameters, all models are applied on the different data sets in this study, PD predictions are made, and the predictive power and calibration of the models is evaluated. Validation results of model estimation based on the in-sample, out-of-sample, and out-of-time data set are shown in Table 1.6, Table 1.7, and Table 1.8 respectively[5].

The relative improvements in AUC (ΔAUC) and Brier Score (ΔBrier) of the ML methods demonstrate a better fit to the in-sample data across almost all horizons and feature sets than the logistic regression.

When validated on the out-of sample data, ML methods provide improved predictive power, especially for market features. Both *XGB* and *RF* show a 3.8% higher AUC for the market set at a default time-lag of 1 month and improved calibration (lower Brier Scores). The predominance of these methods can also be observed for the remaining feature sets and default time horizons. The out-of-sample validation can be viewed as a robustness check of the generalization of the models and a functional check of over-fitting control.

The focus of this paper is also on out-of-time validation (see Table 1.8), since PD prediction into the future is of major concern for academics and practitioners. Tree-based models with the total feature set outperform logistic regression significantly in terms of AUC for all default horizons: the *GB*, *XGB*, and *RF* models result in an AUC improvement of 3.3 to 4.1 percent for the default prediction for the one-month lag. In summary, we confirm our hypothesis that ML methods dominate logistic regressions in corporate default predictions due to non-linear relations between fundamental and market features and default outcomes.

ML models are best for the market feature set, with a maximum AUC improvement of 6.6% for *GB* and *RF* compared with *LR* (24-month horizon). The *NN* model performs best on the market features, i.e., between 2.0% and 4.5% improvement of AUC. This confirms our hypothesis that the ML method dominance is more pronounced for shorter default horizons as the non-linearity decreases as information fades over time.

[5] We tested other ML methods including lasso and ridge regularized logistic regression, k-nearest neighbor method, and stand-alone decision trees, with consistent results. We thus restrict the long list of methods to the models displayed. Validation results for models based on longer time-lags of 36, 48, or 60 months are consistent with the performance of PD models for the 24-month horizon. All analyses are available upon request.

Table 1.6: Prediction model performance: in-sample fit

		Total features				Fundamental features				Market features			
Lag	Model	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier
1 month	*LR*	0.9066		0.9893		0.8894		0.9926		0.8533		0.9981	
	NN	0.9233 ***	1.8%	0.9894	-1 bp	0.8917 ***	0.3%	0.9933	-7 bp	0.8866 ***	3.9%	0.9982	-1 bp
	GB	0.9392 ***	3.6%	0.9792 ***	102 bp	0.9139 ***	2.8%	1.0603	-682 bp	0.9089 ***	6.5%	0.9884 ***	97 bp
	XGB	0.9564 ***	5.5%	0.9739 ***	155 bp	0.9198 ***	3.4%	0.9869 ***	58 bp	0.9197 ***	7.8%	0.9897 ***	84 bp
	RF	0.9938 ***	9.6%	0.9094 ***	807 bp	0.9517 ***	7.0%	0.9841 ***	86 bp	0.9379 ***	9.9%	0.9838 ***	143 bp
6 months	*LR*	0.8831		1.0116		0.8560		1.0123		0.8191		1.0153	
	NN	0.8943 ***	1.3%	1.0108 *	8 bp	0.8580 ***	0.2%	1.0128	-5 bp	0.8434 ***	3.0%	1.0143 ***	9 bp
	GB	0.9072 ***	2.7%	1.0057 ***	58 bp	0.8831 ***	3.2%	1.0088 ***	34 bp	0.8579 ***	4.7%	1.0120 ***	32 bp
	XGB	0.9397 ***	6.4%	0.9961 ***	154 bp	0.8928 ***	4.3%	1.0096 ***	26 bp	0.8700 ***	6.2%	1.0115 ***	37 bp
	RF	0.9944 ***	12.6%	0.9476 ***	633 bp	0.9298 ***	8.6%	1.0058 ***	64 bp	0.8652 ***	5.6%	1.0119 ***	33 bp
12 months	*LR*	0.8324		0.9902		0.8066		0.9912		0.7759		0.9925	
	NN	0.8502 ***	2.1%	0.9898 *	4 bp	0.8112 ***	0.6%	0.9911	1 bp	0.7868 ***	1.4%	0.9926	-1 bp
	GB	0.8694 ***	4.4%	0.9870 ***	33 bp	0.8355 ***	3.6%	0.9898 ***	14 bp	0.8122 ***	4.7%	0.9905 ***	20 bp
	XGB	0.9142 ***	9.8%	0.9829 ***	74 bp	0.8512 ***	5.5%	0.9896 ***	16 bp	0.8170 ***	5.3%	0.9911 ***	14 bp
	RF	0.9329 ***	12.1%	0.9832 ***	70 bp	0.9246 ***	14.6%	0.9861 ***	51 bp	0.8206 ***	5.8%	0.9905 ***	20 bp
24 months	*LR*	0.7459		1.0082		0.7202		1.0085		0.6833		1.0091	
	NN	0.7764 ***	4.1%	1.0085	-3 bp	0.7262 ***	0.8%	1.0086	-1 bp	0.6980 ***	2.2%	1.0091	0 bp
	GB	0.8066 ***	8.1%	1.0062 ***	20 bp	0.7612 ***	5.7%	1.0078 ***	7 bp	0.7286 ***	6.6%	1.0081 ***	10 bp
	XGB	0.8383 ***	12.4%	1.0055 ***	27 bp	0.8049 ***	11.8%	1.0068 ***	17 bp	0.7385 ***	8.1%	1.0084 ***	7 bp
	RF	0.9107 ***	22.1%	1.0023 ***	58 bp	0.9042 ***	25.6%	1.0044 ***	41 bp	0.7478 ***	9.4%	1.0079 ***	12 bp

Notes: This table summarizes the performance of the prediction models for the in-sample data set. AUC, Brier Score, and the corresponding improvements (ΔAUC and ΔBrier) of the machine learning models versus *LR* are indicated. The metrics are shown for different time-lags of 1, 6, 12, and 24 months. Models contain a set of fundamental features ($-(CASHMTA)$, $-(NIMTAAVG)$, $RSIZE$, $TLMTA$), a set of market features ($-(EXRETAVG)$, MB, $-(PRICE)$, $SIGMA$), or all the considered features. The significance for the test to ascertain whether the AUC (Brier Score) of the model is higher (lower) than the AUC (Brier Score) of the benchmark model *LR* is indicated for the 1% (***), 5% (**) and 10% (*) level.

Table 1.7: Prediction model performance: out-of-sample fit

		Total features				Fundamental features				Market features			
Lag	Model	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier
1 month	*LR*	0.9205		0.9831		0.9024		0.9877		0.8753		0.9923	
	NN	0.9321 ***	1.3%	0.9852	-21 bp	0.9026	0.0%	0.9883	-7 bp	0.8998 ***	2.8%	0.9924	-2 bp
	GB	0.9392 ***	2.0%	0.9840	-9 bp	0.9057	0.4%	1.0716	-850 bp	0.9066 ***	3.6%	0.9950	-27 bp
	XGB	0.9445 ***	2.6%	0.9822	9 bp	0.9098 ***	0.8%	0.9877	0 bp	0.9082 ***	3.8%	0.9880 ***	43 bp
	RF	0.9396 ***	2.1%	0.9876	-46 bp	0.9123 ***	1.1%	0.9873	4 bp	0.9084 ***	3.8%	0.9884 ***	39 bp
6 months	*LR*	0.8831		0.9853		0.8622		0.9851		0.8206		0.9885	
	NN	0.8915 ***	1.0%	0.9837 **	16 bp	0.8635	0.2%	0.9856	-5 bp	0.8416 ***	2.6%	0.9875 ***	10 bp
	GB	0.9008 ***	2.0%	0.9861	-9 bp	0.8735 ***	1.3%	0.9857	-6 bp	0.8515 ***	3.8%	0.9875 *	9 bp
	XGB	0.9058 ***	2.6%	0.9856	-3 bp	0.8767 ***	1.7%	0.9852	-1 bp	0.8544 ***	4.1%	0.9872 ***	13 bp
	RF	0.9019 ***	2.1%	0.9859	-7 bp	0.8737 ***	1.3%	0.9852	0 bp	0.8492 ***	3.5%	0.9868 ***	17 bp
12 months	*LR*	0.8349		1.0517		0.8100		1.0524		0.7658		1.0544	
	NN	0.8481 ***	1.6%	1.0512	5 bp	0.8121	0.3%	1.0523	2 bp	0.7828 ***	2.2%	1.0544	0 bp
	GB	0.8518 ***	2.0%	1.0510 *	7 bp	0.8206 ***	1.3%	1.0517 ***	7 bp	0.7878 ***	2.9%	1.0543	1 bp
	XGB	0.8592 ***	2.9%	1.0509 **	8 bp	0.8236 ***	1.7%	1.0516 ***	8 bp	0.7908 ***	3.3%	1.0538 ***	5 bp
	RF	0.8544 ***	2.3%	1.0509 **	8 bp	0.8276 ***	2.2%	1.0516 ***	8 bp	0.7856 ***	2.6%	1.0537 ***	7 bp
24 months	*LR*	0.7552		0.9504		0.7195		0.9504		0.6831		0.9508	
	NN	0.7718 ***	2.2%	0.9513	-9 bp	0.7249 **	0.8%	0.9507	-3 bp	0.6909	1.1%	0.9508	0 bp
	GB	0.7757 ***	2.7%	0.9506	-2 bp	0.7366 ***	2.4%	0.9503	1 bp	0.7050 ***	3.2%	0.9506	2 bp
	XGB	0.7844 ***	3.9%	0.9502	1 bp	0.7413 ***	3.0%	0.9502 *	2 bp	0.7062 ***	3.4%	0.9506 **	2 bp
	RF	0.7791 ***	3.2%	0.9500 ***	4 bp	0.7419 ***	3.1%	0.9502 ***	2 bp	0.7068 ***	3.5%	0.9505 ***	3 bp

Notes: This table summarizes the performance of the prediction models for the out-of-sample data set. AUC, Brier Score, and the corresponding improvements (ΔAUC and ΔBrier) of the machine learning models versus *LR* are indicated. The metrics are shown for different time-lags of 1, 6, 12, and 24 months. Models contain a set of fundamental features ($-(CASHMTA)$, $-(NIMTAAVG)$, $RSIZE$, $TLMTA$), a set of market features ($-(EXRETAVG)$, MB, $-(PRICE)$, $SIGMA$), or all the considered features. The significance for the test to ascertain whether the AUC (Brier Score) of the model is higher (lower) than the AUC (Brier Score) of the benchmark model *LR* is indicated for the 1% (***), 5% (**) and 10% (*) level.

Table 1.8: Prediction model performance: out-of-time fit

		Total features				Fundamental features				Market features			
Lag	Model	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier	AUC	ΔAUC	Brier(‰)	ΔBrier
1 month	*LR*	0.8804		0.8086		0.8638		0.8153		0.8375		0.8147	
	NN	0.8955 ***	1.7%	0.8104	-22 bp	0.8651	0.2%	0.8163	-13 bp	0.8749 ***	4.5%	0.8167	-25 bp
	GB	0.9093 ***	3.3%	0.8080	7 bp	0.8730 ***	1.1%	0.8594	-541 bp	0.8874 ***	6.0%	0.8172	-31 bp
	XGB	0.9168 ***	4.1%	0.8060 *	31 bp	0.8763 ***	1.4%	0.8121 ***	40 bp	0.8911 ***	6.4%	0.8118 ***	35 bp
	RF	0.9150 ***	3.9%	0.8047 ***	48 bp	0.8785 ***	1.7%	0.8114 ***	49 bp	0.8901 ***	6.3%	0.8122 ***	30 bp
6 months	*LR*	0.8708		0.8067		0.8475		0.8091		0.8173		0.8099	
	NN	0.8763 **	0.6%	0.8071	-5 bp	0.8451	-0.3%	0.8093	-2 bp	0.8395 ***	2.7%	0.8090 ***	12 bp
	GB	0.8862 ***	1.8%	0.8066	1 bp	0.8550 ***	0.9%	0.8074 ***	21 bp	0.8515 ***	4.2%	0.8091 *	11 bp
	XGB	0.8983 ***	3.2%	0.8052 **	18 bp	0.8542 **	0.8%	0.8071 ***	25 bp	0.8621 ***	5.5%	0.8086 ***	16 bp
	RF	0.8979 ***	3.1%	0.8042 ***	31 bp	0.8552 **	0.9%	0.8070 ***	26 bp	0.8524 ***	4.3%	0.8084 ***	19 bp
12 months	*LR*	0.8405		0.7594		0.8177		0.7604		0.7744		0.7608	
	NN	0.8368	-0.4%	0.7589 **	7 bp	0.8128	-0.6%	0.7602 **	3 bp	0.7897 ***	2.0%	0.7611	-3 bp
	GB	0.8520 ***	1.4%	0.7586 **	11 bp	0.8188	0.1%	0.7592 ***	15 bp	0.8069 ***	4.2%	0.7606	3 bp
	XGB	0.8624 ***	2.6%	0.7585 **	12 bp	0.8233	0.7%	0.7592 ***	15 bp	0.8107 ***	4.7%	0.7602 ***	8 bp
	RF	0.8546 ***	1.7%	0.7582 ***	16 bp	0.8213	0.4%	0.7593 ***	13 bp	0.8079 ***	4.3%	0.7603 ***	7 bp
24 months	*LR*	0.7692		0.5647		0.7580		0.5649		0.6925		0.5651	
	NN	0.7864 ***	2.2%	0.5662	-26 bp	0.7540	-0.5%	0.5651	-3 bp	0.7154 ***	3.3%	0.5649 ***	2 bp
	GB	0.8064 ***	4.8%	0.5646	1 bp	0.7768 ***	2.5%	0.5647 **	4 bp	0.7383 ***	6.6%	0.5649 *	3 bp
	XGB	0.7998 ***	4.0%	0.5646	3 bp	0.7796 ***	2.9%	0.5649	1 bp	0.7366 ***	6.4%	0.5651	0 bp
	RF	0.8058 ***	4.8%	0.5644 ***	6 bp	0.7815 ***	3.1%	0.5647 ***	4 bp	0.7384 ***	6.6%	0.5650 *	1 bp

Notes: This table summarizes the performance of the prediction models for the out-of-time data set. AUC, Brier Score, and the corresponding improvements (ΔAUC and ΔBrier) of the machine learning models versus *LR* are indicated. The metrics are shown for different time-lags of 1, 6, 12, and 24 months. Models contain a set of fundamental features ($-(CASHMTA)$, $-(NIMTAAVG)$, $RSIZE$, $TLMTA$), a set of market features ($-(EXRETAVG)$, MB, $-(PRICE)$, $SIGMA$), or all the considered features. The significance for the test to ascertain whether the AUC (Brier Score) of the model is higher (lower) than the AUC (Brier Score) of the benchmark model *LR* is indicated for the 1% (***), 5% (**) and 10% (*) level.

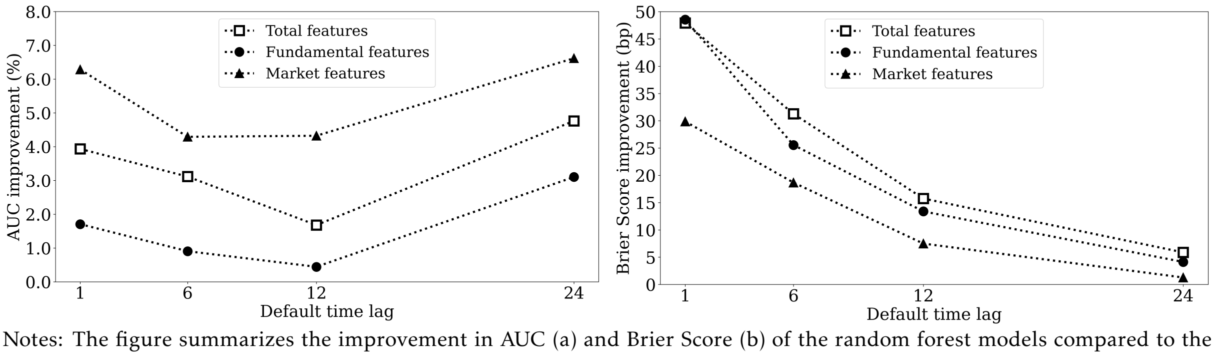
These methods also demonstrate increased prediction power for time-lags of 6 and 12 months. The neural network model *NN* is only slightly better. For the fundamental features, the improvement in AUC of ML methods is limited. While small increases are visible for the 1-month and 6-month lag, no significant AUC improvement on the 12-month horizon can be found. However, these models show significantly better calibration. The *RF* model provides the greatest significant AUC and Brier Score improvements for a 24-month horizon. This confirms our hypothesis that the dominance of ML methods is more pronounced for shorter default horizons as the non-linearity decreases when information fades over time.

Figure 1.8 summarizes the out-of-time AUC and Brier Score improvements for the total features, the market features, and the fundamental features. The random forest models consistently outperform the logistic regression models across all feature sets and default time horizons in terms of AUC and Brier Score. The increase in AUC is particularly large for the market features at both short and long horizons. The random forest model is better able to distinguish between defaults and non-defaults, especially at short horizons where logistic regression is quite competitive. However, the relative AUC increase for longer periods such as 24 months may be due to some residual non-linearity, and discrimination power of the random forest model may not decrease as quickly as with logistic regression. The improvement in model calibration (Brier Score) decreases with an increasing time horizon for all feature sets. This suggests that PD predictions from random forests and logistic regressions both lose calibration quality at the same rate for longer horizons such as 24 months.

Figure 1.8: Out-of-time AUC and Brier Score improvement

(**a**) AUC improvement (%)

(**b**) Brier Score improvement (bp)



Notes: The figure summarizes the improvement in AUC (a) and Brier Score (b) of the random forest models compared to the logistic regression models for different feature sets and default horizons in the out-of-time sample. The model improvements are illustrated for the total features (squares), the fundamental features (circles), and the market features (triangles).

1.5.3 Robustness Check: Non-linearity and Feature Importance of Financial Frictions

Results from both the single-feature and feature-set-based models indicate the presence of non-linearity in the relation between the predicted PDs and the features. To measure the degree of non-linearity of our multi-variable models[6], we apply the concept of accumulated local effects (ALE, see Apley and Zhu (2020)) to our PD prediction framework.

Empirically, we find that variable $RSIZE$ changes the direction of impact in a multi-variable setting as features are correlated. As a result, the uni-variate plots from Figures 1.6 and 1.7 may be misleading as the conditionality of the feature value (note feature value is applied to all observations) and correlations between features are not considered. The presence of feature correlations is also a problem for methods for graphical explanation of feature importance, such as partial dependence plots (see Friedman (2001)). To address these issues, we generate ALE plots following Apley and Zhu (2020) that calculate differences in predictions based on all features for the upper and lower bound of an interval around the considered feature value (local effects). The ALE curve is constructed by accumulating these differences and centering at zero. As a result, each point on the ALE curve (y-axis value) represents the respective difference to the mean prediction[7]. In that way, ALE plots can be interpreted as the influence of a particular feature on the predictions based on the full model, i.e., multiple input features enter the model (”multi-variable case”). This influence is measured over the entire range of feature values, so that differences in the direction and strength of the impact become visible.

For quantification of non-linear effects of a single feature on the linear predictor, an inverse transformation of the default probabilities is required. We apply a logistic inverse transformation so that LR becomes completely linear and we can interpret the feature weights of ML techniques relative to LR[8]. While the Figures 1.6 and 1.7 analyze non-linearity and the relation between the default rate and a single feature (uni-variate case), the ALE plots capture the direction and degree of influence of a feature on the model outcome based on all variables, i.e., multi-variable relations are considered.

[6] Note that Figures 1.6 and 1.7 are based on a uni-variate setting.

[7] For the detailed mathematical description and derivation of ALE plots, we refer to Apley and Zhu (2020).

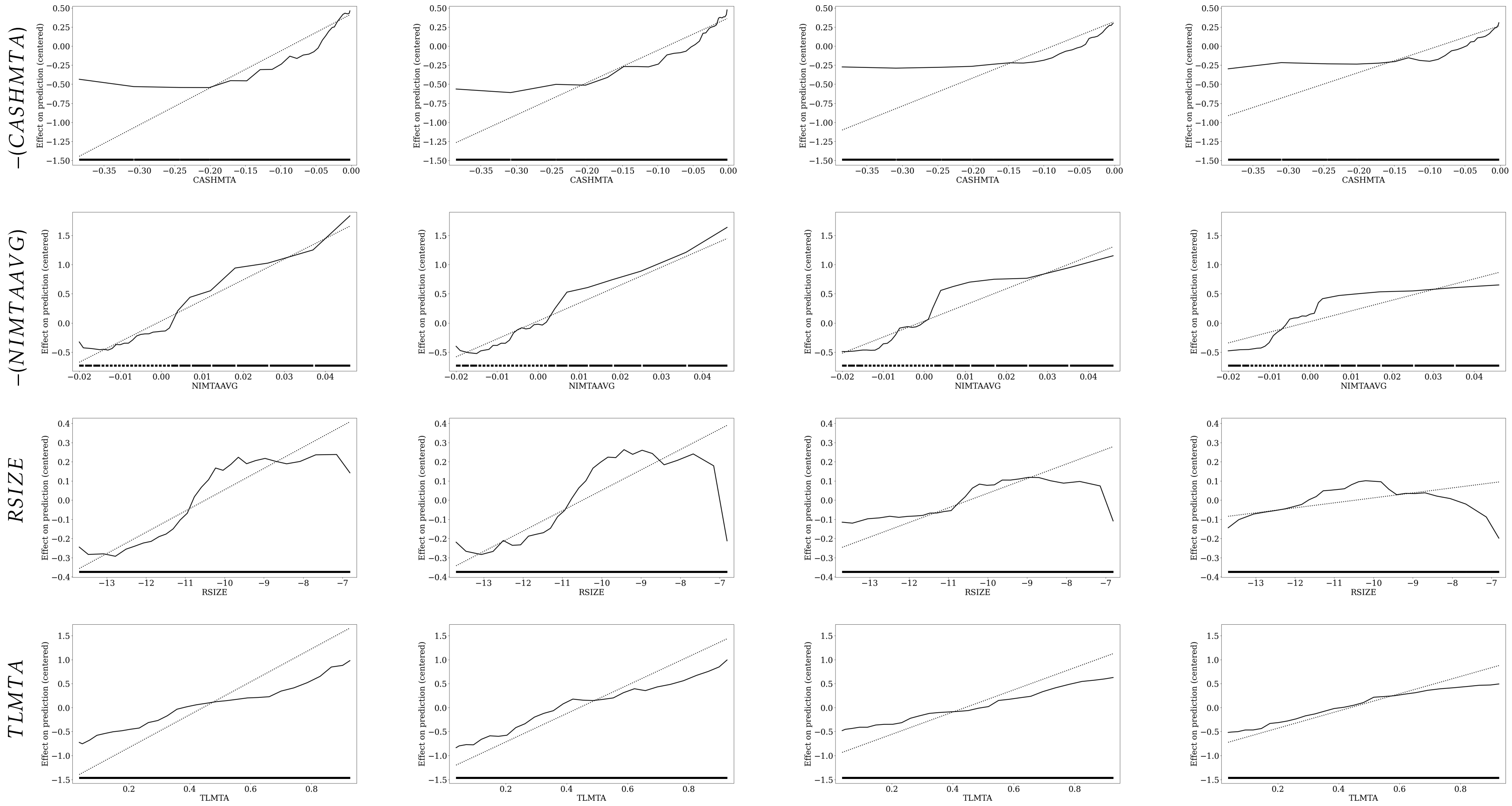
[8] We also measure the effects of features on default probabilities directly for both logistic regression and random forest with consistent results.

Figures 1.9 and 1.10 show the ALE plots for the in-sample data set across the default time-lags (1 month, 6 months, 12 months, and 24 months) for the fundamental and the market features, respectively. The logistic regression models (dashed lines) include the features linearly, and the plots show a linear effect on prediction, i.e., the linear predictor cannot capture non-linear feature effects. The random forest models (solid lines) allow for non-linearity and show a non-linear effect on the predictions for most of the applied features. The black markers at the bottom of each subplot illustrate the distribution of the feature under consideration. The relations are consistent with the uni-variate relations from Figure 1.6 and Figure 1.7 with the exception of $RSIZE$.

Figure 1.9 shows that firms with lower cash holdings in relation to total assets tend to have higher default rates only from a $-(CASHMTA)$ value of -0.2 on, while before there is no effect of lower illiquidity. $-(NIMTAAVG)$ is also positively related to default: the less profitable the company, the higher the default rate on average. The effect of firm size on the default probability prediction in the logistic regression model becomes less pronounced as the default horizon increases as the dashed line becomes flatter. However, the shape of the random forest model becomes positive for smaller firms (lower $RSIZE$ values), indicating decreasing default risk with increasing size at longer horizons (consistent with Campbell et al. (2008)). Note that $RSIZE$ has a different relation to default in a uni-variate setting. A higher leverage ratio ($TLMTA$) increases the risk of default, with this effect becoming more linear the longer the default time-lag.

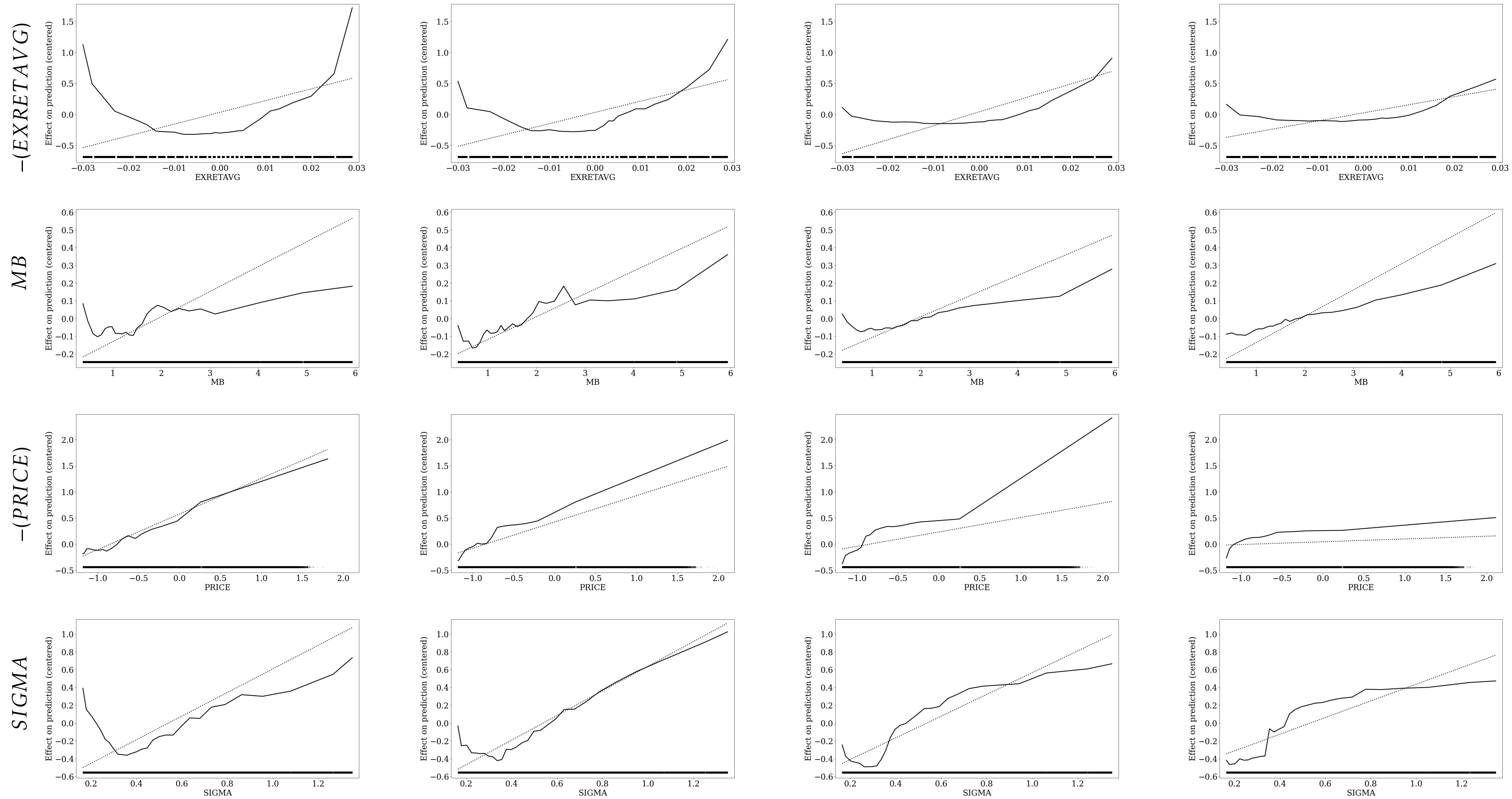
Figure 1.10 shows that the effect of $-(EXRETAVG)$ on PD prediction is u-shaped: companies with clearly lower average returns compared to the S&P500 have higher default risk. The average default risk also increases as the S&P500 index is significantly outperformed. The strength of this non-linear relation becomes less pronounced as the time horizon increases. The same applies to the market-to-book ratio (MB). The effect of MB on PD predictions for companies with rather low or rather high market value in relation to book value of equity is default risk increasing. The default risk increases for lower share prices ($-(PRICE)$). The volatility of the company's stock return ($SIGMA$) shows a u-shape for shorter time horizons, the greater the average default rate.

Figure 1.9: ALE plots: non-linearity and feature importance of the fundamental features by time-lag



Notes: This figure shows the ALE plots of the logistic regression (dashed lines) and the random forest model (solid lines) for the fundamental features of the in-sample data set from 1975 to 2007. The columns contain the subplots for a default time-lag of 1 month, 6 months, 12 months, and 24 months. The black markers along the x-axis of each subplot illustrate the distribution of the feature under consideration. The subplots show that the strength of the non-linear relation between the variables and the default rate decreases with increasing default time-lag.

Figure 1.10: ALE plots: non-linearity and feature importance of the market features by time-lag

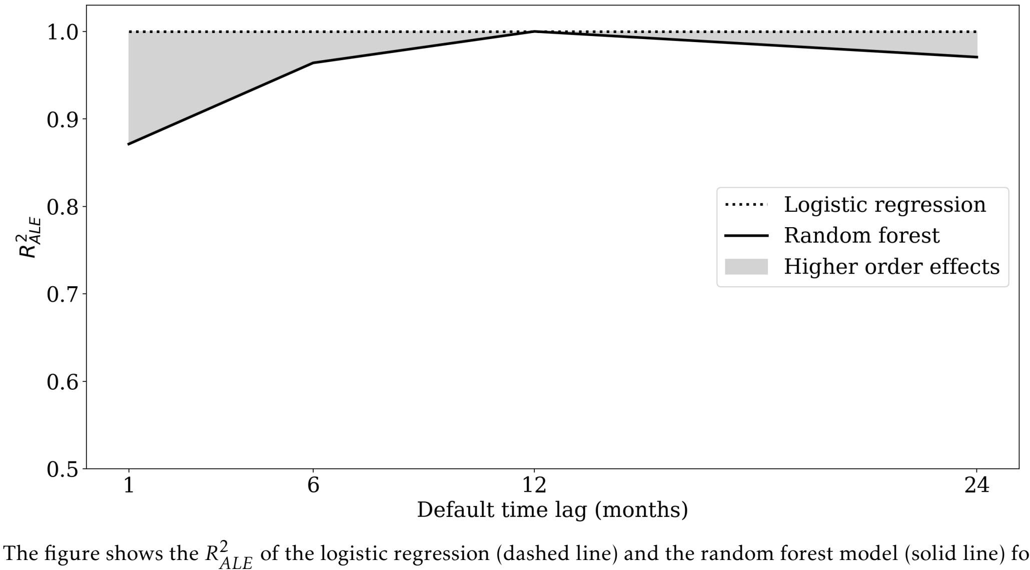


Notes: This figure shows the ALE plots of the logistic regression (dashed lines) and the random forest model (solid lines) for the market features of the in-sample data set from 1975 to 2007. The columns contain the subplots for a default time-lag of 1 month, 6 months, 12 months, and 24 months. The black markers along the x-axis of each subplot illustrate the distribution of the feature under consideration. The subplots show that the strength of the non-linear relation between the variables and the default rate decreases with increasing default time-lag.

In addition to non-linear effects of individual features, i.e., main effects of features on prediction, the presence of higher order effects, i.e., interaction effects may also contribute to increased performance of ML models. To quantify these effects, we compute the Apley and Zhu (2020) R^2_{ALE} measure, which indicates how much of the prediction of the machine learning method can be explained by main effects, that are shown in Figures 1.9 and 1.10. The portion that can be explained by higher order effects, i.e., pairwise interactions or interactions of higher order[9] can then be captured by $1 - R^2_{ALE}$. The gray shaded area in Figure 1.11 visualizes the range between the R^2_{ALE} of the random forest model and that of the logistic regression model, which is equal to one for all horizons, as there are no non-linear effects modeled by LR. The proportion of total higher order effects is 13% for the random forest model and a time-lag of 1 month and decreases with longer default horizons.

In summary, the relation between features and PDs is mainly characterized by marginal non-linear effects. This could also explain the single-digit relative increase in AUC when machine learning methods are used. These methods gain their strength especially when approximating highly complex patterns including interactions. However, the dominance of machine learning methods can be explained by marginal non-linearity as the set of these patterns is limited in this data set.

Figure 1.11: Feature importance - R^2_{ALE}



Notes: The figure shows the R^2_{ALE} of the logistic regression (dashed line) and the random forest model (solid line) for the default time-lags of 1 month to 24 months. The gray shaded area illustrates the level of non-linearity inherent in the higher order effects, which is present at shorter horizons (13% at 1 month) and decreases as the default lag increases.

[9] ALE plots for interaction effects between two or more features are challenging to interpret due the number of plots which is $\frac{1}{2}f(f-1)$ diagrams of second order effects for f features. For example, for our 8 features, 28 diagrams are needed to visualize only the pairwise interactions. Since Figure 1.11 shows that the total fraction of higher order effects is limited and decreases with longer default horizons, we focus on the quantification by $1 - R^2_{ALE}$.

1.6 Discussion

Recent literature on credit risk focuses on the use of machine learning techniques in finance including the prediction of default probabilities. Based on a comprehensive data set covering public U.S. corporate defaults over 46 years, we benchmark different machine learning models in terms of predictive performance with the logistic regression model. We focus on the prediction of default probabilities, and we calibrate the models on a training data set (”in-sample”), validate them on a test data set (”out-of-sample”), and apply them to a forward-looking data set (”out-of-time”).

We find that ML methods dominate logistic regressions in corporate default predictions due to non-linear relations between fundamental and market features and default outcomes. The tree-based ensemble methods like (extreme) gradient boosting (*XGB* and *GB*) and random forest (*RF*) consistently dominate the out-of-time PD predictions over all feature sets. The good performance of these models is in line with the prior literature.

We also find that the dominance of ML methods is more pronounced for market features than for fundamental features. The ML models perform best for the market features (excess return, stock price, stock return volatility, and market-to-book ratio) with a maximum increase in AUC improvement of 6.6% for the models *GB* and *RF*. ML performance on the total features is also better than of logistic regression (*LR*), that partly dominates the ML models on fundamental features (liquidity, net income, firm size, and leverage).

Last, we find that the dominance of ML methods is more pronounced for shorter default horizons as the non-linearity decreases as information fades over time.

As a robustness check, we visualize non-linear effects of market and fundamental features on the PD prediction by drawing accumulated local effects (ALE) plots and find that the strengths of machine learning methods for application to non-linear relations between financial ratios and the probability of default are particularly salient. From ALE plots analysis we also conclude that the set of highly complex patterns, such as interactions between variables, in this data set is limited.

In consequence, machine learning models are most useful for financial predictions if information is market-based and processed in a timely fashion. Resources should be allocated to measure such information and optimize processing times, which may include decreasing reporting and auditing times, publication periods, digitization of data collection and processing, and reduction of human overlap. Fundamental variables, because of their link to accounting, may have minimum lags that cannot be reduced due to accounting standards and other regulations. Instead, other sources of information that are close to the market and timely may be of importance, such as the textual analyses of from press releases, news, or corporate reports.

Suitable application areas of ML can be better defined and its effectiveness increased after analyzing non-linearity accordingly. The better understanding of such effects and its influences on ML models gives practitioners valid arguments for expanding the usage of ML and reduces barriers to entry that may exist. As machine learning increasingly becomes the focus of supervisory authorities, our findings also provide a basis for guiding principles. Further, our research enables banks and regulators to improve the accuracy of probability of default models for loan pricing, loan loss provisioning and capital allocations and hence improve the resilience of the financial system.

Chapter 2

Default Risk and Public Information: A Machine Learning Approach

This chapter corresponds to a working paper with the same name (submitted to the *European Journal of Operational Research*, currently under review).

Abstract

This paper shows how text features generated by advanced language processing models from the forward-looking MD&A sections of corporate reports can be used in both machine learning (ML, e.g., artificial neural networks) and standard models (e.g., logistic regressions) for predicting the default probability (PD) of U.S. corporate bonds. This is of great importance since standard models are still widely used in the financial industry and our approach therefore allows to include public information as additional source of information that would otherwise remain hidden. However, by using Explainable Artificial Intelligence (XAI) techniques, we find that text accounts for about 36% of feature importance beyond common financial ratios such as leverage and improves model calibration by more than 50%. We decipher the factors of this performance gain and quantify the total non-linear impact of (qualitative) public information to be about 43% of the total non-linearity. In addition, interactions between quantitative firm characteristics and qualitative textual features are identified that need to be accounted for in modeling. Our findings provide valuable in-depth insights into relevant drivers of default risk that are important when managing credit risks and making model-based decisions, e.g., in lending or investment processes.

Keywords: OR in Banking; Corporate Default; Explainable Artificial Intelligence; Natural Language Processing; Non-linearity

JEL classification: C53; G21; G33

2.1 Motivation

We show how text features generated by advanced language processing models can be used in both machine learning (ML) and standard models (e.g., logistic regressions), and quantify the importance of qualitative public information (about 36% of feature importance) and its non-linear impacts (about 43% of total non-linearity) in predicting the default probability of U.S. corporate bonds using explainable artificial intelligence (XAI) techniques. Recently, several U.S. government agencies asked financial institutions to comment on the use of ML in their business activities (see U.S. Federal Government (2021)). Also, the European Banking Authority (2021a) invited the financial sector to discuss the application of ML on rating modeling. In weighing the potential risks and benefits of artificial intelligence (AI), particular attention is paid to the explainability of AI, i.e., the need to understand the overall functionality of the methods and models, which are often more ”black-box” in nature than traditional approaches. This demonstrates the growing interest in the application of ML approaches and the importance these methods may gain in the future methodology landscape of the financial industry. However, this requires the knowledge of how to use methods that enable measuring the importance of features and their impact on model results, such as non-linear relations or interaction effects.

Estimating the default probability (PD) of debtors is one of the essential tasks in credit risk management (see, e.g., Ding et al. (2012)). It is recognized in literature and industry (e.g., by regulators) that ML models can be superior to common methods such as logistic regression in predicting default risk. For example, Fraisse and Laporte (2022) find evidence that credit risk management models based on artificial neural networks to predict corporate defaults can reduce regulatory capital requirements. Du Jardin (2021) shows that an ensemble of self-organizing neural networks can be superior to single models when forecasting corporate failure. However, he also recognizes limitations, such as the lack of explainability of the ML-based forecasts (see also du Jardin (2016)), especially hampering its practical usage, and the sole use of financial ratios (see also Geng et al. (2015)).

Corporate reports filed with the Securities and Exchange Commission (SEC) are a source of additional information beyond key financial figures. In general, they include a Management’s Discussion and Analysis (MD&A) section, which is intended to contain forward-looking statements about the firm’s condition and prospects. Therefore, information from these texts may provide additional predictive power when incorporated into forecasting models, as existing literature such as Mai et al. (2019) has been interested to show. However, Mai et al. (2019) again identify the application of XAI to PD prediction models as a point of future research, since in their study the drivers of performance gain through applying ML models remain open and thus the black-box of the trained neural networks remains closed. For the special case of predicting defaults among South American small businesses, Stevenson et al. (2021) evaluate the relevance of loan officers' notes about the lender using a deep learning approach. They apply an advanced natural language processing (NLP) model, labeled BERT (bidirectional encoder representations from transformers, by Devlin et al. (2019)), that they fine-tune to Spanish language and the credit scoring task, which may require expert knowledge to create a training vocabulary. However, due to the dimensionality of the BERT text representations, direct inclusion in standard statistical and machine learning models, such as logistic regression or random forest, is not possible without further dimensionality reduction. To overcome this issue, they use a word frequency approach as an alternative that appears to perform worse than the deep learning model. This lacks a consistent method for text extraction, and a comparative evaluation of the effects of text on PD predictions may not be fully guaranteed. Moreover, in this way the potential of such advanced NLP algorithms cannot be leveraged in standard methods.

These studies reveal a key challenge related to textual information that has not yet been addressed. *How can we incorporate advanced NLP algorithms like BERT into standard models for PD modeling?* Since logistic and linear regressions are still widely used in practice (see, e.g., European Banking Authority (2021a)), the potentials of highly complex and multi-dimensional text representations cannot yet be exploited in important tasks such as risk management. The ubiquitous need for explainability in the use of ML models, which is demanded as a basic prerequisite in literature and practice, further leads to the question of *how to uncover and quantify the influence of quantitative financial ratios and qualitative textual information on predictions for ML models*. Hence, understanding the impact of public information on the probability of default and thus assessing its relevance on PD modeling are of major interest to researchers and industry. Specifically, a financial institution that plans to incorporate text into its traditional PD models requires a methodological framework how to proceed.

This paper seeks to address these issues and makes the following contributions. First, to the best of our knowledge, we are the first to show how to transform text representations generated by the transformer-based NLP model BERT into features that are usable in any machine learning application, but also in traditional approaches of PD modeling such as logistic regression. By applying t-distributed stochastic neighbor embedding (t-SNE, see van der Maaten and Hinton (2008)), we further process the numerical BERT text representations and reduce the dimensionality to obtain modelable numerical text features that are objective, i.e., free of dependencies on expert knowledge such as word lists and generated training samples that would be required for fine-tuning BERT. This enables the inclusion of text especially in methodical landscapes where standard models still prevail. The broadened information base in this way may lead to higher model accuracy and thus may have a positive impact on decision-making processes such as lending, investment, or risk management measures. Second, we decipher the drivers of the performance improvement we find when incorporating public information into artificial neural networks (ANN) to model default risk by use of XAI techniques following Kellner et al. (2022). The ANN using both financial ratios and text as input delivers an increase of AUC by 3.6% (in-sample) and 2.4% (out-of-sample) compared to the logistic regression based on financial ratios only. Model calibration as measured by the Brier Score can be increased by more than 50% (53.6% in-sample and 55.7% out-of-sample), allowing more accurate PD predictions and thus more robust model-based decision making.

We uncover non-linear impacts of text, in particular higher-order effects of qualitative textual BERT features (about 43% of the total non-linearity) and interactions with quantitative firm characteristics. Text accounts for about 36% of feature importance beyond common financial ratios such as leverage. By quantifying and explaining these effects on PD predictions, we contribute to a better understanding of PD models based on machine learning, which often are black-box in nature. This may help to reduce both the limitations to the use of ML cited in the literature and the reservations that may prevail about ML applications in financial firms. The increased predictive power of ML models including text, which can now be explained, may lead to enhancements in the solidity, soundness, and strength of PD modeling frameworks underlying, for example, credit decisions or credit risk mitigation measures.

The remainder of the paper is structured as follows. Section 2.2 gives a brief review of the relevant literature. Section 2.3 presents the data used in this study. Section 2.4 contains the methodology used in this study: the feature extraction and dimensionality reduction approach is explained, the default prediction models are introduced, the explainable AI techniques are presented, and the validation metrics are defined. Section 2.5 provides detailed results and validation figures of the implemented PD models, including the analysis of non-linearity and feature importance of public information. Section 2.6 discusses the findings and concludes.

2.2 Literature Review

In the literature on default risk estimation, quantitative variables have emerged as important indicators of corporate default. Campbell et al. (2008) collect fundamental firm characteristics, i.e., cash liquidity, profitability, size, and leverage, and market characteristics, i.e., share returns in excess of an index return, market-to book ratio, absolute share prices, and share return volatility, for a sample of public companies. By predicting default probabilities over different time horizons, they obtain increased explanatory power, which is crucial for their inferences on the pricing of financially distressed stocks. For portfolios constructed based on analogously built features, Aretz et al. (2018) observe a positive default risk premium for an international sample of bankruptcy filings. Du Jardin (2016) uses financial ratios of French firms in profile-based models for bankruptcy classification and reaches lower misclassification costs. Although financial advisors could benefit from such more precise lending decision rules, du Jardin (2016) acknowledges that the proposed models have a black-box nature that prevents an explanation of the decisions, which may contradict regulatory requirements. Liang et al. (2016) can improve bankruptcy prediction for Taiwan firms by amending financial ratios with corporate governance indicators such as board structure. However, transferability may be limited since the usefulness of these variables depends on the market.

In recent years, increasing computational power has also enabled the use of alternative sources of information, such as the processing of textual disclosures, along with the use of ML approaches to default risk[1]. For example, deep learning models, i.e., neural networks with many layers, are applied to incorporate information from textual disclosures (see, e.g., Mai et al. (2019) or Stevenson et al. (2021)). Prerequisites for feeding text data into models are the ability to extract relevant features from the texts and to reduce the dimensionality of the extracted features. A very common approach is to use pre-defined word lists, such as the lists of positive or negative words in financial contexts proposed by Loughran and Mcdonald (2011), to extract sentiments[2]. For example, the occurrence of words with positive connotations such as ”success” or ”benefit” is used as a measure of the tone or polarity of the analyzed text excerpt. Being very efficient in dimensionality reduction, the strength of incorporating sentiment indices, however, depends on the pre-selection of words to create the word lists. Frankel et al. (2022) argue that the use

[1] There is also recent work on ML applications to credit scoring (see, e.g., Dumitrescu et al. (2022), Kozodoi et al. (2022), Gunnarsson et al. (2021), Li and Chen (2021) or Luo et al. (2020)), mortgage risk (e.g., Sadhwani et al. (2021)), credit card account behavior (see, e.g., Bakoben et al. (2020)), credit loss (e.g., Bastos and Matos (2022)), or Loss Given Default (see, e.g., Nazemi et al. (2022), Bellotti et al. (2021), or Yao et al. (2017)).

[2] Further studies using word list-based text features and sentiment scores on bankruptcy risk prediction include Nguyen and Huynh (2022), Tang et al. (2020), and Ahmadi et al. (2018).

of dictionary-based measures can entail shortcomings, such as omitting important disclosure features or applying them to unintended contexts. In addition, Donovan et al. (2021) state that dictionaries could lead to underestimating the relevance of qualitative information. Instead, they train an ensemble of ML models on a sample of 6,290 firm-years (14,618 firm-quarters) of 10-K report MD&A (conference calls) between 2002 and 2012 to associate the text with credit default swap (CDS) spreads and show improved prediction of credit events for firms with and without CDS spreads. However, the prerequisite of CDS trading severely limits the training sample size, i.e., CDS spreads are not available for more than 90% of long-term debt borrowers according to Donovan et al. (2021), and disclosure properties between companies with and without CDS spreads may differ (see, e.g., Kang et al. (2021), Kim et al. (2018), or Martin and Roychowdhury (2015)). Also, Donovan et al. (2021) did not expand the information base through quarterly MD&A sections from the 10-Q reports.

Not focusing on single words only but considering the context of words and sentences within each type of text, transformer-based methods like BERT (Bidirectional Encoder Representations from Transformers) provide a more in-depth understanding and thus representation of textual data. BERT as proposed by Devlin et al. (2019) is pre-trained on a large corpus using a Masked Language Model (MLM) and next sentence prediction (NSP), i.e., the language representation model learns the structures and contexts of unlabeled data. Then, the model can be fine-tuned for downstream tasks such as question-answering or sentiment extraction using task-specific labeled data. For example, Stevenson et al. (2021) fine-tune BERT to credit scoring and Spanish language with focus on small businesses. Focusing on peer-to-peer lending on the Lending Club platform, Kriebel and Stitz (2022) conduct a horse race of extraction techniques, including BERT adapted for classification, from user-generated text and show that text can significantly improve the credit scoring task. In particular, compared to a full and sometimes rather complex MD&A section, the textual observations studied by Kriebel and Stitz (2022) tend to be rather short, which they believe is one of the reasons that simpler approaches such as average word embedding perform well. For a detailed overview of literature on BERT-based models, especially applications on emotion detection and post-training BERT, see Acheampong et al. (2021).

Fine-tuning BERT in the context of default risk, i.e., labeling text according to default risk-relevant subjects to create a training sample, would require expensive resources such as extensive expert knowledge, analysis time, and computing power. In addition, this may again lead to features that are potentially biased, as subjective opinions on relevance with respect to default risk are co-processed. Since MD&A sections usually contain a wide range of different information and are quite complex in terms of text structure, as much information as possible should be made modelable instead of possibly limited human evaluation.

The above review shows that previous literature has found that dictionary-based measures incorporated in PD models may not be able to fully utilize the information contained in text data. Moreover, there is a gap in the way texts or their high-dimensional but informative representations can be integrated into standard PD models such as logistic regressions, which are still widely used in banking decision-making processes (see, e.g., European Banking Authority (2021a)). We fill this gap and show how text representations extracted by pre-trained BERT can be reduced in dimensionality and thus processed into advanced textual features that are usable as input to both traditional (e.g., logistic regression) and ML default prediction models.

2.3 Data

Our data consists of three sources. First, we retrieve a comprehensive data set of about 2.6 million monthly observations and 1,220 default events of U.S. corporate bonds over a period from 1993 to 2021 from the Moody's (2021a) Default and Recovery Database. Second, several fundamental and market features commonly used in the default risk literature (see, e.g., Campbell et al. (2008), Campbell et al. (2011), or Aretz et al. (2018)) are collected and computed from Refinitiv (2021) Eikon. Third, we download the 10-K and 10-Q filings gathered by SEC (2021) in the EDGAR[3] database to extract the MD&A sections, 65,087 files in total, and calculate textual variables. Table 2.1 provides an overview of all features used in this study.

Bond and issuer-specific variables

The mean default rate and number of observations for the bond and issuer-specific variables are shown in Table 2.2. If Moody's rating is downgraded by at least one notch, the average monthly default rate of a bond increases from 0.0126% to 2.4421%, so a dummy variable (1: downgrade, 0: no downgrade) is included. Industry affiliation has been found to be among the key determinants of credit risk (see, e.g., Acharya et al. (2007)). The lowest monthly default rate is observed in the banking sector (0.0004%), while FIRE (Finance, Insurance & Real Estate) has the highest average monthly default rate (0.1129%)[4].

[3] Via the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database, the U.S. Securities and Exchange Commission (see SEC (2021)) makes publicly available the documents such as annual or quarterly reports that companies are required to file under U.S. law.

[4] As with Krüger et al. (2018), for example, our controls include industry affiliation and we do not restrict our study to specific industries (see, e.g., Bonsall and Miller (2017)) to preserve generalizability as much as possible.

Table 2.1: Table of features

Label	Feature description
Default Indicator	Indicator on default of the firm by Moody's
Downgrade	Indicator of downgrade in rating by Moody's
Industry	Industry classification of firm
Total maturity	Total maturity of bond (in years)
Time to maturity (TTM)	Time to maturity of bond (in years)
$CASHMTA$	Cash and short-term assets relative to market-valued total assets
$NIMTAAVG$	Net income relative to market-valued total assets, avg. of previous four quarters
$RSIZE$	Company's log market capitalization relative to S&P500 index market cap
$TLMTA$	Total liabilities relative to market-valued total assets
$EXRETAVG$	Monthly log stock excess return relative to S&P500 index return, avg.
MB	Market-to-book ratio, calculated as market equity relative to book equity
$PRICE$	Stock price per share, logarithmic and truncated at a cap of $15
$SIGMA$	Standard deviation of a company's daily stock return over previous three months
LM Optimism Indicator	Sentiment from MD&A sections, following Loughran and Mcdonald (2011)
BERT text features	Features extracted from MD&A sections using BERT, following Devlin et al. (2019)

Notes: This table shows the features used throughout this study. The bond and issuer-specific information (default indicator, downgrade, industry affiliation, total maturity, time to maturity) is derived from the Moody's (2021a) Default and Recovery Database. The fundamental variables, i.e., $CASHMTA$, $NIMTAAVG$, $RSIZE$, and $TLMTA$, as well as the market variables, i.e., $EXRETAVG$, MB, $PRICE$, and $SIGMA$, are calculated based on data obtained from Refinitiv (2021) Eikon. Public information in the form of text from MD&A sections of SEC filings (see SEC (2021)) is represented by textual variables (LM Optimism Indicator and BERT text features).

Table 2.2: Bond and issuer-specific variables: Default rate and number of observations

Category	Characteristics	Default rate (%)	# obs.
Downgrade	No	0.0126	2,571,278
	Yes	2.4421	37,585
Industry	Banking	0.0004	447,205
	Capital Industries	0.0341	331,393
	Consumer Industries	0.0157	268,320
	Energy & Environment	0.0671	164,052
	Finance, Insurance & Real Estate	0.1129	696,815
	Media & Publishing	0.0397	50,368
	Retail & Distribution	0.0233	102,867
	Technology	0.0302	241,975
	Transportation	0.0183	65,506
	Utilities	0.0158	240,362
Total maturity	Short-term	0.0954	127,972
	Medium-term	0.0489	1,035,074
	Long-term	0.0410	1,445,817
Time to maturity in years (TTM)	0<TTM≤1	0.0493	257,663
	1<TTM≤2	0.0562	265,253
	2<TTM≤3	0.0536	246,469
	3<TTM≤4	0.0594	223,934
	4<TTM≤5	0.0462	227,182
	5<TTM	0.0414	1,388,362

Notes: This table shows the mean monthly default rate and number of observations for bond- and issuer specific features. *Downgrade* refers to the deterioration of the Moody's rating by at least one notch. The affiliation of the company to a specific industrial sector is controlled by *Industry* dummies. The sample is divided into three categories of *Total maturity*: short-term (up to three years), medium-term (more than three years but less than or equal to ten years), and long-term (more than ten years). The remaining time to maturity (*TTM*) is measured in years from the beginning of the observation month to the date the bond matures and is split into six maturity dummies.

Two time characteristics of bonds are considered (see Table 2.2): Total maturity, i.e., the time from issue to maturity, and remaining time to maturity (TTM) to reflect a possible maturity structure of default risk. Short-term bonds (total maturity up to three years) show higher mean default rates than bonds with longer total maturities. The average monthly default rates for different categories of time to maturity (in years) range from 0.0414% (TTM over five years) to 0.0594% (TTM between three and four years).

Fundamental and market features

Liquidity as an indicator of the company's short-term financial health is assessed by the ratio of cash and short-term assets to market-valued total assets (MTA), $CASHMTA$. The base MTA is calculated as the sum of market value of a company's equity and the book value of its liabilities, and is preferred over the book value of total assets since it may more quickly incorporate new information about the company's prospects (see Campbell et al. (2011)). As a measure of firm profitability, $NIMTAAVG$ depicts the ratio of net income to market-valued total assets (average over previous four quarters). The firm's relative size $RSIZE$ is measured as the logarithmic ratio of the company's market capitalization to the S&P500 index market cap. It is a relevant variable to control for potential size effects, e.g., when scalability is exploited differently. Both numerator and denominator have market values, i.e., provide a co-move, so that the ratio is market- and time-invariant, and thus no longer driven by market movements. Capturing the capital structure of the company, the firm's leverage ($TLMTA$) is calculated as the ratio of total liabilities to market-valued total assets.

These variables are complemented by market features, where the numerator is based on market prices. A measure of performance is the excess return of each firm's stock averaged over the last twelve months ($EXRETAVG$), setting the monthly logarithmic stock excess return in relation to the S&P500 index return. It indicates the ability of the company to outperform the reference index on the stock market. Additionally, the market-to-book ratio (MB) as a measure of both efficiency and growth prospects, and the enterprise's stock price per share ($PRICE$), logarithmic and truncated at a cap of \$15, as distance to default are included. When stock prices fall, a firm is moving toward default. Very low stock prices indicate a short distance to default. Finally, $SIGMA$ expresses the standard deviation of the company's daily stock return over the previous three months, capturing the recent stock performance and its volatility. High volatility may indicate greater likelihood of financial distress. Moderate volatility may point to the company's ability to outperform the reference index on the stock market. Low volatility may indicate a limited ability to trade and thus refinance.

Table 2.3: Summary statistics of fundamental and market features

	Fundamental features				Market features			
Feature	*CASHMTA*	*NIMTAAVG*	*RSIZE*	*TLMTA*	*EXRETAVG*	*MB*	*PRICE*	*SIGMA*
Panel A: Non-default group								
Count	2 607 643							
Mean	0.043	0.005	7.590	0.629	−0.003	2.543	2.668	0.314
Std. dev.	0.040	0.005	1.526	0.238	0.023	2.098	0.122	0.166
Min.	0.002	−0.005	4.581	0.220	−0.051	0.182	2.220	0.134
25%-qu.	0.011	0.002	6.467	0.437	−0.016	1.162	2.708	0.193
50%-qu.	0.029	0.005	7.901	0.623	−0.002	1.846	2.708	0.264
75%-qu.	0.062	0.009	8.655	0.891	0.013	3.125	2.708	0.381
Max.	0.137	0.015	9.923	0.946	0.039	9.130	2.708	0.772
Panel B: Default group								
Count	1,220							
Mean	0.038	−0.001	6.234	0.890	−0.041	0.590	2.592	0.706
Std. dev.	0.027	0.004	1.045	0.136	0.022	0.633	0.205	0.154
Min.	0.002	−0.005	4.581	0.220	−0.051	0.182	2.220	0.134
25%-qu.	0.031	−0.001	4.581	0.946	−0.051	0.417	2.708	0.772
50%-qu.	0.031	−0.001	6.888	0.946	−0.051	0.417	2.708	0.772
75%-qu.	0.031	−0.001	6.888	0.946	−0.051	0.417	2.708	0.772
Max.	0.137	0.015	9.503	0.946	0.039	8.495	2.708	0.772
Panel C: T-test statistics for the difference in means of the default and non-default group								
T-statistic	−4.35***	−43.33***	−31.04***	38.29***	−59.35***	−32.50***	−21.74***	82.50***

Notes: This table shows the summary statistics for the fundamental and market features used in the PD models. The count, mean, standard deviation (std. dev.), minimum (min.), 25%-quantile (25%-qu.), median (50%-qu.), 75%-quantile (75%-qu.), and maximum value (max.) are shown for the non-default group (Panel *A*) and for the default group (Panel *B*). Panel *C* shows the statistics of the t-test for the difference in means of the default and non-default group (see Welch (1947)). The significance is indicated for the 1% (***), 5% (**) and 10% (*) level.

Table 2.3 shows summary statistics for the group of non-defaulted firms (Panel *A*) and that of defaulted firms only (Panel *B*). On average, firms in the default group have lower liquidity, i.e., lower cash reserves that can serve as a short-notice buffer, e.g., for unexpected payments. These companies show also lower profitability, which may indicate ailing corporate health or production at (too) high costs, and are smaller (”too big to fail”). Firms in default also have higher leverage than the non-default group, i.e., tend to have a riskier capital structure.

The firms in the non-default group (Panel *A* of Table 2.3) are more likely to generate excess returns, indicating a strong business model and sales force. The higher average market value relative to book value of the non-defaulted companies may indicate higher growth opportunities and thus lower default risk. A higher (capped) log stock price, combined with lower volatility in stock returns of the non-default group may be associated with a stable (high) valuation of the company on the stock market. The test statistic of the t-test (see Welch (1947)) in Panel *C* indicates that all means are significantly different from those of the other group. The descriptive results are consistent with the literature such as Aretz et al. (2018) or Campbell et al. (2008) and suggest that distressed firms have a different structure in both fundamental and market information.

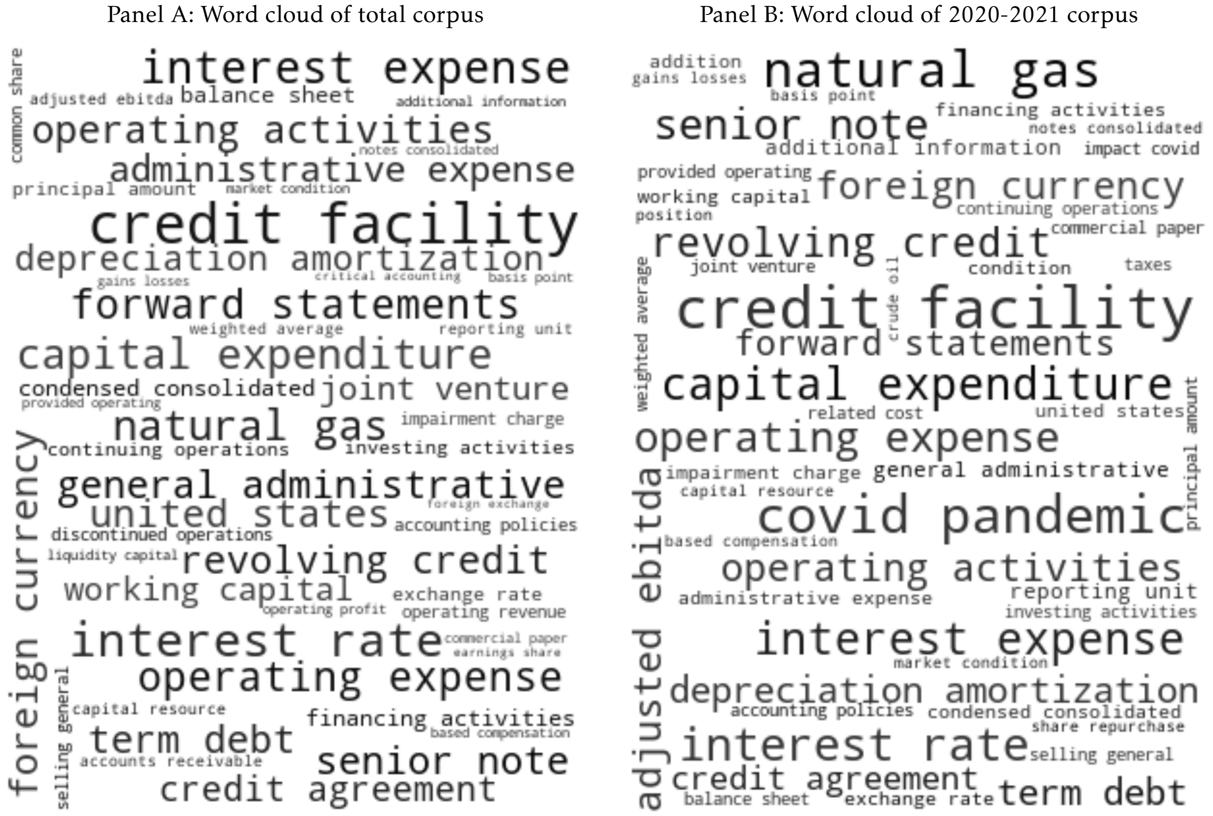
Public information

To complete our database, we download the 10-K and 10-Q filings gathered by SEC (2021) in the EDGAR[5] database, and extract the MD&A sections, in total 65,087 files. Combining the text of all MD&A sections provides a large corpus of words as object of analysis (see also the steps of data pre-processing in Figure 2.2). Graphical analytics tools for word occurrence include word clouds, which display the frequency of words or expressions in a text corpus. The more frequently a term occurs, the more prominent and bold its representation is in the graph. Figure 2.1 shows the 50 most common expressions in the MD&A sections of this study’s total corpus (Panel A) and the corpus based on corporate reports published in 2020 and H1 2021 (Panel B). In both panels, the most salient terms can be divided into financing-related terminology such as ”credit facility” or ”interest expense/rate”, and accounting-related expressions such as ”operating expense” or ”adjusted ebitda”. In addition, energy supply (”natural gas”) stands out as a frequently mentioned topic. Looking more closely at the corpus from 2020 onwards, COVID-19 (”covid pandemic”, ”impact covid”) is among the most frequently occurring topics. Since the MD&A contains management’s view on the firm’s prospects and the factors surrounding them, the

[5] Via the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database, the U.S. Securities and Exchange Commission (see SEC (2021)) makes publicly available the documents such as annual or quarterly reports that companies are required to file under U.S. law.

global pandemic is obviously seen as a key event affecting the operations of many firms.

Figure 2.1: Top 50 most frequent expressions in MD&A sections corpus



Panel A: Word cloud of total corpus

Panel B: Word cloud of 2020-2021 corpus

Notes: This figure shows the word cloud of the 50 most frequent expressions from the MD&A sections of the total corpus (Panel A) and the corpus based on corporate reports published in 2020 and H1 2021 (Panel B). The more prominent and bold an expression is, the more often it occurs in the corpus. The texts are free of numbers and HTML characters as well as stop words, i.e., words that do not add meaning to a sentence, such as ”and”, ”between”, or ”on”.

Data splitting: training and test data set

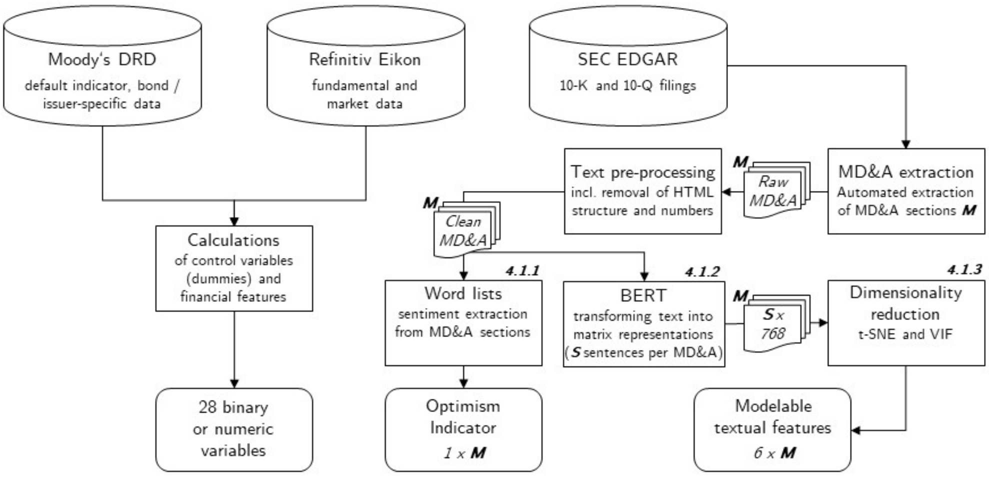
As is common in the literature (see, e.g., Djeundje and Crook (2019), Doumpos et al. (2017), or du Jardin (2016)), we randomly split the entire data set in the cross-section to separate ML model training and validation. Using the ratio 70:30 means that 70% of the data is assigned to the training set (”in-sample”, IS) and the machine learning algorithm learns the structure and context of this data. To validate the fitted model, 30% of the data remains in the test set (”out-of-sample”, OS).

2.4 Research Methodology

2.4.1 Feature Extraction from Textual Disclosures

To transform the extensive information inherent in the textual disclosures, we apply state-of-the-art approaches for extracting features and reducing dimensionality. Figure 2.2 shows the flow diagram of feature preparation, starting with the data sources (Section 2.3). For text processing, the first step after downloading the 10-K and 10-Q reports from EDGAR is to automatically extract the MD&A sections and perform further text pre-processing steps, such as cleaning the text of HTML tags or numbers. As a robustness check, sentiment-based text features using word lists are created (see Section 2.4.1). The feature extraction by using BERT is explained in Section 2.4.1, and dimensionality reduction techniques are presented in Section 2.4.1.

Figure 2.2: Flow diagram of feature extraction and preparation


Moody's DRD
default indicator, bond / issuer-specific data
Refinitiv Eikon
fundamental and market data
SEC EDGAR
10-K and 10-Q filings
MD&A extraction
Automated extraction of MD&A sections M
M
Raw MD&A
Text pre-processing
incl. removal of HTML structure and numbers
M
Clean MD&A
Calculations
of control variables (dummies) and financial features
4.1.1
Word lists
sentiment extraction from MD&A sections
4.1.2
BERT
transforming text into matrix representations (S sentences per MD&A)
M
S x 768
4.1.3
Dimensionality reduction
t-SNE and VIF
28 binary or numeric variables
Optimism Indicator
1 x M
Modelable textual features
6 x M

Notes: This figure shows the flow diagram of feature extraction and preparation. The information from the Moody's DRD and Refinitiv Eikon is further processed by calculating dummy variables and financial ratios, resulting in 28 binary or numeric variables. Texts from EDGAR are processed through MD&A extraction, text pre-processing, and feature extraction steps using word lists (robustness) and BERT (including dimensionality reduction techniques) to obtain modelable text features.

Word Lists

The sentiment of a text usually is measured by using pre-defined word lists, such as the lists of positive or negative words in financial contexts proposed by Loughran and Mcdonald (2011). To construct the optimism indicator (OI), first the occurrence of positively-attributed words such as ”success” or ”benefit”, n_m^{pos}, in the MD&A section m ($m = 1, \ldots, M$) is counted. Then the OI is computed as a measure of the degree of positive tone in the MD&A section m as follows:

$$OI_m = \frac{n_m^{pos}}{n_m^{all}}, \tag{2.1}$$

where n_m^{all} equals the total number of words in MD&A section m.

Transformer-based Model for Natural Language Processing: BERT

A wide range of NLP tasks can be performed using the transformer-based NLP model BERT (bidirectional encoder representations from transformers) by Devlin et al. (2019). The network architecture of the model consists of a stack of transformer encoder layers through which an input text sequence is transformed into weighted vector representations. The key difference from previous language models is the use of a bidirectional self-attention mechanism that enables the capturing of extensive contextual information. Before BERT can be applied to texts, pre-training is required, i.e., the model is trained on a large natural language corpus by optimizing two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)[6]. The output of the pre-trained BERT is a vector representation of size 1x768 for each input sentence. The size of the vector depends on the network structure of BERT, which in the base model consists of 768 hidden layers (see Devlin et al. (2019)). In this way, a text snippet of for example 100 sentences is BERT-transformed into a matrix of 100x768 representing the text and its contextual information[7]. The dimensionality of the text representations is too high for inclusion into logistic regression, for example, especially given the large number of documents in the data set. Hence, dimensionality reduction techniques are applied to obtain features that can be processed in all applied PD models.

[6] In the pre-training phase, the MLM and NSP tasks have two different objectives: The MLM task trains the model for bi-directionality by predicting randomly masked tokens, and is therefore the most important task. The NSP task mainly enhances the model in understanding sentence relations. In a second training phase, BERT can be further fine-tuned to a specific downstream task. This is done by feeding application-specific data into the model and training one or more additional layers on top of the pre-trained network architecture to specialize in the application's topic (see Devlin et al. (2019)). Since the fine-tuning phase requires context-specific labeled data, i.e., expert opinions or automatically (ML) generated opinions on default risk within a considered text, which may introduce biases, fine-tuning is not performed for the purpose of this study.

[7] The text transformation framework was implemented via python scripts. Using an NVIDIA GeForce RTX 3090 GPU with 10,496 cores operating at a boost clock of 1.70 GHz, processing 100 sentences with BERT takes about 9 seconds on average. If the scripts are run on a CPU instead, the processing times may increase significantly, e.g., by a factor of 10 using Intel Core i7-7700 CPU cores with 3.60 GHz.

Dimensionality Reduction Techniques

The BERT text representations provide raw features that are still complex in terms of their dimensions. To reduce dimensionality, we apply t-distributed stochastic neighbor embedding (t-SNE, see van der Maaten and Hinton (2008)). We choose t-SNE as a non-linear dimensionality reduction method because it is able to handle high-dimensional data and outperforms other approaches such as Isomap (see Tenenbaum et al. (2000)), as van der Maaten and Hinton (2008) show. A common but linear technique of dimensionality reduction is principal components analysis (PCA, see Hotelling (1933)). We include Isomap and PCA in our robustness analysis (see Section 2.5). T-SNE is based on stochastic neighbor embedding by Hinton and Roweis (2003) and measures the similarity of high-dimensional objects, assigning higher probabilities to similar objects and lower probabilities to dissimilar objects. By assuming a similar probability distribution over the data in a lower-dimensional space and minimizing the divergence between the distributions, dimensionality of the data can be reduced. Each BERT text representation matrix is reduced to a vector of 1x768 applying t-SNE.

To handle the problem of multicollinearity, we apply the concept of variance inflation factor (VIF), which is calculated as (see James et al. (2013)):

$$VIF(\beta_f) = \frac{1}{1 - R^2_{x_f|x_{-f}}} \tag{2.2}$$

with β_f as coefficient of feature x_f ($f = 1, \ldots, F$). The VIF expresses the proportion of the variance of coefficient β_f as the quotient of the variance of the full model and the variance of the model containing only β_f. The $R^2_{x_f}$ is the R-squared calculated for a regression of the feature x_f on all other features. Not allowing the VIF of each feature to exceed a value of 10, as suggested by James et al. (2013), results in a total of six BERT text features (in addition to the fundamental, market and control variables, see Table 2.1).

2.4.2 Default Prediction Models

Logistic Regression

We use a logistic regression (LR) model based on the fundamental and market features as a benchmark model (see, e.g., Campbell et al. (2008), Giordani et al. (2014), or Butaru et al. (2016)) for comparison with neural network models and models additionally including text features. In an LR, the features (weighted by the parameter vector) linearly enter a non-linear transformation. With N_t as the set of bonds at the start of period $t = 1,\ldots,T$, the default probability of bond i occurring in period t ($i \in N_t$) is defined as

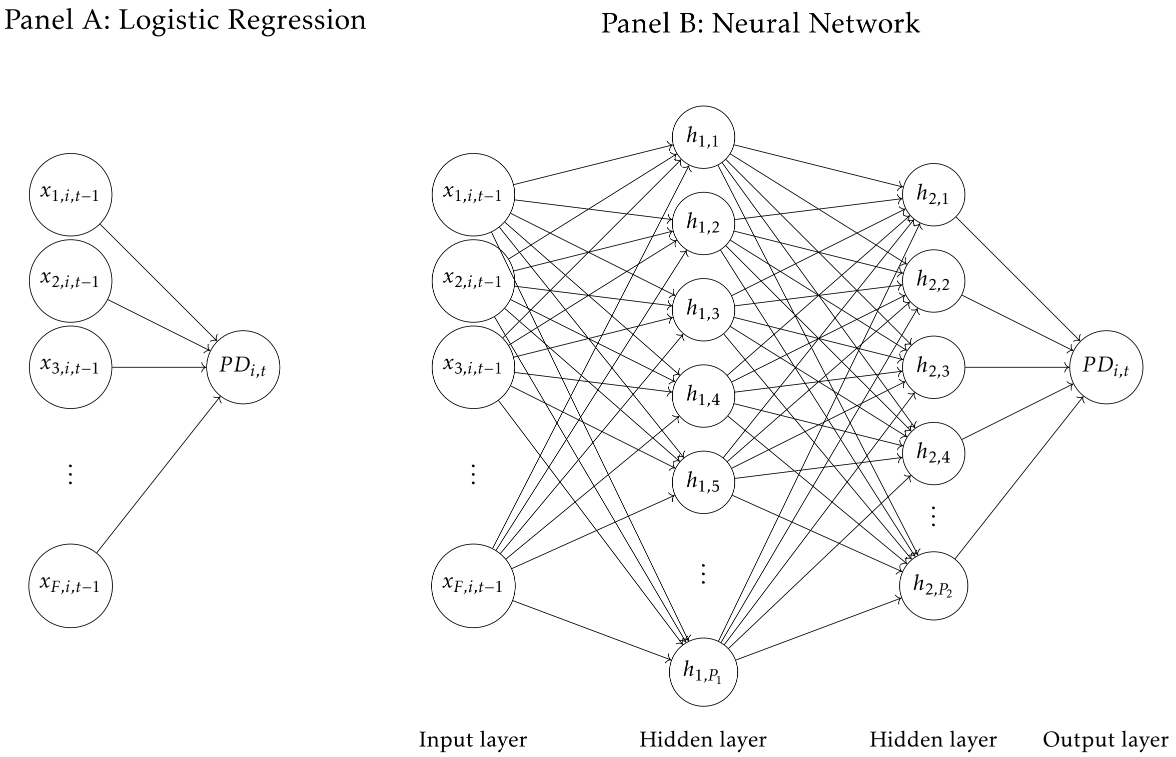
$$PD_{i,t} = P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1}) = \frac{\exp(\boldsymbol{\beta}'\,\boldsymbol{x}_{i,t-1})}{1 + \exp(\boldsymbol{\beta}'\,\boldsymbol{x}_{i,t-1})}, \tag{2.3}$$

where $\boldsymbol{x}_{i,t-1} = (1, x_{1,i,t-1}, x_{2,i,t-1}, \ldots, x_{F,i,t-1})'$ is the feature vector, observed at the end of period $t-1$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_F)'$ is the parameter vector.

Neural Network

By setting up a flexible structure of connected layers of neurons, neural networks (NN) allow non-linear modeling. The features (input layer) are passed to one (or more) so-called hidden layer(s) that non-linearly transform the input and feed the derived features on to an output layer. Figure 2.3 shows the graphical overview of an NN architecture compared to LR (see, e.g., Hastie et al. (2009)). The relation between the PD of bond i in period t and the input feature vector $\boldsymbol{x}$ for LR is displayed in Panel A.

Figure 2.3: Graphical overview — LR vs. NN


Panel A: Logistic Regression

Panel B: Neural Network

Notes: This figure shows the setup of the standard LR (Panel A) and the NN (Panel B). Input to LR and NN is the feature vector $\boldsymbol{x}_{i,t-1}$ of length F. Target variable is the probability of default $PD_{i,t}$ of bond i occurring in period t. The structure of the NN is exemplified for two hidden layers $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ with P_1 and P_2 neurons, respectively.

The same input is used in the neural network, but now the input layer is connected to the output layer via L stacked hidden layers. Each layer l ($l = 1,\dots,L$) consists of P_l neurons $\boldsymbol{h}_l = (h_{l,1}, h_{l,2}, \dots, h_{l,P_l}) \in \mathbb{R}^{P_l}$. Panel B of Figure 2.3 exemplary shows the structure of an NN for $L = 2$, allowing non-linear transformations in the two hidden layers and interactions. A logistic output function is typically used for default prediction. To determine the hyper-parameters of the neural networks applied, such as regularization strength or size of network architecture, a hyper-parameter search is conducted (see Appendix 2.A).

2.4.3 Techniques of Explainable AI

Explainable AI techniques are required to open the black-box of machine learning models and decipher the drivers and nature of the relation between input features and model outputs. Following Kellner et al. (2022), we use gradient information for computing first-order importance and higher-order importance including interactions between features. The approach is based on Horel and Giesecke (2020) and uses the relations learned from the neural network.

The total importance of an input feature x_f is labeled first-order feature importance $FI^{first}(x_f)$ and defined as

$$FI^{first}(x_f) = \frac{1}{C} \cdot \text{sgn}\left(\sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot \left(\frac{\partial P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})}{\partial x_{f,i,t-1}}\right)\right) \sqrt{\sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot \left(\frac{\partial P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})}{\partial x_{f,i,t-1}}\right)^2}, \quad (2.4)$$

based on the gradient for each feature x_f with respect to $P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})$, the default probability of bond i at period t. The normalizing constant C ensures that the importance of all features is equal to 1 in total, i.e., $\sum_{f=1}^{F} \left|FI(x_f)\right| = 1$. The direction on which the feature affects the model output is captured by the sgn$(\cdot)$ operator.

To measure possible interaction effects between input variables, we also compute higher-order feature importance. The degree of non-linear relations of a single input feature x_f is quantified by the second-order feature importance $FI^{second}(x_f)$:

$$FI^{second}(x_f) = \sqrt{\sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot \left(\frac{\partial^2 P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})}{\partial x_{f,i,t-1} \partial x_{f,i,t-1}}\right)^2}, \quad (2.5)$$

employing the second partial derivative of $P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})$ with respect to input feature x_f. The larger $FI^{second}(x_f)$, the greater the non-linearity of feature x_f in terms of influence on PD prediction. The extent of joint effects, i.e., interactions, of two features x_f and

x_g $(f, g = 1, ..., F, f \neq g)$ is measured by $FI^{interaction}(x_{fg})$:

$$FI^{interaction}(x_{fg}) = \sqrt{\sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot \left(\frac{\partial^2 P(D_{i,t} = 1|\boldsymbol{x}_{i,t-1})}{\partial x_{f,i,t-1} \partial x_{g,i,t-1}} \right)^2}. \tag{2.6}$$

The larger $FI^{interaction}(x_{fg})$, the more pronounced the joint impact of features x_f and x_g on the PD prediction. In the case that $FI^{interaction}(x_{fg})$ and $FI^{second}(x_f)$ have values close to zero, the influence of x_f can be assumed to be linear. This feature importance framework enables to identify and quantify non-linear relations between the ML model input and output, besides the direction and size of a single feature's impact.

2.4.4 Model Validation

To assess and compare the models' performances, the alignment of the model predicted PDs and the observed defaults is measured with two validation metrics. For investigation of the discriminatory power, i.e., the ability of the PD model to distinguish between defaults and non-defaults, we use the predominant area under the Receiver Operating Characteristic Curve (AUROC, hereafter referred to as AUC) (see Fawcett (2006) or Kang (2020)), which takes values between 0 and 1. A high AUC value means that the model provides high accuracy in classifying into a default group and a non-default group. The AUC can be calculated as the area under the curve between the hit rate HR (number of correctly predicted defaults relative to the number of defaults) and the false alarm rate FAR (number of non-defaults incorrectly classified as defaults relative to the number of non-defaults) (see Engelmann et al. (2003)):

$$AUC = \int_0^1 HR(FAR)\, d(FAR). \tag{2.7}$$

The relative improvement in AUC of model Z compared to model LR is calculated as

$$\Delta AUC = \frac{AUC_Z}{AUC_{LR}} - 1. \tag{2.8}$$

In addition, we perform a hypothesis test using bootstrapping (see Chava and Jarrow (2004)) to determine whether the AUC for model Z is higher than the AUC of LR. This is done by randomly drawing observations with replacement from the list of default observations and predicted default probabilities, and calculating the AUC values. Then the difference between the mean AUC value for the LR model and the mean AUC value for model Z is calculated. These steps are repeated 2,000 times. Finally, we obtain the probability for the hypothesis that the AUC for model Z predictions is higher than for the predictions of model LR as a percentile rank and report it for different significance levels.[8]

We measure calibration quality, i.e., how well PD model predictions are calibrated to observed default rates, using the Brier Score (BS) (see Brier (1950) or Kruppa et al. (2013)). It is calculated as the average of squared differences between the PD estimates ($P\hat{D}_{i,t}$) and the default observations ($D_{i,t}$):

$$BS = \sum_{t=1}^{T} \sum_{i \in N_t} \frac{1}{N_t} \cdot (P\hat{D}_{i,t} - D_{i,t})^2. \tag{2.9}$$

The better the model calibration, the lower the BS, i.e., the lower the deviation between the PD estimates and the default observations. As with the AUC, a hypothesis test is performed to determine whether the BS for model Z is lower than the BS of reference model LR, using bootstrapping.[9] The relative improvement in BS of the model Z compared to LR is calculated as

$$\Delta BS = 1 - \frac{BS_Z}{BS_{LR}}, \tag{2.10}$$

so that better calibration (lower BS) is expressed by greater values of ΔBS.

2.5 Empirical Results

2.5.1 Evaluation of Predictive Performance

In the following section, we evaluate the predictive performance of the PD models. Table 2.4 shows the coefficients and validation metrics of the logistic regression (1) without text features, (2) with sentiment, and (3) with text features from BERT, based on the total observations of bonds (no sample split). Among the control features, a rating downgrade by Moody's by at least one notch in particular significantly increases default risk in all logit models, indicating the close link between default and rating (see, e.g., Hilscher and Wilson (2017)). Bonds with short total maturity (up to three years) tend to have higher PDs compared to medium-term maturities

[8] We chose this bootstrapping approach to consistently test the differences in AUC and Brier Score of the models. Following Calabrese and Crook (2020), we also apply the test for AUC comparison proposed by DeLong et al. (1988), with consistent results available upon request.

[9] We also apply the test for Brier Score comparison proposed by Redelmeier et al. (1991), with consistent results available upon request.

(more than three but less than or equal to ten years). The closer the bond is to maturity, i.e., the shorter the time to maturity (TTM), the lower the risk of default, since the coefficients increase in absolute terms as TTM decreases. This relation is particularly significant for residual maturities of less than or equal to two years (0<TTM≤1 and 1<TTM≤2, respectively). The observed decline in default risk over time and the relation between shorter maturities to riskier borrowers are consistent with Krüger et al. (2018).

Regarding the fundamental and market features, the signs of the significant[10] coefficients are consistent for the different logit models, showing a constant influence of the features. Consistent with the descriptive statistics in Table 2.3, higher liquidity ($CASHMTA$), bigger market capitalization ($RSIZE$), increased performance in terms of excess stock return ($EXRETAVG$), and an increase (decrease) in market (book) equity compared to book (market) equity (MB) reduce default risk. The probability of default is higher for companies with higher leverage ($TLMTA$) and higher stock return volatility ($SIGMA$). The stock price (subject to the cap and the natural logarithm) has a positive sign for all logit models, indicating lower PDs for firms with a lower stock price ($PRICE$).

Default risk is negatively correlated with Loughran and Mcdonald (2011) sentiment (Model (2) in Table 2.4), which is measured by the optimism indicator, i.e., how positive the tone is in the MD&A section of the company. This suggests that the more positive management's report and outlook on the firm's prospects, the lower the risk of default of the company. The inclusion of this forward-looking information leads to an improved model fit (higher pseudo-R^2), better discrimination between defaulted and non-defaulted bonds (higher AUC), and better calibrated PDs (lower Brier Score), providing evidence that sentiments from MD&As are informative. The transformer-based text features are included in Model (3) and show a significant impact (apart from BERT feature #1). The BERT features can improve the model fit even more and lead to the best calibration of all logit models applied. The already high AUC of Model (1) can only be increased little by inclusion of BERT textual features into logit regression and is outperformed by Model (2). This may suggest that BERT features can help improve PD model results but may not fully unfold when used in logit regression because non-linear effects of features are not captured with logit.

[10] There is an ongoing debate about whether p-values are appropriate because they shrink towards zero as the number of observations increases (see Demidenko (2016)). We address this by analyzing the importance of features following Kellner et al. (2022) in Section 2.5.2.

Table 2.4: Coefficient table of the logistic regression model for the one-month lag

		(1) Base model	(2) LMCD Sentiment	(3) BERT text features
	Intercept	−28.259	−22.788	−29.108
		(1.001)***	(1.015)***	(1.012)***
Rat.	Downgrade	3.746	3.776	3.898
		(0.071)***	(0.072)***	(0.073)***
Industry	Capital Industries	6.802	7.875	7.783
		(0.724)***	(0.728)***	(0.731)***
	Consumer Industries	6.392	7.398	7.370
		(0.734)***	(0.736)***	(0.739)***
	Energy & Environment	6.829	7.673	7.769
		(0.724)***	(0.726)***	(0.730)***
	Finance, Insurance & Real Estate	6.390	7.345	7.309
		(0.710)***	(0.713)***	(0.715)***
	Media & Publishing	6.403	7.180	7.281
		(0.754)***	(0.756)***	(0.762)***
	Retail & Distribution	5.628	6.987	6.745
		(0.740)***	(0.744)***	(0.746)***
	Technology	6.547	7.388	7.537
		(0.726)***	(0.729)***	(0.731)***
	Transportation	6.980	7.804	7.439
		(0.783)***	(0.786)***	(0.795)***
	Utilities	5.282	5.799	6.484
		(0.733)***	(0.735)***	(0.741)***
Total maturity	Short-term	0.758	0.595	0.518
		(0.136)***	(0.140)***	(0.141)***
	Long-term	0.140	0.070	0.027
		(0.082)*	(0.083)	(0.084)
Time to maturity	0<TTM≤1	−0.535	−0.508	−0.499
		(0.137)***	(0.139)***	(0.141)***
	1<TTM≤2	−0.332	−0.396	−0.311
		(0.123)***	(0.125)***	(0.127)**
	2<TTM≤3	−0.244	−0.364	−0.265
		(0.114)**	(0.116)***	(0.117)**
	3<TTM≤4	−0.073	−0.204	−0.059
		(0.114)	(0.117)*	(0.117)
	4<TTM≤5	−0.182	−0.236	−0.123
		(0.121)	(0.122)*	(0.123)
Fundamental features	*CASHMTA*	−11.488	−7.791	−10.838
		(1.150)***	(1.094)***	(1.186)***
	NIMTAAVG	−11.931	6.131	−25.552
		(7.773)	(7.629)	(8.392)***
	RSIZE	−0.134	−0.082	−0.081
		(0.030)***	(0.032)**	(0.034)**
	TLMTA	6.004	4.344	5.254
		(0.439)***	(0.426)***	(0.435)***
Market features	*EXRETAVG*	−13.807	−12.749	−12.121
		(1.759)***	(1.772)***	(1.741)***
	MB	−0.812	−0.663	−0.783
		(0.084)***	(0.084)***	(0.082)***
	PRICE	3.160	1.730	3.037
		(0.197)***	(0.209)***	(0.204)***
	SIGMA	5.599	5.673	6.049
		(0.254)***	(0.263)***	(0.253)***
Text features	LMCD SENTIMENT		−1.799	
			(0.091)***	
	BERT feature #1			−0.001
				(0.003)
	BERT feature #2			−0.031
				(0.002)***
	BERT feature #3			−0.009
				(0.002)***
	BERT feature #4			−0.008
				(0.002)***
	BERT feature #5			0.042
				(0.003)***
	BERT feature #6			−0.012
				(0.003)***
	# observations	2,579,454	2,579,454	2,579,454
	# defaults	1,228	1,228	1,228
	Pseudo-R^2	0.5177	0.5385	0.5492
	AUC	0.9640	0.9661	0.9648
	Brier Score (‰)	0.3677	0.3397	0.3123

Notes: This table shows the coefficients of the logistic regression model for the one-month default time lag. The standard errors are given in parentheses. The significance is indicated for the 1% (***), 5% (**) and 10% (*) level.

A key strength of applying machine learning models such as neural networks is the recognition of non-linear effects that can lead to increased model power. The comparison of model performances of the introduced logit models and neural networks applied only to firm characteristics or including text features is shown in Table 2.5. Panel A depicts the validation metrics for the training sample, while results for the test data set, i.e., data that were not used for fitting the models, are displayed in Panel B. The metrics of model improvement, i.e., ΔAUC and ΔBS, express the relative superiority over Model (1) LR. The relative performance gain from using a neural network for PD modeling compared to logistic regression can be determined by comparing Model (1) and (4), indicating improved model discrimination by 2.0% and a better calibrated model by 51.2%. This suggests that the features show a substantial amount of non-linearity that is modeled by the neural network and enhances results.

Table 2.5: Prediction model performance

Panel A: Training sample

Model	AUC	ΔAUC	Brier(‰)	ΔBrier
(1) LR	0.9624		0.3721	
(2) LR + LMCD Sentiment	0.9646 **	0.2%	0.3431 ***	7.8%
(3) LR + BERT features	0.9633	0.1%	0.3171 ***	14.8%
(4) NN	0.9929 ***	3.2%	0.1927 ***	48.2%
(5) NN + LMCD Sentiment	0.9891 ***	2.8%	0.2130 ***	42.8%
(6) NN + BERT features	0.9967 ***	3.6%	0.1725 ***	53.6%

Panel B: Test sample

Model	AUC	ΔAUC	Brier(‰)	ΔBrier
(1) LR	0.9692		0.3680	
(2) LR + LMCD Sentiment	0.9700	0.1%	0.3383 ***	8.1%
(3) LR + BERT features	0.9687	−0.1%	0.3116 ***	15.3%
(4) NN	0.9883 ***	2.0%	0.1796 ***	51.2%
(5) NN + LMCD Sentiment	0.9899 ***	2.1%	0.2017 ***	45.2%
(6) NN + BERT features	0.9923 ***	2.4%	0.1629 ***	55.7%

Notes: This table summarizes the performance of the prediction models for the training data set (Panel A) and the test data set (Panel B). AUC, Brier Score, and the corresponding improvements (ΔAUC and ΔBrier) over the base model *LR* are indicated. The significance for the test to ascertain whether the AUC (Brier Score) of the model is higher (lower) than the AUC (Brier Score) of the benchmark model *LR* is indicated for the 1% (***), 5% (**) and 10% (*) level.

As already observed for the logistic regressions based on the entire data set (Table 2.4), including texts by sentiment or BERT features in logit can only improve model calibration, but not model discrimination in both the training and test sample. This leads to an analogous interpretation of logit results not fully capturing the information content of texts. Similar, the incorporation of sentiment from the text in the neural network model can only add very limited additional discriminatory power to the test sample (additional gain of 0.1% from Model (4) to (5)). In particular, the model fit (both AUC and Brier Score) for the training sample even deteriorates when sentiment is included in the neural network (5) compared to Model (4). Best models (training and test sample) in terms of both discrimination and calibration are the neural networks including BERT features (6) that can improve the already high fit of model (4) without texts.

Thus, we conclude that the limited performance gain of incorporating dictionary-based sentiment may indicate that dictionary-based sentiment may be too short-sighted for enhancing default risk estimation. This also corresponds to prior studies regarding dictionary-based sentiment (see, e.g., Frankel et al. (2022)). Further, our findings suggest that public information extracted by applying BERT can add additional power to ML models. The neural networks are observed to absorb the informativeness of forward-looking MD&As well. Therefore, a detailed analysis of how BERT text features impact ML predictions and what kind of interactions exist with quantitative features or control variables follows, which contributes to open the black-box of ML models in credit risk applications.

2.5.2 Feature Importance of Forward-looking Public Information

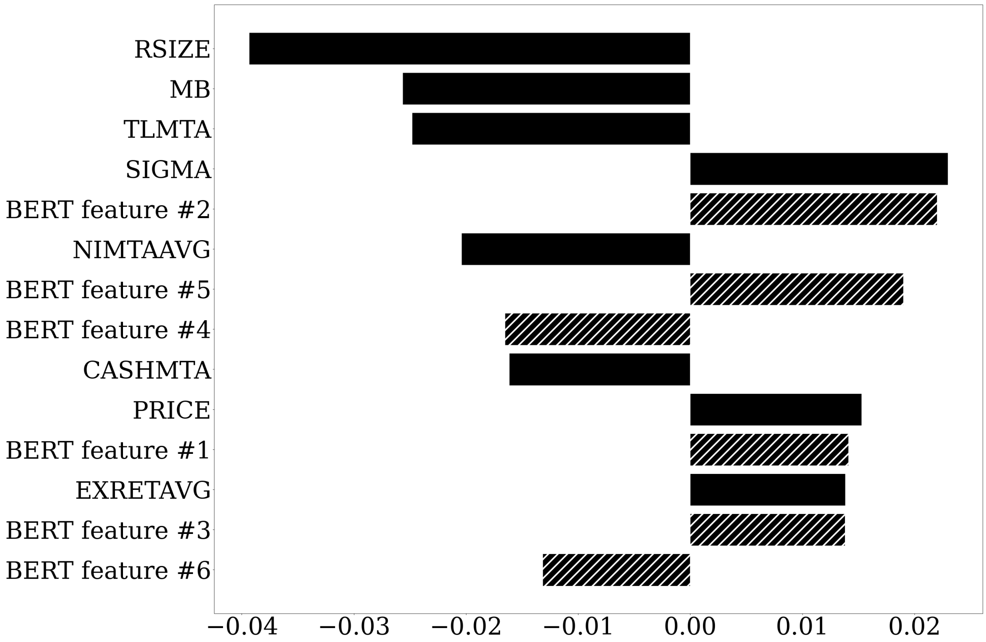
Public information represents a valuable and comprehensive data source that can be retrieved and processed into a learning base for machine learning applications. Researchers have recognized and seized this opportunity to enhance their results such as delivering improved explanatory power or predictability (see, e.g., Frankel et al. (2022) or Donovan et al. (2021)). While methods for extracting modelable features from texts differ, the need to measure the impact of text variables on model output is ubiquitous and should be an essential component when incorporating text. When modeling future-oriented contexts, such as estimating default probabilities, gathering forward-looking information is decisive to this task. The MD&A section of corporate annual and quarterly reports includes such information as management provides an outlook on the firm's prospects and how the current micro and macro situation might affect the company's future.

In this section, we measure the feature importance of the textual features and the quantitative features (fundamental and market variables) when fed as input to neural networks. Following Kellner et al. (2022), we use gradient information for calculating the first-order importance and higher-order importances including feature interactions. For retrieving robust importance measures, we fit the neural network 100 times and compute the average feature importance.

Figure 2.4 summarizes the main effects, i.e., first-order effects of quantitative firm characteristics and qualitative text features. The lengths of the black bars suggest that the size of the company is a very important indicator with a negative value of feature importance, so that default risk decreases with increasing firm size (in terms of market capitalization). This is in line with results from logistic regression (Table 2.4). Also, Campbell et al. (2008) and Aretz et al. (2018) report a negative impact of $RSIZE$ on default probability. Market-to-book ratio (MB) and $SIGMA$ also confirm the direction of impact observed with the LR. $SIGMA$ captures the recent volatility by measuring the company's daily stock return over the previous three months, showing a positive effect. A higher volatility may indicate a greater likelihood to experience financial distress. MB measures efficiency and growth prospects of the company and has a negative impact. Low market-to-book firms are low efficiency firms with high default risk. High market-to-book firms are often start-ups with a high potential to grow but also to fail. Griffin and Lemmon (2002) in conjunction with stock returns, find increased distress risk for firms both with low and high MB values. On average, firms of the non-default group show a significantly higher market value (see Table 2.3), confirming our finding of a negative relation between default risk and growth prospects. Higher leverage ($TLMTA$) indicates lower default risk as the sign of the feature importance is negative. However, also including Figure 2.5 in analysis shows extensive higher-order effects and interactions with leverage that may have influence on the feature sign.

Following size, market-to-book ratio, leverage, and volatility, which literature also acknowledges to be among the most important default risk indicators (see, e.g., Campbell et al. (2008) or Aretz et al. (2018)), further relevant features in Table 2.4 include several text features that underpin importance of public information. BERT feature #2 and #5 have a positive sign indicating an increasing influence on default risk, while BERT feature #4 shows a decreasing effect. Further, the more profitable the firm ($NIMTAAVG$), the lower the probability of default (in line with literature such as Campbell et al. (2008)). Similarly, $CASHMTA$ has a negative relation to default risk (confirmed by Acharya et al. (2012)). Companies with a low level of cash holdings have liquidity constraints that may lead to higher default risk. $PRICE$ shows a positive sign of feature importance, corresponding with the logit coefficient (see Table 2.4). Campbell et al. (2008) report a switching sign of $PRICE$ from negative to positive for increasing time horizons, while Aretz et al. (2018) find no clear sign (mixed impacts for different countries). Among further text features (BERT features #1, #3, and #6), $EXRETAVG$ shows a positive sign, deviating from the corresponding logit coefficient. However, since excess return shows a limited feature importance, its sign may not be fully reliable (in absence of significance levels that cannot be retrieved for neural network feature importance).

Figure 2.4: First-order feature importance

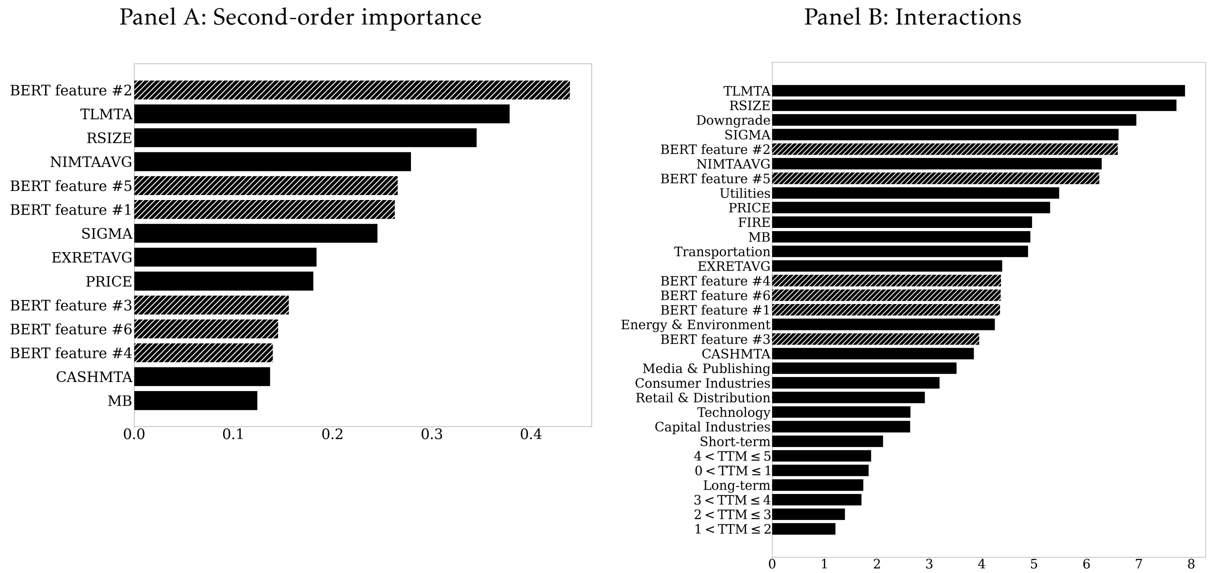


Notes: This figure displays the first-order feature importance of text features (hatched bars) as well as fundamental and market variables (black bars). The strength and direction of importance is computed based on Equation (2.4). The share of the total absolute first-order importance is 35.6% for the text features and 64.4% for the fundamental and market features.

In summary, our findings document that size ($RSIZE$), leverage ($TLMTA$), and stock return volatility ($SIGMA$) are among very important indicators of default risk, which corresponds to results reported by Aretz et al. (2018). Quantitative features account for 64.4% of the total absolute first-order importance. The share of qualitative features is 35.6%, providing evidence that BERT features extracted from forward-looking MD&As are informative.

The model results in Table 2.5 indicate a significant performance increase in applying machine learning models, which are strong in modeling complex patterns and dependencies between variables. Therefore, we measure the degree of non-linearity, i.e., higher-order effects and interactions, in the relation between features and default risk to provide an in-depth analysis of the drivers of this performance gain. Panel A of Figure 2.5 shows that BERT feature #2 has the highest second-order importance of all variables, and further BERT features are among the most important higher-order effects. This provides evidence that public information is related non-linearly to default risk, and that this non-linear relation can be captured by the neural network model. Overall, public information accounts for 42.9% of second-order importance, compared with 57.1% for quantitative features. Also, the portion of interactions with textual variables is high as two BERT features are among the top 7 variables of joint interactions (Panel B). The text features in particular interact with the downgrade indicator (see Table 2.C.1), suggesting that public information about a firm, along with a deterioration in the company's creditworthiness, jointly affects its default risk. Moreover, as the feature with the second-highest second-order effects and the strongest interactions with other variables, the firm's leverage ($TLMTA$) exhibits a significant degree of non-linearity. The top 15 interactions in Table 2.C.1 confirm this observation by listing four interactions of $TLMTA$. In particular, leverage appears to interact with industry affiliation[11] (utilities, media & publishing, and transportation) or rating downgrade. Because of this extensive extents of higher-order effects, the sign of the main effects of $TLMTA$ (see Figure 2.4) may be misleading since effects should be considered in connection. Overall, the considerable portion of non-linearity (second-order effects) and interaction effects show, that public information should be considered as relevant features in relation to default risk.

Figure 2.5: Higher-order feature importance


Panel A: Second-order importance

Panel B: Interactions

Notes: This figure shows the second-order feature importance (Panel A) and the feature interactions (Panel B), based on Equations (2.5) and (2.6), respectively. The hatched bars represent the text features, while the fundamental, market or control variables are shown in black bars.

Study of stylized MD&A sections

To provide an intuitive, human-interpretable example of how information from text can be incorporated into PD models, we write two stylized MD&A sections (see Table 2.6) that are not part of a real report but are completely invented[12]. The first text presents a rather positive

[11] Industry affiliation was not included into the discussion of feature importance because it depends on the company's business activity and cannot be changed by the firm. The first-order feature importance of the control variables is shown in Figure 2.B.1 in the Appendix.

[12] These examples represent stylized texts. We recognize that MD&A sections are typically longer and contain many more aspects of a company's performance or the measures and challenges as assessed by the management. However, these shorter, stylized examples are intended to illustrate the potential impact that forward-looking public information may have on the assessment of a company's default risk.

analysis of the company's current situation and a convincing outlook on future prospects, thus may be labeled as ”optimistic”. The second text excerpt expresses concerns about the status of the company due to endogenous and exogenous factors and gives a restrained preview of subsequent periods, which may be perceived as ”pessimistic”. Assuming that the texts represent the frankly reported state of the company and thus reflect information that may be relevant for credit risk assessment, the company reporting the more optimistic section may be expected to have lower default risk than the pessimistic reporting firm.

Table 2.6: PD predictions for stylized MD&A sections

Type	Text of stylized MD&A section	Predicted PD
Optimistic MD&A Section	*In the last year, we were able to scale the sales of our products and decrease our production costs due to improvements in purchasing strategy. We are confident that the current boom in demand will provide advantages and chances to our business model and strengthen our financial health. The focus on new product development and innovative solutions gives us the chance to expand our market share and profits in the future.*	0.0016%
Pessimistic MD&A Section	*The current economic conditions combined with recent changes in regulatory requirements continue to pose a major challenge to our business. We had to accept a substantial decline in sales due poor developments in our core segment, which requires us to raise unscheduled debt and jeopardizes planned investments. Since the outlook for our long-term rating has been revised from stable to negative, we expect a substantial increase in our cost of debt.*	0.0058%

Notes: This table shows the text of the stylized optimistic and pessimistic MD&A Sections and the resulting monthly PD predictions of the neural network model including BERT text features extracted from these sections. Median characteristics in terms of financial and control variables are assumed.

To measure the effect of text standalone, we assume median characteristics in terms of financial and control variables and predict the probability of default incorporating the text features from the optimistic and pessimistic MD&A sections using the ANN model. We find that the default probability predicted for the pessimistic reporting firm is more than three times that of the optimistic reporting firm (see Table 2.6), which may result, among other things, from the consciously exaggerated stylized texts. However, abstracted from the example, the information content of the texts may add valuable additional information as model input beyond financial variables, that may increase the accuracy of PD models.

Robustness

We perform a series of checks to demonstrate that our findings are robust. All results are available upon request. First, we repeat our analysis with the second- and third-best architecture of the hyper-parameter search. The results are similar, while model performance decreases. Second, we apply cross-validation and validation on an out-of-sample data set to check the generalization of the models, including a functional check for over-fitting control. Third, we initialize each neural network 100 times and report the average feature importance. Overall, the fluctuations are very small. This may indicate that the relation is quite stable. Fourth, we apply further dimensionality reduction techniques such as Isomap or principal component analysis to the extracted text features. This has a decreasing effect on neural network model performance, while dimensionality is too high for use in the logistic regression model.

2.6 Discussion

We show how public information in the form of MD&A sections from SEC filings can be transformed into modelable text features using the transformer-based NLP approach BERT and dimensionality reduction techniques. The key strength of this procedure is that we can feed the text as input features into both machine learning and conventional models such as logistic regressions. This is an important contribution being highly relevant especially to financial industry, as logistic and linear models are still widely used in Finance. Hence, our findings may help to both incorporate public information into traditional models in use and provide a basis for model comparison when evaluating the implementation of machine learning models incorporating textual features.

Based on a comprehensive data set covering default events from U.S. corporate bonds over a period from 1993 to 2021, we decipher the drivers of performance improvement when incorporating public information into artificial neural networks for modeling default risk. Using explainable AI, we find evidence that public information matters for explaining default, as textual features extracted from MD&A sections are related non-linearly to default risk and are among the most relevant variables, accounting for 36% of total feature importance. We uncover these non-linear impacts of text, in particular higher-order effects (about 43% of total non-linearity) and interactions between qualitative textual BERT features and quantitative firm characteristics. Our results suggest that these effects need to be considered when incorporating public information. By quantifying the importance of features and degree of non-linearity, we shed light on the black-box of ML approaches such as artificial neural networks and contribute to a better understanding of machine learning-based PD models.

Studies such as Campbell et al. (2008) or Aretz et al. (2018) have identified fundamental firm variables (e.g., liquidity or leverage) and market variables (e.g., stock returns in excess of an index return or market-to-book ratio) to be well suited for predicting default risk. Consequently, a firm may actively decrease its PD by, for example, increasing liquidity or deleveraging. In contrast, the content of corporate reports may not be entirely at the discretion of the reporting entity, since public company disclosures are usually subject to legal and regulatory requirements. Therefore, the contents of the report cannot be fully controlled and a firm may not be able to alter its PD directly by altering the text. However, by using text as an additional variable and evaluating its effect on predictions in this way (see, e.g., Berger et al. (2020)), we show that public information accounts for a substantial fraction of the total feature importance (about 36%) when modeling PD. Hence, text may be interpreted as a proxy for non-observable information that would otherwise remain hidden (see also Netzer et al. (2019)), while it shows to be highly relevant for default risk prediction.

In summary, this knowledge may enable financial analysts to make more accurate predictions by exploiting text information as we find increased PD model discrimination (out-of-sample increase of 2.4%) and calibration (out-of-sample increase of 55.7%). Disclosure text as a proxy of non-observable information can supplement the information base for PD models in addition to quantitative financial ratios, allowing financial institutions that focus on collecting such information to allocate their resources more effectively. In addition, measuring the importance of features and uncovering non-linear impacts of text will improve understanding the outcomes of machine learning models, which may help ensure that decisions made based on those model results, such as lending or risk mitigation decisions, remain fair and informed. As a consequence, our findings can help make the use of machine learning methods in credit risk management more meaningful and secure, as explainability and interpretability become part of the toolkit.

2.A APPENDIX | The Hyper-parameter Search

In order to find the optimal choice of hyper-parameters for PD prediction, cross-validation and parameter tuning are performed. We conduct a five-fold cross-validation on the training data set for every neural network applied. The cross-validation splits the training data randomly into five parts of approximately equal size. The respective model is fitted on four (i.e., one part less) parts and the fifth part is predicted measuring the resulting prediction error. The permutation of the fifth prediction part leads to five prediction errors, which are averaged over all folds to assess the total performance of the selected hyper-parameter combination (see Rösch and Scheule (2020)). Table 2.A.1 displays the grid for each hyper-parameter to optimize (Panel A) and the validated hyper-parameters per model (Panel B).

Table 2.A.1: Setup of hyper-parameter search and validated parameters

Panel A: Search grid		Panel B: Validated hyper-parameters		
Parameter	Distribution	NN	NN + LMCD Sentiment	NN + BERT features
Learning rate	$U^c \sim [0.000001, 0.01]$	0.000749	0.005025	0.001104
Lambda L1	$U^c \sim [0.000001, 0.01]$	0.000021	0.000001	0.000025
Dropout	$U^c \sim [0.20, 0.50]$	0.206457	0.452609	0.213604
Hidden layer	$U^d \sim [1, 3]$	3	1	3
Multiple	$U^d \sim [1, 3]$	3	3	3

Notes: This table shows the grid for the hyper-parameter search (Panel A) and the final validated values for the hyper-parameters per model (Panel B). U^c denotes the continuous uniform distribution, whereas U^d denotes the discrete uniform distribution. Avoiding overfitting is of key importance, and so we place emphasis on regularization parameters (L1) and different designs of dropout layers. The network architecture employs a baseline structure of halving the number of neurons over the hidden layers, following Gu et al. (2020). The minimum number of neurons in the first hidden layer is 32.

Following Gu et al. (2020), we assume that the number of neurons is halved across the hidden layers, i.e., 32 neurons in the first hidden layer and 16 in the second hidden layer, and so on. Therefore, we validate a multiple of a basic structure instead of validating the actual number of neurons. We take 32 neurons as the minimum for the first hidden layer. So for a multiple of 1 and three hidden layers we have 32-16-8 neurons. If we use a multiple of 3, we get 128-64-32. If only two hidden layers and a multiple of 2 are chosen, we get 64-32 as number of neurons. This approach provides us with great flexibility, but also ensures an efficient method for validating the depth of the neural network. To avoid overfitting, we also use early stopping, which stops training when validation loss increases by a certain number of iterations (called patience). In this study, we use a patience of 50, a maximum number of 500 epochs, and a batch size of 512.

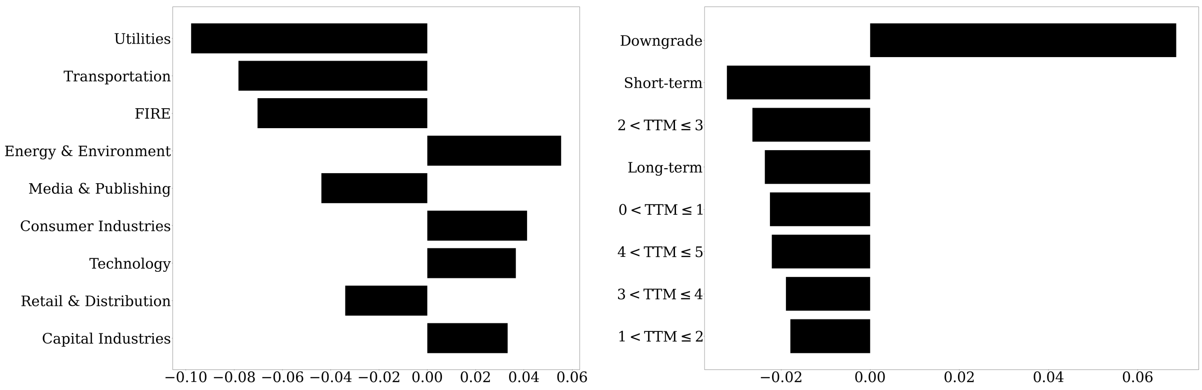
2.B APPENDIX | Control Variables

Figure 2.B.1 shows the first-order feature importance of the control variables. Panel A contains the importance of the industry controls, with reference category *Banking*. Panel B shows the importance of the bond-specific control variables. The reference category for total maturity of the bond (short-term: up to three years, medium-term: more than three years but less than or equal to ten years, long-term: more than ten years) is *medium-term*. The reference category for remaining time to maturity (*TTM*) in years is *5<TTM*.

Figure 2.B.1: First-order feature importance of control variables

Panel A: Industry controls

Panel B: Bond-specific controls



Notes: This figure displays the first-order feature importance of the control variables: industry classifiers (Panel A) and bond-specific features (Panel B).

2.C APPENDIX | Most Important Interactions

Table 2.C.1 shows the top 15 most important interactions of two features. Firm's leverage ($TLMTA$) in particular interacts with industry affiliation (utilities, media & publishing, and transportation) or rating downgrade. Also, three text feature interactions (BERT features #5 and #2) are among the list of top 15.

Table 2.C.1: Top 15 interactions

Variable 1	Variable 2	Joint Importance
$TLMTA$	Utilities	0.846
$RSIZE$	Downgrade	0.787
BERT feature #5	Downgrade	0.763
$RSIZE$	Utilities	0.723
$RSIZE$	FIRE	0.722
$SIGMA$	Downgrade	0.628
$SIGMA$	Utilities	0.615
$TLMTA$	Media & Publishing	0.610
BERT feature #2	Downgrade	0.598
MB	Utilities	0.570
$TLMTA$	Downgrade	0.562
$PRICE$	Retail & Distribution	0.554
$NIMTAAVG$	Downgrade	0.553
$TLMTA$	Transportation	0.543
BERT feature #2	Transportation	0.522

Notes: This table shows the joint importance of the top 15 interactions between features (other than interactions between industry controls only, bond-specific controls only, and industry and bond-specific features.

Chapter 3

The Impact of Qualitative Information on Corporate Creditworthiness

This chapter is a joint work with Maximilian Nagl[*] and Daniel Rösch[†], and corresponds to a working paper with the same name (it has been under review by the *Journal of Accounting and Economics*).

Abstract

This paper shows the non-linear impact of (qualitative) public information and (quantitative) firm variables on corporate creditworthiness. Topics extracted from the forward-looking MD&A sections of corporate filings contribute significantly to the explanation and prediction of credit ratings. We decipher the drivers of this performance gain using Explainable Artificial Intelligence (XAI) techniques and uncover higher-order effects and interactions between quantitative firm characteristics and qualitative textual topics. Stress-testing the topic distribution reveals that firms at the change-point between investment grade and non-investment grade tend to be more susceptible to changes in MD&A content. Our findings provide valuable insights into components of corporate creditworthiness that are relevant to all stakeholders who rely upon credit ratings.

Keywords: Explainable Artificial Intelligence; Rating Prediction; Structural Topic Modeling; Textual Analysis

JEL classification: C53; G21; G33; M40; M41

[*] University of Regensburg, Chair of Statistics and Risk Management, Universitätsstraße 31, 93040 Regensburg, Germany, email: maximilian.nagl@ur.de

[†] University of Regensburg, Chair of Statistics and Risk Management, Universitätsstraße 31, 93040 Regensburg, Germany, email: daniel.roesch@ur.de

3.1 Introduction

We investigate whether the forward-looking Management's Discussion & Analysis (MD&A) sections of quarterly and annual corporate filings provide valuable information for determining the creditworthiness of a company, which is commonly measured by credit ratings of major rating agencies. This is of major concern for various stakeholders in the economy, such as governments, financial institutions, regulators, employees, and the company itself.

3.1.1 Motivation

Ratings enable all participants in the financial markets to easily assess the risk properties of individual securities or corporates using a single and commonly known scale. In addition, ratings are extensively used in regulation and private contracting, and serve as a way to quantify and manage risk, see, e.g., Becker and Milbourn (2011). In general, they are used in investment decision processes, company valuations, allocation of regulatory capital, or determination of interest payments, see, e.g., Hilscher and Wilson (2017). For example, credit ratings are at the core of regulatory rules, especially for pension funds or money market funds, as they are only allowed to hold investment grade-rated securities. Additionally, the amount of regulatory capital required for financial institutions also depends heavily on the assigned credit rating. Hence, ratings depict a key instrument for reducing information asymmetry in financial markets, so that they are regarded as important in particular by regulators, legislators, issuers, and investors. Understanding ratings can therefore be understood as essential for the proper functioning of the financial system, see, e.g., Becker and Milbourn (2011) or Bonsall and Miller (2017). Credit ratings are also relevant for other common investors, as rating changes influence the long-run stock returns of firms, see, e.g., Dichev and Piotroski (2001) or Frankel et al. (2022). Rating triggers included in debt contracts directly affect terms such as collateral or pricing provisions when the associated rating changes (see, e.g., Kraft (2015)). In summary, ratings are important for almost any facet of the financial economy.

An important strand of research has focused on the determinants of credit ratings. Commonly used in the literature are quantitative measures such as accounting and market-based variables, see, for example Blume et al. (1998), Alp (2013), and Baghai et al. (2014). Following the outline of Moody's (2021b), rating agencies use quantitative variables to determine credit ratings, but also qualitative information such as market position, business strategy, competitive advantages, management quality, and salient features of the entity being rated, which could be contained in textual financial disclosures such as the MD&A sections. However, the impact of financial disclosures, particularly whether they provide independent additional information beyond the common quantitative drivers, is considerably less investigated when targeting the creditworthiness of companies. Moreover, the literature argues that these disclosures became increasingly less readable, which hampers investors from processing the information when making investment decisions, see, e.g., Li (2008), Miller (2010), Lehavy et al. (2011), Lawrence (2013), Lang and Stice-Lawrence (2015), Bonsall and Miller (2017), and Bonsall et al. (2017). Therefore, how to process textual information is becoming more important and requires advanced methods.

Along with the discussion *which* information to use, also the discussion on *how* the information is connected to creditworthiness is important. Recent literature suggests that in many financial disciplines, non-linearity and interactions play an important role. For example, when focusing on stock, bond, or hedge fund returns, from recent papers by Gu et al. (2020), Freyberger et al. (2020), Bianchi et al. (2021), or Wu et al. (2021) it is evident that non-linearity captured by machine learning methods can explain and predict a larger portion of return variation. Similar findings for earnings management can be found in Alhadab and Nguyen (2018) or Thanh et al. (2020), or for credit risk parameters, see, e.g., Sopitpongstorn et al. (2021) or Kellner et al. (2022). With respect to credit ratings, the vast majority of the literature uses linear models, see, e.g., Blume et al. (1998), Alp (2013), Baghai et al. (2014), Behr et al. (2018), Bonsall et al. (2017), and Donovan et al. (2021).

Summarizing our motivation, this paper is the first to allow non-linear relations and interactions between common and novel drivers of firms' creditworthiness. Furthermore, we are the first to consider non-linear and joint impacts of qualitative data.[1]

[1] Donovan et al. (2021) focuses on rating downgrades in part of their study and find their sentiment measure to have significant linear impact. However, we differ from this paper by allowing the qualitative information to have non-linear and joint impacts. As shown in the robustness analyses in Section 3.5, we find our approach to be superior to the sentiment measure of Donovan et al. (2021).

3.1.2 Main Results

Before outlining the related literature and presenting our methodologies in detail, the following section provides a high level summary of the main findings. We consider qualitative information using the MD&A section of corporate SEC-filings as a determinant of corporate creditworthiness. We apply an Artificial Neural Network (ANN) extension to model ordinal dependent variables, following Cao et al. (2020), which we refer to as Ordinal Artificial Neural Network (OANN) hereafter. This allows modeling any kind of non-linearity and interactions in the relation between the input variables and the latent response variable y^*, which can be understood as the latent creditworthiness of the company. Ratings are gathered from the Moody's (2021a) Default and Recovery Database from 1994 to 2020 and common quantitative drivers are calculated, following Blume et al. (1998), Alp (2013), and Baghai et al. (2014), among others. For the text data, a Structural Topic Model (STM) following Roberts et al. (2016) is used to convert the text data into topic probabilities. These quantify how likely an MD&A section is to contain a specific topic, such as financial risk, real estate, or sustainability. Therefore, the focus of this paper is on the content of the MD&A section and the evaluation whether the topics addressed in the MD&A section help explain and predict corporate creditworthiness.

We split our sample into a training period (1994Q3 to 2018Q4) and an out-of-time testing period (2019Q1-2020Q4) to test whether the text data can also help predict future credit ratings. In total, four models are compared. Table 3.1 shows the mean absolute error (MAE) of the rating predictions for both periods and all four models.[2]

Table 3.1: Mean absolute error of rating predictions

Training Period: 1994Q3 -2018Q4 — Testing Period: 2019Q1-2020Q4

	Ordered Logit	Ordered Logit + Text	OANN	OANN + Text
Training	0.566	0.537	0.504	**0.439**
Testing	0.536	0.492	0.494	**0.414**

The baseline Ordered Logit model, labeled as *Ordered Logit* relates the common quantitative drivers linearly to y^*. We then add the topic probabilities of the STM (*Ordered Logit + Text*) as additional explanatory variables. Following Table 3.1, the inclusion of text data relatively reduces the MAE by 5% for the training data and 8% for the testing data. Therefore, text data improve the rating prediction assuming a linear connection of topic probabilities and creditworthiness. Moreover, we estimate the novel neural network using only the quantitative

[2] SEC filings (10-K and 10-Q) have been available through EDGAR since 1994. As we use a lag of two quarters for the text data, the first rating information is gathered from 1994Q3.

drivers (*OANN*) and subsequently add the topic probabilities (*OANN + Text*) as explanatory features. The difference between *OANN* and *Ordered Logit* quantifies the amount of non-linearity and interactions in the common quantitative drivers. We find a performance increase of 11% in training and 8% in testing. This shows that non-linearities and interactions are present in the common quantitative determinants. Comparing *OANN* and *OANN + Text*, we observe a relative increase in performance of 13% for the training data and 17% for the testing data. This means that allowing non-linear connections of text data and creditworthiness can lift much more performance compared to a linear connection. Overall, incorporating textual data and allowing for any kind of non-linearities in the latent creditworthiness of companies via *OANN + Text* increases the performance in explaining and predicting the creditworthiness of firms by 22.7% in training and testing compared to *Ordered Logit*. Moreover, in all comparisons, the increase in performance is greater in the testing sample, implying that textual data is particularly important for predicting the future creditworthiness of firms.

In light of the additional findings in Section 3.5, our contribution to the literature is manifold. First, this paper shows the importance of features extracted from the texts including non-linear effects, and the level of interactions with commonly used financial variables such as leverage, firm size, or research & development expenses. Second, key findings related to the text data include structural differences in topic (probability) distributions across rating classes, suggesting differential effects of text on creditworthiness. Third, the empirical analysis documents that text data is especially important for firms at the change-point between investment grade and non-investment grade ratings. This may imply that the MD&A section can make the difference to receive an investment grade rating or not, providing valuable insights for stakeholders and investors in firms at the change-point. Fourth, we also find evidence that processing such data requires machine learning techniques to achieve the best possible results. The best performance in predicting ratings is obtained by applying an Ordinal Artificial Neural Network (OANN), which allows modeling all kinds of non-linearity and interactions in the relation between the input variables and the corporate creditworthiness.

The remainder of the paper is structured as follows. Section 3.2 gives a brief review of the relevant literature. Section 3.3 presents the data used in this study. Section 3.4 contains the methodology, i.e., an introduction to Structural Topic Modeling and Ordinal Artificial Neural Networks. Section 3.5 provides detailed results for the rating models applied. Section 3.6 discusses the findings and concludes.

3.2 Literature Review

The assessment of creditworthiness is at the beginning of many tasks and actions of stakeholders, including investment decision processes, company valuations, the allocation of economic and regulatory capital, or the determination of interest payments (see, e.g., Hilscher and Wilson (2017)). While financial institutions often prepare internal ratings for (future) customers targeting their actual risk of default, rating agencies such as Moody's or Standard & Poor's issue external ratings as independent opinions on the entity's creditworthiness. A common and consistent terminology, i.e., letter grades from "Aaa" to "C", provides a comprehensible valuation metric (see S&P Global Ratings (2019)) that is clearly understandable and allows potential rating deteriorations or improvements (including their outlook) to be tracked directly.

Because of this importance of ratings as a measure of creditworthiness for investors and other stakeholders who consider them in decision making, several studies examine the evolution of corporate credit ratings and their relation to changes in rating standards. Worse ratings or an accumulation of downgrades could be understood as declining credit quality of corporate debt, but Blume et al. (1998) note that this could also be seen as a consequence of stricter rating standards. In line, Alp (2013) finds evidence for a structural break in 2002, following the bursting of the Dot-com bubble and leading to more stringent ratings. He also documents that tighter rating standards are associated with lower default rates. Baghai et al. (2014) conclude that rating agencies became increasingly conservative in their assessment of credit risk over the observation period of 1985 to 2009, as average corporate ratings deteriorated by three notches. Behr et al. (2018) analyze the impact of government regulation on rating stability and possible changes in rating importance over time. As a further factor related to rating agencies decision process, Bonsall et al. (2017) measure the influence of a firm's management capabilities on its credit rating and document that higher managerial abilities are associated with lower credit risk assessment. Basu and Naughton (2020) examine the effect of changes in financial statement recognition on corporate credit ratings. Since rating agencies make quantitative adjustments to firms' accounting figures, revised accounting standards can have a direct impact on the company's credit rating. Hung et al. (2022) find evidence that global rating agencies with large market power tend to have stricter rating standards due to heightened reputational concerns.[3]

[3] Further papers in this context, concerning the development in rating standards or the independence of rating agencies, are Cafarelli (2020), Hwang and Kim (2020), Berwart et al. (2019), Bonsall et al. (2018), Kedia et al. (2017), Dilly and Mählmann (2016), or Kedia et al. (2014).

When determining credit ratings, rating agencies generally make use of a variety of available information sources. Besides non-public information provided by the company in confidence, such as statements on internal capital allocation or budgets and forecasts (see SEC (2003)), this includes publicly available data such as corporate disclosures. Of central importance are quantitative information, for example retrieved from accounting statements, that characterize the firm in terms of key financial figures. In addition, rating agencies also consult qualitative information in their rating process, which may include market position, business strategy, competitive advantages, management quality, and salient features of the entity being rated (see, e.g., Moody's (2021b)). Such data could be contained in textual financial disclosures such as the MD&A sections of corporate SEC reports.

Research on the composition of ratings and their key determinants has been subject of interest for a long time, see, e.g., Horrigan (1966), West (1970) or Pogue and Soldofsky (1969) as early contributions on constituents of bond ratings. The latter conclude that despite the importance of factors like expert judgments for the bond rating process, financial and operating ratios such as leverage and profit have significant explanatory power for rating differences. Besides targeting the relation between credit ratings and quantitative accounting ratios (see Kaplan and Urwitz (1979) or Blume et al. (1998)), studies of qualitative information relevant to the rating process, such as the quality and readability of corporate disclosures, contribute to the discussion of rating components (see, for example, Bonsall and Miller (2017)).

With the increasing computation power that has emerged in recent years, such qualitative information treasures can be tapped more and more, as research on text extraction methodologies and applications in Finance shows. For example, Friberg and Seiler (2017) construct indices of uncertainty and risk from the word lists of Loughran and Mcdonald (2011) based on 10-K reports filed to the SEC to explain corporates financial policies. Muslu et al. (2015) find that companies that disclose more forward-looking sentences in MD&A sections of their 10-K forms tend to be poor at releasing information prior to the filing date, which is also reflected in abnormal stock returns after the filing. Durnev and Mangen (2020) note that textual disclosures of companies are among the determinants of corporate investment. Cohen et al. (2020) measure the impact of structural and linguistic changes in quarterly and annual text filings on the companies' future returns. There seems to be a correlation, but investors appear to be missing the opportunity to take advantage of it. Also targeting the link between returns and sentiment extracted from disclosures, i.e., conference-call dates and 10-K filings, Frankel et al. (2022) find that ML methods such as random forests dominate dictionary-based approaches.

For a very similar set of data, Donovan et al. (2021) find that conference call transcripts that capture the interaction of the firms' managers with market participants provide important information about variations in a borrower's credit risk.

Alternatively, the approach of topic modeling is comparatively less used in the financial literature. Bao and Datta (2014) apply latent Dirichlet allocation (LDA) topic modeling, a simplified version of our approach, to measure risk types from textual disclosures in Section 1A of 10-K filings. They document that the majority of disclosed risk types are not informative. Moreover, Dyer et al. (2017) use LDA to investigate marked trends, such as length, boilerplate, stickiness, or redundancy, in full 10-K filings. We contribute to this line of literature by employing an advanced topic model to value the impact of the MD&A section on firm's creditworthiness.

This review on current literature states that former studies have examined the link between quantitative information and ratings. However, rating agencies consider qualitative text information in their rating processes, and there is limited evidence in the literature on the importance of this data for corporate ratings. This paper contributes to filling this gap by showing the impact of qualitative information on corporate creditworthiness. We extract topics from the forward-looking MD&A sections of corporate reports and incorporate them together with quantitative firm variables. By applying Ordinal Artificial Neural Networks, we allow non-linear relations and interactions between these drivers of firms' creditworthiness.

Our results suggest that public information contributes significantly to explaining and predicting credit ratings. Using Explainable Artificial Intelligence techniques, we uncover the drivers that lead to these performance improvements. Our findings provide valuable insights into the components of corporate creditworthiness that are important to all stakeholders who base their decisions on credit ratings.

3.3 Data

In this study, we collect a comprehensive data set of about 50,000 quarterly observations of U.S. corporate bonds over a period from 1994 to 2020[4]. Target variable is the bond rating (Aaa to Caa-C) obtained from the Moody's (2021a) Default and Recovery Database (see Table 3.2).

Table 3.2: Rating distribution in the data set

Panel A: Number									Panel B: Percentage							
Year	Aaa	Aa	A	Baa	Ba	B	Caa-C	Total	Aaa	Aa	A	Baa	Ba	B	Caa-C	Total
1994	26	71	225	147	59	91	2	621	4.2	11.4	36.2	23.7	9.5	14.7	0.3	100.0
1995	20	56	231	148	56	95	14	620	3.2	9.0	37.3	23.9	9.0	15.3	2.3	100.0
1996	18	45	238	200	80	196	15	792	2.3	5.7	30.1	25.3	10.1	24.7	1.9	100.0
1997	12	58	271	274	106	245	28	994	1.2	5.8	27.3	27.6	10.7	24.6	2.8	100.0
1998	12	59	289	343	132	289	43	1,167	1.0	5.1	24.8	29.4	11.3	24.8	3.7	100.0
1999	10	56	290	412	160	371	81	1,380	0.7	4.1	21.0	29.9	11.6	26.9	5.9	100.0
2000	17	57	305	443	192	425	99	1,538	1.1	3.7	19.8	28.8	12.5	27.6	6.4	100.0
2001	17	49	267	521	238	385	133	1,610	1.1	3.0	16.6	32.4	14.8	23.9	8.3	100.0
2002	16	54	273	564	282	398	169	1,756	0.9	3.1	15.5	32.1	16.1	22.7	9.6	100.0
2003	29	37	354	565	272	425	149	1,831	1.6	2.0	19.3	30.9	14.9	23.2	8.1	100.0
2004	37	41	344	636	269	461	134	1,922	1.9	2.1	17.9	33.1	14.0	24.0	7.0	100.0
2005	26	39	345	675	283	487	136	1,991	1.3	2.0	17.3	33.9	14.2	24.5	6.8	100.0
2006	30	43	334	633	313	460	134	1,947	1.5	2.2	17.2	32.5	16.1	23.6	6.9	100.0
2007	31	39	336	619	314	437	147	1,923	1.6	2.0	17.5	32.2	16.3	22.7	7.6	100.0
2008	10	40	359	637	331	427	226	2,030	0.5	2.0	17.7	31.4	16.3	21.0	11.1	100.0
2009	9	36	344	655	303	413	252	2,012	0.4	1.8	17.1	32.6	15.1	20.5	12.5	100.0
2010	12	43	357	665	347	465	207	2,096	0.6	2.1	17.0	31.7	16.6	22.2	9.9	100.0
2011	10	43	371	731	351	545	170	2,221	0.5	1.9	16.7	32.9	15.8	24.5	7.7	100.0
2012	11	35	370	781	386	555	154	2,292	0.5	1.5	16.1	34.1	16.8	24.2	6.7	100.0
2013	12	30	376	806	369	598	169	2,360	0.5	1.3	15.9	34.2	15.6	25.3	7.2	100.0
2014	12	34	413	823	463	545	178	2,468	0.5	1.4	16.7	33.3	18.8	22.1	7.2	100.0
2015	13	36	410	822	430	531	226	2,468	0.5	1.5	16.6	33.3	17.4	21.5	9.2	100.0
2016	17	40	368	807	443	522	202	2,399	0.7	1.7	15.3	33.6	18.5	21.8	8.4	100.0
2017	12	33	388	843	457	503	144	2,380	0.5	1.4	16.3	35.4	19.2	21.1	6.1	100.0
2018	12	36	382	872	454	470	110	2,336	0.5	1.5	16.4	37.3	19.4	20.1	4.7	100.0
2019	14	35	384	879	412	449	100	2,273	0.6	1.5	16.9	38.7	18.1	19.8	4.4	100.0
2020	6	16	141	348	162	133	56	862	0.7	1.9	16.4	40.4	18.8	15.4	6.5	100.0
Total	451	1,161	8,765	15,849	7,664	10,921	3,478	48,289	0.9	2.4	18.2	32.8	15.9	22.6	7.2	100.0

Note: This table presents the distribution of the Moody's ratings (bond quarters) per year over time. The data set counts 48,289 bond quarters in total. Panel A reports the number of bond quarters in the rating categories Aaa to Caa-C. The percentage of rating grades is displayed in Panel B.

Several accounting and market features, a total of 16 control variables (see Baghai et al. (2014) and Table 3.3) are collected from Refinitiv (2021) Eikon. To complete our database, we download the 10-K and 10-Q filings gathered by SEC (2021) in the EDGAR[5] database, and extract the MD&A sections. Our final data set is constructed combining the rating information, the financial control variables, and the textual MD&A information.

[4] We collect a great number of control variables that are not available for all observed firm quarters or text filings. For example, in computing Beta or recognizing EBITDA we lose some firm quarters. The final number of firm quarters, however, remains representative compared to existing literature.

[5] Via the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database, the U.S. Securities and Exchange Commission (see SEC (2021)) makes publicly available the documents such as annual or quarterly reports that companies are required to file under U.S. law.

Table 3.3: Control variables

Abbreviation	Definition
Leverage	The value of long- and short-term debt relative to the value of total assets (TA) as a measure of the company's leverage
Cash	The ratio of cash or marketable securities and TA
Debt/EBITDA	The long- and short-term debt relative to EBITDA
Interest Coverage	EBITDA relative to total interest
Profit	EBITDA divided by total sales
Profit vola	Volatility of EBITDA/Sales (last 4 quarters)
Size	Log of the book value of assets, in constant dollars
PPE	Net property, plant, and equipment divided by total assets
CAPEX	Capital expenditures divided by total assets
Operating Margin	Operating income divided by sales
R&D	Expenses of research and development divided by TA
Retained Earnings	Retained Earnings divided by TA
Dividend Dummy	1, if companies paid dividends
Debt/Ebitda Dummy	1, if Debt/Ebitda is lower than 0
Beta	Stock's beta, calculated from a market-model regression with daily returns (annually) using the Fama-French market index
Ivol	Idiosyncratic risk of the company, calculated as the annual root mean squared error from a regression of daily stock returns on the returns of the Fama-French market index

Note: This table introduces the description of the employed quantitative variables. We cover various accounting-based metrics, such as leverage, cash, or size of the company. Furthermore, we include market-based measures, such as the stock's beta and idiosyncratic risk.

Comparing the summary statistics of the investment grade group (ratings of BBB and above) with the non-investment group (see Table 3.4), significant differences can be observed. Bonds rated BB and below have higher leverage and higher debt relative to EBITDA, suggesting that firms choosing a riskier capital structure tend to be less well rated than equity-financed companies. A higher EBITDA to Sales volatility of worse rated companies indicates non-constant profits which could lead to the firm's business model to be perceived as not sufficiently robust by rating agencies. On average, firms with lower ratings have a higher capital commitment in PPE or capital expenditures (relative to total assets). Higher Beta and higher idiosyncratic risk also indicate higher-risk business activities, which are assigned lower ratings. Bonds in the investment grade group have slightly lower liquidity (Cash/Assets), lower EBITDA to Sales, lower book value of assets, lower operating income to sales, lower R&D expenses to assets, and lower retained earnings to assets.

Table 3.4: Summary statistics of features

	Panel A: Investment grade (Count: 26,226)					Panel B: Non-investment grade (Count: 22,063)					Panel C: T-statistic
Feature	Mean	25%	Median	75%	Std. dev.	Mean	25%	Median	75%	Std. dev.	
Beta	0.698	0.282	0.741	1.055	0.512	0.735	0.037	0.748	1.279	0.717	6.58 ***
CAPEX	0.014	0.005	0.011	0.019	0.014	0.015	0.004	0.009	0.017	0.018	2.58 ***
Cash	0.071	0.011	0.035	0.095	0.092	0.069	0.015	0.043	0.095	0.079	-2.68 ***
Debt/EBITDA	13.543	5.273	8.896	15.783	18.139	20.746	8.709	14.802	23.124	24.736	36.84 ***
Interest Coverage	11.749	2.601	5.815	9.153	15.983	8.748	1.640	4.619	8.440	12.829	-22.46 ***
Ivol	0.015	0.010	0.013	0.018	0.008	0.029	0.016	0.023	0.033	0.022	90.46 ***
Leverage	0.313	0.210	0.299	0.398	0.144	0.460	0.310	0.429	0.577	0.216	89.03 ***
Operating Margin	0.149	0.077	0.139	0.222	0.171	0.049	0.019	0.071	0.137	0.252	-51.63 ***
PPE	0.344	0.097	0.247	0.567	0.290	0.352	0.119	0.288	0.547	0.271	3.29 ***
Profit	0.263	0.136	0.217	0.344	0.185	0.194	0.077	0.141	0.255	0.191	-40.00 ***
Profit vola	0.045	0.014	0.028	0.052	0.062	0.059	0.014	0.030	0.063	0.087	20.27 ***
R&D	0.003	0.000	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.004	-23.97 ***
Retained Earnings	0.282	0.095	0.257	0.450	0.286	-0.008	-0.118	0.051	0.212	0.397	-93.27 ***
Size	23.007	22.087	22.882	23.836	1.267	21.683	20.881	21.642	22.501	1.154	-119.12 ***

Note: This table introduces the summary statistics for the features used in the rating models. The count, mean, 25%-quantile, median, 75%-quantile, and standard deviation (std. dev.) are shown for the investment grade group (Panel A) and the non-investment grade group (Panel B). Panel C of the table shows the statistics of the t-test for the difference in means of both groups (see Welch (1947)). The significance is indicated for the 1% (***), 5% (**) and 10% (*) level.

3.4 Methods

When dealing with high dimensional text data, a very common approach is to use a predefined word list, count the occurrence of these (positive or negative) words, and compute a so-called sentiment index to capture the polarity of the text data, see, e.g., Loughran and Mcdonald (2011). However, there are two challenges involved. First, one must define these words a priori or use a predefined word list. Since words can have different meanings in different contexts, the information content critically depends on choosing the appropriate words. Second, the word list cannot take into account the different language use of different actors or texts from different segments. Ideally, one needs to define a set of words for each segment, e.g., industry sector, and each application, e.g., textual data of corporate filings as outlined by Loughran and McDonald (2016).

Unsupervised learning methods such as Principal Component Analysis (PCA) or word embeddings, on the other hand, allow more flexibility and greater information content. However, these approaches are black boxes in the sense that we cannot assess which words are important or how textual information is incorporated into the model. PCA, for example, reduces the words into orthogonal components that completely lose their interpretability. Moreover, it is not straight forward to control for overfitting in the training sample in the sense that the estimated components also recover the same amount of variance for the test sample. Word embeddings transform each word into a word vector. These vectors cannot be used directly in standard regression techniques unless they are aggregated into a total document vector, see, e.g., Li et al. (2021).

3.4.1 The Structural Topic Model

Topic models, such as the Structural Topic Model (STM) by Roberts et al. (2016), provide a good balance between flexibility and interpretability and additionally control for overfitting. They reduce the dimension of high-dimensional text data to a small number of topics, similar to the number of components of a PCA, but allow us to evaluate which words are characteristic for these topics. Therefore, we can reduce the dimensionality of our text data while maintaining human-interpretable outcomes. Furthermore, the STM allows us to control for differences between industry sectors, which may be important in our application, see, e.g., Miller (2017) or Huang et al. (2018).[6]

Following Roberts et al. (2016), topic models are based on two central assumptions. First, all documents (in our case, the MD&A sections) are a mixture of latent topics. Every topic is represented with a certain probability in each document, which is called topic prevalence. Hence, the topic prevalence sums up to 1 for each document. Second, each topic is a mixture of words, meaning that every word occurs with a certain probability in that topic, which is called topic content and sums up to 1 for each topic. We can add information from our data, e.g., industry classification, which may influence how likely a topic occurs in a text corpus and which words express that topic. Hence, we can account for different values of topic prevalence across

[6] Earlier variants of topic models also exist, such as the Correlated Topic Model (CTM) by Blei and Lafferty (2007) or the Latent Dirichlet Allocation (LDA) by Blei et al. (2003). The STM is a recent extension of these approaches, as it addresses many of their drawbacks. The LDA assumes independence of topics, which is doubtful in empirical analysis. The CTM provides an extension in this regard by allowing correlated topics. However, LDA and CTM tend to identify less informative topics because they ignore the existence of idiosyncratic language within the text data, such as industry-specific language. This would result in topics which are representative for different industries and therefore their information can be easily captured by industry dummies. The STM can distinguish between different language styles and thus is likely to provide more additional information.

industries, e.g., a company from the Energy & Environment sector might be more likely to write about sustainability than a company from the Technology sector. In addition, we can also consider different words describing the topic, e.g., a company from the Energy & Environment sector may use different words to write about sustainability compared to a company from the Technology sector.

The main parameter of interest in our analysis is the topic prevalence of every MD&A section, i.e., the probability of each topic to occur in an MD&A section.

Following Roberts et al. (2016), we consider a topic k in document i ($k = 1,\dots,K$; $i = 1,\dots,N$).[7] Let $\boldsymbol{z}_i$ be the multinomial distributed assignment of the topics for document i:

$$\boldsymbol{z}_i \sim Multinomial_K(\boldsymbol{\theta}_i), \tag{3.1}$$

where $\boldsymbol{\theta}_i \in [0,1]^K$ is a vector which contains the probabilities of all K topics to occur in document i (the topic prevalence).

Moreover, consider a word v ($v = 1,\dots,V$), given a topic k in document i. Let $\boldsymbol{w}_i$ be the multinomial distributed assignment of words, given topic k in document i:

$$\boldsymbol{w}_i \sim Multinomial_V(\boldsymbol{\psi}_{i,k}), \tag{3.2}$$

where $\boldsymbol{\psi}_{i,k} \in [0,1]^V$ contains the probabilities of all V words to be generated by topic k in document i (the topic content).

A set of parameters of interest is the topic prevalence for which a Logistic Normal distribution is used:

$$\boldsymbol{\theta}_i \sim LogisticNormal_{K-1}(\boldsymbol{\eta}_i \cdot \boldsymbol{\lambda}, \Sigma), \tag{3.3}$$

where $\boldsymbol{\eta}_i \in \mathbb{R}^P$ is the covariate vector with the identifiers (dummies) for the P industries for each document i. $\boldsymbol{\lambda} \in \mathbb{R}^{P \times K-1}$ is the weight vector for the covariate matrix $\boldsymbol{\eta}_i$.[8] The linear predictor $\boldsymbol{\eta}_i \cdot \boldsymbol{\lambda}$ is transformed using the logistic cdf to ensure that each topic probability is between zero and one. $\Sigma \in \mathbb{R}^{K-1 \times K-1}$ is a full rank covariance matrix to allow for correlations among topics.

[7] As we use only observations with an available MD&A section, the number of documents coincides with the number of firm-quarter observations. Therefore, the index i can interchangeably be used for document or observation.

[8] Note that because the topic prevalence sums up to 1 for each document, we need to model only $K-1$ probabilities.

Another set of parameters of interest are the probabilities of the topic content which are modeled as

$$\psi_{i,k,v} = \frac{\exp(m_v + \kappa_{k,v} + \kappa_{\eta,v} + \kappa_{\eta,k,v})}{\sum_{v=1}^{V} \exp(m_v + \kappa_{k,v} + \kappa_{\eta,v} + \kappa_{\eta,k,v})}, \tag{3.4}$$

where $\psi_{i,k,v} \in [0,1]$ is the probability of word v to be generated by topic k in document i. m_v is the baseline probability based on the log-transformed rate of word v in the overall corpus. $\kappa_{k,v}$ is the topic-specific deviation from m_v. $\kappa_{\eta,v}$ is the covariate-specific deviation from m_v and $\kappa_{\eta,k,v}$ is the interaction between $\kappa_{k,v}$ and $\kappa_{\eta,v}$.

The STM is estimated using a variational expectation-maximization algorithm following Roberts et al. (2016) and is implemented in the R-package *stm* by Roberts et al. (2019). The number of iterations is set to 500, which ensures convergence in our application. The number of topics is determined in a data-driven manner, as outlined in Section 3.5.

3.4.2 The Ordinal Artificial Neural Network

To analyze the determinants of ratings, we use an Ordered Logit model and its extension via an artificial neural network where the ratings are modeled as a function of firm characteristics and text data. Both approaches assume a latent variable $y_i^* \in \mathbb{R}$, which can be understood as the latent creditworthiness underlying the rating process for observation i ($i = 1,\ldots,N$).

The firm characteristics and text data are directly linked to y_i^* and thus influence the latent creditworthiness. Formally, the Ordered Logit model for y_i^* with $i = 1,\ldots,N$ is given by

$$y_i^* = \boldsymbol{\beta} \cdot \boldsymbol{x}_i + \boldsymbol{\phi} \cdot \boldsymbol{\theta}_i + \epsilon_i \tag{3.5}$$

and

$$R_i = \begin{cases} 0 & \text{if } y_i^* \in (-\infty, \mu_1) \\ 1 & \text{if } y_i^* \in [\mu_1, \mu_2) \\ \vdots & \\ 5 & \text{if } y_i^* \in [\mu_5, \mu_6) \\ 6 & \text{if } y_i^* \in [\mu_6, \infty) \quad . \end{cases} \tag{3.6}$$

The vector $\boldsymbol{x}_i \in \mathbb{R}^F$ contains the F quantitative firm characteristics for observation i ($i = 1, \ldots, N$), including dummies for the industries. Coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{1\times F}$ accounts for the strength of influence of the firm characteristics. Vector $\boldsymbol{\theta}_i \in \mathbb{R}^{K-1}$ contains the topic prevalence from above for observation (document) i. Since the topic prevalence sums up to 1 for each observation, we omit one topic to remove redundant information and avoid multicollinearity. Coefficient vector $\boldsymbol{\phi} \in \mathbb{R}^{1\times K-1}$ accounts for the strength of each topic's influence on the latent creditworthiness y_i^*. The random error $\epsilon_i \in \mathbb{R}$ is assumed to follow a logistic distribution.[9] Variable $R_i \in \{0, 1, 2, 3, 4, 5, 6\}$ denotes the long-term rating of firm i, where 0 corresponds to a Caa-C rating and 6 to Aaa. Additionally, $\{\mu_1, \ldots, \mu_6\}$ are partition points to link the latent variable to the rating categories. Hence, the Ordered Logit model directly and linearly links the firm characteristics and text data (via topic prevalence) to the latent creditworthiness.

In the machine learning literature, the focus on modeling ordinal outcomes has steadily increased. Cao et al. (2020) provide a simple but very efficient approach to modeling ordinal outcomes via artificial neural networks while ensuring rank consistent probabilities. For a discussion of other approaches to modeling ordinal outcomes with ANNs and their drawbacks, we refer to Cao et al. (2020). We label this Ordinal Artificial Neural Network as OANN. The OANN allows for non-linearities and interactions in the latent creditworthiness of each firm.

It starts with F firm characteristics in vector $\boldsymbol{x}_i$ and $K-1$ topic probabilities in vector $\boldsymbol{\theta}_i$ for observation i ($i = 1, \ldots, N$) as inputs, similarly to the above Logit model. Hence, this input layer has dimension $F + K - 1$. The network then consists of L stacked hidden layers, where each layer l ($l = 1, \ldots, L$) consists of M_l neurons $\boldsymbol{h}_l^{(i)} = (h_{l,1}^{(i)}, h_{l,2}^{(i)}, \ldots, h_{l,M_l}^{(i)}) \in \mathbb{R}^{M_l}$ determined by an affine combination of neurons in the previous layer which is composed of an arbitrary (non-linear) activation function σ, given as

$$\boldsymbol{h}_l^{(i)} = \sigma\Big(\boldsymbol{W}_l \boldsymbol{h}_{l-1}^{(i)} + \boldsymbol{b}_l\Big), \tag{3.7}$$

with $\boldsymbol{W}_l \in \mathbb{R}^{M_l \times M_{l-1}}, \boldsymbol{b}_l \in \mathbb{R}^{M_l}$ as parameters, usually referred to as weights and biases. Note that each neuron is a function of the inputs and therefore different for each observation i ($i = 1, \ldots, N$).

The latent creditworthiness is derived from the last hidden layer L with weights $\boldsymbol{W}_L \in \mathbb{R}^{1\times M_{L-1}}$ and bias $b_L \in \mathbb{R}$. Therefore, one neuron is specified to define the latent creditworthiness

[9] Alternatively, one could use an Ordered Probit model, which assumes a normal distribution of ϵ_i. However, the differences are negligible in our application.

$y_i^{*,OANN} \in \mathbb{R}$ as linear function

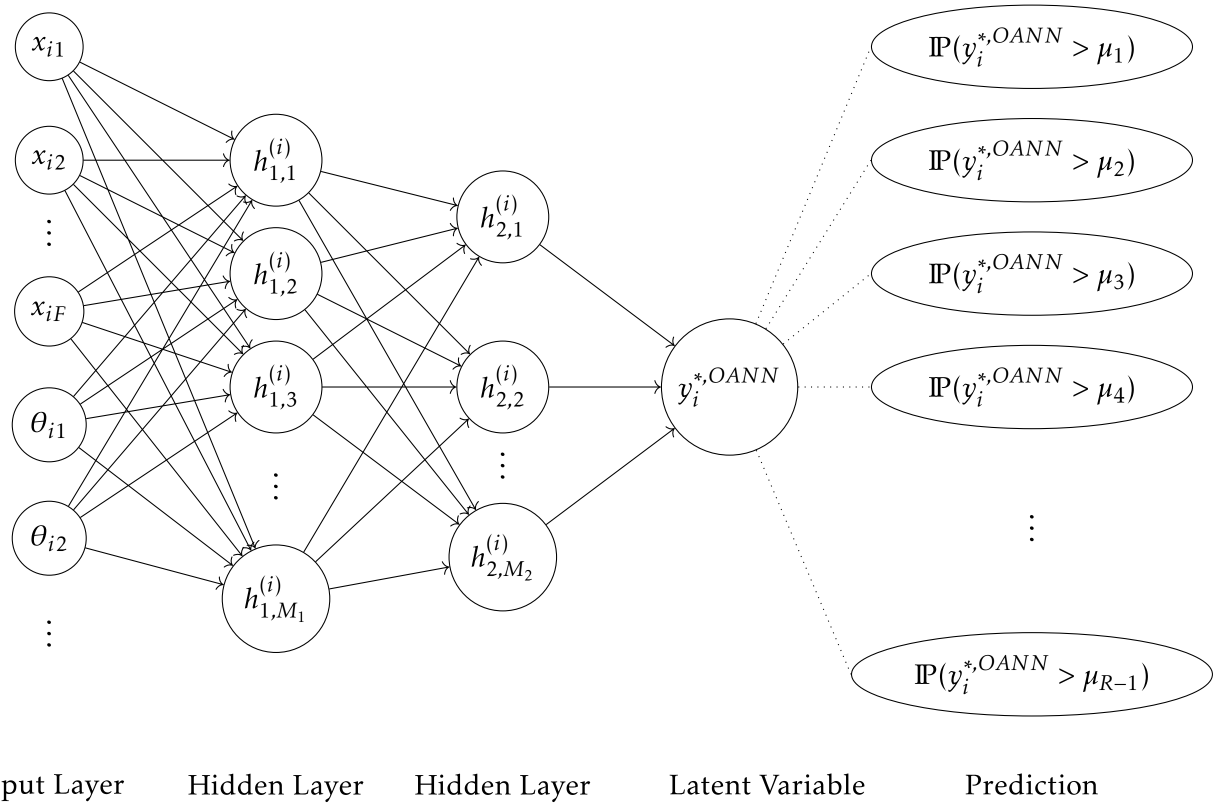
$$y_i^{*,OANN} = \boldsymbol{W}_L \boldsymbol{h}_{L-1}^{(i)} + b_L. \tag{3.8}$$

To obtain the final probabilities for every rating class, the last layer defines $R-1$ output nodes each with a sigmoid (logistic) activation function, yielding

$$\boldsymbol{P}_{R_i} = \ sigmoid(y_i^{*,OANN} + \boldsymbol{b}_{L+1}), \tag{3.9}$$

where $\boldsymbol{b}_{L+1} \in \mathbb{R}^{R-1}$ contains $R-1$ bias terms. $\boldsymbol{P}_{R_i} \in \mathbb{R}^{R-1}$ is a vector with the probabilities of $y_i^{*,OANN}$ being larger than the thresholds $\mu_1, \ldots, \mu_6$. Cao et al. (2020) show that adding these bias terms generates rank consistent probabilities. For a graphical illustration of a network with two hidden layers, see Figure 3.1. Neural networks have become widespread over the past years. However, a major issue is still the lack of interpretability. Therefore, this paper employs techniques used by Kellner et al. (2022) to provide a high degree of interpretability. The approaches focus on the "learned" relations of the neural network and are therefore estimated

Figure 3.1: Graphical overview — OANN



Note: This figure shows the overall structure of an Ordinal Artificial Neural Network. It starts with an Input Layer which contains the quantitative and qualitative information of the firms. These are subsequently processed by non-linear transformations up to the latent creditworthiness $y_i^{*,OANN}$. Finally, we derive the probability of falling into a rating class by adding the rating class specific biases to $y_i^{*,OANN}$ and transform this value with the sigmoid cdf.

using the training data. In summary, three different measures are used to explain the OANN. The first-order feature importance $FI^{First}(\xi_s)$ quantifies the overall importance of an input variable ξ_s, where $s = 1, \dots, F + K - 1$, i.e., for all F firm characteristics and $K - 1$ topic probabilities.

The first-order feature importance is given by:

$$FI^{First}(\xi_s) = \frac{1}{C} \operatorname{sgn}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial y_i^{*,OANN}}{\partial \xi_{is}}\right)\right)\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial y_i^{*,OANN}}{\partial \xi_{is}}\right)^2}. \tag{3.10}$$

$FI^{First}(\xi_s)$ is the feature importance of covariate ξ_s with respect to the latent creditworthiness $y_i^{*,OANN}$, and C is a normalizing constant that ensures that $\sum_{s=1}^{F+K-1} |FI^{First}(\xi_s)| = 1$. The sgn$(\cdot)$ operator defines the direction in which the feature drives the prediction. All variables must be standardized, e.g., by mean-scaling, to allow comparison. The gradients are squared to avoid cancellation of positive and negative values. $FI^{First}(\xi_s)$ quantifies how a change in ξ_s influences $y_i^{*,OANN}$. In statistics, this is commonly known as (average) marginal effect of ξ_s. Additionally, we calculate the second partial derivative with respect to the same input feature to determine the (single) non-linear impact $FI^{Second}(\xi_s)$ as

$$FI^{Second}(\xi_s) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}\left(\frac{\partial^2 y_i^{*,OANN}}{\partial \xi_{is}\partial \xi_{is}}\right)^2}, \tag{3.11}$$

Moreover, we may suspect that some input features have a joint impact, i.e., interact with each other. Therefore, $FI^{Joint}(\xi_{st})$ quantifies the strength of joint effects of two variables $s, t = 1, \dots, F + K - 1$ (interactions) as

$$FI^{Joint}(\xi_{st}) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}\left(\frac{\partial^2 y_i^{*,OANN}}{\partial \xi_{is}\partial \xi_{it}}\right)^2}. \tag{3.12}$$

Since we are more interested in whether there is a joint impact and how strong it may be, we neglect the direction of impact.

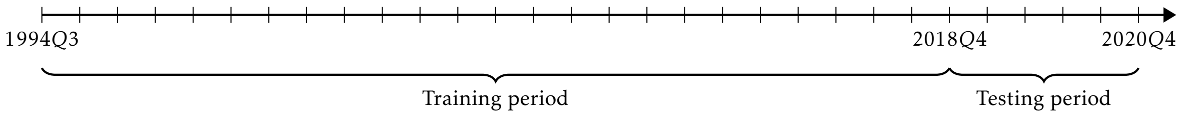
When $FI^{Joint}(\xi_{st})$ and $FI^{Second}(\xi_s)$ are close to zero, we can conjecture the absence of single non-linear and joint impacts of the input variables. This leaves only a linear impact, which corresponds to the linear Ordered Logit model.

3.5 Results

3.5.1 Structural Topic Model

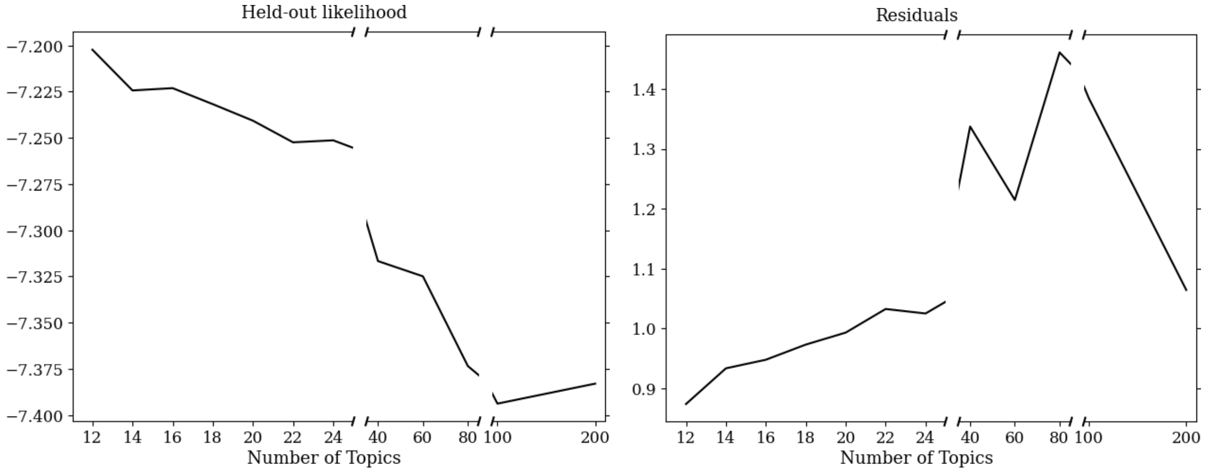
The main results in terms of performance are already presented in Section 3.1.2. Overall, we document a superior performance of the *OANN + Text*, which implies non-linearity and interactions to be present in quantitative and qualitative variables. We split our sample along the time axis into a training sample from 1994 to 2018 and the testing sample in 2019 and 2020, as outlined in Figure 3.2. This gives us the opportunity to evaluate the drivers for the sample from 1994 to 2018 and, subsequently, provide an out-of-time test to verify that the learned relations can also be recovered in future periods and are not a result of spurious overfitting.

Figure 3.2: Timeline



Similar to the PCA, we need to choose the number of topics (components) in advance. For PCA, a level of explained variance is usually set to determine the number of components. Similar metrics exist for the STM.

Figure 3.3: Validating the number of topics



The first is the held-out likelihood, which controls for overfitting of the STM, i.e., selecting too many topics. This metric is calculated by holding back a subset of the texts, fitting the STM for the remaining samples, and computing the (predicted) likelihood for the held-out set. The higher this likelihood, the lower the overfitting. The second is the residual value introduced by Taddy (2012). This metric is based on the idea that the STM is correctly specified when the likelihood in the multinomial model implies a variance of 1 in the residuals. Therefore, the residual metric calculates the sample variance of the fitted STM, which should then be close to 1. Figure 3.3 shows the calculated metrics for various numbers of topics. Please note that the x-axis does not scale linear. Similar to other unsupervised dimensionality reduction methods, there is no unique solution for selecting the number of dimensions. Starting with the held-out likelihood, the values are very similar, but for more than 40 topics, the likelihood decreases. Focusing on the residual, we observe a value close to 1 around 20 topics. Therefore, we use 20 topics for further analysis.[10]

Table 3.5 shows the ten most expressive words for every topic. Usually, one is tempted to give these topics a definitive label. However, labeling is in the eye of the beholder and thus may vary from reader to reader. Therefore, we only highlight those words in bold that we believe give an indication of the meaning of the topic. However, there are some topics where the meaning may be clearer, such as Topic 11 with words like *crude*, *nymex*, *gas*, *pipelines*, *wti*, *drill*, and *oil*, which could refer to fossil energy. Furthermore, Topic 19 contains words such as *lenders*, *secured*, *borrowers*, *refinancing*, *prepay*, *guarantors*, *loans*, *paydown*, and *undrawn*, all of which are about finance-related subjects. Similar clearer meaning may also be found in Topic 2 (audit-related), Topic 5 (real estate-related), Topic 6 (legal-related), Topic 10 (finance-related), Topic 12 (airline-related), Topic 13 (finance-related), Topic 14 (biotech-related), Topic 16 (energy-related), Topic 17 (digitalization-related), and Topic 20 (foreign finance-related). One could argue that the potential meaning of these topics is related to certain industries in some cases and therefore they serve as industry control variables. However, these topics go beyond the traditional industry classification as they are allowed to occur simultaneously in one document. For example, a company in the capital industry may write about energy or biotechnology related topics when discussing investment opportunities or potential concerns with current investments. Similarly, a technology company may write about digitalization as its core business and at the same time write about legal concerns and its financial situation. All of these meanings can be captured with the different topic probabilities for each document.

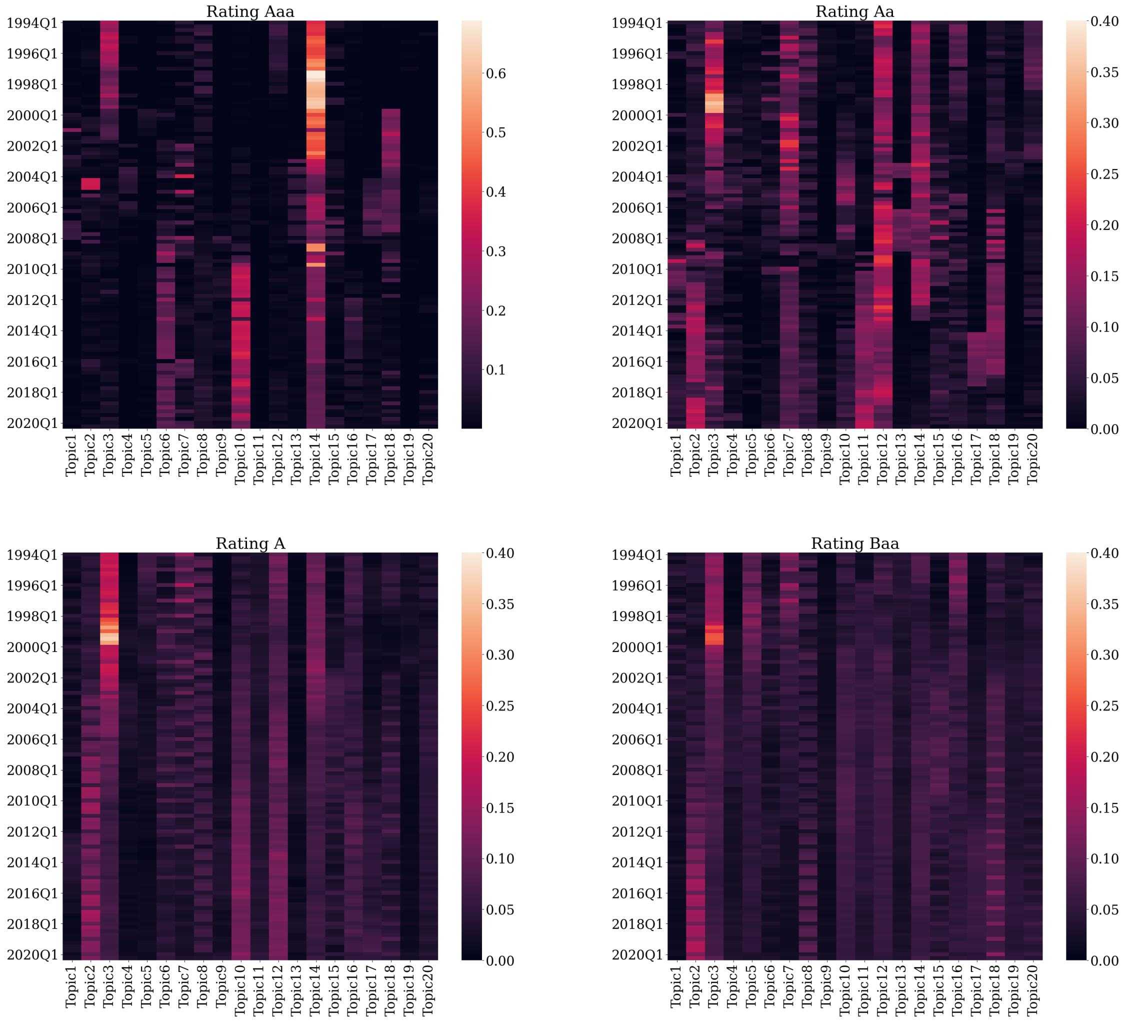
[10] We also tested a lower and higher number of topics, but the results are very similar. Therefore, we only report the results when using 20 topics.

Table 3.5: Expressive words

Topic 1	xbrl; registrants; filer; cal; text; htm; domain; certify; certifies; antidilutive;	**Topic 11**	**crude**; **nymex**; **gas**; **pipelines**; **wti**; tiered; lateral; **drill**; shallow; **oil**;
Topic 2	**gaap**; **reconciliations**; **condensed**; isolation; **adjusted**; rounding; measures; **unaudited**; understand; **comparability**;	**Topic 12**	percent; **cabin**; **airline**; configuration; **flight**; **navigation**; **aviation**; **overhaul**; **flying**; **cargo**;
Topic 3	**mainframe**; readiness; compliant; **interfaces**; contacted; prioritization; **chips**; **coding**; **computer**; **microprocessors**;	**Topic 13**	**bondholder**; **insurer**; inaccuracies; approves; freely; emerged; **warrants**; **certificates**; **guarantors**; **repayable**;
Topic 4	seamless; react; **optimism**; harm; **specializing**; prioritization; **chips**; coding; computer; **lifestyle**;	**Topic 14**	**chemistry**; **sensing**; **ipr**; converters; **syndrome**; standardization; **oncology**; **biopharmaceutical**; formulations; **biotechnology**;
Topic 5	**payoffs**; **condominiums**; **village**; increment; **occupancies**; **renovation**; **parcels**; leasable; **reit**; **undeveloped**;	**Topic 15**	**deficiency**; **incidents**; hierarchy; **corroborated**; **volatilities**; **sec**; **framework**; selecting; rarely; approximates;
Topic 6	**misrepresentation**; **injunctive**; **alleges**; estates; **putative**; **motion**; **plaintiffs**; enrichment; **compel**; **enforced**;	**Topic 16**	**crude**; **oils**; **refinery**; restarted; explosion; **ethanol**; **tonnage**; **gasoline**; railroads; **propylene**;
Topic 7	**lawfully**; knew; director; **unenforceable**; message; **undersigned**; **accountants**; personally; **controller**; exhibit;	**Topic 17**	**cancellable**; **tablets**; audio; **smartphones**; **entertainment**; **video**; **subscribers**; **membership**; **streams**; **networks**;
Topic 8	predicts; ended; months; **cautioned**; administrative; **disclaim**; **expressions**; big; percentage; salary;	**Topic 18**	**uncollectible**; characteristic; sensitivity; **portfolios**; retrospective; triggers; disciplined; **derecognized**; **retrospective**; inherently;
Topic 9	satellite; overruns; churn; depreciate; deployed; constructed; wireline; wireless; disadvantage; sprint;	**Topic 19**	**lenders**; **secured**; **borrowers**; **refinancing**; extinguishment; **prepay**; **guarantors**; **loans**; **paydown**; **undrawn**;
Topic 10	**fiscal**; **deflationary**; **buybacks**; reoccur; predicated; **underfunded**; **repatriating**; indices; sluggish; **multicurrency**	**Topic 20**	**recapture**; **vacated**; **pervasive**; **endangerment**; coalition; **midwest**; **escalations**; **locational**; facilities; rejection;

To give a brief overview of the variation in these topics over time, we plot the quarterly mean topic probabilities for each rating class from 1994 until 2020. The values for 2019 and 2020 are predictions based on the fitted STM from 1994 to 2018. Therefore, we get an understanding which topics were frequently discussed at a certain period of time.

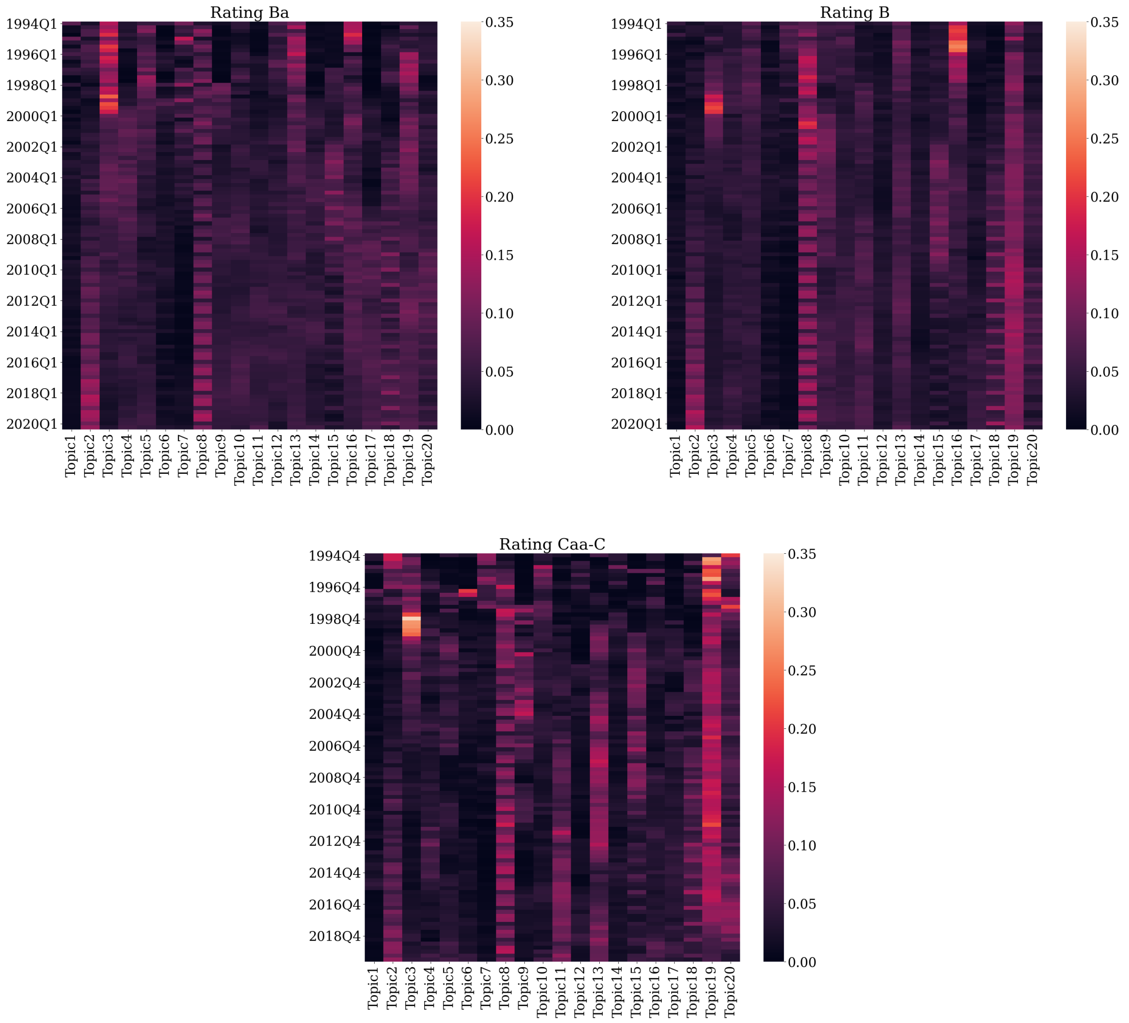
Figure 3.4: Topic probabilities over time for every rating class — investment grade



Note: This figure provides the topic probabilities over time for each rating class of the investment grade group. The probabilities are calculated by taking the quarterly average of all available topic probabilities in every rating class.

Interestingly, Topic 3 (mainframe, chips, computer) has a high proportion for all investment grade ratings except Aaa before the bursting of the Dot-com bubble. This may be plausible since at that time products such as computers and chips were of major interest to the vast majority of companies and therefore may be heavily discussed in the (forward-looking) MD&A section. For Aaa, we observe high probabilities of Topic 14 (chemistry, ipr, biotechnology) early on and of Topic 10 (fiscal, buybacks, repatriating) more recently. In addition, Topic 6 (injunctive, alleges, motion, plaintiffs) is more prevalent. The picture for Aa is more diverse as we observe large probabilities for Topic 2 (gaap, condensed, unaudited), Topic 12 (airline, cabin, aviation), and Topic 11 (crude, gas, oil). This may be plausible in that the letter two topics can be related, as fossil energy may be of high interest when it comes to airline-related subjects. The picture for A and Baa is more similar as the probability for Topic 2 (gaap, condensed, unaudited) increases considerably for more recent time periods.

Figure 3.5: Topic probabilities over time for every rating class — non-investment grade



Note: This figure provides the topic probabilities over time for each rating class of the non-investment grade group. The probabilities are calculated by taking the quarterly average of all available topic probabilities in every rating class.

Similar to investment grade group, Topic 3 (mainframe, chips, computer) has a high probability for all non-investment grade ratings before the bursting of the Dot-com bubble. Hence, this topic is dominant for all rating classes around this time. For Ba, we observe high proportions of Topic 8 (cautioned, disclaim, expressions) in conjunction with Topic 18 (uncollectible, portfolios, derecognized), both of which may express concerns rather than opportunities. For B, a large share of probability goes into Topic 2 (gaap, condensed, unaudited) in more recent times, and

Topic 8 (cautioned, disclaim, expressions) is strongly represented throughout the whole sample period. Furthermore, Topic 19 (lenders, secured, guarantors) becomes more likely over time. In the lowest rating category, we find a large mix of topics, such as Topic 8 (cautioned, disclaim, expressions), Topic 11 (crude, gas, oil), Topic 13 (bondholder, insurer, warrants), Topic 19 (lenders, secured, guarantors), and Topic 20 (recapture, midwest, escalations). In summary, we may argue that finance related-topics are more prevalent in lower ratings than in investment grade ratings, in particular, and the words associated with these topics may have more negative connotations.

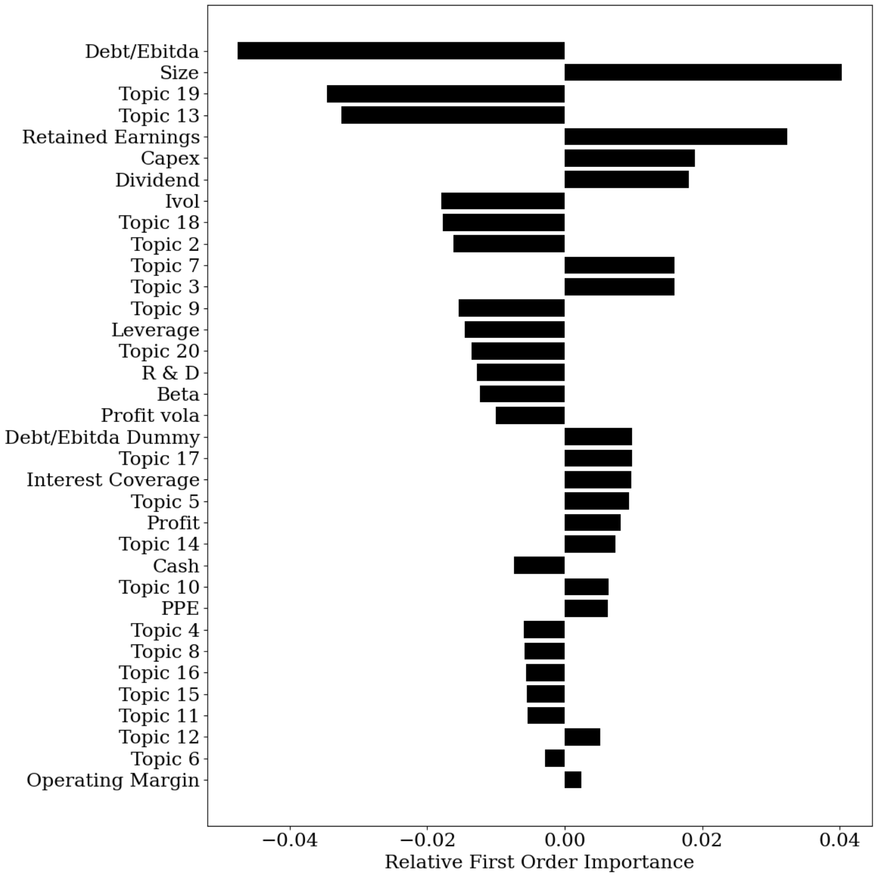
3.5.2 Ordinal Artificial Neural Network

As outlined in Section 3.4, we omit one topic to avoid multicollinearity. For the remainder of this section, we drop Topic 1 as it has a very low probability over the full period and the meaning of this topic is not very clear. The hyperparameters of the two OANNs with and without text data are validated using cross-validation. A detailed description of the hyperparameter selection can be found in Appendix 3.A.

Next, we turn to the factors that affect the creditworthiness of firms. We begin with the first-order feature importance $FI^{First}(\xi_s)$, which quantifies the overall importance of a single variable, followed by the second-order $FI^{Second}(\xi_s)$ and the joint feature importance $FI^{Joint}(\xi_{st})$, both of which provide insight into the actual shape of the relation between drivers and creditworthiness. In the remaining section, we discuss the importance of the OANN with text data. The feature importance for firm characteristics and industry classifiers is virtually the same for the OANN without text data. This may imply that the text data provide independent information beyond these two sets of drivers. For the first-order importance, we illustrate the importance of the industry classifier in Appendix 3.B. Overall, the industry classification accounts for roughly 50 % of the overall importance, and the direction of impact is economically plausible. For example, the Utilities industry has a large positive importance, which means that we can expect a higher creditworthiness in this sector. This makes sense as these companies are usually backed by governments or states, which is likely to increase their creditworthiness.

Figure 3.6 shows the importance of the remaining variables. Overall, the firm characteristics account for roughly 30% and the topic probabilities for about 20%. This result may be expected as we observe a substantial boost in performance when text data is included in the OANN. Positive feature importance implies higher creditworthiness and vice versa.

Figure 3.6: First-order feature importance



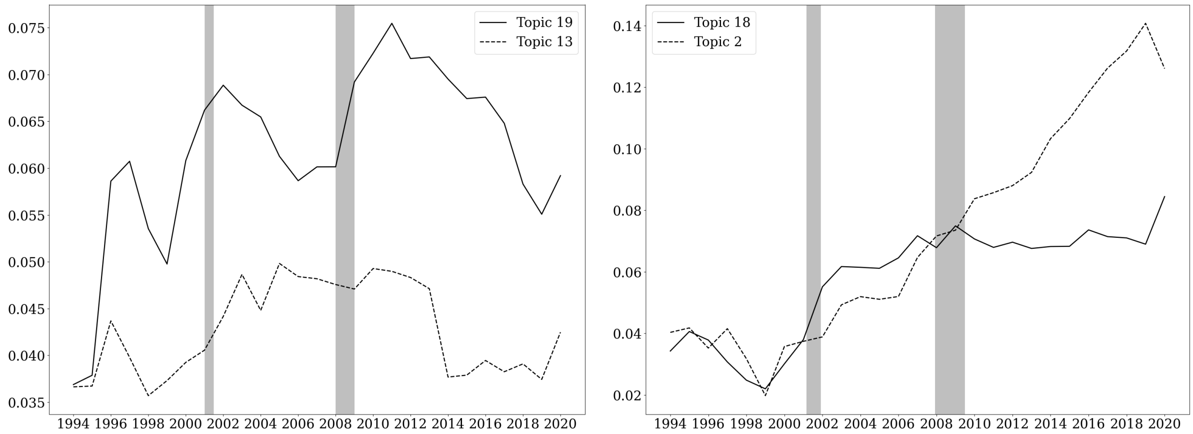
Note: This figure displays the first-order feature importance $FI^{First}(\xi_s)$. The larger the value, the greater the impact of feature ξ_s on the latent creditworthiness $y*$. A positive value indicates that an increase of ξ_s increases the creditworthiness of a company and vice versa. The sum of all absolute values of $FI^{First}(\xi_s)$ is equal to 1.

The single most important variable is Debt/EBITDA, which is commonly known as a good proxy for financial health. The negative sign means that an increase of this ratio (decrease of financial health) results in worse ratings. This variable is also considered by others as an important driver, see, for example, Alp (2013) or Baghai et al. (2014). The variable Size is the second most important variable with a positive sign, suggesting a higher creditworthiness of large firms. This finding is consistent with the results of Baghai et al. (2014). Interestingly, the third and fourth most important variables are Topic 19 (lenders, secured, guarantors) and Topic 13 (bondholder, insurer, warrants), each with a negative sign, which means that the higher the probability of these two topics, the lower the creditworthiness of the company. This recovers the descriptive analyses in Figures 3.4 & 3.5 as we observe these topics quite frequently in lower ratings and considerably less in higher ratings. The next important variable is Retained Earnings with a positive sign, which is in line with Alp (2013), Baghai et al. (2014) and Dorfleitner et al. (2020). This is followed by the CAPEX/Assets variable, which quantifies how much the company spends on acquiring, upgrading, and maintaining physical assets.

We find a positive sign, implying a higher creditworthiness for firms with larger CAPEX. Alp (2013) finds that CAPEX (significantly) decreases the creditworthiness of firms, whereas Baghai et al. (2014) finds that it (significantly) increases the creditworthiness. Dorfleitner et al. (2020) find a (non-significant) negative relation. Thus, the evidence in the literature is mixed. However, the mixed results may also point towards a non-linear relation, as the assumed linear relation changes sign. Dividend with a positive sign and Ivol with a negative sign follow, both in line with the literature. Subsequently, we find five topics as important drivers. In total, there are eight topics in the 15 most important variables, highlighting that the corporate creditworthiness is driven by qualitative data, such as MD&A sections.

Figure 3.7 shows the average annual probability of the four most important topics over time. The gray areas indicate crisis periods according to the NBER-dated recessions.

Figure 3.7: Topic probabilities over time

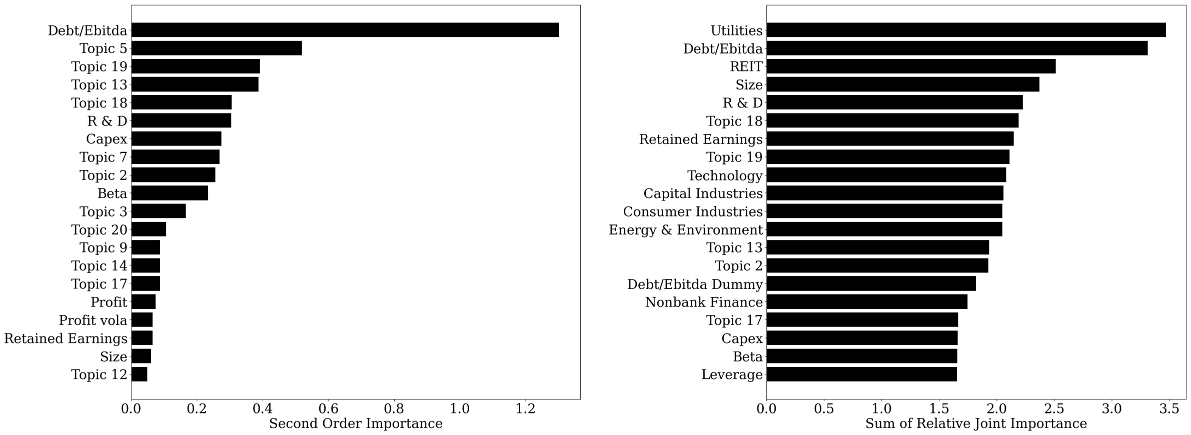


Note: This figure shows selected topic probabilities over time. To indicate recession periods, we follow the NBER crisis dummies. These time periods are illustrated by gray shaded areas.

Interestingly, the probability of Topic 19 (lenders, secured, guarantors) increases sharply before the bursting of the Dot-com bubble and during the Global Financial Crisis. This may indicate that firms more actively debate finance-related subjects around heightened times. This is also true for another crisis that is not captured by the NBER recession periods but is well known. We observe an increase in the probability for Topic 19 (lenders, secured, guarantors) also in the 1997 Asian financial crisis, where companies feared that global financial contagion could lead to a worldwide economic meltdown, see, e.g., Carson and Clark (2013) or Li et al. (2017). This peak can also be observed for Topic 13 (bondholder, insurer, warrants), which increases sharply after the two crisis periods of 1997 and 2001. In more recent times, both topics are at low levels, but they begin to rise in 2020, likely related to the ongoing COVID-19 pandemic. The probability of Topic 18 (uncollectible, portfolios, derecognized) increases from the early 2000s until the

Global Financial Crisis, and then stagnates until 2020, when it increases again. A more constant picture can be observed with Topic 2 (gaap, condensed, unaudited), which increases steadily from 1999 on. This may be due to the SEC's encouragement to discuss accounting policies in particular.

Figure 3.8: Higher-order importance



(a) Second order

(b) Interactions

Note: This figure shows the top 20 marginal non-linear and joint impacts on the firm's creditworthiness. Values of zero for both metrics would indicate marginal non-linear impacts or no joint impacts, respectively.

Baghai et al. (2014) include as a robustness of their results all variables with quadratic and cube terms to allow for non-linearities. They report only a small increase of performance. However, our results suggest a substantial impact of non-linearities and joint effects. Therefore, we analyze where this increase in performance comes from. Figure 3.8 reports in Panel (a) the top 20 marginal degrees of non-linearity ($FI^{Second}(\xi_s)$) and in Panel (b) the amount of interactions with all other variables ($FI^{Joint}(\Xi_{st})$). Please recall that a value of 0 implies linearity or no interactions, respectively. Among the 20 most non-linear variables, we find only eight firm characteristics, implying that there is only a small degree of non-linearity overall. However, we find that Debt/EBITDA shows a substantial degree and is, overall, the most non-linear variable. With a considerably less order of magnitude, we find R&D, CAPEX/Assets, Beta, Profit and its vola, and Size among the top 20. This recovers to some extent the findings of Baghai et al. (2014). In their robustness analysis, they do not include R&D, Beta, Size. Therefore, only one variable (Debt/EBITDA) shows a considerable degree of non-linearity, while the remaining variables appear to be largely linear even in our analysis. This finding also explains the small increase in performance that occurs when the text data is included in the linear Ordered Probit model. We observe that especially the topic probabilities entail an overall large amount of non-linearities, and thus the text data is highly non-linearly related to the creditworthiness of firms, which cannot be accounted for by the linear model.

Panel (a), however, does not answer the question of where the performance increase comes from. Some evidence may be found in Panel (b). We observe that especially the industry controls exhibit a high level of interactions with other variables, accounting for eight of the 20 most important interactions. Firm characteristics account for seven of the most important interactions and the text data account for five out of 20. All values for $FI^{Joint}(\pi_{st})$ are substantially different from zero, suggesting a large number of interactions entailed. Overall, the industry dummies account for roughly 28% of the interaction effects, the text for about 25%, and the control variables for 37%. This means that it is mainly joint effects that drive the performance improvement rather than marginal non-linearity. This also underlines the OANN approach, as these types of joint effects do not need to be specified in advance, but are intrinsically modeled. Setting up a linear Ordered Probit model and estimating and testing all possible interactions and their forms, e.g., quadratic, cubic, etc., would increase the computational cost. Table 3.C.1 in the Appendix 3.C shows the most important interactions. We find the largest single interaction between R&D and Topic 18 (uncollectible, portfolios, derecognized), which may show a relation between investments in research and development and words expressing realized losses, e.g., uncollectible and derecognized. Moreover, this topic is negatively related to creditworthiness, i.e., the higher the probability that a company talks about this topic, the lower its expected creditworthiness. Furthermore, among the top 10 interactions, we find five topics included. This further underlines the conclusion that textual data is related non-linearly and with joint effects to corporate creditworthiness.

For the remainder of this section, we focus on the impact of changes in topic probability. More specifically, we ask how a change in the mix of topic probabilities affects the (cumulative) probability of receiving an investment grade rating, i.e., we focus on $\mathbb{P}(R_i \geq Baa)$. For each rating class, we take the median across the entire training sample for each firm characteristic and topic probability. This serves as the baseline scenario. We then change some topic probabilities while holding the firm characteristics constant.

Table 3.6 shows the baseline scenario in the first row. For example, a firm with the median characteristics of a Baa-rated firm has a 90.8% probability of receiving an investment grade rating. Subsequently, we change two topic probabilities simultaneously, as we need to ensure that the probability over all topics is equal to 1. We choose the topic probability with the highest negative (Topic 19) and positive (Topic 3) impact on corporate creditworthiness, according to Figure 3.6. In the first two scenarios, we observe that the changes in topic probabilities mainly affect ratings at the change-point between investment grade and non-investment grade,

Table 3.6: Impact of probability changes

	Aaa	Aa	A	Baa	Ba	B	Caa-C
Baseline Scenario	99.5	98.7	98.1	90.8	41.7	11.7	3.4
Topic 3 - 5% — Topic 19 + 5%	99.3	98.2	97.4	88.4	36.6	10.1	3.1
Δ	−0.2	−0.5	−0.7	−2.4	−5.1	−1.6	−0.3
Topic 3 + 5% — Topic 19 - 5%	99.6	99.1	98.6	93.3	49.1	14.2	3.9
Δ	0.1	0.4	0.5	5.1	7.4	2.5	0.5
Topic 19 = 99%	89.8	78.1	75.9	55.4	15.0	5.1	0.2
Δ	−9.7	−20.6	−22.2	−35.4	−26.7	−6.6	−3.2
Topic 3 = 99%	99.5	99.2	98.8	95.4	83.0	53.3	17.8
Δ	0.0	0.5	0.7	4.6	41.3	41.6	14.4

Note: This table shows the cumulative probability of receiving an investment grade rating, i.e., we focus on $\mathbb{P}(R_i \geq Baa)$, given the median firm characteristics of every rating class. The baseline scenario applies the median topic probabilities for every rating class, whereas the remaining rows show selected changes in the topic distribution.

leaving very high and low ratings virtually unchanged. For example, if we change the topic probabilities of Topics 3 and 19 by only 5%, the probability of a Baa-rated company receiving an investment grade rating decreases (increases) by 2.4 (5.1) percentage points. For a firm rated Ba, we observe a decrease (increase) of 5.1 (7.4) percentage points. For companies with a very good or poor rating, the changes are less than 1 percentage point. The last two rows show a rather extreme scenario where 99% of the probability mass is shifted to one topic and the remaining mass is evenly distributed across all other topics. When we set the probability of Topic 19 to 99%, we observe serious changes, especially for investment grade ratings. The probability of (still) obtaining an investment grade rating for a Baa-rated firm falls by 35.4 percentage points to 55.4%. Therefore, the probability of receiving a *non-investment grade* rating becomes quite likely. When we set the probability of Topic 3 to 99%, we observe changes especially for non-investment grade ratings. For example, the probability of receiving an investment grade rating jumps to 83% for a Ba-rated company, making the probability of an upgrade very high.

To give a more intuitive example of how the MD&A sections are connected to the ratings, we redo our analysis of Table 3.6 with two self-written, imaginary texts. The first text reflects a promising perspective and expected future growth, so we may label this text as ”optimistic”. The second text writes about financing problems, increased expenses, and indebtedness. Therefore, we may label this one as ”pessimistic”. From an economic perspective, we may expect that the first text impacts the rating positively as the management discusses great achievements and prosper outlooks for the future. On the contrary, financial problems and an increase of debt are discussed in the second text. Therefore, the disclosed problems may reduce the (future) creditworthiness of this firm.[11]

Optimistic MD&A Section:

In the past fiscal year, we increased the selling of our products in any segment. Our business income in the retail segment increases our profit and strengthens the competitive advantage of our brand. We expect growth and profit to increase in the future compared to this year.

Pessimistic MD&A Section:

Unfortunately, the rate we have to pay on our revolving debt increased our expenses and our indebtedness. Furthermore, we need to issue revolving notes to fund our operations. Overall, we observe a sharp increase in our credit and borrowing expenses.

Table 3.7 shows the impact of the two stylized MD&A sections on the cumulative probability of obtaining an investment grade rating $\mathbb{P}(R_i \geq Baa)$. This analysis is similar to that in Table 3.6, but the topic distribution is now derived from the two self-written texts. Therefore, we can directly see and interpret the relation between different MD&A sections and the creditworthiness of the companies. Moreover, we add the average 2020 credit spreads for each category to highlight the impact that the optimistic and pessimistic texts may have on the median firm's cost of capital.

Table 3.7: Stylized example of MD&A sections

	Aaa	Aa	A	Baa	Ba	B	Caa-C
Optimistic	99.5	98.8	98.6	95.2	73.1	32.1	9.1
Pessimistic	93.1	84.7	83.5	62.7	18.4	5.9	2.4
Δ	−6.4	−14.1	−15.1	−32.5	−54.7	−26.2	−6.7
Credit spread (%)	0.79	0.92	1.18	2.01	3.87	5.82	12.61

Note: This table shows the cumulative probability of receiving an investment grade rating, i.e., we focus on $\mathbb{P}(R_i \geq Baa)$, given the median firm characteristics of every rating class. To evaluate the impact of different MD&A sections on $\mathbb{P}(R_i \geq Baa)$, we derive the topic probabilities of the optimistic and pessimistic statement and use them to predict the altered values of $\mathbb{P}(R_i \geq Baa)$. The mean credit spread per rating category is calculated as the average of the option-adjusted spread for 2020 of the ICE BofA U.S. Corporate Indices (AAA, AA, Single-A, BBB) or U.S. High Yield Indices (BB, Single-B, CCC & lower) retrieved from ICE Data Indices (2021).

The analysis in Table 3.7 clearly shows the magnitude of the impact of public information on corporate creditworthiness. Focusing on the optimistic MD&A section (first row), we observe comparable probabilities to those in Table 3.6 for investment grade ratings, but considerably

[11] We are aware that MD&A sections are usually quite longer and contain many more aspects of business strategy or financial planning. However, these stylized examples are intended to showcase how the OANN connects the MD&A section to the rating in a human intuitive way. Therefore, these two texts are not complete MD&A sections, but focus on key aspects of the MD&A sections.

higher probabilities for non-investment grade ratings. Therefore, especially low-rated firms show to be sensitive towards changes in their MD&A sections. For a Ba-rated firm, the probability of an investment grade rating increases to 73.1%, which implies that this positive MD&A section can shift the assessment of its creditworthiness to investment grade level. Table 3.7 also shows the credit spreads for each rating in the last row. Given the rating upgrade, this stylized company could cut its credit spread by almost half (i.e., from 387 basis points to 201 basis points). Hence, the changes in the text data can have direct implications for a company's cost of debt. Evaluating the pessimistic MD&A section, considerably larger changes in $\mathbb{P}(R_i \geq Baa)$ are observed. The biggest change in investment grade ratings can be observed for Baa-rated firms, down to 62.7%. For all non-investment grade ratings, it becomes very unlikely to obtain an investment grade rating if they report such a pessimistic MD&A section. The Tables 3.6 & 3.7 show that MD&A sections can have a significant impact on credit ratings, particularly on the difference between investment grade and non-investment grade ratings for companies near this change-point.

This link between public information and corporate creditworthiness may be especially important for stakeholders of the companies to determine the (expected) risk of their investment. Tables 3.6 & 3.7 show that changes in the text data can have severe impacts on the (future) creditworthiness of a company. As we lag our text data by two quarters, the stakeholder can anticipate the impact on the company's rating. This is particularly important because all quantitative drivers are backward-looking, i.e., based on historical observations, but the text data are forward-looking, i.e., contain the management's perspective, see, e.g., Li (2010) and Muslu et al. (2015). Therefore, our approach allows stakeholders to incorporate the forward-looking information of MD&A sections in their risk assessment.

Robustness

To show that these findings are robust, we perform a battery of checks. All results are available upon request. First, we use a different lag structure of the text data, e.g., a lag by three quarters and by four quarters. The results are similar. However, the overall performance decreases. Therefore, we argue that two lags are most suitable in our analysis. Second, we rerun our analysis with the second and third best architecture of the hyperparameter searches. Third, we estimate the linear Ordered Logit as an OANN without any hidden layer to eliminate the possibility that the performance difference is due to the different optimization procedure. Fourth, we also use different cut-points in the timeline to evaluate the performance of the OANN with text data in more detail. To do this, we estimate the STM for the data from 1994 to 2014, fit all models, and predict the next two years. Subsequently, we estimate an STM from 1994 to 2016, fit all models and predict the next two years. Overall, the performance metrics are stable for all sub-periods, and the main conclusions remain the same. Fifth, we initialize each OANN 100 times and report the average. Overall, the fluctuations are very small. This may indicate that the relation is quite stable and the signal-to-noise ratio is rather high with respect to ratings and their determinants. Sixth, one might argue that we could use the standard Softmax output in our neural network, which is a non-linear extension of the well-known Multinomial Logit model. However, this would neglect the ordinal nature of ratings. We obtain worse results using a Softmax output than with the OANN structure. Moreover, we use a standard ANN with only one output node and a mean-squared-error loss, which can be considered as a non-linear version of the standard linear regression model, commonly referred to as OLS. The ratings are treated as metric variables, similar to Baghai et al. (2014), and the final rating prediction is integer rounded. The results are similar to the Softmax output and worse than the OANN results. Therefore, the OANN provides the better fit in terms of performance metrics while allowing parsimonious interpretation of important (non-linear) drivers. Seventh, we include information from our text data via the commonly used dictionary of Loughran and Mcdonald (2011). The performance differences of the models with and without texts are very small. This may suggest that a one-dimensional incorporation of the text data with a sentiment index cannot account for the non-linear relation with the latent creditworthiness of firms. This is similar to the findings of Frankel et al. (2022) and Donovan et al. (2021), who find that complex models for text extraction outperform dictionary approaches, especially in predicting stock and CDS returns, respectively. Eighth, we include the approach of Frankel et al. (2022) and Donovan et al. (2021) who derive a text-based index using machine learning approaches. The performance metrics are better than the one obtained using the Loughran and Mcdonald (2011), but worse than with our approach. Ninth, we include the GDP growth rate, the S&P 500 growth rate, and the News Based Economic Political Uncertainty (EPU) index to rule out the possibility that the text data account for macroeconomic variation or uncertainty related measures. Interestingly, the inclusion of the GDP growth rate worsens the testing performance. This may be due to the shacked-up macro economy during the COVID-19 crisis. Apart from all robustness checks, the appropriately processed text data provide additional information beyond the common quantitative variables and measures of the economic surrounding and uncertainty.

In summary, this section investigates the impact of text data on corporate creditworthiness in great detail. We find that text data have a comparatively high importance, which implies that they provide information beyond common quantitative drivers. Subsequently, the paper focuses on the shape of relation of these drivers to the creditworthiness of firms. Text data in particular shows a high degree of marginal non-linearity, whereas quantitative variables are predominantly linear. Moreover, the analysis reveals a very large number of interactions between variables, and especially with respect to text data. This explains to a large extent the increase in performance when using text data in a non-linear way. Finally, a scenario analysis indicates that firms at the change-point between investment grade and non-investment grade ratings react sensitive to changes in the mix of topic probabilities. We see that text data play a crucial role for assessing a firm's creditworthiness. This can be attributed to the fact that rating agencies focus especially on this type of information and companies use the MD&A section to provide additional information about their creditworthiness.

3.6 Conclusion

This study examines the influence of the forward-looking Management's Discussion & Analysis (MD&A) section of quarterly and annual corporate filings on determining the creditworthiness of a company, commonly expressed by credit ratings of major rating agencies. As ratings are highly relevant for all groups of participants in the financial markets (see, e.g., Bonsall and Miller (2017) or Becker and Milbourn (2011)), our findings provide valuable insights into components of corporate creditworthiness. Topics from the MD&A sections are extracted using a Structural Topic Model (STM) following Roberts et al. (2016) to convert the text data into topic probabilities, i.e., to quantify how likely it is that MD&A section contains a particular topic such as financial risk, digitalization, or sustainability. In combination with commonly used firm characteristics such as leverage, profit, or size, the extracted topics are fed into an Ordinal Artificial Neural Network (OANN) (see Cao et al. (2020)). This approach allows modeling the rating as ordinal dependent variable, taking into account any kind of non-linearity and interactions in the relation between the input variables and the latent creditworthiness of the company.

Public information in terms of MD&A topics contribute significantly to explaining and predicting credit ratings. Using Explainable Artificial Intelligence (XAI) techniques, we uncover several topics such as finance or regulation among the most important variables in terms of marginal influence (first-order importance). In addition, our results show higher-order effects of firm characteristics such as Debt/EBITDA and, in particular, interaction effects with topics. This underlines that textual data is related non-linearly and with joint effects to corporate creditworthiness. Applying machine learning methods to the analysis of non-linear effects is computationally effective and resource-efficient because interactions can be modeled and evaluated implicitly. Changes in MD&A content appear to affect investment grade and non-investment grade companies differently. A stress test of the topic distribution across texts shows that firms at the change-point between investment grade and non-investment grade ratings react sensitively to changes in their MD&A. Therefore, our findings provide valuable insights into components of corporate creditworthiness that are relevant to all stakeholders who rely upon credit ratings, such as governments, financial institutions, or stakeholders in general. Using our approach, they are able to derive the impact of changes in the forward-looking MD&A section on future ratings. In addition, our findings can be confirmed by a series of robustness checks.

3.A APPENDIX | Hyperparameter Selection

The search of the hyperparameter is inspired by Gu et al. (2020). We use a random search algorithm to draw 500 combinations of hyperparameter sets. Furthermore, we use advanced activation functions, which turned out have favorable properties.

Advanced activation functions

Activation functions are an important part when training deep neural networks. We all know that standard activation functions such as sigmoid and tanh frequently suffer from the so-called vanishing gradient problem, i.e., earlier layers learn much slower than later layers, due to their small gradients. The update of weights relies on gradient information processed by the chain rule. This has the implication of multiplying many small values to compute gradients for early layers in a deep neural network. It can be shown that the gradient (error signal) decreases exponentially with the number of layers and, thus, early layers learn much slower or even stop learning in deep neural networks (Hochreiter (1991)). A common solution is the Rectified Linear Unit (ReLU) activation function, which reduces the vanishing gradient problem considerably (Hochreiter (1998)). However, it suffers from the dying ReLU problem, i.e., the activation becomes inactive and only output 0 for any input, see Lu et al. (2020) for an overview. Therefore, several alternative activation functions are proposed to avoid the vanishing gradient problem and avoid the dying ReLU problem as well. Table 3.A.1 illustrates the advanced activation functions used in this study.

Table 3.A.1: Advanced activation functions

Activation	Formula	Original Paper
Softplus	$\log(\exp(x)+1)$	Dugas et al. (2001)
Swish	$x \cdot \text{sigmoid}(\beta \cdot x)$	Ramachandran et al. (2017)
GeLU	$x \cdot \Phi(x)$	Hendrycks and Gimpel (2020)
Mish	$x \cdot \tanh(\text{softplus}(x))$	Misra (2020)

Note: This table shows the advanced activation functions used in this study. The first column describes the name, the second column contains the mathematical expression, and the last column shows the original paper where the activation function was introduced. This paper uses extensions to common activations such as tanh and sigmoid to avoid the vanishing gradient problem. Furthermore, all four activation functions avoid the dying ReLU problem and provide higher performance than the original ReLU activation function. We opt for $\beta = 1$ using the standard Swish activation formulation.

The hyperparameter search

Following Gu et al. (2020), we assume that the number of neurons halves over the hidden layers, i.e., 32 neurons in the first hidden layer, 16 in the second hidden layer, and so on. Hence, instead

Table 3.A.2: Setup of the hyperparameter search

Parameter	Distribution
Learning Rate	$U^c \sim [0.000001, 0.001]$
Lambda L1	$U^c \sim [0.00001, 0.05]$
Dropout	$U^c \sim [0.20, 0.50]$
Hidden Layer	$U^d \sim [1, 4]$
Multiple	$U^d \sim [1, 8]$
Activation Function	Softplus, Swish, Mish, GeLU

Note: The table shows different values for the hyperparameter search. U^c denotes the continuous uniform distribution, whereas U^d denotes the discrete uniform distribution. Avoiding overfitting is of key importance, and so we place emphasis on regularization parameters (L1) and different designs of dropout layers. . The network architecture employs a baseline structure of halving the number of neurons over the hidden layers, following Gu et al. (2020). The minimum number of neurons in the first hidden layer is 32.

of validating the actual number of neurons, we validate a multiple of a baseline structure. We assume 32 neurons as minimum for the first hidden layer. Hence, for a multiple of 1 and four hidden layers, we have 32-16-8-4 neurons. Using a multiple of 2 we get 64-32-16-4. If only two hidden layers and a multiple of 2 is selected, we have 64-32 as number of neurons. This approach gives us a great flexibility, but ensures an efficient way to validate the shallowness of the neural network. To avoid overfitting, we also use an *Early Stopping*, which stops the training if the validation loss increases a selected number of iterations (so-called *patience*). In this paper, we use a *patience* of 75, a maximum number of 5,000 epochs and a batch size of 64.

For the training sample of 1994 to 2018 we obtain the following hyperparameters for *OANN* and *OANN + Text*, see Table 3.A.3:

Table 3.A.3: Final values of the hyperparameter search

Parameter	OANN	OANN + Text
Learning Rate	0.00048	0.00037
Lambda L1	0.03208	0.02747
Dropout	0.47468	0.38128
Hidden Layer	2	3
Multiple	2	1
Activation Function	Mish	Swish

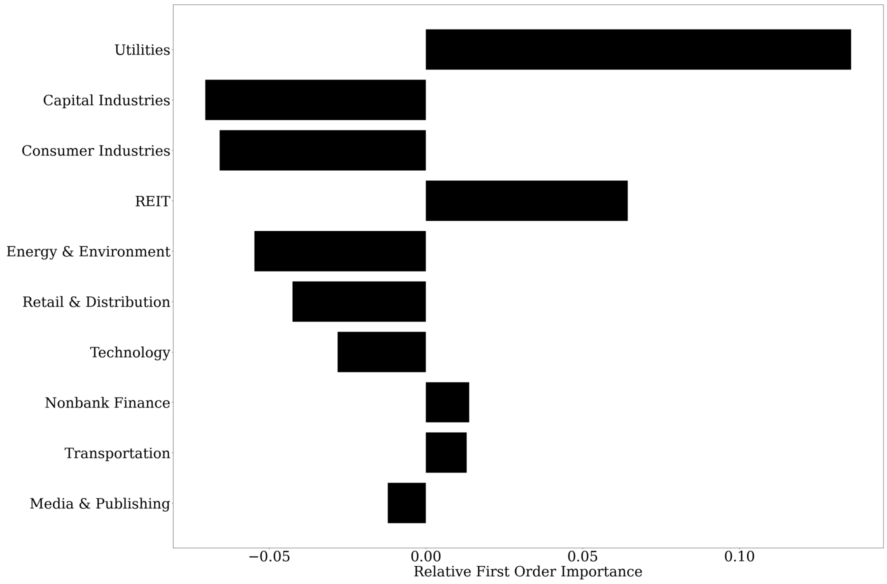
Note: The table shows final values for the hyperparameter search. As avoiding overfitting is of major concern, we put much emphasis on regularization parameters (L1) and different designs of Dropout Layers. The network architecture employs a baseline structure of halving the number of neurons over the hidden layers, following Gu et al. (2020). The minimum number of neurons in the first hidden layer is 32.

The interpretation of these hyperparameters is not straightforward, but we can see that the *OANN + Text* prefers a deeper neural network compared to the *OANN*. With respect to the activation functions, we observe that the *OANN* prefers the Gaussian approximation of the ReLU, whereas *OANN + Text* prefers the logistic approximation.

3.B APPENDIX | Industry Classifier

Figure 3.B.1: First-order feature importance

(**a**) Industry dummies



Note: This table shows the first order feature importance for each industry classifier. The larger the value, the greater the impact of feature ξ_s on the latent creditworthiness $y*$. A positive value indicates that an increase of ξ_s increases the creditworthiness of a company and vice versa. The sum of all absolute values of $FI^{First}(\xi_s)$ is equal to 1.

Figure 3.B.1 shows the estimated values of the first order feature importance for the industry classifiers. The reference category is *Banking*. From an economic perspective, we may conclude the estimated signs are reasonable. For example, we observe that *Utilities* has a higher creditworthiness. This is plausible since utility companies are usually backed by governments or states and, thus, have a higher creditworthiness.

3.C APPENDIX | Most Important Interactions

Table 3.C.1: Top 10 interactions

Variable 1	Variable 2	Joint Importance
Topic 18 (uncollectible, portfolios, derecognized)	RandD_Assets	0.291
Debt_Ebitda	Debt_Ebitda_Dummy	0.288
Debt_Ebitda	Ebitda_Sales_vola	0.269
Topic 2 (gaap, condensed, unaudited)	Technology	0.219
Topic 19 (lenders, secured, guarantors)	Utilities	0.211
Topic 3 (mainframe, chips, computer)	Utilities	0.208
BookValue	Utilities	0.206
Debt_Ebitda	Consumer Industries	0.203
Topic 13 (bondholder, insurer, warrants)	Capital Industries	0.196
Debt_Ebitda	Utilities	0.192

Note: This table shows the 10 most important interactions in our analysis. If the estimated values are greater than 1, we can detect a joint impact.

Table 3.C.1 shows the most important interactions in our empirical analysis. Interestingly, the most important one includes a topic derived from the MD&A sections. We observe a joint impact of Topic 18 (uncollectible, portfolios, derecognized) with the RandD_Assets measure. This implies that the impact of investments in Research and Development on the creditworthiness of a company is connected to the words of Topic 18 (uncollectible, portfolios, derecognized). Therefore, not only the amount of investments may be important, but also the way the company describes these investments in the MD&A section. Overall, we find in 5 out of 10 important interactions qualitative data to be included. This underlines the importance of qualitative data for determining a firm's creditworthiness.

Conclusion

Summary

This thesis presents applications of machine learning (ML) methods in credit risk management and extends the literature on understanding ML-based model results through use cases of explainable artificial intelligence (XAI) techniques in default risk and credit ratings modeling. Thus, the methodological toolbox available to credit risk managers is complemented by instructive examples of how to use innovative methods and approaches to achieve better interpretability when operating ML models. This is among the key requirements articulated by regulatory authorities, financial institutions, risk specialists, and researchers alike in light of the growing interest and desire to deploy ML in business (see Introduction). The independent research papers introduced in this thesis highlight three use cases of ML in the area of credit risk management, with particular emphasis on quantification of model input feature importance and explainability of model outcomes.

The first research paper *Revisiting the Relation between Corporate Default and Financial Frictions with Machine Learning* (see Chapter 1) shows that financial frictions have non-linear impacts on corporate defaults as they become more binding, and that this non-linearity decreases as information fades. In the second research paper *Default Risk and Public Information: A Machine Learning Approach* (see Chapter 2), text feature extraction from Management's Discussion & Analysis (MD&A) sections of corporate filings is presented and higher-order effects of public information in modeling default risk are uncovered by using XAI techniques. The third research paper *The Impact of Qualitative Information on Corporate Creditworthiness* (see Chapter 3) concludes that public information in the form of topics from MD&A sections contributes significantly to explaining and predicting credit ratings, i.e., the creditworthiness of companies.

Discussion and Outlook

This thesis addresses highly topical issues regarding the use, explainability, and interpretability of machine learning approaches for modeling parameters of credit risk, in particular default probabilities and credit ratings. As a result, the deployment of machine learning in credit risk management may be further promoted by providing those responsible for assessing its potentials and deciding on implementation with valid arguments, in particular increased transparency and justifiability of model outcomes. The solution approaches discussed are transferable to a wide range of ML applications and scale up the methods available to, for example, financial analysts or risk specialists, thus contributing to expand the toolkit with state-of-the-art approaches.

The assessment of the importance of quantitative and qualitative information in predicting default probabilities using ML methods, which is part of this thesis, uncovers key drivers of default risk and may be of great interest to credit risk professionals. In addition, the study of credit ratings though machine learning in this thesis provides valuable insights into components of corporate creditworthiness that are relevant to all stakeholders who rely on ratings, such as financial institutions, regulators, governments, employees, and the company itself.

Public information in the form of MD&A texts is a valuable source of data that, once sufficient effort has been expended on its collection and preparation, can provide additional power for models to explain or predict relations associated with the reporting companies. The prevalent thesis has demonstrated this for the cases of incorporation in default risk and credit rating models, but the inclusion of texts may also be useful in other contexts, such as in the determination of loss given default. Moreover, the methodological guidance of this thesis for transforming raw text into modelable text features, including how to use dimensionality reduction techniques or extract topics, enables natural language processing of a wide spectrum of text types, including newspaper articles, headlines, social media messages, or any kind of disclosure.

In summary, this knowledge may enable financial institutions to better allocate their resources in terms of cost-benefit trade-offs when deploying ML methods. It may support businesses mitigating risks arising from the increased use of ML, help banks meeting regulators' requirements for the deployment of AI, and guide supervisors monitoring ML models, thereby solidifying the stability of the financial system in an era where machine learning is further advancing.

Bibliography

Acharya, V., S. A. Davydenko, and I. A. Strebulaev (2012). Cash Holdings and Credit Risk. *The Review of Financial Studies 25*(12), 3572–3609.

Acharya, V. V., S. T. Bharath, and A. Srinivasan (2007). Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. *Journal of Financial Economics 85*(3), 787–821.

Acheampong, F. A., H. Nunoo-Mensah, and W. Chen (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review 54*, 5789–5829.

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

Ahmadi, Z., P. Martens, C. Koch, T. Gottron, and S. Kramer (2018). Towards Bankruptcy Prediction: Deep Sentiment Mining to Detect Financial Distress from Business Management Reports. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 293–302.

Alhadab, M. and T. Nguyen (2018). Corporate diversification and accrual and real earnings management: A non-linear relationship. *Review of Accounting and Finance 17*(2), 198–214.

Alp, A. (2013). Structural Shifts in Credit Rating Standards. *The Journal of Finance 68*(6), 2435–2470.

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(4), 1059–1086.

Aretz, K., C. Florackis, and A. Kostakis (2018). Do Stock Returns Really Decrease with Default Risk? New International Evidence. *Management Science 64*(8), 3821–3842.

Baghai, R. P., H. Servaes, and A. Tamayo (2014). Have Rating Agencies Become More Conservative? Implications for Capital Structure and Debt Pricing. *The Journal of Finance 69*(5), 1961–2005.

Bakoben, M., T. Bellotti, and N. Adams (2020). Identification of credit risk based on cluster analysis of account behaviours. *Journal of the Operational Research Society 71*(5), 775–783.

Bank of England (2020). The impact of Covid on machine learning and data science in UK banking. *Bank of England*, 1–15. Available at: https://www.bankofengland.co.uk/quarterly-bulletin/2020/2020-q4/the-impact-of-covid-on-machine-learning-and-data-science-in-uk-banking.

Bao, Y. and A. Datta (2014). Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science 60*(6), 1371–1391.

Bastos, J. A. and S. M. Matos (2022). Explainable models of credit losses. *European Journal of Operational Research (forthcoming)*. Available at: https://doi.org/10.1016/j.ejor.2021.11.009.

Basu, R. and J. P. Naughton (2020). The Real Effects of Financial Statement Recognition: Evidence from Corporate Credit Ratings. *Management Science 66*(4), 1672–1691.

Beaver, W. H., M. F. Mcnichols, and J.-W. Rhie (2005). Have Financial Statements Become Less Informative? Evidence from the Ability of Financial Ratios to Predict Bankruptcy. *Review of Accounting Studies 10*, 93–122.

Becker, B. and T. Milbourn (2011). How did increased competition affect credit ratings? *Journal of Financial Economics 101*(3), 493–514.

Behr, P., D. J. Kisgen, and J. P. Taillard (2018). Did Government Regulations Lead to Inflated Credit Ratings? *Management Science 64*(3), 1034–1054.

Bellotti, A., D. Brigo, P. Gambetti, and F. Vrins (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting 37*(1), 428–444.

Berger, J., A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel (2020). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing 84*(1), 1–25.

Berwart, E., M. Guidolin, and A. Milidonis (2019). An empirical analysis of changes in the relative timeliness of issuer-paid vs. investor-paid ratings. *Journal of Corporate Finance 59*, 88–118.

Bharath, S. T. and T. Shumway (2008). Forecasting Default with the Merton Distance to Default Model. *The Review of Financial Studies 21*(3), 1339–1369.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond Risk Premiums with Machine Learning. *The Review of Financial Studies 34*(2), 1046–1089.

Bianchi, D., M. Büchner, T. Hoogteijling, and A. Tamoni (2021). Corrigendum: Bond Risk Premiums with Machine Learning. *The Review of Financial Studies 34*(2), 1090–1103.

Blei, D. M. and J. D. Lafferty (2007). A correlated topic model of Science. *The Annals of Applied Statistics 1*(1), 17–35.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Blöchlinger, A. and M. Leippold (2011). A New Goodness-of-Fit Test for Event Forecasting and Its Application to Credit Defaults. *Management Science 57*(3), 487–505.

Blochwitz, S., A. Hamerle, S. Hohl, R. Rauhmeier, and D. Rösch (2005). Myth and reality of discriminatory power for rating systems. *Wilmott Magazine 2005*(1), 2–6.

Blume, M. E., F. Lim, and A. C. Mackinlay (1998). The Declining Credit Quality of U.S. Corporate Debt: Myth or Reality? *The Journal of Finance 53*(4), 1389–1413.

Bonsall, S. B., J. R. Green, and K. A. Muller (2018). Are Credit Ratings More Rigorous for Widely Covered Firms? *The Accounting Review 93*(6), 61–94.

Bonsall, S. B., E. R. Holzman, and B. P. Miller (2017). Managerial Ability and Credit Risk Assessment. *Management Science 63*(5), 1425–1449.

Bonsall, S. B., A. J. Leone, B. P. Miller, and K. Rennekamp (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics 63*(2-3), 329–357.

Bonsall, S. B. and B. P. Miller (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies 22*, 608–643.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*, 1–3.

Butaru, F., Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance 72*, 218–239.

Cafarelli, A. (2020). Creditworthiness risk over years: The evolution of credit rating standards. *Journal of Corporate Accounting & Finance 31*(4), 48–59.

Calabrese, R. and J. Crook (2020). Spatial contagion in mortgage defaults: A spatial dynamic survival model with time and space varying coefficients. *European Journal of Operational Research 287*(2), 749–761.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008). In Search of Distress Risk. *The Journal of Finance 63*(6), 2899–2939.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2011). Predicting Financial Distress and the Performance of Distressed Stocks. *Journal of Investment Management 9*(2), 14–34.

Cao, W., V. Mirjalili, and S. Raschka (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters 140*, 325–331.

Carson, M. and J. Clark (2013). Asian Financial Crisis. *Federal Reserve History*, 1–3.

Chava, S. and R. A. Jarrow (2004). Bankruptcy Prediction with Industry Effects. *Review of Finance 8*(4), 537–569.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794, San Francisco, California, USA. Association for Computing Machinery.

Cohen, L., C. Malloy, and Q. Nguyen (2020). Lazy Prices. *The Journal of Finance 75*(3), 1371–1415.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics 44*(3), 837–845.

Demidenko, E. (2016). The p-Value You Can't Buy. *The American Statistician 70*(1), 33–38.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*, 4171–4186.

Dichev, I. D. and J. D. Piotroski (2001). The Long-Run Stock Returns Following Bond Ratings Changes. *The Journal of Finance 56*(1), 173–203.

Dilly, M. and T. Mählmann (2016). Is There a "Boom Bias" in Agency Ratings? *Review of Finance 20*(3), 979–1011.

Ding, A. A., S. Tian, Y. Yu, and H. Guo (2012). A Class of Discrete Transformation Survival Models With Application to Default Probability Prediction. *Journal of the American Statistical Association 107*(499), 990–1003.

Djeundje, V. B. and J. Crook (2019). Identifying hidden patterns in credit risk survival data using Generalised Additive Models. *European Journal of Operational Research 277*(1), 366–376.

Donovan, J., J. Jennings, K. Koharki, and J. Lee (2021). Measuring credit risk using qualitative disclosure. *Review of Accounting Studies 26*, 815–863.

Dorfleitner, G., J. Grebler, and S. Utz (2020). The Impact of Corporate Social and Environment Performance on Credit Rating Prediction: North America versus Europe. *Journal of Risk 22*(6), 1–33.

Doumpos, M., K. Andriosopoulos, E. Galariotis, G. Makridou, and C. Zopounidis (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research 262*(1), 347–360.

du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research 254*(1), 236–252.

du Jardin, P. (2021). Forecasting corporate failure using ensemble of self-organizing neural networks. *European Journal of Operational Research 288*(3), 869–885.

Dugas, C., Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia (2001). Incorporating Second-Order Functional Knowledge for Better Option Pricing. In *Advances in Neural Information Processing Systems*, 472–478.

Dumitrescu, E., S. Hué, C. Hurlin, and S. Tokpavi (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research 297*(3), 1178–1192.

Durnev, A. and C. Mangen (2020). The spillover effects of MD&A disclosures for real investment: The role of industry competition. *Journal of Accounting and Economics 70*(1), 101299.

Dyer, T., M. Lang, and L. Stice-Lawrence (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics 64*(2-3), 221–245.

Engelmann, B., E. Hayden, and D. Tasche (2003). Measuring the Discriminative Power of Rating Systems. *Deutsche Bundesbank Series 2: Banking and Financial Supervision*(1), 1–24.

Erel, I., L. H. Stern, C. Tan, and M. S. Weisbach (2021). Selecting Directors Using Machine Learning. *The Review of Financial Studies 34*(7), 3226–3264.

European Banking Authority (2021a). EBA Discussion Paper on Machine Learning for IRB Models. *European Banking Authority Paris, France,* 1–29. Available at: https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models.

European Banking Authority (2021b). Risk Dashboard: Data as of Q3 2021. *European Banking Authority Paris, France,* 1–49. Available at: https://www.eba.europa.eu/risk-analysis-and-data/risk-dashboard.

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

Fraisse, H. and M. Laporte (2022). Return on Investment on Artificial Intelligence: the Case of Bank Capital Requirement. *Journal of Banking & Finance (forthcoming).* Available at: https://doi.org/10.1016/j.jbankfin.2022.106401.

Frankel, R., J. Jennings, and J. Lee (2022). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science (forthcoming).* Available at: http://pubsonline.informs.org/doi/10.1287/mnsc.2021.4156.

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

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Friberg, R. and T. Seiler (2017). Risk and ambiguity in 10-Ks: An examination of cash holding and derivatives use. *Journal of Corporate Finance 45*, 608–631.

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

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance 77*(1), 5–47.

Geng, R., I. Bose, and X. Chen (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research 241*(1), 236–247.

Giordani, P., T. Jacobson, E. von Schedvin, and M. Villani (2014). Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios. *Journal of Financial and Quantitative Analysis 49*(4), 1071–1099.

Goldstein, I., C. S. Spatt, and M. Ye (2021). Big Data in Finance. *The Review of Financial Studies 34*(7), 3213–3225.

Griffin, J. M. and M. L. Lemmon (2002). Book-to-Market Equity, Distress Risk, and Stock Returns. *The Journal of Finance 57*(5), 2317–2336.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gunnarsson, B. R., S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research 295*(1), 292–305.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Hendrycks, D. and K. Gimpel (2020). Gaussian Error Linear Units (GELUs). *arXiv:1606.08415*.

Hillegeist, S. A., E. K. Keating, D. P. Cram, and K. G. Lundstedt (2004). Assessing the Probability of Bankruptcy. *Review of Accounting Studies 9*, 5–34.

Hilscher, J. and M. Wilson (2017). Credit Ratings and Credit Risk: Is One Measure Enough? *Management Science 63*(10), 3414–3437.

Hinton, G. and S. Roweis (2003). Stochastic Neighbor Embedding. *Advances in neural information processing systems 15*, 857–864.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München 91*(1).

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6*(2), 107–116.

Horel, E. and K. Giesecke (2020). Significance Tests for Neural Networks. *Journal of Machine Learning Research 21*(227), 1–29.

Horrigan, J. O. (1966). The Determination of Long-Term Credit Standing with Financial Ratios. *Journal of Accounting Research 4*, 44–62.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417–441.

Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science 64*(6), 2833–2855.

Hung, M., P. Kraft, S. Wang, and G. Yu (2022). Market Power and Credit Rating Standards: Global Evidence. *Journal of Accounting and Economics (forthcoming)*. Available at: https://doi.org/10.1016/j.jacceco.2021.101474.

Hwang, D. and Y. Kim (2020). Revisiting the time-varying credit rating policy: a new test of procyclicality. *Applied Economics Letters 27*(10), 810–815.

ICE Data Indices (2021). LLC ICE BofA US Indices Option-Adjusted Spread, Corporate Index (AAA, AA, Single-A, BBB) and High Yield Index (BB, Single-B, CCC & Lower) [BAMLC0A1CAAA - BAMLH0A3HYC]. *Retrieved from FRED, Federal Reserve Bank of St. Louis*.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Kang, J. K., C. D. Williams, and R. Wittenberg-Moerman (2021). CDS trading and nonrelationship lending dynamics. *Review of Accounting Studies 26*(1), 258–292.

Kang, S. (2020). Model validation failure in class imbalance problems. *Expert Systems with Applications 146*, 113190.

Kaplan, R. S. and G. Urwitz (1979). Statistical Models of Bond Ratings: A Methodological Inquiry. *The Journal of Business 52*(2), 231–261.

Kedia, S., S. Rajgopal, and X. Zhou (2014). Did going public impair Moody's credit ratings? *Journal of Financial Economics 114*(2), 293–315.

Kedia, S., S. Rajgopal, and X. A. Zhou (2017). Large shareholders and credit ratings. *Journal of Financial Economics 124*(3), 632–653.

Kellner, R., M. Nagl, and D. Rösch (2022). Opening the black box – Quantile neural networks for loss given default prediction. *Journal of Banking & Finance 134*, 106334.

Kim, J. B., P. Shroff, D. Vyas, and R. Wittenberg-Moerman (2018). Credit Default Swaps and Managers' Voluntary Disclosure. *Journal of Accounting Research 56*(3), 953–988.

Kozodoi, N., J. Jacob, and S. Lessmann (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research 297*(3), 1083–1094.

Kraft, P. (2015). Do rating agencies cater? Evidence from rating-based contracts. *Journal of Accounting and Economics 59*(2-3), 264–283.

Kriebel, J. and L. Stitz (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research (forthcoming)*. Available at: https://doi.org/10.1016/j.ejor.2021.12.024.

Krüger, S., D. Rösch, and H. Scheule (2018). The impact of loan loss provisioning on bank capital requirements. *Journal of Financial Stability 36*, 114–129.

Kruppa, J., A. Schwarz, G. Arminger, and A. Ziegler (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications 40*(13), 5125–5131.

Lang, M. and L. Stice-Lawrence (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics 60*(2-3), 110–135.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics 56*(1), 130–147.

Lehavy, R., F. Li, and K. Merkley (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review 86*(3), 1087–1115.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics 45*(2-3), 221–247.

Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings–A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research 48*(5), 1049–1102.

Li, K., J. Lockwood, and H. Miao (2017). Risk-shifting, equity risk, and the distress puzzle. *Journal of Corporate Finance 44*, 275–288.

Li, K., F. Mai, R. Shen, and X. Yan (2021). Measuring Corporate Culture Using Machine Learning. *The Review of Financial Studies 34*(7), 3265–3315.

Li, Y. and W. Chen (2021). Entropy method of constructing a combined model for improving loan default prediction: A case study in China. *Journal of the Operational Research Society 72*(5), 1099–1109.

Liang, D., C.-C. Lu, C.-F. Tsai, and G.-A. Shih (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research 252*(2), 561–572.

Loughran, T. and B. Mcdonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research 54*(4), 1187–1230.

Lu, L., Y. Shin, Y. Su, and G. E. Karniadakis (2020). Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics 28*(5), 1671–1706.

Luo, J., X. Yan, and Y. Tian (2020). Unsupervised quadratic surface support vector machine with application to credit risk assessment. *European Journal of Operational Research 280*(3), 1008–1017.

Mai, F., S. Tian, C. Lee, and L. Ma (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research 274*(2), 743–758.

Martin, X. and S. Roychowdhury (2015). Do financial market developments influence accounting practices? Credit default swaps and borrowers' reporting conservatism. *Journal of Accounting and Economics 59*(1), 80–104.

McKinsey (2020). Global survey: The state of AI in 2020. *McKinsey Digital*, 1–13. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020.

McKinsey (2021). Global survey: The state of AI in 2021. *McKinsey Digital*, 1–11. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021.

Miller, B. P. (2010). The Effects of Reporting Complexity on Small and Large Investor Trading. *The Accounting Review 85*(6), 2107–2143.

Miller, G. S. (2017). Discussion of "the evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation". *Journal of Accounting and Economics 64*(2-3), 246–252.

Misra, D. (2020). Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv:1908.08681*.

Moody's (2021a). Default & Recovery Database (DRD). *Moody's Analytics, Inc.*. Available at: https://www.moodysanalytics.com/product-list/default-and-recovery-database.

Moody's (2021b). Procedures and Methodologies Used to Determine Credit Ratings | Exhibit 2. *Technical Report*. Available at: https://www.moodys.com/researchandratings/rating-methodologies.

Mullainathan, S. and J. Spiess (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives 31*(2), 87–106.

Muslu, V., S. Radhakrishnan, K. R. Subramanyam, and D. Lim (2015). Forward-Looking MD&A Disclosures and the Information Environment. *Management Science 61*(5), 931–948.

Nazemi, A., F. Baumann, and F. J. Fabozzi (2022). Intertemporal defaulted bond recoveries prediction via machine learning. *European Journal of Operational Research 297*(3), 1162–1177.

Netzer, O., A. Lemaire, and M. Herzenstein (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *Journal of Marketing Research 56*(6), 960–980.

Nguyen, B.-H. and V.-N. Huynh (2022). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society 73*(1), 102–121.

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

Petropoulos, A., V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting 36*(3), 1092–1113.

Pogue, T. F. and R. M. Soldofsky (1969). What's in a Bond Rating. *Journal of Financial and Quantitative Analysis 4*(2), 201–228.

Ramachandran, P., B. Zoph, and Q. V. Le (2017). Searching for Activation Functions. *arXiv:1710.05941*.

Redelmeier, D. A., D. A. Bloch, and D. H. Hickam (1991). Assessing predictive accuracy: How to compare Brier scores. *Journal of Clinical Epidemiology 44*(11), 1141–1146.

Refinitiv (2021). Refinitiv Eikon. *Refinitiv*. Available at: https://eikon.thomsonreuters.com/index.html.

Reisz, A. S. and C. Perlich (2007). A market-based framework for bankruptcy prediction. *Journal of Financial Stability 3*(2), 85–131.

Roberts, M. E., B. M. Stewart, and E. M. Airoldi (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association 111*(515), 988–1003.

Roberts, M. E., B. M. Stewart, and D. Tingley (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software 91*(2), 1–40.

Rösch, D. and H. Scheule (2020). *Deep Credit Risk - Machine Learning with Python*. Amazon.

Sadhwani, A., K. Giesecke, and J. Sirignano (2021). Deep Learning for Mortgage Risk. *Journal of Financial Econometrics 19*(2), 313–368.

Sariev, E. and G. Germano (2020). Bayesian regularized artificial neural networks for the estimation of the probability of default. *Quantitative Finance 20*(2), 311–328.

SEC (2003). Report on the Role and Function of Credit Rating Agencies in the Operation of the Securities Markets. *U.S. Securities and Exchange Commission*, 1–45.

SEC (2021). Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database. *U.S. Securities and Exchange Commission*. Available at: https://www.sec.gov/edgar/search-and-access.

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

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

Sopitpongstorn, N., P. Silvapulle, J. Gao, and J.-P. Fenech (2021). Local logit regression for loan recovery rate. *Journal of Banking & Finance 126*, 106093.

S&P Global Ratings (2019). Guide to Credit Rating Essentials: What are credit ratings and how do they work. *Standard & Poor's Financial Services LLC*, 1–22.

Stevenson, M., C. Mues, and C. Bravo (2021). The value of text for small business default prediction: A Deep Learning approach. *European Journal of Operational Research 295*(2), 758–771.

Taddy, M. (2012). On Estimation and Selection for Topic Models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics 22*, 1184–1193.

Tang, X., S. Li, M. Tan, and W. Shi (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting 39*(5), 769–787.

Tenenbaum, J. B., V. de Silva, and J. C. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*(5500), 2319–2323.

Thanh, S. D., N. P. Canh, and N. T. T. Ha (2020). Debt structure and earnings management: A non-linear analysis from an emerging economy. *Finance Research Letters 35*, 101283.

Tirole, J. (2010). *The Theory of Corporate Finance*. Princeton, N.J.: Princeton University Press.

Traczynski, J. (2017). Firm Default Prediction: A Bayesian Model-Averaging Approach. *Journal of Financial and Quantitative Analysis 52*(3), 1211–1245.

Tsai, C.-F., Y.-F. Hsu, and D. C. Yen (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing 24*, 977–984.

U.S. Federal Government (2021). Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning. *Federal Register 86*(60), 16837–16842. Available at: https://www.federalregister.gov/documents/2021/03/31/2021-06607/request-for-information-and-comment-on-financial-institutions-use-of-artificial-intelligence.

van der Maaten, L. and G. Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*(86), 2579–2605.

Welch, B. L. (1947). The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika 34*(1/2), 28–35.

West, R. R. (1970). An Alternative Approach to Predicting Corporate Bond Ratings. *Journal of Accounting Research 8*(1), 118–125.

White, M. J. (1989). The Corporate Bankruptcy Decision. *The Journal of Economic Perspectives 3*(2), 129–151.

Wu, W., J. Chen, Z. B. Yang, and M. L. Tindall (2021). A Cross-Sectional Machine Learning Approach for Hedge Fund Return Prediction and Selection. *Management Science 67*(7), 4577–4601.

Yao, X., J. Crook, and G. Andreeva (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research 263*(2), 679–689.

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