Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

Jørgen Bolkan and Fredrik Mack Rørvik

Application of XAI in Default Prediction of Commercial Real Estate Companies

Master's thesis in Financial Economics
Supervisor: Snorre Lindset
June 2024

NTNU
Norwegian University of
Science and Technology

Jørgen Bolkan and Fredrik Mack Rørvik

Application of XAI in Default Prediction of Commercial Real Estate Companies

Master's thesis in Financial Economics
Supervisor: Snorre Lindset
June 2024

Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Preface

This Master thesis was done as part of the Masters in Financial Economics at the Norwegian University of Science and Technology during the spring of 2024. It has been an eventful journey with many ups and downs that marks the end of many years of studying for the both of us.

Ambitious as we were, we chose to create our own research problem, without fully knowing the challenges involved. It has been a steep learning curve involving many frustrating nights staring at the same dataset or the same python error message, not knowing how to move forward. However, we were able to pull through and write a thesis that reflects our interests and our expertise regarding machine learning and default prediction.

We would like to thank our supervisor Snorre Lindset for his constructive feedback. We would also like to thank Frank Isaksen at SNN who encouraged us to dive deep into the CRE sector, Mirko Drazic at FundingPartner who inspired us to use machine learning, and ENIN for being helpful in providing us with high quality data. We are sincerely grateful for your contribution and domain expertise. Furthermore, we would like to thank our families and friends for their continuous support during the process of writing this thesis. We would specifically like to thank Kari Bolkan Øverland for being willing to proof-read our thesis.

Abstract

The field of machine learning (ML) and explainable artificial intelligence (XAI) has developed rapidly the last ten years. This development has generated increased interest in the use of complex machine learning models to more accurately assess the credit risk of investments. However, despite machine learning techniques being flexible in their implementation, we observe a lack of development focusing on sector-specific default prediction. We therefore develop an extreme gradient boost (XGBoost) supervised machine learning model specifically for the default prediction of small and medium sized commercial real estate (CRE) companies. CRE is chosen due to the heavy exposure banks have towards this sector and the importance it holds for a country's financial stability. We implement the XAI technique SHAPley additive explanations (SHAP) for post-hoc interpretation of the XGBoost, and prove that the XGBoost provides higher performance compared to a traditional logistic regression (LR) model. Performance is evaluated and compared using ROC-AUC and PR-AUC. Model training is done with increasingly large datasets from 2012-2021, starting with 2012-2017, where testing is done at a one year forward horizon from 2018-2022. Using global SHAP values we find that the features Total Liabilities / Total Assets, the Electricity Price, and Operating Income / Total Assets contribute the most clearly in increasing default probability. In addition, we implement local SHAP values to analyze the features associated with the default prediction of two randomly chosen firms. Overall, we find that several sector-specific features are important for the default prediction in the CRE sector, which emphasizes the need for further sector-specific credit risk research in the future.

Sammendrag

De siste ti årene har vi sett at maskinlæring og forklarbar kunstig intelligens har hatt en tydelig og hurtig utvikling. Denne utviklingen medfører økt interesse og mulighet for å benytte komplekse maskinlæringsmodeller til å mer presist vurdere kredittrisiko av investeringer. Til tross for at maskingslæringmetoder er fleksible i implemetering, så observerer vi en mangel på utvikling som fokuserer på sektorspesifikk konkursprediksjon. Vi utvikler derfor en ekstrem gradientforsterkende veiledet maskinlæringsmodell (XGBoost) spesifikt for konkurspredikering av små og mellomstore næringseiendomsselskap. Næringseiendomsselskap ble valgt på grunn av bankene sin høye eksponering mot denne sektoren samt viktigheten sektoren har for et lands finansiell stabilitet. For å forklare modellen implementerer vi SHAPley additive forklaringer (SHAP) for post hoc fortolkning av XGBoost, og beviser at XGBoost gir høyere ytelse sammenlignet med en tradisjonell logistisk regresjonsmodell (LR). Vi evaluerer og sammenligner resultatene gjennom ROC-AUC og PR-AUC. Modellen ble trent ved bruk av rullende perioder der vi gradvis øker treningssett fra 2012-2021, der vi starter med 2012-2017. Vi utvikler modellen til å predikere ett år i forveien ved bruk av test sett som inneholder suksessivt ett år om gangen fra 2018-2022. Globale SHAP verdier viser at variablene Sum Gjeld / Sum Eiendeler, Elektrisitets Pris og Sum Driftsinntekter / Sum Eiendeler bidrar mest til en økning i sannsynlighet for konkurs. Videre, bruker vi lokale SHAP verdier for å vurdere hvilke variabler som er mest tilknyttet sannsynligheten for konkurs i to tilfeldige valgte selskaper. Samlet sett finner vi at flere sektorspesifikke variabler er viktige for konkurrsprediksjonen i næringseiendomsektoren, noe som understreker det videre behovet for ytteligere sektorspesifikk forskning på kredittrisiko i fremtiden.

Table of Contents

List of Figures

List of Tables

1 Introduction and Motivation

The small and medium sized enterprises (SMEs) in Norway represent 99% of all registered companies, amounting to almost 50% of the value creation in the Norwegian economy (SSB, n.d.-b). In relative terms, the commercial real estate (CRE) sector is the sector with the biggest concentration of both employees and created value within SMEs compared to larger companies in the same sector (NHO, 2024). In fact, 98% of the value created in the CRE sector comes from SMEs, making them a crucial puzzle piece for the economy as a whole. However, CREs require large amounts of capital investment, much of which is financed by debt, making companies in the CRE sector heavily leveraged. The capital structure of CRE companies, as well as the generally higher credit risk associated with smaller businesses, therefore makes accurate credit risk analysis of SMEs in the CRE sector a necessity for potential creditors. Furthermore, CRE is by far the largest sector by bank loans in Norway (SSB, n.d.-a) and has historically been the cause of large financial losses for banks in poor economic times (Hagen et al., 2018). These financial losses would suggest that traditional credit risk models fail to accurately capture the true credit risk present in the CRE sector. In order to capture the inherent risks specific to the CRE sector, we believe new models need to be developed that are better able to capture the dynamics and nuances of this specific sector.

Credit risk is considered the possibility of taking on a loss due to a borrower's failure to repay a loan or meet a contractual obligation. Such a failure will only occur if there is an interruption in the cash flow of the debtor, which most often occurs due to insolvency. As such, credit risk analysis is often concerned with the probability of default (PD) and consequently the loss taken upon default, called loss given default (LGD). Although, LGD is an important aspect of credit risk, we will not include it as a part of this thesis, and will instead focus exclusively on PD prediction. A lot of research has gone into creating models aimed at preemptively identifying firms that will default by looking at their financial ratios, doing a structural analysis of their capital structure, or more recently through machine learning (ML) algorithms. Recent research would suggest that developing such models specifically for SMEs is a necessity (E. I. Altman and Sabato, 2007), and that using ML models to capture a diverse set of characteristics of SMEs

is a good developmental direction[1].

The development of ML models in the banking industry has been slow and stagnant (European Banking Authority, 2022). This slow development specifically relates to the implementation of black-box ML models. Such models are able to accommodate for larger and potentially unstructured datasets, and can extract information from more complex non-linear relationships (Linardatos et al., 2020). But black-box models are difficult to interpret because they provide results without explanations. Therefore, the European Banking Authority (EBA; European Banking Authority, 2022) among others are critical to the practical implementation of black-box models in default prediction[2]. Modern advances in ML techniques, however, has led to the development of explainable artificial intelligence (XAI) techniques. These techniques include post-hoc methods of interpreting results (Angelov et al., 2021) that can be applied in combination with black-box ML models to both improve performance and provide interpretable results.

The extreme gradient boost (XGBoost; Chen and Guestrin, 2016) is one such black-box ML model that has been used plentiful in recent default prediction studies. Together with the XAI technique Shapley additive explanations (SHAP; M. S. Lundberg and Lee, 2017; S. M. Lundberg et al., 2019), the XGBoost has seen excellent default predictive performance (Son et al., 2019; Nguyen et al., 2023). The SHAP can be applied to quantify the impact each input feature has on the model prediction. For credit risk analysis, this implies that we can understand which variables are the biggest cause for any given estimated default probability. The SHAP can provide both global feature importance scores for an entire dataset and local feature importance scores for individual observations (S. M. Lundberg et al., 2019). In other words, for any given model output we can assign an importance value to all input variables which would allow us to explain for instance which facet of a given company is the biggest contributor to the predicted PD.

Following the recent developments within XAI and ML, but the corresponding lack of focus on sector-specific PD models, the aim of this thesis will be the following:

[1] See Ciampi et al., 2021 for a comprehensive review of SME default probability studies and their proposed future direction of the field.

[2] These concerns regard adherence to regulations defined by the EBA in their capital requirements regulation (CRR) for the internal rating based (IRB) approach, such as human judgement (European Banking Authority, 2024a) and external understanding (European Banking Authority, 2024b.

1. Develop an extreme gradient boost supervised ensemble-based machine learning model for predicting the probability of default in Norwegian small and medium sized commercial real estate companies at a one year forward time horizon.

2. Implement explainable artificial intelligence techniques by using Shapley additive explanations to interpret the probabilities of default estimated through the extreme gradient boost model.

The development of such a model will complement the limited, but existing research on the application of ML and XAI in predicting the PD of Norwegian companies (see Paraschiv et al., 2023). Given that there is no agreed upon theory for why companies go bankrupt, research tends to be highly empirical and sector dependent (Perboli and Arabnezhad, 2021). This thesis will focus its research on SMEs in the CRE sector in Norway, as sector specific research can increase model accuracy and better accommodate for sector specific parameters (E. I. Altman and Sabato, 2007).

We are going to compare three different models in this thesis. The first is an XGBoost using SHAP feature selection to find the 18 most important features from a set of 189 total features. The second will be a standard logistic regression (LR) utilizing the four most important features in order to reduce problems concerning dimensionality. Finally, the third model will be an XGBoost modeled using the same four features used in the LR, the aim of which is to see whether the XGBoost performs as well using fewer features, or if the added complexity of a significantly larger feature set increases performance. Both the XGBoost and the LR are supervised ML models, implying we are going to make a training set to teach the models, and a test set to validate. The validation and comparisons of the models will be done using traditional classification evaluation metrics for machine learning. These metrics include the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve. These metrics will be discussed further in the theory section.

The thesis will be structured in the following way. The second section will be dedicated to explaining the history of credit risk analysis and bankruptcy prediction, the different approaches used, as well as key advances in machine learning modelling of credit risk. The third section will introduce some theory concerning how bankruptcy is defined, some crucial ML techniques, and lastly present typical metrics used for evaluating binary classification tasks. The fourth section

delves into the methodology of our chosen models, and their underlying frameworks. The fifth section concerns how we collected our data, what data we are specifically working with, and how we apply our chosen methods. The sixth, and penultimate section will explain the analysis including the results and findings related to our research problem. Lastly, the seventh section will conclude the thesis, provide some limitations of the findings, and discuss future research directions that can build on our presented results.

2 Background

Research on the credit risk of SMEs in Norway is fairly limited. There has been some studies in recent years (Andersen et al., 2021; Paraschiv et al., 2023), however few of these have attempted to implement ML modelling (see, Paraschiv et al., 2023). ML models have been found to be accurate in predicting default of SMEs in other countries, and as such there is an inherent need to assess their potential in the Norwegian market as well (see e.g., Perboli and Arabnezhad, 2021; Schalck and Yankol-Schalck, 2021). Given the large number of SMEs in Norway, several studies suggest that traditional statistical bankruptcy prediction models may be less effective (E. I. Altman and Sabato, 2007; Ciampi et al., 2021). Therefore, exploring ML techniques could potentially significantly enhance the predictive accuracy of SME default models.

2.1 The History of Default Prediction modelling

As with all fields of finance, the modelling of credit events is always developing with the advancements of technology. From the 1930's to the 1960's all studies used univariate, single factor models where the only variable was a significant financial ratio (see Beaver, 1966). The first paradigm shift came during the end of the 60's when the first multivariate study was published in the form of Altman's Z-score (E. I. Altman, 1968). Then, in 1974, in the backdrop of the successful development of the Z-score, Merton published his famous model on the basis of Black and Scholes option-pricing theory (Merton, 1974). The Merton model, as opposed to the Z-score, implements economic theory in its modelling. However its performance is not strictly better than the Z-score due to the struggles of its implementation, and the strong underlying assumptions that come with the theoretical framework (Das et al., 2009). For instance, the Merton model only accommodates for zero-coupon bonds, and only calculates PD at maturity (Merton, 1974). The model has been built upon in more recent models to incorporate more complex

aspects of bankruptcy and to loosen the assumptions imposed. These additions include *distance to default* (DD) models which measures the PD as a function of how far the company's asset value is from a defined default boundary at any given time (see Vassalou and Xing, 2004). Other models incorporate *first passage time* and argue that bankruptcy occurs whenever a company's asset value drops below a given boundary, most often defined through bankruptcy covenants (Black and Cox, 1976). In Norway specifically, a model which incorporates sector-specific re-organization boundaries to accommodate for differences in liquidity across industries, has also been studied (Andersen et al., 2021).

Following the turn of the decade and distraught by the lack of statistical theory underpinning the Z-score, Ohlson (1980) developed his own model. This model is a conditional probability model using logistic regression, whereby the output, called the O-score, itself is an isolated probability of default. Thus, the model is self-contained and is not based on exogenous arbitrary default boundaries such as those found in Altman's Z-score [1]. Today, the O-score remains among the most popular methodologies for bankruptcy prediction, yet it too has its own complications. For instance, it tends to be inaccurate at long time horizons (longer than 12 months) and the underlying assumptions of the true loss function distribution have been proven to be wrong, and extreme value theory, able to better capture extreme events could be more applicable (Perboli and Arabnezhad, 2021).

The next paradigm shift in credit risk analysis came with the introduction of Neural Network (NN) models in the 1990's and early 2000's, the first models that implemented machine learning techniques (see Lacher et al., 1995 and Efrim Boritz and Kennedy, 1995). NNs are inspired by the workings of the human brain with nodes acting as neurons which are connected by edges, modelling synapses. An NN consists of at least three layers of nodes, an input layer, one or more hidden layers, and an output layer. The hidden layer is where the algorithm that transforms a set of input variables into a given predicted output is based. This algorithm transforms the data using some internal weighting, and passes it through an activation function which seeks to best learn the complex, non-linear relationship that exists between the input variables and the company's solvency. A trade-off for being able to model these complex relationships is that the internal activation function becomes impossible to interpret, and what occurs in the

[1] Altman's Z-score sets a boundary condition whereby companies with a score higher are predicted as bankrupt, and companies with a score lower are considered healthy. This boundary is exogenously set based on initial findings in Altman's research, and subsequently is not based on dynamics of a specific company, but based on the initial dataset used by E. I. Altman, 1968.

hidden layer is not seen or understood by even the developer himself (Lacher et al., 1995). As such, there was a growing concern of whether it was the algorithm in the model or the humans behind it that made the final decision, for instance, to credit applications (Addo et al., 2018).

Despite the multitude of default prediction studies applying different paradigms, there is a lack of consensus regarding which models perform the best (Wang et al., 2014). Several different studies have found results that severely contradict each other (see Bauer and Agarwal, 2014). This finding implies that the field is highly empirical, and that the best methodologies including models, variable selection and out-of-sample forecast horizons can vary across countries, sectors, and firm sizes. For instance, even though E. I. Altman et al. (2020) find that logistic regression performs better than even some ensemble-based machine learning techniques, other studies have found the opposite to be true (Perboli and Arabnezhad, 2021; Alfaro et al., 2008; Jones et al., 2017), and Wang et al. (2014)) even reported no significant differences between the two methods.

Even within the different statistical models and across various ML models, studies have come to different conclusions. Bauer and Agarwal (2014) for instance find several different studies that argue either in favor of accounting-based models (Reisz and Perlich, 2007), contingent claim models (Hillegeist et al., 2004) or hazard models (Shumway, 2001; Campbell et al., 2008). This lack of coherence also holds true regarding the plentiful of ML models, where neural networks (E. I. Altman et al., 2020), random forests (Petropoulos et al., 2020), or extreme gradient boosts (Nguyen et al., 2023) have been shown to more accurately predict bankruptcy. As such, one should be careful with assuming the transferability of methodologies across countries, between sectors and across various firm sizes, and any conclusions should be done with the backing of empirical analysis.

In the view of these methodological complications, machine learning techniques, specifically ensemble-based models such as extreme gradient boosting can be more readily replicated across various data sets (Jones et al., 2017). These models require less intervention for data preparation, have less stringent requirements for variable selection, and have a less rigid model architecture compared to other bankruptcy prediction models. In general, ensemble-based models have proven to achieve higher accuracy across both longitudinal and cross-sectional data in predicting corporate failure (Jones et al., 2017).

Ensemble methods utilize multiple weaker ML models together to achieve a higher comprehensive predictive performance. Several different ensemble-based models have been used in bankruptcy prediction studies, including AdaBoost (Alfaro et al., 2008), LightGBM (De Lange et al., 2022), Random Forest (Petropoulos et al., 2020) and XGBoost (Nguyen et al., 2023; Perboli and Arabnezhad, 2021). Across most reviews of the performance of these models, XGBoost seems to perform the best (see Nguyen et al., 2023; Son et al., 2019). In addition, the framework and system packages for the XGBoost are well developed and easy to use for quick implementation (Chen and Guestrin, n.d.). The XGBoost has been used across several different fields of research, including bankruptcy prediction, and is well developed for binary classification analysis (Chen and Guestrin, 2016). Binary classification is the task of classifying a set of data into two mutually exclusive categories, which in the case of bankruptcy is *bankrupt* and *solvent*. The XGBoost will be further discussed in the methodology section.

2.1.1 Default Prediction of SMEs

Research pertaining to credit risk, specifically in small and medium sized enterprises has seen an exponential rise since the 2008 financial crisis (Ciampi et al., 2021). This increase in studies is in part due to the substantial financial hit SMEs took during these times, and in part due to the implementation of a new Basel III framework. Basel III was developed to increase financial stability by imposing higher regulations and rules on the risk taken by banks to avoid another collapse as seen in 2008 (Bank for International Settlements, 2017). However, a since noted effect of the increased minimum capital requirements and liquidity imposed on banks is the reduced financing available for SMEs (Marek and Stein, 2022; Angelkort and Stuwe, 2011).

Unlike large corporations, which often easily secure debt financing due to robust historical results and future profit projections, SMEs more often face challenges in attracting investors and managing debt, especially during periods of rapid economic or industry-specific downturns. Furthermore, external events which either increase the likelihood of default or decrease it, such as macro economic events or financial interventions like government bailouts or central bank policies, have a greater impact on smaller companies compared to larger ones (Ciampi, 2015). Therefore, credit risk modelling of SMEs can often be quite different in comparison to larger companies, which is part of the problem with traditional credit risk models such as Altman's Z-score (E. I. Altman, 1968). The Z-score specifically was developed on the basis of financial ratios in large companies, and may not reflect the nuances found in SMEs. Financial ratios

that predict bankruptcy in large companies may not apply to SMEs, either due to data scarcity or the significant impact of minor accounting changes on financial ratios, which can mislead the models (Ciampi and Gordini, 2013). Moreover, several studies have demonstrated that the performance of generic corporate models solely using financial ratios is notably poor in predicting SME defaults, suggesting that simply relying on quantitative data is insufficient (E. I. Altman and Sabato, 2007; Ciampi, 2015).

Instead of solely relying on financial ratios, several studies have tried to incorporate qualitative, macroeconomic, and sector-specific features to increase performance of SME PD models (Ciampi et al., 2021; Ciampi, 2015). Qualitative features include information on accounting flags, CEO duality, and the individuals on the board of directors. Although these qualitative features have been proven to increase model performance, they are also substantially more difficult to acquire good data for (Ciampi et al., 2021). Macroeconomic and sector-specific features, however are easier to implement. Macroeconomic features have been included in several studies that have demonstrated an increase in overall model performance due to their implementation (E. I. Altman et al., 2017; Filipe et al., 2016). On the other hand, there seems to be a lack of focus on research pertaining to individual sectors. The reasons for this lack of research in specific sectors is unknown to us, but it could potentially be due to limited access to bankruptcy data when exclusively focusing on individual sectors, which can decrease model performance. Due to this lack of research on individual sectors, Andersen et al. (2021) argue that this should be the next step taken in credit risk analysis.

2.2 The Commercial Real Estate Market in Norway

Although, there are many sectors of interest in Norway, perhaps the most connected to the overall financial stability is the commercial real estate sector. Commercial Real Estate (CRE) encompasses real estate that is exclusively used for business purposes, the aim of which is to generate a profit through either capital gain or rental agreements (European Systemic Risk Board, 2015). It generally consists of four different categories of real estate, being offices, retail spaces, hotels, and industrial and logistic properties (Hagen, 2016). A lot of CRE is heavily financed by debt, mostly provided by banks through loans (Hagen et al., 2018). Therefore, banks are heavily exposed to the CRE sector as a whole. This exposure has historically caused large losses for banks, both in Norway and internationally, during poor economic times such as the financial crisis in 2008 (Hagen, 2016).

Part of the reason for banks' willingness to debt finance these companies so heavily is that they take collateral in the properties themselves. In fact about 80% of the market value of the properties banks have collateral in are owned by CRE companies (Bjørland et al., 2022). The bulk of the market value of these properties falls into the *office* category, which represents about 40% of the market value of all CRE properties (Hagen, 2016). But, the value of office spaces can change drastically depending on macro-economic variables (Bjørland et al., 2022). An implication of such a high sensitivity to macro-economic factors in the value of office spaces, is an increased credit risk for the credit service providers (CSPs), such as banks.

As a measure to reduce potential losses in banks during economic crisis, and indirectly increase financial stability, regulatory capital requirements such as Basel III have been introduced (Bank for International Settlements, 2017). One study found that CRE was the first sector to experience a cut in available debt from banks as a consequence of new capital requirements introduced after the 2008 financial crisis (Bridges et al., 2014). Consequently, while these regulations aim to bolster the overall stability of the economy, they can inadvertently place a strain on the SMEs that contribute significantly to economic growth, such as small and medium sized CRE companies (Bjørland et al., 2022).

Therefore, in order to effectively make use of available sources of investment, SMEs in the CRE sector may need to adopt more robust risk management practices and enhance their financial transparency (European Systemic Risk Board, 2015). Such practices could aid banks and other CSPs to more accurately identify the risk inherent in CRE companies, subsequently leading to more available capital. This increased financial transparency may become a necessity, as European regulatory systems themselves have noted the lack of available data and information regarding agents in European CRE markets (European Systemic Risk Board, 2015). In order to accommodate for such an increase in information, newer default prediction models would have to be developed, specifically for SMEs in the CRE sector.

Although recent studies on credit risk seems to have increased its focus on SMEs at large, sector-specific models are still lacking. As Andersen et al. (2021) mention in their development of a structural model for credit risk analysis in Norwegian SMEs, there is potential for further optimization by focusing on individual sectors. Therefore, by modelling CRE companies separately from SMEs in general, it could become possible to incorporate more nuances of the CRE market that could aid in increasing transparancy and thus better predict defaults. Such a task

could more easily be done using ML techniques which are able to handle a diverse set of feature inputs in a flexible way (Jones et al., 2017).

2.3 Interpretability of Machine Learning Models

The major problem with implementing complex ML models concerns compliance with international rules and regulations. According to the EBA, any estimate of the PD set by an institution needs to be "plausible and intuitive" (European Banking Authority, 2024b). Other concerns relates to the level of human understanding and judgement involved (European Banking Authority, 2024a), as well as the structure of the theoretical framework and the underlying assumptions underpinning the model (European Banking Authority, 2024c). A common definition of black-box ML models is that they "do not explain their predictions in a way that humans can understand" (Rudin, 2019, p. 1). As such, black-box models cannot by themselves be used by the CSP, as the CSP lack a sufficient understanding according to the EBA of the bankruptcy prediction produced by the model. One solution to this interpretability issue is the application of explainable artificial intelligence (XAI) techniques, which were developed to increase the "understanding and interpretation of the behavior of AI systems" (Linardatos et al., 2020, p. 2). The application of XAI has gained traction in recent years due to the increase in unstructured data sets, big data, and the use of ML techniques such as deep learning (Angelov et al., 2021). The notion of XAI methods is to provide an accurate intermediary between the complex nature of the ML systems and the novel understanding of humans (Linardatos et al., 2020). Through this intermediary, which S. M. Lundberg et al. (2019) call approximated explanation models, the aim is to make interpretation of the relationship between input features and model output comprehensible for humans.

To evaluate how good XAI methods are at providing adequate interpretable results for humans to understand, Phillips et al. (2021) proposed four distinct criteria. These four criteria are individually called explanation, meaningful, accuracy, and knowledge limits. *Explanation* encompasses the principle that explainable ML methods must provide explanations in the form of evidence and reasons for its output. *Meaningful* implies that the explanation provided must be understandable by human standards, and provide meaningful insights to their comprehension of the model output. *Accuracy* states that the approximated model output must accurately reflect the original model output. Lastly, *knowledge* limits entails that the XAI methods must be aware of its own limits, and provide feedback whenever it operates outside of its designated area of

expertise.

These criteria are relatively new, and the degree to which various XAI methods comply with them is yet difficult to determine. Despite the lack of such testing, the SHAP framework has seen frequent implementation in credit risk analysis (European Banking Authority, 2023). The SHAP performs especially well when interpreting the PD estimated using decision tree ensemble-based models, such as the extreme gradient boost (XGBoost; Nguyen et al., 2023) and the light gradient-boosting machine (LightGBM; Andersen et al., 2021). It has been proven to interpret these models in a way that is intuitive for humans to understand (M. S. Lundberg and Lee, 2017) and it provides meaningful insights for understanding the relationship between feature inputs and output prediction. Furthermore, the SHAP is highly accurate when interpreting locally, and provides consistent explanations for predictions (M. S. Lundberg and Lee, 2017;S. M. Lundberg et al., 2019). Yet, most applications of the SHAP is at the global level as a method of increasing model accuracy. Not enough applications is seen at the local level as a way of individual feature interpretations (Nguyen et al., 2023). Applying SHAP as a way of interpreting local PD predictions is essential, especially in heterogeneous markets such as commercial real estate, where companies can differ substantially as to the individual risk they are exposed to (European Systemic Risk Board, 2015).

3 Theory

In order to accurately predict what facets of a company are the drivers behind bankruptcy, we need to understand specifically how bankruptcy works. Therefore, this section firstly relates to how bankruptcy is defined in the literature, which includes how the bankruptcy process works, and how that relates to the credit risk experienced by a CSP. Furthermore, some principles behind ML that enables the increased performance of our chosen model is discussed. Lastly, the standard evaluation metrics used for evaluating the performance of classification ML models are presented.

3.1 Definition of Bankruptcy

In a perfect capital market, as presented by Modigliani and Miller (1958), the capital structure of a company is irrelevant. However, a perfect capital market does not exist most notably due

to taxes, bankruptcy costs, and other market frictions. Because interest is tax-deductible, there is an optimal capital structure for a company that balances debt and equity. If we assume taxes to be present, the value of a company increases proportionally with the amount of debt added due to a tax shield on interest payments, up to the point where the interest payments exceed the company's EBIT. In essence, to maximize firm value, one would increase debt to the maximum capacity. However, in practice there are other market frictions which ultimately reduces firm value when debt is increased too much, most notably the risk of bankruptcy.

The risk of bankruptcy brings costs associated with the possibility of defaulting on outstanding debt. Direct bankruptcy costs such as lawyer and accountant fees reduce the value of the assets left in the firm. Indirect bankruptcy costs caused by financial distress, whether or not a firm is declared bankrupt, may be of substantial size (Hillier et al., 2019, p. 589). Especially relevant for CRE companies are the indirect costs related to fire sale of assets. In the hope of avoiding bankruptcy a CRE company may try to sell off properties to try and avoid bankruptcy proceedings. The lack of liquidity due to CRE being a heterogeneous product (Bjørland et al., 2022) may cause significant liquidity discounts when a distressed firm needs to sell quickly. Not only may the indirect bankruptcy costs pose a threat to the CRE companies, but also for creditors such as banks, especially if the loan to the CRE company was a significant asset to the bank.

Bankruptcy is generally referred to as the legal process a firm is put in because they cannot fulfill their outstanding debt or other obligations (E. I. Altman, 1968, Warner, 1977). When a firm is leveraged by debt there are two parties involved, the debtor and the creditor. In Norway, the bankruptcy process starts when either the debtor or creditor files for bankruptcy as an appeal filed to the district court where the firm is registered. For a company to be declared bankrupt, it must be insolvent. Insolvency can occur in two forms: either the company cannot meet its financial obligations as they come due, or the value of its assets does not exceed the value of its liabilities (Altinn, 2024). These two types of insolvency are referred to as cash flow insolvency and balance sheet insolvency, respectively. More formally, cash flow insolvency occurs when a firm cannot meet its financial obligations at their due dates, and balance sheet insolvency refers to the situation where liabilities exceed total assets (Uhrig-Homburg, 2005).

An important nuance is that insolvency and bankruptcy do not necessarily occur simultaneously. Creditors may treat a firm as if it were insolvent even when it is not, due to a high

probability of future insolvency. This behavior can lead to situations where the decision to file for bankruptcy is influenced by subjective judgment (Jackson and Scott, 1989). As such when deciding whether to declare bankruptcy, it is crucial to consider whether the firm is experiencing economic or financial distress. Economic distress occurs when the firm cannot generate sufficient revenue to cover its operational expenses. Financial distress is caused by excessive leverage, where the interest costs prevent the firm from achieving positive earnings (Onakoya and Olotu, 2017). For a CRE company, the combination of high leverage and high capital intensity, makes them especially vulnerable to increased interest payments or a drop in rental prices (Hagen et al., 2018). It is therefore crucial to consider both the financial health and economic health of a firm when evaluating its overall stability. A firm is considered financially healthy if it can pay its debts on time, while it is considered economically healthy if it can produce goods or services efficiently and profitably. Notably, a firm might be financially distressed, but still have a viable business model and the potential to recover and be profitable in the long run. Conversely, it might be able to pay its debt, but lack an efficient business model (Adler, 1997).

3.2 Machine Learning

A creditor will be most concerned with the financial distress of a company. As such, they will be interested in the company's ability to repay their debts in the future. Therefore, they would implement default prediction forecasts of the companies at the individual level. As previously mentioned, machine learning (ML) is especially suitable for such an analysis. ML is a subset of artificial intelligence which focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit human programming (El Naqa and Murphy, 2015). These tasks involve recognizing patterns, making predictions, or classifying data. Typically, ML algorithms are categorized along four different approaches: supervised, semi-supervised, unsupervised, and reinforcement learning (Sarker, 2021). The models used in this thesis utilize supervised machine learning algorithms, and we will therefore not elaborate on the other three. Supervised ML involves giving a model a patterned set of inputs and their corresponding true output. Then, the model is trained such that it is able to learn these patterns to subsequently identify them when working with new data not seen before. The aim is to generalize these patterns and make the model understand and correctly predict an output for a new dataset, solely based on a set of input features (Sarker, 2021). Therefore, supervised ML algorithms have datasets split in two, one training set to identify the patterns, and one test set to validate

them.

Supervised ML can be applied for two distinct applications: regression analysis and classification analysis (Nasteski, 2017). A key difference between them lies in the output of the model. Regression analysis exists in the real space, and thus the model output is continuous. Classification analysis on the other hand, produces some categorical (discrete) output defined by preexisting categories, or some probability of belonging to that category. In relation to bankruptcy prediction, classification analysis is naturally used, as the motivation lies in identifying the probability of going bankrupt. Specifically, bankruptcy prediction is considered binary classification analysis, whereby the model output is a probability of default. A threshold is then defined to categorise the company as either solvent or bankrupt depending on the PD.

A problem that arises in supervised ML concerns how well the model captures actual relevant data features and not noise. This relates to overfitting, a concept in ML whereby the model effectively reduces the loss in the training data, but not on the test data (Tian and Zhang, 2022). Overfitting can be caused by having overly-complex models that are fitted too well on the training data. This complexity makes the model good at classifying training data observations, but lacks the power to classify out-of-sample observations. In relation to bankruptcy prediction, this noise causes a lot of company misclassifications, and thus poor model performance. In order to cope with these problems concerning overfitting, several measures have been introduced, some of which will be discussed here (Ying, 2019).

3.2.1 Regularization

They aim of a supervised ML model is to identify patterns in the training dataset that it can apply to out-of-sample observations. As the field has developed, more advanced and complex techniques have been developed to ensure the model does not fit too much noise. These techniques include regularization technology which is used to increase the model generalization by imposing a penalty on the complexity of the model (Tian and Zhang, 2022). Two standard regularization techniques are the L1 (Lasso[1]) and L2 (Ridge) regularization parameters. Both L1 and L2 regularization is used to reduce the effect of individual features on the model as a whole which reduces overfitting. They are added onto a model loss function, creating an objective function. L1 regularization can be defined as

[1]Least Absolute shrinkage and selection operator.

$$L(x) = l(x) + \alpha \sum ||\beta_m|| \tag{3.1}$$

where $L(x)$ is the objective function, $l(x)$ is the original loss function, α is the L1 regularization parameter, and β_m is a model coefficient for feature m. When optimizing the objective function this model constraint encourages the absolute value of the individual feature coefficients to be smaller. The L1 regularization parameter often leads to coefficient values of zero, effectively negating that feature from the model. This effect can be especially beneficial when working with high-dimensional data including many irrelevant features (Ng, 2004). As such, L1 regularization effectively simplifies the model by optimizing the number of relevant features which makes it easier to interpret model outputs (Schaaf et al., 2019).

L2 regularization is also used to reduce the effect of individual features on the entire model. The typical formula for L2 regularization can be expressed as

$$L(x) = l(x) + \lambda \sum ||\beta_m^2|| \tag{3.2}$$

where all parameters are the same as in the L1 regularization case, and λ is the L2 regularization parameter. Because L2 regularization introduces a polynomial model constraint (instead of an absolute one), the coefficients are not directly set to 0, simply minimized. The effect of L2 regularization is again to reduce the effect of individual features on the entire model, thus reducing overfitting. Both L1 and L2 regularization are model constraints, that limit the possible feature coefficient combinations in different ways. They are additions to any model that changes the optimal weights assigned to individual features. In addition to regularization techniques, overfitting can also be reduced by using several less complex models in combination, which in ML are called ensembles.

3.2.2 Ensemble Methods

Ensemble methods for machine learning are simply ways of combining several models together into an overarching model that performs better (Dietterich, 2000). “A necessary and sufficient condition for an ensemble of classifiers to perform better than any of its individual members is if the classifiers are accurate and diverse” (Hansen and Salamon, 1990 as cited in Dietterich, 2000, p. 1). A model by itself is accurate if it performs better than simply guessing, and by using a

variety of such models, the ensemble model can more easily capture the diverse and complex facets of a dataset. There are three types of standard ensemble techniques most commonly used in ML today, called bagging, stacking and boosting (Divina et al., 2018; Dietterich, 2000). All three have been used in the development of a range of different ensemble methods. The model developed for this thesis is an XGBoost, which by default utilizes a boosting method. Therefore, we will not go into depth about stacking and bagging. How boosting is used specifically in the XGBoost will be discussed in the methodology section.

Boosting is a way of minimizing missclassification of observations through the use of successive models that sequentially correct the prediction errors. The final prediction, $\hat{y}_n$, is a weighted combination of all individual model predictions (see figure 3.1). By assigning a weight to either a correct or a faulty classification, the model will eventually learn more about a given observation using a set of input features. Boosting is based on the concept that several weak models together will eventually perform better than one strong model alone (Dietterich, 2000).

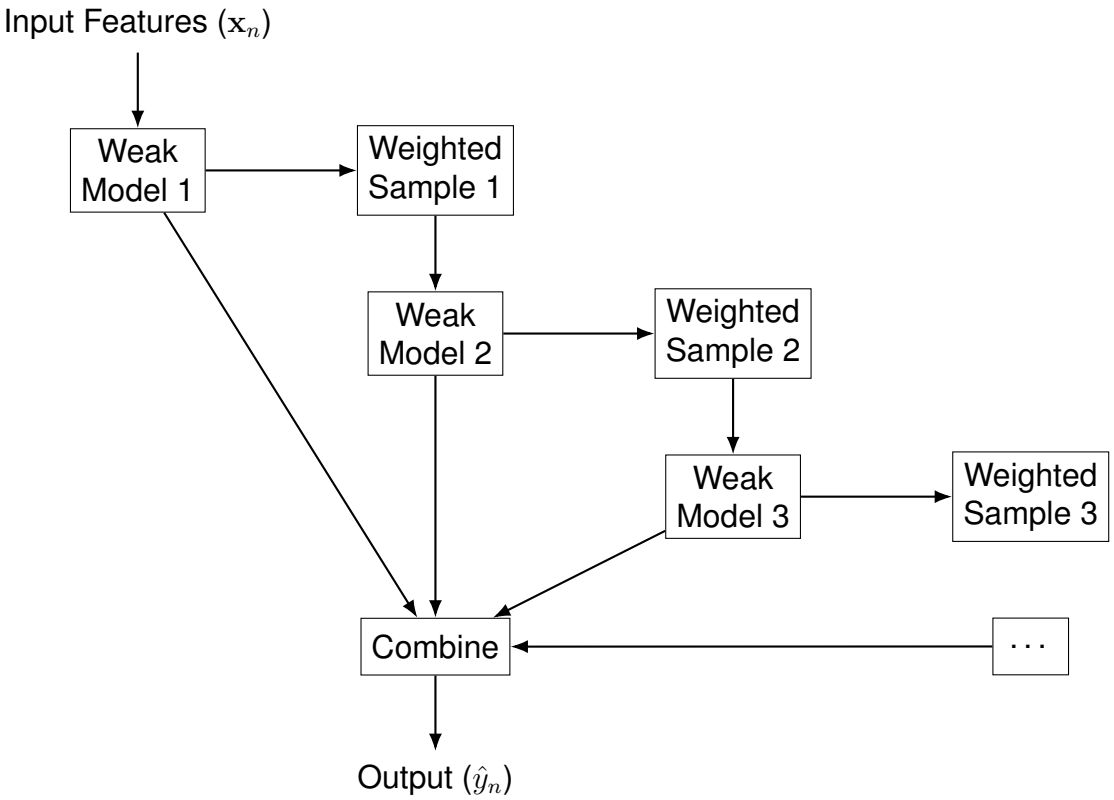


Figure 3.1: Example of the boosting process. Different ensemble-based models use different individual models, but they all combine using some weighed aggregation of individual predictions.

Regularization techniques and ensemble-methods are two common ways of counteracting overfitting in a supervised ML algorithm. In addition, specific techniques are also embed-

ded into individual model frameworks themselves, developed to create better performing algorithms.

3.3 Evaluation Metrics

In order to evaluate how the models perform, several ratios and methods have been developed to compare model performance along a range of metrics. These metrics relate to how well the models reduce either type I or type II errors. In order to avoid any confusion as to what type I and type II errors are in ML and statistical literature, we define the type I and type II errors in a ML context. In the context of binary classification of the solvency of companies using machine learning, type I errors occur when a solvent company is predicted to become bankrupt according to the model. This is called a false positive, as the company is positively identified as bankrupt, despite it actually being solvent. The opposite is called a false negative, or a type II error, whereby an actual bankrupt company is misclassified as solvent.

When working with binary classification tasks such as bankruptcy prediction, because the output is a probability, a threshold is defined. This threshold is defined such that companies that have a higher predicted PD, are categorized as bankrupt. As such, type I and type II errors are highly sensitive to the threshold level set in the model, and different threshold levels could drastically change model performance. The threshold can therefore be an asset as it can be set to increase model performance in the direction of preference according to some evaluation metrics.

Recall, precision and accuracy are three evaluation metrics used to capture different aspects of a ML models' performance. *Accuracy* captures how well the overall model is at classifying all companies correctly as either solvent or bankrupt. This measure is therefore more applicable when working with balanced datasets, as having an overwhelming majority of solvent companies will often make the accuracy high. In the case of working with imbalanced data sets, this measure will therefore capture less important aspects of the models. Accuracy is measured by the following:

$$Accuracy = \frac{\textit{True Positives} + \textit{True Negatives}}{\textit{True Positives} + \textit{True Negatives} + \textit{False Positives} + \textit{False Negatives}} \tag{3.3}$$

where *true positives* are correctly classified bankrupt companies and *true negatives* are correctly classified solvent companies.

Precision is a metric used to measure how many of the predicted bankrupt companies are actually going bankrupt. This metric is therefore important when the misclassification costs of false positives are high. If a model has low precision, then it implies that a lot of solvent companies are identified as bankruptcies. Precision is defined as

$$Precision = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}} \tag{3.4}$$

The last evaluation metric to be introduced is *recall*. Recall may be the most important metric for bankruptcy studies, as the ability to correctly identify bankrupt companies is generally of greater importance than correctly identifying solvent companies. In general, misclassifying a bankrupt company as solvent may lead to a loss that is significantly larger than the opportunity cost of not granting a loan to a solvent firm. Recall measures how many of the positive instances (bankrupt firms) are actually predicted as bankrupt by the model. Recall is defined as

$$Recall = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}} \tag{3.5}$$

Despite recall being of the highest interest for the models in this thesis, there still needs to be a balance between recall and precision, as having too low of a precision will lead to large opportunity costs.

Both precision and recall are sensitive to the threshold defined. A higher threshold means fewer companies are categorized as bankrupt, meaning less false positives, increasing the precision. However it also means more false negatives, as more bankrupt companies are categorizes as solvent, and thus a lower recall. The opposite is true should the threshold be set too low. A trade-off therefore needs to be made such that model performance is optimized. In general we want a model that correctly classifies both solvent and bankrupt companies over a large range of thresholds.

There are two common methods to evaluate classification models according to their average performance over all thresholds. These are the Receiver Operating Characteristic (ROC) curve

and the Precision-Recall (PR) curve. Both measures are quantified using the area under the curve (AUC) of their respective distribution curves, where an AUC value of 1 represents a perfect model. The ROC captures the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). The TPR is the same as the recall, whereas the FPR is the rate of negative cases predicted falsely as positive, defined as

$$FPR = \frac{\textit{False Positives}}{\textit{False Positives} + \textit{True Negatives}} \tag{3.6}$$

The FPR shows the frequency of solvent companies that are predicted as bankrupt. The ROC plots the ratio of recall to FPR for each threshold, going from a threshold of 0 to a threshold of 1. The ROC-AUC is thus a measure of how well a model can classify its positive class without misclassifying too many of its negative class. The ROC-AUC is subsequently not affected by the defined threshold set in the model, since the ROC iterates over all possible thresholds.

The PR curve conversely plots the ratio of precision to recall at different threshold levels. In most cases a high precision will correspond to low precision and vice versa, thus the PR-AUC in essence measures how well on average a model is able to account for both. The PR curve is most notably used when working with highly imbalanced datasets in favor of negative observations, which typically lead an inflation in the number of false positives. Thus, similarly to the ROC curve, the PR curve iterates over all thresholds, but it can better capture the increased prevalence of negative cases seen in imbalanced datasets.

4 Methodology

The methods used to create the models for this thesis are the XGBoost and the logistic regression (LR). Since both are supervised ML models, this involves training the models on a given training set, and then validating one year ahead. In order to best interpret the results found using the XGBoost, we incorporate Shapley Additive Explanations values as an explainable machine learning (XAI) framework.

4.1 Extreme Gradient Boost

The extreme gradient boost (XGBoost) is a supervised ensemble-based ML model developed by Chen and Guestrin (2016), utilizing the ensemble method of boosting, with decision trees acting as weak learners. A simple decision tree contains decision nodes, which splits companies based on their feature input values, and leaf nodes which are terminal nodes corresponding to where the companies end up in each decision tree iteration. Each new decision tree is optimized at the individual leaf node level through the use of gradient descent algorithms. Specifically, the model is optimized with respect to the probability of default predicted in the previous weak learner, similar to other boosting models. The XGBoost framework also incorporates a range of regularization techniques, including L1 and L2 regularization hyperparameters, as well as column subsampling and a shrinkage factor (Chen and Guestrin, n.d.).

Consider a dataset, $\mathcal{D} = \{\mathbf{y}, \mathbf{X}\}$, where $\mathbf{y} = \{y^n\} \in \{0, 1\}^N$ denotes the true classification values of 0 (solvent) and 1 (bankrupt) for $n = 1, ..., N$ company years. $\mathbf{X} = \{x_m^n\} \in \mathbb{R}^{N \times M}$ is a matrix of $m = 1, ..., M$ feature input values for each company year, n. Furthermore, let the bankruptcy prediction be calculated iteratively through $k = 1, ..., K$ decision trees acting as weak learners, defined by the function $f_k(\mathbf{X})$. Each decision tree contains $j = 1, ..., J$ leafs which each company year is mapped onto. The overall model is mathematically defined as

$$\hat{y}_n = f_0 + \eta \sum_{k=1}^{K} f_k(\mathbf{x}_n), \quad f_k \in \mathcal{F} \tag{4.1}$$

where $\mathcal{F} = \{f(\mathbf{X}) = w_{q(\mathbf{X})}\}(q : \mathbb{R}^M \mapsto \{1, ..., J\}, w \in \mathbb{R}^J)$ is the classification and regression tree (CART) space, containing a set of *K* decision trees[1]. $\hat{y}_n$ is the predicted probability of default found in the model. The function f_0 represents the initial guess of the model, which is defined as the average probability of default in the dataset[2]. η is the shrinkage factor used for reducing the effect of individual decision trees on the final ensemble-model prediction, which aids in reducing overfitting. The mapping function $q(\mathbf{X})$ maps all company years onto their corresponding leaf node using their respective feature input values. For each new iteration in the boosting process, the decision tree, f_k, that minimizes the following regularized learning

[1] See Breiman et al., 2017 for an introduction to the use of decision trees in classification tasks.

[2] Both $\hat{y}_n$ and f_0 are expressed as log-odds. In order to transform them into probabilities we have to apply the logistic function. For ease of understanding, we will continue to regard it as a probability until otherwise mentioned.

objective is fitted:

$$\mathcal{L}^{(k)} = \sum_{n=1}^{N} l(y_n, \hat{y}_n^{(k-1)} + f_k(\mathbf{x}_n)) + \Omega(f_k) \tag{4.2}$$

where $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{J} ||w_j^2||$ is a regularization addition. $\Omega(f_k)$ is a penalization term which is a function of a pruning parameter, γ, the L2 regularization parameter, λ, the number of leaf nodes in a given tree, *J*, and the corresponding leaf weights, w_j. How exactly a given leaf weight is calculated can be seen in appendix A. In addition, the L1 regularization parameter, α, can also be added to (4.2) similar to (3.1). The model framework defines $\alpha = 0$ (Chen and Guestrin, n.d.) by default, which is why it isn't included in (4.2). The pruning parameter is set to reduce the complexity of the decision trees such that the most informative splits are retained and the regularization parameter is set to reduce the sensitivity of the overall model to individual observations, thereby avoiding overfitting. Lastly, l represents the loss function. The loss function used in this thesis will be the logistic loss function (also called the cross-entropy loss function) defined as

$$l(y_n, \hat{y}_n^{(k-1)}) = -[y_n \log \hat{y}_n^{(k-1)} + (1 - y_n) \log 1 - \hat{y}_n^{(k-1)}] \tag{4.3}$$

where $\hat{y}_n^{(k-1)}$ is the predicted probability of default in the previous decision tree iteration.

The loss function, (4.3) uses a greedy split finding algorithm, whereby each splitting node is added iteratively in a way that best reduces the losses. Thus, the added tree, f_k, that most reduces the above loss function, (4.3), is found node by node, until further splitting fails to significantly improve the loss reduction. Overall, the optimal added tree can be found using the *gain* of the model, where for each new decision node, the feature and feature value that leads to the highest gain is chosen. As the regularization parameter λ increases, the value of the gain is reduced. Thus, λ ensures that in order to add complexity to the model, there has to be a sufficient increase in gain to justify adding more nodes. Another aspect of the decision tree creation is called *cover*. Cover ensures that features used for decision nodes affect enough observations. If too few observations are considered for a splitting node, then the split will not happen. Both gain and cover are embedded feature importance measures in the XGBoost system. The average gain of a feature reflects how much loss reduction is on average caused

each time a given feature is used in a decision node. Similarly, the average cover reflects the average number of observations affected by a given feature each time it is used.

4.2 Logistic Regression

The benchmark technique used to compare against the XGBoost in this thesis will be a LR. The function $\hat{v}(\mathbf{x}^t) \in \mathbb{R}^M$ is the predicted log-odds of bankruptcy for each company on the basis of the input features for each year, $t = 1, ..., T$. $\hat{v}(\mathbf{x}^t)$ is defined linearly by the set of input features weighted by their individual coefficients, defined as

$$\hat{v}(\mathbf{x}^t) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m^t \tag{4.4}$$

where $\beta_0 \in \mathbb{R}$ is the intercept in the absence of all other input features and $\beta_m \in \mathbb{R}$ are the slope coefficients or weights of each individual input feature. Because these values are still in the form of log-odds, we need to map them onto the probability space which can be done using the following logistic function[3]:

$$p(\hat{v}(\mathbf{x}^t)) = \frac{1}{1 + e^{-\hat{v}(\mathbf{x}^t)}} \tag{4.5}$$

The LR and the XGBoost both use the logistic loss function (4.3) for optimal classification. However, the XGBoost is optimized with respect to the prediction from the previous decision tree in a boosting process ($\hat{y}_n^{(k-1)}$). On the other hand the LR is optimized using a Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS; see Gao and Reynolds, 2006) optimization algorithm, which searches for the coefficient weights, β_m, that minimizes the loss function. It fits the model based on some initial guesses, and then updates it consecutively according to the direction that most reduces the losses across the entire dataset in a gradient descent manner. Once the model parameters are optimized, the LR model is tested on a validation test, similarly to the XGBoost.

A LR is inherently limited by its rigid structure, and strict assumptions. These assumptions include linearity between feature inputs and the log-odds of the output, no multicollinearity,

[3] As mentioned, the XGBoost (and also the SHAP which will be discussed shortly) are also expressed in the form of log-odds. We therefore note that the logistic function can also be applied to their respective outputs to transform the output into probabilities. One simply needs to exchange $\hat{v}(\mathbf{x}^t)$ with $\hat{y}_n$.

and independence of errors (Stoltzfus, 2011). Such assumptions can limit the model in the face of non-linear relationships, and in general make the model less useful in modelling complex patterns. In addition, the logistic regression works poorly when faced with extreme outliers, missing values, and too many input features (Stoltzfus, 2011). Therefore, we employ a range of techniques in order to process and adapt our dataset to best accommodate for these assumptions, which are explained in the data preparation section.

4.3 SHAPley (SHAP) Additive Explanations

As mentioned, in order to effectively understand what factors of a company leads to banruptcy, we employ XAI techniques. Specifically, we are going to use the SHAP framework developed by S. M. Lundberg et al. (2019). SHAP is an XAI method based on game theory which can be applied to interpret the results of any ML model (M. S. Lundberg and Lee, 2017). The benefit of the SHAP is that it is both global in that it can explain the feature attributes of entire models and local in that it can provide explanations for single predictions. This is beneficial for credit risk modelling because it enables the interpretation of the model output for individual companies, and as such makes providing explanations for credit risk evaluations possible. A specific version of the SHAP, called the Tree SHAP was specifically developed to provide these explanations when used together with tree ensemble models such as the XGBoost (S. M. Lundberg et al., 2019). The Tree SHAP applies the Shapley Values developed in game theory to find a unique solution to a proposed model developed as a unification of previous additive feature attribution methods (AFAM) [4].

AFAMs interpret ML model outputs, such as those found in an XGBoost, as a sum of weighted importance values attributed to each input feature. They can be viewed as explanation models that interpret complex models through the use of approximation. For instance it takes the prediction output from the XGBoost and decomposes it into a weighted combination of the individual input features. For any given feature, m, a given importance value or SHAP value, ϕ_m^n, is attributed to represent the influence that feature has on the model output for that company year. The SHAP model output, $g(z')$ which is the log-odds of the predicted bankruptcy probability for that company year, n, is defined as

[4] These are LIME, DeepLIFT, and Layer-Wise Relevance Propogation.

$$g(z') = \phi_0 + \sum_{m=1}^{M} \phi_m^n z'_m \tag{4.6}$$

where z' is a set of 0's and 1's corresponding to the the same set of feature inputs used in the original XGBoost. The individual z'_m are binary classification values representing whether a given feature is present or missing, taking the value 1 if present and 0 if missing. ϕ_0 is a base value attributed to the prediction in the case where all features are missing from the model, and can be viewed as the same initial prediction as in the XGBoost, which is often just the average probability of default in the dataset. Figure 4.1 shows an example of how the additive process works to achieve a final predictive value for a given company.

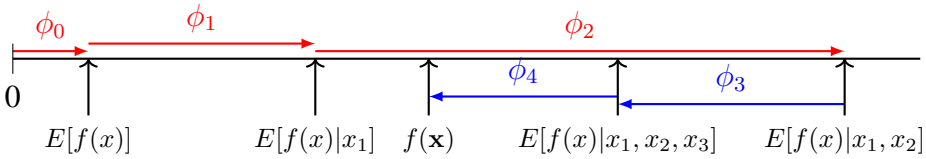


Figure 4.1: This is a local interpretation figure which showcases how a prediction for a single company is calculated. Here, we see that the first two features of the company increase the PD, and the last two features decrease the PD. The final output is the expected PD conditional on all four features, given by $f(\mathbf{x})$. The biggest contributor to this result is ϕ_2, since it has the highest SHAP value.

Since not all companies will have available data for all features, it is important that features that are not present are not given an importance value for those companies. This is a critical property of the unique solution found using the SHAP, called *missingness*. The other two important properties are *local accuracy* and *consistency*. Local accuracy is a property which states that the sum of all present importance values should be equal to the ML model output. This just means that $g(z') = \hat{y}_n$, essentially stating that when attributing importance to all features, the underlying model should be explained fully. Lastly, consistency simply states that the attributed value assigned to a feature can never decrease if a model is changed such that that feature has a bigger impact on the model output. M. S. Lundberg and Lee (2017) argue that the SHAP is the first AFAM that adheres to all three properties, which are individually found in earlier AFAMs.

Specifically, SHAP values are calculated iteratively, whereby the original model is fitted for every possible subset of feature inputs. Then, the average predicted value of the model using all subsets not containing the feature of interest is compared to the predicted value of the model using all features, which is just the original model output. By comparing these predictions,

the SHAP values will reveal how important a given feature is for the overall model prediction. Specifically, we calculate the following equation:

$$\phi_m^n = \sum_{S \subseteq \mathcal{M} \setminus \{m\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{m\} - f_x(S)] \tag{4.7}$$

where $f_x(S) = f(h_x(z')) = E[f(x)|x_S]$ is defined as the expectation of the model output conditional on the set of feature inputs S. S is a subset of feature inputs which excludes the feature of interest, m, in each iteration: $S \subseteq \mathcal{M} \setminus \{m\}$, where $\mathcal{M}$ is the set of all feature inputs. The mapping function h_x is responsible for mapping the set of 0's and 1's found in z' onto the decision tree such that it reflects which features are missing and which are present in each subset. Thus the expected predicted output, $f(x)$, is conditional on the subset of features present, x_S. The impact of a given feature can now be found by comparing $f_x(S \cup \{m\})$ with $f_x(S)$, which are explanation model predictions with and without the feature of interest, respectively. Additionally, $|S|$ is the number of features in the subset S, and M is still the total number of features in the model.

In addition to local SHAP values which are used to identify important feature inputs for individual companies, the SHAP framework also has global SHAP values. Global SHAP values are calculated by averaging the absolute value of all local SHAP values across the entire dataset. By taking the absolute value, importance is assigned to both negative and positive SHAP values, which ensures that both features which are important for predicting bankruptcy and for predicting solvency are considered. Global SHAP values, Φ_m, are defined as

$$\Phi_m = \frac{1}{N} \sum_{n=1}^{N} |\phi_m^{(n)}| \tag{4.8}$$

Global SHAP values have often been used in the literature as a way of increasing the model performance of black-box ML models (Nguyen et al., 2023; Alfaro et al., 2008). In this study, the aim will be to use global SHAP values in conjunction with local SHAP values to both increase interpretability of individual predictions and to better visualize the overall model per-

formance.

5 Data

This section describes the process of data collection, provides some descriptive statistics, and details the data preparation needed for the analysis. Data collection shows how the different raw data was collected and specifically what features we focus on. Bankrupt and solvent companies will have separate descriptive statistics, the purpose of which is to gain insight into the differences between the two categories. Data preparation involves detailing how missing values and outliers are dealt with. Specifically how the financial ratios are used to create new features is described in feature engineering. Lastly, how we ended up splitting the dataset for the training and validation as well as the process of selecting features for the XGBoost and the LR will be discussed.

5.1 Data Collection

In order to acquire a fitting dataset to be used in further analysis, there are several steps that need to be completed. Firstly, this involves the identification of the relevant companies we are interested in predicting bankruptcy for. Secondly, it involves gathering financial statements, macro data, and sector specific data in order to create relevant financial ratios and features. The variables created from this data collection are to be used in our models as feature inputs.

5.1.1 Financial Variables

Our dataset subject to analysis are financial statements of commercial real estate (CRE) firms characterized as small medium enterprises. The European Union's definition of SMEs is measured by two factors: (1) Staff headcount or (2) Turnover or balance sheet total (Directorate-General for Internal Market and SMEs, n.d.). We filter our companies by balance sheet total as it better represents the scale of firms within CRE. Thus, we include only enterprises that have a balance sheet total between €2m and €43m. We include enterprises that reach this balance sheet total criteria at least once during our time period. The time period of interest for our analysis is 2012-2022, with annual frequency on the data, and the proprietary data is downloaded from Enin AS. We construct a longitudinal panel data set using bankrupt and solvent companies. Initially, we sort the data by their respective organization number as identification (id)

and accounting year (year). To identify commercial real estate companies we filter by NACE codes. We sort by the industrial area code "L - real estate activities". And we only include the industrial codes 68.100 and 68.209 (BRREG, n.d.), which ensures that there are no real estate management firms, real estate brokers and house cooperatives.

ENIN enables us to filter out bankrupt and solvent companies by using "corporate flags". We use the flags "Registered bankrupt" and "Bankruptcy defendant". By only using these two flags, companies that ceased to exist due to other events than bankruptcy are avoided; such as demerger and merger & acquisitions activities. We then set our time period, and download the organization numbers that have met this criteria. After this download, a company batch is uploaded to ENINs Application Programming Interface (API). We do the same process for our solvent companies. Now, by using the API we are able to download the income statement and balance sheet of our chosen CRE companies, as well as their geographical location and establishment date.

The data set consists of 14,371 unique CRE companies. Of these companies, 678 registered as bankrupt during the time period 2012-2022. The last 13,693 are considered solvent, which implies that the dataset is highly imbalanced, with a defaulting frequency of only 4.8%. Because we are interested in analysing default probabilities of potential investments, we have already excluded all subsidiaries in the final dataset. Enin's corporate flags enables us to remove subsidiaries by filtering out companies with the flag "corporate group". In general, debt financing is done through non operational holding companies and not subsidiary companies that are created for individual CRE projects. By removing subsidiaries we keep the analysis focused on investments where there is potential risk for the creditor.

5.1.2 Macro and Sector Specific Variables

In addition to financial variables, we include macro and sector specific variables that we believe are relevant in predicting bankruptcy. For example, a rapid increase in interest rates will increase financial expenses, potentially leading to struggles for highly leveraged firms. A stock index consisting of real estate companies may reflect the sentiment of the real estate market as a whole, and could therefore improve our model. The variables were collected from various sources, and then transformed to represent change over time. The aim of these transformations was to acquire dynamic measures of the macro economy and the CRE sector, as previous studies

have found such features to be of interest (Ciampi et al., 2021). The implementation of these variables enables a deeper analysis of how changes in the economy as a whole can impact firms depending on their individual financial situation. More about this under feature engineering.

Below we list the different sources we collected macro economic and sector specific variables from:

- Annual transaction volume (MNOK) for commercial real estate - Akershus Eiendom AS (Eiendom, n.d.).
- NOK 5 year swap rate 6month (NOKAB6O5Y) - Eikon Financial Database.
- Annual electricity price (NOKøre/kWh) for Norwegian price regions - (Nord Pool, 2024).
- VINX Real estate Index NOK GI (SE0004388676) - NASDAQ (Nasdaq, n.d.).
- Unemployment rate - SSB (Statistisk sentralbyrå (SSB), 2024).
- GDP growth for Norway - Trading Economics (Economics, 2024).

These specific variables where chosen as either measures of the sentiment in the economy as a whole or as a measure of the sentiment in the CRE sector. A change in transaction volume represents a change in liquidity, and should affect the risk premium required by CSPs, and subsequently be reflected in the default probability. Furthermore, an increase in transaction volume can be associated with a higher willingness to invest in CRE. The interest rate swap is a 5-year contract where one party pays a fixed interest rate, determined by market conditions and outlook, and the other party pays a floating rate tied to the 6-month NIBOR. The swap rate represents the forward-looking view of the economy as a whole. We use the 5-year swap rate because CRE companies typically use contracts with this rate as a tool to hedge interest rate costs. Utilities are some of the largest variable costs for CRE properties, and electricity price for each price region in Norway is therefore included. We use the business address to identify the county the CRE company operates in, then we map the county to the corresponding electricity price region.

The VINX is a real estate price index containing some of the most traded real estate firms in the Nordics, and similarly to the transaction volume should provide interesting insights into

our bankruptcy prediction study[1]. The unemployment rate is included to take into account that an increase in unemployment may result in a reduced occupancy rate for CRE properties, thus reducing revenue. Lastly, the annual GDP growth rate of Norway may be interpreted as a proxy for the business cycle, and a decline in the GDP growth rate may reflect an increase in the number of defaulting companies.

5.2 Data Description

Although, the features mentioned above were included in our initial model, not all features proved to be relevant for the model we used in the final analysis. Out of 189 features, only 18 features have been selected using SHAP feature selection, which will be discussed in the feature selection section. In order to better understand how the final features differ between bankrupt and solvent companies we will first discuss some descriptive statistics of our data. The descriptive statistics are calculated from the entire dataset, consisting of the years 2012-2022. We show separate statistics for bankrupt and solvent companies in order to visualize fundamental differences between the two categories. In comparing bankrupt and solvent CRE companies, we will highlight three different financial dimensions that are typically discussed in the literature (E. I. Altman et al., 2015): profitability, liquidity, and solvency.

Table 5.1 shows a range of descriptive statistics for the selected features for all companies. Features that are financial statement items are denoted in units, trend variables in percentage and financial ratios are denoted in decimals. We can clearly see that there are large outliers on several of the financial variables. Very large values of ratios like Capital Employed / Fixed Assets (CE/FA), Interest Coverage Ratio (ICR) and Operating Income / Total Assets (OI/TA) are examples of this. As mentioned previously, we we would like to keep these outliers for the XGBoost model as it is able to specifically categorize the companies using their missing values. Missing values may be a sign of poor book-keeping which tends to be a sign of financial distress in small companies.

5.2.1 Profitability

Profitability concerns how effectively a company is at generating revenue using its existing assets. For CRE companies this can be reflected through how well they generate rental income

[1]The VINX contains mostly Swedish firms with one Danish and one Islandic, but no Norwegian firms. However, we believe it still represents an overall sentiment of the CRE market in the Nordics as whole, and therefore also the Norwegian market.

Table 5.1: Descriptive Statistics of Bankrupt and Solvent Companies

(a) Descriptive Statistics of Bankrupt Companies

Feature	mean	std	min	25%	50%	75%	max
Equity	763546.18	17167135.44	-301831985.00	-240318.50	144445.00	1215483.00	264030583.00
Capital Employed/Fixed Assets	-816.12	251161.28	-3181707.00	0.74	0.99	1.19	9514944.00
SWAP 36M MA	1.91	0.59	1.39	1.44	1.65	2.31	3.23
Loss Buffer	14401.00	162538.92	-1396659.50	-0.05	0.04	1.39	999990.00
Retention Ratio Trend 1Y	630.08	39979.93	-30434.78	0.00	0.00	0.00	2544444.44
Operating Income/Total Assets	-63.18	4384.08	-289507.60	0.02	0.12	0.76	3677.55
Retained Earnings	-447183.34	16683998.36	-312144985.00	-733931.00	-12000.00	363420.50	263846993.00
Total Assets	10117928.79	25124781.73	-7601000.00	708866.00	3549511.00	10175470.00	547452000.00
Working Capital/TEBIT	30.69	3997.32	-47626.77	-2.06	-0.26	1.08	244472.80
Total Assets Trend 2Y	38974.98	1116574.40	-12137660.00	-26.33	-0.53	42.20	41287029.41
Transferred Equity/Total Assets	0.00	12.51	-548.79	-0.09	-0.00	0.05	357.90
Interest Coverage Ratio	3028.65	183107.29	-4815532.00	-1.24	1.00	5.77	6209574.50
Current Ratio	8.10	168.01	-7229.85	0.20	0.90	2.36	4854.17
Electricity Price	41.81	38.65	9.35	27.34	33.63	48.25	264.97
Operating Income Trend 1Y	240.23	2128.44	-5273.42	-33.19	1.34	50.00	62314.87
Total Liabilities/Total Assets	2503.74	61952.18	-17047.00	0.63	0.91	1.10	1802583.50
Retained Earnings Trend 1Y	2385.09	176384.39	-2010194.00	-43.79	1.75	46.23	9740925.00
Loan To Value Trend 1Y	14506344.58	730426184.67	-3377.51	-9.62	0.00	8.17	38566392925.21

(b) Descriptive Statistics of Solvent Companies

Feature	mean	std	min	25%	50%	75%	max
Equity	16407388.39	36448300.96	-77617502.00	1871413.50	5353148.50	14269092.50	582963217.00
Capital Employed/Fixed Assets	5381.48	372439.38	-33.94	1.00	1.09	1.51	26365193.00
SWAP 36M MA	1.67	0.32	1.39	1.44	1.56	2.05	3.23
Loss Buffer	7739.27	167751.49	-999990.00	0.99	3.22	14.48	15052944.50
Retention Ratio Trend 1Y	-123.00	8961.32	-491356.04	0.00	0.00	0.00	178660.12
Operating Income/Total Assets	0.15	0.38	-0.10	0.03	0.09	0.16	13.48
Retained Earnings	12676629.89	31340422.97	-109527281.00	845603.50	3794651.00	11063264.00	582323217.00
Total Assets	30820565.35	54248772.96	55.00	7421166.75	12906777.50	28753892.50	853778885.00
Working Capital/TEBIT	-14.68	1451.11	-140267.63	-0.09	0.69	2.77	12368.79
Total Assets Trend 2Y	2773.88	182842.87	-99.88	-5.42	2.31	25.45	17568658.70
Transferred Equity /Total Assets	-0.16	14.14	-1239.55	0.00	0.03	0.08	4.06
Interest Coverage Ratio	26114.65	1049815.63	-1131074.67	1.62	4.68	43.43	90266034.00
Current Ratio	136.16	6507.74	-8212.35	0.99	2.88	9.54	627849.00
Electricity Price	79.43	81.10	9.35	30.31	48.38	94.62	264.97
Operating Income Trend 1Y	1861.75	129396.94	-29331.26	-6.08	2.40	16.79	12254900.00
Total Liabilities/Total Assets	1.21	42.26	-0.12	0.17	0.53	0.82	3132.53
Retained Earnings Trend 1Y	220.92	10591.47	-109973.74	-5.26	5.15	22.56	904085.19
Loan To Value Trend 1Y	422.79	10517.69	-293.83	-12.03	-2.82	3.16	719407.71

from their properties, typically denoted as operating income to total assets (OI/TA). In our descriptive data shown in table 5.1 we see that the median OI/TA is higher for bankrupt companies compared to solvent companies (0.12 versus 0.09). Whilst at first glance this suggests that bankrupt companies are more profitable, this interpretation can be misleading. The distribution of OI/TA for bankrupt companies is more variable, with a lower extreme minimum (-289 507) and higher extreme maximum (3 677) values, resulting in a high standard deviation of 4 384. In contrast, solvent companies have a more consistent distribution, with a minimum value of -0.10, a maximum value of 13.48, and a standard deviation of 0.38. These descriptive statistics indicate that the apparent higher profitability of bankrupt companies is influenced by their wider range of values, and one could argue that trusting the median in this case is not necessarily the correct assumption.

5.2.2 Liquidity

Liquidity reflects the degree to which companies can meet their debt obligations in the short term. A common way of measuring liquidity is using the current ratio. We can see in table 5.1 that solvent companies generally are more liquid represented by their median Current Ratio of 2.88, which indicates that solvent companies in our dataset have nearly three times more current assets to cover its current liabilities. In contrast, for the bankrupt companies, the median is only 0.90. A current ratio of 0.90 suggests that the median bankrupt company does not have the capital on hand to cover its short-term debts.

5.2.3 Leverage & Solvency

Measures of leverage and solvency are generally used to get a better grasp at a company's overall financial stability, their capital structure, and their ability to repay debt in the long run. A typical way of measuring leverage is by calculating how much debt exists in the company, often measured through total liabilities to total assets (TL/TA). A median TL/TA ratio of 0.91 for bankrupt companies indicates that the median company in this group is highly leveraged, with liabilities nearly equaling its total assets. Such high leverage exposes these companies to significant interest rate risk. In contrast, solvent companies have a median TL/TA ratio of 0.53, meaning that only 53% of their total assets are financed by liabilities. This significantly lower leverage suggests that solvent companies are in a more stable financial position and are less vulnerable to interest rate fluctuations.

The interest coverage ratio (ICR) reflects how well a company's income covers its interest expenses, and is therefore often used to measure solvency. The ICR described in table 5.1 accounts for total operating income and income from group companies and subsidiaries. Solvent companies have a median ICR of 4.68, indicating that their operating income is nearly five times their interest expenses. In contrast, bankrupt companies have a median ICR of 1.0, indicating that their earnings before interest and taxes (EBIT) are just sufficient to cover their interest costs. An ICR of 1.0 signals financial distress, as these companies are barely managing to meet their interest obligations.

5.2.4 Feature Correlation

As described above, there are several interesting differences between solvent and bankrupt companies in our dataset. Despite these differences, when looking at individual accounts by them-

selves it would be difficult to determine whether a company is about to declare bankruptcy or not. As such, the aim of these descriptions is not to conclude, but simply to highlight some key differences between bankrupt and solvent companies in the CRE sector. The degree to which the variables presented in table 5.1 correlate with each other and with the output variable, *bankruptcy*, can be seen in figure 5.1.

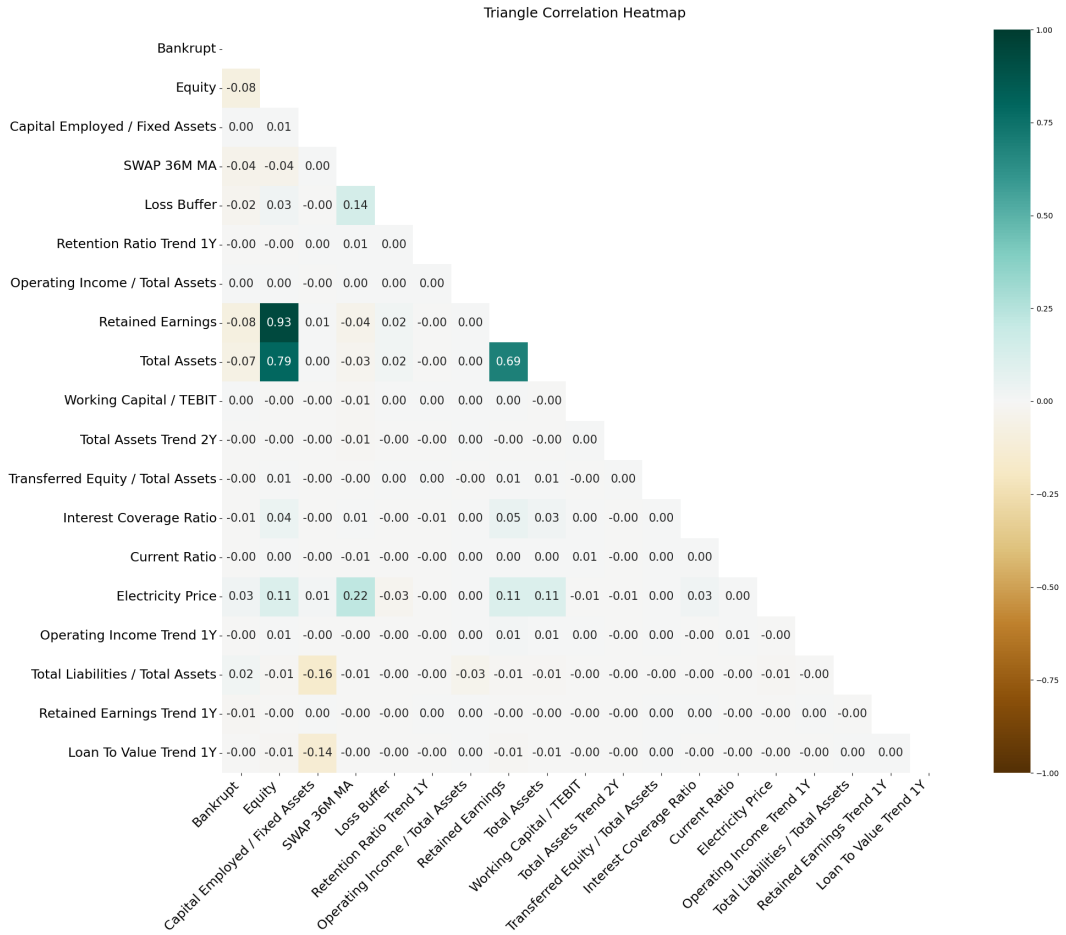


Figure 5.1: Heatmap of correlations between features and the target variable. The colour scale on the vertical axis to the right explains the colour coding. Blue indicates a high positive correlation, and brown indicates a high negative correlation. Most features are not correlated.

Figure 5.1 is a heatmap that displays the linear correlation both between the 18 selected input features and with bankruptcy. The intensity of the cell color represents the degree of correlation, as can be seen on the scale on the right. Ratios that share the same numerator or denominator will have high correlations, such as Retained Earnings (RE) and Equity (E), as well as RE and Total Assets (TA). The target variable "Bankrupt" does not have any significant linear correlation with the input features, indicating that there might be non-linear relationships present, which would support the use of non-linear models such as the XGBoost.

5.3 Data Preparation

Data preparation involves making sure that the dataset is adapted to the models used for analysis. For instance, for the LR, outliers and missing values are problematic, and we therefore need to handle them accordingly. Depending on which ML method is used, we need to make sure we have taken necessary precautions in order for the analysis to be reliable and accurate.

5.3.1 Data Cleaning

Outliers and missing values are present in the financial ratios that we have created. An outlier can be defined as a datapoint that is significantly different from the remaining data (Aggarwal, 2017). In the last year of activity it may be reasonable to expect that the financial reporting is deprioritized. Failure to maintain proper books and records is a common problem for distressed companies (Floyd et al., 2020). Some of the missing values and outliers may therefore be assumed to be arising from poor financial reporting. We need to be mindful when removing outliers and imputing missing values, as bankruptcy could be considered an outlier by itself since it is considered a rare event.

Handling of Missing Values

We need to decide if the missing values are Missing at Random (MaR), Missing Completely at Random (MCR) or Not Missing at Random (NMR). Since we believe that missing values may be a symptom of a company in financial distress, we therefore believe that the missing values are NMR. Missing values that are NMR can provide valuable information to our XGBoost model, as the model has the ability to handle missing values as a separate category and learn the optimal direction to assign for missing values during a split. We will therefore not impute the missing values in the dataset used for our XGBoost model. On the other hand, the LR model cannot use missing values. Therefore we will impute missing values for features in the LR model by using the feature median across all observations for the specific year they are observed in.

Handling of Outliers

The XGBoost itself is not very affected by outliers due to the use of decision trees as weak learners. The splitting nodes split the data based on absolute values, and do not care whether a given feature value is far away or close to the given split value. Since an outlier is a datapoint

far from other observations, this wont affect the split. We have therefore, chosen to keep the outliers in the models as a way of separating bankrupt from solvent companies.

Similar to missing values, the LR is sensitive to outliers, and cannot handle them very well. We therefore use winsorization to manage the outliers, and reduce their effect on the model prediction. Winsorization is a procedure where extreme values of a given feature are replaced by the values closest to them. We defined the boundary as the 1st and 99th percentile for small and large outliers respectively.

5.3.2 Feature Engineering

To be able to gain deeper insights into the relationships between the features and our target variable, as well as align the frequency of our data, we perform a procedure called feature engineering. Feature engineering is the process of transforming raw data to new features with the purpose of improving the predictive performance. Usually it is done by a researcher with domain expertise or through iterative trial and error (Nargesian et al., 2017).

From the financial statements we create a set of financial ratios. These financial ratios are divided into categories, the most relevant of which, as mentioned, are profitability, leverage and liquidity ratios. The initial feature set is partly based on previous studies on credit risk analysis (E. I. Altman, 1968), partly based on specific studies on SME default prediction (Ciampi et al., 2021), and lastly based on the findings that macro-economic and sector-specific data potentially lead to increase model prediction accuracy (Filipe et al., 2016). All initial features are presented in tables B.1 and B.2 in appendix B.

In addition to the financial ratios themselves, we also expand the financial ratio set by calculating trend variables. The trend variables are calculated by percentage change the last one, two and three years. For example, the percentage change last year in Total Assets (TA) is calculated by:

$$Total\ Assets\ Trend\ 1Y = \frac{(TA_t - TA_{t-1})}{TA_{t-1}} \tag{5.1}$$

The features we create from macro and sector specific variables have the purpose of exploring new relationships to see if they have any predictive power. From SWAP rates we calculate the quarterly return and create variables that represent a cumulative return for the last 12, 24 and

36 months. We also include 12, 24 & 36 months Moving Averages (MA) features for the swap. Our assumption is that after a significant rise in interest rates - evident by the 250 basis points rise from 2019 to 2022 - the financial risk remains relevant. This persistence is attributed to the time it takes for companies to fully experience the impact of the new interest rate levels, due to their mix of fixed and floating rate bank loans. Using VINX35 index's quarterly returns we create 12, 24 & 36 months of accumulated returns. We also create features, similarly to the SWAP, at the same intervals for a MA for the VINX35. The purpose of these features is to see whether the aggregated fluctuations can help represent the evolution of market values in the commercial real estate market. Lastly, transaction volume with annual frequency is transformed into change last year, change last two years and change last three years. We believe these transformations of transaction volume have the capability to reflect access to financing for CRE companies where an increase (decrease) in transaction volume may indicate loosening (tightening) of credit.

5.3.3 Splitting of Dataset

The standard procedure in supervised machine learning is to split the dataset into a training set and a test set. When performing this split it is important to reduce the information leakage between the training and test datasets. Information leakage occurs when the trained model is trained using information it is supposed to predict. If we simply were to randomly distribute observations into the training set and the test set, there would most likely be similar company years in both. This would lead to contamination of the training set for two reasons: (1) the training set has access to future information of companies included in the test set; and (2) the training set incorporates the same macroeconomic and sector-specific feature values as the test set, because these are identical across all companies for a given year. Such information leakage would lead to inflated model performance in both the test and validation datasets as the model is trained and validated on similar data.

In order to avoid such information leakage, we therefore apply a modified sliding window technique. Here we annually expand the training dataset and keep the test set the same length. As such, we both validate the model through annual iterations, and improve the model through increasingly informative training datasets. To ensure a similar ratio of bankrupt-to-solvent companies across all relevant years, we undersampled solvent companies to increase the ratio of bankrupt-to-solvent companies. Undersampling was specifically done as we obtained spurious evaluation results due to the overpresence of solvent companies in some cases

where the bankruptcy frequency was lower. We undersample such that we have a somewhat consistent ratio of bankrupt-to-solvent companies for each year. Table 5.2 displays the distribution of bankrupt and solvent companies across each validation window. We wish to see whether the later windows, where the model is trained on a larger set of companies leads to increased model performance.

Table 5.2: Unique Solvent and Bankrupt Companies by Year

	2018		2019		2020		2021		2022	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Bankrupt	138	113	251	91	342	92	434	98	532	146
Solvent	4760	1688	6448	1621	8069	1750	9819	1660	11479	2214
Balance	2.90%	6.69%	3.89%	5.61%	4.24%	5.26%	4.42%	5.90%	4.63%	6.59%

5.3.4 Feature Selection

After performing the feature engineering we will have a considerable amount of features. As the feature set gets bigger, there are several issues that can arise. More features implies a more complex model, which leads to concerns regarding the curse of dimensionality. The curse of dimensionality refers to the multiple problems that can arise when working with many features, such as overfitting, multicollinearity, and data sparsity (N. Altman and Krzywinski, 2018). Additionally, more features leads to more computationally complex models that take longer to finalize. We therefore wish to find the optimal set of features regarding both performance and computational time. Some of the initial features may be redundant and others may not provide much predictive power. Feature selection is thus performed as removing redundant and unnecessary variables can improve interpretability, shorten the learning time, simplify modelling as well as provide a model that is better at generalizing patterns (Guyon and Elisseeff, 2003).

XGBoost

For the XGBoost, we perform a feature selection by training the XGBoost model on the full feature set during the time period 2012-2021. SHAP values are then calculated and ranked from highest to lowest mean SHAP value. Based on these SHAP values we rank the features from most important to least important. Using the 30 most important variables, we start at the feature with the least impact, and then recursively remove one feature at a time. After each iteration we train the model and assess the performance on the test set, 2022 using the ROC-AUC score. After the performance, measured by the ROC-AUC score, drops below 95% of

the initial performance the iterations stop. The variable set with the highest ROC-AUC score is then selected to be the final feature set.

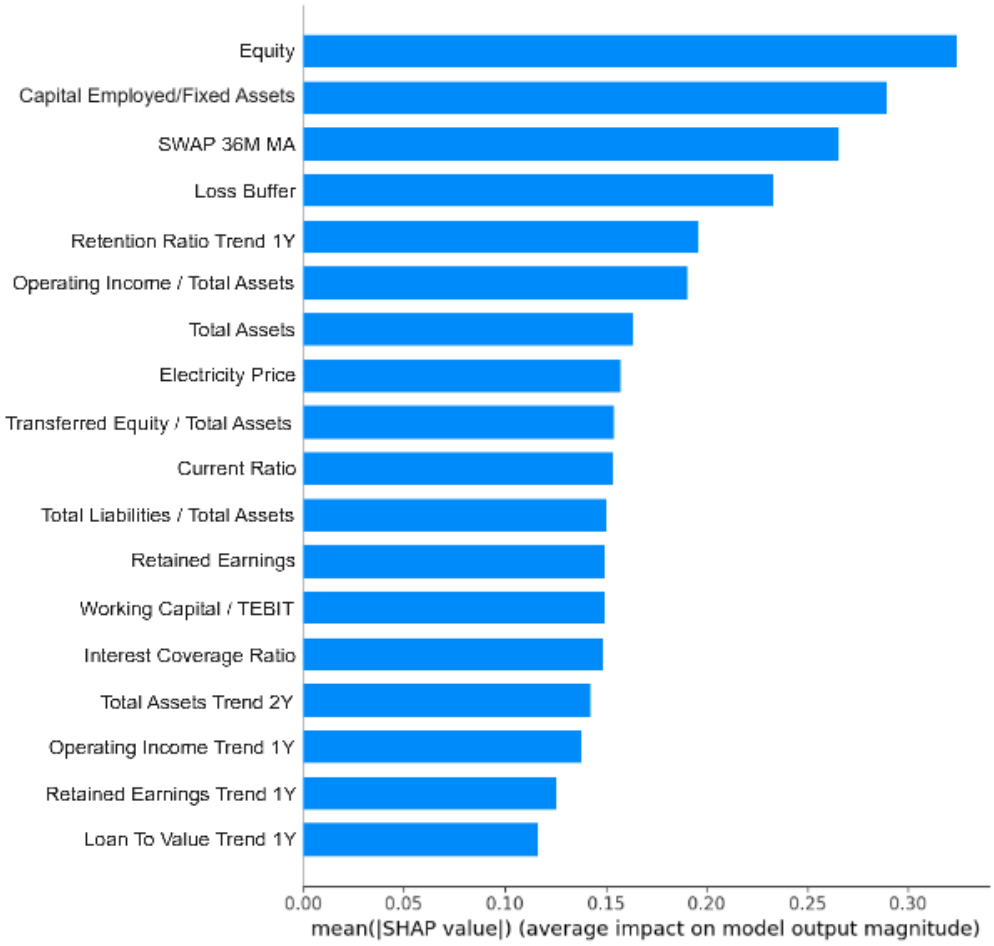


Figure 5.2: Displays the 18 most important features for default prediction of SMEs in the CRE sector according to our XGBoost model. The features are ranked according to their mean absolute SHAP values, from highest to lowest, where the mean absolute SHAP value is in log-odds. Feature names are on the left axis.

In figure 5.2 we see the 18 most important features for the XGBoost model according to the features' average absolute SHAP value. These 18 features represent the variables that have the biggest impact on the model prediction. They were the final 18 features found using the iterative process mentioned above.

Logistic Regression

To make the LR model more parsimonious, we start with reducing the number of features. The initial feature set consists of the 18 best features from the XGBoost with SHAP feature selection. Continuing the feature selection process we remove features that are highly correlated in order

to reduce the bias caused by multicollinearity. This is done by calculating the variance inflation factor (VIF), where variables that score above 10 are removed.

After the VIF procedure, we rank the features based on the mean absolute SHAP values found previously. Recursively we eliminate the features that have the least impact. Again, features are eliminated until the ROC-AUC score declines below 95% of the initial score. The features that are present once this optimization has been completed will be included in the final LR model. We are left with four features that yield the best ROC-AUC score. These features are: *Equity, SWAP 36M MA, OI/TA and ICR*.

Once the optimal feature sets are found for the XGBoost and LR we fit a third model to be evaluated in the analysis. This will be an XGBoost trained and tested using the LR feature set, namely the features *Equity, SWAP 36M MA , OI/TA and ICR*, henceforth called the XGBoost (LR features). We wish to see how the XGBoost (LR features) performs under similar constraints to a LR. The aim of this XGBoost (LR features) will be to compare results with the XGBoost (18 features), and to visualize how much the added complexity gained from 14 more features aids in model performance.

6 Analysis and Results

In this section, we present and discuss the results obtained. The performance of our three models will be evaluated using the metrics presented in the evaluation metrics section. In order to fit the data as best as possible to our models, we implement hyperparameter tuning for optimal performance. Hyperparameter tuning is done for each window in the rolling window analysis, both for the XGBoost (18 features) and the XGBoost (LR features)[1]. Subsequently, two types of analysis' are done. Firstly, we use rolling windows with the purpose of evaluating how the three models perform at various time periods and sample sizes. Secondly, we will evaluate the XGBoost (18 features) model trained from 2012-2021 and tested in 2022 using the SHAP framework. SHAP will be used to explain our XGBoost model's feature importance globally by assessing the average absolute SHAP value of each respective feature. Additionally, we will evaluate the same model locally by sampling two companies from the test set, one bankrupt and one solvent. Local SHAP explanations are then used on the test set, similarly to the global

[1]Refer to Appendix C for specifics on the optimal hyperparameter values across all rolling windows for both the XGBoost (18 features) and the XGBoost (LR features).

SHAP explanations. Usually, the SHAP explanations are used on the training set to be able to interpret what features are the most important in the trained model. Since we have companies observed over time, we want to isolate one year so that we are able to make the most accurate interpretation as possible. We therefore chose the test set, as we want to evaluate and explain the model performance one year ahead, similarly to how its done in practice in banks.

6.1 Rolling Window Performance

To evaluate the models we use both ROC-AUC and PR-AUC over the set of rolling windows found in table 5.2. From figure 6.1 we see that the performance of the models increase as the data for the training set increases. According to the ROC curve, XGBoost (18 features) outperforms the LR and XGBoost (LR features) in all years. The ROC-curve evaluates overall model performance without a strong sensitivity to class imbalance, as it measures the trade-off between TPR (recall) and the FPR across different thresholds. The XGBoost (18 features) is better at detecting true positives with little compromise to how good it is able to avoid false positives. The performance of LR and XGBoost (LR features) decline in 2022. In 2022 geopolitical tensions rose which led to soaring energy prices that put pressure on businesses. Energy prices are represented by electricity prices in the XGBoost (18 features) model. This feature has the ability to capture the increased level of operating costs in CRE. Since the parsimonious models only have four features, non of which are electricity prices, it seems they lack the ability to fully capture the macroeconomic trends seen in 2022, which may be the cause of their declining performance in the 2022 test set compare to previous test sets.

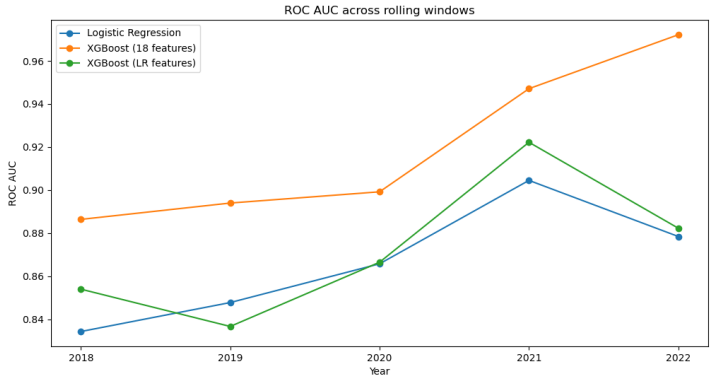


Figure 6.1: Line plot of ROC-AUC scores for test sets from 2018 to 2022. Each line represents one model evaluated using rolling windows. Year is plotted on the x-axis with AUC score on the y-axis

In figure 6.2 we see the same trend as observed in figure 6.1, increasing the amount of

training data increases the performance of the models. Using the ROC-AUC one could come to the conclusion that the XGBoost (18 features) is superior in all years. However, based on the PR-AUC we see that from 2018-2020, all models perform similarly. The PR curve measures the tradeoff between precision and recall at different thresholds, which indicates that even though the XGBoost (18 features) is superior in general, shown by the ROC-AUC, for the specific purpose of identifying the positive class, it is equal to the other two models. In 2021 the parsimonious model, XGBoost (LR features), actually outperforms the more flexible XGBoost model in terms of balancing precision and recall. In 2022, however, the XGBoost (18 features) excels and is superior compared to the other models with a PR-AUC score of 0.74, compared to 0.44 for the XGBoost (LR features) and 0.34 for the LR model, which again showcases that the more flexible model with more features is better at adapting to large economic changes. This flexibility is especially crucial for a model aimed at predicting default among CRE companies that, as mentioned, can cause large losses to banks during financial crisis.

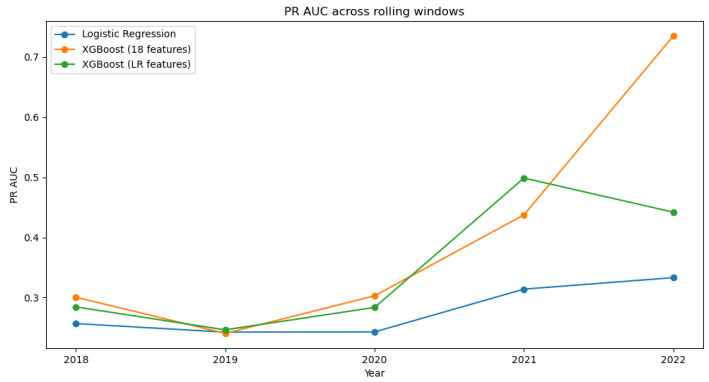


Figure 6.2: Line plot of PR-AUC scores for test sets 2018-2022. Each line represents one model evaluated using rolling windows.

Figure 6.3 depicts the ROC and PR curve for the three models using the year 2022 as the test set, and the years 2012-2021 as the training set. The charts in the figure provides a more detailed summary of how the different models perform across all thresholds. The ROC curve, displayed in the left chart of figure 6.3 shows that the XGBoost (18 features) outperforms the other two models at all thresholds, with an ROC-AUC score of 0.97. A random model has a ROC-AUC score of 0.5, indicated by the red line. Such a case implies that bankrupt and solvent companies are simply categorized at random. The ROC-AUC score is biased due to the heavy imbalance between positive and negative instances in the train and test set. Therefore it may be misleading to interpret the performance of our model solely based on this chart. In addition,

assessing the PR curve may provide a more nuanced picture.

The PR curve displayed in the right chart of figure 6.3 shows that XGBoost (18 features) still outperforms the LR and XGBoost (LR features) model. The XGBoost (18 features) has a PR-AUC score of 0.74 compared to 0.44 for the XGBoost (LR features) and 0.33 for the LR. We observe that the XGBoost (18 features) is convincingly better at identifying bankrupt companies (recall) while still maintaining an ability to classify positive cases correctly. All models are significantly better than a random guesser, represented by the red line.

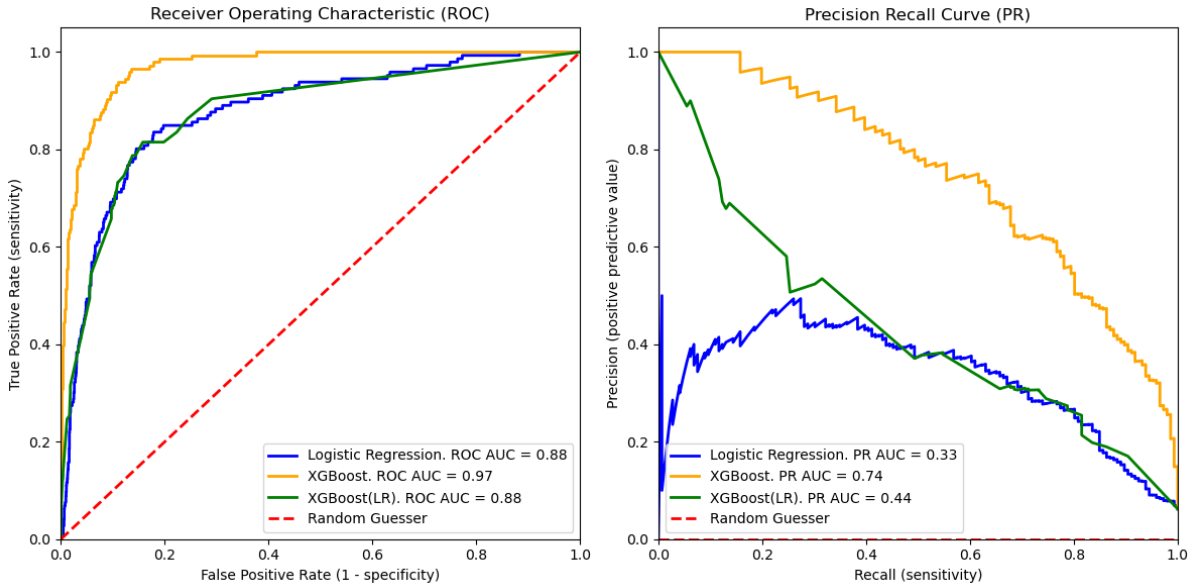


Figure 6.3: ROC- and PR-curve for the three models in the test year 2022.

Based on the PR curve seen in figure 6.3, we observe some unusual behavior from the LR model. Although, the specific reason for this unusual behavior is unknown, it could be due to problems regarding how well the data is fitted in the model. The limited number of macroeconomic features in the LR feature set may imply that the model fails at identifying crucial features associated with bankruptcy in 2022, not previously seen in the training set. At the same time, the same erratic behavior is not seen in the XGBoost (LR features), suggesting some relationships may be non-linear, something the LR cannot capture. These results again highlight the flexibility of the XGBoost model framework, able to use the same feature set as the LR to more precisely categorize bankrupt firms. In essence, there is less noise in the XGBoost (LR features) compared to the LR model, and we observe less false positives, displayed by the higher precision at most threshold levels in the PR curve.

In table 6.1 we can see how the recall, precision, and accuracy change as a function of

the defined thresholds. This shows how sensitive the models are regarding what classifies as a bankrupt company. What initially becomes obvious is the tradeoff between recall and precision. The lower the threshold, the more companies are classified as default, and we therefore observe a high recall. On the other hand, a higher thresholds means more companies are classified as solvent, which increases precision. Lastly, we also observe that accuracy is less reflective of the threshold levels. However, because we have a biased dataset towards solvent companies, a higher threshold tends to increase accuracy. We can observe partially why we found lower ROC-AUC and PR-AUC values in 2022, as both the XGBoost (LR features) and the LR struggle to correctly identify solvent companies even as thresholds increase, as reflected through their low precision ratios.

Table 6.1: Metrics for the three models using 2022 as the test set.

	Metric	XGBoost	XGBoost (LR)	Logistic Regression
Threshold 1.0%	*Recall*	*0.98*	*1.00*	*0.94*
	Precision	*0.26*	*0.06*	*0.10*
	Accuracy	*0.83*	*0.06*	*0.50*
Threshold 1.5%	*Recall*	*0.93*	*1.00*	*0.91*
	Precision	*0.36*	*0.06*	*0.13*
	Accuracy	*0.90*	*0.06*	*0.63*
Threshold 2.0%	*Recall*	*0.86*	*1.00*	*0.86*
	Precision	*0.47*	*0.06*	*0.17*
	Accuracy	*0.93*	*0.06*	*0.74*
Threshold 2.5%	*Recall*	*0.79*	*1.00*	*0.80*
	Precision	*0.55*	*0.06*	*0.26*
	Accuracy	*0.95*	*0.06*	*0.85*
Threshold 3.0%	*Recall*	*0.71*	*1.00*	*0.52*
	Precision	*0.62*	*0.06*	*0.38*
	Accuracy	*0.96*	*0.06*	*0.92*
Threshold 4.0%	*Recall*	*0.62*	*1.00*	*0.30*
	Precision	*0.75*	*0.06*	*0.44*
	Accuracy	*0.96*	*0.06*	*0.93*
Threshold 5.0%	*Recall*	*0.49*	*0.86*	*0.16*
	Precision	*0.80*	*0.19*	*0.40*
	Accuracy	*0.96*	*0.76*	*0.93*
Threshold 6.0%	*Recall*	*0.40*	*0.71*	*0.12*
	Precision	*0.85*	*0.31*	*0.40*
	Accuracy	*0.96*	*0.88*	*0.93*

6.2 SHAP Explanations

In this section we employ the SHAP explanations to interpret our models both at the global and local levels. Variable importance plots, bee swarm plots and decision plots are used to interpret the model globally, and waterfall plots are used to interpret the models locally. All

the figures made using the SHAP library are based on the test set in 2022 using the XGBoost with 18 features. Keep in mind that the SHAP method when used for post-hoc interpretation of binary classification models, expresses its output as log-odds. Log-odds is used to reflect the incremental impact of each feature on the overall likelihood of default.

6.2.1 Global Explanations

Variable Importance Plot

Figure 6.4 displays the average marginal contribution each feature has on the predicted likelihood of default. Compared to the feature selection, the order of features has changed as we have retrained the model with less features. We note that the 36 month moving average of the SWAP rate, Capital Employed / Fixed Assets, Loss Buffer, Retained Earnings and the Current Ratio are the top five most important features. In the top five features in the test set, we observe only one macro economic variable, two financial ratios, and two balance sheet items. Moving outside top five there are several trend variables, most notable is Total Assets Trend 2Y, Loan To Value Trend 1Y and Retained Earnings Trend 1Y.

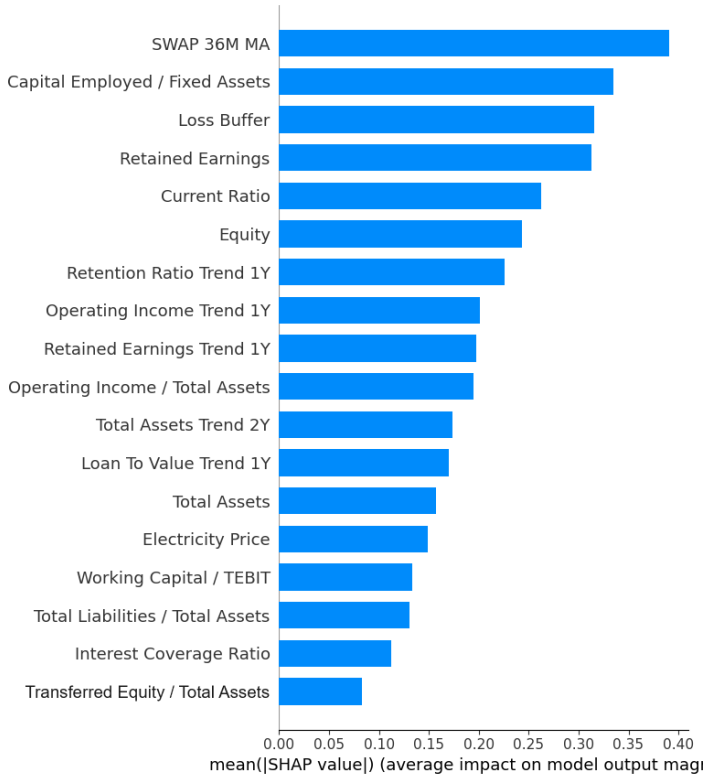


Figure 6.4: Variable importance plot displaying absolute average SHAP values for the most important features. Feature are displayed to the left of each bar. Feature importance are sorted in descending order. X-axis is displayed using the mean of absolute SHAP values in log-odds.

Summary Plot

Figure 6.5 plots the distribution of SHAP values, Φ_m, for each feature over all companies, both bankrupt and solvent, in 2022. For a meaningful interpretation of this plot, it is desirable to have a separation in the distribution between blue and red dots, representing high and low feature values. Such a separation is most notably observed in features such as Total Liabilities / Total Assets (TL/TA) and Operating Income / Total assets (OI/TA). In both cases, we see that high feature values is associated with a larger SHAP value, seen on the x-axis. Thus, increased leverage, seen through TL/TA, and increased profitability, seen through OI/TA leads to increased an increased PD in our model. Although the latter seems counter intuitive, it is reflective of the higher median OI/TA of bankrupt companies compared to solvent companies observed in our dataset. Subsequently, it could reflect a decrease in asset value as opposed to an increase in operational income, which is typical of bankrupt companies whose assets are liquidated.

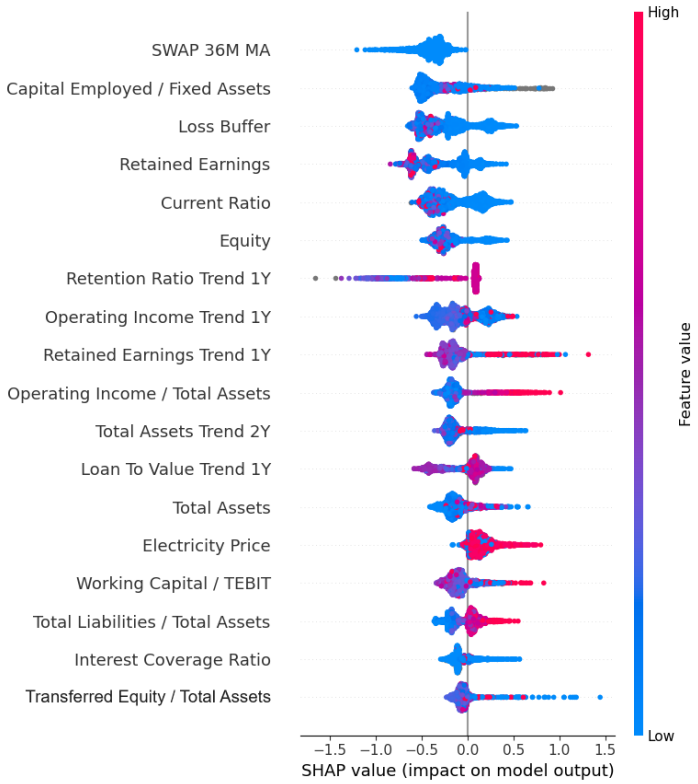


Figure 6.5: Each dot in the figure represents an observation. The observations are color-graded where blue (red) observations represent a low (high) feature value. Observations are distributed along the x-axis depending on the SHAP value. Positive (negative) SHAP values are indicative of an increase (decrease) in probability of default. SHAP values are measured in log-odds.

There are several features, however that are more difficult to interpret by themselves using

figure 6.5. These features include the Loan to Value Trend 2Y (LTV) and Working Capital / Total Assets (WC/TA). Both features have red and blue clusters of observations intertwined. Such clustering suggests that a high feature value can lead to a decreased PD for one company, whilst at the same time leading to an increased PD for another company. We also note that the swap always decreases the PD for all companies. An economy with increasing interest rates tends to be positive for CRE companies that focuses on the rental markets, and which benefit from higher rental income from their properties. Another interesting insight is the impact on PD that the different feature values of Electricity Price has. Electricity Price varies across the different price regions a company is situated in. High electricity price, illustrated by a red observation increases the PD in our model. The few observations of a company that is in a region with low prices sees a low to zero impact on the model output. Given the relatively high electricity prices observed in 2022 compared to previous years, such an observation seems intuitive.

The feature Capital Employed / Fixed Assets provides some valueable information about the XGBoost's ability to model missing values. Missing values, displayed by grey observations, seem to increase the probability of default for this feature .

Decision Plot

In order to visualize how each feature affects bankrupt and solvent companies differently, we separate bankrupt and solvent companies and interpret them separately. In figure 6.6 we randomly sample ten actual defaulted companies from the test set. The threshold for classification is represented by the grey line, which is also the intercept of our model, where all the predictions start from (ϕ_0 as seen in (4.6)). Starting from the bottom, as we add features we can see how the different features change the PD (here represented by log-odds on the x-axis). The threshold is around -3.3 in log-odds (3.55% when converted to PD), and we can see from figure 6.6 that there are two misclassifications.

From figure 6.6 we can see which specific features contribute to the misclassification. For the left most misclassified observation, hereby denoted MC1, it seems like the model struggles to use the features to clearly decide if MC1 is bankrupt or not. By tracking the line we can observe that the one year trend of retention ratio and loan to value increase the PD, whilst the equity seems to decrease the PD again back to the initial prediction. We see that the company's

Figure 6.6: Decision plot shows how different SHAP values of a feature change the predicted log-odds of default. Each plotted line explains a single prediction. All the observation lines start from the intercept on the x-axis. The intercept is the models bias, which is the expected value. The figure should be read from the first feature row and up. As we accumulate SHAP values by adding a feature, the log-odds of default of a company will increase or decrease depending on the features' SHAP value. We convert the log-odds to probability using the logistic function.

ability to withstand losses, indicated by the feature value of the loss buffer, significantly decreases the log-odds and thus also the company's PD. When the 36month moving average of the swap is also taken into account, the company is eventually misclassified. The final log-odds of this misclassification equals a 1.6% PD. It is worth noting that depending on the sample gathered from bankrupt and solvent companies, the order of the features with the most importance will vary.

Figure 6.7 displays the decision plot for solvent companies, where there is one misclassification, hereby referred to as MC2. MC2's cause of misclassification seems to be the features Electricity Price, Retained Earnings trend one year and Equity. Further we see that especially the ratio of OI/TA increases PD. Using this plot in conjunction with the Bee Swarm plot, we can interpret that MC2 most likely has a high ratio of OI/TA. As mentioned, normally a higher ratio of OI/TA would imply higher operating efficiency, but in the context of a firm in distress it could be that the value of TA is very low, causing a high ratio. Should we observe a situation like this in practice, we would examine the books of the company itself to see what could be the cause behind the high OI/TA.

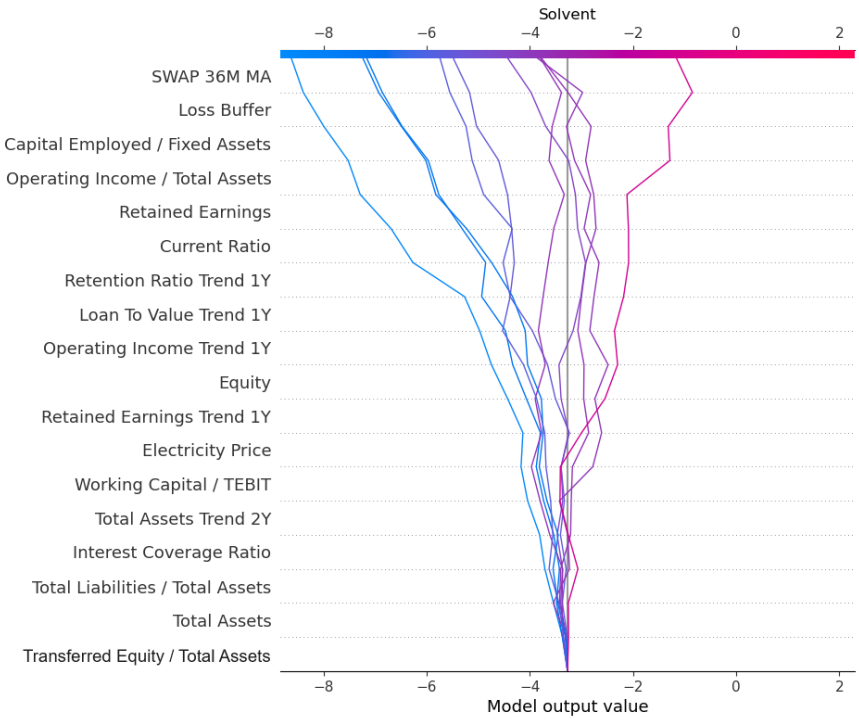


Figure 6.7: Decision plot shows how different SHAP values of a feature change the log-odds. Each plotted line explains a single prediction. All the observation lines start from the intercept on the x-axis. The intercept is the models bias, the expected value. The figure should be read from the first feature row and up. As we accumulate SHAP values by adding a features, the log-odds of a company will increase or decrease depending on the features SHAP value.

6.2.2 Local Explanations

SHAP enables us to interpret a prediction of a single company. This is called a local explanation. Local explanations can be utilized to explain a single defaulting company and see which features are the most important in contributing to the probability of default. Conversely it could also be used to determine which facets of a company are the most solid, contributing the most to a company's solvency.

Waterfall Plots

Figure 6.8 shows the waterfall plot for a bankrupt company. The plots display how each feature either increases the PD or decreases the PD, represented by red and blue bars respectively. The feature value contributing the most to classifying this company as bankrupt is Transferred Equity / Total Assets. The feature value of -0.403 for the bankrupt company suggests that in this particular year (2022), a substantial loss is transferred to equity. We can say that the company experiences cash flow insolvency. The second variable contributing the most to the default of our company is Total Liabilities / Total Assets. A feature value of 1.595 shows that

the company has more liabilities than it has assets, indicating balance sheet insolvency, which a negative equity further confirms. These two ratios, together with a reduction of 44% in assets over the last two years further affirms the conclusion that the company should be classified as default.

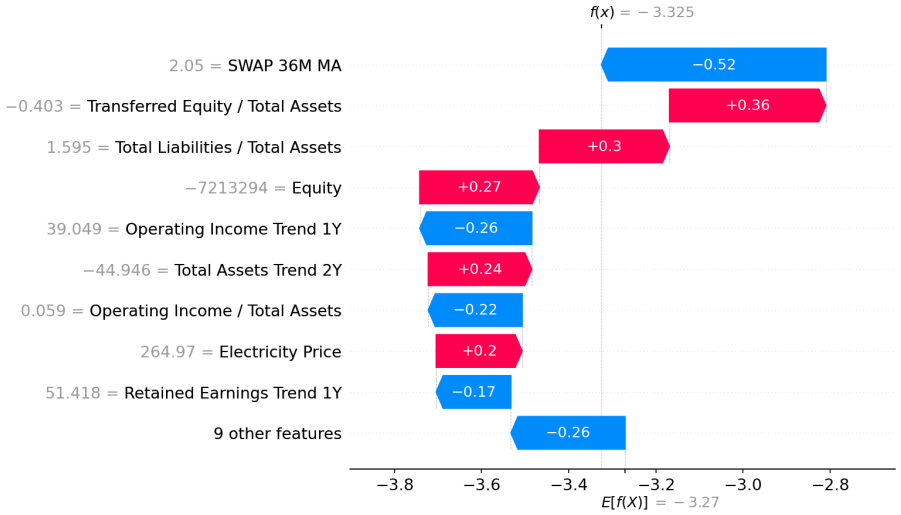


Figure 6.8: Local SHAP values for a bankrupt company. The intercept is denoted by E[f(X)], which is the expected log-odds in our test set. f(x) is the output prediction, displayed in log-odds. Output must be converted to obtain probability of default. X-axis is displaying SHAP values in log-odds. Features names with their corresponding value is listed to the left of their respective bar.

Despite the aforementioned financial items, we see that several features decrease the company's predicted PD. Most notably the swap rate, but also OI/TA and their retained earnings trend the last year. The degree to which the OI/TA can be trusted seems to be questionable however. Whether or not the company ends up being classified as bankrupt eventually depends on the defined classification threshold.

Figure 6.9 displays a waterfall plot for a solvent company. A reduction of 3.6% in loan to value shown by the feature Loan To Value Trend 1Y is reducing the log-odds the most for our company. This reduction of LTV in the last year tells us that a decrease in the leverage of the firm reduces the PD. A ratio of 1.001 for Capital Employed / Fixed Assets indicates that the company is managing its properties well and using them effectively to generate returns. Together with a high level of retained earnings the company appears to be financially solid. On the other hand, a low current ratio of 1 indicates that they may be vulnerable in the liquidity perspective. One misleading result from the plot is that a loss buffer of 0.576 increases the log-odds, and hence the PD. A loss buffer of 0.576 indicates that the company can have a loss of 57.6% before the equity is lost, assuming that last years operating income is the same next

year.

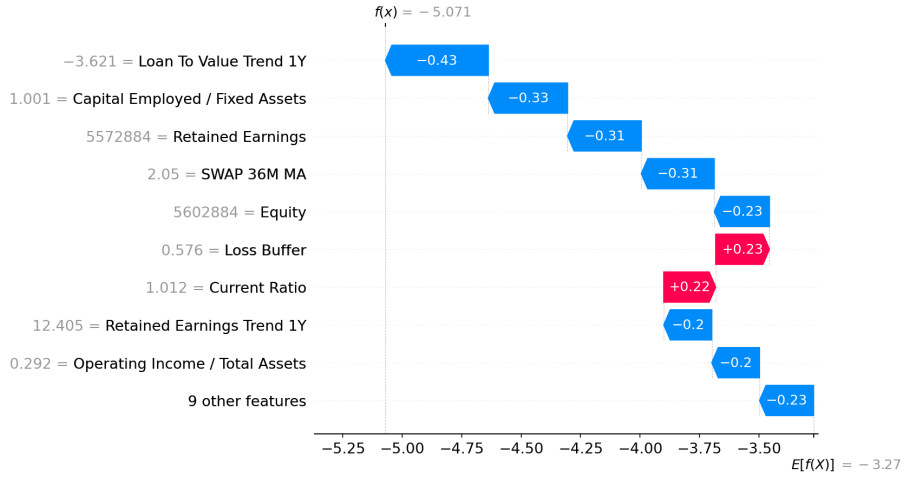


Figure 6.9: Local SHAP values for a solvent company. The intercept is denoted by E[f(X)], which is the expected log-odds in our test set. f(x) is the output prediction, displayed in log-odds. Output must be converted to obtain probability of default. Features names with their corresponding value is listed to the left of their respective bar.

We can note that, whilst our developed model for the most part correctly identifies companies as either solvent or bankrupt, there are crucial flaws. This fact can be seen with the aforementioned loss buffer and its misleading effect seen in figure 6.9. It can also be seen that whilst SHAP can decompose a predicted PD into its features, it cannot interpret how changes in financial ratios are expressed. For instance, regarding our musings concerning the OI/TA and its counter intuitive effects, the SHAP cannot enlighten us as to what financial item causes the effect. It is therefore important to look at the accounts themselves to understand exactly why the SHAP regards a given feature as important for a specific company.

7 Conclusion

In this thesis we present the argument that sector-specific default prediction models need to be the next step taken within credit risk analysis. In consideration of this argument, we presented the following two research problems:

1. Develop an extreme gradient boost ensemble-based machine learning model for predicting the probability of default in Norwegian small and medium sized commercial real estate companies at a one year forward time horizon.

2. Implement explainable artificial intelligence techniques by using Shapley additive explanations to interpret the probabilities of default estimated through the extreme gradient

boost model.

Commercial real estate (CRE) was chosen as it is a sector that banks are heavily exposed towards (SSB, n.d.-a). This exposure has historically caused both large losses for banks and large financial instabilities in the economy as a whole during financial crisis (Hagen et al., 2018). Therefore, as a first step in the implementation of ML in sector-specific PD prediction, we decided to focus on uncovering the associated risks with investing in CRE companies.

In order to develop an XGBoost with high predictive performance, we used SHAP feature selection to identify the 18 most important features. Then, we trained and tested the model in a modified rolling window from 2012-2022. The XGBoost model was benchmarked against a logistic regression (LR) that implemented recursive feature elimination to reduce the number of features further from 18 to four. This reduction in features for the LR was done to avoid potential issues regarding high dimensionality, typical to linear models. We lastly also modeled an XGBoost with the same feature set as the LR, to test how the two models performed when constrained to the same few features.

We evaluated the performance of our models using common classification evaluation metrics including ROC-AUC and PR-AUC. These evaluation metrics were applied to all test sets in a rolling window fashion, from 2018 to 2022. Based on the ROC-AUC, the XGBoost (18 features) outperformed the other two models during each annual window, which indicates that the XGBoost (18 features) more often correctly identified defaults, and also less often misclassified solvent companies as bankrupt. Based on the PR-AUC, however, all three models performed somewhat equally from 2018-2021. Then, in 2022 the XGBoost (18 features) again outperformed the other two models. The reason for this increase in performance from the XGBoost (18 features) is believed to come from its higher complexity which better captures the macro economic turmoil witnessed in 2022.

We furthermore used SHAP in order to interpret the results observed from the XGBoost (18 features) when trained from 2012-2021, and tested in 2022. SHAP was applied to the testing dataset in order to decompose the most important input features for the final predictions. Both global and local SHAP measure were implemented. Global measures allow us to get a better grasp at the average importance of features across all tested companies. We implemented the local measures for one randomly classified solvent company and one randomly classified

bankrupt company to visualize the individual reasons for their respective classifications.

At the global level, the interest rate, modeled as a 36 month moving average swap proved to be the most important feature. When the model was tested in 2022, this feature exclusively decreased PD, most probably due to the effect increasing interest rates have on CRE companies that primarily operate with renting out properties. In addition to the swap, the financial ratios *capital employed to fixed assets*, *loss buffer*, *retained earnings* and the *current ratio* proved to be important. Specifically for identifying defaulting CRE companies in 2022, we observed that high feature values for the features *Total Liabilities to Total Assets, Operating Income to Total Assets* and *Electricity Price* most clearly contributed to increasing a firms PD.

At the local level, our findings related to the most important features are less relevant, as we chose companies at random. The implementation of local feature importance was to showcase how the SHAP framework can not only be implemented at a global level as seen in many previous studies (e.g., Nguyen et al., 2023; Perboli and Arabnezhad, 2021) but also at the individual level to viably interpret feature importance. The aim of this part of the analysis was to demonstrate how the SHAP can be used to interpret results in accordance with the regulations presented by the European Banking Authority (European Banking Authority, 2023; European Banking Authority, 2022). Furthermore, we aimed to visualize how sector-specific features can influence the predicted PD even at individual company levels.

7.1 Future Directions and Data Limitations

We realize that the model we developed for this thesis is lacking in some aspects that could be further improved in future iterations. For instance, the implementation of qualitative data on companies including information on the board of directors, accounting flags, and management capabilities could be added. As mentioned during the background, these variables have been noted to be of interest in the credit risk analysis of SMEs in previous studies (e.g., Ciampi et al., 2021; Filipe et al., 2016). We, however lacked a good source for the implementation of these variables, and therefore chose to disregard them for our model. We do believe that they could increase the models' performance in the future however. We wisely chose to focus on the mother company within a corporate group as they are of the highest relevance for CSPs. However, should company group structure be modeled, it could provide relevant insight from a credit risk analysis perspective.

Furthermore, we noticed some aspects of our results that are counter intuitive to our understanding of financial markets. For instance, some ratios such as the OI/TA have no clear interpretation, and the effects on PD displayed in the summary plot was unusual for some features. Should another similar model be developed, there should be stricter regulations as to what features to include. Additionally, the LR model struggled with unusual feature relationships more than initially thought, which may have undermined some of the results presented.

In order to further build upon the model presented in this thesis, there are two additional aspects that could be incorporated. These aspects include loss given default (LGD) and misclassification costs. Our model is independent of scale, however from a creditors perspective the relevance of correctly modelling credit risk increases the bigger the potential investment. As such identifying the LGD of potential investments is crucial. In addition, employing misclassification costs for false positives and false negatives can make defining a threshold for bankruptcy classification easier. For instance, for some companies and some investments, the alternative cost of not investing in solvent companies could be larger than the LGD of investing in bankrupt companies.

Bibliography

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2). https://doi.org/10.3390/risks6020038

Adler, B. E. (1997). A theory of corporate insolvency. *NYUL Rev.*, *72*, 343. https://www.nyulawreview.org/wp-content/uploads/2018/08/NYULawReview-72-2-Adler.pdf

Aggarwal, C. C. (2017). An introduction to outlier analysis. In C. C. Aggarwal (Ed.), *Outlier analysis* (pp. 1–34). Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3_1

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems*, *45*(1), 110–122. https://doi.org/https://doi.org/10.1016/j.dss.2007.12.002

Altinn. (2024). Konkurs i aksjeselskap [Accessed: 2024-05-24]. https://info.altinn.no/starte-og-drive/avvikling-sletting-og-konkurs/konkurs/konkurs-i-aksjeselskap/

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2978933

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2015). Financial and non-financial variables as long-horizon predictors of bankruptcy. https://doi.org/10.2139/ssrn.2669668

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. *Journal of International Financial Management; Accounting*, *28*(2), 131–171. https://doi.org/10.1111/jifm.12053

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2020). A race for long horizon bankruptcy prediction. *Applied Economics*, *52*(37), 4092–4111. https://doi.org/10.1080/00036846.2020.1730762

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for smes: Evidence from the u.s. market. *Abacus*, *43*(3), 332–357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

Andersen, B. P., Rundhaug, M., & de Lange, P. E. (2021). Utilizing structural models to evaluate probability of default for norwegian stock-based firms. In *Bidrag innen kundeverdi og marked* (pp. 129–158). Universitetsforlaget. https://doi.org/10.18261/9788215055596-2021-07

Angelkort, A., & Stuwe, A. (2011). *Basel iii and sme financing*. Friedrich-Ebert-Stiftung Zentrale Aufgaben.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Bank for International Settlements. (2017). High-level summary of basel iii reforms [Accessed: 2024-05-28]. https://www.bis.org/bcbs/publ/d424_hlsummary.pdf

Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? a comprehensive test. *Journal of Banking & Finance*, *40*, 432–442. https://doi.org/doi.org/10.1016/j.jbankfin.2013.12.013

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71–111. https://doi.org/doi.org/10.2307/2490171

Bjørland, C., Hjelseth, I. N., Mulelid, J. H., Solheim, H., & Vatne, B. H. (2022). *Næringseiendomsmarkedet - ikke lenger en "svart boks"* (*Staff Memo* No. 6/2022). Norges Bank.

Black, F., & Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance*, *31*(2), 351–367. https://doi.org/doi.org/10.1111/j.1540-6261.1976.tb01891.x

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. https://doi.org/10.1201/9781315139470

Bridges, J., Gregory, D., Nielsen, M., Pezzini, S., Radia, A., & Spaltro, M. (2014). The impact of capital requirements on bank lending. https://doi.org/10.2139/ssrn.2388773

BRREG. (n.d.). Industrial codes [Accessed: 21-05-2024]. https://www.brreg.no/en/business-2/industrial-codes/?nocache=1716109523936

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899–2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x

Chen, T., & Guestrin, C. (n.d.). R package - understanding your dataset with xgboost [Accessed: 21-05-2024]. https://xgboost.readthedocs.io/en/stable/R-package/discoverYourData.html

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable system. https://doi.org/10.1145/2939672.2939785

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of italian firms. *Journal of Business Research*, *68*(5), 1012–1025. https://doi.org/10.1016/j.jbusres.2014.10.003

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking sme default prediction: A systematic literature review and future perspectives. *Scientometrics*, *126*(3), 2141–2188. https://doi.org/10.1007/s11192-020-03856-0

Ciampi, F., & Gordini, N. (2013). Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of italian small enterprises. *Journal of Small Business Management*, *51*(1), 23–45. https://doi.org/10.1111/j.1540-627X.2012.00376.x

Das, S. R., Hanouna, P., & Sarin, A. (2009). Accounting-based versus market-based cross-sectional models of cds spreads. *Journal of Banking & Finance*, *33*(4), 719–730. https://doi.org/10.1016/j.jbankfin.2008.11.003

De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable ai for credit assessment in banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifiers systems* (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Directorate-General for Internal Market, E., Industry, & SMEs. (n.d.). Sme definition [Accessed 30-04-2024]. https://single-market-economy.ec.europa.eu/smes/sme-definition_en

Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., & Torres, J. F. (2018). Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, *11*(4). https://doi.org/10.3390/en11040949

Economics, T. (2024). Norway full year gdp growth [Accessed: 2024-06-03]. https://tradingeconomics.com/norway/full-year-gdp-growth

Efrim Boritz, J., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications*, *9*(4), 503–512. https://doi.org/10.1016/0957-4174(95)00020-8

Eiendom, A. (n.d.). Data - transaksjonsvolum [Accessed: 16-05-2024]. https://akershuseiendom.no/markedsinnsikt/data/transaksjonsmarked?sector=Transaksjonsvolum&subSector=Volum

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine learning in radiation oncology: Theory and applications* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

European Banking Authority. (2022). Discussion paper on machine learning for irb models [Accessed: 2024-05-19]. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf

European Banking Authority. (2023). Follow-up report on machine learning for irb models [Accessed: 2024-05-24]. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

European Banking Authority. (2024a). Interactive single rulebook [Accessed: 2024-05-28]. https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/12420

European Banking Authority. (2024b). Interactive single rulebook [Accessed: 2024-05-24]. https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/16023

European Banking Authority. (2024c). Interactive single rulebook [Accessed: 2024-05-28]. https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/12421

European Systemic Risk Board. (2015, December). Report on commercial real estate and financial stability [Accessed: 2024-05-24]. https://www.esrb.europa.eu/pub/pdf/other/2015-12-28_ESRB_report_on_commercial_real_estate_and_financial_stability.pdf

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

Floyd, J. J., Phillips, M. W., Shea, W. F., Lashway, R. W., & Manning, L. (2020). *Distressed company insights report* (Insights Report). Floyd Advisory. https://www.floydadvisory.com/content/uploads/2020/05/Floyd-Advisory_Distressed-Company-Insights-Report.pdf

Gao, G., & Reynolds, A. C. (2006). An improved implementation of the lbfgs algorithm for automatic history matching. *SPE Journal*, *11*(01), 5–17. https://doi.org/10.2118/90058-pa

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*, 1157–1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web

Hagen, M. (2016, March). Næringseiendom i Norge. https://www.norges-bank.no/contentassets/093fda53ce45407aba78d88a97243e10/aktuell_kommentar_6_2016.pdf?v=09032017123525

Hagen, M., Hjelseth, I. N., Solheim, H., & Vatne, B. H. (2018). *Bankenes utlån til næringseiendom - en kilde til systemrisiko?* (*Staff Memo* No. 11/2018). Norges Bank.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9*(1), 5–34. https://doi.org/10.1023/b:rast.0000013627.90884.b7

Hillier, D., Ross, S., Westerfield, R., Jaffe, J., & Jordan, B. (2019). *Corporate finance, 4e* [589]. McGraw Hill.

Jackson, T. H., & Scott, R. E. (1989). On the nature of bankruptcy: An essay on bankruptcy sharing and the creditors' bargain. *Virginia Law Review*, 155–204. https://scholarship.law.columbia.edu/faculty_scholarship/385/

Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, *44*(1-2), 3–34. https://doi.org/10.1111/jbfa.12218

Lacher, R. C., Coats, P. K., Sharma, S. C., & Fant, L. F. (1995). A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, *85*(1), 53–65. https://doi.org/10.1016/0377-2217(93)E0274-2

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lundberg, M. S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, (30). https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles. https://doi.org/10.48550/arXiv.1802.03888

Marek, P., & Stein, I. (2022). Basel iii and sme bank finance in germany. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4261450

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470. https://doi.org/10.2307/2978814

Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American economic review*, *48*(3), 261–297. https://www.jstor.org/stable/1809766

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. S. (2017). Learning feature engineering for classification. *Ijcai*, *17*, 2529–2535. https://doi.org/10.24963/ijcai.2017/352

Nasdaq. (n.d.). Instrument : Se0004388676 [Accessed: 09-05-2024]. https://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004388676

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*(51-62), 56. https://www.researchgate.net/publication/328146111_An_overview_of_the_supervised_machine_learning_methods

Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. *Proceedings of the Twenty-First International Conference on Machine Learning*, 78. https://doi.org/10.1145/1015330.1015435

Nguyen, H. H., Viviani, J.-L., & Ben Jabeur, S. (2023). Bankruptcy prediction using machine learning and shapley additive explanations. *Review of Quantitative Finance and Accounting*. https://doi.org/10.1007/s11156-023-01192-x

NHO. (2024). Tall og fakta om smb [Accessed: 07-05-2024]. https://www.nho.no/tema/sma-og-mellomstore-bedrifter/tall-og-fakta-om-smb/#part2

Nord Pool. (2024). Day-ahead prices [Accessed: 2024-06-11]. https://data.nordpoolgroup.com/auction/day-ahead/prices?deliveryDate=latest&currency=NOK&aggregation=Yearly&deliveryAreas=NO1,NO2,NO3,NO4,NO5

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

Onakoya, A. B., & Olotu, A. E. (2017). *International Journal of Economics and Financial Issues*, *7*(3), 706–712. https://dergipark.org.tr/en/pub/ijefi/issue/32021/354317

Paraschiv, F., Schmid, M., & Wahlstrøm, R. R. (2023). Bankruptcy prediction of privately held smes using feature selection methods. *Available at SSRN 3911490*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3911490

Perboli, G., & Arabnezhad, E. (2021). A machine learning-based dss for mid and long-term company crisis prediction. *Expert Systems with Applications*, *174*, 114758. https://doi.org/10.1016/j.eswa.2021.114758

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, *36*(3), 1092–1113. https://doi.org/10.1016/j.ijforecast.2019.11.005

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence* (Report). National Institute of Standards and Technology (U.S.) https://doi.org/10.6028/nist.ir.8312

Reisz, A. S., & Perlich, C. (2007). A market-based framework for bankruptcy prediction. *Journal of Financial Stability*, *3*(2), 85–131. https://doi.org/10.1016/j.jfs.2007.02.001

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Schaaf, N., Huber, M., & Maucher, J. (2019). Enhancing decision tree based interpretation of deep neural networks through l1-orthogonal regularization. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 42–49. https://doi.org/10.1109/ICMLA.2019.00016

Schalck, C., & Yankol-Schalck, M. (2021). Predicting french sme failures: New evidence from machine learning techniques. *Applied economics*, *53*(51), 5948–5963. https://doi.org/10.1080/00036846.2021.1934389

Scikit-learn. (2024). Sklearn.model$_s election.randomizedsearchcv$ [Accessed: 2024-06-16].

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

Son, H., Hyun, C., Phan, D., & Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, *138*, 112816. https://doi.org/10.1016/j.eswa.2019.07.033

SSB. (n.d.-a). 08116: Finansforetak. utlån, etter låntakernæring (mill. kr) 2009m05 - 2024m03 [Accessed: 07-05-2024]. https://www.ssb.no/statbank/table/08116

SSB. (n.d.-b). 14150: Føretak, etter næringshovedområde (sn2007), organisasjonsform og storleik (k) 2008 - 2021 [Accessed: 07-05-2024]. https://www.ssb.no/statbank/table/14150

Statistisk sentralbyrå (SSB). (2024). Labour force, employment, unemployment and man-weeks worked, by sex, age and type of adjustment. break adjusted figures 2006m01 - 2024m04 [Accessed: 2024-06-03]. https://www.ssb.no/en/statbank/table/13760/tableViewLayout1/

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, *80*, 146–166. https://doi.org/10.1016/j.inffus.2021.11.005

Uhrig-Homburg, M. (2005). Cash-flow shortage as an endogenous bankruptcy reason. *Journal of Banking & Finance*, *29*(6), 1509–1534. https://doi.org/10.1016/j.jbankfin.2004.06.026

Vassalou, M., & Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, *59*(2), 831–868. https://doi.org/10.1111/j.1540-6261.2004.00650.x

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361. https://doi.org/10.1016/j.eswa.2013.09.033

Warner, J. B. (1977). Bankruptcy costs: Some evidence. *The journal of Finance*, *32*(2), 337–347. https://doi.org/10.2307/2326766

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Appendix A: The XGBoost Decision Tree Calculation

This appendix demonstrates how the XGBoost uses decision trees and boosting to predict a probability of default for a given company. Given the example companies A-H and their corresponding feature values presented in table A.1, we are able to fit the decision tree seen in figure A.1.

Company	ICR	D/E	VA/TS	Bankrupt
A	0.8	4	0.2	1
B	0.5	1.8	0.4	1
C	2.2	0.5	1	0
D	1	2.1	0.6	0
E	1.5	1.2	1.2	0
F	2.5	0.7	1.5	0
G	0.7	3	0.5	1
H	1.2	6	0.8	1

Table A.1: Companies used as examples in *XGBoost Decision Tree Example* figure

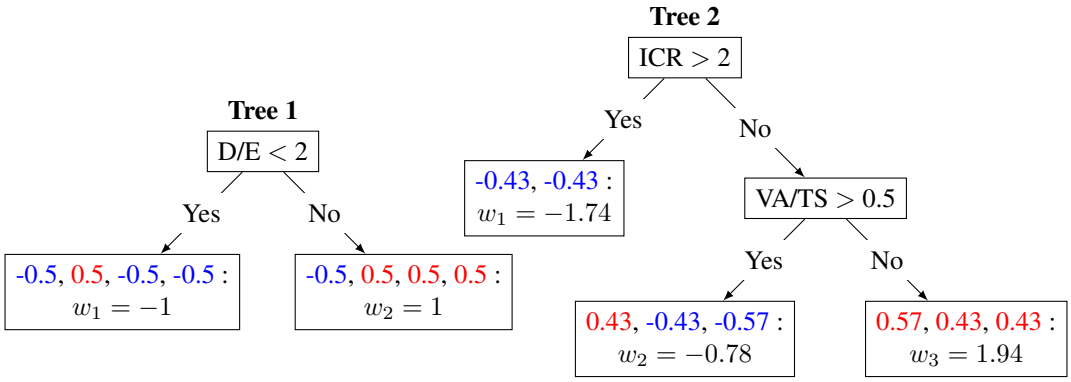


Figure A.1: Here D/E is the debt-to-equity ratio, ICR is the interest coverage ratio, and VA/TS is the asset turnover ratio. The leaf weights are calculated naively with $\lambda = 0$ and an initial prediction of 0.5. Since this is a binary classification model, they are in the form of log(odds), and when converted using the logistic function will represent probabilities. Blue output number values represent non-bankrupt company prediction errors, and red number values represent bankrupt company prediction errors. See appendix A for the entire calculation process. This is a simplification, and does not necessarily qualify as a valid tree in an XGBoost, but suffices for a visual explanation. See Chen and Guestrin, 2016 for specifics.

When optimizing equation 4.2 with respect to f_k, we get the following optimized leaf weights.

$$
\begin{aligned}
w_j^* &= -\frac{\sum_{n\in I_j} g_n}{\sum_{n\in I_j} h_n + \lambda} \\
where \quad g_n &= \partial_{\hat{y}^{(k-1)}} l(y_n, \hat{y}^{(k-1)}), \\
and \quad h_n &= \partial^2_{\hat{y}^{(k-1)}} l(y_n.\hat{y}^{(k-1)})
\end{aligned}
\tag{A.1}
$$

where g_n is the gradient of the objective function with respect to the previous prediction, and h_n is the corresponding hessian. These are found using second order Taylor approximation and is how the weights were found in figure A.1. Both the gradient and the hessian are iterated over the instance set I_j, which is the set of all companies contained within a leaf j. It now becomes obvious that defining $\lambda > 0$ will decrease the effect of individual observations on the entire prediction.

When applying equation A.1 to the logistic loss function 4.3, then this simply becomes:

$$
\frac{\sum_{i\in I_j} y_n - \hat{y}_n^{(k-1)}}{\sum_{i\in I_j} \hat{y}_n^{(k-1)}(1-\hat{y}_n^{(k-1)}) + \lambda} \tag{A.2}
$$

where the numerator is the sum of residuals, and the denominator is regarded as the summed binomial variance of the previous prediction, $\hat{y}_n^{(k-1)}$, over all instances in a given leaf plus the L2 regularization parameter. Equation (A.2) was applied when calculating the leaf weights in figure A.1. Following the calculation of the leaf weights, the logistic function (4.5) for $\hat{y}_n$ can be used to calculate the probabilities.

Using (A.2) we can calculate the leaf weights found in figure A.1. For instance for Tree 1, leaf 1, the residuals are $(-0.5, 0.5, -0.5, -0.5)$. Setting the initial guess to 0.5 since there are 50% bankrupt companies in the test dataset, and $\lambda = 0$, we get the following optimal leaf weight:

$$
w_1^* = \frac{-0.5 + 0.5 - 0.5 - 0.5}{4 * 0.5 * 0.5} = \frac{-1}{1} = -1 \tag{A.3}
$$

To find the probability that any firm in this leaf will go bankrupt we take the log-odds of the initial prediction, $log(\frac{0.5}{1-0.5})$. The reason is that we want to transform the initial guess, which is a probability, onto the real coordinate space since the leaf weights also exist here. So that when

we combine the initial guess with the leaf weights in the final prediction, they exist in the same space. In this case the log-odds of the initial guess are just 0, but it will change when working with imbalanced datasets which have more or less bankrupt companies. We add this to the leaf weights multiplied by some shrinkage factor, often very small. In figure A.1, the shrinkage factor is set to be 0.3, which is the default value seen in Chen and Guestrin (n.d.). In equation A.4 we see the predicted probability of bankruptcy for any company with a debt-to-equity ratio lower than 2 according to the first decision tree in figure A.1.

$$\frac{e^{0.3*-1}}{1+e^{0.3^{-1}}} = 0.43 \tag{A.4}$$

For the three solvent companies found in this leaf, we see that this predicted value is closer to the true value of 0. For the bankrupt company, however, this prediction is further from the true value of 1, meaning that this company has a higher chance of being misidentified as solvent now than after the initial guess. However, across the entire leaf, the average loss has been reduced, which is the aim of such a simple model.

Appendix B: Complete Feature Set

Here we present all base features used in the initial feature set and their corresponding description. These features were then used to create 189 total features using feature engineering.

Feature	**Description**
Age	Accounting year - Year established
County	Norwegian ”Fylke” as of 01.01.2024
Interest Coverage Ratio	(Operating Profit + Share of Profit Subsidiaries) / Interest expenses
Total Liabilities / Total Assets	Total liabilities/total assets
Current Ratio	Current assets / current liabilities
Return On Assets	Net income / Total Assets
Return On Equity	Net Income / Equity
EBIT / Financial Expenses	Operating Profit / Financial Expenses
Current Liabilities / Financial Expenses	Current liabilities / financial expenses
Total Assets	Total Assets
Operating Income	Operating Income
Working Capital / EBIT	(Current Assets - Current Liabilities) / Earning Before Interest & Taxes
Working Capital / TEBIT	(Current Assets - Current Liabilities)/ (Operating Profit + Share of Profit Subsidiaries)
Operating Income / Total Assets	Operating income / total assets
Net Working Capital / Total Assets	(Current Assets - Current Liabilities) / Total Assets
Equity / Total Liabilities	Total Equity / Total Liabilities
Equity / Total Assets	Total Equity / Total Assets
Quick Ratio	(Current Assets - Inventories)/Current Liabilities
Current Liabilities / Total Assets	Current Liabilities / Total Assets
Fixed Assets / Total Assets	Fixed Assets / Total Assets
ROCE	(Operating Profit + Share of Profit Subsidiaries) / (Fixed Assets + Net Working Capital)
Transferred Equity	Transferred Equity
Retained Earnings	Retained Earnings

Table B.1: Variable Descriptions (Part 1)

Feature	Description
Transferred Equity / Total Equity	Transferred Equity as Percent of Total Equity
Transferred Equity / Total Assets	Transferred Equity as Percent of Total Assets
Current Liabilities	Short Term Debt
Debt / Operating Income	Total Liabilities/ Operating Income
Debt / Total Income	Debt / (Total Operating Income + Share of Profit Subsidiaries)
EBITDA/EBIT	Earnings Before Interest Taxes Depreciation & Amortization / Earning Before Interest & Taxes
EBITDA/ Operating Income	Earnings Before Interest Taxes Depreciation & Amortization / Operating Income
Write Down of Assets / Total Assets	Write downs / Total assets
Capital Employed / Fixed Assets	(Fixed Assets + Current Assets - Current Liabilities) / Fixed Assets
Debt/EBITDA	Debt / Earnings Before Interest Taxes Depreciation & Amortization
EBITDA/ Current Liabilities	Earnings Before Interest Taxes Depreciation & Amortization / Current Liabilities
Operating Income / Total Assets	Total Operating Income / Total Assets
EBIT/ Operating Income	Earnings Before Interest & Taxes / Operating Income
Loan To Value	Total long term debt / Total fixed assets
Loss Buffer	Balance Sheet Loss Buffer. Measure of how much equity could be lost before insolvency
5 Year Swap	SWAP5YNOK6M
VINX35	35 largest real estate companies in the Nordics
GDP Growth NO	GDP Growth Norway
Inflation NO	Consumer Price Index NO SSB
Electricity Price	Electricity price for norwegian counties are mapped to their respective price region
Transaction Volume	Total annual transaction volume in NOK for all commercial real estate in Norway

Table B.2: Variable Descriptions (Part 2)

Appendix C: Hyperparameter Tuning

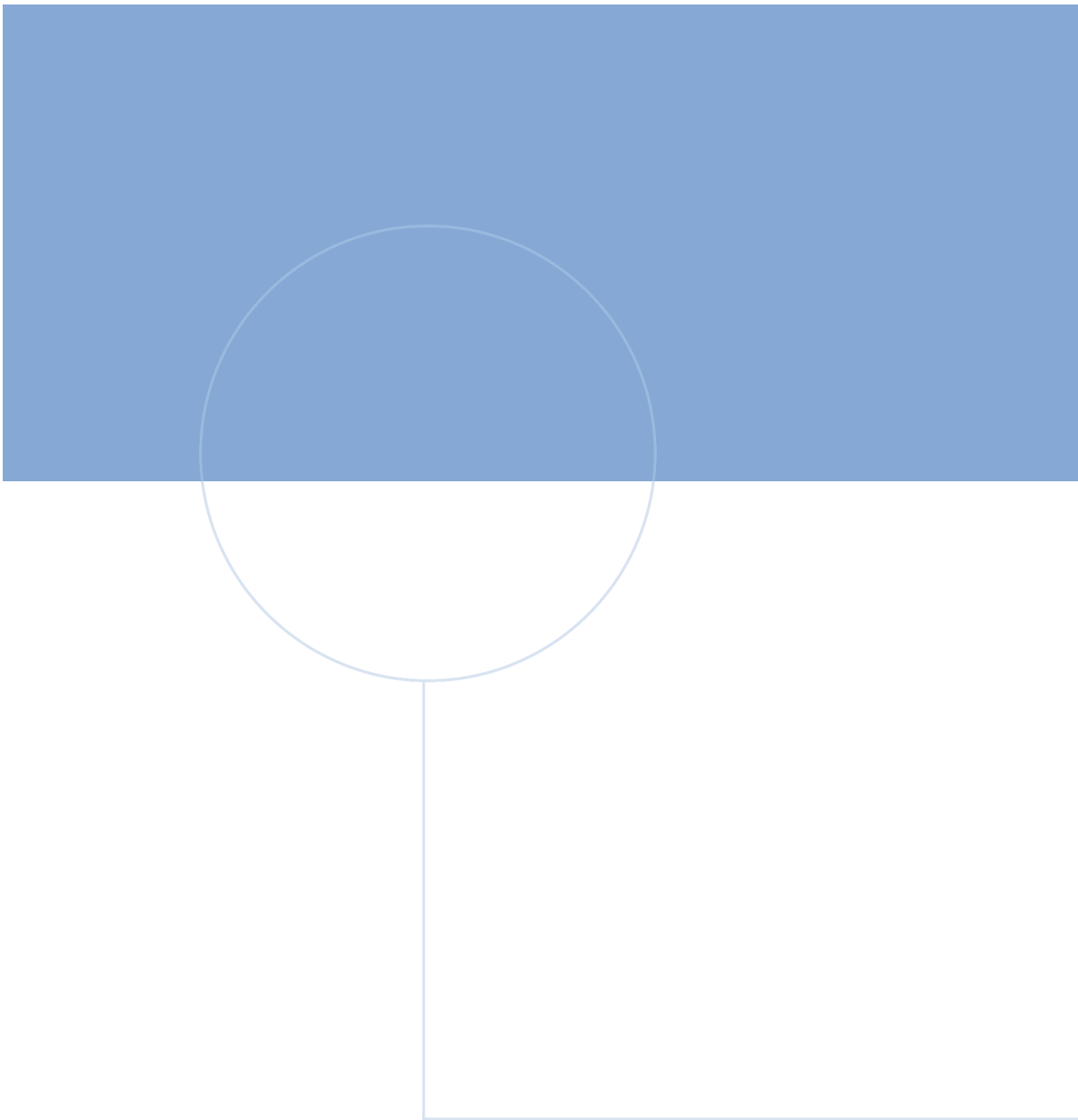
In this appendix we present the hyperparameters that were used in each rolling window. The hyperparameters are found using Randomized Search CV (see Scikit-learn, 2024). We refer to the documentation of the XGBoost for specifics on the description of each hyperparameter (see Chen and Guestrin, n.d.).

Table C.1: XGBoost (18 features) Hyperparameters

Feature	**2018**	**2019**	**2020**	**2021**	**2022**
learning_rate	0.0321	0.0878	0.0713	0.1792	0.0722
subsample	0.5836	0.8758	0.5373	0.7366	0.7628
max_depth	2	7	6	6	4
n_estimators	131	90	180	132	173
gamma	0.3449	0.2361	0.0044	0.0026	0.0851
reg_alpha	1.0238	1.9923	1.7302	0.8917	0.7348
reg_lambda	1	1	3	1	2
colsample_bytree	0.6529	0.9135	0.9655	0.8831	0.9508
min_child_weight	1	1	2	4	2

Table C.2: XGBoost (LR features) Hyperparameters

Feature	**2018**	**2019**	**2020**	**2021**	**2022**
subsample	0.75	1.0	0.5	1.0	0.75
reg_lambda	1	3	3	2	2
reg_alpha	0.5	1.0	0.5	1.5	0.0
n_estimators	200	200	50	200	200
min_child_weight	2	4	4	4	1
max_depth	1	2	1	1	1
learning_rate	0.2275	0.2275	0.0825	0.155	0.01
gamma	0.4	0.3	0.1	0.3	0.3
colsample_bytree	0.8	0.6	0.6	0.6	0.7



NTNU

Norwegian University of
Science and Technology