universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Explainable Machine Learning for Credit Risk Assessment: A Comparative Analysis of Model Performance and Interpretability

verfasst von / submitted by

Julia Ornatowski, B.A.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 974
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Banking and Finance
Betreut von / Supervisor:	ao. Univ.-Prof. Mag. Dr. Andrea Gaunersdorfer
Mitbetreut von / Co-Supervisor:	Dipl.-Inform. Dipl.-Volksw. Dr. Christian Westheide

Abstract

Trotz des Potenzials von maschinellen Lernmodellen im Finanz- und Wirtschaftsbereich stellt ihre undurchsichtige Natur ein Hindernis für ihre umfassende Umsetzung dar, insbesondere in stark regulierten Domänen. Ein Bereich, der das Potenzial Künstlicher Intelligenz (KI) daher noch nicht vollständig ausschöpft, ist die Kreditrisikobewertung. KI kann aufgrund ihrer Fähigkeit, große Datenmengen zu verarbeiten und komplexe Muster zu erkennen, präzisere Vorhersagen ermöglichen und das Risikomanagement verbessern. Gleichzeitig ist es entscheidend, transparente und faire Entscheidungsprozesse sicherzustellen. Fortschritte im Bereich der "erklärbaren KI" (XAI), also der Entwicklung von maschinellen Lernmodellen, die klare und verständliche Erklärungen für ihre Entscheidungen liefern, eröffnen neue Möglichkeiten, Transparenz zu schaffen und die Faktoren, die zur Entscheidungsfindung beitragen, zu verstehen.

Diese Arbeit untersucht die Anwendung von maschinellen Lernmodellen in der Kreditrisikobewertung und analysiert den Kompromiss zwischen Modellgenauigkeit und Erklärbarkeit. Insbesondere wird untersucht, ob einfachere Modelle komplexe Modelle in der Vorhersage von Insolvenzen kleiner und mittelständischer deutscher Unternehmen übertreffen können. Im theoretischen Teil liegt der Fokus auf der Bedeutung eines verantwortungsvollen Technologieeinsatzes in der Kreditrisikobewertung, insbesondere im Kontext der erklärbaren KI. Im empirischen Teil werden verschiedene maschinelle Lernmodelle hinsichtlich ihrer Vorhersagegenauigkeit und ihrer Fähigkeit, den Einfluss von Eingangsvariablen auf die vorhergesagten Ergebnisse aufzuzeigen, verglichen. Die Ergebnisse deuten darauf hin, dass neuronale Netzwerke und Random-Forest-Modelle in der Vorhersage von Kreditausfallrisiken andere Modelle übertreffen. Allerdings liefern die angewendeten Erklärungstechniken keine eindeutigen Ergebnisse darüber, welche Variablen maßgeblich für die Vorhersagen der Black-Box-Modelle verantwortlich sind. Daher kann es in bestimmten Anwendungsfällen sinnvoll sein, einfachere Modelle wie Entscheidungsbäume zu wählen, die klare und interpretierbare Ergebnisse liefern, selbst wenn dies mit einer geringfügig geringeren Vorhersageleistung einhergeht.

Abstract

Despite the many potential applications of machine learning models in finance and economics, their black-box nature has limited their widespread adoption, particularly in highly regulated domains. One field that has yet to fully maximize the potential of Artificial Intelligence (AI) is credit risk assessment. On one hand, AI can be highly useful due to its ability to process large amounts of data and identify complex patterns, leading to more accurate predictions and better risk management. On the other hand, maintaining transparent and fair decision-making processes is crucial. Progress in the field of 'Explainable AI' (XAI), which refers to the development of machine learning models that provide clear and understandable explanations for their decisions, offers new opportunities to achieve transparency and comprehend the factors contributing to the decision-making process.

This thesis investigates the application of machine learning models in credit risk assessment and evaluates the trade-off between model accuracy and explainability. Specifically, it focuses on whether simpler models can outperform more complex ones in predicting the bankruptcy of small and medium-sized German companies. The theoretical section emphasizes the importance of responsible technology usage in credit risk assessment, particularly in the context of explainable AI. In the empirical section, the performance of different machine learning models is compared based on their predictive accuracy and their ability to uncover how input variables affect predicted outcomes. The results suggest that Neural Network and the Random Forest models outperform other models in predicting credit default risk. However, while post-hoc explainability techniques were applied, they did not provide conclusive results as to which variables were primarily responsible for the black box models' predictions. Therefore, a simpler choice of a Decision Tree, which generates clear and interpretable results, may be worthwhile, even if it results in slightly lower predictive performance.

Table of Contents

Chapter 1

Introduction

Technological advancements, including the availability of data, an increase in computing power, and progress in research, have made Artificial Intelligence (AI) technology a popular decision-making tool for financial institutions and financial research in the academic field to estimate the credit default risk of households and institutions (Cao, 2020). The application of AI presents opportunities such as the standardization of lending practices, increased inclusion of participants through lower costs for credit assessment, and a general improvement in the quality of risk management decisions. However, as many popular AI models lack transparency and explainability - two important factors for establishing reliable technology - their widespread usage should be treated with caution (Misheva et al., 2021).

Early generations of machine learning models[1] were relatively easy to interpret and explain, but the emergence of Deep Learning[2] has resulted in increasingly opaque prediction models. This opaqueness is closely tied to their success, as Deep Learning systems can effectively accommodate large parametric spaces comprising hundreds of layers and millions of parameters. Consequently, these models are often referred to as 'Black Box models', since it is often unclear how the model arrives at its predictions or how it weighs different input variables. The lack of transparency poses a significant challenge to interpreting these models, deeming them unsuitable in certain decision-making processes. (Barredo Arrieta et al., 2020; Rudin and Radin, 2019).

Credit default risk modeling, in particular, raises ethical concerns such as the lack of consumer protection, data privacy issues, and discriminatory predictions. Machine learning-based credit models typically rely on a variety of risk drivers, making it

[1]The term 'Machine Learning' refers to a subfield of AI, more specifically algorithms that automatically learn patterns in data with the use of mathematical and statistical tools (Helm et al., 2020).

[2]'Deep Learning' is a further subfield of machine learning models. It mainly describes neural networks with a multitude of layers, which transform input data and result in complex and abstract representations (Goodfellow et al., 2016)

challenging to identify and expose the dominance of any single component in the final credit rating. This can lead to potentially unethical predictions going undetected (Langenbucher and Corcoran, 2021).

The demand for powerful yet transparent and interpretable models, which can justify and legitimize automated decisions, led to the research field of 'Explainable and Interpretable AI'. This field aims to address the shortcomings of black-box statistical machine-learning models by developing mechanisms that allow for a deeper understanding of the model mechanics. The ultimate goal is to make these models more transparent and traceable, without sacrificing the effectiveness of current-generation AI systems (Holzinger et al., 2022a; Bazarbash, 2019; Murdoch et al., 2019).

This work aims to investigate the extent to which these explainability and interpretability techniques deliver on their promise of making black-box classifiers for credit risk predictions traceable, without sacrificing predictive power. Additionally, this study seeks to enhance the selection of suitable classification models in the field. The first part will comprise a theoretical literature review, which will elaborate on the relevance of Explainable AI (XAI) in credit risk management, the risks of discrimination in AI-based credit scoring, and how XAI models can both mitigate these risks and enhance the application of machine learning in the field.
The empirical part of the thesis tests and compares the performance of interpretable models and XAI methods. Based on 2018 data for approximately 450 000 German companies, their default in the year 2019 will be predicted. The needed information is retrieved from the database 'Amadeus' (Amadeus, 2023), containing data of small and mid-sized companies, and fed into a matrix of over 100 variables used for the model predictions. Default is indicated using a binary target value for training different machine learning models, where "1" indicates "Insolvency" and "0" represents "Solvency". The performance and explanatory power of five machine learning models are compared and ranked based on several performance evaluation metrics to determine whether complex models truly perform better.

The results suggest that there is no one-size-fits-all solution for credit risk prediction. While complex models seem to perform better in regard to prediction accuracy, they might not be the best option in every application scenario. Despite their ability to better accommodate non-linear interactions, factors such as computational cost should be taken into consideration when selecting appropriate models, particularly for large datasets. Additionally, significant emphasis should be placed on preprocessing, as performance is heavily dependent on the treatment of data prior to modeling. Notably, XAI can be a helpful tool during the data preparation phase.

Moreover, it should be noted that a limitation of this study is the lack of an evaluation metric for explainability. As the concept of explainability is related to psychology and can be manifested in various forms such as visuals, text, and more, the selection of a technique should depend on the specific use case and the intended audience. Therefore, the choice of an XAI technique should be tailored to the specific task at hand.

Chapter 2

Motivation

2.1 Definition of Explainable AI (XAI)

To understand why XAI is relevant to the credit risk sector, the term itself must first be specified: 'Explainable Artificial Intelligence' can be defined as "a field of Artificial Intelligence that promotes a set of tools, techniques, and algorithms that can generate high-quality interpretable, intuitive, and human-understandable explanations of AI decisions"(Das and Rad, 2020). Misheva et al. (2021) add that its purpose is to produce models that "offer an acceptable trade-off between explainability and predictive utility and enable humans to understand, trust, and manage the emerging generations of AI models." As understanding applied models is not only relevant to data scientists and machine learning engineers but increasingly to the users of algorithms who expect explanations behind automated decisions, the concept of explainability is strongly tied to the user experience and the research field of 'Human-Computer Interaction'[1].

While the distinction between interpretability and explainability is not always clear in the literature, Barredo Arrieta et al. (2020) provide a helpful framework. They define 'interpretability' as a "passive characteristic of a model that refers to the level at which a given model makes sense for a human observer", relating to the model´s transparency. In contrast, explainability is an "active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions". XAI is therefore linked to post-hoc explainability, covering techniques that are applied to transform non-interpretable models into explainable ones (Barredo Arrieta et al., 2020). Fuhrman et al. (2022) use the term 'explainability' when referring to measures that are applied to "explain and improve

[1] "Human-Computer Interaction (HCI) is the study and the practice of usability. It is about understanding and creating software and other technology that people will want to use, will be able to use, and will find effective when used."(Rogers et al., 2015)

the AI system", and 'interpretability' to understand the algorithm output.
The lack of a clear definition of explainable and interpretable AI presents a problem that spills over to the analysis of the effectiveness of different XAI techniques. There is no clear way of operationalizing explainability and comparing it across different models. As the psychological and human interaction implications are beyond the scope of this thesis, the distinction between interpretability and explainability will mainly serve as a categorization of different techniques presented in section 3.1 and chapter 4. This distinction will differentiate between interpretable models which provide explanations without the need for additional steps ; and explainability techniques that can be applied to black-box models to provide an explanation about them (Guidotti et al., 2018).

2.2 The Relevance of XAI in Credit Risk Management

The financial sector has been quick to identify the potential of AI technologies, including their use in credit decisions based on analyzing a borrower's financial data to predict their ability to repay a loan. However, there is a risk of misuse or abuse of AI if it is not regulated properly, which can lead to dysfunctional credit lending systems that hinder individual and macroeconomic prosperity (Cafiso, 2022).
The demand for XAI methods and architectures is generally driven by three factors that contribute to ethical decision-making: the need for "more transparent and traceable models", "techniques that enable humans to interact with and interpret them", and "trustworthy inferences" (Vilone and Longo, 2021). The next section will describe how these factors are relevant in the context of credit lending, discuss how AI can be misused in lending decisions, and explain how XAI can both mitigate such risks and enhance the application of ML in the field (Barredo Arrieta et al., 2020).

2.2.1 Risk of Discrimination in AI-Based Credit Scoring

Compared to non-automated risk assessment by experts, quantitative models are generally praised for their ability to assess risk without relying on possibly flawed human intuition (Krainer, 2005). Quantitative models have been used in the field for a long time without generating significant controversy about their fairness. However, with the introduction of AI in automated decision-making, the discourse changed: In 'traditional' credit scoring models, humans had to manually decide which factors determined credit scores and how these factors were weighted. This approach was not only common in determining household credit risk but also applied to pro-

duce scores for determining a company´s financial health using metrics such as the 'Altman Z-score', which uses financial ratios to predict the probability of default (Altman et al., 2017). In contrast, AI models are able to autonomously choose and weigh those factors. This creates an incentive to provide even larger data pools and allow the system to do the work on its own. Especially in the case of household lenders, the larger data pool no longer just includes an individual's financial history but also 'alternative data'. This consists of information from a person's social media profile, mobile phone records, or online shopping history. All this information is fed into the model, which provides a score for the individual but does not necessarily explain how the score was derived (Langenbucher, 2020).
While granting loans based on certain borrower characteristics such as gender, race or religion is prohibited by law in the E.U, those regulations do not necessarily shield consumers from discriminatory lending practices. The underlying problem with AI-based models goes beyond excluding explicit characteristics. The risk of discrimination or disparate impact on protected groups increases with the size of the data pool used by machine learning algorithms. This is because algorithms can find complex correlations between input variables, some of which may have no obvious relationship to a person's financial tendencies or responsibility. A credit scoring algorithm might, for example, have data about which website the applicant is shopping at and what messenger app they are using, which is then inputted into a model that uses those two variables in combination to predict that an individual is more likely to default on their credit. At the same time, those two variables might approximate a certain race, even though 'race' is not included as a variable in the model. With high-dimensional and complex black-box AI models, the person responsible for the models and hence the credit scores, might not even be aware of the protected class having an indirect influence on the model's prediction (Langenbucher and Corcoran, 2021). To avoid segregation by factors like gender, location, origin, or race it becomes necessary to understand the model's decisions and the processes behind some widely used black-box algorithms (Torrent et al., 2020a).

The European Commission has recognized the potential for discrimination in AI-based credit scoring and has released (Commission et al., 2019) to address the issue. However, the policies remain preliminary and do not provide specific action plans (Smuha, 2019).
While the EU provides directives prohibiting discrimination in certain cases such as access to housing and other goods and services, they may not be comprehensive enough to cover credit scoring. As a result, standardization may be necessary for credit scoring and loan providers to fall under these directives (European Parliament and Council, 2004).

A more recent proposal for an 'AI Act' has been introduced by the EU to establish a regulatory framework aimed at preventing harmful practices that contradict fundamental Union values, including respect for human dignity, freedom, equality, democracy, and the rule of law. This framework includes provisions for protecting Union fundamental rights, such as the right to non-discrimination, data protection, and privacy, and could potentially address discriminatory practices in AI-based credit scoring (Langenbucher and Corcoran, 2021; European Commission, 2021).
Lastly, the General Data Protection Regulation includes an article on automated decision-making that provides individuals with the right to an explanation of automated inferences (Vilone and Longo, 2021).
Even though ethical and regulatory constraints exist for using AI for credit risk assessment, the prevailing vagueness and potential loopholes in the regulations emphasize the importance of further research on how these requirements should be technically implemented (Hacker and Passoth, 2021).

2.2.2 Regulating AI in the Banking Sector: Mitigating Systemic Risks and Ensuring Technical Transparency

The regulation of AI in the context of credit default is not only important for ensuring technically transparent implemented models which shield the individual. It is equally relevant for controlling systemic risks in the banking sector. Inadequate risk models have been cited as a significant contributor to the magnitude and spread of the financial crisis in 2008, failing to account for the potential for widespread and cascading failures. Thus, this area has been at the forefront of developing internal compliance and quality regimes, which are now also being considered for AI regulation. The current regulatory regime is still based on the long-standing use of econometric and statistical models to anticipate risk in the banking sector and the existing regulations addressing systemic risks are based on laws enacted by the EU in response to 2008 (Hacker and Passoth, 2021).
Credit institutions are legally required to provide "robust risk monitoring and management systems". More specific regulations require "banks to validate the score quality (accuracy and consistency) of models for internal rating and risk assessment, via continuous monitoring of the functioning of these models". Additionally, "statistical models and other mechanical methods for risk assessments must have good predictive power, input data must be vetted for accuracy, completeness, appropriateness, and perceptiveness, models must be regularly validated and combined with human oversight" (Hacker and Passoth, 2021).
It should however be noted that these rules only apply to institutions holding a banking license. Other institutions tied to the banking sector, which are not di-

rectly tied to the regulations, are excluded. This for example includes credit rating agencies. The existing regulations should therefore be seen as a foundation to build on and which to extend to other relevant parties (Hacker and Passoth, 2021).

Independent of the scope of institutions addressed by the regulations, there still exists a need to translate those normative constraints into concrete technical model requirements, which should be built through a close collaboration between a variety of research fields spanning from regulators and banking experts to statisticians and psychologists.

Chapter 3

XAI for credit risk prediction

3.1 Technical Requirements

This section provides an attempt at translating the legal and ethical obligations presented in chapter 2 into broad technical requirements for the application of AI in credit risk management:
Firstly, models must provide high accuracy, which can be verified by testing the models when applying supervised learning algorithms.
Secondly, explanations of which features were relevant in a given prediction must also be provided. This serves the purpose of validating that predictions can be generalized beyond the test set and enabling human review of individual cases. In theory, explainability techniques make it possible to investigate how variables affect individual predictions, what the overall importance of predicting variables is, as well as how combinations of variables affect both metrics.
To understand how XAI can address the outlined problems and requirements, the following section illustrates how different methods and their capabilities can be categorized.

3.1.1 Local vs. Global Explainability in AI

From a technical standpoint, explanations can be categorized as either 'local' or 'global', depending on whether they explain individual model decisions or instead characterize the entire model. Local explainability is critical for verifying specific cases, whereas global explainability provides trust that the model behaves reasonably across various domains and scenarios (Hacker and Passoth, 2021). This distinction between local and global explanations translates to two levels of user trust: local explanations address whether individual predictions are trustworthy enough to take action, while global explanations instill confidence that the model's behavior is sound (Misheva et al., 2021).

In the context of AI regulation, local explainability can be used to verify that individual credit decisions are not based on discriminatory variables, while global explainability can be applied to ensure that models make decisions based on factors that do not contribute to systemic risks, such as excessive leverage.

3.1.2 Model-Based and Post-Hoc Techniques

As discussed in section 2.1, there is a differentiation in the field of explainable AI between models that are inherently interpretable and those that require additional XAI techniques for their explanation. The former, also known as the 'model-based' approach, involves building models that inherently provide explanations as part of their decision-making process, and are thus also known as 'ante-hoc' interpretable models. These models provide mappings of inputs to outputs that are visible and understandable to the user. In contrast, the 'post-hoc approach' focuses on deriving explanations for complex models after they have already been trained using XAI techniques. (Murdoch et al., 2019; Holzinger et al., 2022a)

3.1.3 Key Considerations in Selecting Suitable Models and Techniques for Credit Risk Assessment

The distinction between model-based and post-hoc explainability techniques discussed in the previous section highlights the different approaches to achieving interpretability in AI models. Abstracting from specific use cases, a trade-off is often observed when building machine learning models: Interpretable models which are simple and linear can often be more easily understood by humans but might not provide sufficient predictive power. Highly non-linear models, such as neural networks with many layers, often provide better performance but are too complex for humans to understand (Morocho-Cayamcela et al., 2019). Morocho-Cayamcela et al. (2019) provide a general overview of this trade-off between accurate predictions and the explainability of the underlying mechanics, which is depicted in figure 3.1. The figure shows that models such as Neural Networks rank or Random Forest rank among the highest when it comes to prediction power, while models such as Regressions or Decision Trees are highly interpretable but less accurate in their predictions.
As stated in chapter 1, the black-box character of many machine learning models makes their application unsuitable for predicting credit risk. However, the effectiveness of explainability techniques cannot be quantitatively compared to interpretable models in an exhaustive manner due to the subjective nature of explainability and the diverse needs of various stakeholders. For the purpose of this study, the focus is on comparing to which extent different models can identify the key drivers behind credit risk default classifications. Based on Murdoch et al. (2019) the introduced

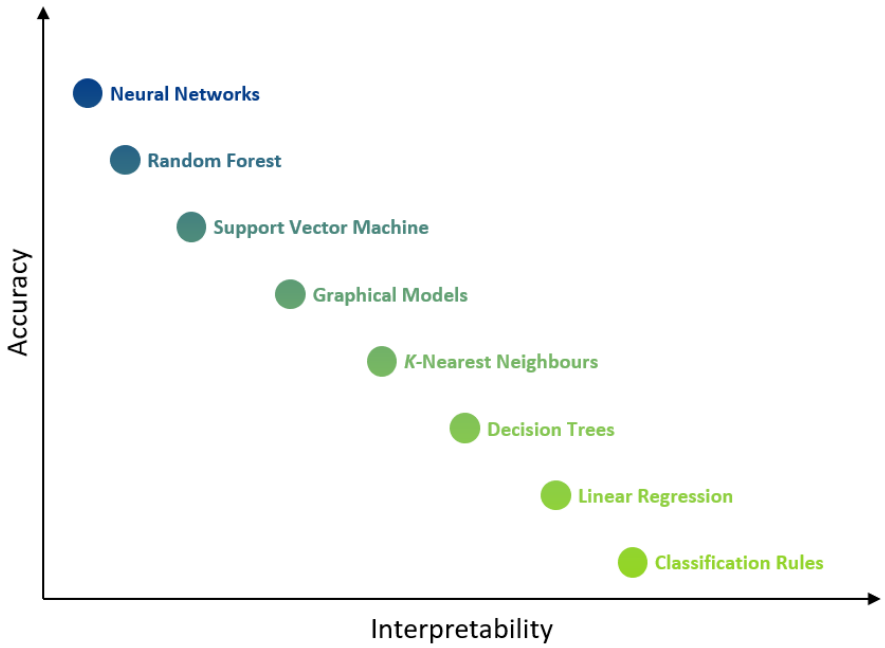

Figure 3.1: Trade-Off between predictive and descriptive accuracy of machine learning models
(Morocho-Cayamcela et al., 2019)

models will therefore be mainly compared regarding their predictive and descriptive accuracy:

Predictive accuracy refers to how well a model can make accurate predictions on new data, or how well it can approximate underlying data relationships. This is typically evaluated using performance tests such as cross-validation, where the model is trained on a subset of the data and tested on the remaining data to measure its ability to generalize (Murdoch et al., 2019).

Descriptive accuracy refers to how well the model's learned relationships can be interpreted and understood. It involves analyzing the model's coefficients, feature importances, or other interpretable measures to gain insight into how the model is making predictions. Descriptive accuracy can be evaluated through local or global analysis of the fitted model (Murdoch et al., 2019).

However, these performance metrics should always be evaluated in the context of the specific problem being addressed and the underlying relationships present in the data. Comparing performance metrics across different contexts or problems can be misleading and may not accurately reflect the true performance of a model. The empirical implementation of these metrics for the case of credit risk prediction is described in 6.4.

3.2 Model Selection

The increased interest in the field of explainable artificial intelligence has led to an excessive amount of context- and domain-specific methods and approaches aimed at interpreting machine learning models. Consequently, it becomes difficult to select the most relevant ones as strengths and weaknesses differ depending on the data it is applied to as well as the objective classification problem (Vilone and Longo, 2021). In their paper "Machine learning techniques for credit risk evaluation: a systematic literature review" Bhatore et al. (2020) review the application of machine learning techniques to credit scoring, which involves using a classifier to distinguish between creditworthy and credit-unworthy customers based on historical data. They find that several machine learning techniques are commonly used for credit scoring, including Neural Networks, Support Vector Machines, naive Bayes, Markov Chain, hidden Markov models, Bayesian Networks, Decision Trees, Bayesian ensembles, HLVQ-C, and various hybrid and ensemble models. Among these techniques, the most popular is the neural feed-forward network, despite its lack of interpretability. Therefore, for this thesis, the selection of XAI methods will be narrowed down based on several sources, which do not stem from research in the field of credit risk management. Those include "Explainable AI Methods - A Brief Overview" (Holzinger et al., 2022b) and "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (Barredo Arrieta et al., 2020). Both sources provide an overview of all XAI methods knowingly available at the point of the respective paper's release relating to the field of credit risk management. However, as the implementation of all methods would be out of scope, the selection was further narrowed down to the most popular XAI toolboxes and repositories on GitHub[1], as presented by Holzinger et al. (2022b). Moreover, some models had to be discarded due to insufficient available computing power in the face of maintaining comparability of the models when applied to a large amount of data.
The following chapter describes the selected interpretable models and explainability techniques. The former is comprised of a Logistic Regression model, that will mainly serve as a benchmark, a Decision Tree Classifier as well as the Explainable Boosting Machine algorithm. Next, two post-hoc techniques, namely SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are presented. These techniques are applied to two black-box classifiers, namely a Neural Network and a Random Forest model. As these models do not give insight into their internal, their mechanics are not elaborated on but can be further explored using sources such as " An Introduction to Multilayer Perceptron

[1] "GitHub is a web-based version control repository hosting service that provides developers with a distributed revision control and source code management (SCM) functionality of Git, in addition to its own features (Al-Qutaish, 2020)."

Algorithm and Its Applications in Clinical Decision-Making" (Zhang et al., 2018) or "Random Forest" (Breiman, 2001).

Chapter 4

Models

4.1 General Model

For each company, n, the response variable Y_n indicates if it will default (become insolvent) in the next year based on the model's prediction. $Y_n = 1$ indicates that the company is insolvent, while $Y_n = 0$ indicates no default. The explanatory input variables X_n are a vector of data from the previous year that are assumed to influence Y_n. A detailed description of the data is provided in section 6.2.

4.2 By-Design Interpretable Models

Models that are inherently explainable can be labeled as transparent. Using models which give access to their internals, is usually the easiest way to ensure interpretability. The most commonly used models for this purpose are regression and tree-based models, which are often sufficient when dealing with less complex problems (Mi et al., 2020; Barredo Arrieta et al., 2020).

4.2.1 Logistic Regression

Logistic regression models are a commonly applied method providing high interpretability (Bussmann et al., 2020). However, due to their statistical simplicity and the high dimensionality of the input data, poor predictive performance is to be expected. Nevertheless, they serve as a commonly used reference point for comparison against more complex models (Gu et al., 2020). Furthermore, as banks do not disclose their exact loan decision models, classic linear models serve as the closest approximation for the banking sector (Addo et al., 2018).
Logistic regression models are an extension of linear regression to classification problems, where the output of a linear equation is transformed using the logit function

to obtain the probability of the outcome. Specifically, for a binary outcome variable, the logistic regression model can be expressed as:

$$\ln\left(\frac{p_n}{1-p_n}\right) = \alpha + \sum_{j=1}^{J} \beta_j x_{nj} \tag{4.1}$$

Here, p_n indicates the default probability of company n, and the explanatory variables for each company n are contained in the J-dimensional vector $\mathbf{x}_n = (x_n 1, \ldots, x_{nJ})$. α denotes the intercept and β_j the jth regression coefficient. After estimating α and β_j, the default probability is estimated as follows:

$$p_n = \frac{1}{1 + \exp(\alpha + \sum_{j=1}^{J} \beta_j x_{nj})} \tag{4.2}$$

The linear regression learns the coefficients (β) for a weighted sum of feature inputs. The output of the linear regression is normalized between 0 and 1 using the sigmoid function 4.2. After obtaining the probabilities, depending on the output value, classes can be assigned. If $p \geq 0.5$, class is assigned as 1, and if $p < 0.5$, class is assigned as 0 (Bussmann et al., 2021).
As with the linear regression, the output is the weighted sum of input variables. Betas are optimized to minimize the loss function, and the coefficients are easy to interpret: For example, if x_1 - a continuous feature - is increased by one unit, the output beta is increased by β_1. Therefore, the impact of x_1 on the output y can be easily understood when looking at changes in the input features, and the model provides full transparency on how predictions are calculated, and the contribution of inputs to the outputs is known.
Apart from their poor performance for high-dimensional tasks, the major drawbacks of logistical regression models are the requirement of strong statistical assumptions and the non-allowance for non-linear effects(Gu et al., 2020; Mi et al., 2020).

4.2.2 Decision Trees

Another commonly used intrinsically interpretable method is the application of classification trees, which have gained popularity as a machine-learning technique for combining predictor interactions. Trees, in contrast to linear models, are entirely nonparametric and have a different logic from conventional regressions. Classification and Regression trees are able to accommodate non-linear relationships between variables and are unreliant on any assumptions about the underlying probability distributions and functional forms of the relationship between the predictor variables and the outcome variable. In general, trees work by finding clusters of observations that act similarly to one another. The data is divided numerous times depending

on various feature cutoff levels. The dataset is divided into many subsets, with each instance belonging to a separate subset. The intermediate subsets are known as internal nodes or split nodes, while the final subsets are known as terminal or leaf nodes. Based on average outcomes in nodes within the training data, predictions can be made. (Gu et al., 2020)

There are different ways of growing trees, differing in aspects like the tree's potential structure (such as the number of leaves per node), which criteria are used to identify splits when to terminate splitting, or how the basic models within leaf nodes are estimated. According to Molnar (2019), CART, short for 'Classification And Regression Trees algorithm', is the most widely used technique for tree induction, therefore it is used to explain the underlying mechanics. Most other tree algorithms can be interpreted similarly. The relationship between feature variables x and the predicted outcome y can be represented as follows:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I x \in R_m \tag{4.3}$$

Each observation belongs to one leaf node, denoted by a subset of R_m. The function $I_{x \in R_m}$ is called the identity function, returning 1 if the observation belongs to the subset of R_m and 0 otherwise. If an observation falls into a particular leaf node R_l, the predicted outcome is $\hat{y} = c_l$, with c_l denoting the average value of all training observations within the leaf node R_l.

In CART, subsets are created by determining the best cutoff point for a feature that minimizes the Gini index of the class distribution of y (for classification tasks). The Gini index quantifies the probability of misclassifying an example by measuring the impurity of a node, or the degree to which all observations in the node belong to the same category. The more observations belong to the same class, the less impure the node is. Therefore the optimal cutoff point creates subsets that differ as much as possible in the target outcome. The algorithm continues this search-and-split recursively in both new nodes until a stop criterion is reached. Possible criteria are a minimum number of instances that have to be in a node before the split or the minimum number of instances that have to be in a terminal node.

The visual representation of the trees makes the algorithm interpretable as a set of "if else rules", allowing for transparently following the reasoning process and determining which category would be more likely if a certain variable of a given instance had a different value. The global importance of a feature can be determined by going through all splits and measuring how much it contributes to a reduction of the Gini index in relation to the parent node and scaling all importance values to 100, measuring the overall model importance.

Local explanations for individual predictions can be generated with Tree Decomposition, which involves breaking down the decision path into components, where each component corresponds to a specific feature. This allows for tracking a decision through the tree and explaining a prediction by examining the contributions added at each decision node. If only the root node were used to make a prediction, it would simply predict the mean of the outcome of the training data. Moving down the tree, based on the next node, each split adds or subtracts a term from this sum. A final prediction is reached by following the path of observation and adding to the formula:

$$\hat{f}(X) = \overline{y} + \sum_{d_1}^{D} split.contribu(d, x) = \overline{y} + \sum_{y=1}^{p} feat.contrib(j, x) \tag{4.4}$$

The outcome of a prediction consists of the mean of the target outcome and the addition of the contribution of D splits between the root and the final node. Since a feature can occur in more than one split, all contributions of all p features are added to determine their overall contribution for a given observation (Molnar, 2019).
One significant drawback of decision trees is their limited ability to model linear relationships, as they rely on approximating such relationships with step functions, which can be inefficient. Moreover, the interpretability of decision trees is heavily dependent on the size of the tree, with shorter trees being easier to understand. However, ensemble methods such as random forests can be employed to overcome some of these limitations, by combining multiple trees into a more complex model. While these methods offer improved performance, they often come at the cost of reduced interpretability, making them less suitable for certain applications where transparency is critical (Molnar, 2019).

4.2.3 Explainable Boosting Machine

The explainable boosting machine (EBM) algorithm is the last inherently interpretable model applied in this thesis. It was specifically developed as a method that is able to reach the predictive accuracy of popular methods like random rorest or boosted trees, while still remaining explainable.
It is based on the generalized additive model (GAM). This regression model allows for non-linear relationships between the target variable and each individual feature, by decomposing the model into a sum of smooth functions of the features, each of which only affects one feature. Since each feature has its own smooth function or "shape function" $f(j)$, it can be plotted and interpreted. The generalized additive

model has the following form:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) \tag{4.5}$$

where β_0 represents the intercept, g the link function (logistic function for classification), and f_i the individual shape functions. With the introduction of EBM, techniques such as bagging and gradient boosting can be used to formulate f_i as ensembles of trees. By introducing more complex shape functions, the model improves both predictive and descriptive power. The model can be improved even further with GA^2M. The model uses the FAST algorithm to choose K feature pairs (k_1, k_2) (where K is a small number), to include as pairwise interactions in the model:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{k=1}^{K} f_k(x_{k_1}, x_{k_2}) \tag{4.6}$$

Pairs are chosen greedily, starting with the pair with the highest marginal contribution to the model performance and continuing until adding interactions does not further improve the model´s performance. Given that their interactions can be visualized as heatmaps, the algorithm remains interpretable (Chen et al., 2022; Nori et al., 2017).

4.3 Post Hoc Methods applicable to Black Box Models

If a machine learning model is not inherently transparent, there are separate methods that can be employed to explain its decisions. Post-hoc methods, also known as model-agnostic methods, do not rely on access to the internal workings of the model (such as the learned parameters like weights and biases in neural networks) and instead work with the model's output. As a result, they can be applied to a wide range of models and are highly flexible (Holzinger et al., 2022b).
Two popular techniques used to explain the decisions of black-box models, which do not provide easy-to-interpret insights into how they arrive at predictions, are Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). These techniques can be applied to various models and help to provide a better understanding of how these models arrive at their decisions. The following sections are based on Ribeiro et al. (2016), Lundberg and Lee (2017) and Molnar (2019).

4.3.1 LIME: Local Interpretable Model-Agnostic Explanations

LIME (Local Interpretable Model-Agnostic Explanations) is a widely used post-hoc explanation technique that can be applied to any black-box model and many types of data. It generates local explanations by fitting a surrogate model that approximates the complex model in the local neighborhood of a specific input data point. The surrogate model is selected from a class of potentially interpretable models, such as linear regression and its variants.
The formal equation for LIME explanations is as follows:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{4.7}$$

Here, x represents an input data point, f represents the complex underlying model, G represents a family of potentially interpretable models, and $g \in G$ is the selected surrogate model. The first term, $\mathcal{L}(f, g, \pi_x)$, is the loss function used to fit the surrogate model to the complex model in the local neighborhood of the input data point. The second term, $\Omega(g)$, is a complexity penalty that ensures the surrogate model stays simple.
The loss function $\mathcal{L}$ quantifies the difference between the predictions of the complex model and the surrogate model, weighted by a proximity measure π_x that defines the local neighborhood around the input data point. LIME generates new data points by randomly perturbing the input data point and uses them to fit the surrogate model. The proximity measure is used to weight the loss function based on how close the new data points are to the input data point.
The complexity penalty Ω is used to select a simple surrogate model that is interpretable. LIME typically uses sparse linear models as the family of interpretable models, which aim to produce as many zero-weighted inputs as possible. This can be achieved using regularization techniques such as Lasso regression.
The original LIME paper (Ribeiro et al., 2016) uses a locally weighted squared loss function that quantifies the distance between the labels from the complex model and the surrogate model. The proximity measure is defined using an exponential kernel that assigns higher weights to data points that are closer to the input data point. This ensures that the surrogate model is locally faithful to the complex model in the local neighborhood of the input data point.
Overall, LIME provides a way to efficiently generate local explanations for individual input data points but it does not have the ability to generate global explanations which is a disadvantage in the context of applying it to credit risk predictions.

4.3.2 SHapley Additive exPlanations: SHAP

Another popular XAI method used to explain machine learning models is the SHapley Additive exPlanations (SHAP) approach, which unlike LIME can also generate global explanations.

This method is based on cooperative game theory and provides an economic foundation for explaining machine learning models by breaking down individual variable contributions to the prediction. The approach can be applied to any underlying machine learning model, including neural networks. Shapley values were first introduced in 1951, way before machine learning became a popular tool and represent the average contribution of each feature to the prediction. The main idea behind SHAP is to compare the performance of a prediction with and without a specific feature to determine its individual contribution. However, interactions between features also need to be considered, which requires analyzing subsets of features when removing individual features. The contributions of single features are then calculated for each subset and averaged over all contributions to determine the marginal value of a variable to the prediction (Bussmann et al., 2021; Lundberg and Lee, 2017).

In the context of machine learning, Shapley Values are mainly used to explain local predictions. However, it is also possible to generate global explanations by aggregating over individual predictions. The values are calculated as follows: Equation 4.8 represents the Shapley value for a feature i, specifically a specific feature value (e.g. company type: small shop).

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{4.8}$$

The input for calculating the Shapley value ϕ_i is the black-box model f, as well as the input data point x (a vector of features for one observation). The first step is to iterate over all possible subsets $z' \subseteq x'$ to account for interaction effects (e.g., only a subset of age and BMI). Then, we obtain the black-box model output $f_x(z')$ with the feature i, and the output $f_x(z' \setminus i)$ without the feature of interest. The difference between these two values tells us how the feature i contributed to the prediction in the specific subset (marginal value). This is done for each possible combination, so each permutation of subsets. Each combination is also weighed according to how many features out of the total number of features are in that subset. For the features that are excluded, random values from the training dataset are used as input, so their relevance is sampled out and becomes random.

Calculating all subsets is computationally expensive, as there are 2^n subsets, where n is the number of features. Therefore, the original SHAP paper by Lundberg and Lee (2017) presents the basic idea of approximating the Shapley values instead of calculating all combinations. For instance, Kernel SHAP samples feature subsets

and fits a linear regression model based on these samples. The variables in this linear regression model indicate whether a feature is present or absent, and the output value is the prediction. After training, the coefficients can be interpreted as the approximated Shapley values. This approach is similar to LIME, but in the case of SHAP, proximity is neglected, and samples are instead weighted according to how much information they contain, as large and small correlations tell us the most about the contribution of features. Besides KernelSHAP, there exist other approximation techniques like TreeSHAP or DeepSHAP, which are used for tree-based models or deep neural networks, respectively. These techniques, however, are no longer model-agnostic but can use the model internals to boost the performance when calculating Shapley values.

Chapter 5

Literature Review/ State of Research

5.1 Literature Review

There exist many studies that both predict bankruptcy through machine learning models and evaluate the application of the models with respect to explainability. The research presents differing results with respect to how efficient different methods are:

Misheva et al. (2021) applied post-hoc explanation techniques LIME and SHAP to a credit scoring model developed on a US-based peer-to-peer lending platform, directly connecting borrowers with individual investors who are willing to lend them money. They trained and tested five classifiers, including logistic regression, two tree-based models, a support vector machine, and a simple neural network, with a focus on an economic interpretation of the results obtained using both methods. The authors found that both LIME and SHAP consistently produced explanations that aligned with financial logic. Moreover, they observed stable feature importance, which remained consistent regardless of the test size. The authors highlighted the need for further research to account for practical constraints associated with implementing different techniques, such as computing power and run time.

Similarly, Bussmann et al. (2020) use Shapley values to explain the factors that contribute to predicting bankruptcy risk for small- and medium-sized companies. They employ a post-hoc approach, combining Shapley values with XGBoost, a gradient tree-boosting learning model, and compare the results to a logistic regression model. The authors report a significant improvement in predictive accuracy with the XGBoost model, but it is unclear how both model were optimized. In addition to the Shapley value approach, a network analysis of the underlying factors could also provide insights into the relationships among variables and help identify key

drivers of bankruptcy risk.
In a subsequent paper Bussmann et al. (2021) the authors propose a methodology that combines network analysis and machine learning to identify key drivers of credit risk and to provide transparent and interpretable explanations for credit risk assessments. They apply this methodology to a real-world dataset of small and medium-sized companies and show that it can improve the accuracy and transparency of credit risk models.
In their study, Torrent et al. (2020a) also compared the performance of several machine learning models, including logistic regression, random forests, and gradient boosting, to predict credit risk. The authors studied the credit risk of individuals in the context of consumer credit. Specifically, they used data from a Spanish financial institution that provides personal loans to individuals. They evaluated the models using several performance metrics, such as accuracy, precision, and recall and found that the boosted models outperformed the other models. Furthermore, they used resampling techniques, such as oversampling and undersampling, to handle the problem of imbalanced data sets. Finally, they provided an economic interpretation of the model's predictions by analyzing the impact of each feature on the risk score.
Addo et al. (2018) implemented different machine learning models, including logistic regression, random forest, boosting, and deep learning models, to conduct binary classifications of credit risk. The paper does not provide specific information about the data set used in their study. They focused on predictive performance and compared the model performance based on the ten most important features of these models but applied them to a separate data set. The authors found that depending on the business model and provided data, a pool of models must be considered, and tree-based models turned out to be more stable than multilayer neural networks. They also noted that complex models do not necessarily outperform simpler approaches, and standard performance metrics may not always be suitable, depending on the models being compared.
These inconclusive results show that choosing suitable methods depends on the application scenario and data at hand.

Chapter 6

Implementation

6.1 Data Sources

In Chapter 2, the theoretical part mostly focuses on the risks of applying machine learning to predict creditworthiness in the context of private household borrowing. However, due to limited access to data used to predict the creditworthiness of individuals, the empirical part focuses on predicting firm-level credit default to ensure a sufficiently large and realistic database. The analysis can be replicated on different types of borrower data, including other institutions and private households.
The necessary information was obtained from the 'Amadeus' database (Amadeus, 2023), which contains data of small and mid-sized companies, including financials, financial strength indicators, and corporate structures. Unlike other databases, Amadeus includes information about private companies not listed on the stock market, providing a sufficiently large number of observations. This information was gathered in a feature matrix. The database contains information about the insolvency of companies, however, to get a more accurate dataset, the German insolvency register was web-scraped for all commercial register numbers during a given timeframe to identify all defaulted companies, which were used as a binary target value for training different machine learning models to indicate the bankruptcy status of each company (Insolvenzbekanntmachungen, 2022). Legally, all bankrupt companies must be listed in an insolvency register with a company registration number, enabling the matching of all insolvent German companies with company information obtained from the Amadeus database (für Justiz, 2020). Upon combining the two datasets, only approximately 10% of the insolvent companies could be found in the Amadeus dataset. Therefore the presented results are based on the target from the initial dataset. Results achieved using the Insolvency-register data can be found in the appendix.

6.2 Data Description

The initial panel dataset contained approximately 4.5 million observations of German companies and included 131 different variables. As the majority of the observations were recorded between 2017 and 2019, the objective was to predict the probability of bankruptcy for the year 2019 based on information from the previous year, i.e., 2018. This led to a reduced dataset size of approximately 450 000 observations. While it was possible to convert the panel dataset from long to wide format, with yearly observations treated as separate variables, this approach would have weakened the analysis results and made it more difficult to identify relevant predictors. The dataset included various variables, such as:

- **FIAS, IFAS, TFAS, OFAS, CUASCOLL, CRPE, EXOP, CURR, SHLQ, SOLR, GEAR, PPE, TPE, SCT, ACE, SFPE, WCPE, TAPE**: Financial ratios related to assets and liabilities
- **DEBT, CASH, CAPI, ONCL, LOAN, CRED, OCLI, WKCA, NCAS**: Various financial metrics related to debt, cash, capital, and other financial factors
- **ENVA, EMPL, OPRE, OPPL, FIRE, FIEX, FIPL, PLBT, TAXA, EXRE, EXEX, EXPT, DEPR, INTE, RD, RSHF, RCEM, RTAS, CFOP, PRMA, GRMA, ETMA, EBMA, NAT, IC, STOT**: Various financial and operational metrics related to the company's activities
- **DATEINC**: The date on which the company was incorporated
- **TYPE_***: Categorical variables related to the type of company
- **CONSOL_***: Categorical variables related to the level of consolidation of the company's financial data
- **INDEPIND_***: Categorical variables related to the company's independent credit rating
- **COMPCAT_***: Categorical variables related to the size of the company
- **QUOTED_***: Categorical variable related to whether the company's shares are publicly quoted.

A full description of the variables can be found in Appendix B. The next section describes how the data was further preprocessed in order to accommodate machine learning model requirements.

6.3 Preprocessing

6.3.1 Dimensionality Reduction

To reduce the number of relevant predictors, non-informative variables were removed. For each pair of predictors, the second variable by order of appearance was removed if the correlation (after Pearson) exceeded a threshold of 95 percent, reducing the number of variables to 100. As touched upon in the introduction and implemented in other studies (Kuhn et al., 2013; Torrent et al., 2020b), explainability techniques can also be used for dimensionality reduction purposes. However, this implies that models have to be applied first to select potentially important features.

6.3.2 Oversampling

Out of a total of 450 000 company observations in 2018, the data indicates that less than 1000 companies were illiquid in the following calendar year (2019), resulting in a highly imbalanced dataset. Imbalanced data can affect the behavior of the algorithm and result in overfitting the majority class, thus favoring its input variables. This issue can be addressed using either over- or undersampling techniques. Since undersampling may miss important entries from the training data in real-life applications, oversampling was preferred [1]. Although there exist complex algorithms such as SMOTE used in a similar study conducted by Addo et al. (2018), they require large computational power. For this thesis, random oversampling was used, which involves randomly selecting examples from the underrepresented class and creating duplicates (Torrent et al., 2020a).

6.3.3 Categorial Data

To include categorical variables One-Hot-encoding was used to convert the corresponding features into dummies. Categories with less than 100 instances were combined into 'Other' and categorial variables with more than 10 variable expressions were removed.

6.3.4 Missing Values

The dataset had a high imbalance and a significant number of missing values, making it impractical to exclude all observations with missing values. A preferred approach for filling missing values would consider the distribution of individual variables and their relationship with other features. However, multivariate techniques such as

[1] Undersampling may be sufficient for comparing the strength of different models.

KNNImputer can be computationally expensive. Thus, the study utilized the IterativeImputer from the scikit-learn library. By default, IterativeImputer employs linear regression to impute missing values, assuming that the missing values are distributed randomly. However, this study could not verify this assumption.

6.3.5 Feature Scaling

To ensure that the dataset could be used with certain learning algorithms, feature scaling was applied to all models except for decision trees and random forests. The scaling was done to normalize the range of values for the features, which had different scales, and to center the data around a mean of zero. For explainable boosting machines, feature scaling helps to reduce the risk of overfitting. Similarly, in the case of neural networks, feature scaling helps to prevent the model from favoring features with higher variances (Torrent et al., 2020b).

6.3.6 Hyperparameter-Tuning and Sample Splitting

To optimize the performance of the machine learning algorithms and prevent overfitting, hyperparameter tuning is performed. Hyperparameters are values set by the user that control how the algorithm operates, such as the learning rate or regularization strength. The specific hyperparameters to be tuned depend on the model, such as penalization parameters in logistic regression or the number of trees in random forests (Feurer et al., 2018). As optimal hyperparameters can vary depending on the data, the hyperparameters for each model are tuned using RandomSearchCV from sklearn. This package randomly tests possible combinations of hyperparameters defined in a parameter space using a five-fold cross-validation scheme, which is a commonly used technique for hyperparameter tuning. The best set of hyperparameters is then selected based on the performance of the model on the validation sets. The exactly selected hyperparameters for each model can be found in appendix A.

6.4 Performance Evaluation

Machine learning models are typically selected based on their ability to make accurate predictions on previously unseen data, without necessarily making strong assumptions about the data distribution, as opposed to traditional statistical methods. Thus, their evaluation focuses on prediction performance, rather than how well they fit the data. (Ij, 2018)
Common metrics for measuring the performance of binary classification problems include accuracy, precision, recall, and F1-score. These metrics are based on comparing ratios between predicted and target ratios and can be visualized using a

confusion matrix, which shows the number of true positive, true negative, false positive, and false negative predictions made by the model (Japkowicz, 2006). After estimating the default probability and classifying each company in the dataset, the prediction accuracy is measured and compared for each model. The data is split into "training" and "testing" subsets, and models are built using the "training" data. The "test" dataset is then used to compare the model's prediction $\hat{Y}_n$ and the actual value Y_n (Bussmann et al., 2021). To account for the imbalanced distribution of the target variable, a 50:50 train-test split is selected to measure the accuracy of predicting the minority class more reliably.
Companies that were predicted to be insolvent but are not in the real data are called false positives (FP), and true positives (TP) represent correctly predicted defaulted companies. Companies that were predicted to still be solvent/not default but in reality defaulted/are insolvent are represented by false negatives (FN), while true negatives (TN) are the insolvent companies that were also predicted as such (see 6.1). The performance metrics used for analyzing the results can be calculated

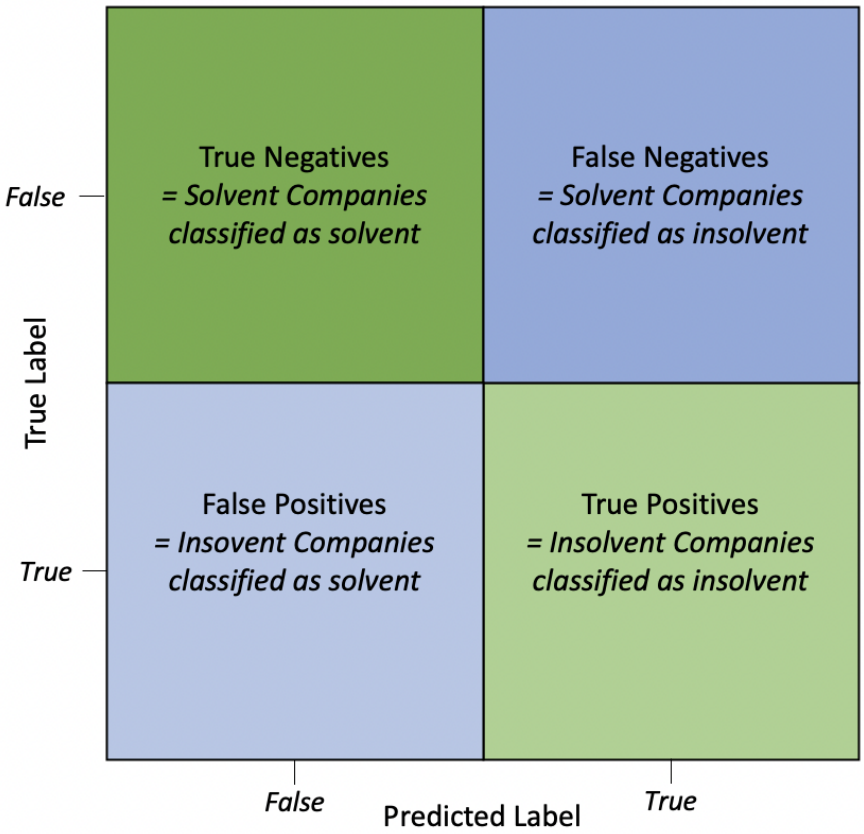

Figure 6.1: Confusion Matrix

as follows:

- $Accuracy = \frac{TP+TN}{P+N}$

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{P}$
- $FPRate = \frac{FP}{N}$

Most performance metrics have limitations. For instance, accuracy is an intuitive evaluation metric as the goal is to reduce classification mistakes, but it does not differentiate between false positives and false negatives, which can cause issues with imbalanced datasets. Precision and recall are metrics that can address this issue by providing information on the model's ability to correctly identify positive examples and its ability to avoid false positives. However, unlike accuracy, precision and recall do not account for how well a classifier determines that a negative example is indeed negative, as they do not rely on the proportion of observations labeled as negative and actually being negative. All metrics should therefore be evaluated in combination (Japkowicz, 2006).
When predicting bankruptcy, the most important performance metric depends on the use case, but generally, the model should aim to keep both false positives and false negatives as low as possible. The F1-score measures the model's accuracy taking into account both precision and recall, forming the harmonic mean:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6.1}$$

In the case of black-box models, verifying these metrics under domain-specific knowledge can be challenging. In this case, Explainable AI (XAI) has an advantage as it allows retracing how the model arrived at its predictions or decisions, which is crucial for identifying and addressing potential flaws (Japkowicz, 2006).
The importance of features provides valuable insight into the variables that are most relevant for the model's predictions. This can be helpful in identifying potential biases or inaccuracies in the input data that could affect the model's performance metrics. Additionally, local and global interpretability techniques can aid in identifying inconsistencies and areas for improvement in the model, ultimately also leading to better predictive performance.

6.5 Constructing and Interpreting the Models

All classifiers were tuned, trained, and fitted using the Scikit-learn software machine learning library for Python (Garreta and Moncecchi, 2013). Visualizations of local and global explanations were generated using InterpretML, an open-source Python package that provides interpretability methods for glassbox models (linear models,

rule lists, GAM) and explainability techniques for blackbox models (Nori et al., 2017). SHAP was implemented using the SHAP-package (Lundberg and Lee, 2020).

Chapter 7

Results

Table 7.1 provides a comparison of the predictive performance metrics for each of the five models used in the study[1]. The Logistic Regression model achieved the poorest results in regards to the F1 score, mainly stemming from wrong predictions of the majority class. Given the imbalanced dataset, the Decision Tree Classifier, Explainable Boosting Machine, and Random Forest models had high accuracy (99.9%) but varied in precision, recall, and F1 score. Optimized for the F1 score, the Neural Network did not achieve the highest F1 score at 93%, but generated the best results at predicting the minorty class, with a recall value of 99%. When optimized for other parameters the overall F1 score could however be improved up to 98.6%. Therefore it can be concluded that overall, the Neural Network model had the best predictive performance, followed by the Decision Tree. It should be noted, that these results were achieved using the model parameters shown in appendix A and can differ, if those are changed. The results of predicting default based on the data gathered in the German insolvency register can be found in Appendix D.

Table 7.1: Comparison of Predictive Performance Metrics across Models

Model	Accuracy	Precision	Recall	FP Rate	F1 Score
Logistic Regression	0.996	0.446	0.981	0.004	0.814
Decision Tree Classifier	0.999	0.915	0.928	0.0002	0.9617
Explainable Boosting Machine	0.999	0.830	0.923	0.0006	0.937
Random Forest	0.999	0.998	0.816	0.000	0.949
Neural Network	0.999	0.765	0.990	0.001	0.931

As for the descriptive results, the local linearity assumption in the case of LIME may make SHAP a more robust choice for making local predictions without assumptions about the data distribution. However, LIME may be a useful choice in cases where good domain knowledge is available and quicker runtime is desired. The differences in results between LIME and SHAP may only be partially eliminated, as the sam-

[1]The values are rounded to three decimal places.

pling methods differ with LIME creating synthetic data and SHAP sampling from the dataset.
Ultimately, using interpretable techniques may be the most reliable option to eliminate mistakes. In this case, given a large amount of available data, the decision tree classifier produces very good prediction results that are indisputable.
The following sections elaborate on the predictive results and compare the models based on their predictive performance.

7.1 Logistic Regression

7.1.1 Predictive Performance

The Logistic Regression model was able to achieve comparably good results when optimized for the F1 score, with a Recall value of 98%. It however performed comparably worse than other models in predicting the majority class, resulting in the overall lowest F1 score. Despite the low optimal choice of C, it´s strength might stem from the fact, that the model does not overfit the data as much as other tested models. The confusion matrix, as shown in Figure 7.1, summarizes prediction results in absolute prediction values.

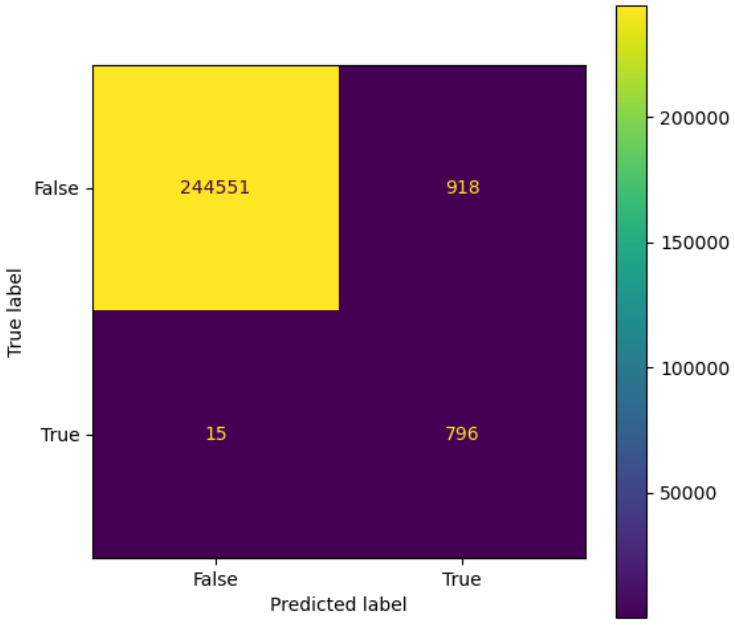

Figure 7.1: Performance Evaluation of Logistic Regression Model

7.1.2 Descriptive Performance

Since Logistic Regression is a linear model, the feature importance can be directly obtained from the corresponding coefficients.

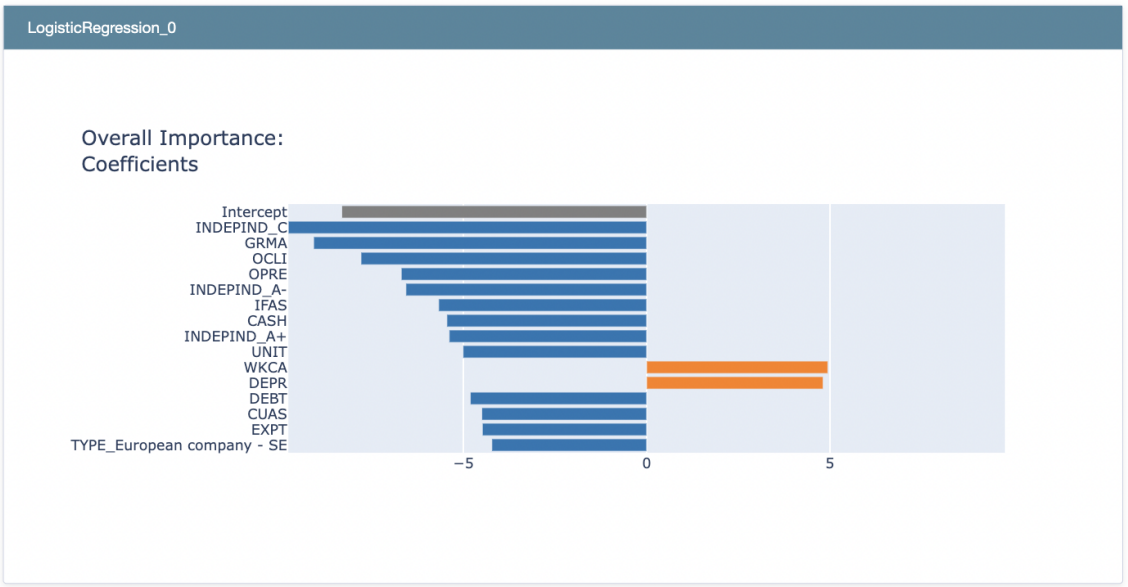

Figure 7.2: Top 15 Predictor Variables Ranked by Absolute Importance in Logistic Regression Model

The independent credit rating (INDEPEND_*) is the most important predictor in the model, followed by the Gross Margin (GRMA), Other Current Liabilites (OCLI) and Operating Revenue (OPRE). The model's L1 penalty, which was automatically chosen as the optimal penalty term, encourages sparsity by driving some coefficients to zero, resulting in a parsimonious set of explanatory variables. As a result, the model uses only a few important predictors to explain the target variable. Figture 7.2 presents the top 15 predictor variables ranked by their absolute importance in the logistic regression model, including the prediction direction of each feature.

Given that no hypotheses about the variable importance were formulated and similar metrics are not applicable to the other models, no significance tests were conducted. Although the Logistic Regression model captures the overall importance of different predictors for the model prediction, interpreting individual predictions can be challenging due to a large number of predictor variables and their varying values across observations. InterpretML provides an interactive display of the results, enabling the review of individual cases and the contribution of each predictor variable to the model's prediction for that case. This can provide insights into the model's behavior in specific instances and help improve our understanding of the model's weaknesses. As no specific feature interactions term were included in the model, there is no need to further examine feature interaction terms.

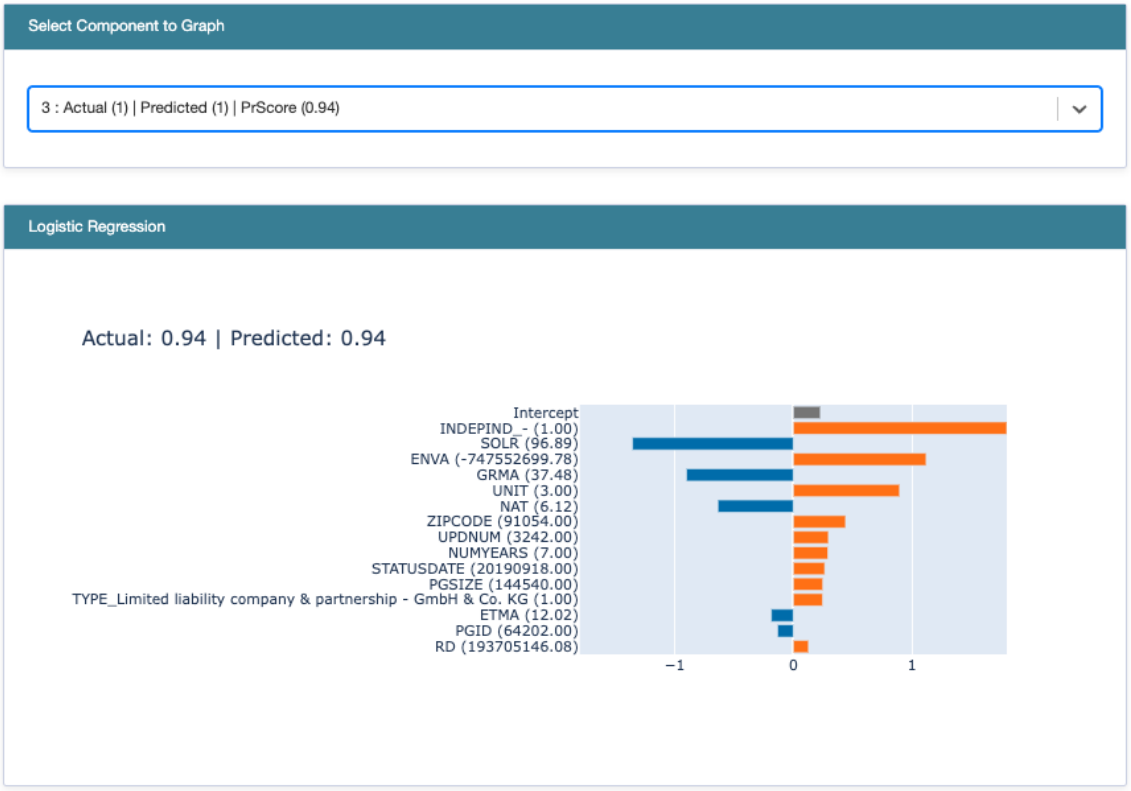

Figure 7.3: Logistic Regression - Local Explanation Output generated using InterpretML

7.2 Decision Tree Classifier

7.2.1 Predictive Performance

The Decision Tree Classifier demonstrated a improvement in overall predictive performance, achieving nearly perfect results, with 99.95% of instances being correctly classified. Of all instances classified as insolvent, 92.8% were actually insolvent, and only 0.002% of solvent companies were incorrectly classified as insolvent. The classifier's F1 score of 96% indicates good overall performance. The confusion matrix 7.4 summarizes the absolute prediction results.

7.2.2 Descriptive Performance

The classification tree model uses different predictor variables compared to logistic regression, as shown in table 7.2. This table displays the top 15 predictor variables ranked by importance in the decision tree classifier. The importance score of each feature is calculated based on its contribution to the overall performance of the decision tree classifier. The higher the score, the more important the feature is in determining the classification of instances. It places less emphasis on independent credit rating and gives more importance to Profit and Loss Before Taxes (PLBT).

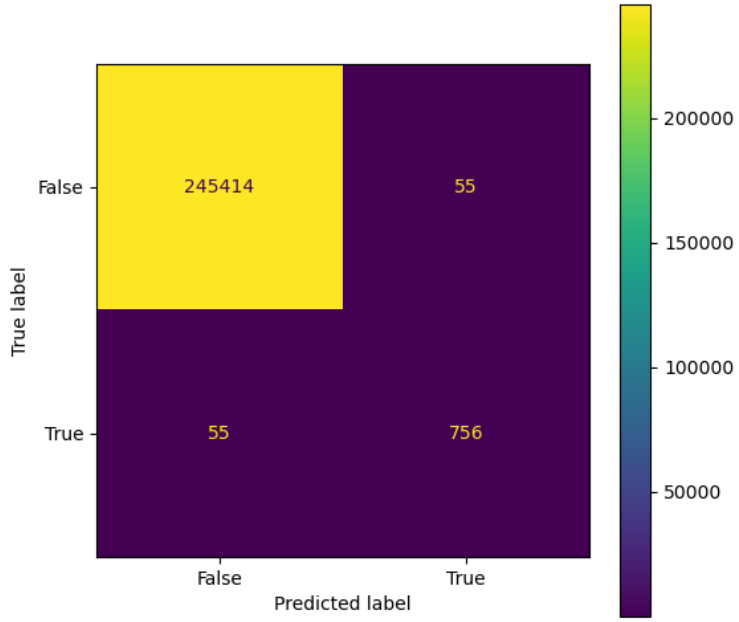

Figure 7.4: Performance Evaluation of the Decision Tree Classifier

Being a non-linear model, it doesn't allow for a direct determination of the direction of the relationship between a feature and the target. However, visualizing the tree can provide insights into the decision-making process that splits the data into different groups. Since the optimal parameters result in a deep tree model, following the results can become tedious. The entire decision tree is visualized in appendix C. When comparing the first two interpretable models, there is a small trade-off to consider.: While Logistic Regression provides a direction of the relationship, making its explanation more informative, the decision tree offers a more complex understanding of the data. Even within the tree structures, one has to choose between more shallow trees, which are easier to interpret, and trees with more depth that provide better predictive results.

The figure 7.8 shows how InterpretML can be used to retrace individual local predictions, with the orange path indicating the prediction process. For visualization purposes, the figure depicts a refitted tree with a smaller depth. The Decision Tree Classifier automatically accounts for feature interactions through the splits. The visualization shows that as subsequent splits depend on nodes closer to the root, results always rely on a combination of features. This is further shown by the fact that features can show up multiple times in a tree, accounting for different types of interactions.

Feature	Score
PLBT	0.781
INDEPIND_-	0.081
GRMA	0.077
RD	0.013
IFAS	0.007
TFAS	0.006
ENVA	0.006
EXOP	0.006
SCT	0.005
EXEX	0.004
OPPL	0.003
TAXA	0.003
CRPE	0.002
NAT	0.002
EXPT	0.001

Table 7.2: Top 15 Predictor Variables Ranked by Importance in Decision Tree Classifier

7.3 Explainable Boosting Machine

7.3.1 Predictive Performance

The EBM (Explainable Boosting Machine) did not perform as well as the Decision Tree Classifier in classifying solvent companies, resulting in a weaker F1 score. The confusion matrix in Figure 7.6 shows a larger number of incorrectly classified companies compared to the decision tree classifier.
Although the EBM correctly classified 99.9% of the companies in the dataset, of the companies that the model predicted to be insolvent, only 83% actually were. The model identified 92.4% of all the insolvent companies in the dataset. The model's performance in classifying solvent companies was also good, with only 0.06% being incorrectly classified, which is lower than for the decision tree classifier.

Descriptive Performance

The 15 most important variables for predicting the classification of companies using the EBM model are shown in Figure 7.7, and as with the decision tree, PLBT appears to be the most significant predictor. It is worth noting that the importance values represent the mean absolute contribution or score that each feature or interaction makes to predictions averaged across the training dataset, with contributions weighted by the number of samples in each bin and the sample weights (Team, 2021). The mechanics behind interpreting EBMs are not as simple as with logistic regressions or decision trees, but with InterpretML, results become tractable for individual

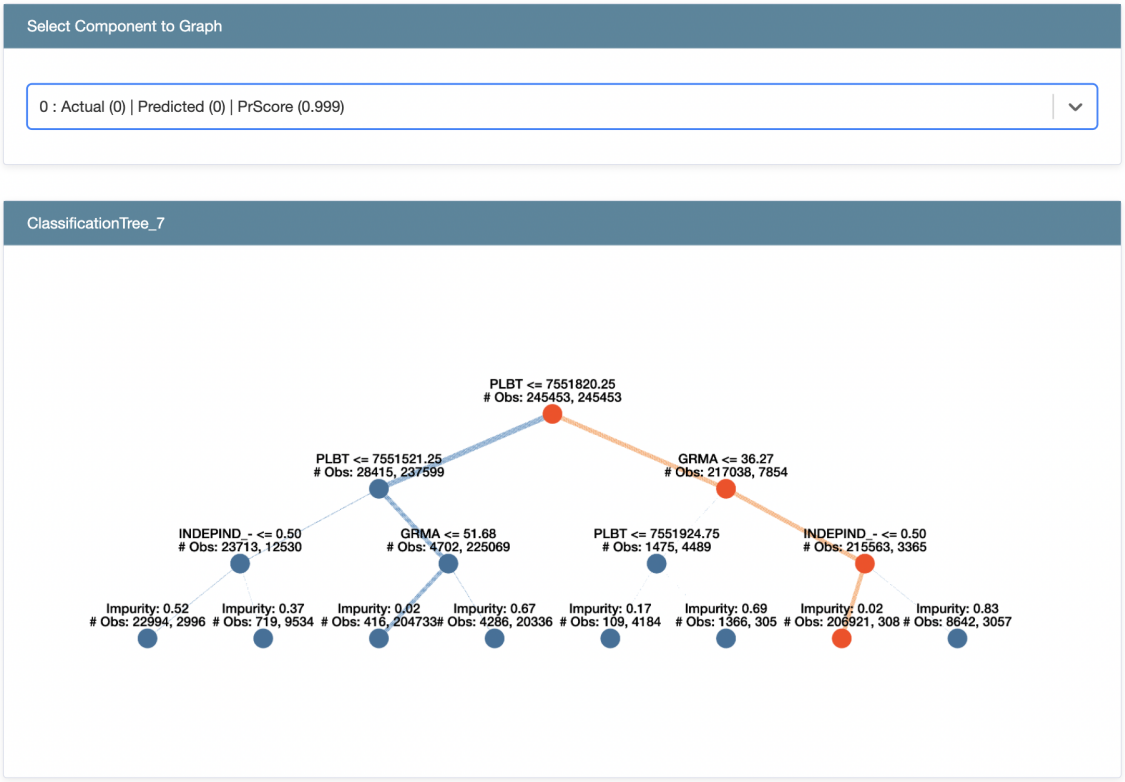

Figure 7.5: Decision Tree Classifier - Local Explanation Output generated using InterpretML

instances. Given that the internal mechanics account for feature interactions but are not easily interpretable, surrogate models would have to be used to comprehend relationships between variables, which do not directly utilize the EBM's internals. In conclusion, while the EBM is promoted as a "Glassbox Model", providing insights into its internals, this only relates to the fact that overall feature importances can be generated without using surrogate models in combination with predefined functions provided by InterpretML. Given that the documentation about how the model is constructed and the lower prediction power in comparison to the Decision Tree Classifier, there are might exist better model candidates to predict company default.

7.4 Random Forest

7.4.1 Predictive Performance

The Random Forest model was able to correctly classify 99.9% of all observations, only making one false positive classification. The recall value, however, tells us that only 81.6% of all companies classified as insolvent were actually insolvent. This also resulted in a F1 score of 94.9%. The predictive performance is shown in figure 7.13

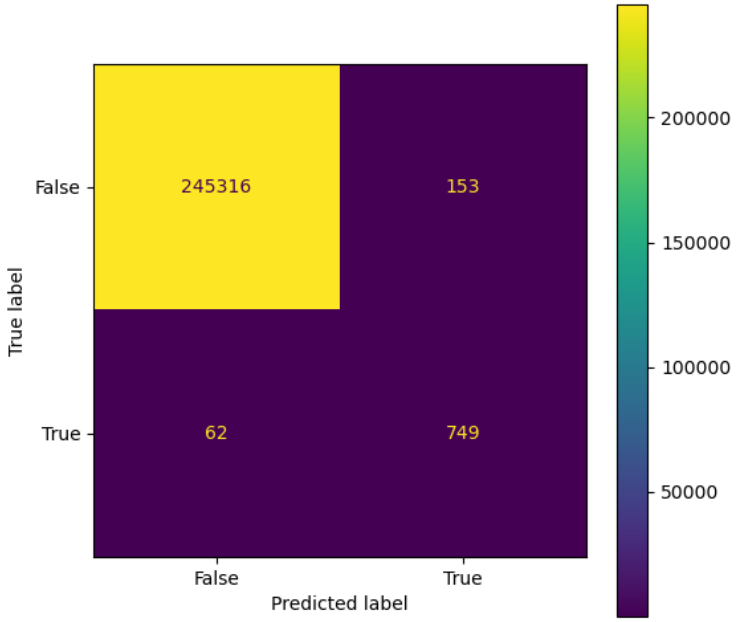

Figure 7.6: Performance Evaluation of Explainable Boosting Machine

Despite its use of multiple decision trees to combine outputs, the random forest model underperformed in comparison to the simple decision tree classifier. This could be due to overfitting, as the model achieved perfect accuracy during training but performed poorly during testing. This is a common issue when dealing with imbalanced datasets, which was the case in this study.
Although the parameter space included the optimal decision tree classifier parameters, more in-depth optimization is necessary to improve the model's performance. This could involve strategies such as adjusting the class weights or using resampling techniques to balance the dataset, or experimenting with different model architectures and hyperparameters.

7.4.2 Descriptive Performance

As a black box model, Random Forests do not give insight into model internals and require techniques such as LIME or SHAP to generate explanations.

LIME

InterpretML provides an easy-to-implement approach to LIME, returning visualized results of which variables were used for individual predictions. Figure 7.10 shows the output for a misclassified instance, where the model falsely predicted solvency.

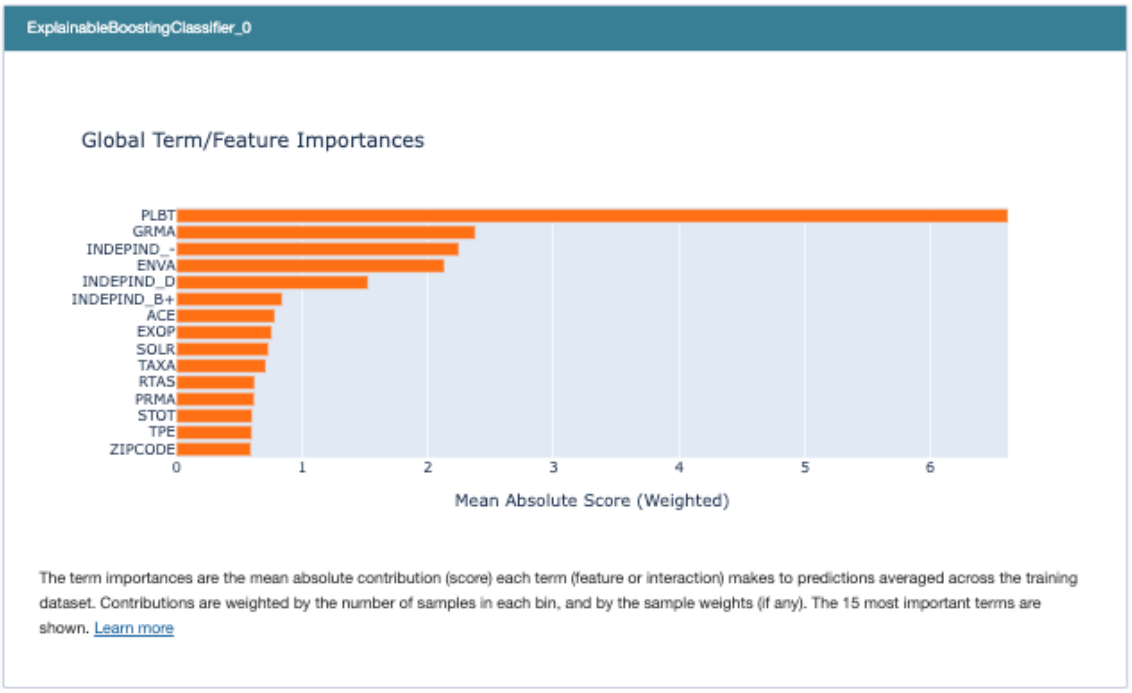

Figure 7.7: Top 15 Predictor Variables Ranked by Absolute Importance in Explainable Boosting Machine

It is evident that the prediction results were very close. According to LIME, the profit drove predictions more towards solvency, while the credit rating contributed largely to the company almost being classified as insolvent.

SHAP

The SHAP package, like LIME, enables the creation of local explanations. A significant advantage of using SHAP for analyzing model predictions is the availability of intuitive visualizations. As shown in figure 7.11, SHAP allows for the identification of the significant role played by the gross margin (GRMA) in the classification of insolvency, while the company's profit (PLBT) drove the prediction towards insolvency.

Although insights gained from local explanations are valuable in justifying mistakes and identifying areas for improvement, it is concerning that SHAP and LIME provide different results for local explanations, rendering the results unreliable. One potential cause of this discrepancy is that LIME assumes local linearity, while SHAP does not have this requirement. Additionally, the techniques rely on different sampling methods to generate the data used for fitting the local model.

Unlike LIME, SHAP can also be used to show global feature importance. With almost 250 000 instances in the training data set, a reduced model was used based on a smaller sample and using approximation techniques to drive down run time, sacrificing some of the accuracy. Figure 7.16 shows that the most important fea-

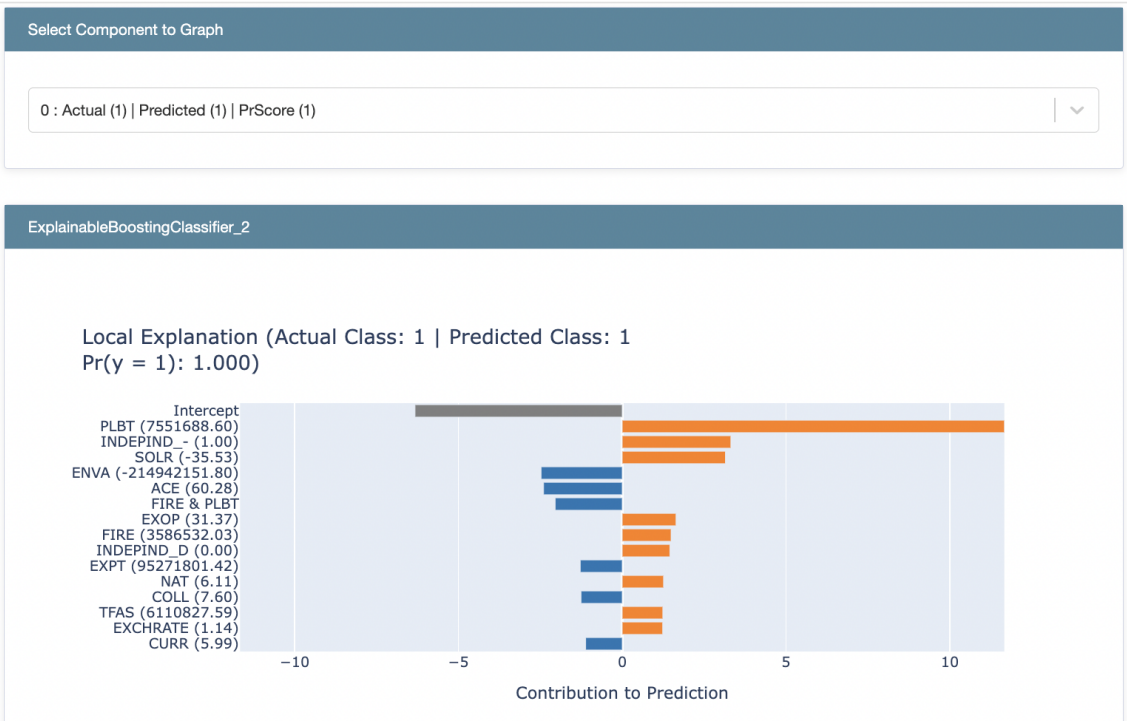

Figure 7.8: Explainable Boosting Machine - Local Explanation Output generated using InterpretML

tures have both positive and negative effects on the model output, with the blue part of the bars indicating the magnitude of the feature's impact driving it towards solvency, while red indicates an impact towards insolvency predictions.
SHAP also enables the analysis of feature interactions in a given model. If any domain-specific assumptions are present or if the model contains fewer features than the current analysis, these interactions can be visualized effectively.

7.5 Neural Network

7.5.1 Predictive Performance

The Neural Network model achieved an accuracy of 99.98%. The precision score is 76.54%. Out of all the companies classified as insolvent, only one was actually solvent. The F1 score was 93.14%, which given the optimization for the F1 score was not the highest. Nevertheless, due to its strength in predicting insolvency, it can be viewed as the most powerful model and it can be concluded that on a predictive level, deep learning models show the best performance in this particular analysis.

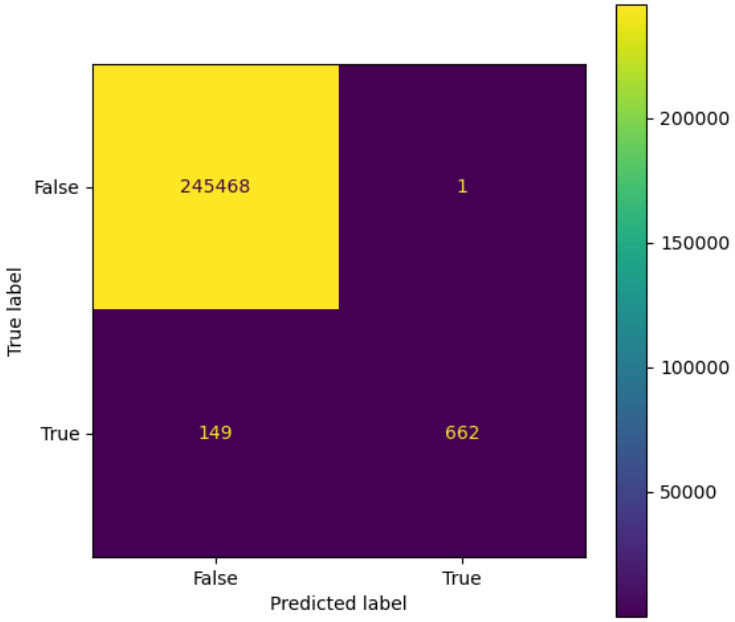

Figure 7.9: Performance Evaluation of Random Forest

7.5.2 Descriptive Performance

LIME

Just as for the Random Forest model, LIME was used to generate an explanation of a mispredicted instance. In this case for analyzing the case where a company was falsely classified as insolvent. LIME suggests that the independent credit rating was the main driver for the false prediction. The generated output is shown in figure 7.14.
For this particular case, it would be helpful to analyze if potential feature interactions drove the false classification, as for example, the company size drove the classification more towards solvency. This however is not possible using LIME.

SHAP

The local explanation generated using SHAP also identifies company size and credit rating as the primary prediction drivers. However, it yields different results regarding the strength of these predictors compared to the local explanation generated using LIME. This discrepancy can again be potentially attributed to LIME's linearity assumption and the different sampling techniques used in both approaches.
Additionally, SHAP identifies other predictors that were not identified by LIME, indicating that SHAP may be better at capturing non-linear relationships between

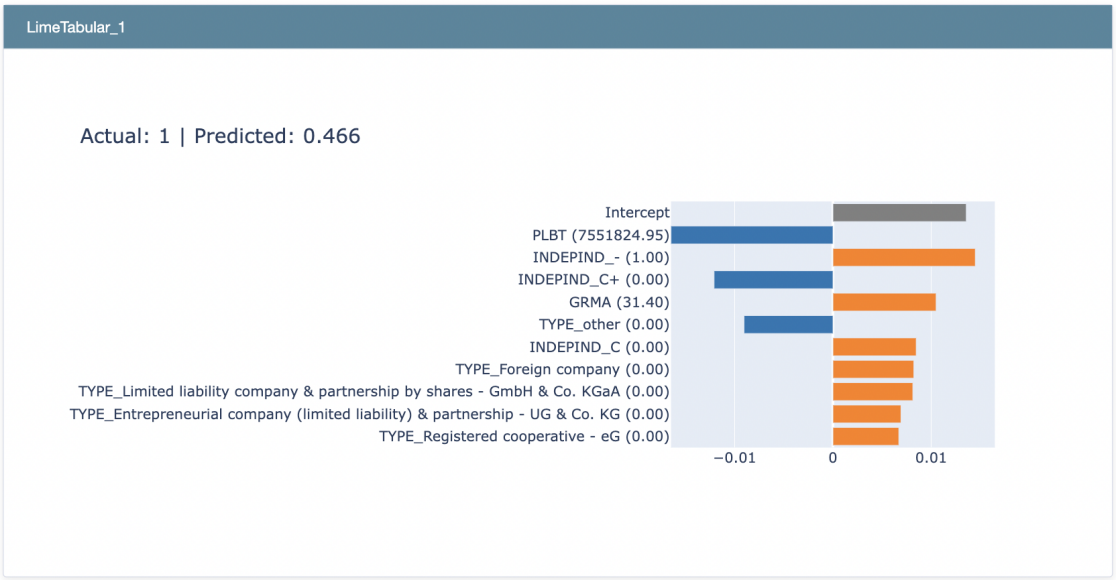

Figure 7.10: Local explanation of an Insolvent Company classified as Solvent Value using LIME (Random Forest)

Figure 7.11: Local Explanation of an insolvent company classified as solvent value using SHAP (Random Forest)

features. The differing results between the two methods highlight the importance of comparing and evaluating the outputs of different local explanation methods before drawing conclusions about feature importance and their relationship with the model's predictions.

SHAP was also used to generate insights on the global feature importance, which is shown in figure 7.15. These results should be treated with caution, as it is based on a sample of the data, which is likely imbalanced and potentially favoring the majority class. It indicates, that enterprise value (ENVA) was by far the most important overall prediction driver for both classes.

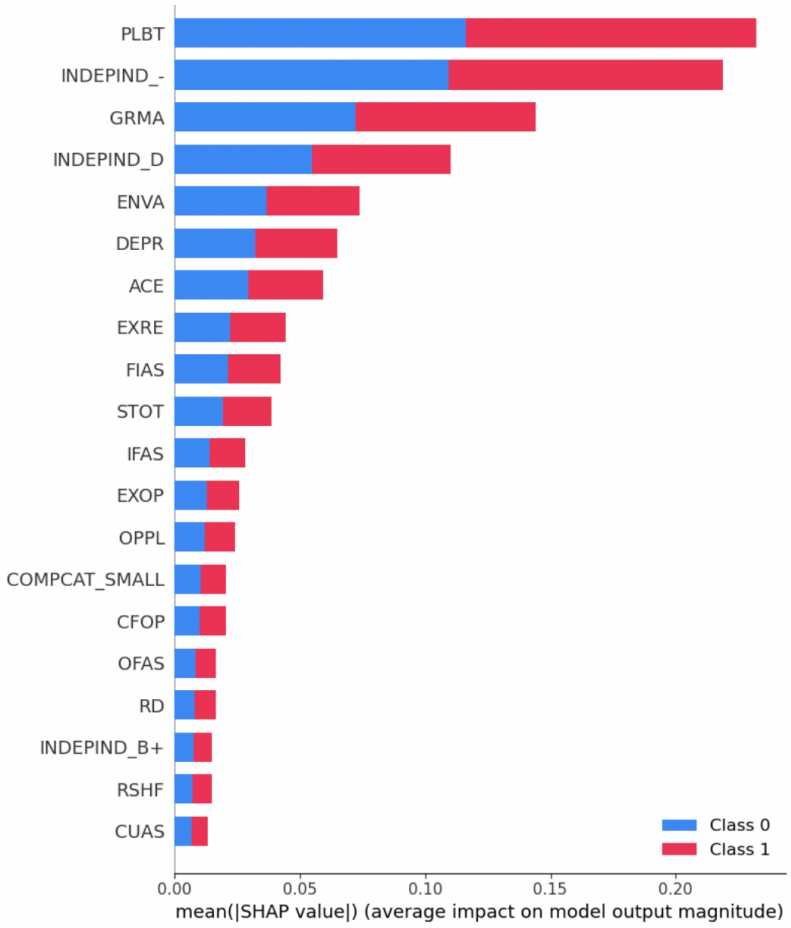

Figure 7.12: Global Explanation of a random forest using SHAP

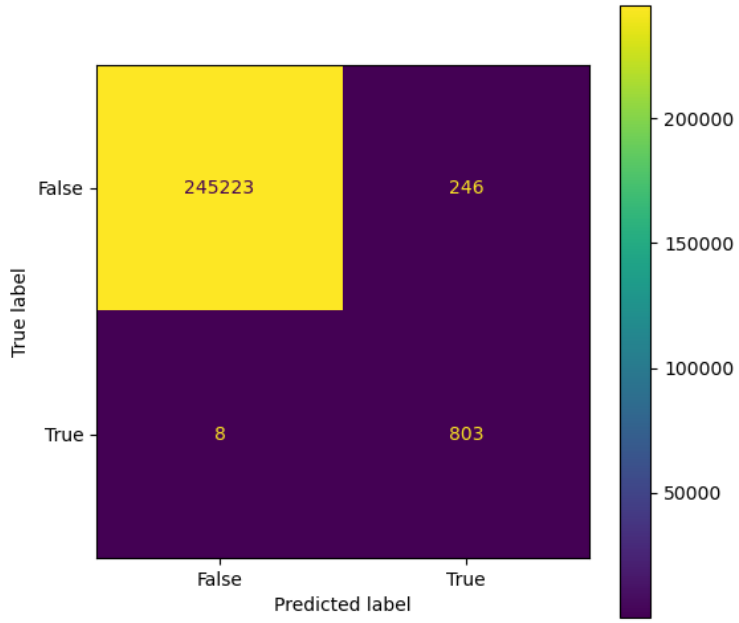

Figure 7.13: Performance Evaluation of Neural Network

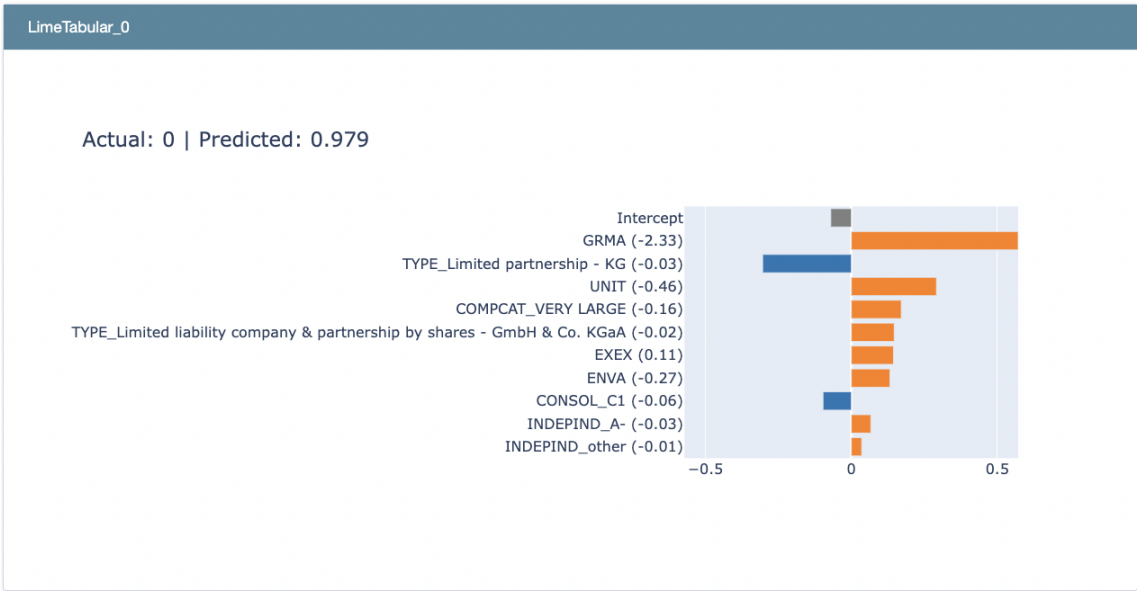

Figure 7.14: Local explanation of an Insolvent Company classified as Solvent Value using LIME (Multi-Layer Perceptron)

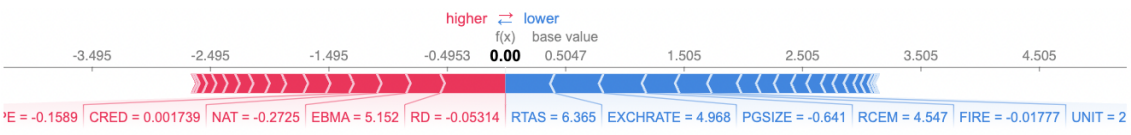

Figure 7.15: Local Explanation of an insolvent company classified as solvent value using SHAP (Multi-Layer Perceptron)

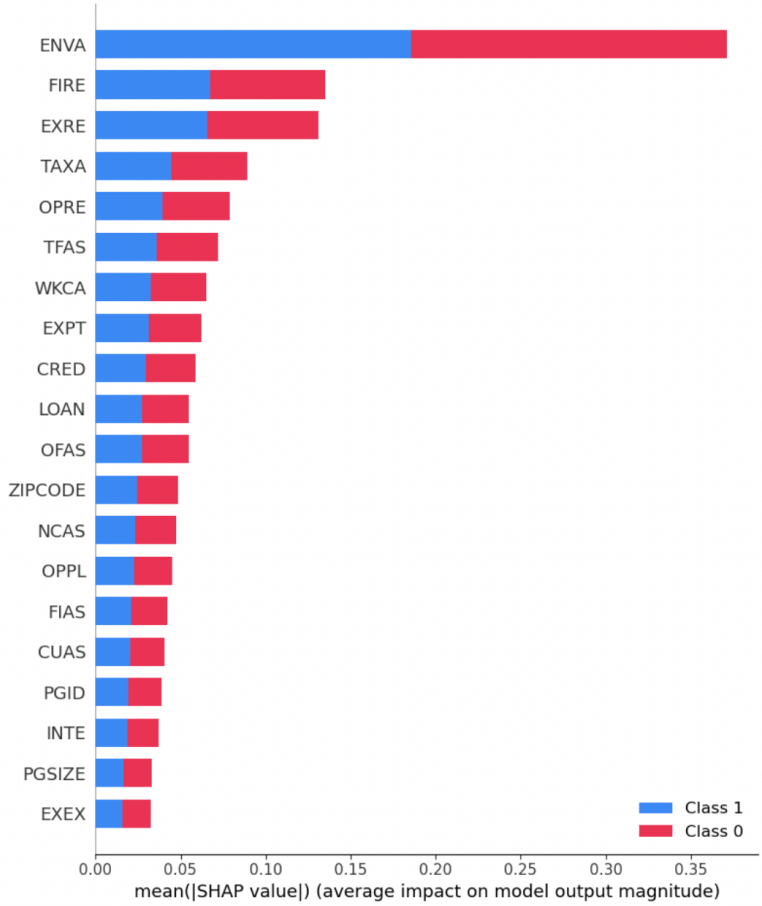

Figure 7.16: Global Explanation of a Multi-Layer Perceptron using SHAP

Chapter 8

Conclusion

8.1 Key Findings

This study has shown that although machine learning models have many potential use cases in finance and economics, their black-box nature limits widespread usage in fields where explanations are critical. To address this issue, explainability techniques can be employed. The choice of a reliable method to assess credit risk should be based on the application's specific requirements, as models behave differently based on the dataset.

Explainability techniques can aid the model development process by assisting in data preprocessing and variable selection, thereby detecting and correcting biases in training datasets and ensuring that only meaningful variables influence the output of the model. They can also help narrow down important predictors and improve model performance by dealing with noisy data, outliers, and redundant features.Barredo Arrieta et al. (2020)

When it comes to explaining black box models, LIME and SHAP are two popular techniques, each with their own strengths and weaknesses. LIME is easy to implement and computationally efficient, making it suitable for big datasets. However, it can only produce local explanations tailored to individual observations, which may not be generalizable. On the other hand, SHAP can produce global explanations and explicitly consider feature interactions. It requires high computational power to handle the data well but provides a more complete picture of the model. Understanding and retracing the surrogate model calculations can be more difficult than with LIME. Therefore, the choice of the right explainability method depends on the specific use case.

One concerning finding in this application is that LIME and SHAP generate different results, which makes it unsuitable for ethical or legal requirements. It cannot be ruled out, that these differences occurred to the limited availability of computing

power.
It should also be noted that while XAI techniques offer a first step towards understanding the internals of black box models, they cannot extract causal relationships, verify the quality of predictors, or distinguish between causal and non-causal effects Murdoch et al. (2019).
In summary, explainability techniques can be a first step towards meeting ethical or regulatory requirements in finance and economics, but their implementation requires a close inspection of the model's internals, including how variables might interact. Moreover, preprocessing and variable selection plays a crucial role in improving model performance, especially for large datasets with many potential predictors. Finally, choosing the right explainability method requires careful consideration of the tradeoffs between local and global explanations, interpretability, and computational efficiency.

8.2 Limitations and Further Research

A major drawback was the limited availability of computing power. This affects the efficiency of tuning models and therefore the predictive accuracy. It also affects to which extent global explanations can be produced. In a professional context, better computers usually eliminate this issue.
The study presents insights into the application of machine learning models to predict bankruptcy in a specific context, but there are several ways in which this research can be extended to provide more comprehensive results: To further explore the robustness of the models, they should be validated using different years, both attempting to predict bankruptcy further in the future and for different years, and potentially different geographical areas.
Further This research only made predictions based on one year of predictors, as dealing with several time periods introduces the issue of time-dependent data making the model very complex. While there exist robust approaches to combine data from multiple years and company characteristics to predict a company´s financial performance in the future, such approaches often require both strong economic assumptions and result in highly opaque predictive models, such as conducted in Brandt et al. (2009).

There are several key areas for further research in the field of Explainable AI in general, that affect its applicability to the economic field: Firstly, there is a lack of technical guidelines for implementing reliable models that adhere to ethical and regulatory standards. Interdisciplinary collaboration is needed to ensure appropriate techniques are selected for each use case and audience, as explainability can be

manifested in many ways, such as visuals, text, and more.
Secondly, it is important to apply these models to consumer lending and develop benchmarks for fairness, to assess which variables should be excluded, especially under the consideration of feature interactions.
Thirdly, the absence of a standardized measure for quantifying 'Explainability' makes it difficult to compare different techniques without knowledge of the true underlying economic relationships.

Bibliography

Addo, P. M., Guegan, D., and Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38.

Al-Qutaish, R. (2020). Integration of the github platform in teaching software engineering: A systematic literature review. *Journal of Information Technology Education: Research*, 19:341–363.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., and 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.

Amadeus (2023). Amadeus database. https://amadeus.com/en/insights/data-services/travel-data-services/amadeus-database. Accessed on April 8, 2023.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Bazarbash, M. (2019). *Fintech in financial inclusion: machine learning applications in assessing credit risk*. International Monetary Fund.

Bhatore, S., Mohan, L., and Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4(1):111–138.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2020). Explainable ai in fintech risk management. *Frontiers in Artificial Intelligence*, 3:26.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1):203–216.

Cafiso, G. (2022). Loans to different groups and economic activity at times of crisis and growth. *Oxford Bulletin of Economics and Statistics*, 84(3):594–623.

Cao, L. (2020). Ai in finance: A review. *Available at SSRN 3647625*.

Chen, Z., Tan, S., Nori, H., Inkpen, K., Lou, Y., and Caruana, R. (2022). Using explainable boosting machines (ebms) to detect common flaws in data. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, pages 534–551. Springer.

Commission, E., Directorate-General for Communications Networks, C., and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2018). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 555–567.

Fuhrman, J. D., Gorre, N., Hu, Q., Li, H., El Naqa, I., and Giger, M. L. (2022). A review of explainable and interpretable ai with applications in covid-19 imaging. *Medical Physics*, 49(1):1–14.

für Justiz, B. (2020). Insolvenzbekanntmachungen. https://www.insolvenzbekanntmachungen.de/. Accessed on: April 15, 2023.

Garreta, R. and Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Hacker, P. and Passoth, J. (2021). Varieties of ai explanations under the law. *From the GDPR to the AIA, and Beyond. Available online: https://papers. ssrn. com/-sol3/papers. cfm.*

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., and Ramkumar, P. N. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine*, 13:69–76.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W. (2022a). *xxAI - Beyond Explainable Artificial Intelligence*, pages 3–10. Springer International Publishing, Cham.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022b). *Explainable AI Methods - A Brief Overview*, pages 13–38. Springer International Publishing, Cham.

Ij, H. (2018). Statistics versus machine learning. *Nat Methods*, 15(4):233.

Japkowicz, N. (2006). Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*, pages 6–11.

Krainer, J. (2005). Assessing and managing credit risk: A primer. *Economic Review-Federal Reserve Bank of Atlanta*, 90(3):1–18.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Langenbucher, K. (2020). Responsible ai-based credit scoring–a legal framework. *European Business Law Review*, 31(4).

Langenbucher, K. and Corcoran, P. (2021). Responsible ai credit scoring–a lesson from upstart. com. In *Digital Finance in Europe: Law, Regulation, and Governance*, pages 141–180. De Gruyter.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lundberg, S. M. and Lee, S.-I. (2020). SHAP: Shapley additive explanations. https://github.com/slundberg/shap.

Mi, J.-X., Li, A.-D., and Zhou, L.-F. (2020). Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, 8:191969–191985.

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., and Lin, S. F. (2021). Explainable ai in credit risk management. *arXiv preprint arXiv:2103.00949*.

Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Morocho-Cayamcela, M. E., Lee, H., and Lim, W. (2019). Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions. *IEEE access*, 7:137184–137206.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2017). Interpretml: A unified framework for machine learning interpretability. *Advances in Neural Information Processing Systems*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rogers, Y., Sharp, H., and Preece, J. (2015). *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2):10–1162.

Smuha, N. A. (2019). The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4):97–106.

Team, I. D. (2021). Explainable boosting machine. https://interpret.ml/docs/ebm.html. Accessed: April 11, 2023.

Torrent, N. L., Visani, G., and Bagli, E. (2020a). Psd2 explainable ai model for credit scoring. *arXiv preprint arXiv:2011.10367*.

Torrent, N. L., Visani, G., and Bagli, E. (2020b). Psd2 explainable ai model for credit scoring. *ArXiv*, abs/2011.10367.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Zhang, J., Ji, P., Wang, Y., and Guo, Y. (2018). An introduction to multilayer perceptron algorithm and its applications in clinical decision-making. *Journal of Healthcare Engineering*, 2018:1546246.

Appendix A

Hyperparameter Tuning

Parameter Spaces

Logistic Regression

- Penalty: L2, L1, Elasticnet
- C: np.linspace(0, 1000, 50)
- Solver: newton-cg, lbfgs, liblinear, sag, saga

Parameters after Optimization: C=0.001, penalty='none', solver='lbfgs'

Decision Tree

- 'max depth': [2, 3, 4, 5, 6, 7, 8, 9, 10, None]
- 'min samples split': [2, 5, 10, 15, 20, 25]
- 'min samples leaf': [1, 2, 4, 8, 16]
- 'max features': ['sqrt', 'log2', None]
- 'criterion': ['gini', 'entropy']

Parameters after Optimization:
min samples split': 25, 'min samples leaf': 1, 'max features': None, 'max depth': None, 'criterion': 'entropy'

Explainable Boosting Machine

- 'max bins': stats.randint(5, 50),
- 'max interaction bins': stats.randint(1, 10),
- 'interactions': stats.randint(0, 3),

- 'outer bags': stats.randint(1, 5),
- 'inner bags': stats.randint(1, 5),
- 'learning rate': stats.loguniform(0.001, 0.5),
- 'validation size': stats.uniform(0.1, 0.3)

Parameters after Optimization:
'inner bags': 3, 'interactions': 2, 'learning rate': 0.04777461250491698, 'max bins': 42, 'max interaction bins': 8, 'outer bags': 1, 'validation size': 0.11362241189930644

Random Forest

- 'n_estimators': [100, 200, 500, 1000]
- 'max_depth': [None, 5, 10, 20]
- 'min_sample_split': [2, 5, 10]
- 'min_samples_leaf': [1, 2, 4]
- 'bootstrap': [True, False]

Parameters after Optimization: n_estimators': 1000, 'max_depth': None , 'min_sample_split': 2 ,'min_samples_leaf': 1 , 'bootstrap': True

Multilayer Perceptron

- 'hidden layer sizes': [(10,), (50,), (100,)],
- 'activation': ['identity', 'logistic', 'tanh', 'relu'],
- 'alpha': [0.0001, 0.001, 0.01, 0.1],
- 'learning rate': ['constant', 'adaptive']

Parameters after Optimization:
'learning_rate': 'adaptive', 'hidden_layer_sizes': (100,), 'alpha': 0.0001, 'activation': 'relu'

Appendix B

Variable Descriptions

- IDNR: Identification number of the company
- REPBAS: Basis of preparation of financial statement (unconsolidated in this case)
- CLOSDATE: Closing date of the fiscal year
- CLOSDATE_year: Year of the closing date
- ACCPRA: Accounting principles used to prepare the financial statement (e.g. Local GAAP)
- MONTHS: Number of months the financial statement covers
- UNIT: Monetary unit used in the financial statement
- EXCHRATE: Exchange rate used to convert monetary units
- EXCHRATE2: Another exchange rate used to convert monetary units
- FIAS: Fixed assets
- IFAS: Intangible fixed assets
- TFAS: Total fixed assets
- OFAS: Other fixed assets
- CUAS: Current assets
- STOK: Stocks
- DEBT: Debts
- OCAS: Other current assets

- CASH: Cash
- TOAS: Total assets
- SHFD: Shareholders' funds
- CAPI: Capital and reserves
- OSFD: Other shareholders' funds
- NCLI: Non-current liabilities
- LTDB: Long-term debt
- ONCL: Other non-current liabilities
- PROV: Provisions
- CULI: Current liabilities
- LOAN: Loans
- CRED: Creditors
- OCLI: Other current liabilities
- TSHF: Total shareholders' funds and liabilities
- WKCA: Working capital
- NCAS: Net current assets
- ENVA: Environmental assets
- EMPL: Number of employees
- OPRE: Operating revenue
- TURN: Turnover
- COST: Costs
- GROS: Gross profit
- OOPE: Operating profit or loss
- OPPL: Profit or loss before taxation
- FIRE: Financial revenue

- FIEX: Financial expenses
- FIPL: Profit or loss from financial activities
- PLBT: Profit or loss before taxation
- TAXA: Taxation
- PLAT: Profit or loss after taxation
- EXRE: Extraordinary revenue
- EXEX: Extraordinary expenses
- EXTR: Profit or loss from extraordinary activities
- PL: Profit or loss
- EXPT: Extraordinary profit or loss
- MATE: Material costs
- STAF: Staff costs
- DEPR: Depreciation and amortization
- INTE: Interest received
- RD: Research and development expenses
- CF: Cash flow
- AV: Asset value
- EBIT: Earnings before interest and taxes
- EBTA: Earnings before taxes and amortization
- RSHF: Return on shareholders' funds
- RCEM: Return on capital employed
- RTAS: Return on total assets
- CFOP: Cash flow from operating activities
- PRMA: Profit margin
- GRMA: Gross Margin

- ETMA: Earnings before Taxes and Minority Interest
- EBMA: Earnings Before Interest, Taxes, Depreciation, and Amortization
- NAT: Net Assets
- IC: Inventory
- STOT: Short-Term Debt
- COLL: Collateral
- CRPE: Credit Period
- EXOP: Export Operations
- CURR: Currency
- LIQR: Liquidity Ratio
- SHLQ: Shareholder's Equity
- SOLR: Solvency Ratio
- GEAR: Gearing Ratio
- PPE: Property, Plant, and Equipment
- TPE: Total Property, Plant, and Equipment
- SCT: Sales and Cost of Goods Sold
- ACE: Average Collection Efficiency
- SFPE: Sales per Fixed Asset
- WCPE: Working Capital per Employee
- TAPE: Total Assets per Employee
- CURRENCY: Currency Code
- NAT_PRIM_CODE: National Primary Industry Code
- NACE_PRIM_CODE: NACE Primary Industry Code
- NAICS_CORE_CODE: NAICS Core Industry Code
- USSIC_CORE_CODE: USSIC Core Industry Code

- SD_ISIN: ISIN Code
- SD_TICKER: Ticker Symbol
- ACCNR: Accounting Number
- OFNBR: Official Number
- NAME: Company Name
- NAME_NAT: Company Name in National Language
- ADDRESS: Company Address
- ADDRESSv_NAT: Company Address in National Language
- ZIPCODE: Company Postal Code
- CITY: Company City
- CITY_NAT: Company City in National Language
- REGION: Company Region
- REGION_NAT: Company Region in National Language
- COUNTRY: Company Country
- CNTRYCDE: Company Country Code
- PHONE: Company Phone Number
- FAX: Company Fax Number
- EMAIL: Company Email Address
- LSTATUS: Legal Status
- STATUSDATE: Status Date
- TYPE: Company Type
- DATEINC: Date of Incorporation
- DATEINC_year: Year of Incorporation
- QUOTED: Stock Exchange Listing Status
- REPBAS_HEADER: Reporting Basis Header

- CONSOL: Consolidated Status
- TYPACC: Type of Accounts
- LASTYEAR: Last Available Financial Year
- NUMYEARS: Number of Financial Years Available
- UPDNUM: Update Number
- IPSOURCE: Data Source
- INDEPIND: Independence Indicator
- COMPCAT: Company Category
- PGID: Profit and Loss Group ID
- PGNAME: Profit and Loss Group Name
- PGDESC: Profit and Loss Group Description
- PGSIZE: Profit and Loss Group Size
- DATEINC_char: Date of Incorporation in Character Format

Appendix C

Graphics

Decision Tree Classifier

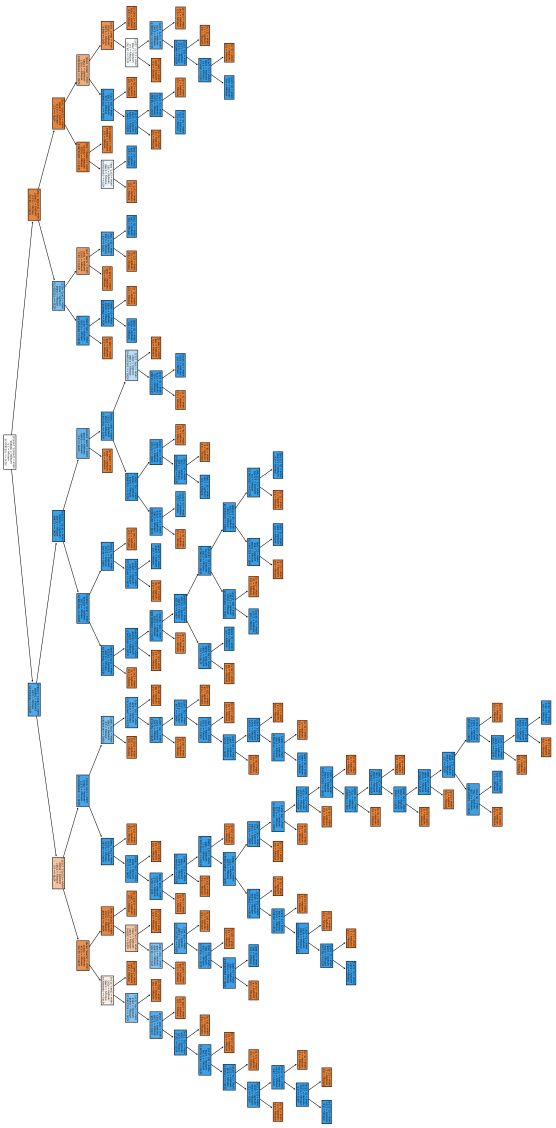

Figure C.1: Fitted Decision Tree Model

Appendix D

Models fitted on Webscraping Data

Logistic Regression

- Accuracy: 0.8012377650002195
- Precision 0.005703715122552798
- Recall 0.6590330788804071
- False Postive 0.1985165122649349
- F1 Score 0.4504112515340299

Decision Tree Classifier

- Accuracy: 0.9963613220383619
- Precision 0.02813852813852814
- Recall 0.03307888040712468
- False Postive 0.0019741730677066615
- F1 Score 0.5142932987603259

Explainable Boosting Machine

- Accuracy: 0.9963613220383619
- Precision 0.02813852813852814
- Recall 0.03307888040712468
- False Postive 0.0019741730677066615

- F1 Score 0.5142932987603259

Random Forest

- Accuracy: 0.9982618619145854
- Precision 0.0
- Recall 0.0
- False Postive 1.3190465931224911e-05
- F1 Score 0.4995650875096647

Neural Network

- Accuracy: 0.9816310406882325
- Precision 0.017557251908396947
- Recall 0.17557251908396945
- False Postive 0.01697612965348646
- F1 Score 0.5113249128389252