UNIVERSITY OF TWENTE

MASTER THESIS

Stakeholder-centric approach to applying machine learning to probability of default models

Author:
Dyon KOK

University Supervisors:
J.O.R. Osterrieder
Dr. M.R. Machado
Company Supervisors:
Leon Dusee, MSc
Dr. Markus Haverkamp

UNIVERSITY OF TWENTE.

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the field of

Financial Engineering and Management

March 6, 2024

Abstract

Dyon KOK

Stakeholder-centric approach to applying machine learning to probability of default models

The financial sector's reliance on Probability of Default (PD) models necessitates an equilibrium between predictive accuracy and interpretability, a balance that is pivotal for maintaining stakeholder trust and adherence to regulatory standards. This thesis rigorously examines the integration of Explainable Machine Learning (XML) into PD modeling, focusing on two advanced techniques: the Explainable Boosting Machine (EBM) and the Generalized Additive Models with Interactions Network (GAMINet). Through a comprehensive comparative analysis, this research evaluates the models against a traditional logistic regression benchmark used within a financial institution, assessing them on explainability, regulatory compliance, performance, and operational feasibility.

The methodology encompasses a detailed examination of model architecture, hyperparameter optimization, and cross-validation processes to ensure robust model evaluation. Despite the anticipation of superior performance, the findings reveal that the newly implemented XML models did not outperform the existing system. However, the study illuminates the significant potential of XML techniques in enriching PD models by augmenting their interpretability without compromising predictive capability. Notably, the research delineates the nuances of each model's approach to balancing complexity with interpretability, highlighting EBM's straightforwardness and GAMINet's capacity to model intricate interactions.

A critical aspect of this thesis is its stakeholder-centric analysis, emphasizing the importance of model transparency and the logical rationale behind feature selection in fostering stakeholder acceptance. The comparative evaluation offers insights into the practical implications of deploying EBM and GAMINet within financial institutions, considering computational resource constraints and the integration challenges with existing IT systems. Furthermore, the thesis discusses the ethical considerations and the models' capacity for bias management, underscoring the necessity of ethically sound modeling practices in financial risk assessment.

Concluding with a forward-looking perspective, the thesis proposes strategic recommendations for the practical integration of XML techniques into PD modeling practices, advocating for a balanced approach that leverages EBM for its efficiency and GAMINet for its analytical depth. Additionally, it outlines promising avenues for future research, including the exploration of Python-SAS integration methods and the potential for advanced feature transformation techniques to unlock new dimensions of model accuracy and interpretability. This research contributes to the evolving discourse on the application of machine learning in finance, aiming to bridge the gap between technological advancement and the sector's regulatory and ethical imperatives.

Acknowledgements

I'd like to start by expressing my gratitude to everyone who took the time to read my thesis, which has played a part in my graduation journey. A big thank you goes out to Leon for believing in me and offering me the chance to work at the financial institution. I'm also grateful to Markus for his valuable feedback and advice on the model development. A heartfelt thanks to all the people who agreed to be interviewed, making time for my questions. Special thanks to Francesco for shedding light on the PD model development process.

I also want to thank my supervisors for their supportive feedback and trust in my abilities.

Lastly, I couldn't have reached this point without the unwavering support of my family and friends, who have all contributed to this achievement in their own ways.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ALE	**A**ccumulated **L**ocal **E**ffects
EAD	**E**xposure **A**t **D**efault
DSRM	**D**esign **S**cience **R**esearch **M**ethod
GAM	**G**eneral **A**lgebraic **M**odel
GDPR	**G**eneral **D**ata **P**rotection **A**ct
IRB	**I**nternal **R**atings- **B**ased
LGD	**L**oss **G**iven **D**efault
LIME	**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
PD	**P**robability of **D**efault
PDP	**P**artial **D**ependency **P**lots
SHAP	**SH**apley **A**dditive ex**P**lanations
XAI	e**X**plainable **A**rtificial **I**ntelligence
XML	e**X**plainable **M**achine **L**earning

1 Introduction

1.1 Research context

1.1.1 Background

Banks have always been central to the economy's health (Gobat, 2012b), bridging the gap between savers and borrowers. As these financial institutions consistently expand their loan portfolios, the inherent risk of their decisions amplifies. The primary role is to take in funds—called deposits—from those with money, pool them, and lend them to those who need funds. Regulations are generally designed to limit banks' exposures to credit, market, and liquidity risks and to overall solvency risk. The bank therefore seeks a balance in making money and also mitigating the risks that come with this (Gobat, 2012a).

However, the journey of striking the balance between lending and mitigating the risk hasn't always been smooth. The 2008 financial crisis unveiled the severe consequences of inadequate risk assessment. Banks, including big financial institutions in the Netherlands, faced tumultuous times and liquidity positions dwindling. This crisis accentuated the imperative of maintaining an appropriate risk rating scenario within a bank's portfolio (BIS, 2017; IMF, 2018).

In response to the turmoil, the European Central Bank (ECB)[1] proposed new guidelines for risk rating models, emphasizing the requisite liquidity standards for banks (European Central Bank., 2019). Around this period, automation and Machine Learning (ML) change various industries, from self driving cars to automation in manufacturing (Schlicht, 2023; Vermesan, 2022) . Yet, the regulatory modelling area of the finance sector, despite being an information-rich industry, lagged in harnessing the full potential of these technologies. The European Banking Authority (EBA) published a discussion paper on the application of machine learning across the industry, structured around guidelines that reflect the EBA's viewpoint on its usage and inquiries regarding how different institutions implement it (EBA, 2021). The feedback received in response to this paper was meticulously examined and encapsulated in a subsequent report (EBA, 2023). This analysis reveals that, although a number of institutions are delving into machine learning, reservations persist about the methodologies employed in its application.

In recent times, the banking sector has witnessed a significant transformation with the integration of automation into numerous operational processes. Modern innovations such as banking chatbots, automated financial reporting, and advanced portfolio analysis are predominantly powered by sophisticated algorithms (Ortaköy and Özsürünç, 2019). This technological shift has contributed to the gradual obsolescence of physical bank branches, as more services migrate online, leveraging automated systems that draw on extensive internal data sources for processing without

[1] https://www.ecb.europa.eu/home/html/index.en.html

the need for human intervention (Clercq, 2023). While studies highlight the efficiency of these systems, noting their reduced tendency for errors like decimal misplacements compared to human operators, a persistent skepticism towards machine-made decisions remains a notable challenge (Association, 2016).

As machine learning (ML) models become increasingly embedded within decision-making frameworks in banking, the necessity for transparency and understandability in these automated decisions escalates. This is where Explainable Artificial Intelligence (XAI) enters the picture, aiming to demystify the "black-box" nature of these algorithms. XAI endeavors to make the workings of ML models interpretable, ensuring their decisions align with economic standards and can be justified logically. The challenge, however, lies in the diversity of the audience: explaining an algorithm's decision-making process to a risk analyst demands a different approach than that used for a retail client. This distinction underscores the importance of developing XAI methods that not only enhance the transparency of ML applications but also tailor explanations to meet the varied needs and understanding levels of different stakeholders. The progression towards more explainable and user-centric ML models represents a crucial step forward in bridging the trust gap between automated systems and their human users, thereby facilitating more informed, transparent, and reliable decision-making in the banking sector.

1.1.2 Problem description

The integration of machine learning (ML) into various sectors is receiving attention, underscored by its significant potential to enhance predictive accuracy and efficiency across different sectors (Pugliese, Regondi, and Marini, 2021; Schröer, Kruse, and Gómez, 2021; Schlicht, 2023). Despite this, its adoption within regulatory credit risk models remains notably limited. Research has consistently demonstrated the superior performance of ML models over traditional approaches; however, their widespread use is hampered by concerns regarding transparency (Addo, Guegan, and Hassani, 2018; Commission, 2023a; Petersson, 2023). To navigate these challenges, an initial step involves conducting a comprehensive stakeholder analysis within a major financial institution, followed by examining the current reporting processes for the more interpretable models currently in use.

In the contemporary financial landscape, ML models, particularly those applied to Probability of Default (PD) prediction, have demonstrated remarkable predictive powers. PD models are crucial for assessing credit risk by determining the likelihood of a borrower defaulting on their commitments. While ML technologies have been shown to outperform some traditional models in terms of accuracy and efficacy, the complexity of these ML models, especially those based on deep learning, renders them "black boxes". Despite their accurate predictions, the opaque nature of these models makes it challenging for humans to understand the rationale behind their decisions, significantly impeding their acceptance and implementation in credit risk assessment (Langer et al., 2021).

Another significant barrier to their adoption is regulatory approval, particularly from the European Central Bank (ECB). The European Commission's recent discussion on AI in financial systems highlighted the use of ML in loan risk assessment as a high-risk activity. It raised concerns about ensuring thorough risk assessment, maintaining high-quality data to minimise risks and prevent discrimination, ensuring traceability through activity logs, providing detailed documentation for compliance, clear user information, appropriate human oversight, and maintaining secure and accurate systems. An anticipatory regulatory framework, expected to draw

parallels with the General Data Protection Regulation (GDPR), is projected to be released in early 2024, setting a comprehensive backdrop for AI applications in finance (Commission, 2023b).

The GDPR, effective since 2018, marked a significant milestone in data protection and privacy within the EU and EEA, aiming to give individuals more control over their personal data and simplifying the regulatory environment for international business. It introduced principles crucial for the financial sector, such as explicit consent for data processing, highlighting the intertwined nature of AI system operations with GDPR guidelines due to their reliance on extensive personal data (Commission, 2018).

The opacity of ML models raises substantial concerns among stakeholders, including credit officers, regulators, and borrowers, who require clear explanations for model outputs to ensure fairness, avoid discriminatory practices, and support decision-making processes. The quest for explainability extends beyond comprehension; it is pivotal for building trust, facilitating decision-making, and promoting model adoption. The obscurity of a model exacerbates the challenge of identifying and correcting biases or errors, potentially leading to unfair or inaccurate credit assessments (Meske et al., 2020; Langer et al., 2021).

Addressing these challenges transcends the development of explainable machine learning (XML) techniques; it involves integrating these techniques into PD models in a way that is insightful and acceptable to all stakeholders. Balancing model accuracy with explainability, meeting diverse stakeholder expectations, navigating communication nuances, and understanding the operational implications of adopting explainable AI (XAI) present intricate challenges that are central to this issue.

Thus, the question arises: How can we effectively balance the predictive capabilities of ML with the imperative need for transparency and explainability, especially in the context of PD prediction?

1.2 Research questions

The main research question discussed in this thesis is:

How can explainable machine learning be used for probability of default models while taking into account stakeholders requirements?

This thesis aims to evaluate the potential of machine learning models in credit risk assessment, focusing on not only surpassing the accuracy of traditional models but also ensuring their explainability for all stakeholders. Given that traditional approaches are well-established in this domain, the adoption of machine learning alternatives hinges on their ability to offer clear performance advantages while maintaining transparency and interpretability. This dual objective addresses the need for technological advancement in credit risk modeling, while also aligning with the regulatory and operational requirements of credit providing organizations. To answer the main research question, the research is divided into five sub-questions. These are listed below with a short motivation and categorized into qualitative and quantitative aspects.

Qualitative Research Questions

1. **How do stakeholders perceive the importance of explainability in machine learning models for PD (Probability of Default)?**

Understanding stakeholders' perspectives is crucial because it reveals the importance of transparency in model outcomes. This can also highlight potential resistance or support for the usage of such models.

2. **What are the challenges and concerns stakeholders associate with non-explainable versus explainable machine learning models in PD?**

   This question seeks to uncover pain points and issues stakeholders might have with machine learning models, shedding light on the potential advantages of explainable machine learning models.

3. **What are appropriate explainable machine learning models for application in PD models?**

   In order to propose a solution to the problem different approaches for the development of machine learning models need to be explored. To scope the research there will be looked into two machine learning approaches. Also because it is stated that "researchers need to be careful to avoid publishing conclusions that one method is better than another based only upon one data set covering a single time period". (Breeden, 2020)

Quantitative Research Questions

1. **How does the accuracy of an explainable machine learning model compare to a traditional model in predicting PD?**

   This is vital for establishing the trade-off, if any, between explainability and predictive accuracy. Stakeholders would want to know if emphasizing explainability sacrifices model quality.

1.3 Research methodology

The research methodolog of the thesis is the CRISP-DM method. The CRISP-DM method is a method to develop a model. In this case it is used to develop a machine learning model. Also the CRISP-DM method is a usefull methodology when creating an artifact (like a model) with several iterations loops. This is usefull for this research because feedback will be received on the model and improvements are made. In this case this will be the development of the explainable machine learning model for the probability of default. (Hotz, 2018)

1.3.1 CRISP-DM Stages

In this section, the different stages of the CRISP-DM are outlined to guide the thesis, Figure 3.9.

- **Business Understanding**: This foundational phase is about understanding the core problem. A deep dive into existing literature, especially into explainable machine learning applied to probability of default models. Simultaneously, understanding stakeholders' needs is vital, ensuring that the final model aligns with their requirements for assessing default probabilities.

- **Data Understanding**: Here, the available data sources are explored, and their suitability for addressing the identified business problem is assessed. A meticulous literature review provides insights into similar research and data sources.

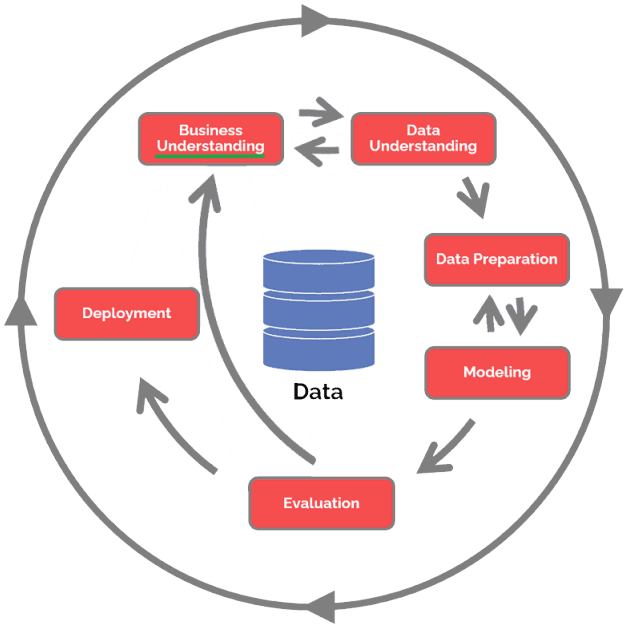


FIGURE 1.1: CRISP-DM Method
(Hotz, 2018)

It is a first step to determine the relevant data and understanding its characteristics, quality, and intricacies.

- **Data Preparation**: This stage focuses on converting raw data into a refined format suitable for modeling. It involves processes like data cleaning, transformation, and feature engineering. Aligning the data preparation with the stakeholders' requirements, as identified earlier, ensures that the resulting dataset is ready for modeling.

- **Modeling**: After preparing the data, the actual model tailored to determine the probability of default is designed and developed. The model's construction must ensure alignment with stakeholder expectations and address the problem highlighted in the Business Understanding phase.

- **Evaluation**: Once the model is in place, its efficacy is tested. Key metrics like accuracy and AUC gauge the model performance. In addition, the model results are compared against existing models used within the financial institution. To conduct robustness checks with cross-validation and synthetic datasets to confirm the model's reliability.

- **Deployment**: Upon satisfactory evaluation, the model is introduced into the business environment. The respective financial institution receives a comprehensive report explaining the model, its development process, and challenges faced. The model itself is not implemented into the financial institution due to time constraints.

Specific Objectives

This section outlines the specific objectives of the thesis, providing a foundation for the subsequent discussion of the research design, Table 1.1.

Literature

To conduct a comprehensive review of existing literature on both risk assessment in banking and the principles and methods of XAI. This will help identify current best practices, gaps, and opportunities for integration. This is also needed to narrow down the research and be able to scope the research down.

Stakeholder Analysis

To identify and understand the primary stakeholders in credit risk assessment. The aim is to elucidate their specific needs, preferences, and concerns regarding algorithmic explanations. This is needed to get a list of requirements for a probability of default model and what the stakeholders want to receive in terms of model explanation. It is not needed to flood them with information.

Model Examination

To evaluate machine learning models currently employed in risk assessments, focusing on the feasibility and effectiveness of applying existing XAI techniques.

Feedback Integration

To refine the XAI explanations based on stakeholder needs, ensuring the explanations are both clear and relevant to the target audience. This is done based on the stated requirements in the stakeholder analysis and the received feedback from the participants.

Recommendation Development

To formulate actionable recommendations for banks and financial institutions. These will guide the integration of XAI techniques in risk assessment models, ensuring both clarity and usefullness for stakeholders.

TABLE 1.1: Research design

Questions	**Objective**	**CRISP-DM Stage**	**Criteria**	**Data gathering**	**Deliverables**
How do stakeholders perceive the importance of explainability?	Understanding stakeholders' perspectives on transparency in model outcomes	Business understanding	Importance of explainability	Literature review, interviews	Interview results
What are the challenges with non-explainable vs explainable models?	Uncover challenges	Business understanding	Find challenges	Literature review	List of challenges to consider
What are appropriate explainable ML models for PD?	Identification of potential ML models	Data understanding and Modelling	Applicability in PD context.	Literature review	List of models
How does explainable ML model accuracy compare to traditional models?	Establish trade-off between explainability and predictive accuracy	Evaluation		Model testing and evaluation	Comparative analysis

1.4 Scope of the Research

This research investigates Explainable Artificial Intelligence (XAI) within the credit risk context, focusing on understanding stakeholder perspectives. It also explores the existing XAI literature, its fundamental principles, and applications across various industries, emphasizing the identification of gaps in understanding stakeholder preferences for XAI within the financial system.

The study then conducts a stakeholder analysis. It identifies key stakeholders in credit risk, such as credit analysts, borrowers, and regulators, and determines their needs and preferences regarding XAI explanations using a limited, representative sample from each group.

After the stakeholder analysis, the research examines representative machine learning models used in risk assessments to evaluate the suitability of applying XAI techniques. It implements basic XAI methods on these models to generate explanations, which it then presents to a subset of stakeholders. The research seeks feedback on the explanations' clarity, utility, and relevance.

The research will make adjustments based on this feedback to improve the clarity and utility of explanations. Notably, the study focuses on using existing XAI techniques rather than developing new ones. It covers two machine learning approaches, comparing them to the current model and stakeholder feedback. Due to the limited stakeholder sample size, the findings may not be universally generalizable. Additionally, the research does not extend to examining the broader impact of XAI explanations on actual lending decisions or the wider financial market.

1.5 Scientific Contribution

Studies focusing on probability of default modeling with the use of machine learning have gained significant attention in recent years. Researchers have explored various machine learning algorithms to predict the likelihood of default in different financial

scenarios. For instance (Obare, Njoroge, and Muraya, 2019) recommended the use of logistic regression in conjunction with supervised machine learning approaches for loan default prediction in financial institutions. They also suggested further research on ensemble methods to enhance prediction accuracy.

(Chong, Labadin, and Meziane, 2022) utilized a supervised machine learning model with logistic regression to predict the probability of default for loans funded through peer-to-peer lending platforms. Similarly, (Mor et al., 2022) employed a supervised machine learning algorithm-based logistic regression to predict loan default risk in the Indian commercial banking sector.

Furthermore, studies like (Coenen, Verbeke, and Guns, 2021) have extensively researched probability of default estimation using machine learning on historical data, particularly in credit risk modeling. Additionally, (Sifrain, 2023) employed machine learning methods such as decision trees, random forests, and bagging to analyze and determine significant factors in predicting default risk in peer-to-peer lending platforms.

Hence, it is evident that considerable research has been conducted in this field. However, previous studies have primarily concentrated on outcomes rather than the stakeholders involved. This study shifts the focus towards stakeholders, particularly emphasizing the crucial role of regulators. Another distinct aspect of this research is its concentration on the corporate default within financial institutions. While studies have explored this area, they have not specifically addressed stakeholder concerns or investigated which machine learning techniques might be most beneficial. Additionally, much of the existing research relies on publicly available data, which, while demonstrating the capabilities of machine learning methods, often overlooks the subtleties of using data from financial institutions. Consequently, conducting research based on data from financial institutions offers a unique opportunity to gain fresh insights into the application of machine learning techniques on such datasets.

1.6 Structure of thesis

The structure of this thesis is organized into six chapters, each focusing on an aspect of the research. Chapter 2 discusses the relevant literature, providing an overview of credit risk, machine learning, and the principles of explainable machine learning. Chapter 3 presents the findings from stakeholder interviews, offering insights into various perspectives on the topic and the introduction of used explainable machine learning models. Chapter 4 details the methodology employed in the research, including data description, preprocessing steps, and the application of machine learning techniques. In Chapter 5, the thesis discusses the results obtained from the machine learning application in Chapter 4. Finally, Chapter 6 concludes the research, summarizing key findings, discussing the implications, and suggesting avenues for future work in this area.

2 Theoretical context

In this chapter, we present an in-depth exploration of key definitions and concepts integral to the structure and substance of this thesis. A thorough examination of credit risk, machine learning, and the interpretability of machine learning models is discussed to establish a robust theoretical base.

2.1 Credit Risk

2.1.1 Introduction

Credit risk, as conceptualized by the Basel Committee on Banking Supervision, is defined as the potential that a bank borrower or counterparty will fail to meet its obligations according to agreed terms. This form of risk is significant for banks, as their primary revenue source is often derived from extending credit to borrowers. Precise quantification of credit risk is crucial for setting equitable interest rates on loans and for the overall financial stability of the banking sector.

Regulatory bodies such as the European Banking Authority (EBA) mandate the assessment and management of credit risk. Regulations like the EU's Capital Requirements Regulation (CRR) require banks to maintain a capital buffer for unexpected losses, ensuring stability even in scenarios of multiple client defaults (Commission, 2023b; EBA, 2019).

2.1.2 Credit Risk Modelling

The quantification of credit risk is visually represented in Figure 2.1, showcasing both expected and unexpected losses on a loss curve. The curve's initial segment represents expected loss, calculated to inform bank pricing policies. The subsequent area symbolizes unexpected loss, significant for regulatory capital reserve determinations. The scope of this thesis is on the regulatory capital and therefore we go into detail for this calculation.

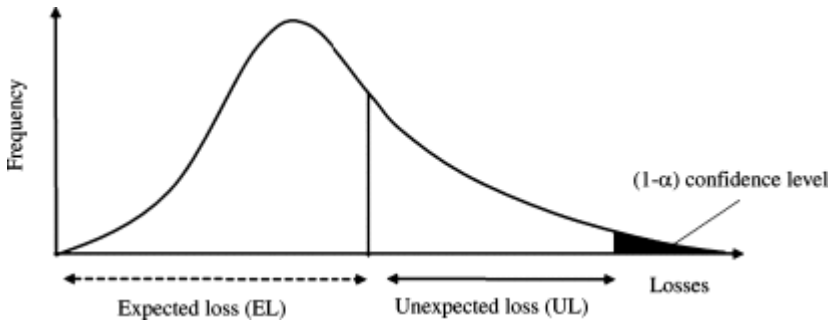


FIGURE 2.1: Credit loss distribution
(Bellini, 2019b)

Credit risk modeling involves three components:

- **Probability of Default (PD)**: The likelihood of a borrower failing to repay a loan, typically defined by a loan being over 90 days overdue. In the estimation oft the PD a one-year time period is taken into account (EBA, 2017)
- **Exposure at Default (EAD)**: The unpaid loan amount at the time of default, usually decreasing as the loan approaches maturity.
- **Loss Given Default (LGD)**: Expressed as a fraction of the EAD, dependent on collateral and representing the unrecouped loan proportion in a default event.

The Expected Loss is calculated using the formula, this is as stated used for pricing policies:

$$\text{Expected Loss} = \sum_i \text{PD}_i \times \text{LGD}_i \times \text{EAD}_i \tag{2.1}$$

The calculation of unexpected loss can also be performed using a specific formula, as mandated by regulatory authorities. This calculation is essential for determining the level of unexpected loss that a financial institution should be equipped to withstand. Within this formula, the Worst Case Default Rate (WCDR) represents the maximum default rate that we are 99.9% confident will not be exceeded. This rate is aggregated across all outstanding accounts, denoted by 'i'. While each component of this formula can be modeled, the primary focus of our analysis is on the Probability of Default (PD) component. This approach emphasizes the critical role of PD in assessing and managing the risk of unexpected loss.

$$\text{Unexpected Loss } = \sum_i (WCDR_{99.9\%,i} - PD_i) \cdot LGD_i \cdot EAD_i$$

By accurately calculating the unexpected loss, a financial institution can determine the necessary capital reserves required to comply with regulatory requirements. Effectively reducing this capital requirement can directly benefit the lending capacities of the financial institution, enabling it to extend more credit while still adhering to regulatory standards. This optimization not only ensures regulatory compliance but also enhances the institution's financial flexibility and lending potential.

Probability of Default (PD)

The distinction between Point in Time (PIT) and Through the Cycle (TTC) is paramount in credit risk modeling within the banking sector. These terms outline distinct approaches for predicting the probability of default (PD) for borrowers (Bellini, 2019a).

- *PIT (Point in Time):* Captures the probability of default over a specific, usually short-term, time horizon, factoring in the prevailing economic conditions. It sensitively responds to the cyclical economic fluctuations. Hence, in a recession, PIT PD would likely increase, while in an economic boom, it would decrease.
- *TTC (Through the Cycle):* Measures the PD over a full economic cycle, averaging out its highs and lows. The resultant TTC PDs remain more stable over time, reflecting the long-term risk over both favorable and adverse economic conditions (Bellini, 2019a).

Both methodologies present their inherent strengths and limitations. The preference between PIT and TTC is predominantly influenced by the specific application

in question and the overarching business and regulatory ambiance of the financial institution. For the regulatory requirement a TTC PD is used based on the more stable data and the usage of the full economic cycle.

2.2 Machine Learning

Machine learning (ML) has the field of credit risk analysis by providing a suite of algorithms and techniques for pattern recognition and data-driven decision-making. In this thesis, we focus on the application of ML to enhance the accuracy of Probability of Default (PD) predictions, a critical component in credit risk management. ML not only augments traditional statistical models but also introduces advanced capabilities to handle complex, non-linear relationships in financial data.

2.2.1 Theoretical Foundations of Machine Learning

Machine learning encompasses various approaches, including supervised, unsupervised, and reinforcement learning. Its mathematical underpinnings involve optimization, statistical inference, and linear algebra. Within a given dataset $\mathcal{D}$, consisting of n samples where each sample is a tuple (x_i, y_i) with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, supervised learning aims to discover a function $f : \mathbb{R}^d \to \mathbb{R}$ that minimizes a loss function L across the dataset.

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \tag{2.2}$$

Here, $\mathcal{F}$ represents the set of possible functions.

2.2.2 Machine learning types

To have a introduction in machine learning we first must define the different possibilities that machine learning have to offer. It is not a on size fits all. First we will explain the types of machine learning techniques. To make the distinction between a classification or a regression task. In figure 2.2 the interesting types of machine learning can be seen. There are more types of machine learning

Supervised Learning

Supervised Learning is a type of machine learning where the model is trained on a labeled dataset. In this approach, the model learns to map input data to the output label through a training process. It uses known input-output pairs to learn a function that can predict the output associated with new input data. This type of learning is widely used for applications like classification and regression(Géron, 2019).

Classification and Regression

In the context of Supervised Learning, two primary tasks are Classification and Regression. Classification involves predicting a discrete label - for instance, determining whether an email is spam or not. Regression, on the other hand, involves predicting a continuous quantity - like forecasting the temperature for tomorrow (Géron, 2019).

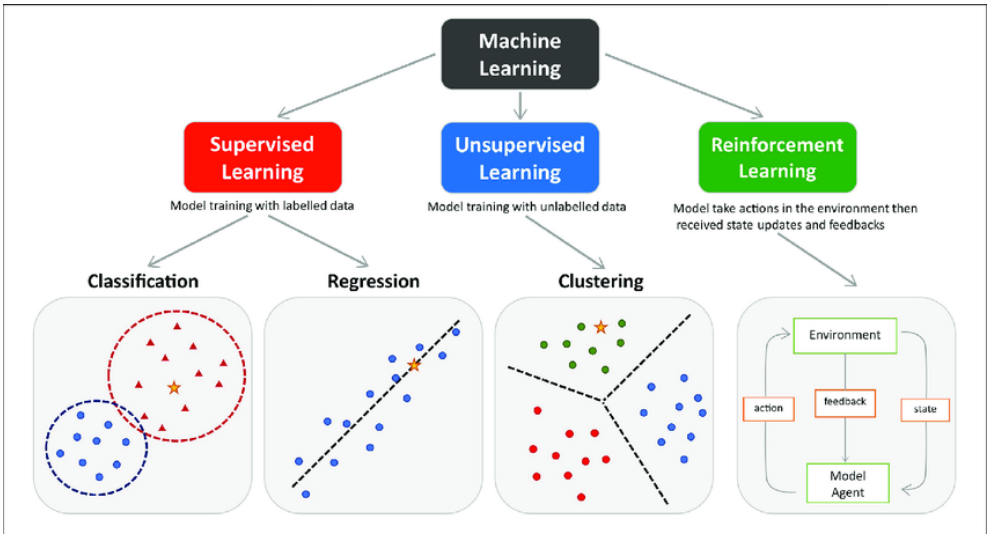


FIGURE 2.2: Overview of machine learning types
(Peng et al., 2021)

Unsupervised Learning

Unlike Supervised Learning, Unsupervised Learning involves training models on data that is not labeled. The goal here is to discover inherent patterns, structures, or features within the input data. Common applications of unsupervised learning include clustering and dimensionality reduction. This learning type helps in understanding and summarizing data sets where the explicit outcome is not known (Géron, 2019).

Reinforcement Learning

Reinforcement Learning is a paradigm of learning where an agent learns to make decisions by performing certain actions in an environment and receiving rewards or penalties in return. It is characterized by the agent's ability to learn its behavior based on feedback from its own actions and experiences rather than from a predefined set of data (Géron, 2019).

Clustering

Clustering is a main technique in Unsupervised Learning where the goal is to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. It is widely used in data analysis for pattern recognition, image analysis, and information retrieval (Géron, 2019).

Ensemble Techniques

Ensemble techniques combine multiple machine learning models to improve performance. These techniques can be used in both Supervised and Unsupervised Learning and are particularly effective in improving the accuracy, robustness, and reliability of predictions. Examples of ensemble methods include Random Forests, Gradient Boosting, and Stacking (Géron, 2019).

2.3 Explainability in Machine Learning

Incorporating machine learning (ML) into fields, such as credit risk analysis, necessitates a comprehensive understanding of model predictions. In this regard, Explainable Machine Learning (XAI) plays a role in ensuring transparency and reliability. To provide clarity on the subject matter and establish the definitions we will adhere to, it is important to delve into the concepts of explainability and interpretability.

2.3.1 Interpretability vs. Explainability

The terms interpretability and explainability are often used interchangeably in literature, yet they hold distinct meanings. Interpretability is a model's passive trait, reflecting how well a human can understand its outputs or behaviour, to transparency. Explainability, in contrast, is an active quality, involving methods a model uses to clarify its internal workings (Barredo Arrieta et al., 2020).

- **Understandability**: This refers to a model's ability to be grasped by humans in terms of its functioning, without delving into its internal algorithmic process (Barredo Arrieta et al., 2020).
- **Comprehensibility**: Relates to a machine learning algorithm's presentation of acquired knowledge in human-understandable forms (Barredo Arrieta et al., 2020). It emphasizes outputs that are symbolically, semantically, and structurally relatable to human deduction.
- **Interpretability**: The ability to explain or make sense of a model's outputs in human terms.
- **Explainability**: Concerns the creation of an interface that accurately represents the decision-maker's processes in a manner understandable to humans.
- **Transparency**: Pertains to a model's inherent understandability, classified into simulatable, decomposable, and algorithmically transparent types.

In these definitions, understandability is key. It involves not just the model's inherent clarity but also the human capacity to comprehend its decisions. Comprehensibility and transparency are closely related to this concept, with the latter focusing on the model's inherent clarity. Hence, understandability, encompassing the model's clarity and the user's comprehension, forms the foundation of XAI, a theme that will be expanded upon in 2.6. This form of explainability will be employed throughout the remainder of this thesis.

2.3.2 Approaches to Explainability

Explainability in ML can be segmented into two broad categories: ante-hoc and post-hoc explanations.

Ante-hoc (Ad-hoc) vs. Post-hoc Explanations

- **Ante-hoc Explanations:** These are built into the model's design, ensuring interpretability from the onset. For instance, linear regression models inherently provide coefficients that directly relate features to the predicted outcome.

- **Post-hoc Explanations:** These techniques aim to interpret complex, often opaque models after they have been trained. They are crucial for understanding models like deep neural networks where ante-hoc interpretability is not feasible.

Explainability Techniques and Tools

To emphasize the focus of this research is more on already interpretable machine learning techniques. The possibilities of tools and techniques can contribute to the explenation of the machine learning model. Therefore we will explain different techniques like partial dependency plots (PDP). Individual conditional expectations (ICE) and S

Partial Dependency Plots (PDP) and Accumulated Local Effects (ALE) Plot

PDPs provide insights into how a feature affects the model's average prediction. ALE Plots offer a local perspective, focusing on feature changes within subsets of data.(Molnar, 2023; Goldstein et al., 2014)

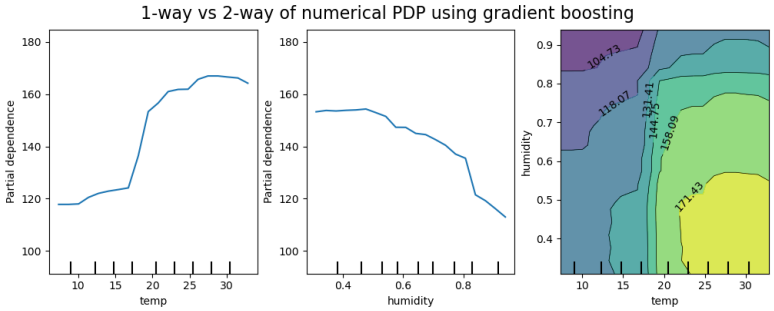


FIGURE 2.3: PDP example
(**empty citation**)

Individual Conditional Expectations (ICE)

ICE plots show how the model's prediction changes for an individual instance, enhancing the understanding provided by PDPs, especially in diverse datasets.(Molnar, 2023; Goldstein et al., 2014)

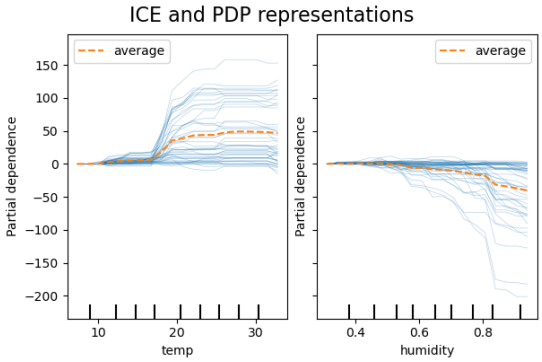


FIGURE 2.4: ICE example plot

SHAP Values

SHAP values, based on game theory, attribute each feature's contribution to a specific prediction. For a model function f and features N, the SHAP value for feature i is:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{2.3}$$

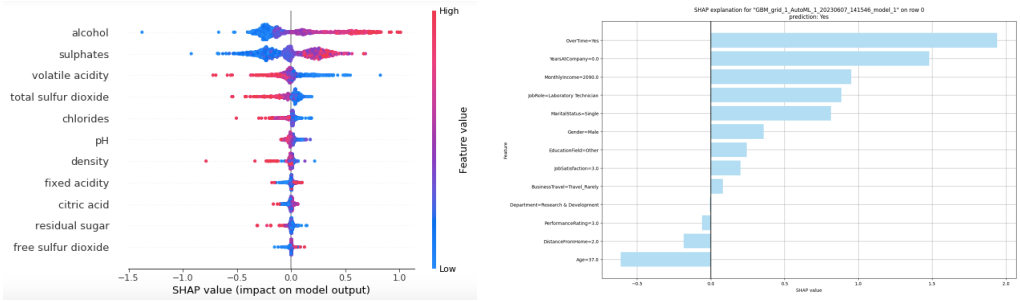


FIGURE 2.5: Shapley values examples
(Hadji Misheva, 2023)

2.4 Audience-dependent Explainable Artificial Intelligence

The integration of machine learning (ML) models into critical decision-making areas, such as credit risk management, has underscored the growing importance of Explainable Artificial Intelligence (XAI). Unlike a one-size-fits-all approach, audience-dependent XAI recognizes the diverse informational needs and backgrounds of various stakeholders (Belle and Papantonis, 2021). This approach tailors explanations to the specific requirements and understanding levels of different user groups, ensuring that ML models are transparent, understandable, and trustworthy across all segments of the financial sector.

Key stakeholders in credit risk management include regulators, who demand detailed explanations to ensure compliance (Commission, 2023b; EBA, 2023); financial analysts and credit officers, who require a mix of technical detail and practical insights ; borrowers and consumers, who seek transparent, jargon-free information ; and data scientists and developers, who need in-depth technical explanations to refine ML models (Belle and Papantonis, 2021). Catering to the unique needs of these groups involves selecting and customizing XAI techniques like LIME, SHAP, and decision trees (Barredo Arrieta et al., 2020), establishing feedback mechanisms. Moreover, developing comprehensive policies and documentation on XAI practices is essential for enhancing transparency and accountability.

2.5 Model validation

Model validation is a critical process in machine learning that ensures the reliability and robustness of predictive models. It involves various techniques and methodologies to assess a model's performance and its ability to generalize to unseen data. Effective validation is key to avoiding common pitfalls such as overfitting, where a model might perform well on training data but poorly on new data. This section

will explore the fundamentals of model validation, including the use of confusion matrices, key performance metrics, and cross-validation techniques. These tools and strategies are essential for evaluating model accuracy, understanding its behavior in various scenarios, and ensuring that the model remains effective and reliable when deployed in real-world applications.

Confusion matrix

A confusion matrix is typically structured as a 2x2 table, representing the outcomes of predictions for a binary classifier. The matrix consists of the following components also visible in Figure 2.6:

- **True Positives (TP)**: These are instances where the model correctly predicts the positive class.
- **True Negatives (TN)**: These are instances where the model correctly predicts the negative class.
- **False Positives (FP)**, also known as Type I error: These occur when the model incorrectly predicts the positive class.
- **False Negatives (FN)**, also known as Type II error: These occur when the model incorrectly predicts the negative class.

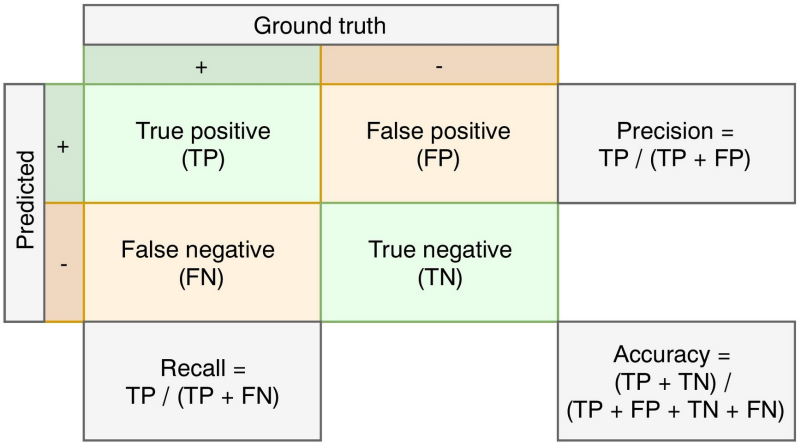


FIGURE 2.6: Confusion matrix
(Jeppesen et al., 2019)

The confusion matrix is used calculating various performance metrics, such as accuracy, precision, recall, and the F1 score. It provides an intuitive understanding of not just the overall accuracy of the model, but also how well it performs in terms of each class.

Key performance metrics

Several performance metrics can be derived from the confusion matrix:

- **Accuracy**: The proportion of total predictions that were correct.
- **Precision**: The proportion of positive identifications that were actually correct.

- **Recall** (or Sensitivity): The proportion of actual positives that were correctly identified.
- **F1 Score**: The harmonic mean of precision and recall.

In the evaluation of predictive models, particularly for classification tasks in machine learning, metrics such as AUROC (Area Under the Receiver Operating Characteristic curve), AUPRC (Area Under the Precision-Recall Curve), and the Gini Coefficient are of significant importance .

- **AUROC**: This metric evaluates a classifier's ability to distinguish between classes, providing an aggregate measure across all possible classification thresholds. An AUROC close to 1 indicates strong class separability, while a value near 0.5 suggests performance equivalent to random guessing. AUROC is particularly useful in imbalanced datasets as it is not influenced by the distribution of classes.
- **Gini Coefficient**: The Gini Coefficient, often used in credit scoring models, is a measure derived from the AUROC. It is calculated as $Gini = 2 \times AUROC - 1$. The Gini Coefficient ranges from -1 (worst) to 1 (best), with higher values indicating a better performing model. A Gini Coefficient of 0 is equivalent to random guessing, mirroring the interpretation of the AUROC. Because this measurement is used in the financial institution this will also be used for the developped model.
- **AUPRC**: The AUPRC is valuable when the positive (minority) class is of primary interest, particularly in imbalanced datasets. It focuses on precision (true positive predictions divided by all positive predictions) and recall (the model's ability to identify all positive instances). A higher AUPRC value implies more accurate identification of positive instances with fewer false positives.

Together, these metrics offer a comprehensive view of model performance. While AUROC assesses general class differentiation ability, the Gini Coefficient provides a scaled interpretation of this differentiation, and AUPRC focuses on the model's effectiveness in identifying positive cases in scenarios with class imbalance.

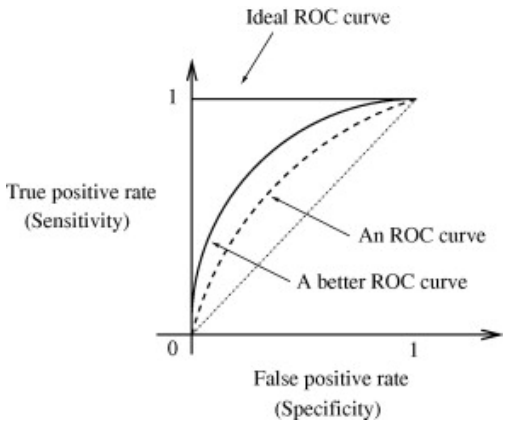


FIGURE 2.7: ROC curve
(Meyer-Baese and Schmid, 2014)

Cross-validation

Cross-validation is a statistical method used in machine learning to evaluate the performance of models. It is particularly useful for assessing how the results of a statistical analysis will generalize to an independent data set. The primary goal of cross-validation is to prevent overfitting, a model's tendency to learn the noise in the training data rather than the underlying pattern.

There are several types of cross-validation, each serving different purposes and suited to different types of data. Some of the most common types include:

- **K-Fold Cross-Validation**: The data set is divided into 'K' number of subsets. The holdout method is repeated 'K' times, with each of the subsets serving as the test set once, and the remaining data as the training set. The results are then averaged over the rounds.

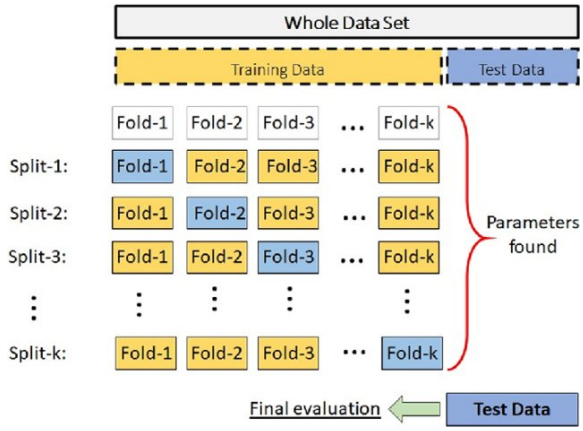


FIGURE 2.8: K-fold cross validation
(Sevinç, 2022)

- **Leave-One-Out Cross-Validation (LOOCV)**: A special case of k-fold cross-validation where 'K' equals the number of data points in the dataset. It involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated on all ways to cut the original sample on a validation set of just one observation, and a training set.
- **Stratified K-Fold Cross-Validation**: Similar to K-fold, but in this method, each fold contains approximately the same proportion of class labels as the original dataset, which can be crucial for imbalanced datasets.

2.6 Conclusion

In this chapter, we have delved into the significant potential of machine learning (ML) for enhancing Probability of Default (PD) modeling within the realm of credit risk assessment. The investigation has underscored the dual nature of ML's impact: while it offers notable improvements in predictive accuracy and efficiency, it simultaneously presents challenges concerning transparency, fairness, and regulatory compliance. Highlighting the importance of explainable artificial intelligence (XAI), we explored various techniques such as Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), Individual Conditional Expectation (ICE), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations

(LIME). These methodologies, coupled with the distinction between ante-hoc and post-hoc explanations, constitute a comprehensive framework for demystifying ML models and fostering a deeper understanding of their decision-making processes.

The exploration of XAI within PD modeling not only aims to enhance model transparency and trust but also seeks to ensure adherence to evolving regulatory standards that prioritize fairness and risk management. As we navigate through the complexities of integrating ML into credit risk analysis, the ongoing development of regulatory frameworks and guidelines emerges as a critical factor in balancing the benefits of technological advancement with the need for robust financial stability.

This chapter contributes to the broader discourse on the application of ML in financial services, highlighting the essential role of collaboration among regulators, financial institutions, and technology providers. Such partnerships are vital for addressing the challenges posed by ML and for crafting a responsible and sustainable approach to leveraging these technologies in improving the accuracy and reliability of PD modeling. Through this collaborative effort, we can better navigate the intricacies of modern credit risk management, ensuring that the financial sector remains resilient amidst the rapid pace of technological change.

3 Stakeholder Perspectives on Explainability in PD Modeling

3.1 Introduction

While the technical powers of ML models offers enhanced predictive accuracy, their complexity raises critical questions about their explainability and transparency. This chapter aims to delve into these questions by examining the perspectives of various stakeholders involved in the PD modeling process. The first question this chapter seeks to address is: **"How do stakeholders perceive the importance of explainability in machine learning models for PD?"** Understanding the viewpoints of stakeholders including financial institutions, regulatory bodies, credit analysts, and borrowers is needed to grasp the broader implications of model explainability. This exploration is not merely an academic exercise; it has practical ramifications in terms of model acceptance, regulatory compliance, and the ethical dimensions of credit risk modeling. Stakeholders' perspectives on the importance of explainability will shed light on the balance between the need for advanced analytical capabilities and the imperative for transparency in decision-making processes. The second important question discussed in the chapter is : **"What are the challenges and concerns stakeholders associate with non-explainable versus explainable machine learning models in PD?"** This is needed in understanding the concerns and potential resistance stakeholders might have towards 'black-box' ML models. By contrasting these concerns with the perceived benefits of explainable models, the chapter will offer insights into the trade-offs and decision-making criteria that stakeholders consider when assessing the potantial for adoption of advanced ML techniques in PD modeling. In this chapter we will also be able to derive a checklist to help in the evaluation phase of the CRISP-DM.

3.2 Interviews

In preparation for the interviews, a careful selection of key stakeholders was undertaken to gain a comprehensive understanding of the various perspectives involved in model development within financial institutions. This selection process aimed to encompass a broad range of roles and responsibilities, ensuring diverse insights into the modeling process. The primary stakeholders identified for these interviews include model developers, who are directly responsible for the creation and implementation of the models; model owners, who oversee the models' lifecycle and integration into the business process; front office risk management professionals, who assess and manage the risks associated with the model; and model validation teams, responsible for ensuring the accuracy and reliability of the models. This diverse group of stakeholders is crucial to provide insight in the model development and its impact on the financial institution. The interview questions and the approach are described in appendix A.

3.2.1 Stakeholders

Internal

- Model (co-)owner Front Office WB Lending CLEC
- Model (co-)owner Risk Management Wholesale Banking Risk
- Data delivery COO Risk Finance
- IT Implementation COO Risk Finance
- Product Area Lead WB Lending CLEC
- Model development FR Model Development
- DD and RRD regulator rator development

External

- DNB
- ECB

3.3 Analysis of Stakeholders' Perspectives on Explainability in PD Modeling

After conducting interviews with various stakeholders in the banking sector, including a model validator, model owner, model developer, and a front office risk professional, we have gathered insightful perspectives on the implementation of machine learning models in Probability of Default (PD) modeling. Notably, our research scope is confined to the stages leading up to model implementation; therefore, the model implementation team itself was not included in these interviews. Additionally, while audit teams and committee regulators were not directly interviewed due to unavailability, their viewpoints are represented and encompassed through an analysis of regulatory view papers. From this research, we have derived insights regarding the critical importance of explainability, as well as the distinct challenges posed by explainable versus non-explainable machine learning models in the context of PD modeling.

3.3.1 Importance of Explainability in Machine Learning Models for PD

- **Regulatory Compliance:** All stakeholders emphasized the necessity of explainability for meeting regulatory demands. Regulatory bodies require clear explanations and justifications for model decisions, making explainability a critical compliance factor.
- **Ethical Responsibility:** There's a strong ethical dimension to using explainable models. These models allow stakeholders to identify and rectify potential biases, ensuring that they do not discriminate against certain customer groups.
- **Operational Transparency:** Explainability is valued for its role in unmasking how model inputs influence outputs. This clarity is essential for compliance with the regulator, internal decision-making and maintaining customer trust.
- **Risk Management:** Understanding the reasoning behind predictions is vital for effective risk management. Explainable models facilitate better risk assessment and mitigation.

3.3.2 Challenges and Concerns with Non-Explainable vs. Explainable Machine Learning Models in PD

- **Complexity vs. Transparency Trade-off:** A significant trade-off between model complexity and transparency is recognized. More complex models may offer higher accuracy but lack transparence. Complex models might also be overfitted. An important feedback from stakeholder interviews is that the accuracy and transparancy need to go hand in hand.This is because without transparancy the model will not be accepted. Therefore a model must surpass a minimal threshold of transparency and on this bases improvements in model performance can be considered.

- **Regulatory Challenges:** Non-explainable models are more challenging in meeting regulatory requirements, since these requirement often mandate a clear reasoning behind predictions.

- **Ethical and Bias Concerns:** There's a heightened awareness that non-explainable models may harbor biases that might lead to unfair discrimination to certain individuals or groups of people. Explainable models are thus preferred for their capacity to detect and correct biases.

- **Internal Acceptance and Trust:** Models lacking in explainability face internal resistance. Decision-makers and risk managers prefer models that can be trusted, due to their transparancy and understandable log and principle of operation.

- **Stakeholder Communication:** Explainable models support clearer communication with various stakeholders, including regulators, internal teams, and customers. The communications with the stakeholder should take place in different formats. For the end users a knowledge document where the use of the model is described would suffice. The model developers, model validators and front office risk managers want to fully understand the model. A difference in communication with risk management, model developers and validators can be made in terms of the level of technical detail. While the model owner is primarily interested in comprehending how alterations in particular features influence the model output, an in-depth understanding of the underlying reasons for these changes may not be necessary. From a risk management perspective, it is crucial that changes in probability of default (PD) align with economic logic and intuition. For instance, the model should not include implausible causalities, such as suggesting that the number of letters in a person's name directly affects their PD. Such insights ensure that the model's outputs are not only accurate but also economically meaningful and interpretable.

In conclusion, stakeholders highlight the importance of balancing the advanced predictive capabilities of machine learning models with the need for explainability in Probability of Default (PD) prediction. This balance is important for ensuring regulatory compliance, upholding ethical integrity, managing risk effectively, and maintaining trust among all participants in the credit lending process. Next, we will delve into the perspectives of external stakeholders.

3.4 Regulatory Perspective

The regulator, as an external stakeholder, is responsible for the acceptance of a model. This section delves into the regulatory perspective on the application of Artificial Intelligence (AI) and Machine Learning (ML) technologies. It is structured around various regulatory frameworks and guidelines issued by leading international bodies, aiming to encapsulate key principles, challenges, and directives outlined in these frameworks. The outcomes of this comprehensive analysis are systematically presented in table 3.1 , offering insights into the evolving regulatory landscape governing AI and ML implementations.(Commission, 2023a; European Central Bank., 2019; Barredo Arrieta et al., 2020; Hottenhuis, 2022)

TABLE 3.1: Summary of Regulatory Guidelines and Principles

Issuing Body	**Act / Guidelines / Article**	**Principles and Findings**
BIS	BCBS Newsletter: Newsletter on AI and ML	- Explainability: Transparency in model design - Governance structure - Implications of ML models
	FSI Insight No 35	- Transparency - Reliability - Accountability
EU	GDPR: Articles 5.1 (c), 5.1 (h), 22 + Recital 77	- Data minimization - Customers' rights - Human intervention requirements
	ALTAI	- Human agency and oversight - Technical robustness and safety - Privacy and data governance
EBA	CRR: Articles 174, 175, 179, 189	- Human oversight - Extensive documentation - Intuitive model design - Senior management comprehension of system design and operations
	Discussion Paper: On ML in IRB Models	- Model complexity evaluation criteria: - Number of parameters - Non-linear relation representation - Data amount for sound estimation - Data utilization for information extraction - Applicability to unstructured data - Recommendations: 1. Avoid unnecessary complexity 2. Ensure correct model interpretation and understanding 3. Establish reliable validation processes

3.4.1 Regulatory Context in the European Union

General Data Protection Regulation (GDPR)

On a European level, cross-industry regulations are established by the European Union, with the General Data Protection Regulation (GDPR) (EU, 2016) being one of the most impactful. Although GDPR significantly affects banks, only a few articles directly pertain to the use of ML.

- Article 5.1(c) emphasizes 'data minimization', posing challenges in ML implementation due to requirements like maintaining five years of data history for risk drivers (CRR, article 180). This constraint limits ML implementation, especially as model complexity and the number of risk drivers increase (Commission, 2018).
- Article 15.1(h) grants customers the right to access meaningful information about logic in automatic decision-making, restricting the use of inexplicable models like deep neural networks.
- Article 22 requires that models allow for human intervention, which is challenging with black box models.
- Recital 71 emphasizes the need for human intervention and explanation in automated decision-making processes.

The EBA discussion paper and the follow up on ML for IRB models (EBA, 2023; EBA, 2021) highlights challenges like interpreting results, ensuring management understanding, and justifying results to supervisors. It proposes evaluating model complexity based on characteristics like the number of parameters and the capacity for non-linear relationships.

3.5 Machine Learning Model Selection Checklist for PD Modeling

Based on the insights gathered from interviews and regulatory documents, we have compiled a checklist to assess if a machine learning model adheres to the essential criteria for approval. Due to the time constraints of this thesis, the implementation aspect of the checklist will not be examined. For each item on the checklist, it is necessary to articulate how and in what sequence the model complies with these specified requirements.

1. **Explainability and Transparency**
    - The model's decisions can be explained and understood (In how far features affect the result)
    - Outputs are transparent, allowing stakeholders to trace how inputs are transformed into predictions.
2. **Regulatory Compliance**
    - The model meets current regulatory standards and guidelines (EBA, 2017).
    - Documentation is available and sufficient for regulatory review.
    - The model can be audited and validated as per regulatory requirements.

3. **Ethical Considerations and Bias Management**
    - The model includes mechanisms to identify and mitigate biases.
    - It ensures fairness and not lead to unfairly discriminate against individuals or groups of people.
    - The model respects ethical guidelines related to AI and machine learning in finance.
4. **Complexity vs. Interpretability Balance**
    - The model balances predictive accuracy with ease of interpretation.
    - Complexity does not overshadow the model's ability to be understood by non-technical stakeholders.
5. **Stakeholder Acceptance and Trust**
    - The model is acceptable to internal decision-makers and risk managers.
    - It is consistent and reliable in its predictions.
6. **Performance and Accuracy**
    - The model demonstrates high predictive accuracy on relevant dataset.
    - It performs well across various metrics (AUROC, Gini and AUPRC)
7. **Data Efficiency and Robustness**
    - The model efficiently handles the available data, avoiding overfitting.
    - It is robust to changes in data patterns and economic conditions.
8. **Operational Feasibility**
    - Implementation of the model is feasible with available IT infrastructure.
    - Required computational resources are within acceptable limits.
9. **Maintenance and Adaptability**
    - The model can be regularly updated and maintained without excessive resource allocation.
    - It is adaptable to changes in the market and regulatory environment.

3.6 Machine learning techniques

This section outlines the machine learning techniques that will be employed in our study. The selection of these models has been informed by the outcomes of our interviews. The interview and regulator shows that we need a interpretable model. Towards the end of the thesis, we will compare the performance of the developed models against the requirements identified from these interviews. We begin with an introduction to General Additive Models (GAM), followed by detailed explanations of two specific machine learning models: Gami-Net and the Explainable Boosting Machine (EBM).

3.6.1 Explainability in machine learning

In order to maintain a focused scope, our research will not encompass all the various types of machine learning techniques. Instead, we aim to concentrate specifically on the aspect of explainability. There is existing research in this field which discusses the trade-off between accuracy and interpretability. By building upon these findings, our study seeks to further explore and understand how this balance can be effectively achieved in machine learning applications (Bussmann et al., 2021; Yang et al., 2023; Langer et al., 2021).
In figure 3.1 the trade-off is shown in a graph. In this thesis we are looking into explainable machine learning techniques and are interested more interpretable models. When looking in the graph you can see logistic regression technique. This is a technique already used in the financial institution to create the model. To keep around the same level of interpretability and look into models that have a higher accuracy we look into the General additive models (GAM) and specifically, explainable boosting machine (EBM) and GAMINET. We wanted to limit the number of machine learning techniques tested and FIGS shows similar results as a normal GAM. EBM and GAMINET are additions towards the GAM and claim to have a higher accuracy while keeping the interpretability.

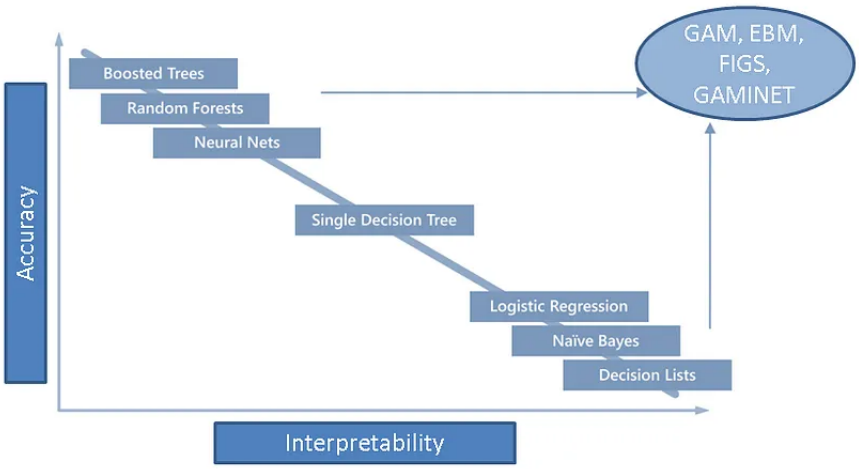


FIGURE 3.1: Accuracy and interpretability trade-off (Baruah, 2023)

3.6.2 Machine learning techniques (Géron, 2019)

Considering the emphasis on ensuring that a Probability of Default (PD) model is both transparent and explainable, our interest gravitates towards GAMI-Net and EBM (Explainable Boosting Machine). The selection of these models is influenced by their intrinsic focus on explainability, as highlighted in the literature (Yang, Zhang, and Sudjianto, 2021; Nori et al., 2019). As both models are fundamentally based on General Additive Models (GAM), it becomes imperative to delve into GAM to grasp the underlying principles of these advanced machine learning techniques.

EBM, in particular, enhances GAM through the incorporation of boosting techniques, a topic we will examine closely. Conversely, GAMI-Net innovates by integrating a neural network approach to refine its training process. The overarching

aim within this thesis is to develop a model that not only meets but exceeds the predictive accuracy of the incumbent logistic regression-based model. To thoroughly understand EBM and GAMI-Net, it is also essential to explore the specific methodologies they employ, including decision trees, ensemble learning techniques, and neural networks. This comprehensive analysis will enable a deeper appreciation of the models' capabilities and their potential applications in enhancing the transparency and efficacy of PD modeling.

Logistic regression

Logistic regression, a fundamental classification algorithm, is widely employed in credit scoring. In logistic regression, the probability of the dependent variable being in a certain category is modeled as a function of the independent variables. This is expressed using the logistic function, which is a sigmoid function that takes any real-valued number and maps it into a value between 0 and 1.

The logistic function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

where z is a linear combination of the independent variables, given by:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3.2}$$

In this model, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ denote the parameters, and $x_1, x_2, \ldots, x_n$ represent the independent variables. Each β value is a coefficient that predicts the influence of its corresponding variable. The algorithm's goal is to achieve the best possible fit of the function to real-world data, striving for the minimum discrepancy between predicted and actual outcomes. An illustration of this fit can be seen in Figure 3.2, where the logistic function models the relationship, and the individual data points represent cases classified as either defaulted or non-defaulted. It's important to note that logistic regression forms the foundation of Probability of Default (PD) modeling and is extensively employed in financial institutions for risk modelling.

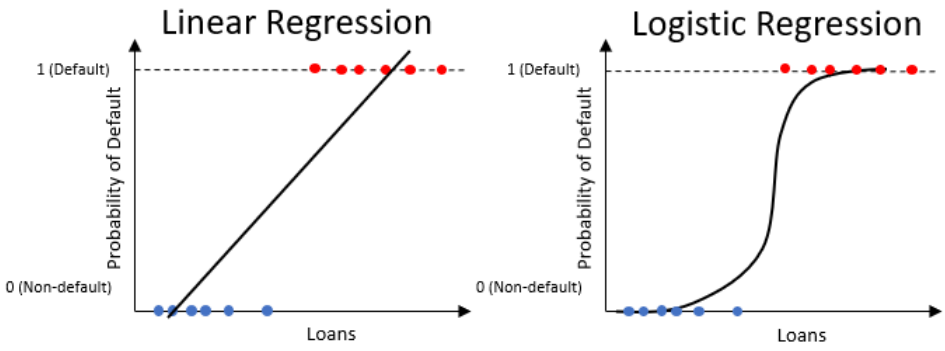


FIGURE 3.2: Linear regression and logistic regression

Decision Trees

Decision Trees categorize data into segments, assigning labels or values to each. A decision tree is built from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). We use statistical measures to choose the most significant feature at each step to split the data.

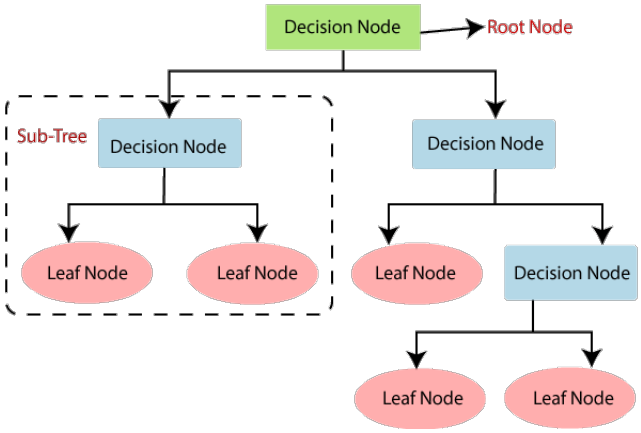


FIGURE 3.3: Decision tree algorithm

Nodes in Decision Trees

- **Root Node**: It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Splitting**: It is a process of dividing a node into two or more sub-nodes based on certain conditions.
- **Decision Node**: When a sub-node splits into further sub-nodes, it is called a decision node.
- **Leaf/Terminal Node**: Nodes that do not split are called leaf or terminal nodes, representing the decision or final outcome.

Splitting Criteria

The decision of making strategic splits significantly affects a tree's accuracy. Different algorithms use different metrics for this:

1. **Gini Impurity** (used in CART algorithm): Gini Impurity measures the disorder of a set. A Gini Impurity of 0 means all elements in the set belong to a single class. The Gini score G for a split is calculated as:

$$G = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.3}$$

   where p_i is the probability of an item with label i being chosen.

2. **Information Gain** (used in ID3, C4.5, and C5.0 algorithms): Information Gain is based on the concept of entropy from information theory. For a dataset S, the entropy is defined as:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3.4}$$

   Information Gain is the change in entropy after a dataset is split on an attribute. It is calculated as the difference between the entropy of the parent node and the sum of the entropies of the child nodes.

Pruning

Pruning is a crucial technique in decision tree algorithms to prevent overfitting, a common issue where a model learns noise in the training data rather than the underlying pattern. Overfitting leads to poor model generalization on unseen data. Pruning effectively reduces the size of the tree by removing parts of the tree that provide little power in classifying instances.

Pruning can be implemented in several ways:

- **Minimum Samples for a Leaf**: This method involves setting a threshold for the minimum number of samples that a leaf node must have. If a split results in a leaf node with fewer samples than this minimum number, the split is not made. This approach helps in reducing the complexity of the tree and thereby, preventing overfitting.
- **Maximum Depth of Tree**: Another method is to set the maximum depth of the tree. The tree is allowed to grow only up to this predefined depth. Limiting the depth of the tree prevents the model from becoming overly complex and learning noise from the training data.
- **Mean Squared Error Reduction**: Particularly in regression trees, pruning can be guided by the reduction in Mean Squared Error (MSE). The process involves evaluating the reduction in MSE that each subtree contributes. Subtrees that contribute minimally to decreasing the overall MSE are pruned. The formula for this is:

$$\text{MSE Reduction} = \text{MSE}_{\text{original}} - \text{MSE}_{\text{pruned}}$$

  where $\text{MSE}_{\text{original}}$ is the Mean Squared Error of the model before pruning, and $\text{MSE}_{\text{pruned}}$ is the MSE after pruning a subtree.

These pruning techniques ensure that the decision tree does not grow too complex, improving its ability to generalize well to new data, while maintaining adequate accuracy on the training data.

Ensemble Learning Techniques

Ensemble learning is a machine learning paradigm where multiple models (often called "weak learners") are trained to solve the same problem and combined to get better results. The main principle behind ensemble learning is that a group of weak learners can come together to form a strong learner, thereby increasing the accuracy of the model. Two popular ensemble learning techniques are Boosting and Bagging. The working of bagging and boosting can be seen in figure 3.4

Boosting

Boosting is an ensemble technique that aggregates several weak learners to create a strong learner. The fundamental concept of boosting involves sequentially training predictors, with each new predictor focusing on correcting the errors of its predecessor. The process starts by training an initial model on the dataset, which then makes predictions. If there are misclassifications, the weights of these instances are increased for the next round of training. A new model is then trained on this adjusted dataset and makes predictions again. This cycle of training and adjusting weights based on the previous model's performance continues until an ensemble of models is formed. In the final ensemble model, each individual model contributes

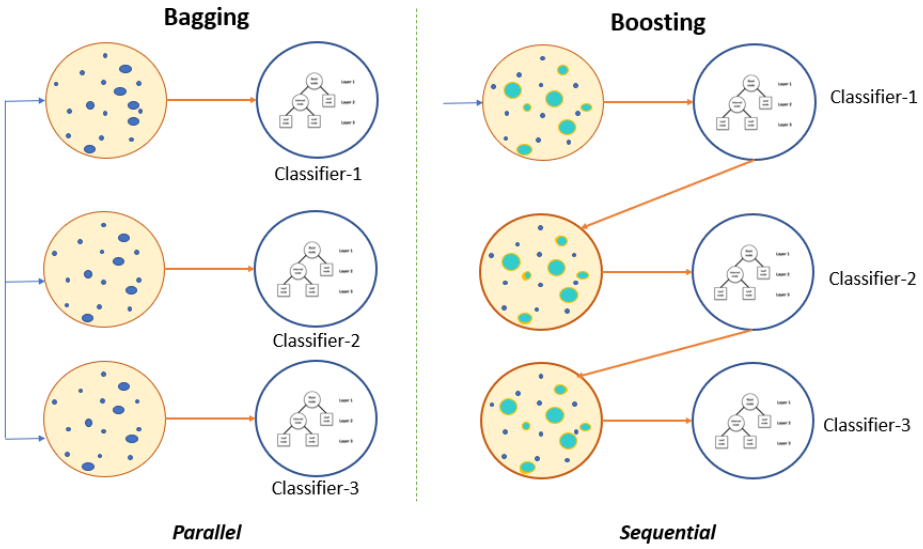


FIGURE 3.4: Bagging and boosting
(Akturk, 2021)

a weighted vote to predict the output, with the weights reflecting the model's accuracy. Two of the most widely used boosting algorithms are AdaBoost (Adaptive Boosting) and Gradient Boosting. Boosting is applicable to both regression and classification problems and is especially effective in binary and multi-class classification scenarios, significantly enhancing the accuracy of the models.

Bagging

Bagging, or Bootstrap Aggregating, is an ensemble learning technique that enhances the stability and accuracy of machine learning algorithms. It involves creating multiple subsets of the original dataset via bootstrap sampling, which is random sampling with replacement. Each subset is then used to train a separate model. The predictions of these individual models are combined to form the final output. For regression problems, this combination is typically the average of the outputs, while for classification problems, it's a majority voting system. The primary advantage of bagging is its ability to reduce variance and prevent overfitting, making it particularly effective with decision trees. While often associated with decision trees, bagging can be applied to various types of methods and is effective in both regression and classification problems. By training on diverse subsets, bagging ensures that the models are robust and stable.

Neural Networks

Deep learning models, a subset of neural networks, process inputs through multiple layers to capture intricate data patterns, making them suitable for multifaceted financial data. A neural network is composed of layers of interconnected nodes or neurons, each linked by weights that represent the strength of connections. These layers are typically organized into three types also shown in Figure 3.5:

- **Input Layer**: The first layer that receives the input signal to be processed.
- **Hidden Layers (Middle layer**: One or more layers where computations are performed through a system of weighted connections. These layers extract and process features from the input.

- **Output Layer**: The final layer that produces the output of the network.

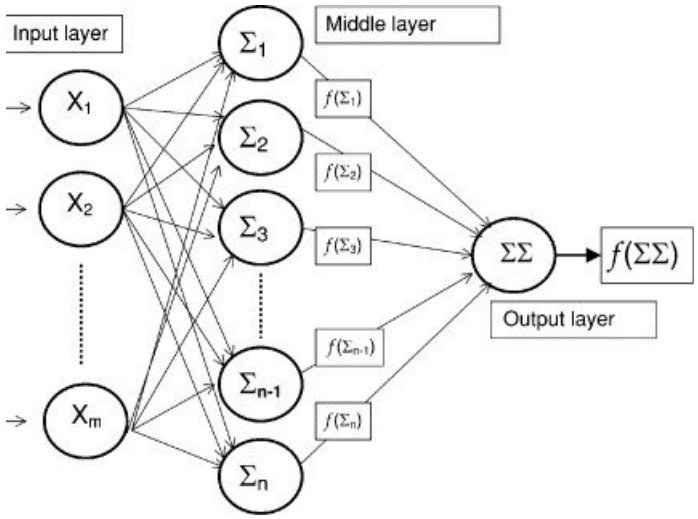


FIGURE 3.5: Basic working of Neural Network
(Demirtaş and Dalkılıç, 2021)

Functioning of a Neuron

Each neuron in a network receives input, processes it, and passes an output to the next layer. The operation of a neuron can be described mathematically as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{3.5}$$

Here, x_i represents the input values, w_i are the weights, b is the bias, and f is the activation function that introduces non-linearity into the output of a neuron.

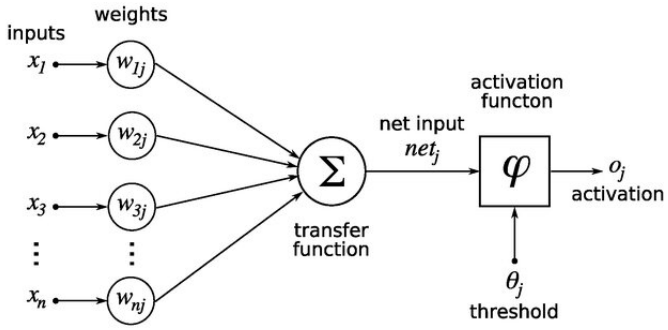


FIGURE 3.6: Basic working of Neural node
(Demirtaş and Dalkılıç, 2021)

Learning Process

The learning process in neural networks involves adjusting the weights of connections based on the error in predictions. The most common learning algorithm is backpropagation combined with an optimization technique like gradient descent.

Backpropagation

Backpropagation involves the following steps:

1. Forward pass: Input is passed through the network to obtain the output.

2. Loss computation: The error between the actual output and the predicted output is calculated using a loss function.
3. Backward pass: The error is propagated back through the network, and the weights are adjusted according to how much they are responsible for the error. The adjustments are made using the gradient descent algorithm to minimize the loss function.

Activation Functions

Activation functions are crucial in neural networks as they introduce non-linear properties to the network. Common activation functions include:

- **Sigmoid**: $\sigma(x) = \frac{1}{1+e^{-x}}$
- **ReLU** (Rectified Linear Unit): $f(x) = \max(0, x)$
- **Tanh** (Hyperbolic Tangent): $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Types of Neural Networks

There are various types of neural networks, each suited for different tasks:

- **Feedforward Neural Networks**: The simplest type, where the data moves in one direction from input to output.
- **Convolutional Neural Networks (CNNs)**: Primarily used in image processing and computer vision.
- **Recurrent Neural Networks (RNNs)**: Suitable for sequential data like time series or natural language.
- **Deep Neural Networks (DNNs)**: Characterized by having multiple hidden layers.

3.6.3 Generalized Additive Models (GAMs)

This section introduces Generalized Additive Models (GAMs), a concept in the field of machine learning that balances the trade-off between model accuracy and interpretability. GAMs extend linear models by incorporating non-linear functions, allowing for a more nuanced understanding of complex relationships in data.

Theoretical Basics

Traditional linear regression models take the form of 3.1. In contrast, GAMs generalize this approach:

$$G(Y) = a + W_1F_1(X_1) + W_2F_2(X_2) + \cdots + W_nF_n(X_n) + C \tag{3.6}$$

Here, $F_n(X_n)$ are non-parametric, smoothing functions, and $G(Y)$ is a link function connecting the expected value of Y to the input features.

Components of GAM

Smoothing Functions F_n The F_n are unique functions for each input feature, commonly implemented as regression spline functions. These splines, or basis functions, enable the model to capture complex non-linear relationships between predictors and the response variable.

Regression Splines Regression splines are a combination of basis functions, expressed as:

$$F_n(X_n) = \sum_i W_i B_i(X_n) \tag{3.7}$$

where B_i are the basis functions and W_i are the corresponding weights.

Link Function $G()$ The link function $G()$ maintains a linear relationship between the target variable and the functions of the input features. This is essential in cases where the relationship is inherently non-linear, for example, using a logit function in binary classification.

Advantages of GAMs

GAMs offer several advantages:

- Enhanced ability to model non-linear relationships.
- Additive nature allows isolation of individual feature impacts. This means that the model expresses the output as a sum of the effects of each individual feature. Unlike some complex models where feature interactions can make it hard to interpret the impact of each feature, in GAMs, each feature contributes independently to the final prediction. This allows for a clear understanding of how each feature affects the outcome.
- Flexibility to control function smoothness based on data complexity. If a feature has a simple, linear relationship with the target variable, the corresponding function in the GAM will be more linear (less smooth). Conversely, if the relationship is more complex (e.g., non-linear), the function can be adjusted to be smoother to capture this complexity. It ensures that the model is neither too rigid (underfitting) nor too flexible (overfitting), which is essential for accurately capturing the underlying patterns in the data.

Limitations and Extensions

While GAMs effectively capture non-linear feature relationships, they do not inherently include interaction terms, so the combinede effect of two features. This limitation is addressed in other models like Explainable Boosting Machines (EBMs), which will be discussed in subsequent section of this thesis.

3.6.4 Explainable Boosting Machines

Explainable Boosting Machines (EBMs), a development from Microsoft Research, offer a promising approach in machine learning, particularly in balancing the trade-off between predictive accuracy and model interpretability. EBMs extend the concept of Generalized Additive Models (GAMs) by integrating interaction terms, thus evolving into tree-based Generalized Additive Models with Interaction terms (GA2Ms).

EBMs employ gradient-boosted ensembles of bagged trees 3.6.2. This approach is renowned for its effectiveness in complex predictive modeling scenarios, striking a balance between accuracy and explainability (Nori et al., 2019).

Mathematical Framework

The underlying mathematical representation of EBMs can be described as follows:

$$y = g^{-1}\left(\beta_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j)\right) \tag{3.8}$$

Here, y represents the target variable, g^{-1} is the link function, β_0 denotes the intercept, $f_i(x_i)$ signifies the main effect of feature x_i, and $f_{ij}(x_i, x_j)$ illustrates the interaction effect between features x_i and x_j.

Training Process

Training an EBM involves constructing small trees sequentially, focusing on individual or pairs of input features. The process includes:

1. **Sequential Tree Building**: This process involves constructing small trees for each feature x_i, computing residuals, and then building subsequent trees on these residuals with different features. The process is iterative, with each new tree aiming to correct the errors (residuals) made by the previous trees. A key aspect of this methodology is the use of a minimal learning rate to ensure stability. The learning rate, a critical hyperparameter in this context, controls how much each tree contributes to the final model. A smaller learning rate means that each tree has a limited influence, requiring more trees to model complex relationships but enhancing the model's stability and generalization ability. It effectively slows down the learning process to prevent overfitting, allowing the model to learn more nuanced patterns in the data.

2. **Creation of Contribution Graphs**: After building the trees, a contribution graph for each feature is developed, serving as a mapping between each feature value and its contribution to the final prediction. An example of the developped graphs can be seen at the bottom of 3.7.

Pairwise Interactions

In EBMs, the handling of pairwise interactions, represented by $\sum f_{ij}(x_i, x_j)$, involves a two-step process. Initially, the model focuses on fitting the main effects, which are the individual contributions of each feature. Once these main effects are determined, they are 'frozen'. This 'freezing' means that their values are fixed and no longer updated during the subsequent modeling steps. This approach allows the model to isolate and accurately capture the individual impacts of each feature before addressing the interactions between them.

After freezing the main effects, the model then calculates the residuals. These residuals represent the error or the portion of the target variable not yet explained by the main effects. The next step is to model the pairwise interactions to further reduce these residuals. This process is illustrated in Figure 3.8, which demonstrates how a pair of features, in this case $f_a \times f_b$, are trained together. The model then attempts to explain the residuals using other interaction pairs. This approach highlights the method by which EBMs manage to capture and interpret the interactions between different feature combinations.

For identifying and selecting the most significant pairwise interactions, EBMs often employ an algorithm like FAST (Fast And Simple Tree). FAST is a heuristic algorithm designed for efficiency and effectiveness in high-dimensional spaces. It

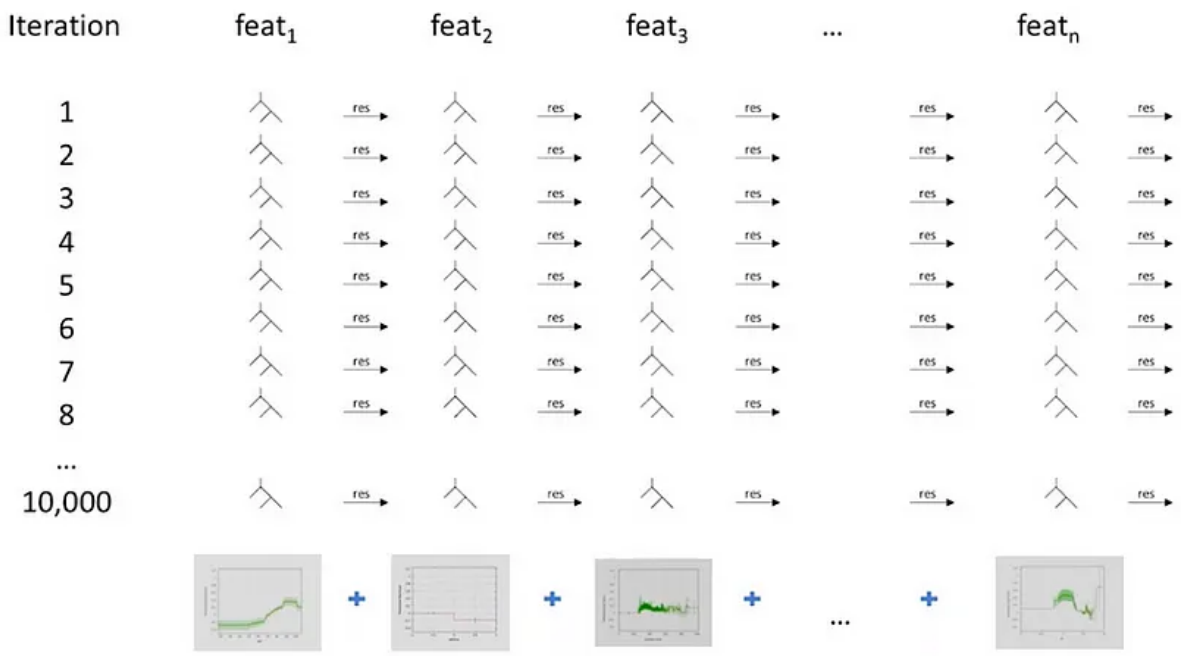


FIGURE 3.7: EBM Training features
(Caruana et al., 2015)

ranks potential pairwise interactions based on their predictive power or contribution to reducing the residuals. Then, it selects the top-ranked interactions for inclusion in the model. By training trees on these selected interactions, EBMs can capture complex relationships between pairs of features while maintaining interpretability and computational efficiency.

This approach of freezing main effects and then using FAST to handle pairwise interactions ensures that EBMs provide a balance between accuracy (by capturing both main effects and key interactions) and interpretability (by simplifying the model structure and focusing on the most relevant interactions).

Interpretation and Predictive Inference

The final prediction in an EBM is an aggregate of contributions from each feature's graph, processed through the link function g. This reliance on lookup values for inference enhances the model's speed.

Challenges and Limitations

EBMs, while powerful, have certain limitations. One challenge is the complexity of interpretation, which can arise from unusual patterns in contribution graphs, often attributable to outliers. As EBMs evaluate all main effects and predetermined interactions, the model's complexity can increase significantly. Furthermore, there can be an overlap in the effects of main and interaction terms, which might obscure the distinct contributions of individual features. This overlapping can complicate the understanding of how each feature independently influences the model's predictions.

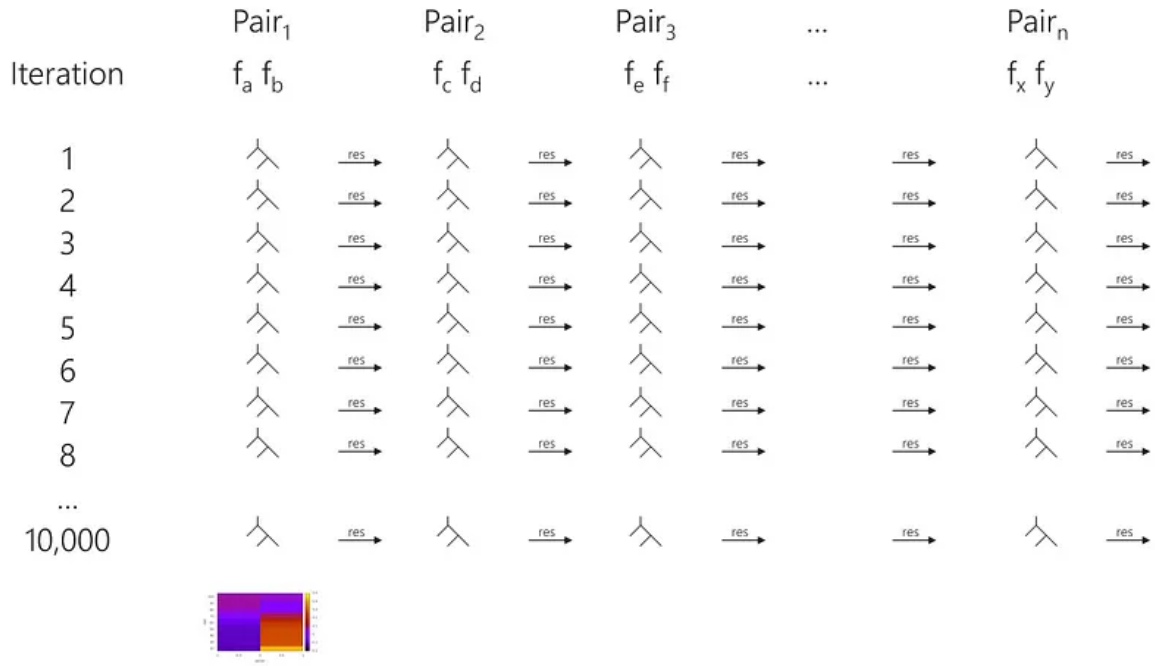


FIGURE 3.8: EBM Training interaction
(Caruana et al., 2015)

3.6.5 Generalized Additive Models with Structured Interactions (GAMI-Net)

GAMI-Net, as detailed in (Yang, Zhang, and Sudjianto, 2021), is a sophisticated deep learning model that combines multiple additive subnetworks. Each subnetwork represents either a main effect or a pairwise interaction of input variables, additively combined to produce the final output. The mathematical formulation of GAMI-Net is expressed as follows:

$$\text{y(Output)} = \text{Bias} + \sum_{j} h_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k) \tag{3.9}$$

where $h_j(x_j)$ denotes main effect, which uses the risk drivers as input represented as x_i, and $f_{jk}(x_j, x_k)$ represents interaction pairs of risk driver. Which are both used as input as can be seen in the two nodes in the figure and then used as input for the neural network representing the interaction pairs.

GAMINET Constraints

GAMI-Net integrates three primary constraints to enhance model interpretability and manage complexity:

1. **Sparsity Constraint**: This constraint is vital for maintaining computational efficiency and interpretability, especially when dealing with a large number of inputs and interactions.

2. **Hereditary Constraint**: The concept of **heredity** in machine learning, especially in models dealing with interaction effects, is grounded in the principle that interactions between features should be considered only if each of the interacting features independently contributes to the predictive power of the

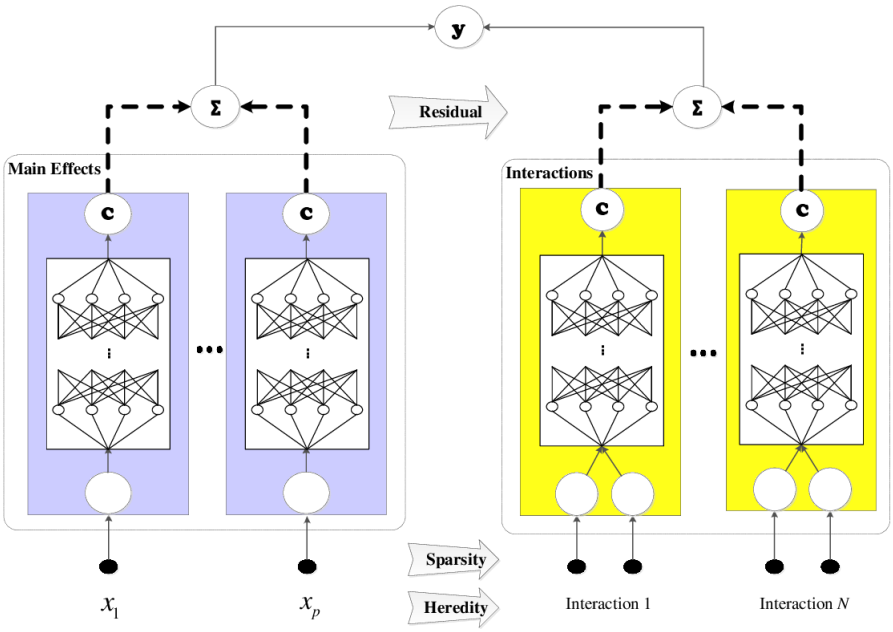


FIGURE 3.9: Gami-Net
(Yang, Zhang, and Sudjianto, 2021)

model. In GAMINet, heredity guides the inclusion of interaction terms. An interaction term between two features is included if, and only if, both features individually show significant main effects. Applying the heredity principle can lead to models that are less complex and more interpretable. It helps to ensure that the model does not become overly complex by including spurious interaction terms, thereby aiding in the prevention of overfitting.

3. **Marginal Clarity**: Marginal Clarity is a design principle that ensures the orthogonal decomposition of each feature's impact within a model like Gami-Net. Orthogonal decomposition means that the contribution of each individual feature (and their interactions) to the model's output is separated in a way that they don't overlap or influence each other, enhancing both the interpretability and reliability of the model.

Model Training Process

The training of GAMI-Net is structured into three distinct stages:

1. **Training Main Effects**: This stage involves estimating each main effect subnetwork and then applying a pruning process 3.6.2 based on the variance contributions of these effects.

2. **Training Interaction Effects**: After the main effects are established, interaction effects that satisfy the hereditary constraint are trained and subsequently pruned using similar methods.

3. **Fine-tuning Network Parameters**: The final stage involves jointly retraining all active subnetworks, incorporating marginal clarity regularization to enhance model interpretability and performance.

Interpretability of GAMI-Net

GAMI-Net facilitates interpretability through several approaches:

1. **Importance Ratio (IR)**: This metric quantifies the contribution of each input feature to the model's overall predictions.
2. **Global Interpretation**: The model employs visual tools like plots and charts to depict the relationship between input features and the target variable.
3. **Local Interpretation**: GAMI-Net provides a detailed explanation for each prediction, breaking down the influence of input data points on the output.

Verbose

The **verbose** setting in machine learning models is a toggle for the level of detail provided by the model during training. When verbose is enabled (typically set as *True*), the model outputs extensive information about its training process. This includes progress updates, performance metrics, and potential warnings.

Verbose output is particularly beneficial for debugging and closely monitoring the model's learning process. It allows for an in-depth understanding of how the model evolves over time and helps in identifying any issues that might arise during training. In a production environment or scenarios where detailed execution information is not necessary, the verbose level can be reduced or completely turned off to minimize output.

Activation Function

The choice of an activation function in neural network architectures is crucial, as it determines the non-linear transformation applied to input data and significantly influences the model's capacity to learn complex patterns. A commonly used activation function is the Rectified Linear Unit (ReLU), defined as $f(x) = \max(0, x)$. ReLU introduces a piecewise linear non-linearity, which is effective for efficient training and particularly adept at modeling complex relationships due to its linear behavior in the positive domain.

In contrast, the sigmoid function, expressed as $f(x) = \frac{1}{1+e^{-x}}$, smoothly maps inputs to a range between 0 and 1, enabling smoother function curves but potentially leading to the vanishing gradient problem. This issue arises in deep networks where the multiplication of small derivatives during backpropagation results in exponentially smaller gradients, slowing down the learning process significantly.

In the development of the Gami-Net model, both ReLU and sigmoid activation functions will be evaluated to determine their efficacy in training and their impact on the model's performance. This comparative approach will help in identifying the most suitable activation function for capturing the intricacies of the data in our specific context.

3.7 Conclusion

This chapter has provided an in-depth exploration of various advanced machine learning models, including Explainable Boosting Machines (EBMs), Generalized Additive Models with Structured Interactions Networks (GAMI-Net). Each of these models has been dissected to understand their structure, training processes, and

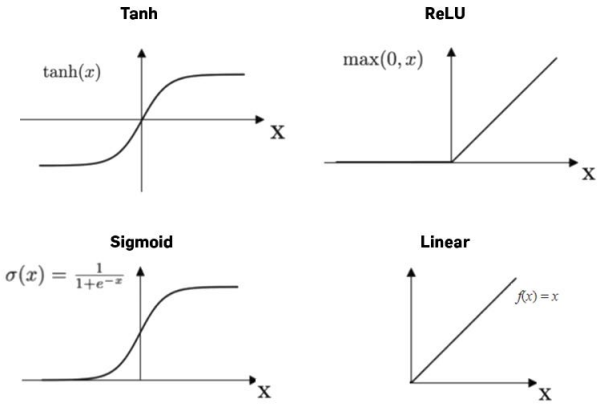
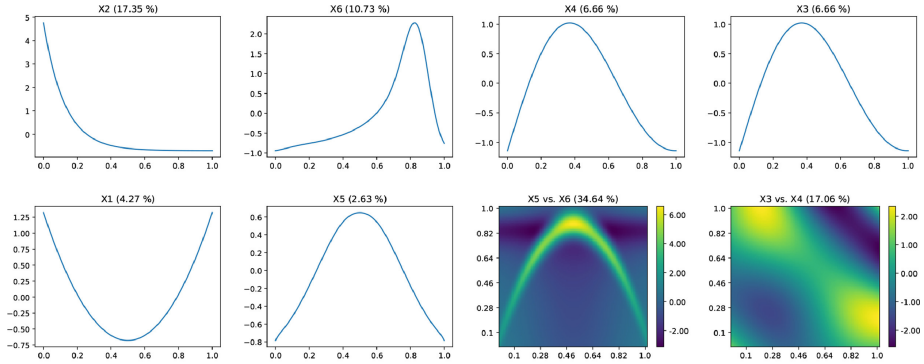
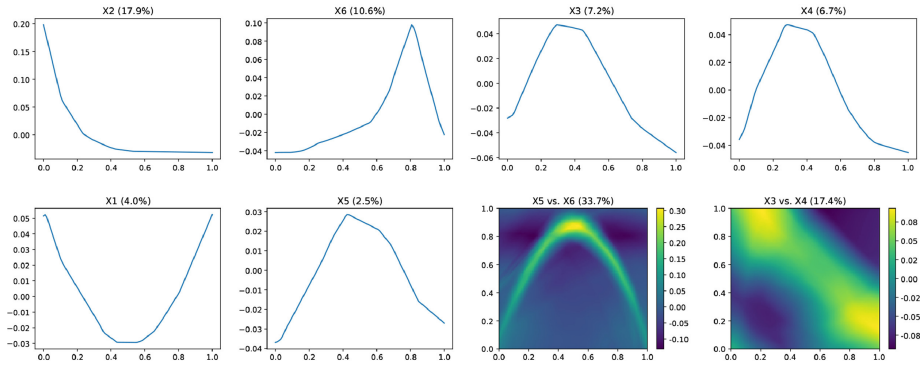


FIGURE 3.10: Activation functions

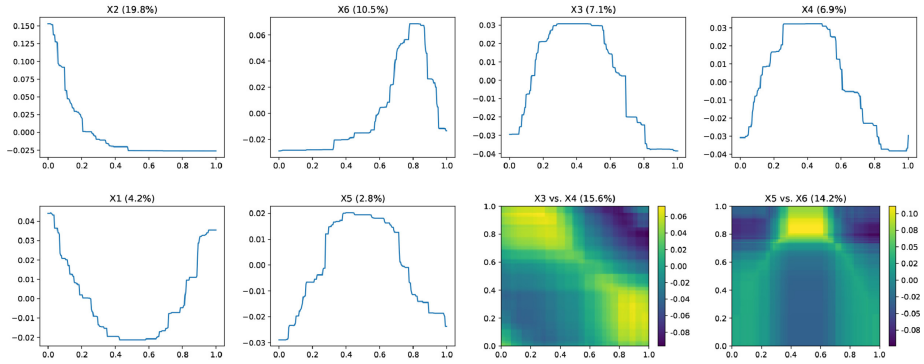
intrinsic capabilities for interpretability and predictive accuracy. The upcoming sections of this thesis will leverage these advanced machine learning approaches to predict the probability of default. The choice of these approaches is motivated by their enhanced predictive performance, ability to handle complex nonlinear relationships, and, importantly, their interpretability. This interpretability is crucial in the financial domain, where understanding the reasoning behind predictions is essential for risk assessment, regulatory compliance, and strategic decision-making. Figure 3.11 provides a comparative illustration between the original formula representations and their respective reproductions using GAMI-Net and Explainable Boosting Machine (EBM) models. The top row of the figure displays the original formula, serving as the reference benchmarks. In contrast, the bottom row showcases the outcomes of the GAMI-Net and EBM models, which have been trained to approximate these original formula. This allows for a direct visual assessment of the models' efficacy in capturing the intricate patterns and relationships inherent in the original mathematical expressions. This chapter also expressed a checklist for the evaluation phase in the CRISP-DM stages. The checklist is scheduled for use later in the thesis, following the creation of two separate models: one based on EBM and the other on GAMI-Net.



(a) Ground Truth



(b) GAMI-Net



(c) EBM

FIGURE 3.11: Comparison of Models
(Yang, Zhang, and Sudjianto, 2021)

4 Modelling

In this chapter, we go into phases CRISP-DM framework as applied to our project. We begin with an exploration of the Data Understanding phase, where we will analyze and familiarize ourselves with the dataset's characteristics and nuances. Following this, we will embark on the Data Preprocessing steps, which involve cleaning, transforming, and preparing the data for subsequent modeling. Each of these phases is essential for laying the groundwork for effective data analysis and model development. In this chapter also the modelling phase of the CRISP-DM is described.

4.1 Data selection and preparation

This thesis focuses on a dataset of wholesale customers of large corporates from a financial institution. The raw dataset is extensive, comprising 763 features and 1,363,256 rows, indicating its significant size and complexity. It's important to note that this dataset is the same one utilized by the model developers, and many of the features were specifically crafted for their proprietary model. In this discussion, each feature will be referred to by a numerical identifier, such as 'Feature 1', 'Feature 2', and so on, to maintain anonymity. Additionally, a description of these features will be provided to aid in understanding their roles and characteristics within the dataset.

4.1.1 Description of the data

The following groups were identified in the data set:

1. Static: any customer information that does not typically change over time, like country, industry and Environmental and Social Risk (ESR).
2. Financial statements: any customer information from the annual reports, like balance sheet and profit and loss.
3. Behavioural: any customer information that is for the financial institution specific and changes over time, like watchlist, historical defaults and utilisation information.
4. Qualitative: any customer information that is based on expert judgment within financial institution (e.g. assessment of quality of management).
5. External data: data from external providers, not all on customer level
6. Old model outcomes: for comparison purposes.

The target variable for this is the Default flag of 12 months. As we said in the earlier chapter a company is in default when the payment is 90 days past due. However this is not directly used in the model because then the prediction needs to be

made on the likelihood that the company will go into default in the upcoming one year period. The financial institution has therefore already made a new feature in the dataset with this taken into account. Therefore our target variable to predict will be the Default 12 months.

As is already discussed before it is a highly unbalanced dataset. The percentage of default defined as defaulted 12 months. The percentage is only 1, 3%

4.1.2 Data pre-processing

Data quality and preprocessing are fundamental steps in machine learning practices. Successful preprocessing ensures that the dataset is a reliable and suitable source for applying ML algorithms. We will also apply the data pre-processing steps for our dataset. This is visualized in the following figure 4.1

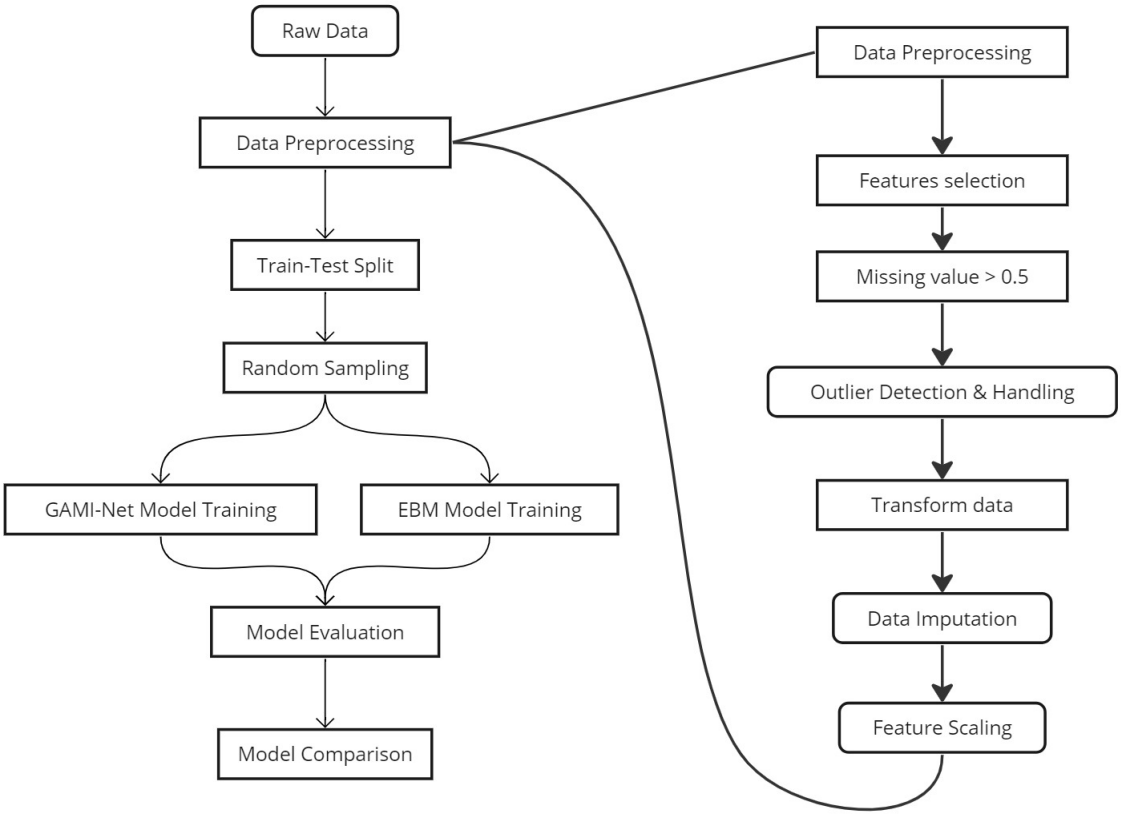


FIGURE 4.1: Model development

4.1.3 Data Cleaning and Reduction

This section will address common issues in data cleaning and discuss techniques for data reduction. It will also express the techniques used for the dataset of this thesis.

Data Cleaning

Data cleaning is the first step post data acquisition. While it might seem redundant if the data is clean, in reality, data is rarely clean. Data cleaning includes handling missing data, eliminating outliers, removing noise, and correcting inconsistencies.

- **Missing Data**: Missing values, often due to faulty sampling or limitations in data acquisition, cannot be ignored as they may lead to inaccurate conclusions.
- **Noise Elimination**: Noise refers to random variations in data. Techniques like noise polishing and filters are suggested for mitigation.
- **Inconsistencies**: Inconsistent data, such as differing notations for the same entity, must be standardized for accurate ML model performance. This is the case for some qualitative risk drivers in the data. A mapping for these is applied in order to give them the right ordinal number and to get rid of the inconsistencies. For example no dot or with a dot while the qualitative rating is the same.

Data Reduction

After cleaning, data reduction comes into play, involving feature selection, feature extraction.

Feature Selection: This process involves identifying relevant features and discarding non-informative ones, enhancing model performance and interpretability. Methods such as missing value ratio, low variance filter, high correlation filter, and Random Forest feature importance are utilized to identify the most relevant features. These steps are all done now in order to reduce the dataset quickly in order to be able to perform preprocessing steps. This was not possible with the entire dataset on the laptop provided.

Before handling missing data, initial preprocessing of the dataset involved feature selection methods to decrease the number of variables and enhance model performance. The process began by omitting features that were engineered by the creators of the model as well as principal indicators that provided no additional insight. Further reduction was implemented to consolidate financial reporting data spanning three years into single features, to compensate for any gaps in year-specific data. Lastly, numerous descriptive variables that served only to explain other variables were also minimized. 4.2

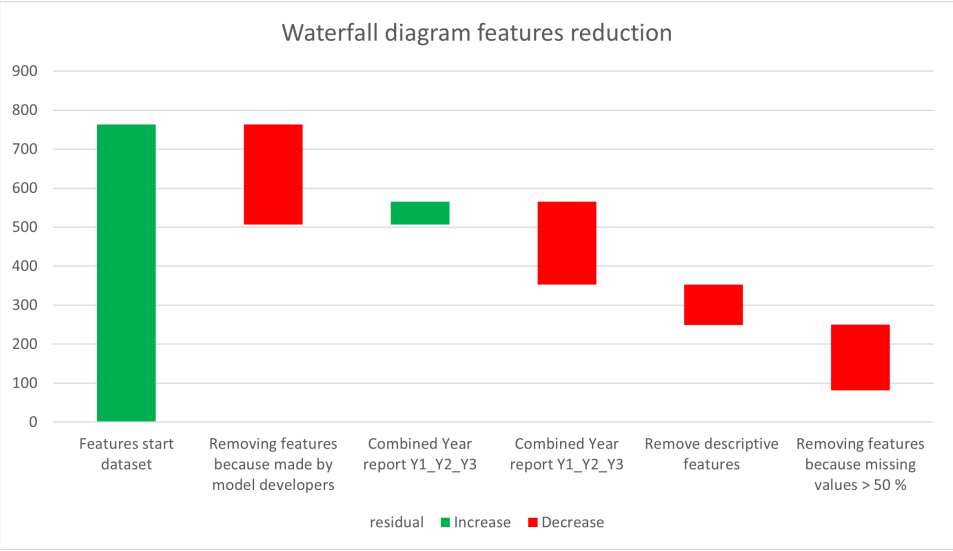


FIGURE 4.2: Waterfall diagram

Missing Value Reduction

The dataset underwent a missing value reduction process to minimize the impact of incomplete data on model performance. This step involved assessing the proportion of missing values within each feature and excluding those with a significant percentage of missing data. The rationale behind this approach is grounded in the premise that features with excessive missing values may introduce bias or noise, detracting from the model's ability to learn meaningful patterns. Consequently, features were retained based on a predetermined threshold for acceptable missing value ratios, ensuring a cleaner, more reliable dataset for analysis. This resulted in a dataset with 82 features.

High Correlation Filter

To further refine the dataset, a high correlation filter was applied to identify and remove features exhibiting substantial inter-correlation. This technique is predicated on the understanding that highly correlated features contribute redundant information, potentially obscuring the model's interpretative capacity and inflating its complexity unnecessarily. By calculating Pearson's correlation coefficient between pairs of features, those with coefficients surpassing a specified threshold indicated a high degree of redundancy. From each pair of correlated features, one was removed, thereby reducing feature redundancy without compromising the dataset's integrity. features with more than 0.9 correlation are reduced. The correlation matrix can be seen in B.5.

The Imbalanced Dataset Problem

Credit risk management often involves dealing with imbalanced datasets, especially in binary classification problems. These datasets are typically dominated by one category, leading to a large number of non-defaulting clients and a smaller number of defaulting ones. This imbalance can result in inadequate prediction models, particularly for the minority class. Looking at the dataset of this thesis with a default rate of 1.3%, we can conclude that our dataset is also unbalanced. There is some discussion into how to handle this unbalance. No clear conclusion on what is best is found in literature. We will therefore perform different procedures (Mazumder, 2021). However we will also use no treatment for handling the imbalance. To also be able to compare these results.

To address this challenge, several solutions have been proposed, including:

- **Random Over-Sampling (ROS)**: ROS involves replicating samples from the minority class to balance it with the majority class, thereby improving the performance of the ML model. However, it may lead to overfitting and does not introduce new information.
- **Random Under-Sampling (RUS)**: RUS reduces the size of the majority class to balance the dataset. While simple to implement, it risks losing important information due to the removal of majority class samples.

Random undersampling showed the best results in the different performance metrics and we will therefore also continue using this in the rest of the thesis.

Handling outliers

One critical pre-processing step in machine learning projects is the effective treatment of outliers. Outliers are atypical data points that differ significantly from other observations and can arise due to various reasons, including measurement errors, human errors, or simply as extreme yet valid observations. The challenge lies in striking a balance between identifying data errors or noise and recognizing genuine, albeit unusual, patterns. **Impact of Outliers**: If not addressed properly, outliers can substantially impair the learning ability of an algorithm. They can skew the results, leading to biased or inaccurate models. Therefore, careful consideration is required to determine whether an outlier represents a data error or a significant data point worth including in the analysis.

In this thesis, we employ the Interquartile Range (IQR) method for outlier detection. The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of a dataset's attribute. We use box plots to visualize the distribution of numerical values, showing the Median, the 25th and 75th percentiles. The whiskers of the box plot extend to represent $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, helping to identify data points that are potential outliers. After this the outliers are set to the maximum of these ranges.

Data imputation

The performance and reliability of machine learning models are significantly influenced by how missing values within the dataset are managed. It has been observed that the proposed models exhibit limitations in directly handling missing data, necessitating a strategic approach to missing value treatment. According to (Badr, 2019), identifying the nature of missing data within the dataset is a critical preliminary step. This process involves a detailed examination of individual features to classify the type of missingness they exhibit.

Types of Missing Values

Missing data can broadly be categorized into three types, each with distinct implications for data analysis and imputation strategies:

- **Missing Completely at Random (MCAR):** The likelihood of data being missing is the same across all observations. In this case, the missing data is independent of both observed and unobserved data.
- **Missing at Random (MAR):** The propensity for data to be missing is not random, but any missingness is fully accounted for by variables where complete information is available. Here, the missing data depends on the observed data but not on the missing data itself.
- **Missing Not at Random (MNAR):** The missingness is related to the unobserved data, implying that the reason data is missing may be related to its hypothetical value.

Understanding the mechanism behind missing data is crucial for selecting the appropriate imputation technique that will yield the most reliable model performance.

Imputation Methods Explored

Several imputation methods were explored to address missing values, tested cross-validation process to identify the most effective approach:

1. **Mean/Median Imputation:** Replacing missing values with the mean or median of the observed values in the same feature. This method is straightforward and often effective for MCAR data.

Feature scaling

In the process of preparing our dataset for the Generalized Additive Models with Interactions Network (GAMINet) and the Explainable Boosting Machine (EBM), we opted for Min-Max scaling as our normalization technique, guided by the insights provided by (Chong, 2023). This approach is particularly pertinent given that both models incorporate logistic regression principles to some extent. While the EBM model does not strictly require feature scaling at the initial stages of training, its application does not adversely affect the model's performance. Consequently, to maintain uniformity in data preprocessing across both models, Min-Max scaling was uniformly applied.

Min-Max scaling is a normalization strategy that rescales the numerical values of features to a standard range, typically between 0 and 1. This technique is especially beneficial for ensuring that no single feature disproportionately influences the model due to its scale. The mathematical expression for Min-Max scaling is delineated as follows:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{4.1}$$

where:

- X_{scaled} represents the scaled value.
- X denotes the original value.
- X_{min} and X_{max} are the minimum and maximum values observed for the feature across the dataset, respectively.

The rationale behind selecting Min-Max scaling for this study is twofold. Primarily, it accommodates the computational preferences of GAMINet and EBM, both of which are optimized for inputs that have been normalized. Secondly, it establishes a consistent preprocessing framework across the models, thereby enhancing the reliability and comparability of our comparative analysis.

Train-test Split

A step in the model development process is the partitioning of the dataset into training and testing subsets, a practice fundamental to the validation of machine learning models. For this study, the dataset was divided using an 80/20 split, allocating 80% of the data for model training and the remaining 20% for testing. This distribution was chosen to ensure that the models had access to a substantial volume of data for learning, thereby facilitating the development of robust and generalizable predictive algorithms.

4.2 Model Building Strategy

Two types of machine learning strategy are used. As explained in the earlier chapter we will use the EBM and the GamiNET.

We start the models training with the proposed hyperparameters by the developers and change some hyperparameters in order to fine tune the model. These results can be seen in C

4.2.1 Model Architecture

At its core, the EBM utilizes a boosting technique to sequentially train an ensemble of simple models, with each model focusing on improving the prediction accuracy in areas where previous models have performed poorly. Each of these simple models, typically shallow decision trees, captures the effect of single or interactions of multiple features. By summing up these individual contributions, EBMs provide an interpretable model that can be visualized and understood by humans (Microsoft, n.d.).

4.2.2 Hyperparameter Choices

The performance and interpretability of an EBM can be significantly influenced by the choice of hyperparameters. Key hyperparameters include:

- **Learning Rate:** Determines the step size at each iteration of the boosting process. A smaller learning rate requires more trees to model the data but can lead to a more accurate and stable model. After testing multiple learning rates the improvement for lower learning rates is best for 0.0001. To decrease the learning rates does not show high marginal increases in the performance.
- **Number of Trees:** Controls the number of boosting rounds or the number of simple models to train. A higher number of trees can capture more complex patterns but may lead to overfitting. Therefore this is tested to not overfit. The standardized number was sufficient to not overfit the model.
- **Maximum Tree Depth:** Limits the depth of each decision tree in the ensemble. A depth of 1 (decision stump) ensures maximum interpretability, while a greater depth allows for capturing interactions between features. To balance this number multiple numbers have been tested and with taking interpretability into account a number of 3 is used in the final model.
- **Min Samples Leaf:** The minimum number of samples required to be at a leaf node of a tree. This parameter helps control overfitting by providing a constraint on the granularity of the learned functions. It is tested if a larger values reduces the fitting of the model. This effect is minimal and therefore this is set to a minimum. This helps in smoothing the curves for the interpretability.

4.2.3 Model Architecture

GAMINet leverages a neural network framework to capture the complex and nonlinear relationships between features and the target variable. Its architecture consists of two main components:

4.2.4 Hyperparameter Choices

The configuration of GAMINet involves several hyperparameters that influence its performance and interpretability:

- **Number of Neurons:** Determines the capacity of the network, affecting its ability to model complex relationships. The input has been set to the values of the developers.
- **Learning Rate:** A crucial parameter for the training process, the learning rate controls how quickly the model updates its weights. Tuning this parameter is essential for balancing convergence speed and stability. To balance this
- **Interaction Strength:** A key hyperparameter unique to GAMINet, which controls the extent to which interaction effects are modeled. Adjusting this parameter allows the practitioner to prioritize between capturing additive effects and exploring complex interactions.

For the GAMINet model, specific hyperparameters were carefully selected to optimize the training process and model performance. A learning rate of 0.01 was chosen to accelerate the training process. While a faster learning rate might potentially compromise the model's accuracy, evidence from other research suggests that any reduction in performance is likely to be minimal. This balance between speed and accuracy was deemed optimal for the scope of this study (Yang, Zhang, and Sudjianto, 2021).
Additionally, the Adam optimizer was selected for use during the neural network's training phase. Renowned for its efficiency in handling large datasets and complex architectures, the Adam optimizer facilitates a more effective convergence to optimal features by dynamically adjusting the learning rate. Its ability to navigate the challenges of both sparse gradients and noise makes it particularly suited for refining the neural network within the GAMINet framework.
The number of epochs will be set to ensure sufficient training time without causing overfitting.

Python was selected as the programming language for the machine learning models in this research due to its robust capabilities in managing large datasets and its comprehensive support for data manipulation and analysis. The preprocessing of data was effectively conducted using Python, taking advantage of its powerful data processing libraries. Furthermore, the wide array of machine learning libraries available in Python, including those pertinent to the algorithms used in this thesis, significantly influenced the decision. PyCharm was chosen as the integrated development environment (IDE) for this project, offering a conducive and efficient workspace for Python development with its rich set of features and tools.

5 Results

In Chapter 5, we share the outcomes of our newly developed models, starting with their results and the features that make them explainable. Next, we compare these models to the checklist we created in Chapter 3. This step helps us see how well each model meets our set criteria, highlighting their strengths and areas for improvement. The final section of this chapter scores the two models based on their performance. Looking at the used methodology we now are in the evaluation phase of the CRISP-DM. These scores will be crucial in Chapter 6, where we'll draw our final conclusions for this thesis, combining what we've learned from our models with the broader aims of our research.

5.1 Overview of Results

5.1.1 EBM

The performance of the EBM model, as measured by AUROC and AUPRC, is showcased in Figure 5.1. This figure shows the performance metrics of both the EBM and GamiNet models, offering a comparative insight. As can be seen they both perform similar. But the performance of the Gami-Net is slightly higher when looking in the left graph this can be because of the use of the neural network and discarding less important features. However this is not shown in the precision recall where we can see that this is not of high performance in the precision. This is both in the prediction of the false positives as well as the false negatives.

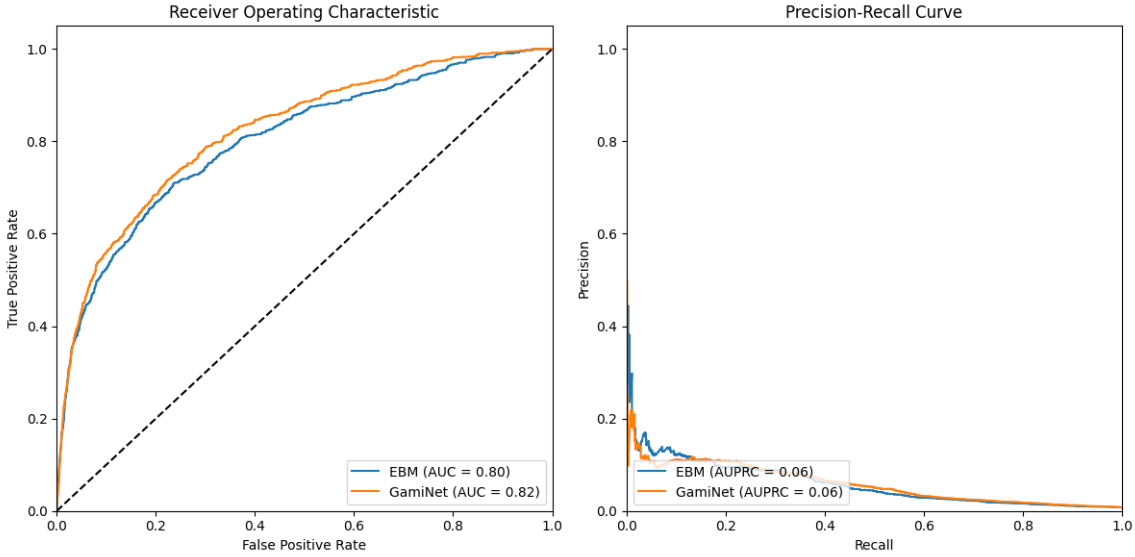


FIGURE 5.1: Performance metrics of EBM and GamiNet

Explainability of Features

A crucial element of our analysis involves discerning the features leveraged by the

model. To achieve this, we delved into feature importance, as depicted in Figure 5.2. This global perspective highlights the relative significance of each feature, providing critical insights for stakeholders. For model users and regulators, this aspect is particularly vital as it enables an evaluation of the model's economic rationale. Essentially, it assesses whether each feature logically contributes to the final prediction. Should a feature's contribution not be readily apparent or logical to both the user and the regulator, its inclusion in the final model may be reconsidered. The analysis of both feature importance and feature functions, as shown in 5.3, is instrumental in facilitating this understanding.

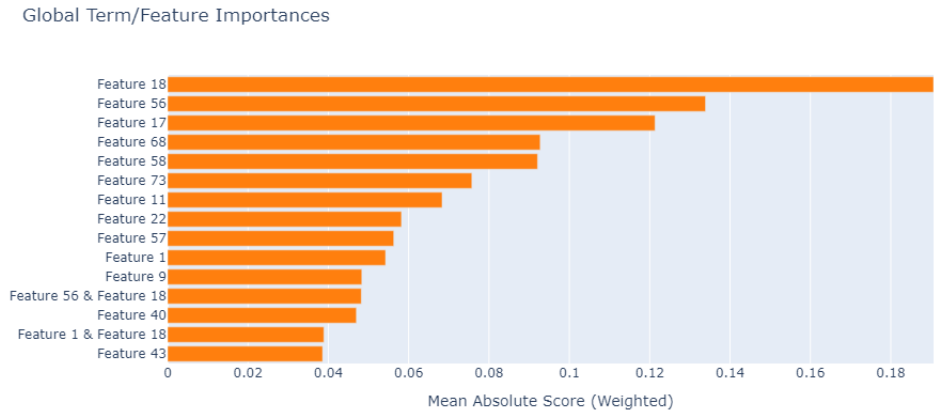


FIGURE 5.2: Feature importance of EBM

The influence of company ratings (Feature 4) on default risk is elucidated in Figure 5.3. The analysis indicates a direct correlation between lower ratings and an increased likelihood of default, a pattern that underscores the pivotal role of company ratings in financial stability assessments.

Moreover, local explanations offer a more granular view of the decision-making process by illuminating the role of individual features in shaping the final decision, as shown in Figure 5.4. These detailed explanations are instrumental in addressing stakeholder inquiries regarding whether a specific company is in default and elucidating the reasons behind such a determination.

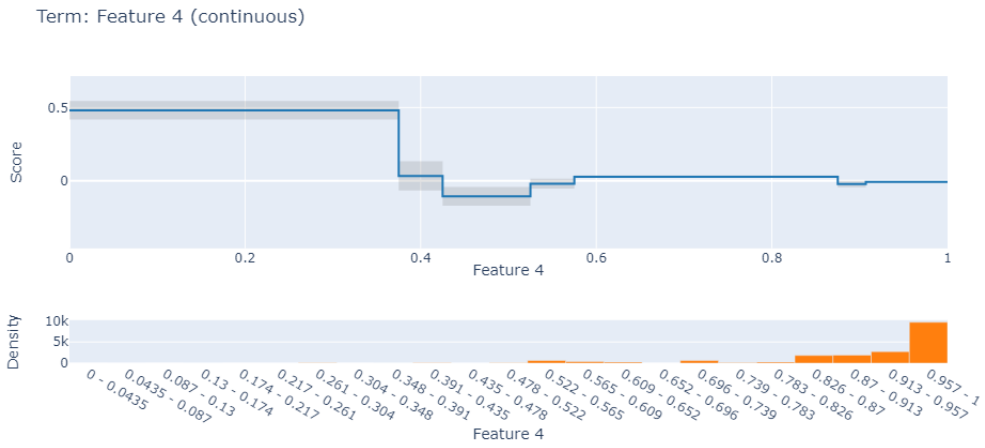


FIGURE 5.3: Feature 4 influence on outcome

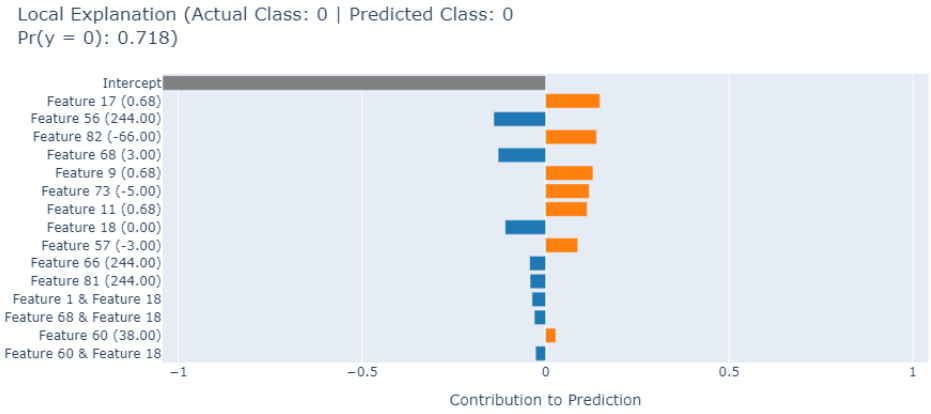


FIGURE 5.4: Local explanation of EBM

5.1.2 Gami-Net

The optimal features and training stages of the Gami-Net model are detailed in Figures 5.5 and 5.6. These figures offer a comprehensive view of the model's development and optimization efforts, showcasing the optimal count of main effects and interactions. This process is crucial for enhancing the model's interpretability by eliminating unused features, thereby streamlining the model and making it more understandable.

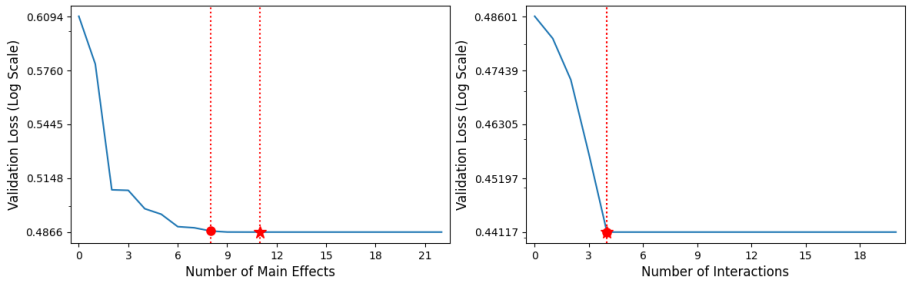


FIGURE 5.5: GamiNet optimal features

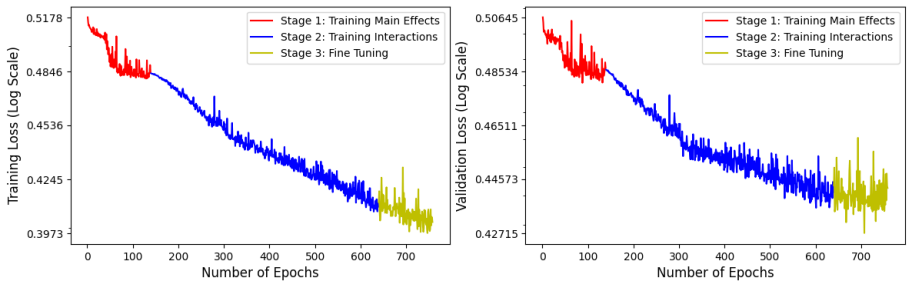


FIGURE 5.6: GamiNet training phases

Explainability of Features

The feature visualization for the GamiNet model, depicted in Figure 5.7, sheds light on how each feature influences the model's predictions. Particularly notable

are the plot discontinuities, which underscore the effects of mean imputation on the model's performance. Similar to the EBM model, GamiNet also emphasizes feature importance. However, a distinct observation is that GamiNet's visual representations tend to exhibit smoother lines without abrupt transitions. This difference stems from the underlying architectures of the two models: EBM is built on a tree-based framework, whereas GamiNet operates on a neural network basis, leading to the smoother transitions observed in its feature visualizations.

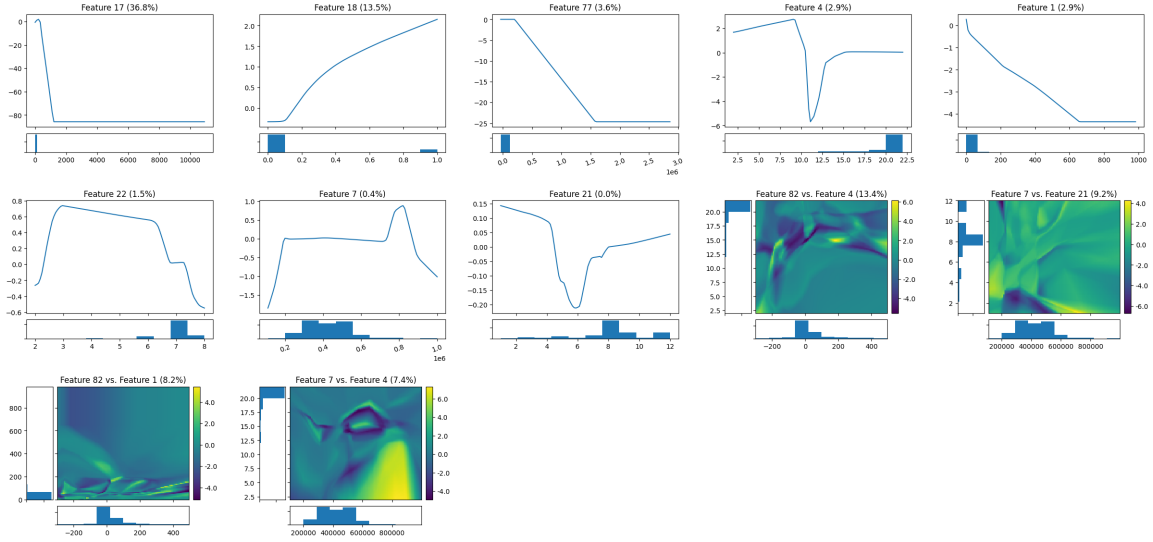


FIGURE 5.7: GamiNet features final model

Enhancing model explainability, especially through the analysis of individual predictions, is crucial. Figure 5.8 offers an in-depth view of how each input feature influences the model's final prediction, underlining the importance of transparency throughout the modeling process. This level of detail is particularly valuable when responding to regulatory inquiries about the rationale behind specific Probability of Default (PD) assignments for individual companies. The figure illustrates the contribution of each feature to the prediction outcome; for instance, feature 22 exerts a negative impact, nudging the prediction closer to 0, whereas feature 1 has a positive influence, steering the prediction towards a default (1). Such insights are not only pivotal for regulators but also for model validators as internal stakeholders, and the model users, who gain a clearer understanding of how specific company characteristics are reflected in the predictions. Incorporating this level of explainability is a significant enhancement over current models and is deemed essential in models developed using machine learning techniques.

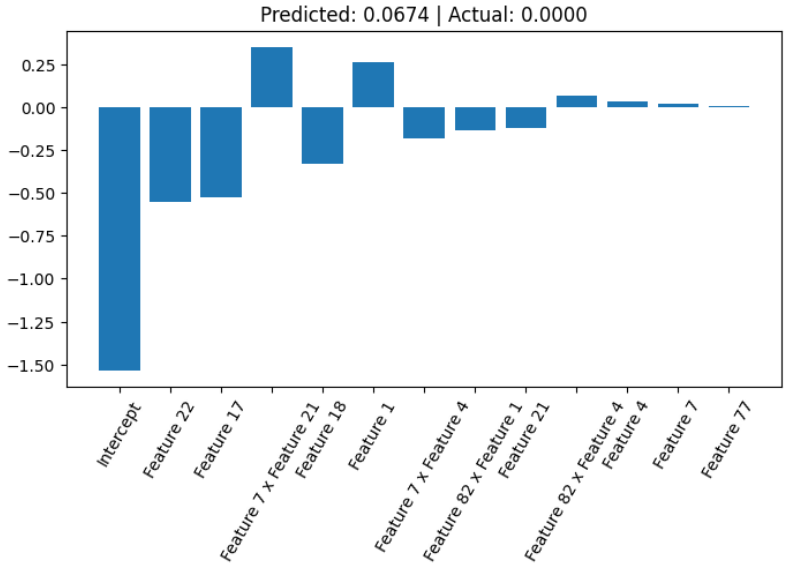


FIGURE 5.8: Local explanation GamiNet

5.2 Validation

The evaluation of both models will be conducted through cross-validation and testing on a random dataset. The outcomes of these evaluations are presented below.

Stratified cross validation
In this section, the validation of both models will be carried out using stratified cross-validation. This method is particularly suited for handling imbalanced datasets, as previously discussed. The findings are detailed in the results section (see Table 5.1). These cross-validation outcomes highlight the EBM model's consistent performance in accurately classifying across various subsets of the data.

TABLE 5.1: Stratified Cross-Validation Results for the EBM Model

Fold	AUROC	AUPRC
0	0.9117	0.2194
1	0.9021	0.1867
2	0.9048	0.2079
3	0.8899	0.1920
4	0.9101	0.1935
Mean	**0.9037** ($\pm$ **0.0077**)	**0.1999** ($\pm$ **0.0120**)

Random dataset test
The method of validating model performance by testing them with random data is a technique designed to assess how models handle non-informative input. This process involves substituting the original training set's features with randomly generated data while retaining the original structure and number of features. The random data, uniformly distributed between 0 and 1, replaces each feature in the dataset, creating a scenario where the input data lacks meaningful information. Subsequently, both models are retrained with this newly generated random dataset to evaluate their performance in conditions where the input data do not provide informative cues.

TABLE 5.2: Stratified Cross-Validation Results for the Gami-Net Model

Fold	AUROC	AUPRC
0	0.69599	0.03345
1	0.69653	0.03352
2	0.69599	0.03365
3	0.69631	0.03454
4	0.68207	0.03210
Mean	**0.6934** ($\pm$ **0.0057**)	**0.0335** ($\pm$ **0.0008**)

Upon implementing this approach, it was observed that the Explainable Boosting Machine (EBM) produced results that were lower than anticipated. Ideally, when models are trained with non-informative (random) data, we would expect their performance metrics, such as accuracy or Area Under the Curve (AUC), to converge towards 0.5, reflecting the outcome of random guessing. The deviation from this benchmark suggests that the EBM might be overfitting to the random data or improperly processing the lack of informative signals.

On the other hand, GAMI-Net displayed results that were almost identical to those obtained from stratified cross-validation. This similarity could be attributed to the imbalance present in the dataset, leading GAMI-Net to predict a higher number of non-defaults. Such behavior in the presence of unbalanced datasets is a critical factor to consider, as it can mask the true predictive performance of a model, making it appear more effective than it actually is when dealing with non-informative data.

TABLE 5.3: Performance Metrics on Random Dataset Test

Model	AUROC	AUPRC
EBM	0.40871	0.00581
GamiNet	0.61005	0.01052

5.3 Comparison with benchmark model

In the evaluation of our machine learning models against the financial institution's benchmark, an intriguing observation emerged regarding the nature of feature engineering employed in the benchmark model. The logistic regression-based benchmark model incorporates sophisticated feature engineering techniques, notably including conditional statements and the application of logistic regression or maximization on raw features. This approach not only enhances the model's performance but also aligns with the conceptual framework of Generalized Additive Models (GAMs). By transforming features in a manner that mirrors the operational principles of GAMs—where the relationship between the dependent variable and each feature is modeled separately—this model inadvertently leverages the interpretive and predictive strengths characteristic of GAMs.

This revelation provides a compelling explanation for the benchmark model's robust performance and offers insights into why the explicitly GAM-based models explored in this study, such as the Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet), did not achieve superior results. The benchmark model's implicit use of GAM-like feature transformations underscores the potency of such techniques in extracting and utilizing the

predictive information embedded within the features. It suggests that the benchmark model's success is not solely attributable to the logistic regression algorithm but is significantly bolstered by the GAM-like feature engineering practices it employs.

The convergence between the benchmark model's feature engineering strategy and the operational essence of GAMs elucidates the foundational role of thoughtful feature manipulation in enhancing model performance. This analysis underscores the critical importance of feature engineering in the development of predictive models, particularly in the financial sector where interpretability and accuracy are paramount. It implies that the integration of GAM principles through feature engineering can substantially contribute to a model's effectiveness, even outside the explicit use of GAM-based algorithms.

5.4 Checklist review on models

5.4.1 Comparative Evaluation of EBM and GAMINet

This section offers a comparative analysis between Explainable Boosting Machines (EBM) and Generalized Additive Models with Interactions Network (GAMINet), based on key criteria that are crucial for interpretable machine learning applications. Our objective is to delineate the distinct characteristics, strengths, and potential use cases for each model. It's important to note that this comparison is intended as a reasoned evaluation, enriched by the knowledge and experience gained through the construction and optimization of the models. The evaluation will utilise a scoring system of low, medium, and high. A high score indicates that a model meets all the criteria listed in the checklist without any shortcomings. A medium score is assigned when a model meets some but not all requirements, suggesting room for enhancement. A low score is awarded when a model fails to meet any of the checklist criteria.

5.4.2 Explainability and Transparency

EBM Score: High **GAMINet Score:** High
Conclusion: Both EBM and GAMINet exhibit high levels of explainability and transparency. EBM provides clear insights into input feature impacts, while GAMINet extends this with neural networks to capture complex interactions, maintaining interpretability. This is showcased in the figures shown before. The only less transparent and explainable part is the training phase but this is not expected to be a problem.

5.4.3 Regulatory Compliance

EBM Score: Medium to High **GAMINet Score:** Medium to High
Conclusion: The interpretability inherent to both models plays a pivotal role in ensuring regulatory compliance, although the degree of compliance also hinges on the thoroughness of documentation and the specifics of each use case. The Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet) both promote transparency in how decisions are made, a crucial factor for adhering to regulatory standards. Furthermore, both models meet the requirements for data usage, utilizing the same dataset that was employed in the development of the existing model, which has already been validated as suitable for such purposes. The capability to clearly articulate the role and impact of each feature

within the models, along with a judicious selection of features to avoid unnecessary complexity, aids in simplifying the final model. This approach not only supports regulatory compliance but also enhances the models' transparency and interpretability, aligning with regulatory expectations for clear and understandable decision-making processes.

5.4.4 Ethical Considerations and Bias Management

EBM Score: Medium to High **GAMINet Score:** Medium to High
Conclusion: The inherent transparency of each model plays a crucial role in identifying and mitigating bias, highlighting the importance of proactive efforts to manage biases and maintain fairness in model predictions. Such measures emphasize the necessity of careful data management and thorough model evaluation processes. While bias management may be less critical for applications involving large corporates, probability of default (PD) modeling is also applied to consumer data, where fairness and bias mitigation become significantly important. It is essential to recognize and address these concerns. Fortunately, both models are equipped to handle these challenges, providing mechanisms to ensure that bias identification and mitigation are integral to their operation.

5.4.5 Complexity vs. Interpretability Balance

EBM Score: High **GAMINet Score:** High
Conclusion: The Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet) both excel in striking a commendable balance between managing complexity and ensuring interpretability. EBM achieves this by distilling complex relationships into more manageable forms using decision trees and Generalized Linear Models (GLMs), while GAMINet leverages neural networks to conduct thorough interaction analyses without compromising on clarity. Notably, these advanced techniques are utilized during the training phase, but the resultant models bear closer resemblance to a linear regression model, providing a clear indication of the influence each feature exerts on the outcome. This approach facilitates the generation of interpretable results, showcasing the contribution of individual features. The complexity inherent in both models is subject to the developer's discretion, allowing for considerable flexibility in model configuration. EBM offers the option to utilize very shallow trees, simplifying its structure, whereas the complexity of GAMINet's neural network can be adjusted by modifying the network's architecture. Such adaptability enhances the models' manageability and allows for a tailored approach to balancing performance with interpretability. Although this customization may slightly impact performance, both models maintain a high level of effectiveness, underscoring their utility in applications where understanding the model's decision-making process is as critical as achieving accurate predictions.

5.4.6 Stakeholder Acceptance and Trust

EBM Score: High **GAMINet Score:** High
Conclusion: The clarity and interpretability offered by both models significantly contribute to enhancing trust and acceptance among stakeholders. The Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet) are designed to provide users with a clear understanding of how

predictions are made. This transparency is achieved by showcasing the functions and contributions of various features within the models, allowing stakeholders to see exactly how each element influences the final outcome. Additionally, the ability to delve into more localized explanations and understand the decision-making process further bolsters stakeholder confidence. This level of insight into the models' operations not only fosters trust but also facilitates broader acceptance across different stakeholder groups, underlining the value of interpretability in complex modeling solutions.

5.4.7 Performance and Accuracy

EBM Score: Medium **GAMINet Score:** Medium to High
Conclusion: While the Explainable Boosting Machine (EBM) showcases strong performance, particularly with structured datasets, the Generalized Additive Models with Interactions Network (GAMINet) is designed to achieve even higher predictive accuracy. Yet, this enhanced accuracy has not been fully demonstrated with the current dataset, leading to a provisional medium evaluation. Enhancements to the dataset could potentially elevate performance levels, as this would reduce the need for extensive assumptions during the modeling process.

Adherence to regulatory guidelines, such as those established by the European Central Bank (ECB), imposes limitations on the extent of data that can be used for training. This regulatory compliance necessitates that the models not only predict accurately but also maintain explainable relationships within the data. A further factor contributing to the medium performance rating is the challenge posed by the dataset's imbalance, which impacted the models' predictive effectiveness. Given that dataset imbalance is a common issue and likely to persist, it's crucial to account for it as an inherent characteristic of the data in future modeling efforts.

5.4.8 Data Efficiency and Robustness

EBM Score: Medium **GAMINet Score:** Medium to High
Conclusion: The Explainable Boosting Machine (EBM) demonstrates efficient data handling and robust performance, though it has a tendency to overfit under certain conditions. Conversely, the Generalized Additive Models with Interactions Network (GAMINet) is capable of capturing complex relationships more robustly, provided it is correctly set up and the data is appropriately managed. The flexibility to select the number of features and their interactions enhances the robustness of both models. In terms of data efficiency, EBM has an advantage, evidenced by shorter training times compared to GAMINet. This suggests that EBM can process and learn from data more quickly, making it a practical choice for scenarios where speed is a priority. Nonetheless, both models are adept at handling various types of data, including numerical and categorical variables. They commonly employ one-hot encoding to manage categorical features, a process which converts these features into a form that can be effectively used by the models. However, any alternative treatment of categorical data should be addressed during the data preprocessing phase, before model training begins. This ensures that both EBM and GAMINet operate on data that is optimally prepared for their specific learning algorithms.

5.4.9 Operational Feasibility

EBM Score: Medium to High **GAMINet Score:** High
Conclusion: Integrating the Explainable Boosting Machine (EBM) into current IT systems presents a moderate challenge, whereas incorporating the Generalized Additive Models with Interactions Network (GAMINet) can demand significantly more computational resources due to its use of neural networks. Despite this, with the right infrastructure in place, both remain viable options. The financial institution primarily utilizes SAS[1], which does support machine learning capabilities. However, the machine learning models discussed are designed for implementation with Python-based libraries, and adapting these models for use within the institution's existing SAS framework may introduce some obstacles. Given the complexities associated with neural network models, it is anticipated that implementing GAMINet would be more challenging than deploying EBM, categorizing the former's integration difficulty as high and the latter's as medium to high.

5.4.10 Maintenance and Adaptability

EBM Score: Medium **GAMINet Score:** High
Conclusion: The Explainable Boosting Machine (EBM), with its more structured framework, tends to be easier to maintain and fine-tune. In contrast, the complexity of the Generalized Additive Models with Interactions Network (GAMINet) poses some challenges in terms of adaptability, although these can be effectively managed with careful monitoring. Training or retraining the GAMINet model typically demands more time than the EBM, making it somewhat less efficient in this regard. However, for models that are already well-established and require less frequent retraining, this becomes a less significant issue. A unique advantage of EBM is its capability to manually adjust the scores or weights assigned to different terms, offering a level of customization not as readily available in GAMINet. This feature of EBM allows for more direct control over the model's behavior, which is particularly beneficial when fine-tuning the model's response to specific factors, a task that proves to be more complex in the GAMINet framework.

Summary: EBM and GAMINet stand out for their balance of interpretability and advanced predictive capabilities, each with unique strengths in modeling complex data relationships. EBM offers a straightforward approach to explainability with a focus on feature impact, making it highly accessible. In contrast, GAMINet excels in capturing intricate interactions and delivering superior predictive performance, albeit with potentially higher computational demands. Both models represent valuable tools in the interpretable machine learning arsenal, suitable for a variety of applications where model understanding and transparency are paramount.

[1] https://www.sas.com/nl_nl/insights/analytics/machine-learning.html

6 Conclusions and Discussion

6.1 Conclusions

In the final analysis of our exploration, the performance of the newly implemented model does not surpass that of the pre-existing model within the financial institution. This finding not only validates the efficacy of the current system but also invites a deeper inquiry into potential enhancements and the specific challenges at hand.

This research aimed to elucidate the role of explainable machine learning (XML) in refining PD models, considering the imperative of stakeholder requirements. It becomes clear that XML methodologies can be seamlessly incorporated into the prevailing modeling framework. Our investigation demonstrates that while XML models may initially lack transparency during their training phase, they significantly regain this attribute once trained, thereby meeting the crucial demand for intelligible and logical feature application. This characteristic is vital for fostering stakeholder trust, predicated on the understandability of the model's predictive mechanisms.

Key Insights: The comparative analysis of the two models underscores the potential of machine learning, especially XML, to innovate PD modeling within the current financial ecosystem. The emphasis by stakeholders on the rationality of feature utilization over the intricacies of the model's training phase underscores a preference for outcome interpretability and the logic behind decision-making processes.

To effectively integrate XML in PD modeling, it is essential to:

- Ensure post-training interpretability, offering clear, logical explanations for predictions to stakeholders.
- Maintain an economically sound and comprehensible rationale for feature selection and utilization within the model.
- Continually explore and integrate XML techniques to discover new insights and improve predictive accuracy without compromising transparency.

In response to how explainable machine learning can be leveraged for PD models while satisfying stakeholder requirements, this study charts a course forward. By harnessing the strengths of XML, we can enhance both model performance and stakeholder engagement, ensuring that advancements in machine learning are seamlessly integrated with the institution's predictive modeling practices. A good conclusion is that it is possible to use machine learning techniques in the current environment.

6.2 Discussion

This section delves into the implications of the findings from the comparative analysis of Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet) in the context of developing Probability of Default (PD) models. The evaluation centered on cross-validation results highlights several

critical aspects of integrating explainable machine learning (XML) techniques within financial modeling frameworks.

Performance Insights: The cross-validation process revealed that while both EBM and GAMINet exhibit high levels of explainability and performance, there are nuanced differences in their applicability to PD modeling. EBM's strength lies in its simplicity and direct interpretability, making it exceptionally suited for scenarios where stakeholders require straightforward explanations of model predictions. Conversely, GAMINet's ability to capture complex interactions offers a deeper, albeit slightly less direct, level of interpretability, suggesting its potential in applications demanding a granular understanding of inter-feature relationships.

Stakeholder Considerations: The analysis underscores the importance of stakeholder requirements in choosing the appropriate XML technique. While both models meet the threshold for transparency and explainability, the selection between EBM and GAMINet should be informed by the specific needs of stakeholders, including regulatory bodies, model developers, and end-users. The necessity for models to provide logical and economically sensible explanations for their predictions cannot be overstated, particularly in the highly regulated finance sector.

Integrating XML Techniques: The findings advocate for a balanced approach to integrating EBM and GAMINet within existing PD modeling frameworks. This involves leveraging EBM for its interpretability and efficiency in scenarios requiring rapid, clear decision-making. Simultaneously, GAMINet's advanced analytical capabilities could be harnessed to enhance model sophistication, especially in complex modeling environments where interaction effects are critical.

6.3 Limitations

This research and the development of the models under study have encountered several limitations that are important to acknowledge. Understanding these limitations provides context for the findings and guides future research directions.

Interpretation of External Stakeholder Perspectives: The analysis of external stakeholders' views is primarily based on documents available online. This approach might not capture the full spectrum of external stakeholder opinions, potentially overlooking some aspects. However, this limitation is considered minimal because the models are subject to stringent regulations. Compliance with these regulations is presumed to meet the expectations of external stakeholders, thereby mitigating the impact of this limitation.

Time Constraints in Model Creation: The time allocated for developing the model, particularly in the aspect of feature development, presented a limitation. While the development team did explore feature development, merely adopting their features without further investigation does not fully explore the potential of machine learning in enhancing the model. The decision to not delve deeper into feature development is expected to impact model performance negatively.

Model Runtime and Computational Resources: All model training, testing, and processing were conducted on a single laptop, leading to extended training times

and limited testing of various methodologies. This was particularly challenging for the imputation of missing values. To accommodate the limited computational resources, adjustments were made to the model's learning rates and the number of iterations, which allowed the model training to complete but might have constrained performance improvements. Utilizing more computational power could enhance the model, although significant leaps in performance are not anticipated.

6.4 Practical Implications and Recommendations

6.5 Potential Future Research Directions

This thesis has laid a foundational understanding of the current state of machine learning applications within financial institutions, particularly focusing on the Explainable Boosting Machine (EBM) and Generalized Additive Models with Interactions Network (GAMINet). While significant strides have been made, there remain opportunities for further research that can enhance the interpretability, efficiency, and accuracy of these models. Below are outlined several promising avenues for future investigation.

Integration of Python Packages into SAS Environments SAS remains the predominant tool within banking sectors, noted for its robust data handling and statistical capabilities. However, it lacks the comprehensive machine learning and advanced analytical packages found in Python. Investigating methodologies for seamlessly integrating Python's extensive libraries into SAS environments represents a crucial research direction. This integration could facilitate the adoption of sophisticated machine learning techniques, such as EBM and GAMINet, directly within the existing SAS-based workflows, significantly advancing the analytical capabilities within financial institutions without disrupting established systems.

Advanced Feature Transformation Techniques Preliminary analysis has demonstrated that the performance of machine learning models, including EBM and GAMINet, can be markedly improved through the application of feature transformations developed by model engineers. Future research should delve deeper into these transformations, exploring both their theoretical underpinnings and practical applications. This investigation could yield insights into how best to preprocess data to enhance model accuracy and interpretability, offering a pathway to more sophisticated modeling approaches that leverage existing knowledge within financial institutions.

Exploration of Diverse Datasets and Data Quality Enhancement The performance of machine learning models is inherently tied to the quality and characteristics of the underlying data. Further research is warranted to assess the adaptability and robustness of models like EBM and GAMINet across different datasets, including those with varying degrees of quality and complexity. This exploration should include rigorous testing of the models on higher-quality datasets and the development of strategies to mitigate the impact of data imperfections. Additionally, extending model evaluations to include diverse financial contexts and datasets can provide a more comprehensive understanding of their applicability and limitations, paving the way for more generalized and robust machine learning solutions in finance.

Using more qualitative risk driver The current model incorporates qualitative risk

drivers primarily derived from the financial institution's internal assessments, adhering to established guidelines to ensure stability and minimize bias. However, a promising avenue for future research lies in the integration of large language models to analyze market sentiment or corporate communications. This approach could potentially uncover novel connections and insights that are not readily apparent through traditional risk assessment methods. Although this would necessitate the collection of new data types, initiating this process could significantly enhance the dataset's richness. By embedding sentiment analysis into the dataset, future models could be developed to leverage these nuanced indicators, offering a more dynamic and informed perspective on risk. This innovation represents a forward-thinking expansion of the model's analytical capabilities, with the potential to refine risk prediction methodologies in the coming years.

Bibliography

Addo, Peter Martey, Dominique Guegan, and Bertrand Hassani (June 2018). "Credit Risk Analysis Using Machine and Deep Learning Models". en. In: *Risks* 6.2. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 38. ISSN: 2227-9091. DOI: 10.3390/risks6020038. URL: https://www.mdpi.com/2227-9091/6/2/38 (visited on 09/04/2023).

Akturk, Mehmet (Apr. 2021). *What Is the "Boosting" Ensemble Method?* en. URL: https://mathchi.medium.com/what-is-the-boosting-ensemble-method-76610a5cb39f (visited on 01/11/2024).

Association, International Communication (2016). *Humans less likely to return to an automated advisor once given bad advice*. en. URL: https://phys.org/news/2016-05-humans-automated-advisor-bad-advice.html (visited on 08/31/2023).

Badr, Will (Jan. 2019). *6 Different Ways to Compensate for Missing Data (Data Imputation with examples)*. en. URL: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779 (visited on 02/12/2024).

Barredo Arrieta, Alejandro et al. (June 2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: perfp (visited on 01/11/2024).

Baruah, Indraneel Dutta (Dec. 2023). *How Do Inherently Interpretable AI Models Work? — General Additive Models*. en. URL: https://indraneeldb1993ds.medium.com/how-do-inherently-interpretable-ai-models-work-general-additive-models-72d0b29daf4b (visited on 01/11/2024).

Belle, Vaishak and Ioannis Papantonis (July 2021). "Principles and Practice of Explainable Machine Learning". English. In: *Frontiers in Big Data* 4. Publisher: Frontiers. ISSN: 2624-909X. DOI: 10.3389/fdata.2021.688969. URL: https://www.frontiersin.org/articles/10.3389/fdata.2021.688969 (visited on 03/08/2024).

Bellini, Tiziano (Jan. 2019a). "Chapter 1 - Introduction to Expected Credit Loss Modelling and Validation". In: *IFRS 9 and CECL Credit Risk Modelling and Validation*. Ed. by Tiziano Bellini. Academic Press, pp. 1–30. ISBN: 978-0-12-814940-9. DOI: 10.1016/B978-0-12-814940-9.00009-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128149409000098 (visited on 01/18/2024).

— (Jan. 2019b). "Chapter 2 - One-Year PD". In: *IFRS 9 and CECL Credit Risk Modelling and Validation*. Ed. by Tiziano Bellini. Academic Press, pp. 31–89. ISBN: 978-0-12-814940-9. DOI: 10.1016/B978-0-12-814940-9.00010-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128149409000104 (visited on 01/11/2024).

BIS (Dec. 2017). "Basel III: Finalising post-crisis reforms". en. In: URL: https://www.bis.org/bcbs/publ/d424.htm (visited on 09/04/2023).

Breeden, Joseph (May 2020). *A Survey of Machine Learning in Credit Risk*. DOI: 10.13140/RG.2.2.14520.37121.

Bussmann, Niklas et al. (Jan. 2021). "Explainable Machine Learning in Credit Risk Management". en. In: *Computational Economics* 57.1, pp. 203–216. ISSN: 1572-9974. DOI: 10.1007/s10614-020-10042-0. URL: https://doi.org/10.1007/s10614-020-10042-0 (visited on 08/22/2023).

Caruana, Rich et al. (Aug. 2015). "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission". en. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney NSW Australia: ACM, pp. 1721–1730. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788613. URL: https://dl.acm.org/doi/10.1145/2783258.2788613 (visited on 01/09/2024).

Chong, Jason (July 2023). *What is Feature Scaling & Why is it Important in Machine Learning?* en. URL: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048 (visited on 02/12/2024).

Chong, Pei Swee, Jane Labadin, and Farid Meziane (2022). "Credit Risk Prediction for Peer-To-Peer Lending Platforms: An Explainable Machine Learning Approach". en. In: *Journal of Computing and Social Informatics* 1.2. DOI: 10.33736/jcsi.4761.2022. URL: https://doi.org/10.33736/jcsi.4761.2022 (visited on 03/08/2024).

Clercq, Paul le (Jan. 2023). *Bankkantoren met een loep zoeken: ABN heeft er nog maar 27*. nl. URL: https://www.rtlnieuws.nl/economie/bedrijven/artikel/5360095/bank-kantoor-abn-amro-ing-bank-rabobank-sns-regiobank-volksbank (visited on 09/04/2023).

Coenen, Lize, Wouter Verbeke, and Tias Guns (2021). "Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods". en. In: *Journal of the Operational Research Society* 73.1. DOI: 10.1080/01605682.2020.1865847. URL: https://doi.org/10.1080/01605682.2020.1865847 (visited on 03/08/2024).

Commission, European (2018). *General Data Protection Regulation (GDPR) – Official Legal Text*. en-US. URL: https://gdpr-info.eu/ (visited on 08/22/2023).

— (Aug. 2023a). *EU AI Act: first regulation on artificial intelligence | News | European Parliament*. en. URL: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (visited on 08/22/2023).

— (June 2023b). *Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future*. en. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (visited on 08/31/2023).

Demirtaş, Naime and Orhan Dalkılıç (Jan. 2021). "CONSISTENCY MEASUREMENT USING THE ARTIFICIAL NEURAL NETWORK OF THE RESULTS OBTAINED WITH FUZZY TOPSIS METHOD FOR THE DIAGNOSIS OF PROSTATE CANCER". In: *TWMS Journal of Applied and Engineering Mathematics* 11, pp. 237–249.

EBA (2017). *Guidelines on PD and LGD estimation (EBA-GL-2017-16)_EN.pdf*. URL: e (visited on 08/22/2023).

— (July 2019). *Capital Requirements Regulation (CRR)*. en. URL: https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/504 (visited on 08/22/2023).

— (2021). *Discussion paper on machine learning for IRB models.pdf*. URL: https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf (visited on 08/22/2023).

EBA (2023). *Follow-up report on machine learning for IRB models.pdf*. URL: https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf (visited on 08/22/2023).

European Central Bank. (2019). *ECB guide to internal models*. en. LU: Publications Office. URL: https://data.europa.eu/doi/10.2866/849746 (visited on 08/23/2023).

Gobat, Jeanne (Mar. 2012a). "Back to Basics: What Is a Bank?: Institutions that match up savers and borrowers help ensure that economies function smoothly". en. In: *Finance & Development* 49.001. ISBN: 9781451922141 Publisher: International Monetary Fund Section: Finance & Development. DOI: 10.5089/9781451922141.022.A011. URL: https://www.elibrary.imf.org/view/journals/022/0049/001/article-A011-en.xml (visited on 12/19/2023).

— (2012b). *Banks: At the Heart of the Matter*. ENG. URL: https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Banks (visited on 09/04/2023).

Goldstein, Alex et al. (Mar. 2014). *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. arXiv:1309.6392 [stat]. URL: http://arxiv.org/abs/1309.6392 (visited on 01/11/2024).

Géron, Aurélien (Jan. 2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition [Book]*. en. ISBN: 9781492032649. URL: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ (visited on 01/18/2024).

Hadji Misheva, Branka (June 2023). *Explainable artificial intelligence*.

Hottenhuis, Wouter (Dec. 2022). "Inherently interpretable Machine Learning for Probability of Default Estimation in IRB Models". en. In.

Hotz, Nick (Sept. 2018). *What is CRISP DM?* en-US. URL: https://www.datascience-pm.com/crisp-dm-2/ (visited on 09/28/2023).

IMF (2018). *A Decade after the Global Financial Crisis: Are We Safer?* ENG. URL: https://www.imf.org/en/Publications/GFSR/Issues/2018/09/25/Global-Financial-Stability-Report-October-2018 (visited on 09/04/2023).

Jeppesen, Jacob Høxbroe et al. (Aug. 2019). "A cloud detection algorithm for satellite imagery based on deep learning". In: *Remote Sensing of Environment* 229, pp. 247–259. ISSN: 0034-4257. DOI: 10.1016/j.rse.2019.03.039. URL: https://www.sciencedirect.com/science/article/pii/S0034425719301294 (visited on 01/11/2024).

Langer, Markus et al. (July 2021). "What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research". In: *Artificial Intelligence* 296. arXiv:2102.07817 [cs], p. 103473. ISSN: 00043702. DOI: 10.1016/j.artint.2021.103473. URL: http://arxiv.org/abs/2102.07817 (visited on 08/23/2023).

Mazumder, Saikat (June 2021). *5 Techniques to Handle Imbalanced Data For a Classification Problem*. en. URL: https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/ (visited on 02/12/2024).

Meske, Christian et al. (Dec. 2020). "Explainable Artificial Intelligence: Objectives, Stakeholders and Future Research Opportunities". In: *Information Systems Management*. DOI: 10.1080/10580530.2020.1849465.

Meyer-Baese, Anke and Volker Schmid (Jan. 2014). "Chapter 6 - Statistical and Syntactic Pattern Recognition". In: *Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition)*. Ed. by Anke Meyer-Baese and Volker Schmid. Oxford:

Academic Press, pp. 151–196. ISBN: 978-0-12-409545-8. DOI: 10.1016/B978-0-12-409545-8.00006-6. URL: https://www.sciencedirect.com/science/article/pii/B9780124095458000066 (visited on 01/11/2024).

Microsoft (n.d.). *Explainable Boosting Machine — InterpretML documentation*. URL: https://interpret.ml/docs/ebm.html (visited on 01/09/2024).

Molnar, Christoph (Aug. 2023). *Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/ (visited on 08/23/2023).

Mor, Surender et al. (2022). "Artificial intelligence and loan default: The case of commercial banks in India". en. In: *Strategic Change* 31.6. DOI: 10.1002/jsc.2529. URL: https://doi.org/10.1002/jsc.2529 (visited on 03/08/2024).

Nori, Harsha et al. (Sept. 2019). *InterpretML: A Unified Framework for Machine Learning Interpretability*. arXiv:1909.09223 [cs, stat]. DOI: 10.48550/arXiv.1909.09223. URL: http://arxiv.org/abs/1909.09223 (visited on 01/18/2024).

Obare, Dominic M., Gladys G. Njoroge, and Moses M. Muraya (2019). "Analysis of Individual Loan Defaults Using Logit under Supervised Machine Learning Approach". en. In: *Asian Journal of Probability and Statistics*. DOI: 10.9734/ajpas/2019/v3i430100. URL: https://doi.org/10.9734/ajpas/2019/v3i430100 (visited on 03/08/2024).

Ortaköy, Selman and Zehra Özsürünç (Jan. 2019). "The Effect of Digital Channel Migration, Automation and Centralization on the Efficiency of Operational Staff of Bank Branches". In: *Procedia Computer Science*. 3rd WORLD CONFERENCE ON TECHNOLOGY, INNOVATION AND ENTREPRENEURSHIP"INDUSTRY 4.0 FOCUSED INNOVATION, TECHNOLOGY, ENTREPRENEURSHIP AND MANUFACTURE" June 21-23, 2019 158, pp. 938–946. ISSN: 1877-0509. DOI: 10.1016/j.procs.2019.09.134. URL: https://www.sciencedirect.com/science/article/pii/S1877050919313055 (visited on 09/04/2023).

Peng, Junjie et al. (Sept. 2021). "Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges". English. In: *Frontiers in Pharmacology* 12. Publisher: Frontiers. ISSN: 1663-9812. DOI: 10.3389/fphar.2021.720694. URL: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.720694/full (visited on 03/08/2024).

Petersson, David (Feb. 2023). *AI transparency: What is it and why do we need it? | TechTarget*. en. URL: https://www.techtarget.com/searchcio/tip/AI-transparency-What-is-it-and-why-do-we-need-it (visited on 09/28/2023).

Pugliese, Raffaele, Stefano Regondi, and Riccardo Marini (Dec. 2021). "Machine learning-based approach: global trends, research directions, and regulatory standpoints". In: *Data Science and Management* 4, pp. 19–29. ISSN: 2666-7649. DOI: 10.1016/j.dsm.2021.12.002. URL: https://www.sciencedirect.com/science/article/pii/S2666764921000485 (visited on 09/04/2023).

Schlicht, Peter (2023). "AI in the Automotive Industry". en. In: *Work and AI 2030: Challenges and Strategies for Tomorrow's Work*. Ed. by Inka Knappertsbusch and Kai Gondlach. Wiesbaden: Springer Fachmedien, pp. 257–265. ISBN: 978-3-658-40232-7. DOI: 10.1007/978-3-658-40232-7_29. URL: https://doi.org/10.1007/978-3-658-40232-7_29 (visited on 09/21/2023).

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez (Jan. 2021). "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science*. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020 181, pp. 526–

534. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.01.199. URL: https://www.sciencedirect.com/science/article/pii/S1877050921002416 (visited on 09/11/2023).

Sevinç, Ender (Jan. 2022). "An empowered AdaBoost algorithm implementation: A COVID-19 dataset study". In: *Computers & Industrial Engineering* 165, p. 107912. DOI: 10.1016/j.cie.2021.107912.

Sifrain, Rocheny (2023). "Predictive Analysis of Default Risk in Peer-to-Peer Lending Platforms: Empirical Evidence from LendingClub". en. In: *Journal of Financial Risk Management* 12.01. DOI: 10.4236/jfrm.2023.121003. URL: https://doi.org/10.4236/jfrm.2023.121003 (visited on 03/08/2024).

Vermesan, Ovidiu (Sept. 2022). *Artificial Intelligence for Digitising Industry – Applications*. en. Google-Books-ID: g_yGEAAAQBAJ. CRC Press. ISBN: 978-1-00-079431-1.

Yang, Wenli et al. (Aug. 2023). "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects". en. In: *Human-Centric Intelligent Systems*. ISSN: 2667-1336. DOI: 10.1007/s44230-023-00038-y. URL: https://doi.org/10.1007/s44230-023-00038-y (visited on 08/23/2023).

Yang, Zebin, Aijun Zhang, and Agus Sudjianto (Dec. 2021). "GAMI-Net: An explainable neural network based on generalized additive models with structured interactions". In: *Pattern Recognition* 120, p. 108192. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2021.108192. URL: https://www.sciencedirect.com/science/article/pii/S0031320321003484 (visited on 01/08/2024).

A Appendix: Stakeholder Interview

1. Objective Definition:

The primary objective of the stakeholder interviews is to delve into the intricacies of developing an explainable PD (Probability of Default) model from a practitioner's perspective. This endeavor is not merely a theoretical exploration but a practical necessity, given the rising complexity of financial systems and the increasing demand for transparency. Through these interviews, the research seeks:

- **Identification of Challenges:** Understand the real-world challenges, both technical and operational, that stakeholders encounter when integrating explainability into PD models. This could range from data limitations to regulatory hurdles.
- **Formulation of Guidelines:** Drawing from the collective expertise and experience of the stakeholders, the aim is to derive actionable guidelines for the development of an explainable PD model. This would serve as a foundational roadmap for financial institutions embarking on similar ventures.

2. Participant Selection:

Criteria:

- **Roles:** Participants should be directly involved in the design, development, implementation, or oversight of PD models in their respective financial institutions. This includes but is not limited to roles like Risk Analysts, Model Developers, Compliance Officers, and Senior Decision-Makers.
- **Experience:** A diverse range of experience levels is sought, from novices who bring a fresh perspective to veterans who provide depth and historical context. However, a minimum threshold of having at least two years of hands-on experience with PD models is set to ensure the relevance of insights.
- **Prior Involvement:** Preference is given to stakeholders who have either spearheaded or been an integral part of initiatives aimed at enhancing the explainability of financial models within their institutions.

Sampling Method:

- **Purposive Sampling:** Given the specialized nature of the topic, a purposive sampling method is employed. This means specifically seeking out individuals who have the expertise and experience relevant to the research objectives.
- **Snowball Sampling:** Recognizing the close-knit nature of the financial modeling community, initial participants are requested to refer other potential interviewees who fit the criteria. This method allows for the discovery of stakeholders who might be less visible but equally insightful.

3. Interview Design:

Questionnaire Development:

- **Literature Review:** A review of existing literature, previous researches, and expert discussions on PD models and their explainability was conducted. This helped in identifying gaps in knowledge, potential areas of exploration, and ensuring our questions tapped into crucial aspects of the topic.
- **Relevance to Objective:** Each question was meticulously crafted to echo the primary research objectives: identifying challenges and formulating guidelines. The questions were open-ended, designed to encourage participants to share their experiences, insights, and recommendations in depth.
- **Stakeholder Feedback:** Preliminary feedback was sought from a small group of stakeholders (not part of the main interviewee pool) to ensure that the questions were clear, relevant, and didn't unintentionally bias the responses.

Pilot Testing:

- A pilot test was indeed conducted with a subset of three stakeholders to evaluate the effectiveness of the interview design. This not only tested the clarity and relevance of the questions but also provided an opportunity to gauge the average duration of the interviews.
- **Feedback Incorporation:** Post the pilot, feedback was analyzed. Questions that were deemed too leading or not eliciting the desired depth of information were refined.

4. Data Collection:

Interview Format:

- Given the global spread of stakeholders and the ongoing trend towards remote collaboration, a hybrid approach was adopted. While face-to-face interviews were conducted where feasible, many interviews were carried out via secure online platforms and over the phone to accommodate participants' preferences and logistical constraints.

Duration:

- On average, each interview lasted between 45 minutes to an hour. The pilot testing played a pivotal role in determining this timeframe, ensuring participants had ample opportunity to share their insights without feeling rushed.

Recording:

- Before the commencement of each interview, participants were informed about the intention to record the session. This was not just for transcription purposes but to ensure accuracy in capturing their insights. All participants provided explicit consent for recording. They were assured that the recordings would be used strictly for research purposes, with all identifying information being anonymized in subsequent analyses and publications.

5. Data Protection and Ethics:

Anonymization:

- To ensure the utmost confidentiality, all personal identifiers, such as names, titles, and specific institutional affiliations, were redacted from the interview transcripts. Each participant was assigned a unique code, which was used throughout the analysis phase. Additionally, all

recordings and transcriptions were stored on secure, encrypted platforms, accessible only by the primary researchers.

Consent: • Prior to the interviews, each participant was presented with a privacy statement and a consent form. This document outlined the purpose of the research, the recording process, and the measures in place to ensure data protection. Participants were given ample time to review the statement and ask any clarifying questions before providing their written consent, signifying their voluntary participation and understanding of the data usage.

6. Data Processing:

Transcription: • Soon after each interview, the audio recordings were transcribed verbatim by a member of the research team. This timely transcription ensured nuances and subtleties of the conversation were captured accurately.

Segmentation: • Each transcription was then broken down into distinct segments, corresponding to different questions and themes that emerged during the interviews. This segmentation facilitated a systematic and thematic analysis of the data.

Validation: • To validate the accuracy of the transcriptions, a two-pronged approach was adopted. Firstly, another member of the research team reviewed a random selection of transcripts against their corresponding recordings. Secondly, a subset of participants was given the opportunity to review and confirm the accuracy of their transcribed statements, ensuring they felt their views were accurately represented.

7. Data Analysis:

Coding Strategy: • **Grounded Theory Approach:** The research commenced with open coding, where segments of data were initially coded based on their core content. This was succeeded by axial coding, which established connections between codes, forming broader categories. The final phase was thematic coding, used to extract overarching themes that encapsulated the essence of the interviews.

Themes and Patterns: • After the coding process, a comprehensive analysis was carried out to discern recurrent themes and patterns. The findings highlighted not only the common challenges and recommendations but also diverse perspectives, nuances, and innovative strategies expressed by participants.

Software Utilization: • For an efficient and enhanced analysis process, specialized software, NVivo, was utilized. This enabled organized coding, facilitated easy theme retrieval, and allowed for visualization of patterns, ensuring the analysis was rigorous and replicable.

8. Validation and Reliability:

Inter-coder Reliability: • To enhance the trustworthiness and consistency of the coding process, inter-coder reliability was introduced. After the primary

researcher conducted the initial coding, a secondary coder reviewed a randomized 20

Iterative Approach:

- Consistent with the grounded theory approach, the data analysis was executed iteratively. As codes and themes surfaced, they were perpetually cross-referenced with the original data to confirm their accurate representation. This iterative process ensured the analysis remained rooted in the stakeholders' perspectives and didn't venture into speculative interpretations.

9. Integration with Larger Research:

Role of Stakeholder Interviews:

- While the stakeholder interviews are a distinct methodology, they are integral to the broader research fabric. The insights from these discussions serve multiple purposes:
  - **Guideline Framework:** The information on challenges, best practices, and stakeholder recommendations are amalgamated to create a comprehensive guideline framework for the development of explainable PD models.
  - **Comparative Analysis:** Stakeholder insights are compared with academic theories and models related to explainability. This comparison illuminates gaps between theoretical and practical aspects, suggesting potential bridges.
  - **Future Research Recommendations:** The variety of perspectives and pioneering strategies shared offer potential avenues for upcoming research in the realm of explainable AI within the financial sector.
  - **Informing Model Development:** The broader research, which delves into the intricacies of explainable PD models, will heavily draw upon insights from these interviews. Feedback on technical challenges, data intricacies, and regulatory concerns will significantly shape the direction and depth of future research endeavors.

A.1 Interview Questions

1. Introduction and General Questions

1. Can you briefly describe your role and your experience with credit risk models, especially PD models?
2. What is your current understanding and opinion of machine learning as applied to finance?

2. Objectives and Business Use

3. What are the primary objectives you aim to achieve with this PD model?
4. How do you envision the application of this PD model in your organization's operations?
5. Who are the intended users of this PD model?

3. Explainability and Transparency

6. On a scale of 1-10, how important is the explainability of the PD model for you? (Where 1 is not important at all and 10 is critically important)
7. In your opinion, why is explainability significant for this model?
8. Are there specific regulations or standards that the PD model needs to adhere to in terms of explainability?
9. How do you perceive the trade-off between model accuracy and model explainability?

4. Technical Aspects and Features

10. Are there specific variables or features that you believe are essential for the model?
11. How do you envision handling missing data or outliers in the PD model?
12. Are there any specific machine learning techniques or algorithms you are particularly interested in?

5. Model Validation and Testing

13. How do you envision the validation and testing process for the PD model?
14. What criteria would you use to determine the model's success or failure?
15. Are there any industry benchmarks or standards you believe the model should surpass?

6. Stakeholder Communication

16. How do you prefer to receive updates or explanations about the model? (E.g., visualizations, reports, meetings)
17. What level of technical detail are you comfortable with in communications about the model?

7. Ethical Considerations

18. How do you view the ethical implications of using machine learning for credit risk modeling?
19. Are there any specific ethical guidelines you believe the PD model should adhere to?

8. Future Outlook and Implementation

20. How do you see the role of machine learning evolving in credit risk modeling over the next 5-10 years?

9. Concluding Questions

21. Are there any other concerns or considerations you believe should be addressed in the development of the PD model?

22. Do you have any references, resources, or individuals you recommend consulting for further insights on this topic?

B Appendix: Data preprocessing

B.1 Feature histograms

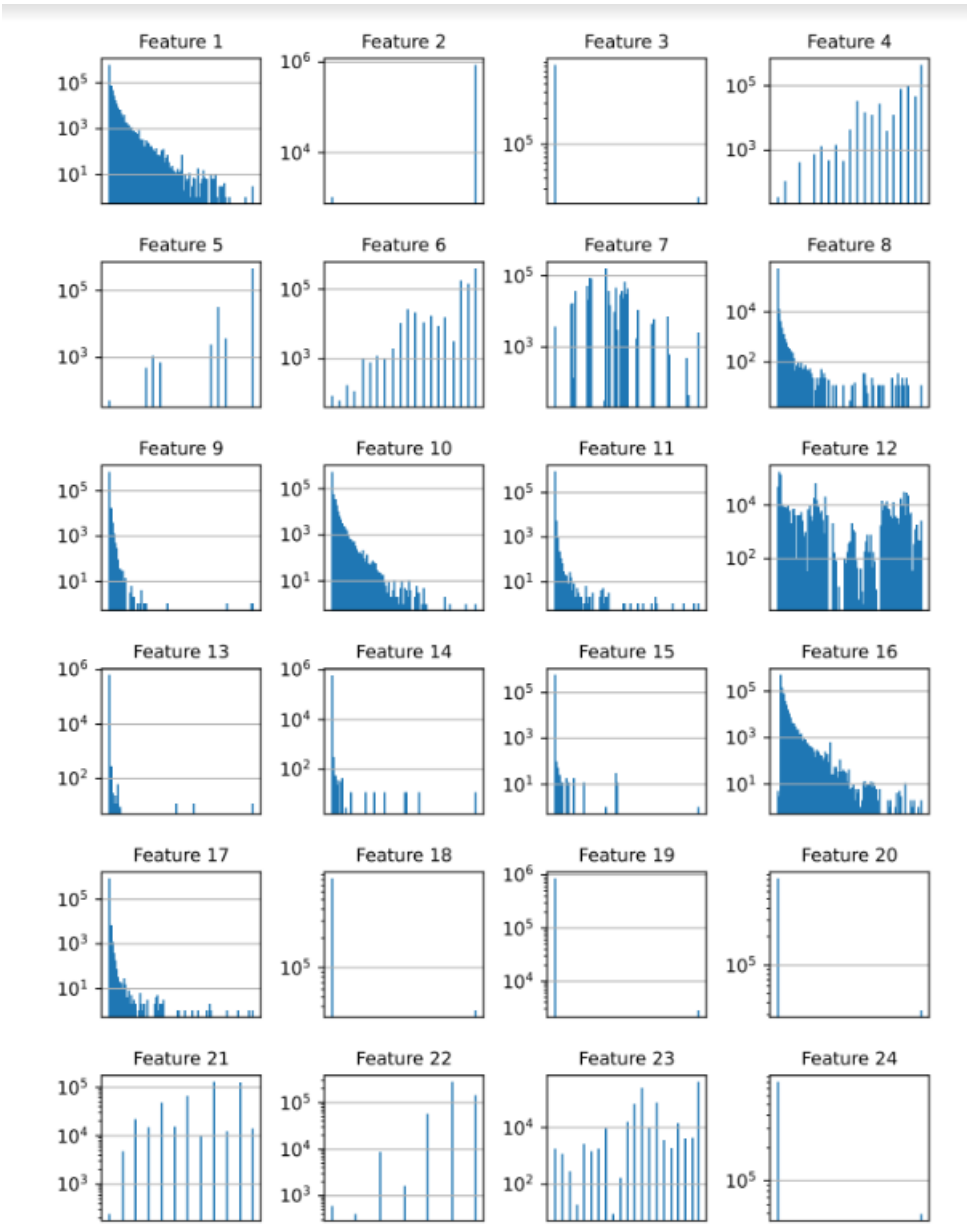


FIGURE B.1: Feature distribution 1

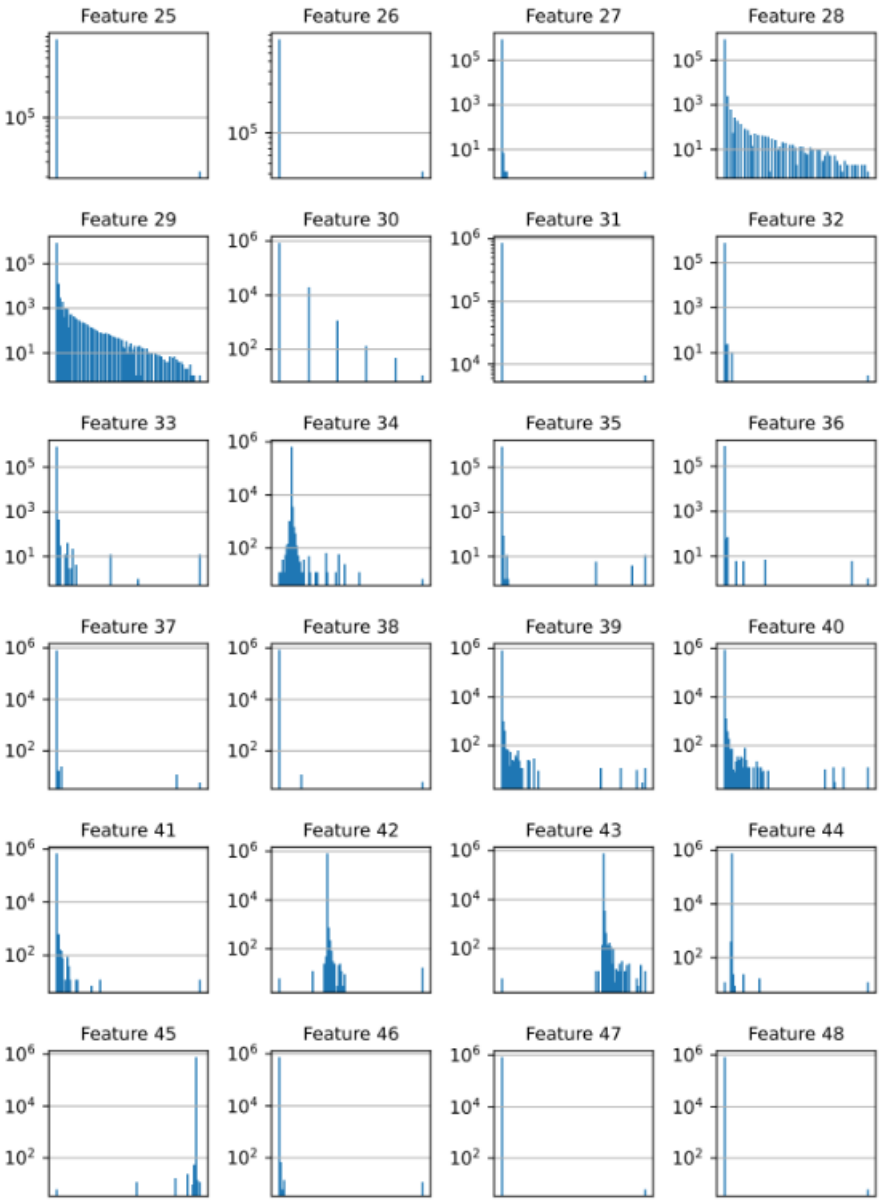


FIGURE B.2: Feature distribution 2

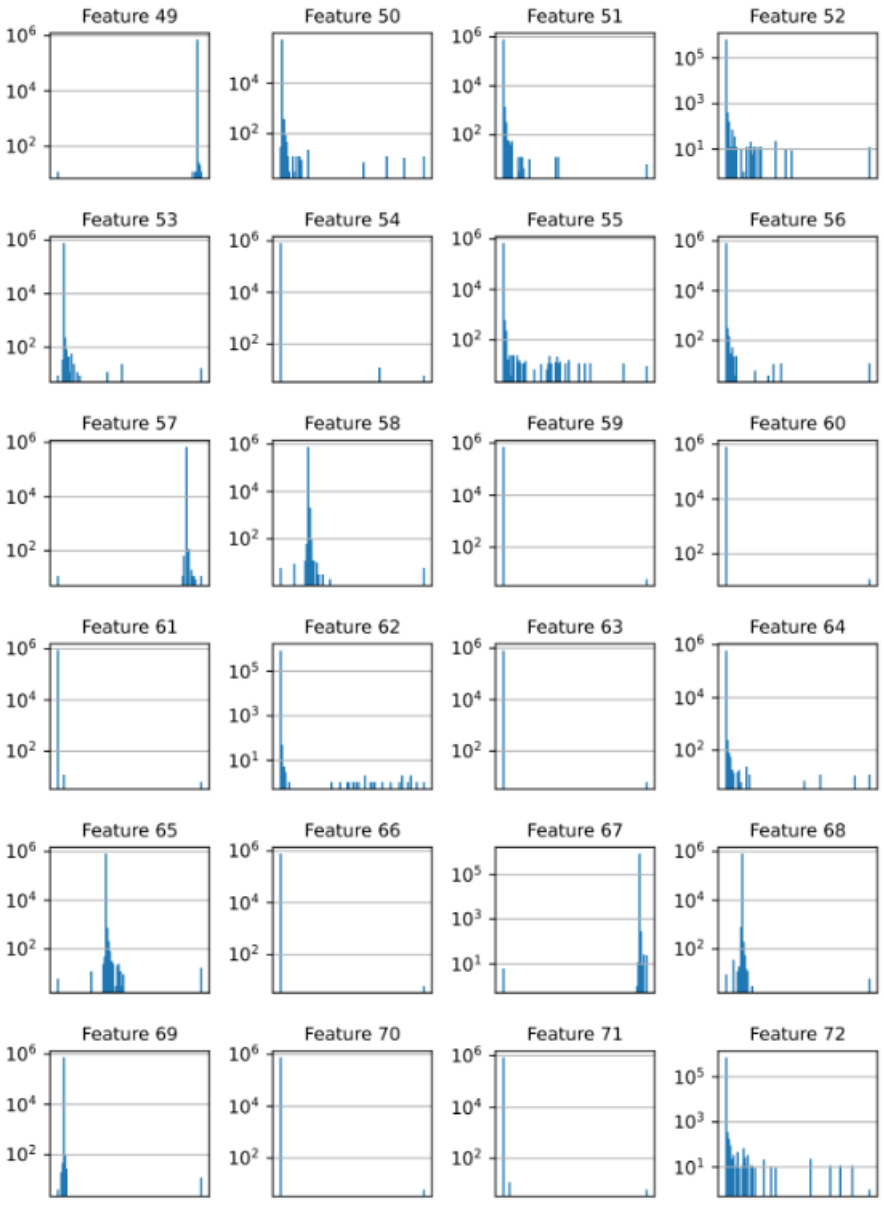


FIGURE B.3: Feature distribution 3

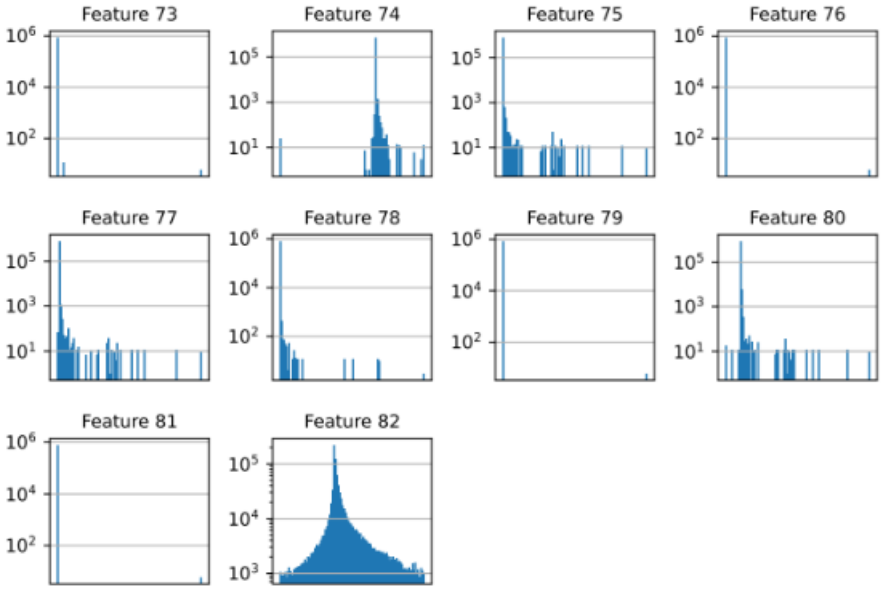


FIGURE B.4: Feature distribution 4

B.2 Correlation matrix

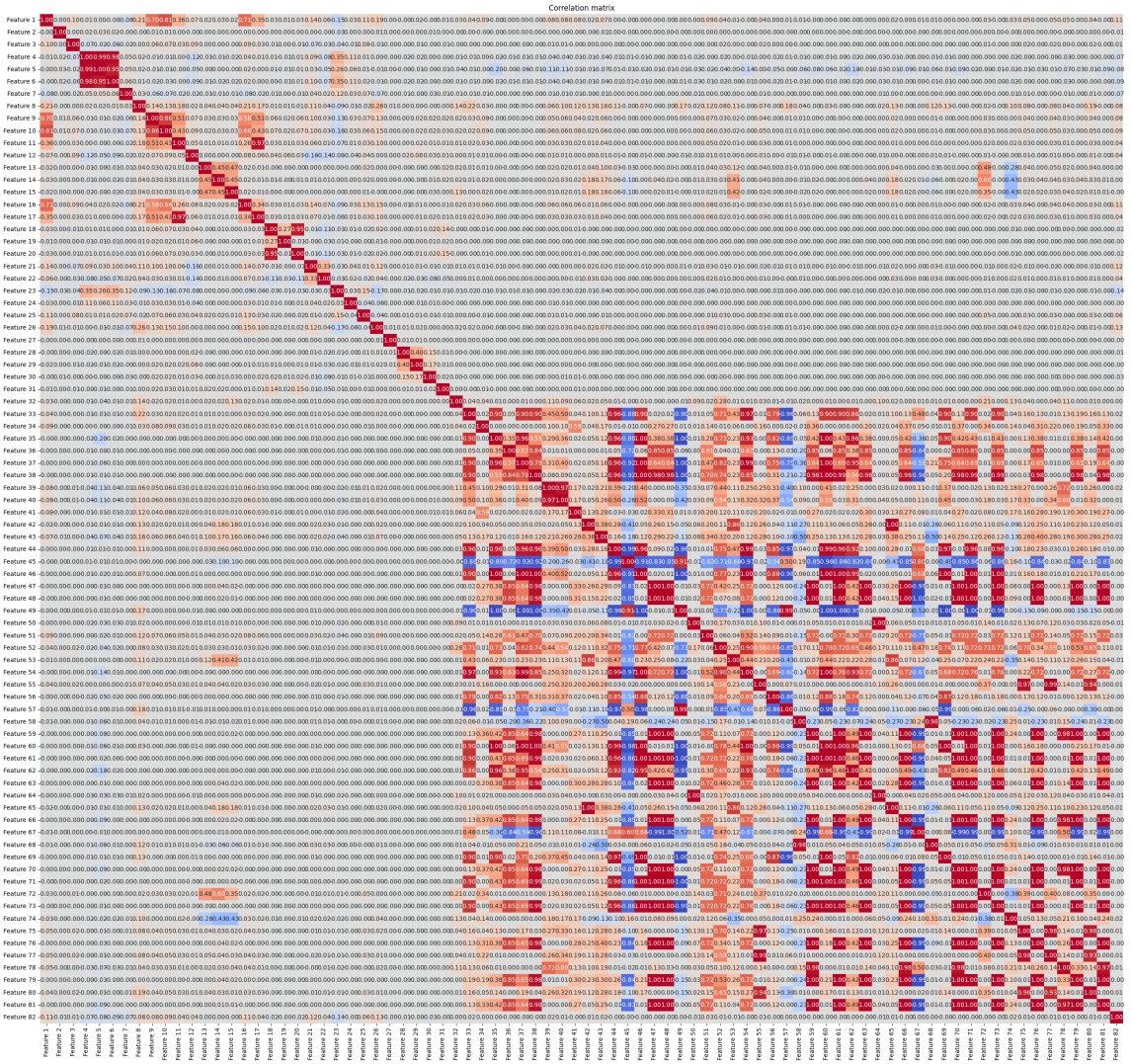


FIGURE B.5: Correlation matrix

TABLE B.1: Feature Selection Overview

Feature	**Status**	**Feature**	**Status**
feature 1	Included	feature 42	Dropped
feature 2	Included	feature 43	Included
feature 3	Included	feature 44	Dropped
feature 4	Included	feature 45	Dropped
feature 5	Dropped	feature 46	Dropped
feature 6	Dropped	feature 47	Dropped
feature 7	Included	feature 48	Dropped
feature 8	Included	feature 49	Dropped
feature 9	Included	feature 50	Dropped
feature 10	Dropped	feature 51	Dropped
feature 11	Included	feature 52	Dropped
feature 12	Included	feature 53	Dropped
feature 13	Included	feature 54	Dropped
feature 14	Dropped	feature 55	Dropped
feature 15	Dropped	feature 56	Included
feature 16	Included	feature 57	Included
feature 17	Dropped	feature 58	Included
feature 18	Included	feature 59	Dropped
feature 19	Included	feature 60	Included
feature 20	Dropped	feature 61	Dropped
feature 21	Included	feature 62	Dropped
feature 22	Included	feature 63	Dropped
feature 23	Included	feature 64	Included
feature 24	Included	feature 65	Dropped
feature 25	Included	feature 66	Included
feature 26	Included	feature 67	Dropped
feature 27	Included	feature 68	Included
feature 28	Included	feature 69	Dropped
feature 29	Included	feature 70	Dropped
feature 30	Included	feature 71	Dropped
feature 31	Included	feature 72	Dropped
feature 32	Included	feature 73	Included
feature 33	Dropped	feature 74	Dropped
feature 34	Dropped	feature 75	Dropped
feature 35	Dropped	feature 76	Dropped
feature 36	Dropped	feature 77	Included
feature 37	Dropped	feature 78	Dropped
feature 38	Dropped	feature 79	Dropped
feature 39	Dropped	feature 80	Dropped
feature 40	Included	feature 81	Included
feature 41	Dropped	feature 82	Included

B.3 Feature exclusion

B.4 Included features

TABLE B.2: Classification of Features by Type

Feature	**Type**
feature 1	Financial
feature 11	Behavioural
feature 17	Behavioural
feature 18	Behavioural
feature 21	Qualitative
feature 22	Qualitative
feature 31	Behavioural
feature 4	Static
feature 40	Financial
feature 43	Financial
feature 56	Financial
feature 57	Financial
feature 58	Financial
feature 60	Financial
feature 64	Financial
feature 66	Financial
feature 68	Financial
feature 7	Behavioural
feature 73	Financial
feature 77	Financial
feature 81	Financial
feature 82	Financial
feature 9	Behavioural

C Appendix: Hyperparameters

C.1 EBM

TABLE C.1: Hyperparameter Testing for `min_samples_leaf`

Min_samples_leaf	Accuracy	Precision	Recall	F1 Score	AUC-ROC
2	0.9785	0.0356	0.0657	0.0462	0.6378
3	0.9785	0.0355	0.0657	0.0461	0.6377
10	0.9785	0.0357	0.0657	0.0462	0.6389
1000	0.9386	0.0413	0.3041	0.0727	0.7122

TABLE C.2: Hyperparameter Testing for `max_leaves`

Max_leaves	Accuracy	Precision	Recall	F1 Score	AUC-ROC
2	0.9718	0.02596	0.07011	0.03789	0.6146
3	0.9785	0.0356	0.06568	0.04617	0.6378
4	0.9759	0.02993	0.06494	0.04098	0.6559
5	0.9789	0.03091	0.05461	0.03948	0.6485

TABLE C.3: Hyperparameter Testing for `max_bins`

Max_bins	Accuracy	Precision	Recall	F1 Score	AUC-ROC
10	0.9627	0.04793	0.1970	0.07710	0.6863
20	0.9737	0.04410	0.1122	0.06331	0.6287
100	0.9632	0.02410	0.0923	0.03822	0.6506
150	0.9738	0.03059	0.0753	0.04351	0.6381
256	0.9785	0.0356	0.0657	0.04617	0.6378

TABLE C.4: Hyperparameter Testing for Learning Rates

Learning Rate	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0.00001	0.9638	0.0745	0.3129	0.1203	0.7421
0.0001	0.9587	0.0733	0.3624	0.1220	0.7566
0.001	0.9789	0.0684	0.1321	0.0902	0.6977

C.2 GamiNet

TABLE C.5: Combined Hyperparameter Testing for Heredity and Learning Rates

Heredity	Learning Rate	Accuracy	Precision	Recall	F1 Score	AUC-ROC
False	0.001	0.6669	0.7856	0.4802	0.5961	0.7346
False	0.01	0.8647	0.0266	0.4797	0.0503	0.7115
True	0.01	0.9557	0.0649	0.3672	0.1103	0.6944
True	0.001	0.9206	0.0461	0.4883	0.0842	0.7903