UNIVERSITY OF TWENTE

MASTER THESIS

Inherently interpretable Machine Learning for Probability of Default Estimation in IRB Models

Author:
Wouter Hottenhuis

University Supervisors:
dr. B. Roorda
dr.ir. W. van Heeswijk

Company Supervisors:
M. Mackay
S. Haro Alfaro

UNIVERSITY OF TWENTE. Deloitte.

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

in the field of

Financial Engineering and Management

July 12, 2022

UNIVERSITY OF TWENTE

Abstract

Master of Science

Inherently interpretable Machine Learning for Probability of Default Estimation in IRB Models

by Wouter HOTTENHUIS

In this thesis, we investigate the topic of inherently interpretable machine learning algorithms for the use in internal ratings-based models. Three different high potential models will be assessed and compared on their applicability for the use in internal ratings-based models, specifically on the probability of default component.

To effectively assess and compare potential machine learning algorithms, a framework is constructed to score the different models. Research on the industry's perspective and the regulatory context showed that there are three main categories on which the models should be evaluated: interpretability, performance, and implementation. These categories are split up into criteria, which are used to score the different models. The status quo in probability of default modelling, a logistic regression model, is also included in the comparison as a baseline. The investigated models are i) Logistic Model Tree (LMT), ii) Generalized Additive Models with Structured Interactions constructed with disentangled feed forward neural networks (GAMI-Net), and iii) Genetic Programming based Symbolic Regression (GPSR). Credit data of the lending platform Lendingclub is used to construct those models and evaluate their performance.

In terms of performance, the LMT and GAMI-Net showed to outperform the logistic regression with respectively an increase of 1.29% and 2.03% in terms of area under the ROC curve. Although the GPSR did not outperform the logistic regression in terms of performance (-1.33%), it has some other interesting qualities that can be proven to be of use in future research. The GAMI-Net sacrificed less in terms of interpretability than the LMT did to get to a better performance. On average, the GAMI-Net scored a 7.4 and the LMT a 6.2 in terms of interpretability, whereas the benchmark logit model scored an 8. The LMT has more disadvantages, which makes it less suitable to be adopted in the IRB model landscape when compared to the GAMI-Net. Additionally, the GAMI-Net shows to have several advantages over the logistic regression. However, the algorithm does make use of neural networks in the construction of the final model, which is the main disadvantage of the GAMI-Net. To conclude, after the LMT as runner-up, the main competitor of the logistic regression model is the GAMI-Net, which seems to have the right balance on the interpretability-performance trade-off.

Acknowledgements

With the completion of this thesis, I fulfill the requirements for the Master of Science degree in Industrial Engineering and Management, with a specialization in Financial Engineering and Management. Therewith, an end has come to my life as a student in Enschede. Looking back at six great years, I am extremely grateful for the time that has passed.

The last six months were dedicated to writing my thesis at the Financial Risk Management team at Deloitte Risk Advisory. I would like to thank the team for offering me the opportunity to write my thesis while being part of their team. I enjoyed many moments together, especially also with my fellow interns in the team. A big thanks to Mats Mackay and Stef Haro Alfaro for the continuous guidance throughout the process. Specifically, Mats helped me with reviewing several drafts and discussing a variety of matters, for which I am very thankful.

From the university, I like to thank my supervisors Berend Roorda and Wouter van Heeswijk. My first supervisor, Berend, showed to be interested in the topic, and was helpful in all discussions we had. My second supervisor, Wouter, was a useful addition in the supervision process as he had valuable feedback in the last stages of the thesis process.

Furthermore, I would like to express my gratitude to my friends from my study and student association, with whom I experienced tremendous valuable moments. Lastly, I am extremely grateful for the continuous support of my parents, my brothers, and Diede during the writing of my thesis.

Wouter Hottenhuis
Amsterdam, 12th of July 2022.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ALTAI	**A**ssessment **L**ist for **T**rustworthy **A**rtificial **I**ntelligence
AUC	**A**rea **U**nder the **C**urve
AUPRC	**A**rea **U**nder the **P**recision **R**ecall **C**urve
AUROC	**A**rea **U**nder the **ROC** curve
BCBS	**B**asel **C**ommittee on **B**anking **S**upervision
BIC	**B**ayesian **I**nformation **C**riteria
BIS	**B**ank for **I**nternational **S**ettlements
CART	**C**lassification **a**nd **R**egression **T**rees
CRR	**C**apital **R**equirements **R**egulation
EAD	**E**xposure **A**t **D**efault
EBA	**E**uropean **B**anking **A**uthority
FSI	**F**inancial **S**tability **I**nstitute
GAM	**G**eneralized **A**dditive **M**odels
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
GPSR	**G**enetic **P**rogramming based **S**ymbolic **R**egression
IIF	**I**nstitute of **I**nternational **F**inance
IML	**I**nterpretable **M**achine **L**earning
IQR	**I**nter**Q**uartile **R**ange
IRB	**I**nternal **R**atings-**B**ased
LGD	**L**oss **G**iven **D**efault
LIME	**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
LMT	**L**ogistic **M**odel **T**ree
ML	**M**achine **L**earning
PD	**P**robability of **D**efault
PDP	**P**artial **D**ependence **P**lot
ROC	**R**eceiver **O**perating **C**haracteristic
SA	**S**tandardized **A**pproach
SHAP	**SH**apley **A**dditive ex**P**lenations
VIF	**V**ariance **I**nflation **F**actor
WCDR	**W**orst **C**ase **D**efault **R**ate
XAI	e**X**plainable **A**rtificial **I**ntelligence
XML	e**X**plainable **M**achine **L**earning

Chapter 1

Introduction

1.1 Research context

1.1.1 Background

Globally, data collection increased exponentially over the last two decades. Industries are making more and more use of the potential that data analytics have. Among others, the financial industry benefits from this data overload: improved customer service or experience, more effective fraud detection, and many other examples can be named in risk estimation and management. To translate this into numbers, the expected compound annual growth rate of the big data banking analytics market is nearly 13%, meaning that within five years the total market value will be doubled from nearly $30 billion to over $60 billion (Petrov, 2022).

All this use of data is enabled by technological innovations and an exponential increase in computational power, allowing us to effectively process these enormous flows of data. In particular, the increase in computational power has showed to be a good match with large volumes of data. Machine learning has proven to outperform existing methods in many domains. These often highly complex algorithms used in machine learning can distinguish faces from each other, are keeping us up-to-date to our preferences on social media, and assists us in driving cars more safely by using autonomously emergency braking and lane assist, for example.

However, as always, along with innovation and new trends, there often is a justifiable lack of trust. Even some resistance to fully adopt these novel concepts is often present, as there are many examples in which machine learning does not deliver on promises. Since within the financial industry machine learning has proven to outperform traditional operations, it is especially the lack of trust in black box models that predominates the discussion on fully adopting this new techniques within the company's models.

1.1.2 Recent development

Recently, the European Banking Authority (EBA) published a discussion paper aiming to get a better understanding of the challenges and opportunities coming from machine learning (EBA, 2021). The paper is specifically aimed at machine learning techniques that have the potential to be applied in the context of internal ratings-based (IRB) models to calculate regulatory capital for credit risk. The motivation for the EBA to publish this discussion paper is caused by the discrepancy between the use of ML in different areas of the financial industry. Illustrative, machine learning algorithms are adopted very quickly in so-called FinTech. However, the adoption of machine learning in credit risk with respect to regulatory purposes is lagging behind.

In a survey of the Institute of International Finance (IIF, 2019) credit institutions reasoned that regulations bound the use of machine learning in practice: "*regulatory requirements do not always align with the direct application of ML models, due to the fact that regulatory models have to be simple, while ML models might be more complex (although not impossible) to interpret and explain*". In that same survey, it was found that credit institutions shifted the use of machine learning away from regulatory purposes, such as capital requirements, towards business-related solutions, such as monitoring outstanding loans (Alonso and Carbó, 2020).

The regulatory burden for applying machine learning techniques in regulatory capital calculations has several reasons. In different areas of the financial service industry, the main objective is accuracy, in which machine learning algorithms have proven to outperform traditional models. Although some find evidence that ML models yield at most similar results compared to traditional benchmark models such as logistic regression (Bacham and Zhao, 2017), more often, others find that several ML models outperform traditional models (Albanesi and Vamossy, 2019; Petropoulos et al., 2019), sometimes even with an improvement for default classification of over 20% compared to those traditional models (Alonso and Carbó, 2020). However, within credit risk management, there are more regulations in place. For example, guidelines require that credit providers do not "*discriminate against protected classes and that consumers are offered explanations for denial of credit*" (Breeden, 2020). As mentioned earlier, these highly complex machine learning algorithms are extremely hard to interpret and explain. Therefore, the inner workings of a model are not clear, resulting in a so-called "black box" model. Without a clear link between inputs and outputs, the overview is missing, and one might not be able to identify issues such as discrimination, or non-intuitive relations between inputs and outputs in credit decision processes.

1.1.3 Motivation

Before zooming in on the research question, we will briefly motivate why it is important to have non-"black box" algorithms in the area of credit risk. As with the healthcare industry, the financial industry is also an industry recognized to be of high priority. In many decision-making processes in these industries, output of models need to be explained, especially when (legal) persons are involved, such as in credit provision practices.

To illustrate this, one can think of a person applying for a mortgage loan. First, when a decision is made about the approval of the loan, the customer may want an explanation for it, which is also a legal obligation of a credit provider. Additionally, one needs to determine the pricing, i.e., the interest on the loan, and also justify that. Next to that, the model developer benefits from interpretable models, when trying to tune the model or trying to solve issues. On top of that, all stakeholders involved want to trust the model, which can be assured by transparency. Another, very significant, aspect that comes with an interpretable machine learning model, is the possibility for regulators and auditors to check the model's inner workings. Does the model, for example, satisfy all legal requirements? Especially this last reason is of high interest, since regulators saw credit institutions shifted the use of machine learning away from regulatory purposes.

In short, interpretable algorithms will benefit 1) the customer with enhancing trust in the decision-making process, 2) the financial institution by allowing for more efficient problem diagnosing and solving, and 3) the regulatory agencies with more effective checking on compliance issues.

1.2 Research questions

The main research question of this thesis is:

> **Which interpretable machine learning algorithms are applicable for the use in IRB models and how do they differ from each other?**

The question above is from a credit providing organization's perspective and focuses on IRB models. This focus means that the model must be inherently explainable, as post-hoc explainable artificial intelligence transpire not to satisfy the applicable guidelines and regulations (this will be discussed in more detail in Chapter 2). The perspective from the credit providing organization is also relevant, as the possible algorithms should also yield these companies benefits. This is among others that algorithms should outperform current models. For the probability of default models, the logistic regression model should be outperformed, which is currently the status quo in IRB models (Triple A – Risk Finance, 2022).

In order to answer the main research question, the research is divided into four subquestions. These are listed below, with a short motivation.

A) **What is the current state of machine learning adoption within IRB models in the industry and in terms of regulations and guidelines?**
Answering this question gives us an overview of the problem at hand. This insight enables us to identify the current problems with implementing machine learning in IRB models. The aim is to also get to know the current regulations and guidelines with respect to machine learning and, to a lesser extent, AI.

B) **What is an appropriate way of comparing machine learning algorithms in terms of applicability for the use within IRB models?**
As stated in the main research question, we want to be able to measure 'applicability in IRB models'. This will include aspects of interpretability and also, for example, classification performance. The deliverable of this subquestion is an assessment framework with criteria which can be used to assess and score different machine learning methods on their applicability for the use within IRB models. This framework will be used in the final subquestion.

C) **What are appropriate machine learning algorithms from the literature for application in IRB models?**
From the literature, we will explore the variety of machine learning algorithms available, and more specifically those models that are explainable by nature.

D) **How do the different algorithms score based on the assessment framework, and what are their key distinctions?**
At last, we can actually compare and contrast the chosen models and highlight their strengths and weaknesses. We will do this by performing a quantitative analysis, based on an open-access credit-lending database. Different models will be developed, fitted, and ultimately, compared to each other and a benchmark model.

1.3 Methodology

The methodology used for this research consists of at least the points listed below. When these steps are followed, we have also answered the research questions (denoted by **RQ X**).

- *Qualitative research*
  We start by addressing literature to get to know the current state of machine learning in credit risk. This will serve as the basis of the theoretical framework used in this thesis. The following topics will be addressed:
  - Credit risk
  - Machine learning (algorithms)
  - **RQ A** - Regulations in credit risk
  - **RQ B** - Finding useful assessment criteria
  - **RQ B** - Composing the assessment framework
  - **RQ C** - Possible ML algorithms for in IRB
- *Quantitative research*
  After identifying promising algorithms, we will use a publicly available data set to evaluate the chosen algorithms. The dataset consists of a peer-to-peer loans, entailing characteristics of the loan, and the applicant of the loan. First, we need to construct the chosen algorithms and build the models. This enables us to answer **RQ D**, where we evaluate the performance of the algorithms with the chosen evaluation criteria. This part of the research will encompass:
  - Data cleaning
  - Data preparation
  - Model development
  - Model tuning
- *Evaluation*
  RQ D - The final step is to evaluate the different models on the basis of the chosen assessment criteria from the qualitative research method. To finalize the thesis, we draw conclusions, state the limitations of the research and give some recommendations for further research.

1.4 Outline

To assist the reader in maintaining an overview throughout the reading of this thesis, we will outline the structure of this thesis. In Chapter 2 we will perform a qualitative research. We introduce some main concepts and definitions for this study, and therewith we define the scope of the research. Subsequently, in Chapter 3 we answer research question A by zooming in on the current situation, including the regulatory context of machine learning in IRB models. After that, we zoom in on how we can measure the appropriateness of machine learning algorithms for adoption in IRB models. In that way we constructed an assessment framework, which is the deliverable of research question B. In Chapter 4 we will answer research question C, in which we will choose machine learning models from the literature that we will compare. Afterwards, the data will be prepared which will enable us to answer research question D in Chapter 5, in which we actually use the assessment framework to compare the algorithms.

Chapter 2

Theoretical Context

In this chapter, we will bring up different definitions and concepts necessary to understand the structure and content of this thesis. Getting more familiar with the topics of credit risk, machine learning, and interpretability of machine learning will enable us to articulate the scope of the research better at the end of this Chapter 2.

2.1 Credit risk

In this thesis, we define credit risk the same as the Basel Committee on Banking Supervision does: "*credit risk is the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms*" (BCBS, 2000). Most traditional and large banks' main source of income comes from providing credit to lenders. That makes that most of the financial risk that banks face is credit risk. However, a bank does not lend out capital for free, therefore it requires some interest on the credit from the borrower. By aggregating these premiums, the bank is able to afford an *expected loss*. Therefore, a bank wants to quantify this credit risk as accurately as possible, to define a fair interest on the loan.

Next to the intrinsic incentive of a bank to quantify credit risk, there is also an external body that compels banks to do so. The European Banking Authority (EBA) has the duty to maintain financial stability within the European Union and is mandated to assess risks and vulnerabilities in the EU banking sector (EBA, n.d.). One way of doing this is by enforcing banks to hold a capital buffer for when there is economic downturn or crises. In contrast to the money held for the expected loss, the regulatory capital is generally held for the *unexpected loss*. These two concepts are depicted in Figure 2.1. In the case that multiple clients will fail to meet their obligations, the bank will still have a capital buffer to continue its operations. This is all enforced in the EU Regulation No 575/2013, the Capital Requirements Regulation (CRR). We will specifically address the regulations regarding ML and AI in Section 3.2, where we address the regulatory context. With regard to this thesis, we will focus on how credit risk is quantified. We will address the calculations of the *expected* and *unexpected loss* after we explain the three main components of credit risk modelling below.

2.1.1 Credit risk modelling

In Figure 2.1 the well-known loss curve is depicted, with the *expected* and *unexpected loss* mentioned earlier. As we can see, the first part of the graph corresponds to the expected loss. Therefore, the bank wants to quantify how many losses they can expect and subsequently base their pricing policy on that. The gray second part corresponds to the unexpected loss, and is, from a regulator's perspective, extremely

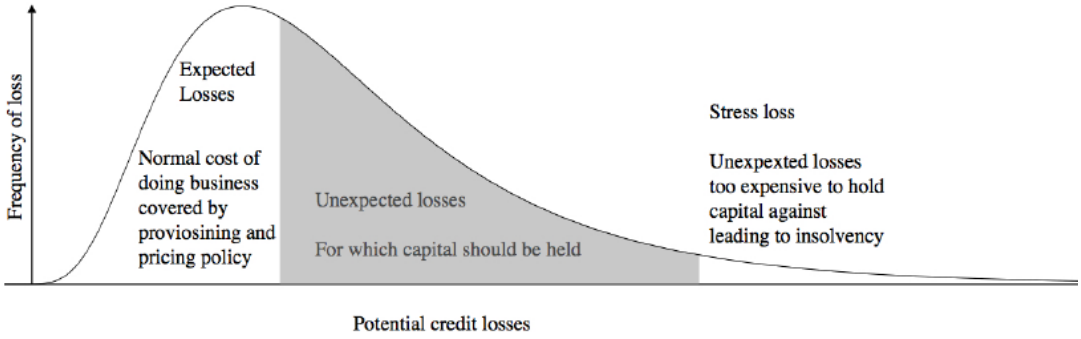

FIGURE 2.1: Total credit loss function with explanations (Gonzalez, Savoia, and Sotelino, 2012).

important to quantify correctly as this is necessary to calculate the capital to be held to ensure stable operations.

The quantification of this risk is done in terms of expected and unexpected losses, which corresponds to the x-axis of Figure 2.1. To quantify and model credit risk adequately, one needs three components to do this accurately.

- **Probability of Default (PD)**
  A default occurs when an outstanding loan is not paid back timely. Generally, a loan is considered to be a default when the obligor is past due more than 90 days. To quantify the risk of not getting paid back the loan and interest as a bank, the bank estimates the probability of default. Given some loan characteristics, macroeconomic variables, and/or characteristics of the obligor, the bank is able to estimate the probability of default. In IRB-models, this is often done with the use of logistic regression. Obviously, the probability of default is a number between 0 and 1. Often, in estimating the PD, a one-year time horizon is taken into account (CRR Articles 160, 163, 179 and 180 (EBA, 2013)).

- **Exposure at Default (EAD)**
  The exposure at default depicts the amount of the loan that is not yet paid back. Often, the closer to maturity of the loan, the lower the exposure at default becomes.

- **Loss Given Default (LGD)**
  The loss given default is defined as a fraction (of the EAD). The loss given default among others depends on the collateral that is used in the lending contract. If, given a default, the bank is able to recover a collateral or a certain amount of the loan, then that will impact the loss given default. A higher recovered amount will correspond to a lower LGD. Ab LGD of 100% means that the full loan is lost when the client defaults (corresponding to no collateral or no recovery of the collateral). Generally, in the European market, the empirical LGD distribution has a U-shape (Vujnovic, Nikolic, and Vujnovic, 2016). That is quite intuitively, namely, most observed LGDs are around low numbers, between 0 and 0.2, or around high numbers, between 0.8 and 1.

Finally, to quantify the credit risk, the three components are multiplied for each asset or portfolio denoted by i, and aggregated to a total amount (Hull, 2007):

$$Expected\ Loss = \sum_i PD_i \cdot LGD_i \cdot EAD_i \tag{2.1}$$

This corresponds to the left white area under the curve in Figure 2.1. Also, the gray area under the curve, the unexpected loss, can be calculated with this formula. For this, we need to quantify one extra concept, namely the Worst-Case Default Rate (WCDR)[1]. This number is calculated with the PD-component and a correlation coefficient between the different i portfolios. It corresponds to the area under the loss curve for which capital should be held. The area is determined with a confidence level, which is often set by the regulator to 99.9%. Only the 0.1% largest losses are not taken into account in the WCDR. Now, the unexpected loss can be calculated by (Hull, 2007):

$$Unexpected\ Loss = \sum_i (WCDR_{99.9\%,\,i} - PD_i) \cdot LGD_i \cdot EAD_i \tag{2.2}$$

As shown above, credit risk consists of multiple components. Each of these components can be modeled by itself, and also their correlation can be modeled and analyzed. To limit the scope of this thesis, we will from this point onward only consider the ***PD-component*** in credit risk modelling. This enables us to focus on a binary classification problem; a default or no default, which is the topic where the application of machine learning arises.

2.2 Machine learning

Another important topic to introduce for this thesis is machine learning. To ensure comparability of this thesis with other research, we will use the definition of the International Organization for Standardization on IT Governance: "*machine learning is a process using algorithms rather than procedural coding that enables learning from existing data in order to predict future outcomes*" (ISO, 2017). Note that machine learning (ML) is a subset of the artificial intelligence (AI) domain. Where AI covers the area of making a computer or machine do the same task as a human (e.g., robotics, or voice-assistants), ML only covers the part where the computers or machines learn from data by the use of a (ML) algorithm.

2.2.1 Classification algorithms

As this thesis will be focusing on the PD-component of IRB models, the scope of algorithms to be considered is reduced to probabilistic classification algorithms (Braak, 2021). That is the case, since we are interested in finding the PD. Differently formulated, we are interested in the probability (probabilistic) that a loan belongs to a certain class (classification), default in this case. To further specify this, we limit ourselves to binary probabilistic classification, since there are only two classes: non-default or default, i.e., binary. Knowing that these are the aspects of ML to take into account when estimating the PD, helps us narrow down the point of focus in the wide ML landscape.

Below, we will introduce some of the most well-known ML algorithms used for binary classification. We want to stress that this is not meant to be a comprehensive list, neither a full-explanation of these algorithms, as that would go beyond the purpose of giving an introduction to these methods. Some ML algorithms are considered to be inherently interpretable, also many are not. There are many alterations to these most commonly used ML algorithms, which need not be treated to introduce the reader to the ML field. We will explicitly treat those highly specific alterations to

[1] This number can be determined with Vasicek's model (see Vašíček, 1987)

those common algorithms after Chapter 3, where we take a deep dive into current regulations, interpretability, and other criteria to evaluate the applicability of ML in IRB models.

Logistic regression

Following the definition of ML stated in the beginning of this section, we should treat logistic regression as an ML algorithm. However, this has been the standard for binary classification for the past decades, and is thus often not treated as 'ML'. Logistic regression makes use of a linear combination of the predictor variables. Mathematically, this linear combination can be written as:

$$f(x_1, x_2, ..., x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.3}$$

Where all β are coefficients that need to be learned by the algorithm, and all x are predictor variables. The β_0 is called the bias term that is included to scale to the outcome. To translate the linear combination of coefficients multiplied by the predictor variables into a binary classification (i.e., choose either class 0 or 1), logistic regression makes use of the logistic function:

$$y = \frac{1}{1 + \exp(-f(x_1, x_2, ..., x_n))} \tag{2.4}$$

This results in an S-shaped curve on the interval $[0, 1]$, see Figure 2.2. In this illustrative example, the predicted variable y, in this case *Probability of Default*, takes only one predictor, namely the *Balance* of a person. In general, a logistic regression learns, based on the available predictors, the best coefficients, β, that maximize the so-called likelihood function of the coefficients. Simply said, the algorithm finds the coefficients that yield the lowest error between the predicted and actual outcome, just like the well-known least squares method. In this thesis, we assume readers are familiar with common topics such as maximum likelihood method, therefore we do not further explain this concept.

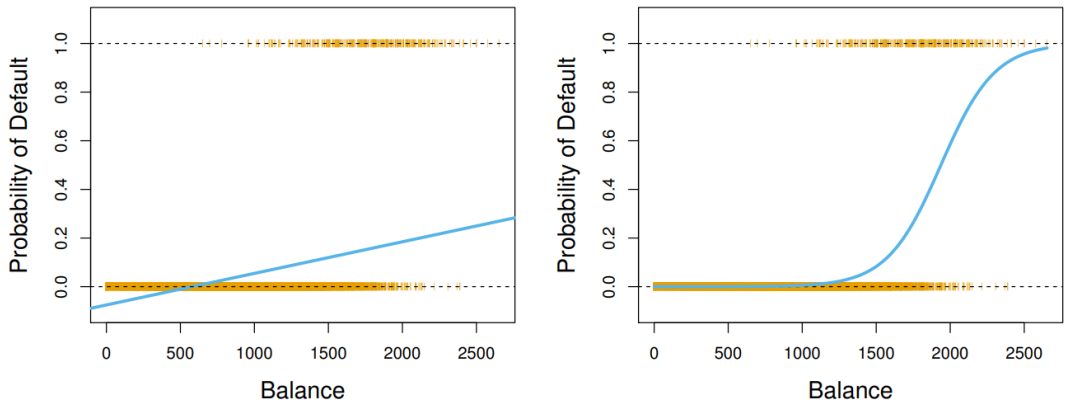

FIGURE 2.2: *Left*: A linear regressor fitted on a binary classification problem, predicting negative probabilities. *Right*: The logistic function mapping the outcomes of a logistic regression to the interval $[0, 1]$ (James et al., 2021).

As we will address at a later stage in this chapter, in Section 2.2.2, the logistic regression is one of the few methods that is perceived as interpretable enough to use

within IRB models. That is, among others, the reason for it being the standard for estimating the PD in the IRB approach.

K-nearest neighbors

The k-nearest neighbors algorithm is a way to classify a data point based on the distance from the features corresponding to that data point and the features of other data points. The K stands for how many neighbors are considered. For example, when considering only 1 neighbor, the data point will be labeled the same as the closest other data point. When $K = 10$, for example, the data point will be labeled the same as the majority of the 10 closest data points, e.g., when 7 of the 10 are labeled 1, the data point will also be classified as 1.

Although this method is often recognized as easy to interpret, it does have two major disadvantages. First, the algorithm's workings are only well-understood and visualizable when there are at most 2 dimensions in the feature space. Second, K-nearest neighbors lacks the ability to produce meaningful probabilities, just as decision trees. It can produce some kind of probability, but that actually is the proportion of neighbors that have that specific class. That is, for the example given in the example above, a proportion of 70%. It is clear, that this cannot be interpreted as a real PD but is more like a height of the certainty that it belongs to that class. With calibration methods, the probabilities can be mapped onto more realistic areas of probabilities to overcome this problem. This is outside the scope of this thesis.

Generalized additive models

Generalized additive models (GAMs) are a bit more sophisticated than logistic regression, but still belong to the more interpretable models. That is because a GAM essentially does the same as a logistic regression. The only difference is that the logit model only allows for linear combinations which are summed (i.e., coefficient β multiplied by x), whereas a GAM sums non-linear combinations, so-called smoothing or shape functions, see Figure 2.3. GAMs are thus additive, just like logit models. Mathematically, it can be written as follows:

$$f(x_1, x_2, ..., x_n) = \beta_0 + g_1(x_1) + g_2(x_2) + ... + g_n(x_n) \tag{2.5}$$

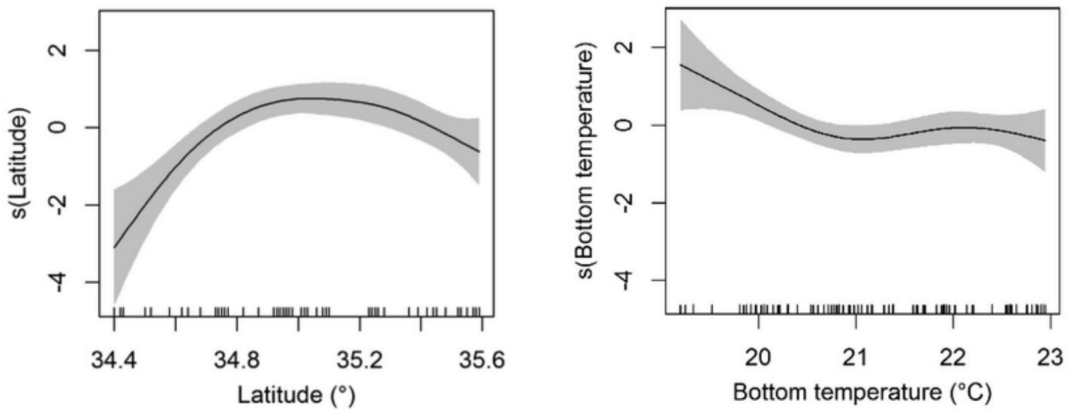

FIGURE 2.3: Examples of shape functions from a GAM (Xue et al., 2018).

In contrast to Equation 2.3, the additive components of Equation 2.5 are now functions, $g_i()$, of the predictor variable x_i. This can either be linear, or non-linear. Again, the final prediction is made, by substituting $f(\mathbf{x})$ in the logistic function (from Equation 2.4).

Support vector machine

Given a training sample of defaults and non-defaults, a support vector machine tries to find a boundary between these two classes in the feature space. It therefore uses a function to maximize the separation between the two classes and the boundary. In a two-dimensional space, the support vector machine is clearly visualizable, but the algorithm also works well for higher dimensional feature spaces. However, it is also a non-probabilistic binary classifier, and therefore not ideal for a direct interpretation of PDs.

Decision trees

Decision trees are tree structured classification algorithms, which are intuitive and easy to understand, even by nonexpert users (Fürnkranz, 2010). It consists of splitting the data at each node. These are binary splits, for example: "$Income > \$5.000$". After each split, another split can be made, and in such a way a tree is grown. See Figure 2.4, on the right side such a small tree is made. On the left side, the so-called decision boundaries are provided. Observe that these boundaries are all vertical or horizontal lines, but the degree of fragmentation increases as the tree size increases. To overcome a tree algorithm to grow a large tree that is sensitive to overfitting, often a tree is pruned. In this way, using a validation data set, one removes leaves nodes, until a certain threshold value of decrease in accuracy is reached, or the desired tree size is established.

A decision tree performs best when, at each leaf node where the final classification is made, either 1 or 0, all the data observations belong to that given class. This is therefore a very transparent ML approach. A disadvantage is that it does suffer from not being able to produce meaningful probabilities that are necessary for a PD model. However, there are different methods and techniques to infer probabilities from the model, but these will not be further explained here.

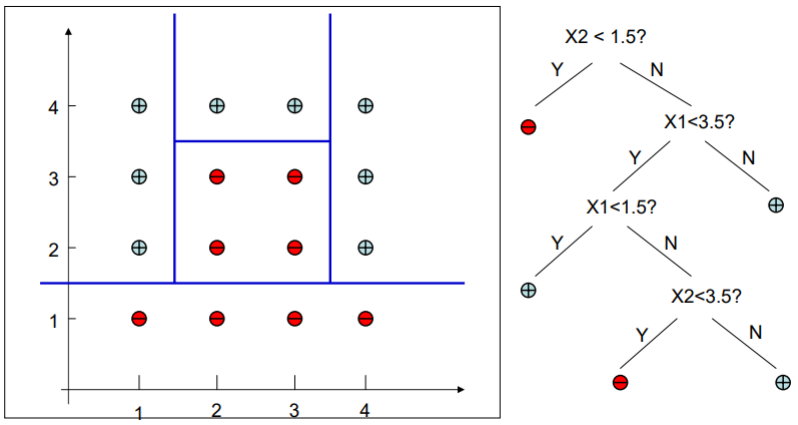

FIGURE 2.4: The working of a classification tree, with the corresponding decision boundaries.

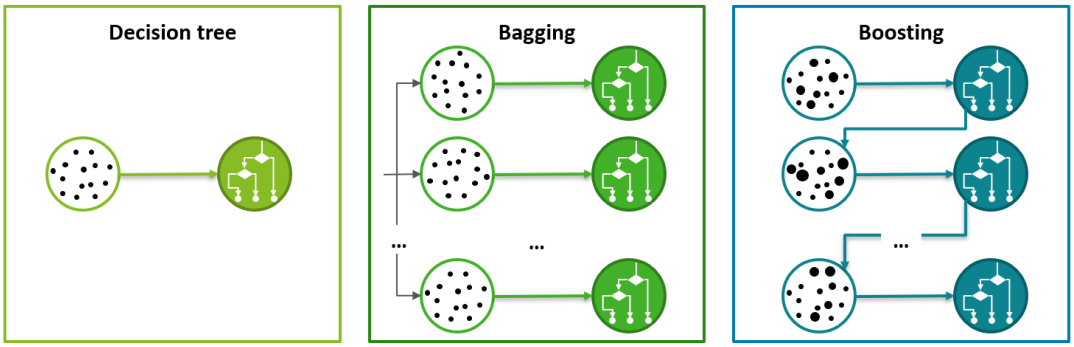

FIGURE 2.5: A comparison of tree-based methods. Left: a regular decision tree that is grown by the dataset. Middle: bagging uses multiple bootstrap samples to grow many trees parallel. Right: boosting makes use of a sequential learning algorithm.

Tree ensemble: bagging

The term bagging is an abbreviation for bootstrap aggregating, and belongs to the ensemble methods within ML. The widely known ML algorithm 'random forest' is an example of bagging. To understand this method, we explain the two terms, bootstrapping and ensemble/aggregating.

Bootstrapping is a sampling method where random samples with replacements are picked. In this manner, it is possible to artificially generate several new datasets from the first dataset. For a random forest, this number is often chosen in the range of 100 to 2000 datasets.

For each bootstrapped dataset, a classification tree is grown. Since each tree is optimized for that specific bootstrap sample, every tree can be different from one-another. If we then want to make a prediction on a new data record, this data record is passed through every individual tree. Following this procedure, each tree has one vote in the final prediction. For example, when 750 of the 1000 trees classify the record as a default (= 1), the final prediction is also a default. It is often shown in literature that the proportion of votes in a random forest actually can be used as an actual PD (Olson and Wyner, 2018). See Figure 2.5 for a conceptual illustration of bagging, and a comparison with a regular decision tree and the boosting algorithm that is treated below.

The disadvantage of random forests is the size of the model. Although built up with the relatively simple classification trees, the size, which is the number of trees, causes the model to be not transparent. On top of that, composing such a large number of trees can be computationally expensive, especially when datasets are large.

Tree ensemble: boosting

Boosting is, just like bagging, an ensemble method, but differs from bagging in the way that it creates the individual trees. The main difference becomes clear from Figure 2.5, where it is shown that there is a sequential flow of information instead of a parallel-wise flow.

The algorithmic procedure is as follows. A decision tree is grown from the initial dataset. From this classifier, the data records that were mislabelled, will be assigned a higher weight for the next iteration. This higher weight that belongs to a data observation has an impact to the learning procedure, specifically when evaluating

a loss function. When the next tree will be grown, also called a weak learner, it incorporates those higher weights from the previous trees, therewith extra focus is put on the data records that are not easy to classify. In that sense, each next decision tree will learn from the mistakes of the previous tree. This continues up to a pre-set number of trees or specific loss criterion, such as no increase in prediction power. The final decision is made with a weighted average of all weak learners, where the weights of the weak learners are based on the error of that weak learner[2] (Zhang, 2019).

Boosting also suffers from being not interpretable. It is a bit more sophisticated than bagging. It is even less traceable, since one should follow hundreds or even thousands of paths from tree to tree to see what happens inside the model. It is also computational more expensive, as it cannot be run in parallel, such as a bagging algorithm.

Deep learning

Deep learning is a method that makes use of a neural network. It consists of several layers of nodes, the first one being the input features. Each node is then connected with every other node of the next layer. Eventually, the final layer consists of one node (in binary classification), fed with values from the previous layer it uses sigmoid, softmax, and/or logistic functions to calculate the PD. A visualization of this design is given in Figure 2.6.

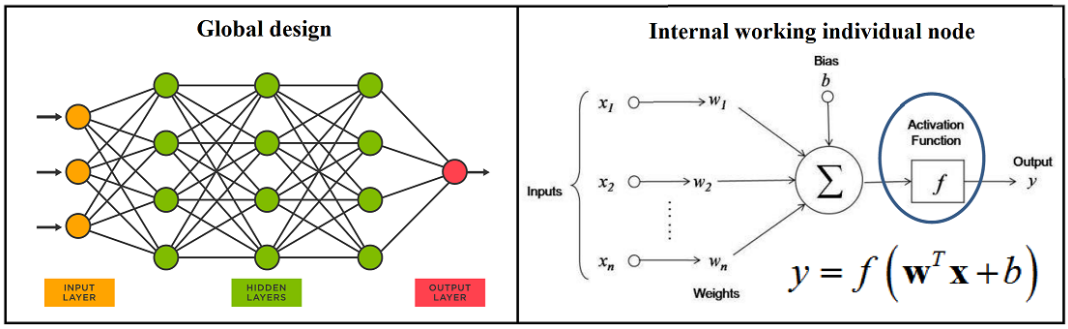

FIGURE 2.6: The structure and inner workings of a deep neural network (based on Kimura et al., 2019; TIBCO, n.d.).

Also visualized in the same figure is the internal working, which is a bit more complicated. Each connection has its own weight. Within each node, the sum of the product of the previous output of a node and its corresponding weight becomes the input for the current node. Within a node, the input plus a bias term is processed in an activation function, which becomes the output of that specific node. The power of a deep neural network is that all the functions formed in the net are differentiable (Sarıgül, Ozyildirim, and Avci, 2019). Given that characteristic, with the use of back-propagation, one is able to adjust each weight in the learning process to minimize a loss function.

Obviously, the inner-workings of a deep learning model consisting of more than three layers is not traceable. Additionally, the possibly high predictive power comes at the cost of transparency. For example, one cannot know in a deep learning model

[2]This is a minimal explanation of the boosting concept. To get a thorough understanding of boosting, we refer the reader to other literature, such as Zhang, 2019.

when the variable *Income* rises, whether the PD goes up or down. In a logistic regression, for example, this is a lot easier, as the corresponding coefficient is the only term having an effect on the outcome. On top of that, a neural network needs to learn many weights in the learning phase, which makes it computationally heavy, and it therefore needs also a lot of data records to be properly trained.

Evolutionary algorithms

Although not an ML algorithm on its own, we will also consider evolutionary algorithms in this section, as they are related to ML and can be used in the learning process. It is, however, a relatively less commonly used concept in ML. Evolutionary algorithms are inspired by Darwin's evolution theory, and can be classified as a heuristic. It makes use of mechanisms like mutation, selection, and recombination to efficiently search the solution space. One specific type of evolutionary algorithm is genetic programming, where the final computed solution is in the form of a computer program, which can also be used in combination with classification problems.

As the name of evolutionary algorithm already implies, they are in some sense different from the 'regular' learning algorithms. The former are used for evolution, for heuristically optimization of solutions in evolution and the latter are to train agents to perform better or to increase their performance (Iqbal, 2015). However, evolutionary algorithms have also been used for learning, also referred to as learning classifier systems (Urbanowicz and Moore, 2009). Given training data, a learning classifier system efficiently searches the solution space, by evolving through each iteration. In that way, it actually acts as a learning algorithm by getting a better fit throughout the learning phase.

We can best illustrate this with an example. Consider the environment, the solution space, where there are living lots of individuals, the possible solution models. The individuals (models) with a higher fitness (better fit with the to-be-predicted outcome) have a higher chance of surviving, just like actual evolution. In each generation, or iteration, the individuals (models) reproduce themselves. In this stage, the concepts of mutations, and crossovers between families find place. This might lead to even better fitted models in the next generations. This continues until a certain stopping criterion is reached. From a possible infinite solution set, one is able to efficiently narrow down to a single well-fitted model.

To conclude this subsection, we present Figure 2.7. This figure is no ground-truth and is not intended to be such. The placement of the algorithms in the plot is a combination of perceptions in the literature. The goal of this figure is to give the reader a feeling of how the different ML algorithms relate to one-another in terms of interpretability and accuracy. In here, the evolutionary algorithms are not placed, as they are more like a heuristic with some learning characteristics of which there is not an unambiguous perception on its interpretability and accuracy within the literature.

2.2.2 Explaining black boxes

It is often the case that people use the term 'black box' as a synonym for machine learning. ML models are referred to as black boxes since they are often not seen as intelligible concepts. Given some inputs; e.g., some features from a loan such as interest rate and maturity, the black box processes it, where after it gives some

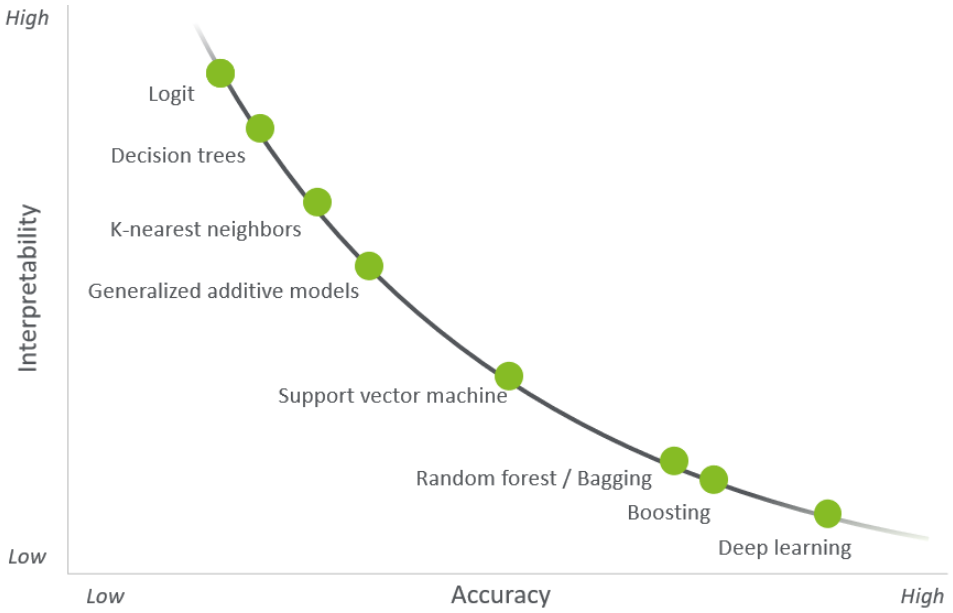

FIGURE 2.7: A conceptual visualization of the traditional trade-off between interpretability and accuracy.

output; e.g., 90% chance on a default. With the ever expanding field and complexity of ML models, these models are thus all labeled as black boxes.

There is, however, an actual difference between ML models and black boxes, since the latter is a subset of the former. Not all ML models are per se black boxes, because some of them are perceived as being simple to understand, for example linear or logistic regression models (EBA, 2021). Although there is no clear distinction between those simple and advanced models, we will in this thesis only focus on the simple, interpretable, models.

Post-hoc explanations versus inherently interpretability

When addressing the interpretability or explainability of machine learning models, often terms such as explainable artificial intelligence (XAI), explainable machine learning (XML), or interpretable machine learning (IML) are mentioned in the literature. Since in the literature the terms explainability and interpretability are used interchangeably, we will not try to distinguish between them in this thesis, and thus also use them mutually.

The above-mentioned terms XAI, XML, and IML are most often referring to model-agnostic post-hoc explanations (Molnar, 2022). That means, after running a black-box model, the post-hoc explainer considers the inputs and outputs of the model, and based on that, generates explanations. Figure 2.8b illustrates this. Well known examples are, for example, Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017), Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin, 2016), or TreeExplainer (Lundberg et al., 2019). Generally, these explanations consist of tricks such as visualizing partial dependence plots, showing counterfactuals, or finding feature interactions. Interestingly, these XAI techniques are also mentioned in the discussion paper of the EBA that we highlighted in the recent development Section 1.1.2. There, the supervisor is cautiously hinting at the potential advantages that XAI techniques can bring in the near future.

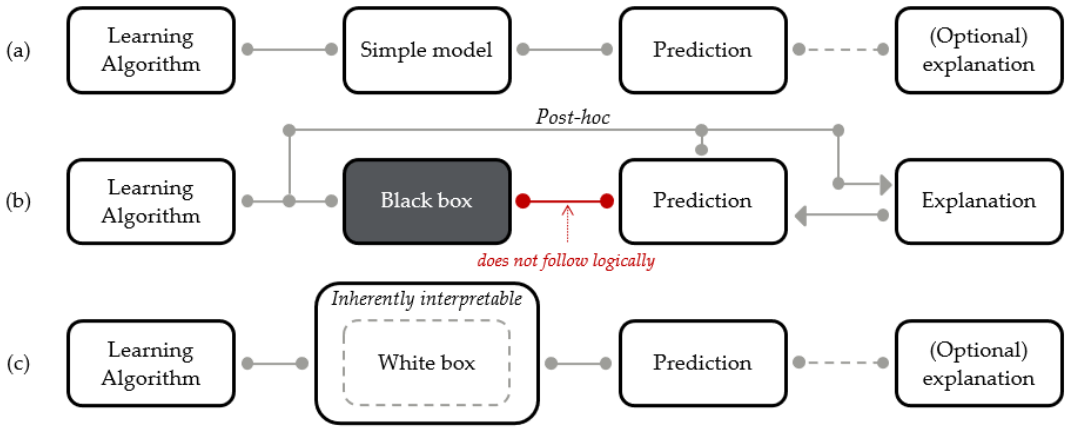

FIGURE 2.8: Conceptual overview of different ML pipelines. (a) Traditional supervised ML such as logistic regression. (b) Standard black box approach, where the result between the model and predictions is unclear, so post-hoc explanation techniques are needed. (c) White box approach, showing an interpretable model, that does not need post-hoc explanation techniques by design

However, the key-disadvantage of XAI techniques, is that their explanations are often not reliable, and can be misleading (Rudin, 2019). The EBA also acknowledge that XAI only helps to partially understand some model. Rudin further explains that explanations of XAI/XML must be wrong. "*They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation*". Having such an extra model, which also does not provide truthful explanations all the time, establishes even more model risk. In short, a post-hoc explainer is only able to approximate some causal relationship that happens in the model, without knowing and showing what actually happens.

Instead of finding a way around black box models with post-hoc explanations, recent research has focused its point of attention on inherently interpretable models. In literature, these are also often recognized as white box models, or referenced to as explainability / interpretability by nature or design. All those elegant variations boil down to the overview that is given in Figure 2.8c. Just as in Figure 2.8a, the model should be well comprehensible for humans, however, in the ideal situation, the white box model leverages advanced ML techniques in order to remain as accurate as the most advanced black box models. Therefore, they are placed separately from the traditional ML in Figure 2.8. To further explain, in Figure 2.8b, we see that the post-hoc explanation is needed, as the prediction does not follow logically, that is, a human cannot make the same prediction using the input parameters due to interpretability issues. Since XAI is sometimes helpful, it is illustrated in Figure 2.8a and 2.8c as optional techniques to be used. Examples of ML algorithms that are inherently interpretable (excluding the traditional linear and logistic regression) are classification and regression trees (CART) (Breiman et al., 1984), k-nearest neighbors, and generalized additive models (GAMs) (Burkart and Huber, 2021).

The increased interest in white box models did not bypass the financial industry. According to the IIF, a growing number of firms are expressing interest in using inherently interpretable machine learning models. N. Bailey, policy advisor for digital

finance at the IIF, also acknowledges this: "*When we first started looking at what methods firms were using for explainability, initially it was for post hoc techniques. But that has changed in the last couple of years, where now there is an understanding that these models are being built from the ground up, and you're not sacrificing performance*" (retrieved from interview of Marlin, 2021). This citation is an example that highlights the relevance of this research.

Global versus local explanations

Since a traditional logistic regression and white box models are interpretable by design, it does not mean that no post-hoc explanations can be applied. Sometimes those explanations can yield some extra benefits for the user or modeler in order to quickly get a rough estimation on how the model performs. This was also indicated in Figure 2.8a and 2.8c. Therefore, we shortly introduce the two main types of explanations.

The main distinction is made between global and local explanations. A global explanation tries to capture the overall working of a model, whereas the local one explains one single instance. In Figure 2.9 we can see the global explanation of an arbitrary model on the right side. This is specifically a feature importance explanation. On average, the attribute `SEX` contributes the most to all predictions in the model. However, if we look on the left side of the same figure, we see that for each record (represented by a small dot) the impact of the individual value of `SEX` of that specific record differs a lot from other records. In this case, your gender will either hugely positively impact the result, or negatively impact the result. So, depending on all other attribute values, you can also come up with a feature importance list per individual. That might show very different results than the global explanation, as in that case you compare an individual prediction with the average prediction. This is thus also a disadvantage of this XAI technique, as the different explanations never do separately explain the full model.

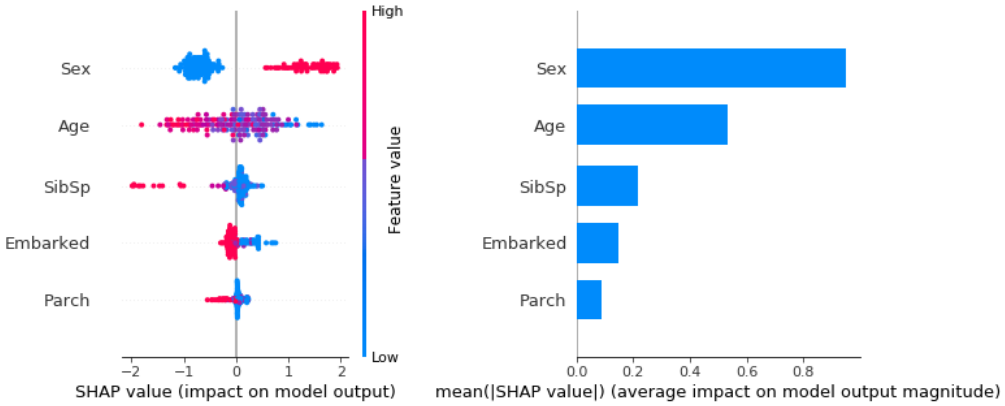

FIGURE 2.9: Example of visualizations of local and global explanations using SHAP (Lundberg and Lee, 2017; Berenbaum, 2020). The left part shows local explanations, where each dot at a feature represents a data point. The right part shows the global feature importance, i.e., on average for each feature.

2.3 Conclusion on theoretical context

To conclude this part, we will shortly recapitulate what is discussed and how we defined the scope of this thesis.

First, we addressed credit risk modelling. Although credit risk modelling covers many different aspects, we have decided to limit our scope to the Probability of Default component of it. The current standard for computing an estimation of the PD is a logistic regression model.

Additionally, we found that there are many ML algorithms that are suitable for a binary classification problem: logistic regression, GAMs, K nearest neighbors, decision trees, bagging, boosting and also deep learning. In choosing an ML algorithm, one should be careful that the output of the model can be a meaningful probability, since it is applied to the expected and unexpected loss function. Many of the ML algorithms with a high predictive power are considered to be black boxes. In this thesis, we will shift away our focus from black box models, and focus on inherently interpretable ML models. These so-called white box models do have the property to be understood by humans, and therefore are more applicable to be used within the IRB models of banks. Also, these models ideally do not carry additional model risk, which XAI-techniques suffer from.

Chapter 3

Assessment Framework Development

In this chapter, we answer research question A: *"What is the current state of machine learning adoption within IRB models in the industry and in terms of regulations and guidelines?"*, and B: *"What is an appropriate way of comparing machine learning algorithms in terms of applicability for the use within IRB models?"*. This encompasses A) getting an overview of the problem at hand, i.e., the current state of ML in the industry and in regulations, and B) developing a framework to assess ML-models' *applicability* for the use in IRB models.

To effectively develop a suitable assessment framework to compare different ML-models, we will investigate different perspectives from important stakeholders in answering research question A. The findings from these perspectives are inputs to construct an assessment framework which considers different interests. First we take the standpoint of banks, the industry perspective. After that, we zoom in on the (financial) regulatory context, especially getting familiar with the current rules and guidelines in the area of ML and AI. With that in mind, the assessment framework to evaluate ML algorithms on their potential to be applied within IRB models is constructed. To conclude this chapter, we will theoretically underpin the method of scoring. This is necessary to evaluate the different ML models and draw intermediate conclusions in Chapter 5.

3.1 Industry perspective

3.1.1 Motivation for the Internal Ratings Based Approach

First, for a better general comprehension, we address why many banks are motivated to develop internal models for the Internal Ratings Based Approach. For credit risk modelling, the EBA accepts two approaches: the Standardized Approach (SA), and the Internal Ratings Based Approach. The SA is a more general approach, whereas the IRB approach allows for a model designed by each bank individually, to be more aligned with the true characteristics of their asset portfolio's. The IRB approach is thus the area where the most advanced and sophisticated models could be used. Using advanced ML models can be leveraged to generate higher prediction accuracy, one of the main advantages of using ML in PD models.

A better PD model in the IRB model landscape has many advantages on its turn. Having better predictions, means that your estimates become closer to reality. Meaning that for many cases, one can do a more accurate risk assessment. Besides, the PD model in the IRB approach is also often used for other use cases, for example pricing or credit acceptance. Improving these operations of a bank can make it both

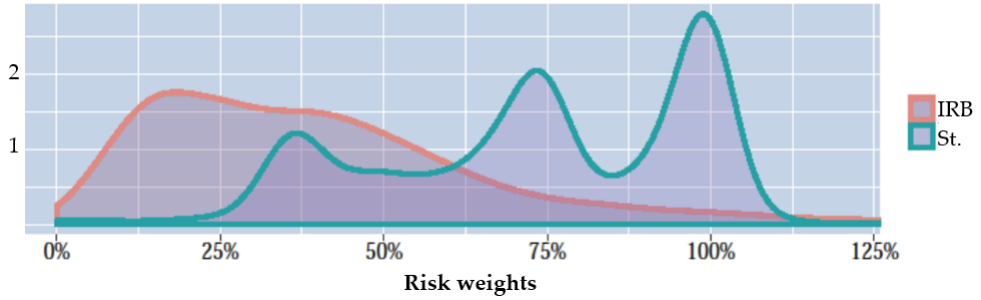

FIGURE 3.1: Average distribution of risk weights of exposures in both the SA and IRB approach (areas under the graphs are both equal to one) (Doeme and Kerbl, 2018).

more profitable, and more competitive. Lastly, for the capital requirements calculations, the SA uses fixed risk-weights for different assets, whereas in the IRB approach generally speaking, the risk weights are much smaller, see Figure 3.1. Lower risk weights results in lower capital requirements, which frees up money to be invested in the company. Although the advantages are clear and promising, there are still some challenges to fully implement ML in the IRB approach.

3.1.2 Challenges for implementing ML identified by the industry

In a survey of the IIF, various challenges of using ML in credit risk management are identified by a large group of respondents (IIF, 2019). In Figure 3.2 the key challenges are presented. Over 80% of the respondents feel that the main bottleneck to implement ML originates at the supervisor. By either a lack of understanding, or by the lack of willingness to adopt new processes, the supervisor is seen as an obstruction. This specific challenge will be zoomed in on in the next section, Section 3.2, when we specifically take the standpoint of the regulator.

Just a little behind the most identified challenge, is the 'difficulty of explaining processes'. Under this challenge, financial institutions report that they have concerns about the transparency, auditability, and interpretability of results.

Another problem that is highly attached with other challenges, is the cost of implementing the new technology. From Figure 3.2, IT infrastructure related problems, and the availability of appropriately skilled staff, are ultimately related to costs. However, these two also have their own origination why it is a challenge. IT related challenges to the implementation of new techniques are for example that more data and processing power are needed. Also, integrated IT infrastructures are needed as more and more models will be connected to each other. This poses some extra limitations, as legacy systems of banks often do not support modern coding languages (IIF, 2019). The availability of appropriately skilled staff also causes costs to rise, but additionally slows down the whole process of implementing ML. Since regulators and financial institutions are both fishing in the same pond, there is not a quick solution for this challenge to overcome.

Finally, it is noted that most of the financial institutions who have ML in production in other areas of their operations, have engaged their supervisor in their application of ML within IRB models.

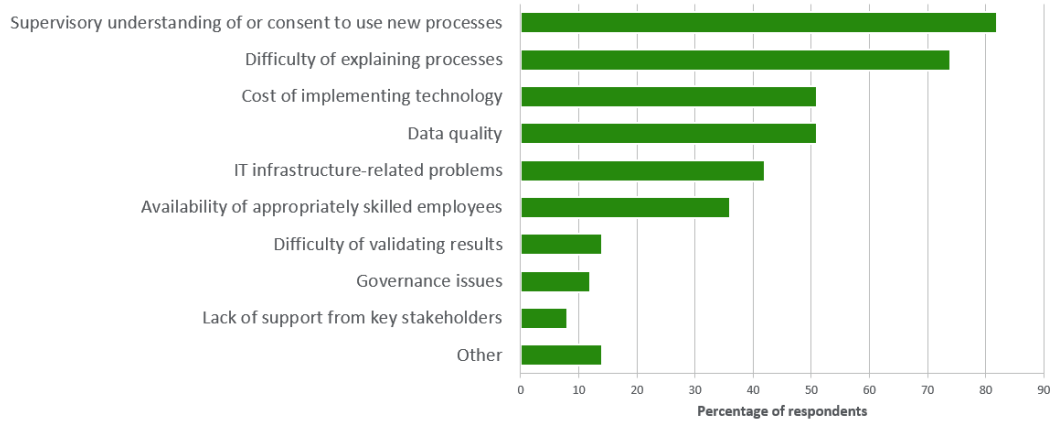

FIGURE 3.2: Key challenges identified by the industry of using ML compared to previously used models (retrieved from the survey conducted by IIF, 2019).

3.2 Regulatory context

In this section, we address the regulatory perspective on ML and AI. We do this in order to get a view on the expectations that regulators and supervisors have in the implementation of ML. As described by the Financial Stability Institute (FSI) of the Bank of International Settlements (BIS) in a recent report (Prenio and Yong, 2021), there is no standard-setting body that develops international guidance or standards in the area of AI governance. Also, "*authorities' views on how these (AI governance related) themes should be implemented are still evolving*".

Since there are no international standards, and many authorities or regulatory bodies are still in the process of developing some guidance for the use of AI, there is not yet a golden standard. In order for us to develop an assessment tool that reflects the general opinion and guidelines, we will address the most significant bodies, their regulations and some (discussion) papers below.

3.2.1 Bank for International Settlements

We start off with the BIS. The BIS is a guiding body, to promote global monetary and financial stability through international cooperation. Other institutions and committees that are part of or closely linked with the BIS are the FSI and the BCBS. The latter has very recently published a newsletter (*Newsletter on artificial intelligence and machine learning*) in which it provides details on its internal discussions regarding AI and ML (BCBS, 2022). In it, it highlights three focus areas for further investigation on the supervisory implications. First, the extent and degree to which outcomes of models can be understood and explained. Secondly, ML model governance structures, including responsibilities and accountability for ML-driven decisions. And thirdly, it wants to further investigate the potential implications of broad usage of ML for the resilience of individual banks and the broader financial stability. Especially the first two focus areas, understanding & explainability and model governance, are important factors to take into account when evaluating the applicability of ML algorithms for IRB models.

Meanwhile, the FSI published their own insights in their report *Humans keeping AI in check – emerging regulatory expectations in the financial sector* (Prenio and Yong, 2021). They have summarized the existing issuance of authorities' expectations and

guidance into five common principles: reliability/soundness, accountability, transparency, fairness, and ethics. Additionally, it is noted that these principles are most of the time also used in assessing the traditional, logistic regression, models. However, the current discussion revolves around how these principles are approached differently in case of ML models. For example, in terms of accountability, for a logistic regression model, a company can assign a single model owner, whereas for an ML model this might not be as straightforward. One can imagine that different aspects of the model, e.g., IT infrastructure and monitoring of several components of the model, should be covered by different employees. The FSI also gives an overview of these regulatory expectations relating to the aforementioned principles, see Table 3.1.

TABLE 3.1: FSI's summary of regulatory expectations relating to the AI common principles (Prenio and Yong, 2021).

Common principle	**Regulation/legislation/guidance**
Reliability/soundness	• Similar expectations as those for traditional models (e.g., model validation, defining metrics of accuracy, updating/retraining of models, ascertaining quality of data inputs). • For AI models, assessing reliability/soundness of model outcomes is viewed from the perspective of avoiding causing harm (e.g., discrimination) to consumers.
Accountability	• Similar expectations as outlined in general accountability or governance requirements, but human involvement is viewed more as a necessity. • For AI models, accountability includes "external accountability" to ascertain that data subjects (i.e. prospective or existing customers) are aware of AI-driven decisions and have channels for recourse.
Transparency	• Similar expectations as those for traditional models, particularly as they relate to explainability and auditability. • For AI models, external disclosure (e.g., data used to make AI-driven decisions and how the data affects the decision) to data subjects is also expected.
Fairness	• Stronger emphasis in AI models (although covered in existing regulatory standards, fairness expectations are not typically applied explicitly to traditional models). • Expectations on fairness relate to addressing or preventing biases in AI models that could lead to discriminatory outcomes, but otherwise "fairness" is not typically defined.
Ethics	• Stronger emphasis in AI models (although covered in existing regulatory standards, ethics expectations are not typically applied explicitly to traditional models). • Ethics expectations are broader than "fairness" and relate to ascertaining that customers will not be exploited or harmed, either through bias, discrimination or other causes (e.g., AI using illegally obtained information).

3.2.2 European Union

On a European level, posing regulations cross-industry, stands the European Union. This institution makes high-impact regulations, such as the well-known *General Data Protection Regulation* (GDPR) (EU, 2016). Although the GDPR does have major implications for banks, there are only a few articles and rules that have a direct link to the use of ML. We will discuss those articles that are of use in constructing an assessment framework to evaluate the applicability of ML in IRB models, and therefore choose to not incorporate, for example, data collection related articles, as this is outside the scope of this thesis.

One important note to make is that some of the articles mentioned below actually do not restrict ML models in the *IRB approach*. That is because in IRB models, capital calculations are done, and data records of individuals do not impact the data subjects itself. In contrast, for *credit acceptance* and *pricing models*, the data subject is impacted by the decision of an ML model. As we will see when we treat the specific regulations in CRR below in Section 3.2.3, PD-models from the IRB approach must also play an essential role for other internal processes, such as credit approval and decision-making processes. To increase the relevance of the assessment framework, we will therefore also cover those regulations that deal with these other processes where the PD-model is also used.

- First, article 5.1(c) states that personal data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'. It is summarized thereafter in the GDPR as 'data minimization', which can yield troubles in implementing ML models. That is because other regulations (CRR, article 180) oblige a bank to have five years of data history for risk drivers. As more complex models also tend to have more risk drivers, this is a limiting factor for the implementation in ML. On top of that, the difficulty of this data collection increases exponentially when a bank uses more risk drivers, since the easy risk drivers, with easily accessible history, are used first (Folpmers, 2021).

- Second, Article 15.1(h) gives the customer the right to access meaningful information about the logic involved in automatic decision-making. This makes that the bank cannot use a deep neural network, that the controller itself can also not explain to a customer in a logical way.

- On top of that, Article 22 states that the customer has the right to obtain human intervention on the part of the controller. Again, this is not easily done with a black box model.

Recital 71, which is a complementary document the regulator writes as an additional explanation on the regulations in a nonbinding language, summarizes the intention of the regulator comprehensively: "[an automated process] should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."

In short, these guidelines and assessment list comprise the following aspects:

Another important view from the EU on AI and ML is written by a High-Level Expert Group on AI (HLEG-AI, set up by EU), who formulated the *Ethics Guidelines for Trustworthy AI* (HLEG-AI, 2019) and the *Assessment List for Trustworthy Artificial Intelligence* (ALTAI) (HLEG-AI, 2020). The former served as non-binding guidelines and was build upon four "Ethical imperatives", which are: respect for human autonomy, prevention of harm, fairness and explicability. The latter, the ALTAI, is a tool that supports the former in terms of actionability, based on seven key requirements:

1. Human agency and oversight

2. Technical robustness and safety
   Including general safety, resilience to attacks, accuracy, and reliability in terms of fall back plans and reproducibility.

3. Privacy and data governance

4. Transparency
   Covering traceability, explainability and communication

5. Diversity, non-discrimination and fairness

6. Environmental and societal well-being
   Consisting of environmental well-being, impact on work and skills, impact on society at large or democracy

7. Accountability
   Including auditability and risk management.

3.2.3 EBA

Recall from Section 2.1 where we addressed the capital requirements that banks need to hold on to. It was also shown in Section 3.1.1, specifically in Figure 3.1, that the use of the IRB model can drastically reduce these capital requirements. Therefore, the EBA's Capital Requirements Regulation (CRR) imposes strict rules on the implementation of ML in an IRB model. Below, the most important, and most restricting rules are itemized.

- *CRR article 174*: the institution shall complement the model by human judgement and human oversight to review model based assignments and to ensure that the models are used appropriately. In this case, the complexity of an ML model can make it difficult to allow for human interference and judgement.

- *CRR article 175*: the institution shall document the design and operational details of its rating systems. It shall provide a detailed outline of the theory, assumptions and mathematical and empirical basis of the assignment of estimating PDs. One can imagine, that for a more complex model, this can become exponentially difficult. On top of that, article 175.4(c) imposes that the bank shall indicate any circumstances under which the model does not work effectively. The more black box a model is, the harder it becomes to find out and to document what can go wrong.

- *CRR article 179*: in quantifying the risk parameters (including PD), the IRB model must be "intuitive". There shall be an easy relationship between risk drivers and risk parameters. A classic logistic regression is in this case allowed, since, a positive coefficient will lead to a higher PD and the other way around. A neural network would not fit this requirement, as there are no relationships directly observable, and it is also not intuitive.

- *CRR article 180*: obliges a bank to have five years of data history for their risk drivers. This article was already treated in the EU regulation which focused on data minimization, which can together pose a problem when more risk drivers will be used in complex models.

- *CRR article 189*: all aspects of the rating and estimation process of the PD shall be approved by the management body and senior management. They shall have a detailed comprehension of its management reports, and a good understanding of the rating systems designs and operations.

Following the industry's need for a more perspicuous standpoint of the EBA with respect to ML, as well as, the EBA wanting to get input on current developments with regard to ML in IRB models, it developed the *EBA discussion paper on machine learning for IRB models* (EBA, 2021). This thesis started off with the introduction of this paper. Now, we highlight the things from within that paper that indicate the standpoint of the EBA. It can be summarized in three identified challenges with complex models: 1) interpreting their results, 2) ensuring their adequate understanding by the management functions, and 3) justifying their results to supervisors. However, when moving towards more complex models, the EBA does recognize that evaluating complexity of the model becomes more important. They propose five characteristics that are useful for this process:

- The number of parameters.
- The capacity to reflect highly non-linear relations between the variables accurately.
- The amount of data required to estimate the model soundly.
- The amount of data from which the model is able to extract useful information.
- Its applicability to unstructured data (reports, images, social media interactions, etc.).

Additionally, they made three distinct recommendations: avoid unnecessary complexity, make sure the model is correctly interpreted and understood, and set up reliable validation processes.

Finally, we present an overview of the regulatory context in Table 3.2. (Note: the central bank of the Netherlands, De Nederlandse Bank, also formulated *General principles for the use of Artificial Intelligence in the financial sector*. The concepts in this paper, formulated by the 'lowest' regulatory body, are already well covered in the other above-mentioned guidelines and regulations.) This regulatory context, together with the industry perspective, answers research question A. More importantly, they form the constructs for the assessment framework.

3.3 Components of the framework

After addressing the perspective of the industry, and the applicable regulations concerning AI/ML in IRB models, we formed an idea on relevant aspects to take into consideration in the assessment framework. In this section, we zoom in on these aspects and eventually answer research question B by presenting the assessment framework.

In order to structure the assessment framework, we make use of four different components that one encounters when developing and implementing an ML model. These are model design, input and output relationship, output of the model, and the model use and implementation. They are discussed below.

3.3.1 Model design

The design of the model focuses on the operations within the model. This excludes all concepts related to the output of the model. It does involve the construction of the model with the algorithm, with the use of the inputs. Specifically, this touches

TABLE 3.2: Regulations and guidelines that impact the use of ML in IRB models.

From	Act / guidelines / article	Principles and findings
BIS	BCBS Newsletters: Newsletter on AI and ML	Continued focus areas: • Explainability: transparency in model design, operation, and interpretability of model outcomes. • Governance structures: including responsibilities and accountability for AI/ML-driven decisions. • Implications of ML models for the resilience and financial stability
	FSI Insight No 35: Humans keeping AI in check	Challenges in implementing the AI-related expectations or guidance (see Table 3.1): • Transparency • Reliability and soundness • Accountability • Fairness and ethics • Addressing regulatory and supervisory challenges through proportionality
EU	GDPR: • Art. 5.1(c) • Art. 15.1(h) • Art. 22 + recital 71	Regulations focus on: • Data minimization • Right for customer to access meaningful information about the logic involved in automatic decision-making. • Right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision of the automated decision.
	ALTAI	Actionable key requirements for AI: 1. Human agency and oversight 2. Technical robustness and safety 3. Privacy and data governance 4. Transparency 5. Diversity, non-discrimination and fairness 6. Environmental and societal well-being 7. Accountability
EBA	CRR • Art. 174 • Art. 175 • Art. 179 • Art. 189	Comprising: • Include human oversight • Extensive documentation • Intuitive model design • Detailed comprehension of senior management on the systems designs and operations
	Discussion paper: On ML in IRB models	To evaluate model complexity: • The number of parameters. • The capacity to reflect highly non-linear relations between the variables accurately. • The amount of data required to estimate the model soundly. • The amount of data from which the model is able to extract useful information. • Its applicability to unstructured data (reports, images, social media interactions, etc.). Additionally, three applicable recommendations: 1) avoid unnecessary complexity, 2) make sure the model is correctly interpreted and understood, and 3) set up reliable validation processes.

upon the interpretability of the model itself. As we have seen in the previous section, and in the summary of it in Table 3.2 a lot of regulations also comment on the interpretability of ML. Throughout literature, the concepts of interpretability, explainability, and transparency are widely used, often with different definitions, but all boil down to the same idea. There is a certain inconvenience in making an assessment framework that should take into account the interpretability of models, since it is an interrelated concept with many other principles and findings in Table 3.2. In an FSI's Insights paper, they highlight this issue as well: "*Transparency of an AI/ML algorithm is a prerequisite to fulfilling some of the other sound AI governance principles*" (Prenio and Yong, 2021). If it is not transparent, we cannot assess its reliability, performance, fairness, or any other topic related to the inner workings of a model. As for this reason, together with the reason that interpretability is not easily measurable in itself, we will decompose interpretability to more measurable concepts. In this way, we can measure the suitability of a model design for the use in IRB models.

Decomposing interpretability

When addressing interpretability in ML, most literature focuses on explanation methods, such as Robnik-Sikonja and Bohanec, 2018 and Barredo Arrieta et al., 2020. However, in the decomposition of this subsection, we explicitly do not focus on explanations of an ML model's prediction. We step away from the concepts of 'XAI', and center the attention on the ability for humans, from a customer to model owner, to understand the model. In the paper of Lipton, 2018 the term transparency is used for this. It considers three hierarchical dimensions in terms of going deeper into the model, focusing on the human-level understanding of a model. First simulatability, which is the transparency at the level of the entire model, decomposability at the level of individual components (e.g., parameters), and at the level of the training algorithm we investigate the algorithmic transparency. Below, these are explained further:

- **Simulatability** is the first level of transparency. It refers to a model's ability to be simulated by a human. Next to the simplicity of a model, also the size of the model is of influence on this dimension. The most simulatable model is a model that is fully understood by a human when it takes the input data and all relevant parameters, and produce a prediction within a reasonable amount of time. Taking size and the simplicity of the computations into one criterion is because the trade-off between these two vary between models and is better captured as one to resemble simulatability. The quantity of 'reasonable' is a subjective notion, however, as Lipton, 2018 describes, given the limited capacity of human cognition, this ambiguity might span only several orders of magnitude, and is therefore justified. In 3.5 we will elaborate more on how this is measured.

- **Decomposability** is the second level of transparency. It denotes the ability to break down a model into several components. The components consist of the inputs, and the computations involved, subject to the parameters of the mdoel. As a prerequisite, the components must be intuitive, e.g., no advanced feature engineering techniques are used to construct input parameters. The parameters and computations can be seen as the interactions that input features have. These interactions should be intelligible, or easily understandable. As an example, Lipton, 2018 mentions that each node in a decision tree corresponds to

a plain text description. Each interaction is therefore easy to understand: "If variable X is larger than 10, and variable Z is lower than 6, then predict 1".

- **Algorithmic Transparency** is the final, third level of transparency, which encompasses the algorithm itself. In entails the learning process that the algorithm uses. For linear models, one can actually understand the shape of the error function or surface. However, some of the larger and more complex algorithms, a strong stochastic property of the learning process can make the algorithm opaque. This makes outcomes less well reproducible. Some examples are heuristic optimization procedures of neural networks, or the heuristic algorithm used in genetic programming (see subsection 2.2.1).

3.3.2 Input-output relationship

The relationship between input and outputs of a model form an important aspect of an ML algorithm. We identified one important criterion to evaluate different algorithms on.

To explain what we mean with input-output relationship, recall the previous paragraph, where we touched upon the decomposability of a model. We can also apply this concept to the input and output relationship. The decomposability of this subject is better known as individual variable contribution. An additive model is much more interpretable than a model in which variables use other operations such as multiplications, or even nonlinear operations. That is because one can, for a given input, give all individual variable contributions for the output. It relates to concepts such as human oversight, mentioned in the CRR articles of the EBA and the ATLAI of the EU. Also, with clear individual variable contributions, the customer is able to challenge the outcome. A fully connected deep neural network is an example of a model that does not have clear input-output relationships. The deep neural network also has nontransparent individual variable contributions to the outcome. The only way to distill some variable contribution from the model is by the use of XAI or partial dependence plots (PDPs). These show the marginal effect of one feature to the outcome, leaving all other variables the same. However, in such a neural network, when changing two variables, the user is almost not able to know what the outcome will be, and in fact cannot easily say whether the prediction is going up or down, even with very logical features such as `salary`. This comes from one of the two main disadvantages from PDPs, namely it assumes independence between variables (which might not be the case, as illustrated above). Another disadvantage is that it is not able to show the heterogeneous effects because PDPs only show the average marginal effects. That is, when for half of the data the outcome is positively impacted, and for the other half the same size negatively impacted, the PDP will show a horizontal line of dependence, since *on average* there is no effect.

Decomposing the effects of one feature (or the interaction between multiple features) on the final prediction, without making use of PDPs, is necessary for the interpretation, but more importantly, also to comply with regulations. The BCBS touched upon this with the implications of ML models for financial stability, and the FSI mentioned the concept of reliability and soundness. It all boils down to having an *economically justifiable relationship between inputs and output*. Within financial institutions, this often implies having a monotonic relationship between input and output when this is expected. The relationships should have an economic rational. For example, when getting a higher income, the PD should always decrease (monotonically decreasing), since a higher income would make the loan provider more

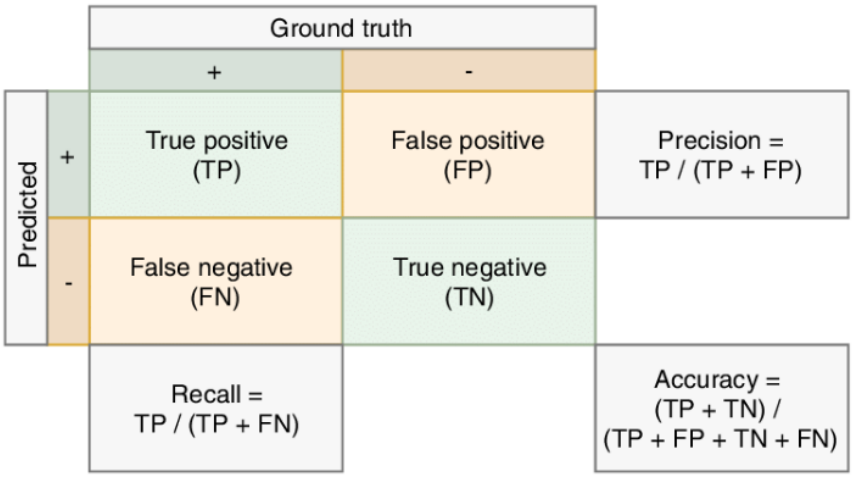

FIGURE 3.3: A confusion matrix (retrieved from Jeppesen et al., 2019).

sure about getting paid back the loan. It is important to note, that this criterion does not imply that all input variables relationship with the output should be restricted. They, however, should make (economically) sense, and where necessary, the model developer should ideally restrict the model to a relationship that is appropriate.

3.3.3 Output

With regard to the output of the model, there are two aspects that were mentioned multiple times in the regulatory context. These are the performance of the model, and the fairness.

Classification performance

As mentioned in the Introduction of this thesis in Section 1.1.2, ML models have often shown to outperform traditional models. This performance is generally measured with a few well known metrics. We make use of the concept Area under the Curve (AUC). This concept will become clear when we will further elaborate on the ”curves we use. The two measures that are used in this thesis make use of two different ‘curves’, namely the Receiver Operating Characteristic curve, and the Precision-Recall curve, which will be elaborated further below.

To explain these concepts, one first needs a thorough understanding of a confusion matrix, which is depicted in Figure 3.3. The matrix is divided in four squares, by the combination of positive (1) and negative (0) classes and the ground truth and predicted value. The two squares in green form the correct predictions: either correctly predicted records of class 1, true positives, or correctly predicted records of class 0, true negatives. The squares on the other diagonal are incorrectly predicted, false negatives or false positives. All predictions of the models are of course probabilities, since we are predicting the PD. That makes the distribution of data records across the confusion matrix highly dependent on the so-called threshold level, the point where we decide to predict 1 if the prediction is higher than the threshold, and 0 if it is lower.

After getting familiar with the confusion matrix, we can zoom in on the performance measures. Starting with the AUC of the Receiver Operating Characteristic (ROC) curve, which together are abbreviated as AUROC. The ROC curve is a graph that shows the performance of a classification model at all threshold levels

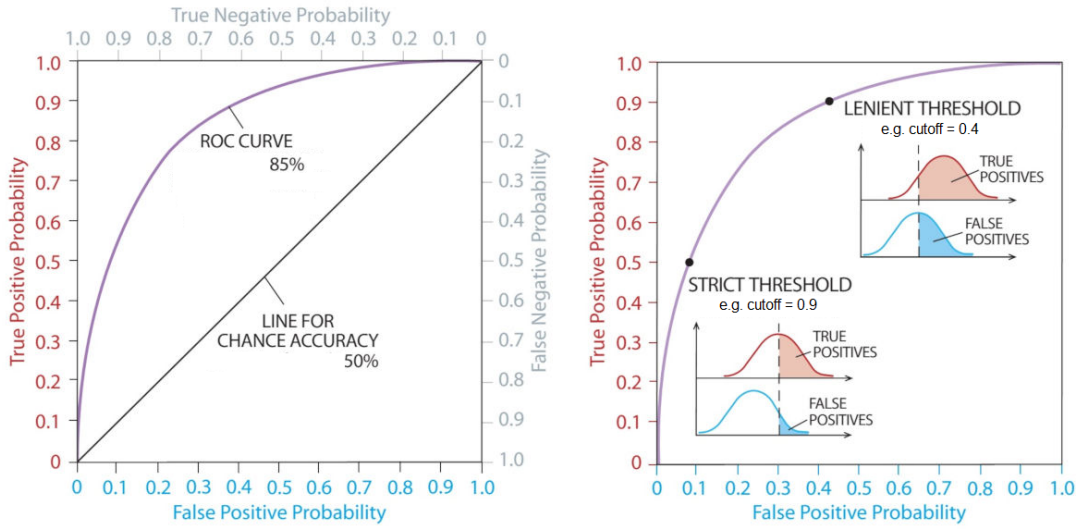

FIGURE 3.4: Example visualizations of the ROC curve, AUC, and true and false positives (based on a figure from Swets, Dawes, and Monahan, 2000). The left figure shows an ROC curve with corresponding AUROC of 85%, The right figure shows how the true positive and false positive rates vary based on the threshold level.

(Narkhede, 2018). Figure 3.4 gives two examples of ROC curves. The curve shows the true positive rate versus the false positive rate for all possible threshold levels. The area that is under the ROC curve, which is the AUROC, is 85% in the left figure. In the right figure of Figure 3.4, we see how choosing a different threshold level will affect the true positive and false positive rates. A threshold level of 0.4 for example, forces the model to classify a prediction as a default when it predicts the PD to be above 40%. The threshold level can also be set to 0.9, shown in the right plot of Figure 3.4 in the lower left corner of the ROC plot. In this case, very little data observations will be classified as default, so there is a low true positive rate, and a low false positive rate. A perfect classification follows the left and upper border of the plot. The area under the curve, the AUROC, would then be equal to 1, since the total area of the 1×1 square is 1. A fully random classifier has 50% of guessing right, and therefore shows a diagonal line, see the left figure of Figure 3.4. The advantage of the ROC curve is that it gives a good indication on how well the model is performing. Next to that, the AUROC is used in many researches, as it has well-defined interpretations of the AUC (probability that a positive is ranked higher than a negative), as well as other concepts such as the distance above the curve (probability that an informed decision is made rather than guessing). However, for this research, we will only focus on the values of the AUROC scores of different models.

Another commonly used performance indicator makes use of the precision recall (PR) curve. Again, the measure is the AUC, which is now abbreviated with AUPRC. Like the ROC curve, the PR curve also resembles all possible threshold levels that can be chosen to predict a class positive, however, now the values plotted resemble the precision and the recall. The precision versus recall is a commonly known trade-off. Their calculations are also shown in the confusion matrix in Figure 3.3. A high precision relates to a low false positive rate, and high recall relates to a low false negative rate (Pedregosa et al., 2011). In Figure 3.5 the ROC and PR curve for two models are depicted to show their differences. One can interpret the PR curve as follows: to predict at least 50% of the defaults correctly (corresponding to a recall

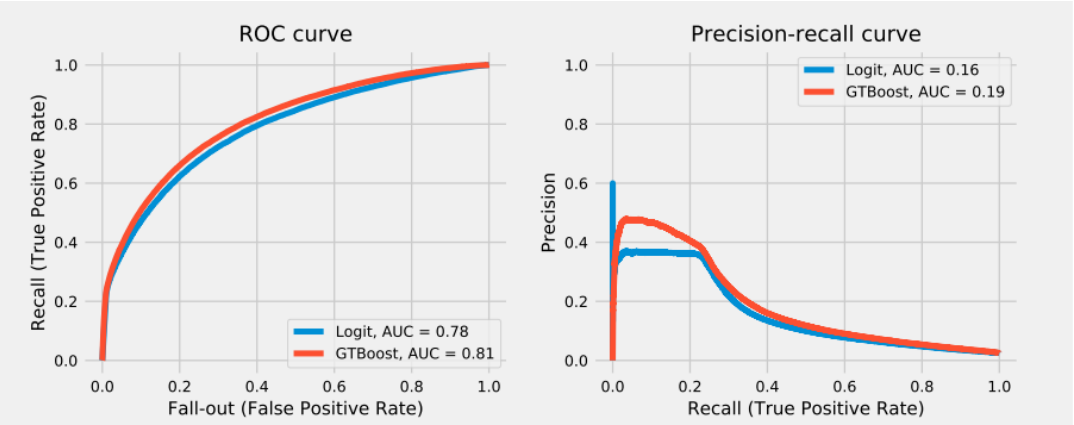

FIGURE 3.5: Example of two ROC curves and the corresponding precision-recall curves.

of 0.5), we need to incur a precision of approximately 10%. That indicates that 90% of the positively predicted will actually not default. The figure clearly illustrates that although the AUROCs are fairly comparable, the PR curve shows that there is an actual difference in the output of the model. Especially in highly imbalanced datasets, the AUPRC is of added value as a performance measure (Sigrist, 2022).

The main difference between the AUROC and the AUPRC is that the AUROC gives a general idea on how well a model is performing, and generally lies between 0.5 (random guessing), and 1[1]. Whereas, the AUPRC is also dependent on the fraction of positives and negatives in the dataset. To illustrate this, one can always say that a model with an AUROC of 0.9 is a very good model, but a model with an AUPRC of 0.9 does not yet tell anything about its performance without knowing the data that it was tested on. To conclude, although the AUPRC is good to compare models fitted on the same data with each other, it does not tell enough on the general predictive power of the model. Therefore, we use the two metrics AUROC and AUPRC in conjunction.

Fairness

Another aspect that should be addressed in the model's output is fairness. Mentions in among others the regulatory documents FSI Insight, the GDPR, and the ALTAI (see Table 3.2) indicate the importance of this aspect. Fairness in ML and AI is also an important topic that is often addressed in literature. However, the fairness of a model can only be assessed when a dataset is available in which features are collected that cannot be used in the model. In this way, one can check whether, based on non-discriminatory features, the model can still discriminate. As this data is not available in open source credit datasets, and since fairness in AI is a large research area in itself, we will not cover the fairness of the model in the assessment framework. Additionally, after all, an inherently interpretable model would theoretically be less sensitive to unforeseen unfairness.

[1]Technically, the AUROC can also be lower than 0.5, but then the model performs worse than random guessing.

3.3.4 Model use and implementation

The last aspect to cover in the assessment framework is the use and implementation of the model in the organization.

The most important aspect is the model governance. This encompasses a large subset of characteristics of a model when it is in use. The degree of 'manageability' of a model is important in this metric. It covers the range from business operations and privacy and security management, to IT governance. These different roles are important to address, but also the policies and procedures underlying it, which can be seen as accountability. The accountability is most simple if only one model owner is in place. However, model governance can also grow to be a task of several people, all safeguarding a specific component, or process in the model. What is also included in this criterion, is the documentation that comes with an algorithm. All these related concepts will be combined in one criterion. That is because these concepts are intertwined, they influence each other. Splitting these up, would not be of added value, and could harm the interpretability of the assessment framework.

Lastly, one might consider the training time for the model as a criterion for under the category of model use and implementation. However, since the model development is only once in several years, and updating of the model is also happening only approximately once per year, this is not a significant criterion to take into account when assessing the models.

3.4 The assessment framework

After taking the components mentioned in this Section 3.3 into account, we design an assessment framework to use for the comparison of several ML algorithms on their applicability in IRB models. This assessment framework is depicted in Table 3.3 below. Therewith, research question B is answered, which enables us to compare different ML-algorithms effectively in Chapter 5.

Multiple criteria, such as `simulatability` and `decomposability` are not straightforward to measure. In the 'explanation' column, for one criterion, one can identify multiple metrics. For `decomposability`, for example, the metrics cover both the number of features and the number of interactions between them. As these metrics are strongly intertwined, we choose to aggregate those metrics to one criterion. In this way, we make sure that the criteria are as much mutually exclusive as possible.

TABLE 3.3: The assessment framework for evaluating the applicability of ML in IRB models.

Subject	Criteria	Category	Explanation	Related concepts
Model design	Simulatability	Interpretability	The ability for a human to achieve the same output, given some input parameters. This criterion covers both the simplicity and the size of the model	Simplicity, sparsity, compactness
	Decomposability	Interpretability	Consisting of the number of features used in the final (regularized) model, and the degree of interactions between those features	Sparsity, size of feature space, feature interactions, non-linearity
	Algorithmic transparency	Interpretability	Covers the full algorithm, including subcomponents of the learning process of the model. This focuses on the ease to get a thorough understanding of the algorithm	Transparency, thorough understanding
Input - output relationship	Economically justifiable relationship between input and output	Interpretability	Economically justifiable relationships often require monotonicity, is this the case? To what extent is it possible to make additional constraints in the learning phase, or post-hoc alterations to the model	Fairness, (economically) justifiably, consistency, soundness, monotonicity, challengeable outcome, human oversight
Output	Performance - AUROC	Performance	Area under the ROC curve	-
	Performance - AUPRC	Performance	Area under the PR curve	-
Model use and implementation	Governance and documentation	Implementation	Covers the full length of how well the model is manageable, whether 'new' problems arise in comparison with traditional models	Manageability, accountability, responsibility, IT security, documentation

3.5 Using the assessment framework: scoring

Comparison of the algorithms is enabled by the assessment framework. All models will be assessed by scoring them on all criteria. As one can see, the criteria are not all straightforward to measure. To decide on how we should score the models, we take a closer look at the literature on scoring different models using an assessment framework.

First, the level of measurement is an important concept to understand. It tells how precisely variables can be recorded. There are four different levels of measurement, see Table 3.4. Starting from the top, the nominal measurement level is the least explicitly defined. Nominal variables, such as `country` and `gender`, are categorical variables that only can be compared between instances of being equal, or not equal. There is no hierarchical order between the values of that variable, in contrast to ordinal variables. This second level of ordinal variables consists of variables such as a likert-scale (*strongly disagree* up until *strongly agree* with 5 levels), and for example `learning ability`, which would be scored from low, to medium, and high. These variables can also be compared in terms of which observation is larger than the other. Thirdly, interval variables are numerical variables without the property of a true zero point. This specific level of measurement is used in for example measuring `temperature` in Celsius or Fahrenheit and `IQ points`. Numerical values can be compared quantitatively and also subtracted and added to get meaningful results, i.e., ranges, for example. Without having a true zero point, these measurements do not allow for multiplication or division. A true zero point is not present in for example the variable IQ, or degrees of Fahrenheit; a zero on these variables do not mean that there is a complete lack of that property. Having a true zero point is only applicable at the ratio measurement level. Examples of this most explicit level are `weight` and `speed`. The mathematical operations of multiplications and divisions make one able to calculate ratios and percentages of the variables. Note that from the most explicit level of measurement, one can always use properties of measurement levels that have a lower explicitly to compare different variables.

TABLE 3.4: Levels of measurement with the respective properties (Bhandari, 2020).

Level	Categories	Rank order	Equal spacing	True zero	Math. oper.
Nominal	X				$=, \neq$
Ordinal	X	X			$=, \neq, <, >$
Interval	X	X	X		$=, \neq, <, >, +, -$
Ratio	X	X	X	X	$=, \neq, <, >, +, -, \times, \div$

In the Assessment framework of Table 3.3 we see that most of the criteria are ordinal variables. One is able to argue why one ML model is more simulatable than the other, but cannot exactly say how much better. The performance measures on the contrary are examples of ratios; the AUROC and AUPRC both have true zero points, namely.

As most of the data is ordinal, we are restricted in what way we measure the different criteria in the assessment framework. Cooper and Schindler, 2014 describe for this type of data two different scaling types, namely rating scales and ranking scales. These measurement scales are often used when measuring the more complex constructs, Cooper and Schindler note.

"*Rating scales are used when participants score an object without making a direct comparison to another object or attitude*" (Cooper and Schindler, 2014). A few examples of these are the Likert-type scale, the Multiple Rating List scale, and Graphical rating

scale. These scales all require a definitive positive or negative statement, with which one can agree or disagree, which makes it not applicable for this thesis. Another example of a rating scale is a numerical scale, in which the numerical scores have equal intervals that separate their numeric scale points. However, this scale needs concepts that are standardized or defined, the numbers should anchor the end-points and points along the scale. For example, in evaluating whether one would buy a new product, the numerical scale from 1 to 7 has well-defined end-points: 'definitely won't buy' and 'would definitely buy'. Concepts of the assessment framework such as simplicity and decomposability do not possess these properties of well-defined anchor points.

The other option is to use ranking scales. Ranking scales are used to directly compare two or more objects. Three different options are possible for these types of scales: Paired Comparison scale, Forced Ranking scale, and a comparative scale. The first one consists of making unique pairs of all options, and comparing these one by one. In our case, with four alternatives, we need to make six paired comparisons. This becomes numerous comparisons when assessing all the models across all different criteria. The forced ranking scale on the other hand just asks the evaluator to rank the alternatives from first to last place. The advantage of this is that fewer comparisons need to be made. Additionally, this type of scaling does not require anchored end-points. Lastly, the comparative scale takes one alternative, to which all other alternatives should be contrasted. In our case, naturally, the logistic regression becomes the reference level. However, since the new models are naturally a bit more complex than the standard logistic regression models, this type of scoring does not satisfy the requirements because the scale does not allow for a well-defined distinguishing between the other alternatives.

Ultimately, following the literature, the forced ranking scale is the best choice for evaluating which ML models are applicable for the use in IRB models by means of the assessment framework. To illustrate how to compare and contrast the alternatives, Figure 3.6 shows an often used visualization for these kinds of problems. In here, one can see which model performs better on which criteria compared to others in just a moment of time. Although the lines between data points may mislead the reader that there is some chronological order, it is often used, as it is more readable in this manner.

However, one large flaw on using this type of ranking scale and visualizations for the scoring of the models on all criteria has the major flaw of 'losing' information. For example, the distance between the number one and number two might be very little, whereas between the number two and number three there is a huge difference.

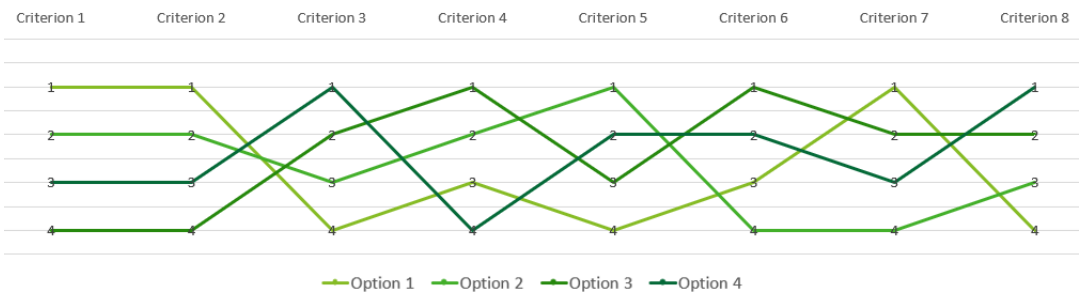

FIGURE 3.6: A simple example of how ranking scales are generally visualized. Only the rank is visualized, and one cannot distill how much better one option is over another.

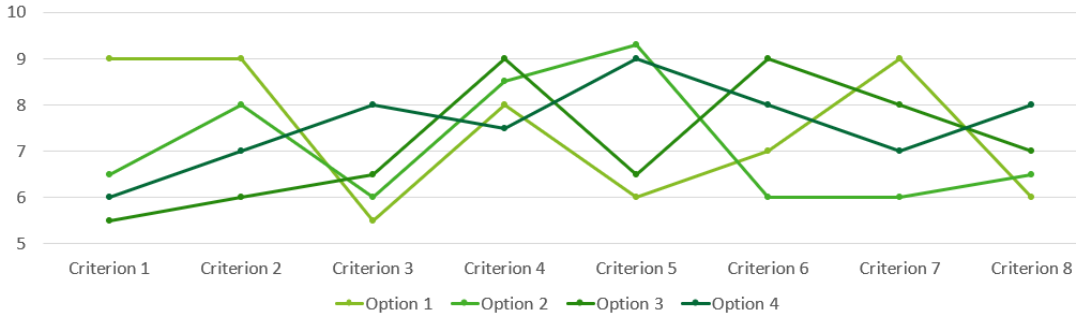

FIGURE 3.7: Example of how rating scales give a better picture, closer to the truth, compared to a ranking scale.

To reduce the loss of information, we will actually make use of a rating scale. Since we do not have anchor points that are generally needed for the rating scale, we will base this rating on the comparison with the benchmark model. In this way, there is no need for a scoring table, but we can still score all models compared to the reference level (which is also used in the well-known Simple Multi-Attribute Rating Technique (Edwards, 1977)). The benchmark model will be assigned a rating that corresponds to what is generally perceived in literature about it. That is done with the use of commonly known grading scale, i.e, from 5: insufficient and 6: sufficient up to 8: good, and 9: very good. After that, we can compare the alternative models with the benchmark model. In this way, we will not exclude important knowledge. The advantage of using a scoring method with the 'grades' mentioned above, is that we can leverage the general intuition about a rating scale from 1 to 10 that is build up throughout the use of the decimal numeral system.

With scoring in this manner, we can see in the illustrative Figure 3.7 that the perception can change of how models are actually scoring. Now, for the same data, Figure 3.7 tells a whole different and more compelling story than Figure 3.6. We can conclude that this way of scoring is of added value in drawing conclusions in this thesis. However, we must note that this way of scoring is contingent upon a subjective opinion of the evaluator. This is the main disadvantage of using a scoring scale.

Lastly, we could choose to use a weighting scheme. However, we do not make use of weightings for the criteria, which is often done, also in the aforementioned Simple Multi-Attribute Rating Technique (Edwards, 1977). That is because when using such methods, one needs to assume that all variables are fully compensatory in making the final prediction, i.e., a low score of an alternative on a certain criterion may be compensated by a high score on another criterion. This assumption is not correct in the assessment framework, and thus making use of rating scales, and scoring each criterion individually and a qualitative comparison based on the assessment framework is the best way to proceed in this thesis.

3.6 Conclusions on the assessment framework development

In this chapter, we found that industry's attention for ML is increasing. However, it shifts focus from the regulatory area towards other non regulated areas in the financial industry. The two main challenges that financial institutions identify are: 1) a lack of supervisory understanding, and 2) difficulty of explaining processes.

Regulators generally acknowledge these problems. They constructed several documents, guidelines, or regulations to ensure a proper use of ML in the financial industry. Concluding, the key topics that need to be taken into account when using ML in IRB models are interpretability, performance, and implementation challenges.

An assessment framework to evaluate the applicability of ML models for PD estimation in IRB models was constructed using the aforementioned topics. The criteria are as follows: simulatability, decomposability, algorithmic transparency, economically justifiable relationship between input and output, AUROC, AUPRC, and governance and documentation.

To make use of the assessment framework, we showed that a rating scale is preferred over a ranking scale. In this way, we are able to differentiate the key qualities and vulnerabilities of the models better when compared to each other.

Chapter 4

Model Selection and Data Preparation

Besides that the assessment framework will enable us to compare ML models with one another, the assessment framework is also a stepping stone to select potential ML algorithms to be compared. Knowing that we need to focus on interpretability, we will in this chapter select several promising ML algorithms. The selection of these algorithms is the deliverable of research question C: *"What are appropriate machine learning algorithms from the literature for application in IRB models?"*. After that, we will describe, clean, and process the data that we will use. This enables us to properly design and tune several ML models with the chosen algorithms. The goal of this Chapter 4 is to choose and tune the models and prepare the data. Eventually, these deliverables are combined with the assessment framework produced in Chapter 3. Together they enable a proper comparison that is addressed in Chapter 5: Results.

4.1 Model selection: inherently interpretable ML

In this section, we zoom in on a very specific part of ML algorithms, namely inherently interpretable algorithms. As described in Section 2.2.2, in the literature this is also known as interpretability/interpretable by design, white box models, etc. In order to come up with a short list of algorithms that we will compare, the literature is addressed.

The current most simple classification techniques comprise classification trees, and the simple logistic regression. Recall from Section 2.2.2 where we mentioned the very basic classification (and regression) tree algorithm CART. Throughout the years, many alterations have been formulated in constructing the tree, amongst others the well known *C*4.5, and *C*5.0 algorithms (Quinlan, 1993). These use rule-sets and include weighted classification errors in the training phase. They, however, have not proven to be very accurate in PD estimation. One seemingly logical step is to combine a classification tree, with logistic regression. The Logistic Model Tree (LMT) was proposed by Landwehr, Hall, and Frank, 2005. The algorithm constructs a tree, and at every leave node, a logistic regression model is fitted. The advantage of this is that features actually can interact, because based on an attribute value, each split will eventually lead to other coefficients being used in the logistic regression. This advantage, together with the fact that the model is a combination of two interpretable models, is the reason the LMT is one of the selected models to be compared.

Another relatively interpretable ML-algorithm is a generalized additive model (recall, abbreviated to GAM). Initially proposed by Hastie and Tibshirani, 1990, GAMs currently come in very many different alterations in recent literature. Firstly,

the way that the functions of single features, also called shape functions, are constructed can be different. The way that this is done does not harm the interpretability of the model itself. Therefore, the shape function construction can be rather complex. Methods used for this are ordinary classification trees, bagging, boosting (Lou, Caruana, and Gehrke, 2012), and also recently deep learning, such as NODE-GAM (Chang, Caruana, and Goldenberg, 2022). A second possible alteration on the regular GAM is that more advanced GAMs also allow for interactions between variables. Recall Formula 2.5, when interactions are added this becomes for example:

$$f(x_1, x_2, ..., x_n) = \beta_0 + g_1(x_1) + g_2(x_2) + ... + g_i(x_i, x_j) + g_n(x_n) \tag{4.1}$$

where $i \neq j$. In this example $g_i(x_i, x_j)$ is the interaction component. One can still comprehend this well, as we can plot the two-dimensional graph corresponding to the interaction using a heatmap, for example. Higher orders of interactions might allow for more predictive power, but immediately suffer from being too complex to comprehend and almost impossible to visualize while staying interpretable. The combination of 1) using a deep neural network in the shape function creation, and 2) allowing for interactions, is exactly what Yang, Zhang, and Sudjianto, 2021 propose. Their GAMI-NET is "*a disentangled feedforward network with multiple additive subnetworks; each subnetwork consists of multiple hidden layers and is designed for capturing one main effect or one pairwise interaction*". We select this specific GAMI-Net algorithm for the model comparison.

As we saw in Figure 2.7, the other ML algorithms such as SVM, bagging, boosting or deep learning are all too 'black box' to be made inherently interpretable. We therefore take a look at one very specific algorithm, which originates from the genetic programming domain, namely symbolic regression. Genetic programming based symbolic regression (GPSR) has showed to outperform several traditional black box models in regression and classification tasks (Orzechowski, Cava, and Moore, 2018). GPSR is a heuristic to efficiently search the infinite possibilities of models that can be build for a classification problem. This algorithm is part of the evolutionary programming domain (cf. Section 2.2.1), and specifically uses all kinds of mathematical expressions to let input parameters interact, to construct a single output value. In each iteration, these mathematical expressions are altered, in order to see what relationships between variables have the most predictive power. In Section 4.3 we will further elaborate on the inner-workings of this model.

Having chosen three models; LMT, GAMI-Net, and GPSR, we are able to make a comparison. In order to effectively contrast the models to the current situation, we will also fit a logistic regression model. That is the benchmark model of this thesis, and it will also be referenced as such throughout the comparison. The three models discussed above are constructed in `Python` with the use of the libraries `lineartree`, `gaminet`, and `feyn`. To conceptually illustrate how these models relate to the traditional ML models shown in Figure 2.7, we added them in Figure 4.1. Note that this is no ground truth, but its purpose is to aid the reader in the conceptual placement of the interpretable ML algorithms on the interpretability-accuracy trade-off graph.

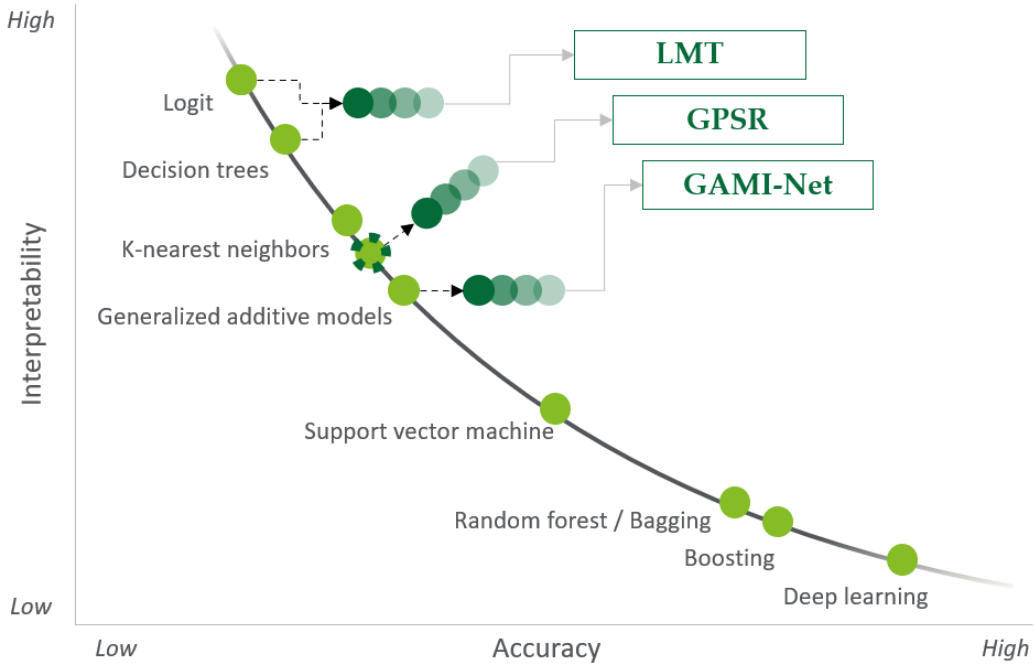

FIGURE 4.1: A conceptual visualization of the interpretability and accuracy trade-off for the algorithms to be assessed.

4.2 Data selection and preparation

4.2.1 Peer-to-peer lending

As mentioned in the methodology in Section 1.3, we will be using a publicly available dataset from LendingClub. LendingClub is a US based peer-to-peer lending platform. A characteristic belonging to peer-to-peer lending is quick and easy transfers of money from (private) investors to lenders. To efficiently connect investors with lenders, the platform makes use of a transparent online marketplace. This data is collected and published by LendingClub, and is a useful source for many researchers. Although this thesis is focused on ML in IRB models, we will use this LendingClub dataset since it is a rich dataset, containing a lot of data records and features and is specifically focused on credit risk. The use of 'real' bank data is not a possibility, as this data contains a lot of personal information, which makes it not allowed to publish this thesis when describing and reporting on that data.

4.2.2 Data description

We use the most recent version of the dataset[1], covering a period from 2007 Q1 up until 2020 Q3. Shortly after that date, LendingClub stopped its peer-to-peer lending branch. To the best of our knowledge, this had no influence on the available loans and investors. The raw dataset contains 2.93 million data records, covering 141 separate features. Each data record represents a loan, having certain characteristics, and most importantly a loan status. The `loan_status` feature is a ten-level categorical variable, of which only two are absorbing states: 'Fully Paid', and 'Charged Off'. These are the loan statuses that are of interest, which respectively translate to 'No Default', and 'Default'. The default rate in this dataset, i.e., the proportion of defaults, is 19.5%. We note that this is a relatively balanced dataset compared with

[1]Retrieved from *Kaggle* on 12th of April, 2022

TABLE 4.1: Overview of loan statuses in the raw dataset

Status	Count	Percentage
Fully Paid	1,497,783	51.20%
Current	1,031,016	35.24%
Charged Off	362,548	12.39%
Late (31-120 days)	16,154	0.55%
In Grace Period	10,028	0.34%
Late (16-30 days)	2,719	0.09%
Issued	2,062	0.07%
Does not meet the credit policy. Status: Fully Paid	1,988	0.07%
Does not meet the credit policy. Status: Charged Off	761	0.03%
Default	433	0.01%
Sum	**2,925,492**	**100.00%**

bank's data on different portfolio. The relatively high fraction of defaults in this LendingClub dataset is another characteristic that belongs to peer-to-peer lending.

4.2.3 Data pre-processing

As described above, there are only two absorbing states[2]. That means that the rest of all the loans, are not yet in a final state and therefore cannot be used in a probability of default estimation model. Current outstanding loans are therefore discarded, see Table 4.1. These include Current loans, Late and In Grace loans, Issued loans, and also Defaulted loans. This latter loan status does not correspond to a 'default' as we defined it in this thesis and is therefore not in an absorbing state and for that reason removed.

After removing these rows, we take a closer look at all the features in the dataset. Since the dataset contains information on the loan that was not known at the point of origination, we need to exclude these features to prevent so-called data leakage. That means that we only use information based on characteristics of the loan and applicant that were known when the applicant applied for the loan. This makes this research and the PD model also applicable for loan acceptance and pricing, besides the use of within IRB models. Next to those post-origination features, multiple features are also removed since they do not contain any valuable information, e.g., the URL to the loan on LendingClub's platform.

Data imputation

After removing the rows and columns that we cannot use, we are left with a dataset with dimensions 1,860,331 x 48. Many columns and rows still contain missing values, which can cause trouble when implementing the models. Many of the features show exactly the same number of missing values. This indicates that a few data records have missing data throughout the feature space, and are therefor not useful in our analysis. After deleting those rows, the column `mths_since_recent_inq` is left with approximately 10% missing values. Disregarding all these data records could have an impact on the validity of the results. Therefore, we choose to impute the data. Imputation can be done in multiple different ways, and there is no 'right' or 'wrong' method for this. To name a few examples, one can impute missing values by the mean, median, or mode. The advantage of this is that it is an easy and fast way to do. However, it has also the disadvantages that a bias is introduced and

[2]Although the two status that include 'Does not meet the credit policy.' are also absorbing states, these are excluded, as they should not have been approved as loans given the current credit policy.

that the distribution of the data can be largely affected. This influences the learning capabilities of the model, and possibly the predictive power of the feature. Another way of dealing with these missing values is by treating them as a separate category. In this way, one can investigate whether `having a missing value on feature 'x'` is a good predictor in the model. Since we are looking at a numerical value (months since most recent inquiry), and the previously mentioned imputation methods do not suffice for this feature, we use another option that is a bit more sophisticated. We fit a linear model to impute the missing data. In this way, we can, based on a few highly positively or negatively correlated features, predict what the value could be. We selected the three features with the highest and three features with the lowest correlations with the feature of concern. By selecting multiple features for the linear model, we make sure that the imputation will not hurt the models' interpretability too much regarding collinearity, see the next section below. Eventually, a linear model was fitted to impute the 10% missing values of this feature. Additionally, to prevent extremely high or negative values, the minimum and maximum values in the data were capped to remain the same as the original training set. A visualization of mean imputation and the chosen linear model imputation is shown in Figure 4.2.

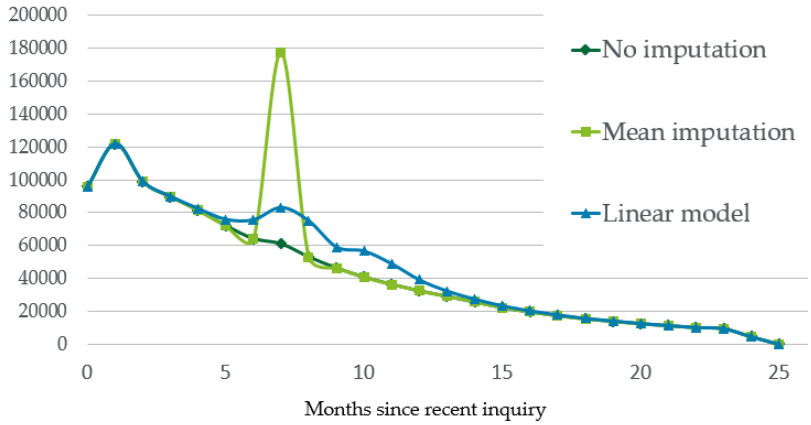

FIGURE 4.2: Data imputation techniques with their influence on the distribution of the data of the feature `mths_since_recent_inq`.

Correlation and collinearity

When trying to predict a certain outcome, ideally, we want predictors with a high correlation with the target variable. However, when two predictor variables have a high correlation with each other, the model's performance and interpretability can suffer from it. That is because the two predictors will move together (i.e., it is not possible to change one predictor and leaving the other remain the same), and therefore we cannot observe the individual linear relationship of the one predictor variable with the target variable. In short, it makes the effects of $X1$ on Y difficult to differentiate from the effects of $X2$ on Y (Goyal, 2021).

To deal with this, we first produced a correlation matrix, enabling investigation of the highly correlated features. See Appendix A, Figure A.1 for the plotted correlation matrix. The features with a correlation larger than 0.8 were selected, see Table A.1, and subsequently investigated with the use of the Variance Inflation Factor (VIF). The VIF is a commonly used expression to detect (multi)collinearity, and

it is calculated by formula 4.2 below:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{4.2}$$

where R^2 is the well-known statistical metric that measures how much of the variance from the target variable is explained by the predictor variable, which is the same as the square of the correlation. To calculate this fraction, one uses a single predictor variable, and make a regression model with the other predictor variables to predict the single other predictor variable. In that way, we can find the R_i^2. When this is high, we observe that the other predictor variables capture the variance of the single variable that we tried to estimate, feature i. This results in a high VIF. A high VIF means that the feature can be disposed without losing significant performance, and simultaneously improving the interpretability and computation time of the model. As a rule of thumb, a $VIF > 5$ is a significant concern, and is therefore disregarded in the model (Menard and SAGE., 2002). This is an iterative approach, outcomes of the VIF in each iteration, and the deletion of the features are given in Table A.2 in Appendix A. In total, six features are disregarded in this step.

Handling outliers

One other pre-processing step that is often performed in ML projects is the treatment of outliers. Outliers should be treated carefully, as they can occur for several reasons, such as measurement errors, human errors, or just extreme observations. Specifically, one needs to find the right balance in distinguishing between data errors/noise, and genuine patterns. However, when not dealt with appropriately, outliers can harm the learning ability of an algorithm significantly (James et al., 2021). A common approach is deletion of data records, or capping the outliers. In this thesis, we detect outliers by using the interquartile range (IQR), which is the difference between the 75th percentile (Q3), and the 25th percentile (Q1) of an attribute. For all numerical values, we plotted a box plot to visualize the Median, the 25th and 75th percentile. Additionally, the box plot's whiskers show $Q1 - 1.5 \cdot IQR$ and $Q3 + 1.5 \cdot IQR$.

With the use of this visualization, and with the use of the fraction of observations that are outliers, we are able to effectively process these outliers. The percentage of outliers is given in Table A.3. Specifically for numerical features with a large number of outliers, deletion, or capping might not be the best option. To illustrate this, Figure 4.3 presents two box plots, one without outliers, and one with many outliers. The red dots, which overlap each other, represent 17% of the data. Additionally, it is observable that Q1, the median, and Q3 overlap each other. From the description of `pub_rec` it is defined as follows "*Number of derogatory public records*". When over 17% of the data has a derogatory public record, it is not sensible to cap or delete these records. We do however think that this can be an explanatory variable and therefore choose to binarize this feature. That means that the data will be binary formatted, where 0 belongs to "*having no derogatory public record*" and 1 means "*having a or multiple derogatory public records*". This is done for multiple features that showed the same behavior. For the remaining features, where the outliers made up less than 10%, the outliers were capped to be maximum $Q3$ or $Q1$, plus or minus $1.5 \cdot IQR$.

We note that although the treatment of outliers used above is a common approach in the literature, it is not the only right perspective on outliers. Some also argue that outliers are actually the data records that are the rare extreme cases from

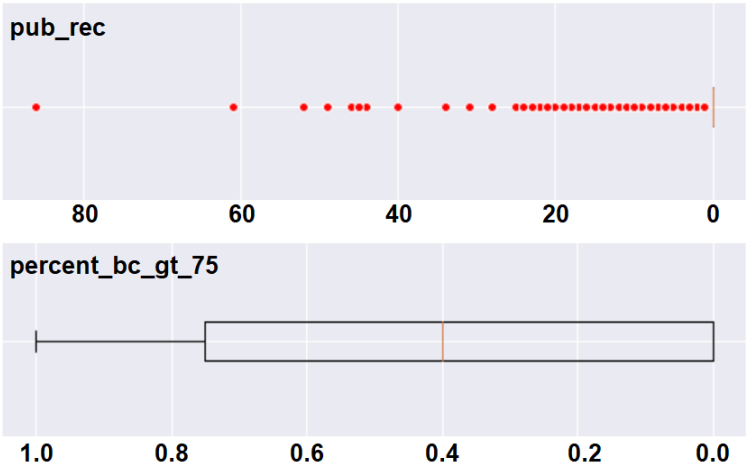

FIGURE 4.3: Box plot of two numerical features in the dataset, one showing no data points outside the whiskers and one with many outliers.

which can be learned most. To follow this perspective, one needs to thoroughly do an extensive research on data outliers, possibly interpreting each outlier individually, to distinguish data errors from genuine patterns. Due to time limitations, this is not possible to do in this thesis, and the research on outliers is furthermore outside the scope of this thesis.

Scaling

One of the final pre-processing steps is making use of a scalar function to scale the data in a way that makes it easier for the algorithm to learn from. This normalization step is often necessary as the range of values of all numerical predictor variables vary widely. Algorithms that make use of the Euclidean distance or gradient descent in the learning process will observe a much faster convergence when all features consist of the same data ranges (Ioffe and Szegedy, 2015).

Several scalars are often used such as mean normalization, a robust scalar, or min-max scaling. The latter has the advantage that it will scale all features to be in the range $[0, 1]$. The min-max scalar makes use of the following formula:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{4.3}$$

where x_i is the instance that is being transformed, and $\min(x)$ and $\max(x)$ are respectively the lowest value of the specific feature, and the highest value. Within this research, the min-max scalar is chosen to be used.

Train-test split

To train our models on the data, and eventually also evaluate the performance of the models, we make use of a train-test split. We split the raw data in an early stage of the data pre-processing, to ensure that we do not infer information from the test set into the processing of the data. In this way, we apply the exact same data processing steps on the test set, only based on information from the train set. To illustrate this, for data imputation, a linear model was trained on the non-missing values of the train set, and subsequently used to predict the missing values of the train set, as

well as the test set. Another example, in which data leakage was prevented, is in the outlier detection and processing step. We only learn the IQR of the train set, and based on that place a cap on the train and test set.

After all the aforementioned pre-processing steps are performed, we formed a training set with dimensions 1,201,802 x 37 and a test set of 515,058 x 37. All 37 features with their corresponding descriptions are added to Appendix A, Table A.5.

4.3 Model tuning

In Section 4.1 we presented the three models. In this section, we will zoom in on these models. We elaborate on the design of each algorithm, the learning technique used, and model tuning. Additionally, depending on the model, some extra data processing operations might be needed in order to make the data applicable for the specific algorithm, and to retrieve the best performance possible. This is discussed below for each model that we investigate: the logistic model tree, the GAMI-Net, and the genetic programming based symbolic regression.

4.3.1 LMT

The LMT builds a tree and fits logistic regression models at each leave nodes. In Section 2.2.1 the normal classification tree was discussed. For all different kind of tree algorithms, the goal is to split the data in such a way that the child nodes are homogeneous in terms of target variable. In contrast to splitting based on the purity of the child nodes, the LMT needs to split the data in such a way that the at the child nodes, the target data is linearly separable. That means that for each extra split that is considered, a logistic regression is fitted on the child nodes. Then these two child notes must perform better than the previous parent node did individually. The algorithm makes use of a cross-entropy loss function in determining whether a split makes it into the final model, by calculating the weighted sum of the training losses over the child and parent nodes. The cross-entropy loss function calculates the log loss in the following way:

$$Loss = -(y \log(p) + (1 - y) \log(1 - p)) \tag{4.4}$$

where y is the target variable, and p the predicted probability. Figure 4.4 shows how this log loss value behaves under different predictions and target variables. One learns from Figure 4.4 that the loss increases more when it further diverges from the actual target.

In order to train the model, one extra pre-processing step was performed. Categorical features cannot be used in the model directly, and are therefore encoded. This is done with a regular dummy encoding, which creates an extra column, for each level in a categorical variable. All these dummy variables are thus binary variables.

Aiming to get the best performance possible and simultaneously minimize the risk of overfitting, we make use of five-fold cross-validation in finding the best hyperparameters. In order to restrict the complexity of the model, we choose to limit the depth of the tree to be at most three. That means at most three consecutive splits, resulting in a maximum of $2^3 = 8$ leaf nodes. On top of that, we want a minimum depth of one, as otherwise the LMT would not make an actual tree, and would behave as a regular logistic regression model. Subsequently, we perform

the cross-validation on the hyperparameter of the minimum fraction of samples at a leaf node together with the regularization strength parameter for the so-called l1-regularization for fitting each logistic regression model in a leaf node. Also, a few parameters were fixed. We instructed the algorithm to use a balanced sample weighting. Especially when there is not an equal distribution of classes, it helps the algorithm by putting more emphasis on the minority class in learning. The way this is done, is by putting more weight on the loss function for the minority class. It is called balanced weighting, as the weights become the inverse of the fraction the class takes up in the whole dataset, e.g., an 80:20 sample size distribution will be respectively weighted with 1.25, and 5. We deliberately choose not to use any sampling techniques (over- or undersampling). These are techniques that artificially increase or decrease the number of observations of respectively the minority and majority class. We do not need to use this artificial resampling technique since the dataset is large, and the class labels show a low imbalance in terms of proportion `default:non-default`. On top of that, several studies and online competitions on the same dataset showed that these techniques did not yield any better performance. To conclude, the results of the model tuning and training will be given in the next chapter, Chapter 5, on the model assessment and results.

4.3.2 GAMI-Net

The GAMI-Net is a GAM which allows for interactions in a two-dimensional space. The shape functions of the GAM are constructed with the use of fully connected deep neural network. Although these networks are not interpretable, the final model will be so, on the contrary. That is because the neural networks are means to compute a shape function. After that, the neural networks can be disregarded, and only the shape functions, which are very intuitive and interpretable by design, will make it into the final model.

Recall from Section 2.2.1 how a neural network is build. In contrast to Figure 2.6, that showed multiple input features for a neural network, the neural networks that are used in the GAMI-Net are only using one input feature for the so-called main effects, and two features for the interactions. In Figure 4.5 the GAMI-Net architecture is depicted. The construction of the GAMI-Net follows three stages: 1) fitting main effects, 2) fitting interactions, and 3) fine-tuning. The left-hand side of Figure 4.5 shows the first stage, three neural networks are fitted, all constructed with only

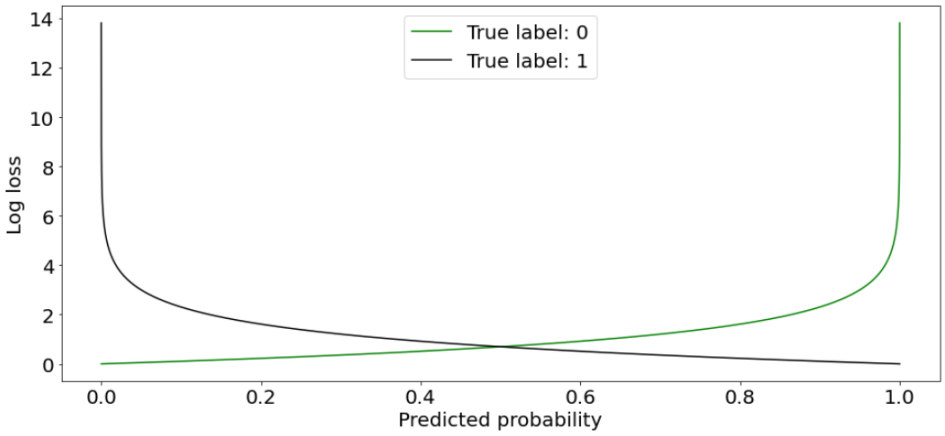

FIGURE 4.4: Log loss values for a binary target variable for different predicted probabilities.

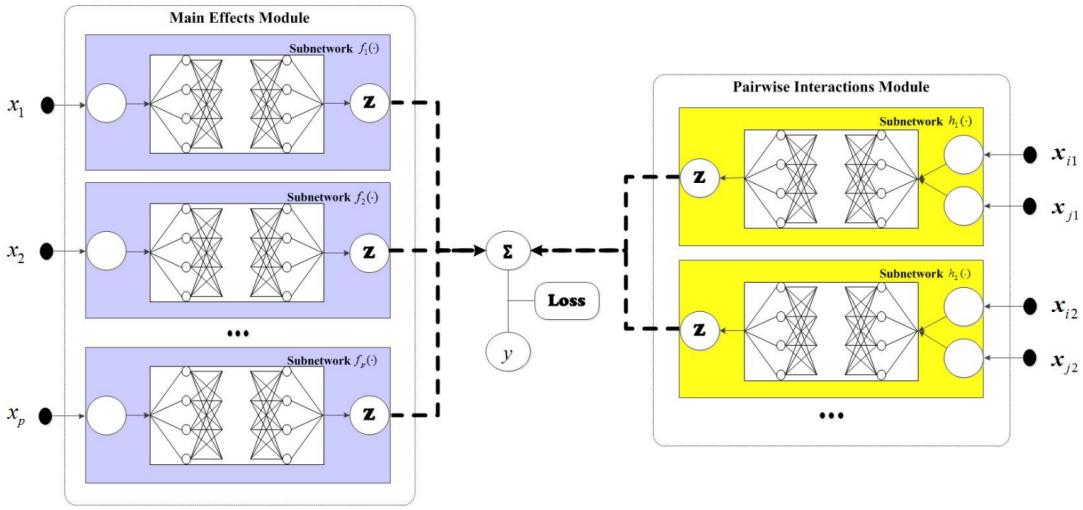

FIGURE 4.5: The architecture of the GAMI-Net (Yang, Zhang, and Sudjianto, 2021).

one input feature x_i and several hidden layers. The right-hand side comprises the second stage, in which two input features are used to interact with each other. The advantage of these two stages is that they produce respectively one-dimensional curves, and two-dimensional surfaces. The curves (shape functions) and surfaces (heatmaps) will ultimately be linearly combined with a bias and inserted in a link function, just like logistic regression, which eventually outputs the PD.

To tune the model, different hyperparameters can be set. In order to give the learning of the shape functions and heatmaps enough 'freedom', we follow the developer's default settings, namely having 5 ReLu-hidden layers with 40 nodes per layer. Due to computational time limitations, this architectural design of the subnetworks is not tuned or validated. Parameters that will be cross-validated are the learning rate and the boolean argument "`heredity`". Yang, Zhang, and Sudjianto, 2021 included this argument to impose extra restrictions to the use of interactions in the model. This heredity concept is also often used in literature covering variable selection techniques (Bien, Taylor, and Tibshirani, 2013; Choi, Li, and Zhu, 2010). When the heredity argument is set to be true, only interactions can be used in the final model when at least one of the interacted variables is also included in the model as a single main effect. However, releasing this constraint can possibly significantly improve the performance of the model when two features only have a strong predictive power when combined. Again, as a result of computational time limitations, we cannot cross validate these parameters by running the whole model multiple times. As an approximation, we use for the cross validation less epochs. See Figure 4.6 for the outcomes of the five-fold cross validation for these parameters.

First of all, we note that a lower learning rate performs better, where there is a marginal performance improvement from 0.01 to 0.001. Remarkably, we note that when the heredity argument is set to false, the model tends to under-perform in contrast to setting it to true. Only for the largest learning rate this differs. That is, because the purpose of using heredity constraint is to help reduce the search space of interactions, and simultaneously make the model structurally more interpretable (Yang, Zhang, and Sudjianto, 2021). Having a higher learning rate, will make the model search the solution space faster, and in that way possibly find interactions that are not allowed with the heredity constraint in place. Naturally, we expect that when running for a longer number of epochs, the results of the models without the

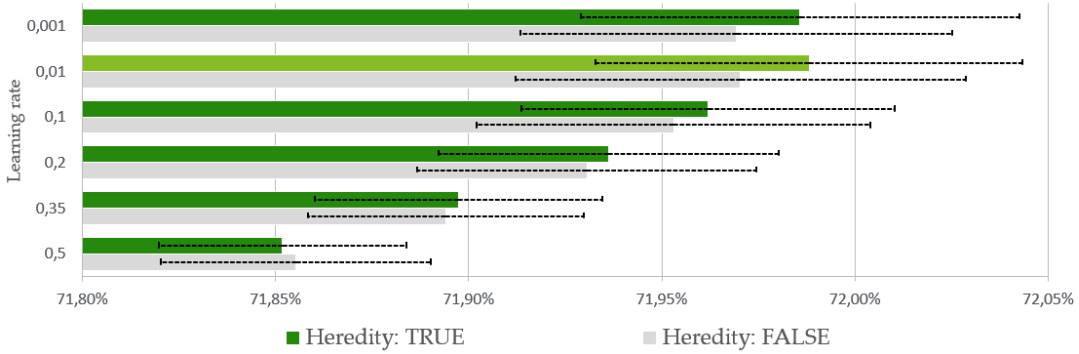

FIGURE 4.6: AUROC scores of five-fold cross validation of the learning rate and heredity parameter in the GAMI-Net. The whiskers show the standard deviation of the cross validated AUROC scores.

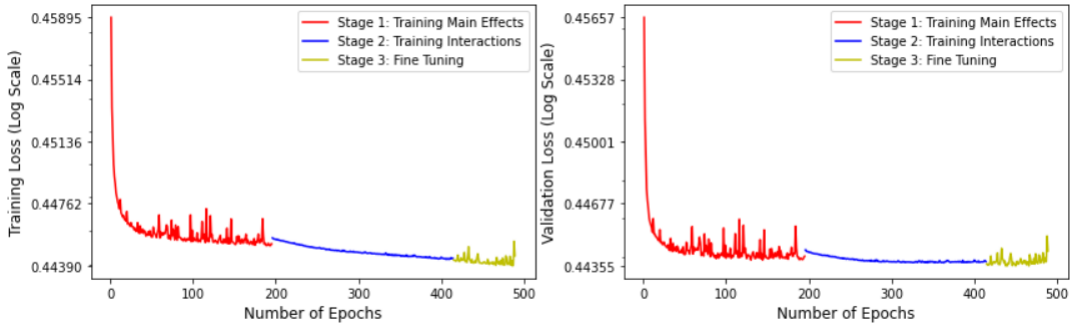

FIGURE 4.7: Learning process of the GAMI-Net in three stages. On the left the training loss is depicted, and on the right the 20% hold-out validation loss is depicted.

heredity constraint will increase and possibly bypass the ones with the constraint. However, we observe that the inclusion of the constraint does not have significant influence on the predictive performance, and therefore choose to include the constraint to increase the interpretability.

The final model will be run with a learning rate of 0.01 and with the heredity constraint. The model makes use of the Adam optimizer to develop the neural network sub-models[3]. It will run for 1000 epochs for each stage, with a stopping criterion if the validation loss does not decrease significantly for a length of 50 epochs. Lastly, the model is given the same balanced weights for the classes as in the LMT.

In Figure 4.7 the learning process in terms of the loss function is visualized for the three aforementioned stages. Notice that there is not a significant decrease of the loss function in stage 2, the interactions training. This gives us an indication of the features in the data being exploratory in itself, and no interactions have outstanding exploratory power.

4.3.3 GPSR

In Section 4.1 we briefly addressed GPSR and explained that it is a combination of symbolic regression and genetic programming. First, symbolic regression is a way

[3]In this paragraph, extremely specific learning settings were mentioned that need a very detailed explanation in order to understand the technical details of it. These explanations are not added to this thesis and are treated outside the scope of this thesis.

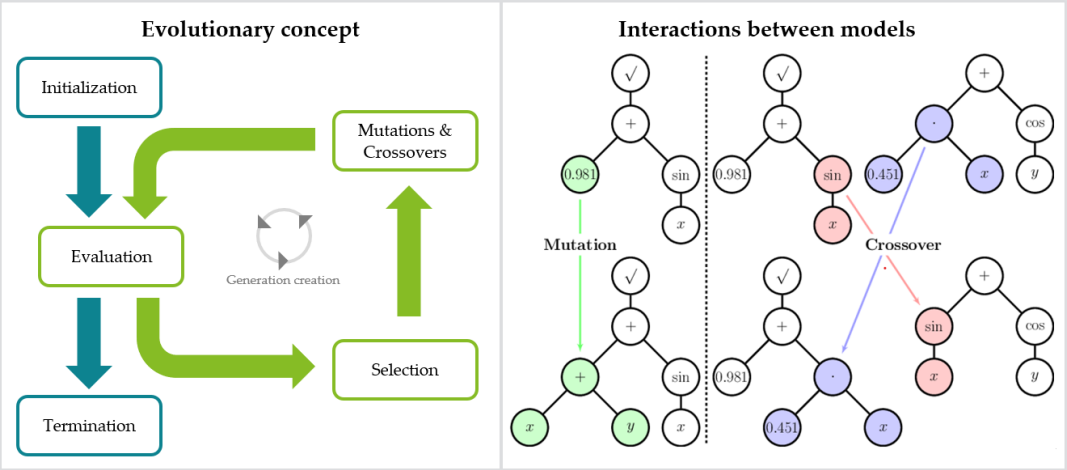

FIGURE 4.8: Evolutionary concept and interactions between models in genetic programming based symbolic regression.

to fit a model by making use of mathematical operations. Instead of fitting only additive components, such as in a linear and logistic regression, we now allow for far more interactions between features. In symbolic regression, addition, multiplication, linear transformation, exponents, logarithms, and so on can be used. One can imagine that when having a variety of mathematical operations, that can be applied to many different features, and also combined in many ways, an infinite solution space is created.

That is where genetic programming's power comes in helpful. Genetic programming is a heuristic to efficiently search an infinite solution space and makes use of the concept of evolution. In Figure 4.8 on the left-hand side, the concept of evolution in model estimation is depicted. The start is by generating multiple random initial models, the initialization. After that, the models are all evaluated, using the loss function that we described earlier. Then the process of generating a new generation takes place. All models are assigned a weight, having a higher weight when a model's loss is lower. This weight resembles the survival chance, just like in evolution, corresponding to the fitness of the individual model. Given these weights, models are randomly selected for the next generation. Before evaluating these models again, two different interactions *between* those models take place. These are depicted on the right-hand side of Figure 4.8. Either some of the operations within a symbolic regression model are altered, which is called a mutation, or two models interchange parts of its model, which is called a crossover. After that, the process of evaluation, selection, and mutations and crossovers starts over again, until no further improvement is found. In this way, the algorithm tries to find near-optimal solutions for the given solution space.

This algorithm does not need any further pre-processing, as it can handle categorical variables, although it also uses dummy encoding within the algorithm. There are not many hyperparameters to tune in this algorithm, as the heuristic is already well-defined without much inference from a developer. What we can control for, is the level of complexity of the final model. We can do this automatically with the use of the build in Bayesian Information Criteria (BIC). It makes use of a likelihood function and adds a penalty term for the number of parameters in the model, reducing the risk of overfitting. Using BIC as criterion will yield the model that should generalize the best to out-of-sample test data (Larsen, n.d.). To familiarize us with

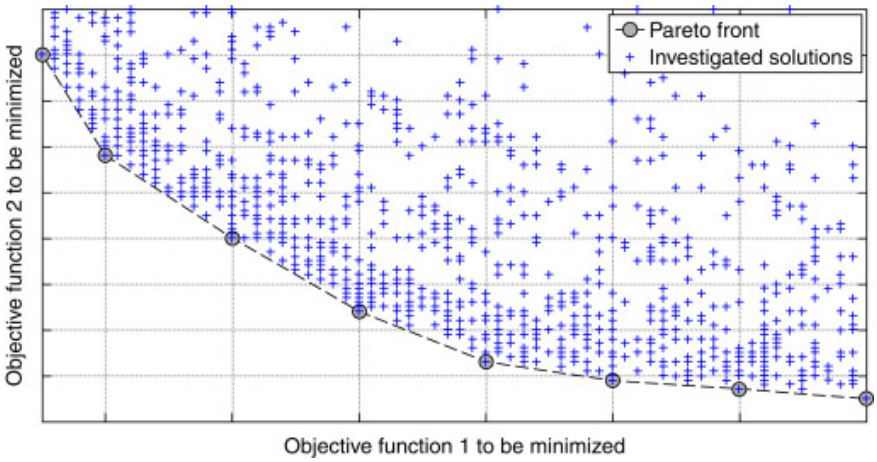

FIGURE 4.9: An illustrative example of how a Pareto front is created with possible solutions.

the trade-off on complexity and performance, we can plot an approximation of the Pareto front. The Pareto front consists of models that cannot be outperformed by any other model in the solution space. For example, given a model i with complexity X_i, and performance Y_i, there is no model j with $X_j > X_i$ when keeping $Y_j <= Yi$. This also holds vice versa (Neumann, 2012). See Figure 4.9 how the Pareto front is created for different possible solutions. In the figure, two functions are being minimized, therefore the most optimal point is in the lower left corner.

We ran the GPSR multiple times, with the use of balanced weighting, and allowing for different types of complexities. The complexity is measured in terms of 'edges', i.e., how many variables are linked to each other in the model. After each run of 10 epochs, the average loss is calculated, which is, just like complexity, an objective function to be minimized. Figure 4.10 shows the results of these experiments as a Pareto front plot. In contrast to the example given in Figure 4.9, this Pareto front is a stepwise function. That is caused by the fact that the complexity of the GPSR is measured in integers, instead of a fully continuous variable.

Looking at the Pareto front of Figure 4.10, we wee that the Pareto front is created by some outliers. We therefore take into account the concentration of the investigated models around the levels of complexity. From that, we can conclude that for the construction of the final model in Chapter 5, we can restrict the learning process to have at most 6 edges, since performance does not increase after that. We will still use the BIC, but restricting complexity reduces the solution space, and therefore improves the computation time.

4.4 Conclusions on model selection and data preparation

To conclude, in Chapter 4 of this thesis, we selected three interpretable ML algorithms that have a high potential to be applicable for the use in IRB models. The three models are the LMT, GAMI-Net, and GPSR. After that, the data of a peer-to-peer loan website was pre-processed. This was done until the point where every algorithm desires its own specific formatting. All models were individually tuned on the training set by making use of cross-validation. We choose to restrict the LMT to a maximum depth of 3 to allow only interpretable models. Also, a minimum depth was used, since otherwise a regular logit model was fitted. Cross-validation

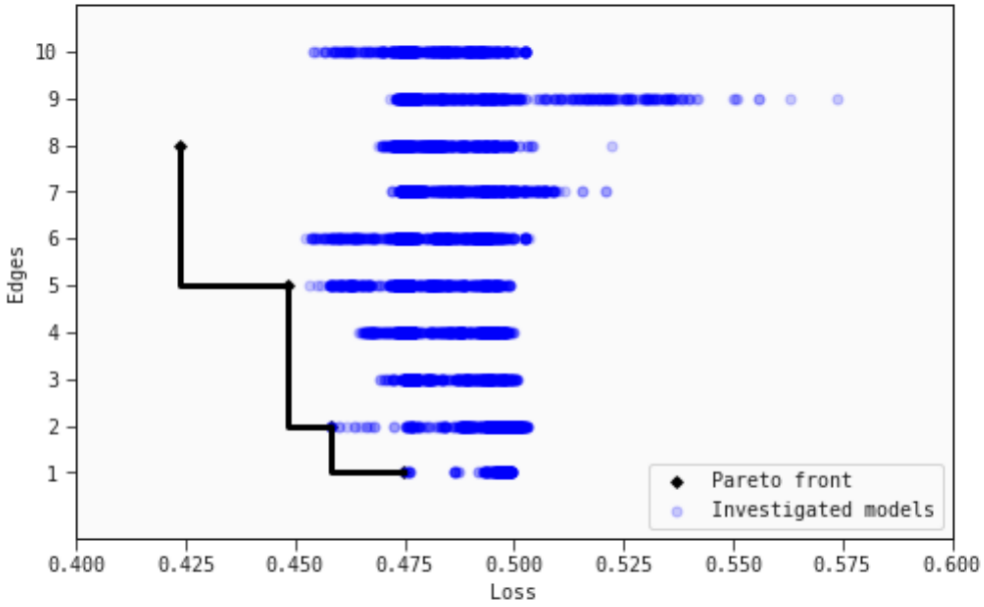

FIGURE 4.10: Pareto front plot of complexity versus loss in the GPSR model. Only a sample of 50% of the investigated models is plotted to illustrate the concept of Pareto optima.

of the GAMI-Net shows that a learning rate of 0.01 yielded the best results. We include the heredity constraint to increase the interpretability of the model. Lastly, experiments with the GPSR shows that the GPSR learning process can be restricted to a maximum complexity of six edges, i.e., the number of interactions and features used.

In the next chapter, Chapter 5, the outputs of the ML models will be shown. For a good comparison, also the benchmark model, a logistic regression model will be included in the analysis of the different models.[4]

[4]The hyperparameters of the benchmark model are also tuned with five-fold cross validation. The results are not described in the thesis, as they are not part of the core of the modelling and analysis.

Chapter 5

Model Assessment and Results

In this chapter, the final models are presented. In Section 5.1 we highlight the most important outputs of each model. These also include explanatory visualizations. These help us to assess the models in Section 5.2. In that section, we assess the models' applicability in IRB models with the use of a ranking scale. A motivation of the scoring on each criterion is also provided. The final deliverable of this chapter is the scoring of our three chosen ML models and the benchmark logistic regression model based on the assessment framework. The content of this chapter gives an answer to research question D: *"How do the different algorithms score based on the assessment framework, and what are their key distinctions?"* Afterwards, conclusions can be drawn in Chapter 6.

5.1 Model output description

For every selected model, we present the outputs of the models in visualizations and complementary explanations. This full description of the outputs is necessary to be able to compare and contrast them in the next section.

5.1.1 LMT

An advantageous property of a tree-method is that the tree can be visualized, see Figure 5.1. Looking at the figure, one sees that in total, six splits are performed and seven logistic regression models are fitted. The node id's depicted in Figure 5.1 are

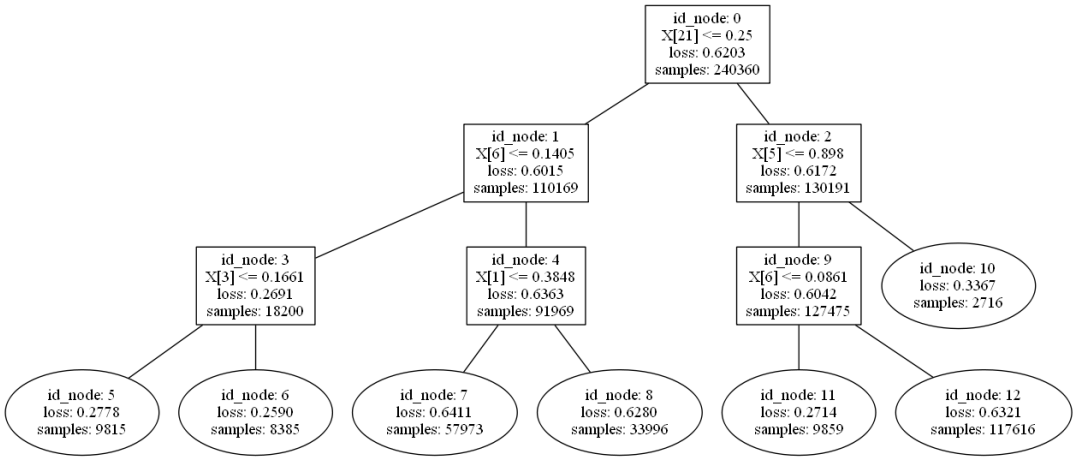

FIGURE 5.1: Visualization of the trained LMT with at each leaf node the number of samples and average loss.

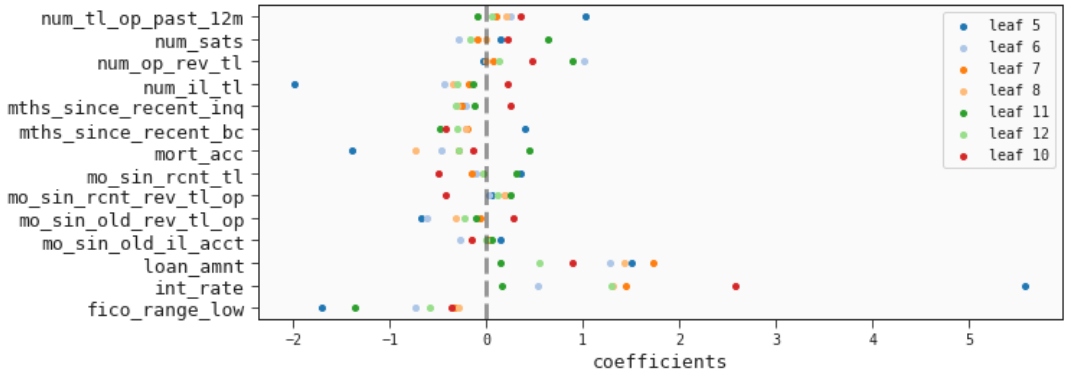

FIGURE 5.2: LMT visualization for a selection of features: coefficients of the fitted logistic models at the leaf nodes of the model tree.

also used in Figure 5.2. That figure shows a snippet of all the coefficients for each feature in the seven different leaf node models. The full image, with coefficients for all 73 (dummy-)features, can be found in Appendix B in Figure B.1. Looking closely at some of the differences in the coefficients in Figure 5.2, and the corresponding splits in Figure 5.1, it is possible to get some insights. For example, in general, the variance across similar coefficients at different leaf nodes is an indication that the LMT has an advantage over a regular logistic regression, since it has identified significant differences on how 'important' a feature is based on a split in the data. However, one needs to be careful in drawing conclusions on this, since all coefficients only combined form a logistic regression model at a leaf node. This implies, that a larger coefficient at one leaf node does not immediately imply that it is more important compared to a second node, since the former could have many (and large) negative coefficients that balance out the effect of the large positive coefficient. This behavior is also observable in for example leaf node 5 (blue), as across all the features, it tends to have higher positive, or higher negative values compared to the rest.

5.1.2 GAMI-Net

The GAMI-Net is developed using the hyperparameters and the results of the cross-validation of the previous chapter. Figure 5.3 shows the size of the model. The left side of the figure plots the number of main effects and the resulting loss value. The higher the number of main effects, the more one can reduce the loss function, although with a chance on overfitting. The same holds for the right-hand side of the figure, in which the number of interactions is plotted against the loss values. The optimal number of main effects and interaction are denoted by the red star. However, taking the sparsity of the model into account, the red dot is more favorable as it allows for a relatively small increase in loss, by significantly reducing the number of features used.

To give an example of how the GAMI-Net is build up, we can easily visualize the additive components of the GAMI-Net. As an illustration, the three most significant main effects, and three most significant interaction effects are plotted in Figure 5.4. For the overview of all main effects and interactions that are included in the model, see Appendix B, Figure B.2 and Figure B.3. The advantage of using shape functions instead of linear functions (that a logistic regression makes use of) immediately becomes clear. For example, in the upper left shape function of `int_rate`, the interest rate. There is no linear relationship between the target variable and interest rate: for

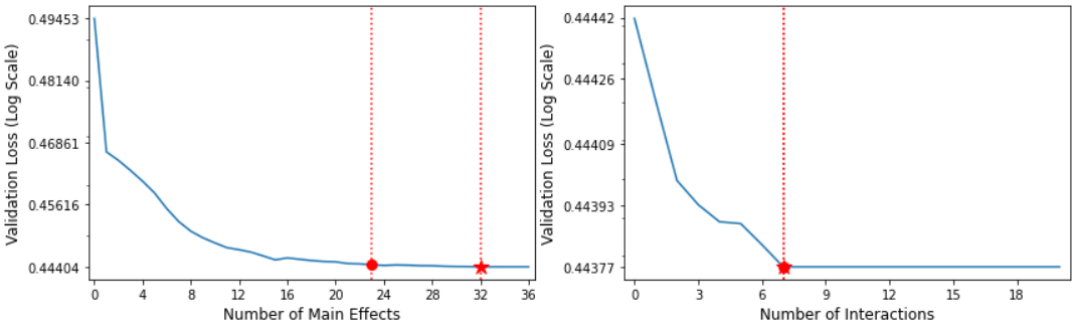

FIGURE 5.3: Validation loss values plotted against the number of features in the GAMI-Net. Left: individual features. Right: number of interactions in the model.

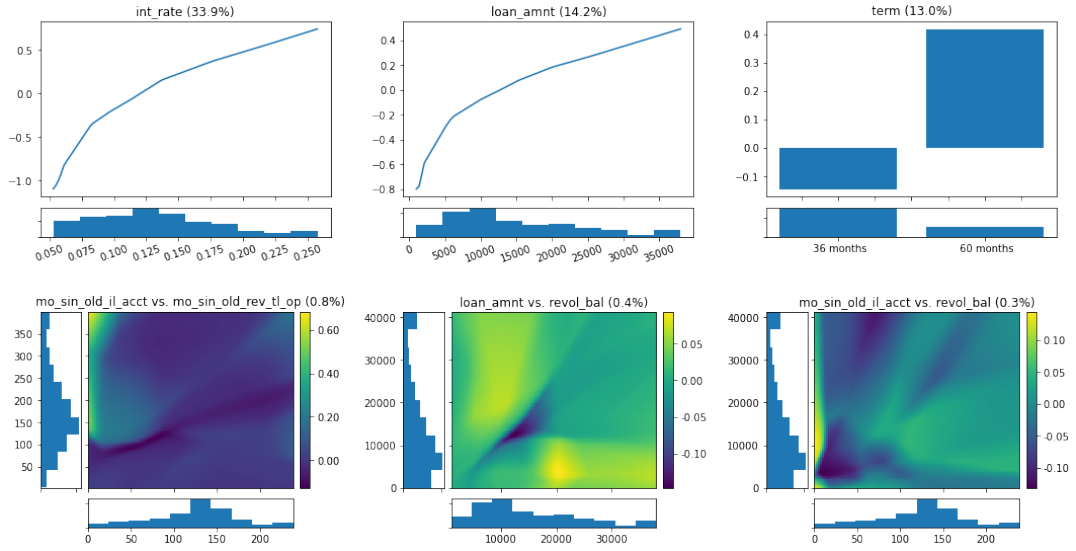

FIGURE 5.4: Plots of the most important main effects (upper row) and interactions (second row) of the GAMI-Net.

instances below 10%, the effect of a low interest rate in predicting the PD becomes increasingly negative. Above 10%, the effect is decelerating positive. This is known as a concave relationship. Apparently, this non-linear relationship is a better predictor than the linear relationship that in a logistic regression can only be used.

On top of each plot in Figure 5.4, the variable or variables are depicted and a corresponding percentage is printed. The percentage represents the amount of variance explained by that specific component and can be interpreted as the importance of the variable(s). These are scaled such that the sum of all percentages is equal to 100%. These global importance values are also depicted in Figure 5.5 for all features and interactions in the model. The insight from Figure 4.7, about the relatively low impact of adding interactions, can also be recognized in Figure 5.5. The seven interaction terms that made it into the final model are all at the tail of the descending importance graph.

Important to note, is that the GAMI-Net itself consists of only the components mentioned above: 19 main effects (shape functions), and 7 interactions (heatmaps). Therefore, the whole learning process, consisting of many neural networks, is not part of the model anymore. The outputs of each neural network is an interpretable visualization. Each visualization is one component for the final model. That makes

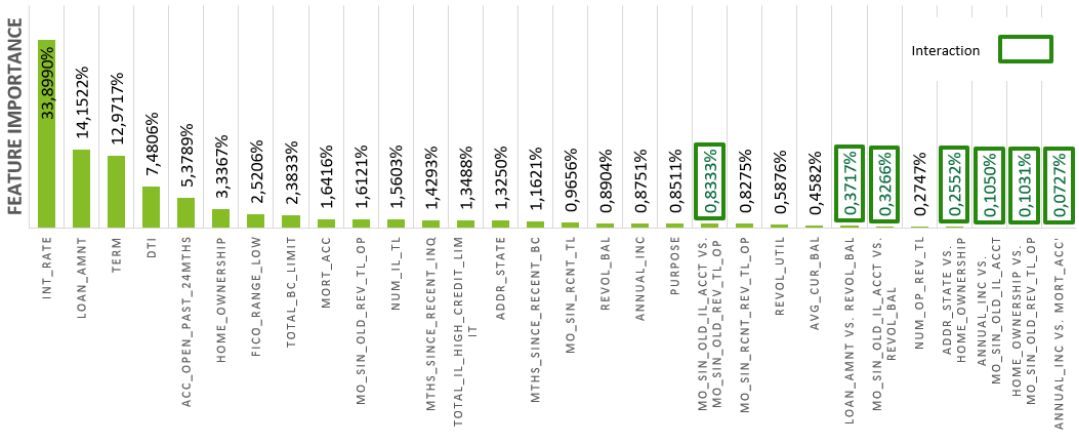

FIGURE 5.5: GAMI-Net's global feature importance, depicted for the main effects and interactions of features.

the GAMI-Net still a very interpretable ML model, while using an advanced ML model, a deep neural network.

Finally, the GAMI-Net has one specific advantage, just like other GAMs, namely a final check on the shape functions can help to identify data issues that were not resolved during the pre-processing. This can be found by unexpected movements in the shape functions. For example, when mean imputation is used for missing values. However, this is only of a real advantage when the data is retrieved from an external source, and the researcher does not know what was done with the data before the acquirement of it, see for example the research of Christensen et al., 2022. Although a bank retrieves and holds its own data about their loans and clients, it is a well conceivable situation that a data team processes the data for the modelling team. Given that there are always communication flaws between departments, one extra safety check is definitely desirable in that situation.

5.1.3 GPSR

We run the GPSR for 50 epochs, evaluating over 40000 models in the learning process. As we found in the previous chapter, we restrict the model to have at most 6 edges. The final model is a simple formula, a combination of features with mathematical operations. The output is given in Figure 5.6. This simplified visualization does not include biases and weights. When including this, the formula that is put into the logistic function becomes:

$$1.8 - 0.71 \cdot e^{2(\text{term}-0.28)^2+1.2(0.19\cdot\text{avg\_cur\_bal}-0.83\cdot\text{int\_rate}+1)^2} \tag{5.1}$$

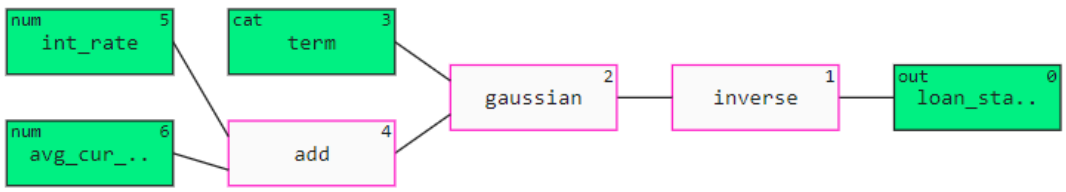

FIGURE 5.6: Visualization of the final GPSR model, excluding bias terms and weights.

where `term`, `avg_cur_bal` and `int_rate` are the features used. For the categorical variable `term`, each category gets assigned a value that the algorithm learned. In this case, for example, a loan term of 36 months will be replaced with -0.1195 and a loan of 60 months gets assigned 0.1858. Additionally, a bias term for both values is included, namely -0.2795.

To illustrate the simplicity of this specific model, we give an example. Given a `term` of 36 months, an `avg_cur_bal` of 0.740951 and an `int_rate` of 0.077337, the exponent term becomes:

$$2 \cdot (-0.28 - 0.12 - 0.28)^2 + 1.2(0.19 \cdot 0.74 - 0.83 \cdot 0.077 + 1)^2 = 2.27$$

Substituting that into Formula 5.1 yields:

$$1.8 - 0.71 \cdot e^{2.27} = -5.12$$

Finally, this is put into the logistic function with a bias weight, resulting in a PD of 0.1187. The mathematical operations are comparable to a regular logistic regression, where only now one needs to use exponents and multiplications. The advantage on the other hand is that the model gets its best performance already by making use of only three features in the feature space with a size of 36 features.

5.2 Model assessment

For each criterion, an explanation will be given for the ranking and scores of the different models. Eventually, this section concludes with an overview visualization, such as in Figure 3.7.

5.2.1 Simulatability

Starting with the simulatability criterion, which was the 'first level of transparency', referring to a model's ability to be simulated by a human. The simulatability focuses on the simplicity but also refers to the total size of the model, i.e., is someone able to take the input data, simulate the model, and produce the same prediction in a reasonable amount of time.

All three models, and the benchmark logistic regression model, use the logistic function to restrict predictions between 0 and 1. Therefore, we focus on the components that are substituted within this function for each model. First, the logistic regression model makes use of only some additive components. Each component is the multiplication of a single coefficient and the value of the feature. The regulated benchmark model has only 16 of the 73 (dummy-)features. The GAMI-Net makes also use of additive components. However, for the 23 variables, and 7 interactions in the model, every graph needs to be investigated to get the right values to be summed. The GPSR scores higher in terms of simulatability than the GAMI-Net, as the GPSR only uses three features. The relatively harder computations (i.e., multiplications and exponents) of the GPSR are outweighed by the many variables of the GAMI-Net, and the burdensome way of getting all the right values out of the graph. The LMT scores last. This stems from that the model uses many features and one needs to perform a large number of operations, i.e., first following the tree, and then handling a logistic regression with many features, as we showed in the previous section and in Appendix B in Figure B.1.

5.2.2 Decomposability

Decomposability is the second level of transparency, as stated in Section 3.3. This denotes to what extent a model can be broken down into several components: inputs, parameters, and the computations. On top of that, the components itself are ideally intelligible.

Again, since all models make use of the logistic function to restrict predictions between 0 and 1, we can neglect that part in the assessment. The most simple model, the logistic regression model, is the easiest to decompose. All components are additive, and all components in itself are only linear transformations of the feature value, using a single coefficient. The GAMI-Net follows the logistic regression closely. It makes use of additive components, which are easily detachable, and they are each also easy to investigate individually. Although a linear transformation of features is a bit easier to interpret, a shape function or heatmap does not compromise much of its interpretability. The computation to come up with the additive component is only a bit heavier (either producing the mathematical outcome of the function, or reading the graph). After the GAMI-Net, LMT follows in terms of decomposability. Although larger in terms of size, the GAMI-Net is very well decomposable. After that, the LMT is most decomposable; the tree is a structure that is very intelligible in itself. That is, since every split can be rephrased in a plain text description (Lipton, 2018). Then, following the tree towards a leaf node, a regular logistic regression model is found which is also well decomposable. The size of the LMT makes it less well to be decomposed compared to the GAMI-Net. At last, the GPSR is a rather lengthy formula, with interactions between the variables that are not easily understandable. One needs to put some effort in and have some strong basis in mathematics to grasp the effect of one variable on the outcome of the final prediction.

5.2.3 Algorithmic transparency

The third level of transparency covers the algorithmic transparency. Therefore, the focus is on the learning process that is used.

Starting off with the regular logistic regression, this is the most well-known and basic algorithm used for learning. It makes use of relatively old statistical knowledge, such as maximizing a likelihood function. Therefore, the learning of the logistic regression can be properly conceived. After that, the LMT is most transparent as an algorithm. With the tree, it splits the data in such a way, that it tries to make buckets of data that are well separable with a logistic regression model. It makes use of an iterative approach to fit the tree, and simultaneously fit the coefficients of the logit models in the leaf nodes. That leaves GAMI-Net and GPSR left. Both make use of complicated algorithms to come up with the final model, which on its turn is relatively transparent. The GAMI-Net makes use of multiple deep neural network, with a batch gradient descent. The output of each neural network is a shape function, containing only one variable, or two for structured interactions. The main advantage of the GAMI-Net's neural networks, is that the neural networks do not involve more than one (or two for interactions) input parameters. In that respect, one can observe what the learning process is doing. The GPSR on the contrary is an algorithm that is less traceable. Since it makes use of a heuristic to search an infinite solution space, it also has a dominant stochastic property. The stochastic learning process, in which each iteration survival of a model depends partly on chance, makes it less tractable

and less reproducible and therefore opaque. For the opaqueness of the GPSR algorithm and the verifiable outputs of the deep neural networks of the GAMI-Net, we note that the GAMI-Net scores higher than GPSR in algorithmic transparency.

5.2.4 Economically justifiable relationships

Coming to the relationship between inputs and output in the assessment framework, we note that an economically justifiable relationship between those parameters must be ensured. As this is a prerequisite, all models do adhere to this. However, the extent to which the relationships are justifiable and the extent to whether they are representing the common perception of the relationship between inputs and output differs across the models.

Often, a economically justifiable relationship restricts the relationship between input and output to be monotonically increasing or decreasing. A logistic regression satisfies this constraint, one should only take care that the direction of the monotonic relationship is correct. However, there are several examples in which a linear relationship is not the most favorable relationship. For example, income is a good indication whether you can pay your bills. However, a linear relationship is probably not the most justifiable or in line with the actual relationship, since an income increase from $\$3,000$ to $\$4,000$ is more important than an increase from $\$10,000$ to $\$11,000$. Another example in the GAMI-Net is `int_rate`. The variable is a good proxy for the risk that is involved with a loan. However, a linear relationship is not suitable, which the GAMI-Net also learned, see Figure 5.4 in the previous section. Additionally, the GAMI-Net algorithm allows for monotonicity constraints within the learning process, while still keeping the flexibility of learning non-linear relationships. With these properties, the GAMI-Net scores better than the logistic regression model. With regard to the GPSR, we note that it can only have monotonic relationships with the output variable, since it consists of one formula with mathematical operations only allowing monotonicity for the feature values inserted (which are in $[0, 1]$). Generally, it therefore comes up with more justifiable relationships than the linear relationships of logistic regression, but does not allow for as much freedom in construction of the relationship as the GAMI-Net does. Lastly, the LMT makes use of linear relationships, just like the logistic regression model. However, due to the splits in the tree, the relationships between inputs and output can change when there is a change in the feature value. As an illustration, one can think of a salary increase. Before the salary increase, the person fell in leaf node i, where the variable `salary` had a large positive coefficient. After the salary increase, the LMT could place the data record in another leaf node where the coefficient is negative, since other positive coefficients level out the negative coefficient of `salary`. Although this is not something that is likely to happen, this is still an inherent disadvantage of the LMT. Therefore, it scores last.

5.2.5 Performance

Now, changing the point of attention to the output of the model; performance is extremely important for the financial institutions, as this is the area where they take most advantage of. With regard to performance, two performance criteria are used, the AUROC and the AUPRC. For the three selected interpretable ML models and the benchmark model we used exactly the same train and test set in order to compare

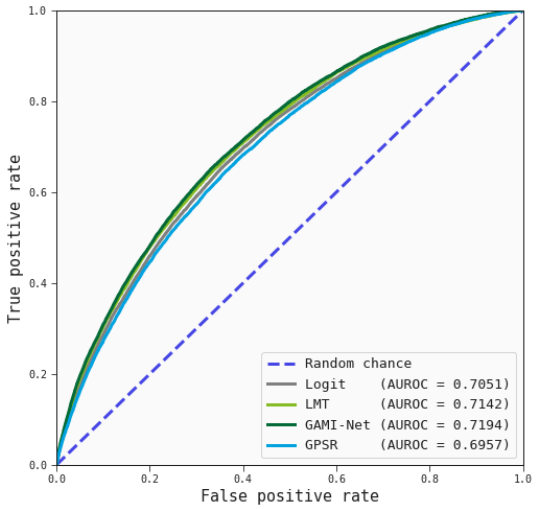

FIGURE 5.7: The ROC curves of the chosen ML models

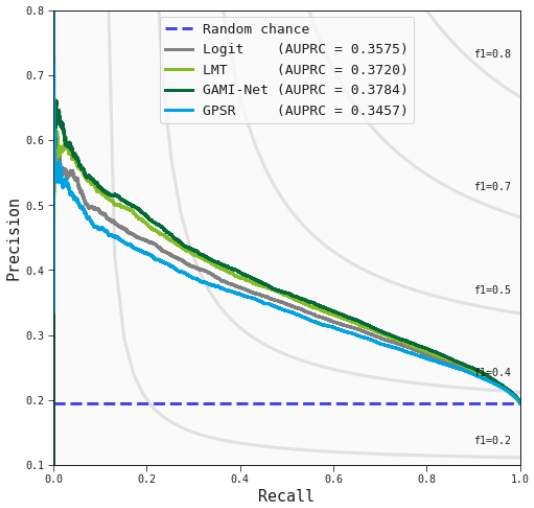

FIGURE 5.8: The PR curves of the chosen ML models

TABLE 5.1: AUC scores for the different ML models.

	Logit	**LMT**	**GAMI-Net**	**GPSR**
AUROC	0.7051	0.7142	0.7194	0.6957
difference with Logit	-	+1.29%	+2.03%	-1.33%
AUPRC	0.3575	0.372	0.3784	0.3457
difference with Logit	-	+4.06%	+5.85%	-3.30%

them effectively[1]. Having four prediction arrays with PDs, and one array of true values containing whether a client defaulted or not gives the ability to plot four ROC curves, and four PR curves. Therefore, see respectively Figure 5.7 and Figure 5.8.

As we see, for both plots, the different lines run past each other. The ROC curves plot in Figure 5.7 shows the same pattern as the PR curves plot from Figure 5.8, namely the lowest line belongs to the GPSR, followed by the benchmark model, next LMT, and shortly followed by the GAMI-Net. Interesting to note are the differences in PR curves. For example around a precision of 0.5, one can choose to use a GAMI-Net instead of the benchmark logistic regression, and double the recall from 0.09 to 0.18, while maintaining a precision of 0.5.

The AUC values for all curves in both plots are given in Table 5.1. For the ROC curve, the differences are a bit smaller, and also harder to differentiate in the plot. However, the PR curves and the corresponding AUPRCs show a larger variance between them. Obviously, the order of the lines from bottom to top described above is the reverse order in terms of performance in the AUC metrics. The final scores in the assessment overview in the next section are calculated using a spread of 4 points. In that way, the differences in performance scores can be divided over the spread of points, with the same distribution as in Table 5.1.

[1]Below we report only the scores that were retrieved by predicting on the test set, that has carefully been excluded in the data processing to prevent data leakage. The test scores correspond to the training scores, therefore there is no sign of under- or overfitting.

5.2.6 Governance and documentation

The last component of the assessment framework focuses on the use and implementation of the models. Related concepts are the governance, documentation, manageability, and responsibility.

The logistic regression is the status quo in the current IRB models. The property of being a traditional, thoroughly researched method, and being the tradition for years, makes it a model with which many are familiar. In Section 3.2.3 we mentioned CRR article 189: "*all aspects of the rating and estimation process of the PD shall be approved by the management body and senior management. They shall have a detailed comprehension of its management reports and a good understanding of the rating systems and operations*". Especially this article thwarts the implementation of new ML models. However, for logistic regression this is not an issue, since in every regular statistics class, logistic regression is treated. One could argue that the LMT would then also score high on this last criterion on the assessment framework, as it makes use of merely logistic regressions. However, there are multiple reasons why the governance of the model is not that easy. First of all, the size makes it hard to get a thorough understanding of all components. Therefore, multiple persons should be involved. On top of that, as was described when discussing the economically justifiable relationships, several branches of the tree can contradict each other in weights given to the same feature. All this makes it certainly less easy to implement and to use on a daily basis. In between the LMT and the logistic regression model, we score the GAMI-Net. The GAMI-Net itself is tractable, consisting of shape functions and interactions. The basic knowledge of fitting log odds from a logistic regression can also be applied in the GAMI-Net. The interactions do not hurt the manageability. The main disadvantage in terms of manageability is the use of deep neural networks to construct the components. Although, these are no real black boxes (because they only have one input variable), it is still necessary to have sufficient knowledge on how to construct these neural nets, and tune them. On last place, the GPSR is placed. This is due to the relatively innovative use of evolutionary programming. The algorithm does not allow for human interference. On top of that, the stochastic process can cause finding completely different models after refitting the model on some new data. The instability caused by the stochastic property makes it less usable for IRB models.

5.3 Overview of the results

In this section, we present the results of our analysis that is made in this chapter. We use the assessment framework, that is developed in Chapter 3, to assess an ML model's applicability for the use in IRB models. With a scoring scale, we can assess the models individually, but it also allows for a thorough comparison based on the criteria. In that way, the framework can be applied by companies wanting to make a decision on which ML model to use in IRB modelling, or to assess which vulnerabilities and strength a model has compared to a benchmark model.

In Figure 5.9, the final overview of the assessment done in the previous subsections is given. We want to stress that the use of this assessment framework with the scoring scale includes the opinion of the one using the assessment framework. We emphasize that this comparison is not a ground truth, but is a thoroughly motivated comparison which also includes the knowledge and experiences retrieved in constructing and tuning all models.

Figure 5.9 presents the three areas of interest: interpretability, performance, and implementation. The scores are depicted in the large circles, and connected to the

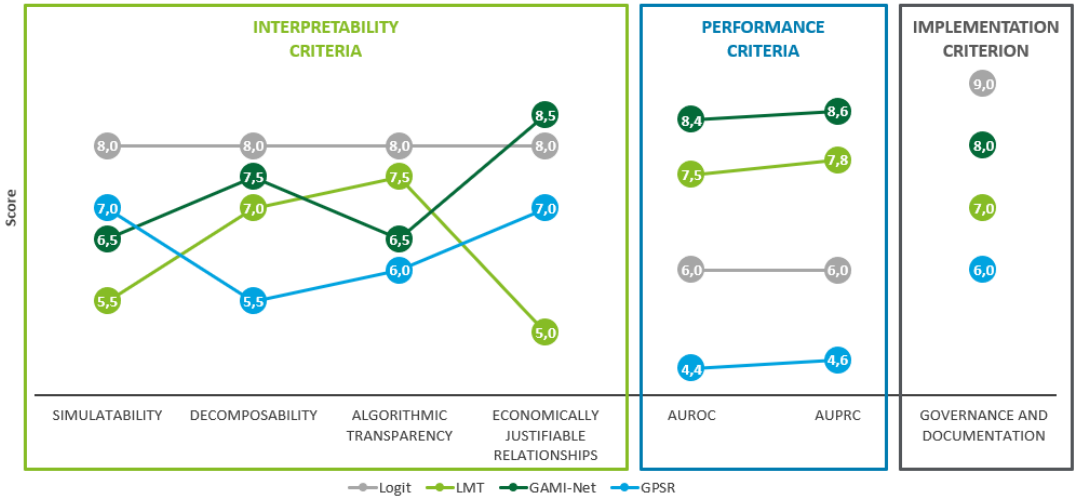

FIGURE 5.9: Evaluation of the applicability of ML models in IRB models. Three selected interpretable ML models are plotted, together with the benchmark, a logit model.

other scores of the same model within the same area of interest for an easier comprehension. We observe that the logistic regression is the best model in terms of interpretability and ease of implementation. On the other hand, it scores relatively low on performance, where we observe that the LMT and GAMI-Net outperform the benchmark. The GPSR is clearly scoring lower than the logistic regression on all three areas of interest, and is therefore not the best candidate to be adopted in IRB models.

On top of the overview of all criteria in the figure mentioned above, we present an overview of averages of each criteria category in Figure 5.10 on the next page. This aggregated score is of extra support in drawing conclusions in the next chapter. We note that averaging scores implies equal weighting, therefore we only make this overview supplementary and only average scores per category.

On an aggregated level, the logistic regression scores an 9.0 in terms of interpretability, whereas the runner-up, the GAMI-Net scores a 7.3. The rest scores below a 6.5 and is therewith less competitive with regard to the logistic regression. However, in terms of performance, the logistic regression scores a 6.0, whereas the LMT and GAMI-Net now clearly outperform the benchmark. The GPSR is in terms of performance of no use in IRB modelling, as it has a score below 5. At last, the implementation criterion is topped by the logistic regression. Followed by the GPSR and LMT, we see roughly the same ranking as on the interpretability criteria.

Figures 5.9 and 5.10 are the final deliverables of the last sub research question of this thesis. Having answered all sub research questions, we are able to draw conclusions and develop recommendations in the final chapter, Chapter 6: Conclusions & Discussion.

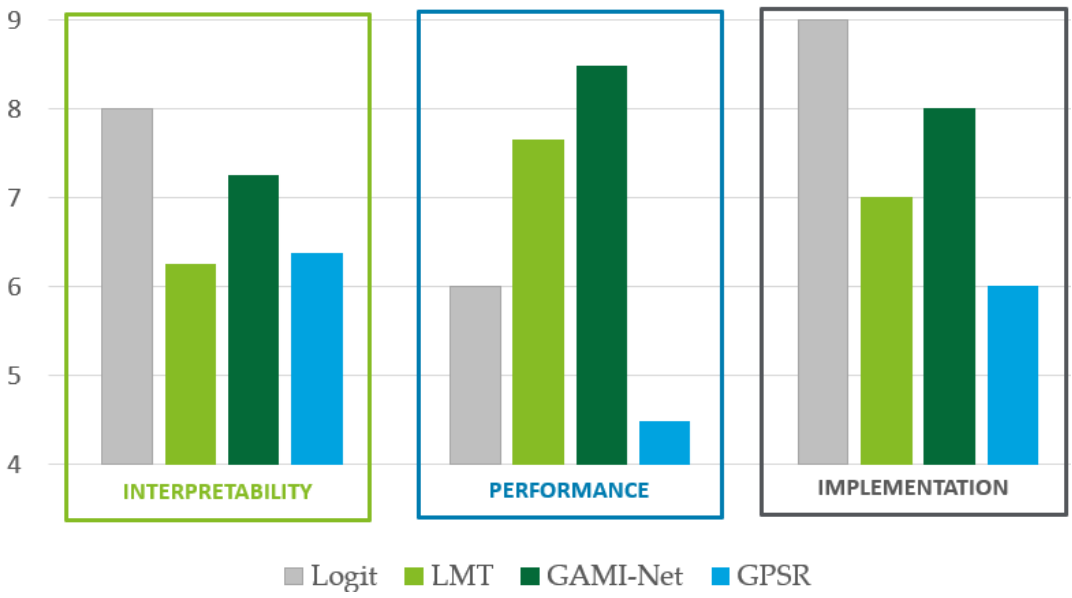

FIGURE 5.10: Final result: averages of each criteria category per model.

Chapter 6

Conclusions & Discussion

6.1 Conclusions

The research subject of this thesis was motivated by the discussion paper published by the EBA, 2021. Advanced machine learning models are perceived as black boxes, which raises concerns at the European Banking Authority. They identified three key challenges that restrict the adoption of machine learning in IRB models: 1) interpreting results of ML models, 2) ensuring adequate understanding by management, and 3) justifying the results to supervisors.

To resolve these issues, this thesis shifts the focus away from 'black-box' machine learning models, and focuses on inherently interpretable ML algorithms. Ideally, the use of inherently interpretable ML algorithms result in transparent models, which subsequently help to ensure an adequate understanding by management and helps in justifying results to supervisors. Therefore, we formulated the following main research question:

> **Which interpretable machine learning algorithms are applicable for the use in IRB models and how do they differ from each other?**

Beginning with the determined scope, we limited this research to the probability of default (PD) component within credit risk modelling, particularly in Internal Ratings Based (IRB) models. Under the IRB approach, banks have more freedom in modelling their own risk, including the modelling of the PD component. Regarding the current state of ML adoption in IRB models, we found that especially within the area where there is an approval needed from supervisors, ML adoption is lacking behind. In fact, for the use of ML in financial institutions, a negative trend in the regulatory area is observed, and a positive trend in many other financial areas. We found that a lot of regulations apply to banks using IRB models. Most of the regulations are opaque, consisting of guidelines that often mostly overlap with each other. The main area of interest from the industry as well as the supervisor is interpretability. However, the industry's attention is also focused on increasing performance with respect to the status quo in PD estimation, a logistic regression.

To effectively compare ML models, there was a need to develop a method to do so. We developed an assessment framework to compare the ML algorithms in terms of applicability for the use within IRB models. The assessment framework focuses on the two main areas of interest: interpretability and performance of an ML model. For the measurement of interpretability of ML in IRB models, we found that one should look at: simplicity, decomposability, algorithmic transparency, and justifiable economic relationships between inputs and output. For the performance assessment, at least the industry standard of the area under the ROC curve should be used. On top of that, the area under the precision-recall curve is of added value

when comparing the performance between models, since these can yield more insights. We also included a third area of interest in the assessment framework, which is the implementation and use of the model within an organization. This last area of interest is specifically important in order to make the assessment framework relevant for the use in practice.

In a literature review, we found three state-of-the-art interpretable ML algorithms to be promising in terms of applicability for the use within IRB models: the Logistic Model Tree (LMT), the Generalized Additive Models with Structured Interactions (GAMI-Net), and the Genetic Programming based Symbolic Regression (GPSR). The LMT is an algorithm that constructs a tree, and adds a logistic regression at each leave node. The GAMI-Net makes use of a deep neural network to construct the components of a generalized additive model. It also allows for interactions of at most two variables for each additive component. Lastly, the GPSR is an algorithm build on the principles of evolution. It heuristically searches the infinite space of all kinds of mathematical operations combined with all available features. After we selected these three models, we consecutively developed, tuned, and evaluated the three aforementioned models and the benchmark logistic regression. We made use of an open source peer-to-peer credit lending dataset from the lending platform Lendingclub.

We used a rating scale to score the models. A rating scale was preferred over a ranking scale, as we could make the differences between models more explicit. This gave us a good overview of the strength and weaknesses of the ML algorithms and how they relate to each other. With this overview, we can answer the final research question, and draw conclusions on this research:

General conclusions on results

- The investigated models all have their strength and weaknesses. In general, we conclude that in terms of interpretability, the logistic regression is not matched by others. Based on our scoring, the logistic regression performs better than the alternatives on all but one interpretability criteria, which results in a score of 8, whereas the runner-up, the GAMI-Net scores a 7.3 and the rest below 6.5. The unparalleled interpretability of a logistic regression is one of the reasons why it is still the status quo. However, by sacrificing a bit of interpretability, alternative, still inherently interpretable ML models show to outperform the logistic regression in terms of predictive power (for a critical reflection on this, see Section 6.2.1). The GAMI-Net shows a significant 2.03% increase in terms of AUROC and the LMT outperforms the logit with 1.29%. The GPSR does not outperform the logistic model.

- From the compared algorithms, the only real challenger of the logistic regression is found to be the GAMI-Net when considering the scores on interpretability, performance, and implementation. This has two important reasons. First, the GAMI-Net is an all-rounder. Looking at the assessment framework and the scores, the GAMI-Net does not have major weak spots, and performs above average on all three areas of interest. Secondly, the GAMI-Net is a good alternative since it least sacrifices interpretability of the researched methods. Specifically because most of the hesitant external parties' documents on ML (e.g., guidelines, regulations, and other statements of supervisors) center around the diminishing degree of interpretability. That makes the GAMI-Net therefore also the most interesting from a regulator's point of view.

Key insights of differences in models

- The LMT shows to outperform the logistic regression in terms of performance. Although there is a positive difference in performance, the LMT sacrifices interpretability when compared to the benchmark. The most prominent weakness of the LMT is that the relationship of an input to the output is not always economically justifiable. The relativity and dependency of coefficients amongst each other at each leaf node makes it extremely challenging to compare leaf node models. This results in the significant reduction in terms of interpretability. The choice to use the LMT in IRB thus depends on the preferences of a company, and how much interpretability it wants to sacrifice for an increase in performance.
- The GAMI-Net is the most favorable option as alternative for the logistic regression. The final model does not suffer from components that are not interpretable. The only debatable flaw is that the GAMI-Net algorithm makes use of deep neural network. In order to be able to confidently use these, the developer of the model needs to have sufficient knowledge on this topic. However, the opaque construction algorithm of the model, does not punish the interpretability of the model itself. In terms of performance, the GAMI-Net is the best of the compared models, and it therefore is a model that has a good balance on the classic interpretability-accuracy trade-off. It is a real challenger of the logistic regression, as it only sacrifices a bit of interpretability, while significantly increasing performance.
- We conclude that the GPSR is not of use within the modelling of the PD component in IRB models. It scores lower on all criteria that it was assessed on compared to the benchmark model. The relatively low performance, and its low degree of manageability, makes it not a viable alternative. However, we must note that the algorithm is innovative in such a way that it searches an infinite solution space efficiently. On top of that, the model does have a remarkably high performance for only using three features.

6.2 Discussion

In the discussion, there is room to reflect on the results, we will elaborate a bit more on the 'why' of the results. On top of that, we discuss the reliability and validity of the research. Another aspect that is treated in the discussion is the contribution to the theory and the relevance in practice of this research. Finally, we will conclude the section, and therewith this thesis, with some recommendations for further research.

6.2.1 Reflection on results

In general, the results make sense in terms of expectations. Below, we will reflect and elaborate on the results and conclusions drawn in this thesis, specifically for two of the three investigated models:

- Regarding the Logistic Model Tree, we found that it did outperform the logistic regression model by sacrificing interpretability significantly. We restricted the learning algorithm to find at least one split in the data in order to fit a real LMT, otherwise it could become a regular logistic regression. During this tuning, it was found that when this restriction was lifted, with the use of BIC

regularization, the LMT favored no split at all in the data. That means, that no tree would grow, and the LMT would be actually the same as the logistic regression. This is a fair indication of that in this specific dataset, the LMT is not favorable.

We can further extend this reasoning, with the fact that the researched dataset only contains peer-to-peer loans, which is a fairly specific type of loan, and can thus be seen as a large portfolio of a bank. Correspondingly, banks actually also have all kinds of portfolios, on which they each fit another logistic regression. Looking at the broader perspective of a bank, we see that the bank actually already has a large LMT heuristically designed on its own, with splits resembling the path that leads each data record to the right portfolio / leaf node. We can thus argue, that a further split in the data could be seen as a redundant step, which sacrifices too much interpretability to improve performance.

- Although the GPSR is found to be not a good alternative, the model has a remarkably high score by only using very few features. That makes the algorithm a very strong algorithm, however, not for the use of estimating the PD in the IRB models. This does not mean that it is of no use for banks. As long as the logit model is the status quo, and a lot of feature engineering is done to pull the maximum performance out of such a model, the GPSR can be of help. The GPSR can construct artificial new features (by combining some features in the existing feature space), and sequentially those can be used in a logistic regression. The efficient way of searching an infinite solution space in terms of models to come up with near-optimal outcome is a promising concept.

Furthermore, it should be noted that although the increase in terms of performance is already convincing, this can only become more in the (near) future. With more and more unstructured data being collected, these investigated models can possibly thrive even more. Because, currently, we are using risk drivers that banks have been using for decades. However, new data features of the loan and borrower that are not yet used currently, can possibly be leveraged with ML. New features could be retrieved from, for example unstructured data, such as social media use and other online behavior, or transaction data, which are extremely many data records on the level of transactions of each client. With the use of these features, one can imagine that the investigated models in this thesis can widen the gap in terms of performance with respect to the logit model. Especially, when data becomes available that together, as interactions, have large predictive power. After all, that is where ML operates best; efficiently finding patterns in data, that traditional methods do not capture.

Lastly, we note that in this thesis we wanted to create a level playing field for all ML approaches and the logistic regression by doing all pre-processing steps on all features that are necessary to let the ML models learn from data. However, we did not use any feature engineering technique, nor did we tune specific features. Since logistic regression has been used for a long time in PD modelling, many feature specific transformations have been learned *by humans* in the process of working with it. For example, the `salary` of a person can be transformed by making use of a logarithm, to better resemble the perceived relationship between the `salary` and the PD. In that way, we could slightly increase the performance of the logistic regression model. However, since we are also not tuning individual features specifically for other ML models, we believe that we made the best level playing field as possible.

After all, the strength of ML is that an algorithm learns how it can use a feature best to yield the best predictions.

6.2.2 Reliability and validity

Part of the discussion comprises the reliability and validity of the research. Since we have been making use of an assessment framework with a scoring scale, we are most interested in the reliability and validity of the way of measurements and the measurements itself. Cooper and Schindler, 2014 mention several aspects for both reliability and validity to assess, which we will mention the most important ones of below.

Reliability is defined as the extent to which the results can be reproduced when the research is repeated under the same conditions. In general, literature such as Cooper and Schindler, 2014 or Middleton, 2019, describe three types of reliability: test-retest, interrater, and internal consistency. We will shortly address those reliability assessment areas. The test-retest reliability assesses the consistency of a measurement across time. This thesis is reliable if it is repeated at a later moment, although we note that new guidelines or regulations can impact the assessment framework. With respect to the iterrater reliability, the measure whether different observers would get consistent results, this thesis shows room for improvement. Although different observers would probably rank the different ML models in the same order, likely their scores would be slightly different. This, however, does not mean that the conclusions are inconsistent, since the fluctuations in scores are not that significant, and the conclusions can be distilled from the ranking. In order to increase the iterrater reliability of this thesis, one could reach out to an expert panel, and aggregate different evaluations to an average score. We touch upon that in the final subsection of this chapter: recommendations for further research. Lastly, the internal consistency is about the measurement itself, for example, whether different parts of a survey are consistent with each other. This last aspect cannot be determined in this research, as the assessment framework is compact form, that does not have iterative questions such as in a large survey.

With regard to the reliability of the outcomes of this thesis, we must also note that in this thesis we shifted our focus to inherently interpretable ML, and completely neglected XAI techniques. However, it could be the case in the near future that the regulator is of the opinion that black boxes in combination with XAI techniques satisfy their requirements sufficiently. In this case, also black box models are applicable for the use in IRB models, which will probably overrule the inherently interpretable ML models treated in this thesis in terms of performance. However, that does not imply that this research is not of significance, as we state in the next section on contributions and relevance.

When assessing the validity of the research, we can distinguish internal and external validity. Starting with internal validity, three types are defined by Cooper and Schindler, 2014: content, criterion-related, and construct validity. Content validity is the degree to which the content of the items adequately represents the universe of all relevant items under study. In this thesis, the content is well validated by taking a perspective from all stakeholders involved, the financial institutions and the supervisor/regulator. On top of that, we made use of most recent literature and guidelines on ML in financial institutions. Criterion-related validity measures the degree to which the predictor is adequate in capturing relevant aspects of the criterion. That means, that what we measure, is also the thing that we are trying to measure. We

covered this mainly by doing further research on concepts described in regulations. For example, we decompose interpretability by investigating all relevant aspects involved. In this way, we were able to come to more measurable criteria. Lastly, in construct validity, we try to answer the question: 'what accounts for the variance in the measure?'. Just like the last aspect of consistency, internal consistency, this is not applicable to our research, as we do not have multiple scores for each criterion. To address this and to exclude bias, one could, as mentioned before, opt for an expert panel.

Finally, next to internal validity, we should also address the external validity. "*The external validity of research findings is the data's ability to be generalized across persons, settings, and times*" (Cooper and Schindler, 2014). This is something that is currently not known, and can be seen as a limitation of this research. Specifically, we used a US-based peer-to-peer lending dataset. That means, although there is no specific reason to assume otherwise, the results of this thesis are location specific. Also, the type of loans is fairly specific, which harms the external validity. Extrapolating the results of this thesis to a 'regular' bank portfolio is not directly possible, since these portfolios are probably a lot different compared to peer-to-peer lending. However, the research, the data processing steps, and the assessment framework itself actually can be directly used by a bank to come to their own conclusions in terms of interpretability, performance, and ease of implementation per model.

6.2.3 Contribution and relevance

In this subsection, we want to shortly highlight this thesis' contribution to theory and its relevance in practice.

Regarding the contribution to the scientific body of knowledge, this thesis can be seen as an exploratory study on the use of ML in IRB models. We distinguish several contributions. First, we introduced a good overview of the current industry's perspective, and the current state of regulations and guidelines on the use of ML in IRB models. On top of that, we constructed a novel assessment framework to compare and contrast ML models on their applicability for the use in IRB models. The concepts used in this framework can also be of use for research on XAI techniques. Finally, the assessment of three state-of-the-art ML models, is a new addition to the literature, and is of added value in the research area of inherently interpretable ML, or, white boxes.

In terms of relevance in practice, this thesis can be of added value for multiple purposes. First, this research, and specifically the assessment framework, can be used as a checklist or tool by financial institutions. It can help them find out whether their proposed ML approach satisfies the criteria of the regulator and their own interests. Additionally, it could help them to define on which criterion or criteria their focus should be placed in convincing the regulator for the use of a certain ML approach, or in the documentation and validation of an ML model. Next to financial institutions, the regulator can use this thesis as a stepping stone to define prudent and clear regulations for the use of ML in IRB models, or at least communicate their expectations clearly. This helps to manage expectations from the financial industry and could ultimately lead to a fair and effective evaluation of ML in IRB models by the supervisor. Lastly, this thesis is a novel way to assess this type of applicability of ML in the industry. It can serve as inspiration for the abovementioned parties, financial institutions, regulators, and supervisors, to keep the conversation going on this hot topic that needs a collaborative attitude of all actors involved.

6.2.4 Recommendations for further research

To finalize this research, we advise on some topics that can be researched in the future. The recommendations for further research cover two areas: 1) technical recommendations to make improvements on or extend this research, and 2) recommendations for further research if we look at this thesis in a broader context:

Technical recommendations for improvements

- Starting with the recommendation that we touched upon earlier, one could find a panel of experts in the field of IRB models and ML. The panel of experts can be used to let them all score the investigated models on the different criteria. Since there is no ground truth for the evaluation of models on relatively vague constructs such as simplicity and decomposability, one can filter out the bias of one opinion by asking multiple experts. On top of that, one can perform a quantitative analysis on the results of the survey that is filled in by multiple experts. In this way, the results are not dependent on one researcher's opinion. This makes it also possible to quantitatively evaluate the aspects of both reliability and validity.

- To increase the relevance and possibly find some other interesting insights, one can train and test the chosen models on different datasets. First, it is useful to validate whether we observe the same performance ranking amongst the different models on different datasets, and possibly different industries. Secondly, looking at the characteristics of datasets, one can gather valuable insights on strength and weaknesses of the models. For example, is a specific ML model very good in highly imbalanced datasets, are they robust in multiple data areas, and do they perform well on extremely high or low dimensional data?

- Since the GPSR is an efficient way to find correlations of high order between features, it can be specifically interesting to look whether it can be used to construct proxy variables, or other artificial new features. Proxy variables can be used in a classification problem when the 'real' feature is not available in the dataset. Artificial new features can be of added value when used in a classic logistic regression, to improve predictive performance without sacrificing interpretability and shifting to new models.

- Next to that, one can take a closer look at the degree of imbalance of the data, and the way of tackling this. Over- and under-sampling techniques have shown to be of no use in other research on Lendingclub data, since the dataset was relatively balanced. However, one could research which algorithms work better on extremely imbalanced datasets by artificially generating different imbalanced datasets. It could, for example, possibly be found that some are very robust for extremely imbalanced datasets, for example, 0.1% default rate, which make them extra interesting for low default portfolios of banks.

- Lastly, an increased of computing power could be of use to further improve the modelling of the algorithm. Specifically, the hyperparameter tuning can be extended thoroughly. First of all, the learning algorithms were instructed to use a balanced weighting scheme. The weight assigned to each class is actually also a hyperparameter that can be tuned. For the GAMI-Net, we used default settings on all neural networks, while these could also be tuned to improve the

performance. Likewise, there are many more hyperparameters that could be tuned closer to the optimal solution.

Further research in a broader context

- Looking beyond the scope of this thesis, one can also dive deeper into the problems that arise when trying to implement ML into the organization. Although many interpretable models are already developed, and ML models have also shown to be promising, it is still not the case that they are adopted in IRB models. What could be interesting to look at, is whether organizational burdens also slow down the process of ML adoption. For example, the fear of change is something that can be woven into the culture of a company. If the company's culture is a bottleneck for the implementation, maybe another approach to solve this problem is interesting to look at.

- Another thing that can be interesting to research is to find out what can be done in the supervisory landscape to increase the adoption speed of innovative techniques such as ML. The point in many complex and automated technologies is that one will be more and more focused on the outputs of the models. Especially, when advanced techniques become the standard in the industry, one should not neglect that a company implementing such a model has many opportunities during selection, testing and training processes to tune a model towards its preferences. In the case of an IRB model, this could be, for example, a strong focus on minimizing capital requirements. It is up to further research how we can ensure a strict supervision on the data processing and model construction in the future, when models become more and more complex. Also within the industry, this is an ongoing concern: "*Adjustments in the risk management framework will need to be made in order to address AI-associated risks. For example, model development assumptions and methodologies, model input, and control measures will all need to be revisited*" (Deloitte China, 2022). But also, the BIS acknowledges that a full validation in the AI governance process becomes unacceptably large: "*Overly onerous regulatory requirements to 'prove' the accuracy of an ML algorithm may not be achievable. (...) In practice, this (model validation) requires a line-by-line review of the source code, a comprehensive analysis of all datasets used, and an examination of the model and its parameters, which some firms view as unachievable*" (Prenio and Yong, 2021). One research area could be to find a right balance between the intensity of governance measures and the impact of the AI solution in place.

Bibliography

Albanesi, Stefania and Domonkos F. Vamossy (Aug. 2019). *Predicting Consumer Default: A Deep Learning Approach*. NBER Working Papers 26165. National Bureau of Economic Research, Inc. URL: https://ideas.repec.org/p/nbr/nberwo/26165.html.

Alonso, Andrés and José Manuel Carbó (Oct. 2020). *Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost*. Working Papers 2032. Banco de España. URL: https://ideas.repec.org/p/bde/wpaper/2032.html.

Bacham, Dinesh and Dr. Janet Zhao (July 2017). *Machine learning: Challenges, lessons, and opportunities in credit risk modeling*. URL: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling.

Barredo Arrieta, Alejandro et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Basel Committee on Banking Supervision (Sept. 2000). "Principles for the Management of Credit Risk". In: URL: https://www.bis.org/publ/bcbs75.pdf.

— (Mar. 2022). "Newsletter on artificial intelligence and machine learning". In: URL: https://www.bis.org/publ/bcbs_nl27.htm.

Berenbaum, Dina (Feb. 2020). *Is it easy to explain? Local explainability*. URL: https://towardsdatascience.com/is-it-easy-to-explain-local-explainability-4f325565210c.

Bhandari, Pritha (July 2020). *Levels of Measurement*. URL: https://www.scribbr.com/statistics/levels-of-measurement/.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani (2013). "A lasso for hierarchical interactions". In: *The Annals of Statistics* 41.3, pp. 1111 –1141. DOI: 10.1214/13-AOS1096. URL: https://doi.org/10.1214/13-AOS1096.

Braak, Lars ter (Dec. 2021). *Introduction to Probabilistic Classification: A Machine Learning Perspective*. URL: https://towardsdatascience.com/introduction-to-probabilistic-classification-a-machine-learning-perspective-b4776b469453.

Breeden, Joseph (Jan. 2020). "Survey of Machine Learning in Credit Risk". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3616342.

Breiman, L. et al. (1984). *Classification and Regression Trees*. Taylor & Francis. ISBN: 9780412048418. URL: https://books.google.nl/books?id=JwQx-WOmSyQC.

Burkart, Nadia and Marco Huber (Jan. 2021). "A Survey on the Explainability of Supervised Machine Learning". In: *Journal of Artificial Intelligence Research* 70. DOI: 10.1613/jair.1.12228.

Chang, Chun-Hao, Rich Caruana, and Anna Goldenberg (2022). "NODE-GAM: Neural Generalized Additive Model for Interpretable Deep Learning". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=g8NJR6fCCl8.

Choi, Nam Hee, William Li, and Ji Zhu (2010). "Variable Selection With the Strong Heredity Constraint and Its Oracle Property". In: *Journal of the American Statistical Association* 105.489, pp. 354–364. DOI: 10.1198/jasa.2010.tm08281. URL: https://doi.org/10.1198/jasa.2010.tm08281.

Christensen, Niels Johan et al. (2022). "Identifying interactions in omics data for clinical biomarker discovery using symbolic regression". In: *bioRxiv*. DOI: 10.1101/2022.01.14.475226. eprint: https://www.biorxiv.org/content/early/2022/05/24/2022.01.14.475226.full.pdf. URL: https://www.biorxiv.org/content/early/2022/05/24/2022.01.14.475226.

Cooper, D.R. and P.S. Schindler (2014). *Business Research Methods*. McGraw-Hill Education. ISBN: 9781259070952. URL: https://books.google.nl/books?id=fIy6DAEACAAJ.

Deloitte China (2022). *Artificial Intelligence for Credit Risk Management*. URL: https://www2.deloitte.com/cn/en/pages/risk/articles/artificial-intelligence-for-credit-risk-management.html#.

Doeme, Zsofia and Stefan Kerbl (Jan. 2018). *Comparability of Basel risk weights in the EU banking sector is questionable*. Tech. rep. VOXEU Centre of Economic Policy Research. URL: https://voxeu.org/article/bank-risk-weights-under-basel-are-not-comparable.

Edwards, Ward (1977). "How to Use Multiattribute Utility Measurement for Social Decisionmaking". In: *IEEE Transactions on Systems, Man, and Cybernetics* 7.5, pp. 326–340. DOI: 10.1109/TSMC.1977.4309720.

European Banking Authority (n.d.). *EBA at a glance*. URL: https://www.eba.europa.eu/about-us/eba-at-a-glance.

— (June 2013). *Capital Requirements Regulation (CRR): Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012*. URL: https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/504.

— (Nov. 2021). *Eba Discussion Paper on Machine Learning for IRB Models*. 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.

European Parliament and Council of the European Union (May 2016). *Council regulation (EU) no 2016/679 - General Data Protection Regulation*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Folpmers, Macro (Dec. 2021). *The Rise of Machine Learning in IRB Models: New EBA Outlook Could Open Door*. URL: https://www.garp.org/risk-intelligence/technology/the-rise-of-machine-learning-in-irb-models-new-eba-outlook-could-open-door.

Fürnkranz, Johannes (2010). "Decision Tree". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pp. 263–267. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_204. URL: https://doi.org/10.1007/978-0-387-30164-8_204.

Gonzalez, Rodrigo, Jose Savoia, and Fernando Sotelino (Apr. 2012). "How to minimize incentives to rating-centered regulatory arbitrage? the current Brazilian

Approach to Minimum Capital Requirementes". In: *Journal of Business and Policy Research* 7, pp. 30–48.

Goyal, Chirag (Mar. 2021). *Multicollinearity in Data Science*. URL: https://www.analyticsvidhya.com/blog/2021/03/multicollinearity-in-data-science/.

Hastie, T.J. and R.J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN: 9780412343902. URL: https://books.google.nl/books?id=qa29r1Ze1coC.

High-Level Expert Group on AI (set up by EU) (Apr. 2019). *Ethics Guidelines for Trustworthy Artificial Intelligence*. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651.

— (July 2020). *Assessment List for Trustworthy AI*. URL: https://airegio.ems-carsa.com/nfs/programme_5/call_3/call_preparation/ALTAI_final.pdf.

Hull, John (2007). *Risk management and financial institutions*. English. Pearson Education Upper Saddle River, N.J, xvi, 500p. ; ISBN: 0132397900.

Institute of International Finance (Aug. 2019). *Machine Learning in Credit Risk Report*. Summary Report 2nd. URL: https://www.iif.com/Portals/0/Files/content/Research/iif_mlcr_2nd_8_15_19.pdf.

International Organization for Standadization (2017). *ISO/IEC 38505-1:2017(en) Information technology — Governance of IT — 3.7: machine learning*. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:38505:-1:ed-1:v1:en.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR*. arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

Iqbal, Zahid (June 2015). *How are "Evolutionary Algorithms" different from "Learning Algorithms"?*

James, Gareth et al. (2021). *An introduction to statistical learning with applications in R*. 2nd ed. Springer. URL: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf.

Jeppesen, Jacob et al. (Aug. 2019). "A cloud detection algorithm for satellite imagery based on deep learning". In: *Remote Sensing of Environment* 229, pp. 247–259. DOI: 10.1016/j.rse.2019.03.039.

Kimura, Nobuaki et al. (Dec. 2019). "Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions". In: *Water* 12, p. 96. DOI: 10.3390/w12010096.

Landwehr, Niels, Mark Hall, and Eibe Frank (May 2005). "Logistic Model Trees". In: 59.1–2, 161–205. ISSN: 0885-6125. DOI: 10.1007/s10994-005-0466-3. URL: https://doi.org/10.1007/s10994-005-0466-3.

Larsen, Emil (n.d.). *Complexity-Loss Trade-Off*. URL: https://docs.abzu.ai/docs/tutorials/python/pareto_front.html.

Lipton, Zachary C (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Lou, Yin, Rich Caruana, and Johannes Gehrke (2012). "Intelligible Models for Classification and Regression". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Association for Computing Machinery, 150–158. ISBN: 9781450314626. DOI: 10.1145/2339530.2339556. URL: https://doi.org/10.1145/2339530.2339556.

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: http:

//papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, Scott M. et al. (2019). "Explainable AI for Trees: From Local Explanations to Global Understanding". In: *CoRR* abs/1905.04610. arXiv: 1905.04610. URL: http://arxiv.org/abs/1905.04610.

Marlin, Steve (Aug. 2021). "Wells touts new explainability technique for AI credit models". In: *Credit risk & modelling Special report 2021 - risk.net*. URL: https://www.risk.net/risk-management/7865541/wells-touts-new-explainability-technique-for-ai-credit-models.

Menard, S. and SAGE. (2002). *Applied Logistic Regression Analysis*. Applied Logistic Regression Analysis no. 106. SAGE Publications. ISBN: 9780761922087. URL: https://books.google.nl/books?id=EAI1QmUUsbUC.

Middleton, Fiona (July 2019). *Reliability vs Validity in Research | Differences, Types and Examples*. URL: https://www.scribbr.com/methodology/reliability-vs-validity/.

Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. URL: christophm.github.io/interpretable-ml-book/.

Narkhede, Sarang (June 2018). *Understanding AUC - ROC Curve*. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

Neumann, Frank (2012). "Computational Complexity Analysis of Multi-Objective Genetic Programming". In: *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*. GECCO '12. Association for Computing Machinery, 799–806. ISBN: 9781450311779. DOI: 10.1145/2330163.2330274. URL: https://doi.org/10.1145/2330163.2330274.

Olson, Matthew Lyle and Abraham J. Wyner (Dec. 2018). "Making Sense of Random Forest Probabilities: a Kernel Perspective". In: *ArXiv* abs/1812.05792.

Orzechowski, Patryk, William G. La Cava, and Jason H. Moore (2018). "Where are we now? A large benchmark study of recent symbolic regression methods". In: *CoRR* abs/1804.09331. arXiv: 1804.09331. URL: http://arxiv.org/abs/1804.09331.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Petropoulos, Anastasios et al. (2019). "A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting". In: *Are post-crisis statistical initiatives completed?* Ed. by Bank for International Settlements. Vol. 49. IFC Bulletins chapters. Bank for International Settlements. URL: https://ideas.repec.org/h/bis/bisifc/49-49.html.

Petrov, Christo (Feb. 2022). *25+ impressive Big Data Statistics for 2022*. URL: https://techjury.net/blog/big-data-statistics/.

Prenio, Jermy and Jeffery Yong (Aug. 2021). *FSI Insights on policy implementation No 35. Humans keeping AI in check - emerging regulatory expectations in the financial sector*. Tech. rep. Financial Stability Institute of the Bank for International Settlements.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1558602380.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144.

Robnik-Sikonja, M. and Marko Bohanec (2018). "Perturbation-Based Explanations of Prediction Models". In: *Human and Machine Learning*.

Rudin, Cynthia (May 2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1, pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Sarıgül, M., B.M. Ozyildirim, and M. Avci (2019). "Differential convolutional neural network". In: *Neural Networks* 116, pp. 279–287. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.04.025. URL: https://www.sciencedirect.com/science/article/pii/S0893608019301315.

Sigrist, Fabio (Jan. 2022). *Demystifying ROC and precision-recall curves*. URL: https://towardsdatascience.com/demystifying-roc-and-precision-recall-curves-d30f3fad2cbf.

Swets, John A., Robyn M. Dawes, and John Monahan (2000). "Math-based aids for making decisions in medicine and industry could improve many diagnoses — often saving lives in the process Better DECISIONS through SCIENCE". In.

TIBCO (n.d.). *What is a Neural Network?* URL: https://www.tibco.com/reference-center/what-is-a-neural-network.

Triple A – Risk Finance (Apr. 2022). *Machine learning for IRB model*. URL: https://www.aaa-riskfinance.nl/publicaties/machine-learning-for-irb-model/.

Urbanowicz, Ryan and Jason Moore (Jan. 2009). "Learning Classifier Systems: A Complete Introduction, Review, and Roadmap". In: *Journal of Artificial Evolution and Applications* 2009. DOI: 10.1155/2009/736398.

Vašíček, Oldřich (1987). *Probability of Loss on a Loan Portfolio*. Tech. rep. KMV.

Vujnovic, Milos, Nebojsa Nikolic, and Anja Vujnovic (Jan. 2016). "Validation of loss given default for corporate". In: *Istrazivanja i projektovanja za privredu* 14, pp. 465–476. DOI: 10.5937/jaes14-11752.

Xue, Ying et al. (Apr. 2018). "Using a new framework of two-phase generalized additive models to incorporate prey abundance in spatial distribution models of juvenile slender lizardfish in Haizhou Bay, China". In: *Marine Biology Research* 14, pp. 1–16. DOI: 10.1080/17451000.2018.1447673.

Yang, Zebin, Aijun Zhang, and Agus Sudjianto (2021). "GAMI-Net: An Explainable Neural Network based on Generalized Additive Models with Structured Interactions". In: *Pattern Recognition* 120, p. 108192.

Zhang, Zixuan (June 2019). *Boosting Algorithms Explained: Theory, Implementation, and Visualization*. URL: https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30.

Appendix A

Data Pre-processing

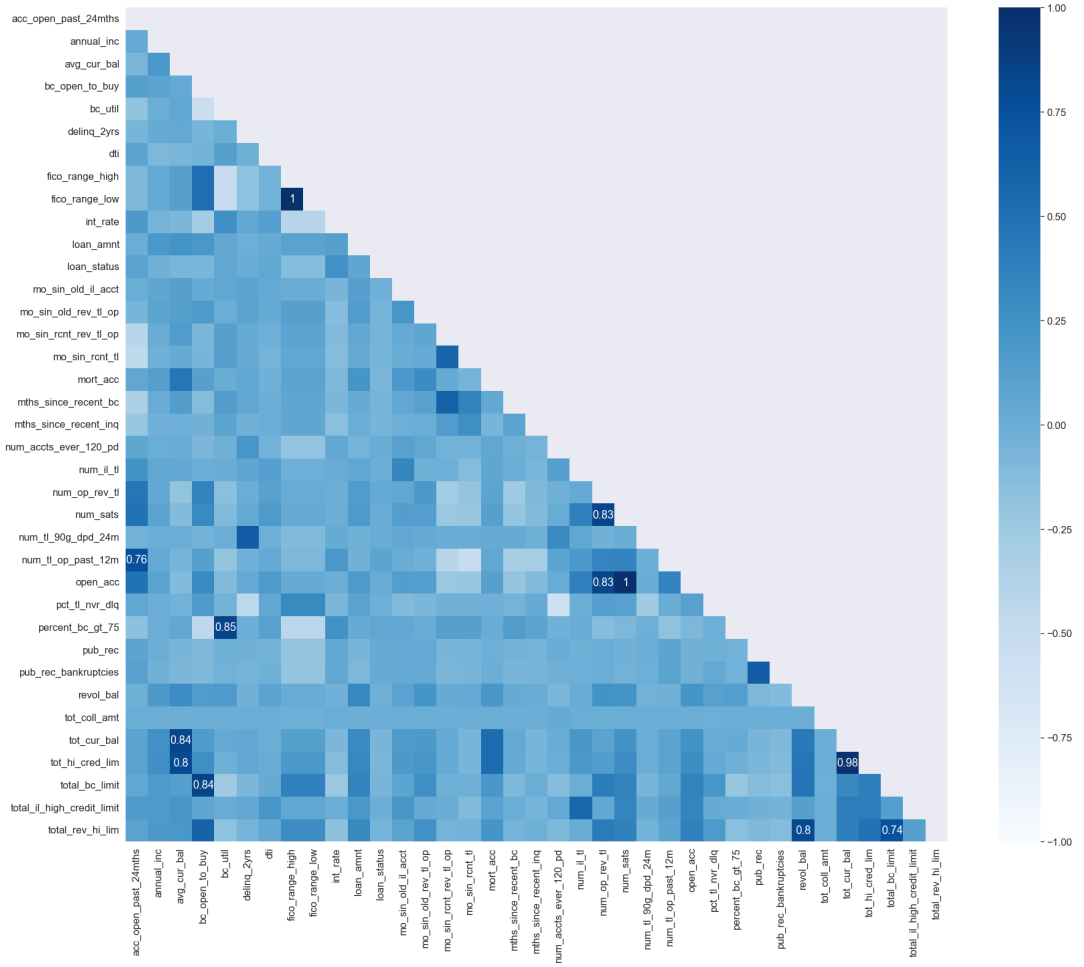

FIGURE A.1: Correlation matrix of the numerical features in the training set.

TABLE A.1: High correlations between features

Feature (1)	**Feature (2)**	**Correlation**
fico_range_low	fico_range_high	1.00
num_sats	num_op_rev_tl	0.83
num_tl_op_past_12m	acc_open_past_24mths	0.76
open_acc	num_op_rev_tl	0.83
open_acc	num_sats	1.00
percent_bc_gt_75	bc_util	0.85
tot_cur_bal	avg_cur_bal	0.84
tot_hi_cred_lim	avg_cur_bal	0.80
tot_hi_cred_lim	tot_cur_bal	0.98
total_bc_limit	bc_open_to_buy	0.84
total_rev_hi_lim	revol_bal	0.80
total_rev_hi_lim	total_bc_limit	0.74

TABLE A.2: Iterative feature deletion based on VIF scores

	Iteration						
High correlated features	**1**	**2**	**3**	**4**	**5**	**6**	**7**
acc_open_past_24mths	2.82	2.82	2.82	2.82	2.82	2.81	2.81
avg_cur_bal	5.38	5.38	5.38	4.58	4.50	4.48	1.22
bc_open_to_buy	9.44	9.43	9.39	9.39	7.23		
bc_util	4.80	4.80	4.80	4.80	4.79	4.13	4.13
fico_range_high	7.06M	1.71	1.70	1.68	1.67	1.61	1.61
fico_range_low	7.06M						
num_op_rev_tl	4.06	4.06	4.05	4.02	3.88	3.87	3.81
num_sats	447.81	447.81	4.45	4.29	4.26	4.25	3.49
num_tl_op_past_12m	2.43	2.43	2.43	2.42	2.42	2.42	2.42
open_acc	449.82	449.82					
percent_bc_gt_75	3.59	3.58	3.58	3.58	3.58	3.55	3.55
revol_bal	7.35	7.35	7.34	6.94	1.96	1.69	1.59
tot_cur_bal	36.83	36.83	36.82				
tot_hi_cred_lim	33.80	33.79	33.78	5.76	5.61	5.60	
total_bc_limit	7.13	7.13	7.12	7.12	7.06	2.02	1.93
total_rev_hi_lim	10.49	10.49	10.48	9.65			

TABLE A.3: Percentage of outliers in the numerical features

Feature	**Percentage of outliers**
acc_open_past_24mths	2.12
annual_inc	4.23
avg_cur_bal	4.58
bc_util	0.0
delinq_2yrs	17.01
dti	0.63
fico_range_low	2.22
int_rate	1.6
loan_amnt	0.96
mo_sin_old_il_acct	2.36
mo_sin_old_rev_tl_op	2.63
mo_sin_rcnt_rev_tl_op	6.91
mo_sin_rcnt_tl	5.14
mort_acc	1.08
mths_since_recent_bc	7.69
mths_since_recent_inq	1.04
num_accts_ever_120_pd	20.65
num_il_tl	3.7
num_op_rev_tl	1.84
num_sats	2.13
num_tl_90g_dpd_24m	4.86
num_tl_op_past_12m	2.46
pct_tl_nvr_dlq	0.0
percent_bc_gt_75	0.0
pub_rec	15.01
pub_rec_bankruptcies	11.25
revol_bal	5.27
revol_util	0.01
tot_coll_amt	13.62
total_bc_limit	5.18
total_il_high_credit_limit	4.53

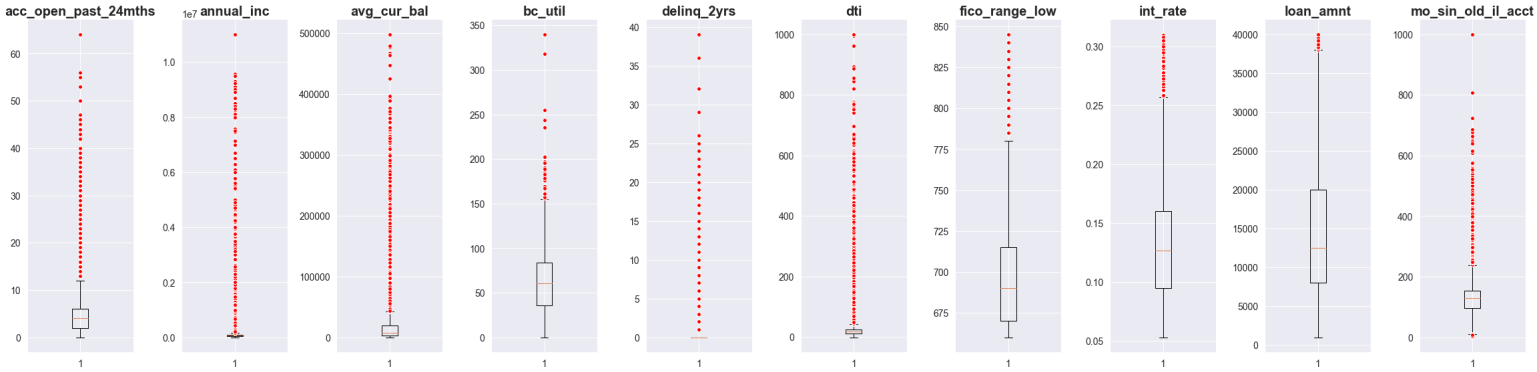

FIGURE A.2: Boxplots of all numerical values (1 of 3).

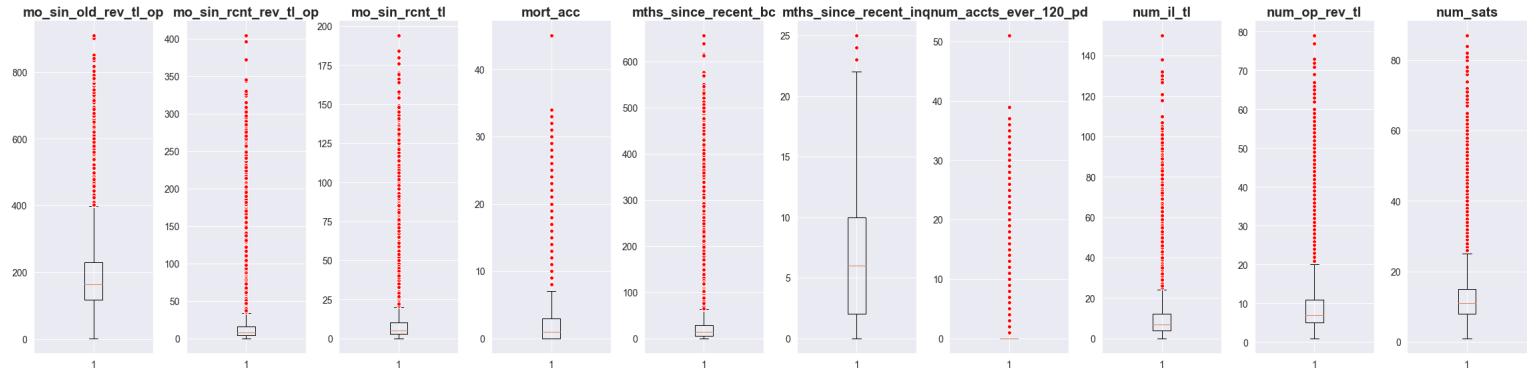

FIGURE A.3: Boxplots of all numerical values (2 of 3).

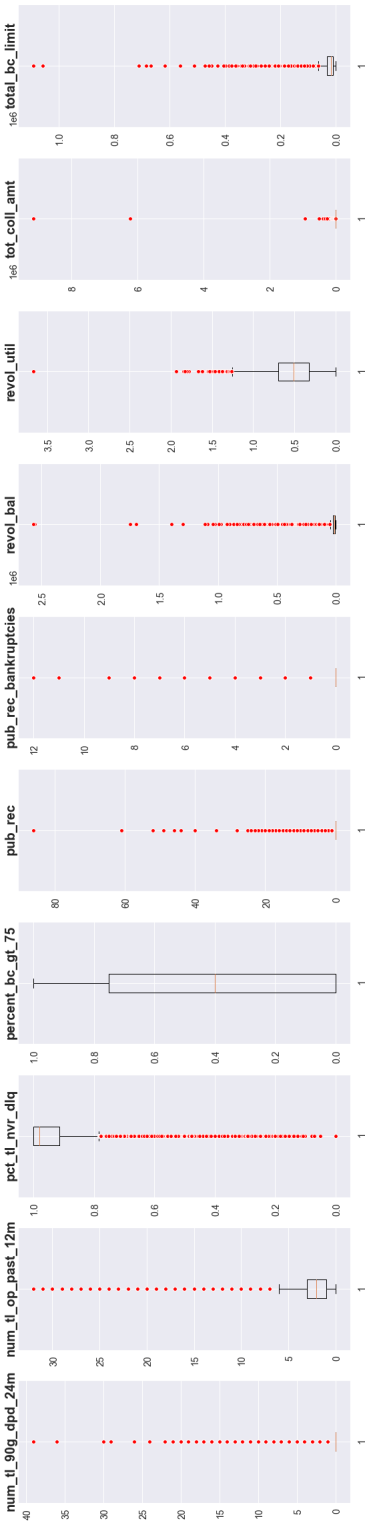

FIGURE A.4: Boxplots of all numerical values (3 of 3).

TABLE A.4: Final variable inclusion and exclusion with motivation

Variable	Inclusion/Exclusion	Variable	Inclusion/Exclusion
acc_now_delinq	Drop: insufficient data	num_accts_ever_120_pd	Keep
acc_open_past_24mths	Keep	num_actv_bc_tl	Drop: post-origination
addr_state	Keep	num_actv_rev_tl	Drop: post-origination
all_util	Drop: post-origination	num_bc_sats	Drop: post-origination
annual_inc	Keep	num_bc_tl	Drop: post-origination
annual_inc_joint	Drop: insufficient data	num_il_tl	Keep
application_type	Keep	num_op_rev_tl	Keep
avg_cur_bal	Keep	num_rev_accts	Drop: post-origination
bc_open_to_buy	Drop: based on VIF	num_rev_tl_bal_gt_0	Drop: post-origination
bc_util	Keep	num_sats	Keep
chargeoff_within_12_mths	Drop: post-origination	num_tl_120dpd_2m	Drop: post-origination
collection_recovery_fee	Drop: post-origination	num_tl_30dpd	Drop: post-origination
collections_12_mths_ex_med	Drop: post-origination	num_tl_90g_dpd_24m	Keep
debt_settlement_flag	Drop: post-origination	num_tl_op_past_12m	Keep
deferral_term	Drop: post-origination	open_acc	Drop: based on VIF
delinq_2yrs	Keep	open_acc_6m	Drop: insufficient data
delinq_amnt	Drop: post-origination	open_act_il	Drop: post-origination
dti	Keep	open_il_12m	Drop: insufficient data
dti_joint	Drop: post-origination	open_il_24m	Drop: insufficient data
earliest_cr_line	Drop: irrelevant	open_rv_12m	Drop: post-origination
emp_length	Drop: irrelevant	open_rv_24m	Drop: post-origination
emp_title	Drop: irrelevant	orig_projected_additional_accrued_interest	Drop: post-origination
fico_range_high	Drop: based on VIF	out_prncp	Drop: post-origination
fico_range_low	Keep	out_prncp_inv	Drop: post-origination
funded_amnt	Drop: post-origination	payment_plan_start_date	Drop: post-origination
funded_amnt_inv	Drop: post-origination	pct_tl_nvr_dlq	Keep
grade	Drop: dep. on int_rate	percent_bc_gt_75	Keep
hardship_amount	Drop: post-origination	policy_code	Drop: irrelevant
hardship_dpd	Drop: post-origination	pub_rec	Keep
hardship_end_date	Drop: post-origination	pub_rec_bankruptcies	Keep
hardship_flag	Drop: post-origination	purpose	Keep
hardship_last_payment_amount	Drop: post-origination	pymnt_plan	Drop: post-origination
hardship_length	Drop: post-origination	recoveries	Drop: post-origination
hardship_loan_status	Drop: post-origination	revol_bal	Keep
hardship_payoff_balance_amount	Drop: post-origination	revol_bal_joint	Drop: irrelevant
hardship_reason	Drop: post-origination	revol_util	Keep
hardship_start_date	Drop: post-origination	sec_app_chargeoff_within_12_mths	Drop: irrelevant
hardship_status	Drop: post-origination	sec_app_collections_12_mths_ex_med	Drop: irrelevant
hardship_type	Drop: post-origination	sec_app_earliest_cr_line	Drop: irrelevant
home_ownership	Keep	sec_app_fico_range_high	Drop: irrelevant
id	Drop: irrelevant	sec_app_fico_range_low	Drop: irrelevant
il_util	Drop: post-origination	sec_app_inq_last_6mths	Drop: irrelevant
initial_list_status	Drop: irrelevant	sec_app_mort_acc	Drop: irrelevant
inq_fi	Drop: post-origination	sec_app_num_rev_accts	Drop: irrelevant
inq_last_12m	Drop: post-origination	sec_app_open_acc	Drop: irrelevant
inq_last_6mths	Drop: post-origination	sec_app_open_act_il	Drop: irrelevant
installment	Drop: dep. on loan_amnt	sec_app_revol_util	Drop: irrelevant
int_rate	Keep	sub_grade	Drop: dep. on int_rate
issue_d	Drop: data leakage	tax_liens	Drop: post-origination
last_credit_pull_d	Drop: post-origination	term	Keep
last_fico_range_high	Drop: post-origination	title	Drop: irrelevant
last_fico_range_low	Drop: post-origination	tot_coll_amt	Keep
last_pymnt_amnt	Drop: post-origination	tot_cur_bal	Drop: based on VIF
last_pymnt_d	Drop: post-origination	tot_hi_cred_lim	Drop: based on VIF
loan_amnt	Keep	total_acc	Drop: post-origination
loan_status	Keep	total_bal_ex_mort	Drop: post-origination
max_bal_bc	Drop: post-origination	total_bal_il	Drop: post-origination
mo_sin_old_il_acct	Keep	total_bc_limit	Keep
mo_sin_old_rev_tl_op	Keep	total_cu_tl	Drop: post-origination
mo_sin_rcnt_rev_tl_op	Keep	total_il_high_credit_limit	Keep
mo_sin_rcnt_tl	Keep	total_pymnt	Drop: post-origination
mort_acc	Keep	total_pymnt_inv	Drop: post-origination
mths_since_last_delinq	Drop: insufficient data	total_rec_int	Drop: post-origination
mths_since_last_major_derog	Drop: insufficient data	total_rec_late_fee	Drop: post-origination
mths_since_last_record	Drop: post-origination	total_rec_prncp	Drop: post-origination
mths_since_rcnt_il	Drop: insufficient data	total_rev_hi_lim	Drop: based on VIF
mths_since_recent_bc	Keep	url	Drop: irrelevant
mths_since_recent_bc_dlq	Drop: insufficient data	verification_status	Drop: irrelevant
mths_since_recent_inq	Keep	verified_status_joint	Drop: irrelevant
mths_since_recent_revol_delinq	Drop: insufficient data	zip_code	Drop: dep. on addr_state
next_pymnt_d	Drop: post-origination		

TABLE A.5: Feature description of the features that are used as inputs for the models.

Variable	Description
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
delinq_2yrs_01	Ever 30+ days delinquency in borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
int_rate	Interest Rate on the loan
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
mo_sin_old_il_acct	Months since oldest bank installment account opened
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_inq	Months since most recent inquiry.
accts_ever_120_pd	Account of user ever gone 120 or more days past due
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_sats	Number of satisfactory accounts
num_tl_90g_dpd_24m01	Ever a account 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts >75% of limit.
pub_rec01	Ever a derogatory public record
pub_rec_bankruptcies01	Ever had publicly record bankruptcies
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
coll_owned	Collections ever owned, binarized
total_bc_limit	Total bankcard high credit/credit limit
total_il_high_credit_limit	Total installment high credit/credit limit

Appendix B

Full Visualizations of Results

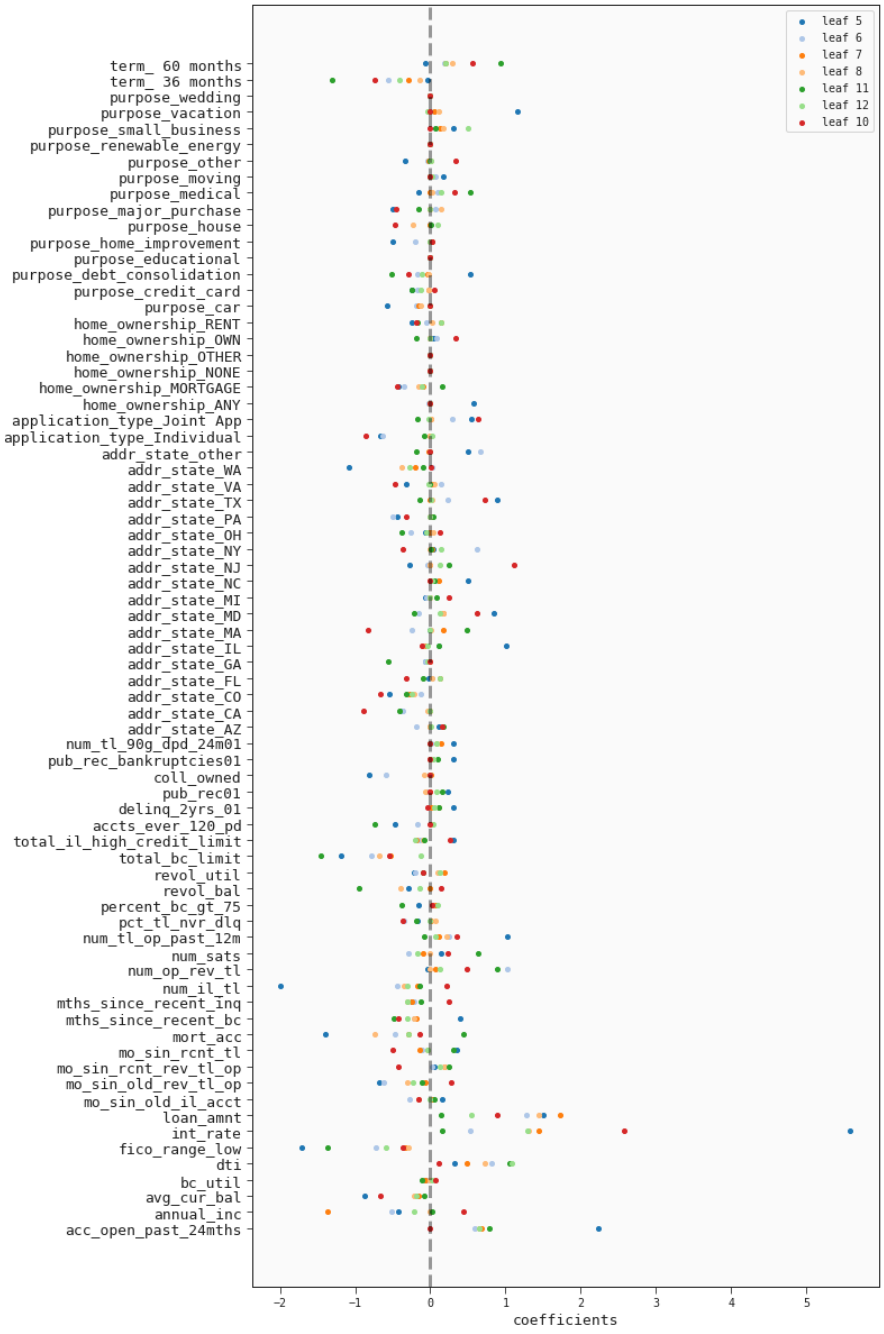

FIGURE B.1: LMT visualization, including all coefficients for the fitted logistic models at the leaf nodes of the model tree.

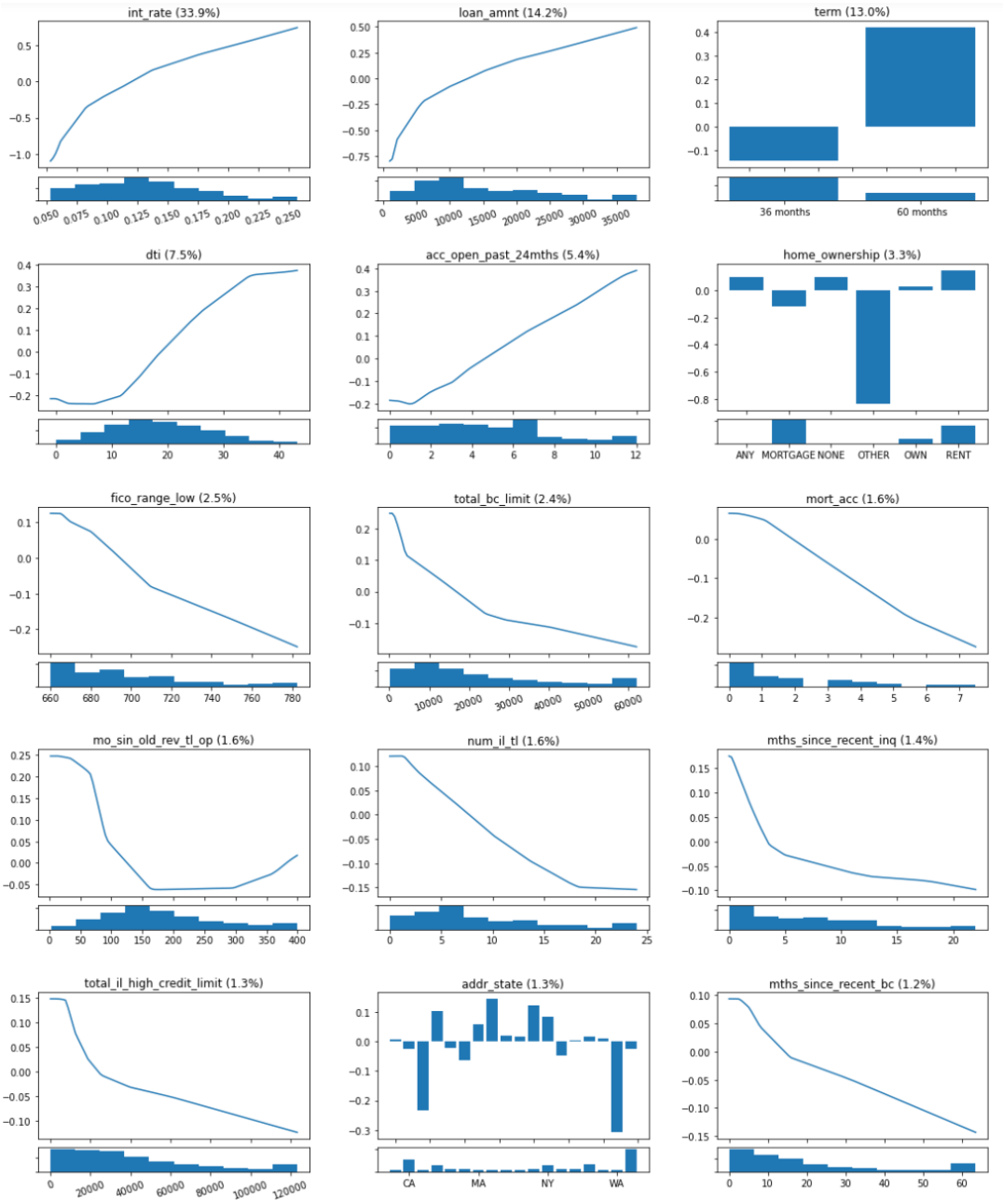

FIGURE B.2: All features and interactions included in the GAMI-Net (1 of 2).

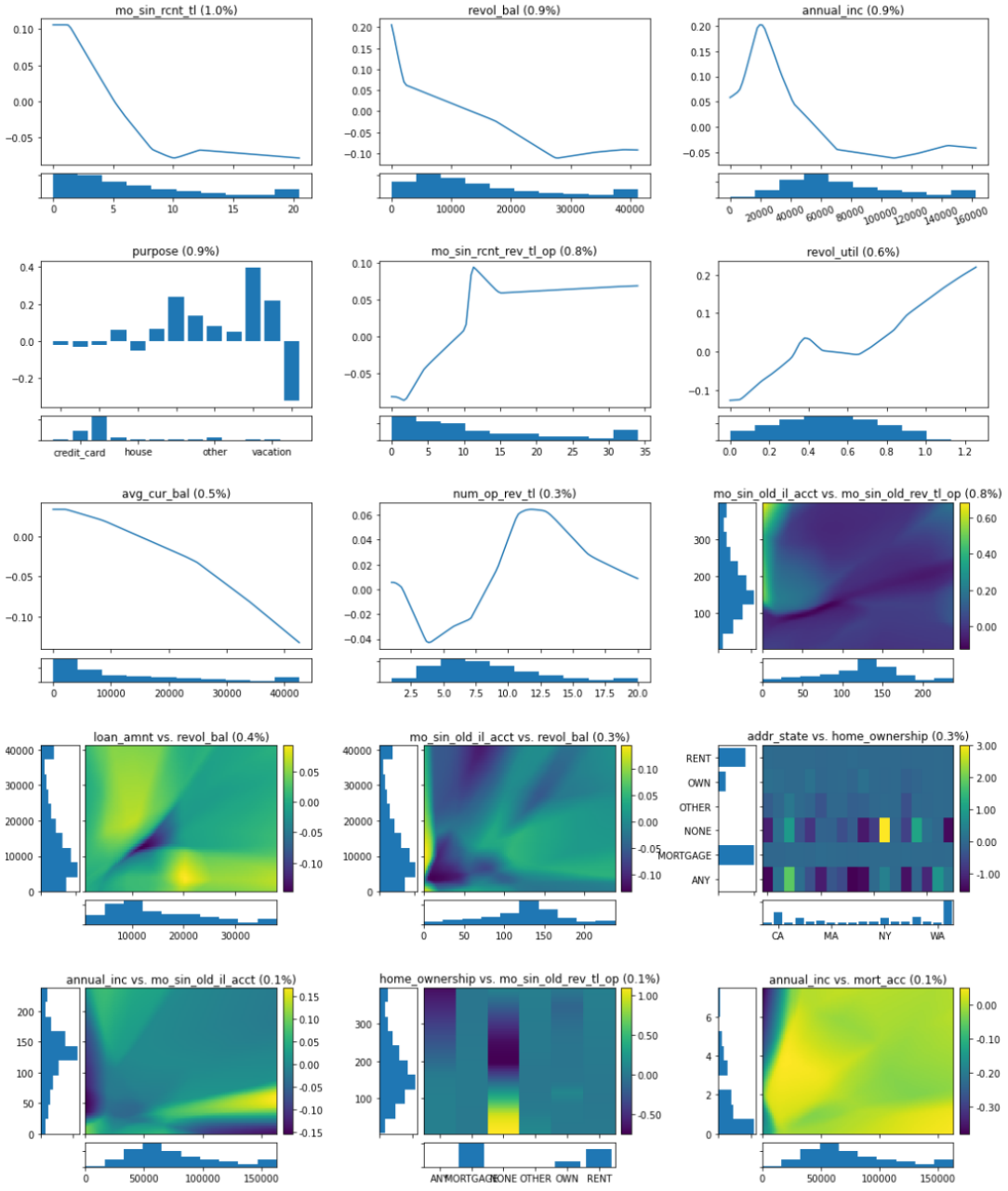

FIGURE B.3: All features and interactions included in the GAMI-Net (2 of 2).