Functional ANOVA for Inherently Interpretable Gradient Boosting in PD Modelling

Sara Kopelman

Master's thesis
2026:E44

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Master's Theses in Mathematical Sciences 2026:E44
ISSN 1404-6342

LUNFMS-3143-2026

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Abstract

For financial institutions, estimating the probability of default is central to credit-risk modelling. This is traditionally done using logistic regression, but machine learning methods have shown promising results because of their ability to capture more flexible relationships. However, the complexity of such models makes them difficult to interpret and consequently difficult to justify to regulators.

This thesis explores the use of functional ANOVA decomposition to construct an interpretable representation of depth-limited gradient-boosted decision trees for credit-risk modelling. The aim is to evaluate whether the resulting decomposition can be regarded as an inherently interpretable model that aligns with regulatory expectations.

The results show that the approach successfully reconstructs the fitted model as a sum of main effects and pairwise interaction effects. Simulations showed strong recovery for both main and interaction effects when the target function is observed directly. Recovery is weaker when the target is a latent score function, but the method still approximately recovers the main effects.

The method is also applied to Polish bankruptcy data, where it provides an interpretable representation of a fitted credit-risk model. The decomposition makes it possible to inspect the effects of individual financial ratios and selected pairwise interactions, and to assess whether these effects are reasonable from an economic perspective. The results indicate that the main effects are more stable and easier to interpret than the interaction effects. Overall, the findings suggest that the proposed decomposition can provide a transparent way to use flexible machine learning models in credit-risk modelling by providing a representation that can be directly inspected and interpreted.

Keywords: *probability of default; credit-risk; gradient-boosting; decision trees; functional ANOVA; interpretability; machine learning*

Popular scientific summary

Can we create a statistical model that is as flexible as possible but also easy to interpret?

That is a question of central importance within predictive modelling, particularly in high-stakes contexts. For financial institutions, for example, estimating the probability of default is of high importance to their lending activities. This is typically accomplished using logistic regression, which is a statistical modelling technique with a long history. Logistic regression is used for a reason. It can achieve good results and the model is easy to use and understand. However, one drawback is that it requires assumptions to be made regarding the structure of the relationships between the variables and the risk before the model is built.

In order to be able to capture any type of structure, we would prefer to make as few assumptions as possible. This is one reason to use more recent modelling techniques, known as machine learning. Machine learning models are characterized by being highly flexible and requiring fewer assumptions compared to traditional methods. However, one drawback is that they are highly complex and difficult to understand, to the point that they are sometimes referred to as "black boxes". Credit-risk modelling is highly regulated, with strict requirements regarding predictive models. The lack of interpretability of machine learning models has therefore led to limited uptake within the industry.

However, there are certain ways to make a machine learning model easier to understand. Often this is done using what is referred to as post-hoc methods, but such methods have the downside of only summarizing what the model does, without allowing us to truly inspect it. We would therefore prefer to have what is known as an inherently interpretable model, which can be understood directly. One such way is through functional decomposition, where we attempt to re-express our model as a sum of functions representing main and interaction effects. However, in order for us to use this sum to interpret the model, we need the decomposition to be unique. One way to accomplish this is to use functional ANOVA, where we choose the functions so that as much as possible is explained by the main effects.

The functional ANOVA decomposition is generally difficult to compute, but if we combine it with a certain category of machine learning models, known as tree ensembles, this can be performed in a relatively fast and efficient way. By doing this, we can construct a flexible machine learning model that is at the same time easier to understand.

For credit-risk modelling, this means that we can examine which risk drivers are most important, how they affect the predicted risk, and whether these effects are reasonable from an economic perspective. In this thesis, the method is studied using both simulated data and data on bankruptcies among Polish companies.

The results show that machine learning models do not have to be treated as black boxes. By combining tree-based machine learning with functional ANOVA, it is possible to construct a flexible model that is also interpretable. This suggests that the method has the potential to allow for a more flexible modelling approach for credit-risk modelling that aligns with legal requirements.

Acknowledgements

This Master's thesis was carried out during the spring semester 2026 within the non-retail credit-risk team at SEB.

I would like to extend my sincere gratitude to my academic supervisor Magnus Wiktorsson for his excellent guidance and interesting conversations throughout the process. I would also like to extend my thanks to my external supervisor Johan Palmquist at SEB for suggesting an engaging thesis topic and for his great advice.

I would also like to extend my thanks to the non-retail credit-risk team at SEB for providing such a welcoming and supportive environment. Finally, I would like to thank Erik Larsson for being a constant source of support through yet another thesis.

Contents

1 Introduction

Probability of default (PD) modelling is of central importance for lending institutions. Traditionally, such models have often relied on statistical methods such as logistic regression, which are relatively transparent but require the functional relationship between risk drivers and default risk to be specified in advance. Machine learning methods, such as gradient-boosted decision trees, provide a more flexible alternative by allowing nonlinearities and interactions to be learned from the data. However, their use in regulated credit-risk modelling remains limited, partly because this flexibility comes at the cost of interpretability.

In recent years, regulatory authorities have clarified that machine learning methods are not ruled out in principle for PD estimation. However, their use places additional demands, particularly with regards to explainability. This creates a central challenge: flexible machine learning models may be useful for modelling credit risk, but they must also be sufficiently transparent to satisfy regulatory expectations.

Many common interpretability methods are post-hoc, meaning that they are applied after a model has been fitted in order to summarize or approximate its behaviour. While such methods can be useful, they have also been questioned in high-stakes settings because the explanation may not be fully faithful to the fitted model. This has motivated interest in approaches where the fitted model itself can be represented in a more interpretable form.

One such approach is functional decomposition, where a multivariate prediction function is represented as a sum of lower-dimensional component functions. Functional ANOVA provides a principled way to construct such a decomposition by imposing centering and orthogonality conditions. In general, however, this decomposition can be difficult to compute, especially when the input variables are dependent.

Recent work has shown that the problem becomes substantially simpler for piecewise-constant functions, such as tree ensembles. In this setting, the functional ANOVA decomposition can be obtained through an iterative purification procedure. This makes it possible to decompose a fitted gradient-boosted tree model into an additive representation consisting of main effects and low-order interaction effects.

The aim of this thesis is to implement this approach and evaluate whether functional ANOVA decomposition of a depth-constrained gradient-boosted decision tree model can provide a useful interpretable framework for non-retail credit-risk modelling, with particular attention to regulatory expectations regarding explainability.

This is studied in two ways. First, simulation studies are used to evaluate whether the method can recover known functional structures under controlled conditions. Second, the approach is applied to a Polish bankruptcy dataset in order to examine whether the resulting decomposition produces effects that are stable, plausible, and interpretable in a realistic credit-risk setting.

The thesis is organized as follows. The background chapter introduces credit-risk modelling, statistical learning, gradient-boosted decision trees, interpretability, and functional ANOVA. The methodology chapter describes the purification procedure and the design of the simulation and empirical studies. The results chapter presents the findings, which are then discussed in relation to interpretability, model stability, and regulatory expectations. Finally, the thesis concludes with a summary of the main findings and possible directions for future research.

2 Theoretical background

This chapter introduces the theoretical concepts needed to motivate and develop the methodology used in the thesis. The chapter begins with credit-risk modelling, since the proposed method is studied in the context of probability of default modelling. It then introduces statistical learning and gradient-boosted decision trees, which provide the predictive modelling framework. The chapter subsequently discusses interpretability in machine learning. Finally, functional ANOVA decomposition is introduced as the main mathematical tool used to represent the fitted gradient-boosted model in an interpretable form.

2.1 Credit-Risk

This section introduces the credit-risk setting considered in the present thesis. The focus is on probability of default modelling, and in particular on why both predictive performance and interpretability are important when such models are used in a banking context.

Credit-risk refers to the risk of financial loss arising from a borrower failing to fulfil its contractual obligations towards a lender. Such losses may occur if the borrower (also referred to as the obligor or counterparty) is unable to repay the principal or interest owed on a loan in time, or deliberately refuses to pay. From the perspective of a lender, such as a bank, it is therefore important not only to assess the likelihood of such an event occurring, but also the potential magnitude of the resulting losses.

For banks, the primary source of credit risk originates from lending activities. However, credit risk may also arise from other financial exposures, such as derivative contracts including swaps and options. These exposures are typically modelled using different methods than traditional credit-risk models for loan portfolios, and will therefore not be considered further in this thesis. [1]

There are several ways to measure and quantify credit risk, but one of the fundamental quantities is the probability that a borrower defaults on its obligations, commonly referred to as the probability of default (PD). Since default is observed as a binary outcome, PD modelling is closely related to binary classification problems.

However, in many applications the objective is not merely to classify borrowers as defaulting or non-defaulting, but to estimate the conditional probability of default itself. Such information may then be used when determining interest rates, pricing, collateral requirements, and other contractual terms.

Historically, credit assessment was pioneered by rating agencies (Anderson, 2021). However, banks have increasingly developed their own internal credit-risk models rather than relying solely on external ratings to assess the credit quality of counterparties. Credit-risk models provide several important benefits for banks. According to Basel Committee on Banking Supervision (1999), such models support more consistent and centralized risk assessment, improved measurement of portfolio and concentration risk, and greater responsiveness to changes in credit quality and economic conditions. They may additionally improve pricing, limit setting, reserve estimation, and portfolio management.

[1] It is worth noting that while both loan portfolios and other counterparty credit exposures constitute credit risk from a banking and regulatory perspective, much of the financial mathematics literature on credit risk focuses on the latter, particularly in the context of pricing traded instruments using continuous-time models. In contrast, the modelling of default risk for loan portfolios is typically framed as a prediction problem and is often referred to as credit scoring or credit rating.

However, arguably the most important use of credit-risk models in modern banking is their role in determining regulatory capital requirements under the Basel framework. While the Basel I framework relied largely on standardized risk weights, Basel II introduced the internal ratings-based (IRB) approach, under which banks may use internal risk estimates as inputs to capital calculations. Under the Foundation IRB approach, banks estimate the probability of default internally, while other risk parameters are prescribed by the regulator. Under the Advanced IRB approach, banks may additionally estimate quantities such as loss given default (LGD) and exposure at default (EAD).

The introduction of the IRB framework substantially increased the importance of internal credit-risk models, since model outputs became directly linked to regulatory capital requirements. Later Basel III reforms, introduced after the financial crisis of 2007–2008, further strengthened capital requirements and model-governance expectations, reflecting increased regulatory concern with model risk and variability in internally estimated risk measures.

Note that the discussion up to this point has discussed the PD model as if it directly produces the final PD estimate. However, within regulatory capital frameworks such as the IRB framework, banks are generally required to estimate the *long-run* probability of default. It is natural that the probability of default is not only influenced by borrower-specific characteristics, but also by the prevailing macroeconomic environment, with some periods resulting in substantially higher default frequencies than others.

Consequently, regulatory PD modelling is often conceptually divided into two separate stages. The first stage is the *risk differentiation* stage, where the objective is to construct a model that is highly capable of distinguishing between high-risk and low-risk borrowers based on observed borrower characteristics and default behaviour in the development sample. This stage primarily focuses on relative risk ranking.

The second stage is the *calibration* or quantification stage, where the model outputs are calibrated, often using macroeconomic information or long-run portfolio statistics, such that the resulting PD estimates align with the long-run observed default frequency. However, how such calibration procedures should be performed is a large topic in itself and falls outside the scope of this thesis. Therefore, the present thesis is solely focused on the risk differentiation stage of PD modelling.

Credit exposures are commonly divided into retail and non-retail credit. Retail credit primarily consists of lending to private individuals, while non-retail credit primarily concerns lending to businesses.[2]

While banks are ultimately interested in estimating the long-run probability of default, the model outputs for individual counterparties are typically expressed in the form of a *rating* for non-retail counterparties or a *score* for retail counterparties.

Although retail and non-retail models may utilize similar statistical methods, important differences exist with respect to the type of information available. Retail credit-risk models have traditionally relied heavily on behavioural and transactional information, whereas non-retail models rely more strongly on financial statement information, ratio analysis, and expert assessment (Anderson, 2021).

Compared to retail credit portfolios, non-retail portfolios are also generally charac-

[2]Non-retail credit also includes portfolios involving more specialized counterparties, such as financial institutions, sovereigns, or other institutional counterparties. These portfolios are often characterized by very small numbers of defaults, or in some cases no observed defaults at all, and are therefore commonly referred to as low-default portfolios (LDPs). Such portfolios typically require substantially different modelling methodologies and are outside the scope of this thesis.

terized by lower granularity, fewer defaults, and substantially smaller datasets. As a consequence, non-retail modelling often faces challenges associated with limited default observations and low-default settings.

However, important differences also exist within the non-retail category itself. In particular, SME (small and medium-sized enterprise) portfolios may exhibit substantially higher default frequencies and greater portfolio granularity than large-corporate portfolios, making them in some respects statistically more similar to retail portfolios. Nevertheless, non-retail models, including SME models, are still generally more reliant on financial statement information and ratio analysis than retail credit-scoring models, which rely more heavily on behavioural, transactional, and bureau data (Anderson, 2021).

The present thesis focuses on non-retail PD modelling.

Historically, corporate credit-risk assessment evolved from expert judgement toward more quantitative approaches, first through financial-ratio analysis and later through statistical models for bankruptcy prediction (Altman, 1968; Bellovary et al., 2007). Early work such as Beaver (1966)'s univariate ratio analysis and Altman (1968)'s Z-score model demonstrated the predictive value of financial ratios, while Ohlson (1980) introduced one of the first applications of logistic regression for bankruptcy prediction.

Because default is observed as a binary outcome, PD modelling is naturally formulated as a probabilistic classification problem. In traditional PD models, the conditional probability of default is commonly modelled as a function of borrower characteristics, often referred to as risk drivers. The dominant statistical specification is logistic regression, where the log-odds of default are assumed to be linear in the predictors.

Note that modern credit-risk systems are typically considerably more complex than a single statistical model. Different modules consisting of different categories of risk drivers, such as financial, behavioural, or qualitative information, may be estimated separately and subsequently combined. Furthermore, expert judgement adjustments or overrides may also be applied to increase or decrease a rating, although such adjustments do not necessarily alter the underlying PD estimate itself (Izzi et al., 2019; Anderson, 2021).

Recently, modern machine learning models[3] have seen increasing use in credit-risk research, where they have often demonstrated improved predictive performance compared to traditional statistical approaches (Ayari et al., 2025). However, despite these promising results, such models have so far seen relatively limited uptake in production credit-risk systems.

There is no explicit regulatory requirement for institutions to use a particular statistical or machine learning modelling approach when developing IRB models. In principle, any modelling approach may be used provided that the resulting models satisfy the regulatory requirements regarding, among other things, discriminatory power, calibration, governance, validation, and documentation.

Nevertheless, despite the increasing success of machine learning methods in academic credit-risk research, institutions have historically been cautious in adopting such methods within regulatory IRB frameworks. One important reason is that many modern machine learning models are substantially more complex and less interpretable than traditional approaches such as logistic regression. While such models may improve predictive per-

[3]The distinction between "machine learning" models and more traditional statistical models is not always clear-cut. In its discussion paper on machine learning for IRB models, the European Banking Authority (EBA) notes that methods such as linear and logistic regression may technically fall under broad definitions of machine learning, but that in financial practice the term is typically reserved for more complex models characterized by large numbers of parameters, substantial data requirements, and the ability to capture highly non-linear relationships (European Banking Authority, 2021).

formance, they may simultaneously make it more difficult to understand the relationship between borrower characteristics and predicted risk.

Both supervisory authorities and the banking industry have identified this trade-off between predictive performance and interpretability as one of the central challenges associated with the use of machine learning methods in regulatory credit-risk modelling. In its discussion paper on machine learning for IRB models, the European Banking Authority (EBA) notes that while machine learning models may improve predictive power, they may also create substantial challenges with respect to interpretability, governance, validation, documentation, and human understanding of the models (European Banking Authority, 2021). Similarly, the European Central Bank (ECB) has emphasized that institutions using machine learning techniques in internal models are expected to maintain adequate explainability, auditability, validation procedures, and governance frameworks, particularly for highly complex or weakly interpretable models (European Central Bank, 2025).

Consequently, the introduction of increasingly complex machine learning models into credit-risk modelling has also increased the importance of model interpretability.

2.2 Statistical Learning

This section introduces the statistical-learning framework used to describe the predictive models considered in the present thesis. This framework also provide the basis for the later discussion of model complexity and interpretability. The discussion is limited to the concepts needed in the following sections; for a more comprehensive treatment of the theory and practice of statistical learning, see Hastie et al. (2009).

Following Hastie et al. (2009), statistical learning studies methods for learning patterns and relationships from data. In supervised learning, a set of predictor variables, also referred to as inputs or features, is observed together with one or more response variables, also referred to as outputs or target variables. In the present thesis, these correspond to borrower characteristics and default outcomes, respectively.

The objective is to learn the relationship between the predictors and the response in order to predict future outcomes or better understand the relationship between them. Conceptually, one may think of this as assuming the existence of some unknown function relating the predictors to the response, which we seek to approximate using observed data.

Broadly speaking, supervised learning problems may be divided into regression and classification problems.

In regression problems, the response variable is quantitative, typically represented as $Y \in \mathbb{R}$. The objective is then to predict a numerical outcome from a set of predictors. In classification problems, the response instead takes values in a discrete set $\mathcal{S}$. In this thesis, we are primarily concerned with the binary case

$$Y \in \{0, 1\},$$

where the objective is to predict whether a counterparty will default. As previously discussed in the context of credit-risk modelling, we are often interested not only in the predicted class label itself, but also in estimating the conditional probability

$$p(x) = \mathbb{P}(Y = 1 \mid X = x).$$

Following Hastie et al. (2009), the supervised learning problem may be formalized as follows. Let X denote a random input vector of dimension p and let Y denote a response

variable with joint distribution $\mathbb{P}_{X,Y}$. We seek to estimate a function $\hat{f}$, where $\hat{f}(X)$ approximates the unknown relationship between X and Y. To quantify the quality of such an approximation, we introduce a loss function $L(Y, \hat{f}(X))$ measuring the discrepancy between the observed response and the prediction. The objective is then to find a function f^* satisfying

$$f^* = \arg\min_f \ \mathbb{E}\Big[L(Y, f(X))\Big]. \tag{2.1}$$

The choice of loss function determines what constitutes a good prediction and depends both on the characteristics of the response variable and on the intended use of the predictions. For example, in regression problems with a quantitative response variable, losses based on numerical distance between observed and predicted values are common, with squared error loss being the most widely used example. Such losses may also target different aspects of the conditional distribution of $Y \mid X$. For example, squared error loss targets the conditional mean, whereas other loss functions may target medians or quantiles (Hastie et al., 2009).

For pure classification, a natural criterion is the misclassification error. However, this loss depends only on the predicted class label and does not account for the uncertainty of the prediction. In credit-risk modelling, the relevant object is instead the conditional probability of default, or at least a score that ranks counterparties by relative default risk.

For probability estimation it is desirable that the loss function is a proper scoring rule[4]. A scoring rule is said to be proper if the expected loss is minimized when the predicted probabilities coincide with the true conditional probabilities (Buja et al., 2005).

The most common loss function for probabilistic binary classification is the cross-entropy loss (also commonly referred to as log-loss), defined as

$$L(y, p(x)) = -y \log p(x) - (1 - y) \log(1 - p(x)). \tag{2.2}$$

In practice, the joint distribution $\mathbb{P}_{X,Y}$ is unknown and only a finite sample of observations is available. The expected loss is therefore typically approximated using the empirical loss. The supervised learning problem thus becomes the problem of finding a function minimizing the empirical loss over the observed sample.

However, minimizing the empirical loss over the observed sample does not necessarily imply good predictive performance on new unseen observations. If the class of admissible functions is essentially unrestricted, there may exist many functions that fit the observed training data well, while behaving very differently away from the observed data points (Hastie et al., 2009). Such functions may fit not only the underlying relationship between the predictors and response, but also random fluctuations and noise specific to the observed sample, a phenomenon commonly referred to as overfitting. Since the objective is predictive performance on unseen data, model performance is commonly estimated using validation data or resampling procedures such as cross-validation, rather than relying only on the training loss.

In order to obtain useful predictions from finite data, the set of admissible functions must therefore be restricted in some way. Such restrictions may be imposed explicitly, for example through a parametric model specification, or implicitly through the learning algorithm itself (Hastie et al., 2009). Regularization provides another common approach, where constraints or penalty terms are introduced to discourage overly complex fitted

[4]Even for classification, misclassification error tends to be unstable and proper scoring rules are often utilized as surrogate criteria (Buja et al., 2005).

functions. This reflects the trade-off between model flexibility and generalization performance: restrictive model classes may reduce variance and improve interpretability, but may fail to capture important nonlinearities or interactions, while more flexible model classes can represent more complex relationships but often require more data and are more difficult to interpret.

A classical example of a restrictive model class for binary probability estimation is logistic regression. Let $Y_i \in \{0, 1\}$ denote the binary response and let

$$p(\mathbf{x}_i) = \mathbb{P}(Y_i = 1 \mid X_i = \mathbf{x}_i)$$

denote the conditional probability of default. In logistic regression, the response is modelled as

$$Y_i \mid X_i = \mathbf{x}_i \sim \text{Bernoulli}(p(\mathbf{x}_i)),$$

where the probability is linked to the predictors through the logistic function

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}. \tag{2.3}$$

Equivalently, the log-odds are assumed to be linear in the predictors:

$$\log\left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{2.4}$$

The parameter vector $\boldsymbol{\beta}$ is commonly estimated by maximum likelihood. Equivalently, this corresponds to minimizing the empirical cross-entropy loss.

This linear log-odds structure makes logistic regression relatively simple and interpretable, which has contributed to its widespread use in credit-risk modelling. At the same time, it makes the model restrictive, since nonlinearities and interactions must be specified explicitly through transformations or interaction terms.

This distinction between restrictive and flexible model classes is closely related to the distinction between what Breiman (2001) refers to as the two cultures of statistical modelling. In the traditional data-modelling culture, one assumes a relatively structured form for the relationship between predictors and response, and the fitted model is often interpreted in terms of this assumed structure. Logistic regression is a typical example of such an approach.

In contrast, the algorithmic modelling culture places greater emphasis on predictive performance and treats the data-generating mechanism as complex and potentially unknown. Models in this tradition typically impose weaker functional-form assumptions and may therefore capture nonlinearities and interactions more automatically. This increased flexibility may improve predictive performance, but it also makes the resulting fitted function more difficult to interpret.

Decision trees provide one way of constructing such flexible function classes by recursively partitioning the predictor space. Ensemble methods, and in particular gradient-boosted decision trees, combine many such trees into a more powerful predictive model. The following section therefore introduces gradient-boosted decision trees, which form the predictive model class studied in this thesis.

2.3 Gradient-Boosted Decision Trees

The previous section described supervised learning as the problem of estimating a prediction function from data. This section considers a particular class of flexible predictive models: gradient-boosted decision trees.

The construction begins with decision trees, which recursively partition the predictor space through binary splits and assign a constant prediction within each terminal region. Each internal node corresponds to a decision rule, while each terminal node assigns a prediction to observations satisfying the corresponding sequence of rules. For classification trees, this prediction is typically a class label, although terminal nodes may also be used to assign class probabilities. A schematic illustration is shown in Figure 1.

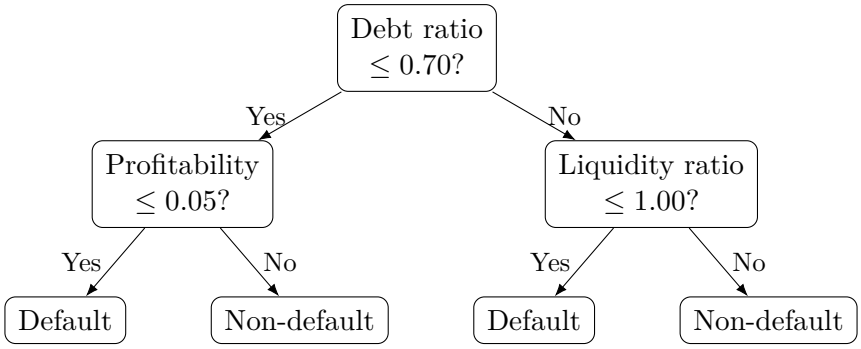


Figure 1: Schematic illustration of a classification tree for default prediction. Each internal node represents a binary decision rule, while each terminal node assigns a predicted class label to observations satisfying the corresponding sequence of rules.

In general datasets, the number of possible tree partitions grows rapidly and the tree must be estimated from data. Following Breiman et al. (1984), this involves choosing splits, deciding when to stop splitting, and assigning predictions to terminal nodes. Splits are typically chosen greedily, by comparing candidate binary splits according to the reduction they produce in a chosen criterion, such as the Gini index or cross-entropy for classification trees and within-node squared error for regression trees (Hastie et al., 2009). Tree complexity is controlled through stopping rules or pruning, while terminal nodes are assigned predictions such as majority classes, empirical class probabilities, or within-node averages.

Decision trees are attractive because they require no global parametric specification and because individual predictions can be traced through explicit decision rules. However, single trees are often unstable, in the sense that small changes in the training data may lead to substantially different tree structures (Hastie et al., 2009). Moreover, capturing complex relationships may require growing the tree so large that its interpretability is weakened. This motivates tree ensembles, where many trees are combined to obtain a more stable and flexible prediction function.

An ensemble method combines several simpler models to form a final prediction function. For tree ensembles, two important families are bagging-type methods, such as random forests, and boosting-type methods. The present thesis focuses on boosting, where the ensemble is built sequentially so that each new tree improves on the previous one.

The theoretical origin of boosting is often traced to Schapire (1990), who showed that several weak learners may be combined to obtain a substantially more accurate predictor. A practical and highly influential realization of this idea was introduced by Freund and Schapire (1997) through AdaBoost, which sequentially fits weak classifiers while giving greater weight to difficult observations. When combined with decision trees, AdaBoost was shown to achieve substantially lower error rates than a single tree (Friedman et al., 2000).

Although early boosting methods were originally formulated in terms of reweighting

observations, later work gave boosting a statistical interpretation as additive modelling and loss minimization. This interpretation is related to generalized additive models, where the linear predictor in a generalized linear model is replaced by a sum of component functions (Hastie and Tibshirani, 1990). For a binary response, an additive logistic model may for example be written as

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \sum_{j=1}^{p} f_j(x_j),$$

where the functions f_j describe potentially nonlinear main effects.

This also motivated LogitBoost, which fits an additive logistic model more directly using the binomial log-likelihood criterion (Friedman et al., 2000). This perspective was generalized by Friedman (2001), who formulated boosting as greedy function approximation. In this framework, a prediction function F is estimated by sequentially adding simple base learners in order to reduce an empirical loss. After M iterations, the fitted model can be written as

$$F_M(x) = F_0(x) + \nu \sum_{m=1}^{M} h_m(x),$$

where F_0 is an initial prediction, h_m is the base learner added at iteration m, and ν is the learning rate or shrinkage parameter controlling the size of each update. In classification problems, the fitted function $F_M(x)$ should not necessarily be interpreted as a class label itself. For binary classification with logistic loss, $F_M(x)$ is instead a real-valued score, commonly interpreted on the log-odds scale. The corresponding predicted probability is obtained by applying the logistic transformation,

$$\hat{p}(x) = \frac{1}{1+\exp\{-F_M(x)\}}.$$

Thus, in probabilistic binary classification, the gradient-boosted model is fitted through an additive score function, while the probability prediction is obtained only after applying the logistic transformation.

At each step, the next base learner approximates the direction in which the current fitted function should be changed to reduce the loss. In gradient boosting, this direction is given by the negative gradient of the loss evaluated at the current fitted values. These quantities are often called pseudo-responses or pseudo-residuals, since they serve as the response values when fitting the next base learner; for squared-error loss they coincide with ordinary residuals (Friedman, 2001).

Hastie et al. (2009) describe how gradient boosting may be implemented using decision trees as base learners. In this setting, each boosting iteration fits a regression tree to the current pseudo-responses. Since each boosted tree only provides one incremental update to the current fitted function, the trees are typically restricted to have a fixed small size rather than being grown as full standalone trees.

Let J denote the number of terminal nodes in each tree. This parameter controls the complexity of the individual base learners. Since a tree with J terminal nodes has $J-1$ splits, it can represent interactions of order at most $J-1$. For example, $J=2$ corresponds to decision stumps and gives an additive model with only main effects. Setting $J=3$ allows two-way interactions (depth-2), while $J=4$ allows interactions up to order three. In practice, relatively small trees are often sufficient, and Hastie et al. (2009) note that values between four and eight terminal nodes are commonly adequate.

Several regularization mechanisms are used to control the complexity of gradient-boosted tree models. The number of boosting iterations M controls how many trees are added; increasing M allows the model to fit the training data more closely, but may eventually lead to overfitting. It is therefore commonly selected using validation data, cross-validation, or early stopping. Another important parameter is the learning rate, or shrinkage parameter, ν, which scales each tree update. Smaller values of ν make the model learn more slowly and typically require more boosting iterations, but often improve generalization. Additional regularization may be introduced through stochastic gradient boosting, where each tree is fitted using a random subsample of the training observations.

The gradient-boosted tree procedure may therefore be understood schematically as in Figure 2.

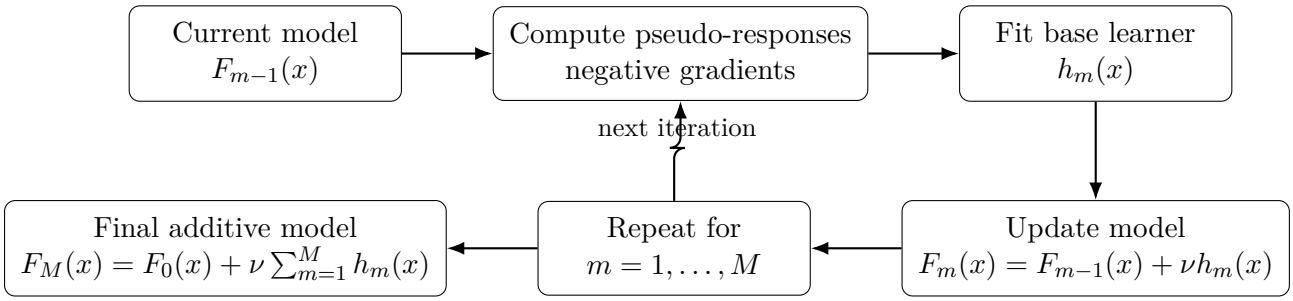


Figure 2: Schematic illustration of gradient boosting. At each iteration, pseudo-responses are computed from the current fitted model, a new base learner is fitted, and the model is updated. Repeating this procedure yields an additive ensemble.

When the fitted tree at iteration m has terminal regions

$$R_{1m}, \ldots, R_{Jm},$$

it can be written as

$$h_m(x) = \sum_{j=1}^{J} \gamma_{jm} \mathbb{I}(x \in R_{jm}),$$

where γ_{jm} is the value assigned to terminal region R_{jm}. With shrinkage, the boosted model is updated according to

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J} \gamma_{jm} \mathbb{I}(x \in R_{jm}).$$

Thus, after M boosting iterations, the fitted model is an additive ensemble of tree functions. Although each individual tree is simple, the full ensemble is generally difficult to interpret directly. The effect of a single predictor may be distributed across many trees and split points, and interactions may arise whenever a tree contains splits on more than one predictor.

Many implementations of gradient-boosted trees have been developed. One particularly influential implementation, and the one used in this thesis, is XGBoost (Chen and Guestrin, 2016). XGBoost is designed as a scalable system for fitting gradient-boosted tree models, using computational techniques such as parallelization and approximate split search to improve efficiency. It also augments the gradient-boosting objective with explicit regularization terms that penalize model complexity. In addition to row and column subsampling, XGBoost uses efficient approximate split-search algorithms and sparsity-aware handling of missing values.

These implementation details are important for the practical success of XGBoost, but they are not the main focus of the present thesis. For the purposes of the following analysis, the central point is that XGBoost fits a regularized additive ensemble of decision trees.

Although gradient boosting has been described above as a form of function approximation, this should not be interpreted as saying that boosted tree models automatically converge to an arbitrary true function. Rather, the target of estimation is constrained by the class of base learners, the loss function, and the regularization used. In the idealized population setting, one may therefore view the fitted function as targeting the best approximation to the unknown regression or score function within the function class made available by the model.

Recent theoretical work has studied XGBoost from this function-class perspective. Ki and Guntuboyina (2026) show that XGBoost can be understood as estimating functions within a structured class generated by sums of bounded-depth regression trees, together with a complexity measure related to the algorithm's regularization.This supports viewing XGBoost as a regularized method for estimating functions within a particular tree-based function class. Consequently, even with unlimited data, the relevant limiting object is not necessarily the true data-generating function itself, but the best approximation to the target within this regularized class. Although these results are developed for squared-error regression, the perspective remains conceptually relevant here because binary gradient-boosted tree classification models still define a real-valued score function, later transformed to a probability through the logistic function.

2.4 Interpretability

As discussed in the previous section, gradient boosting tree models such as XGBoost can form highly flexible prediction functions and often achieve strong predictive performance. However, this flexibility comes at the cost of transparency. While an individual decision tree can often be inspected directly, an ensemble consisting of many sequentially fitted trees is substantially more difficult to understand. The fitted prediction function is mathematically well defined, but its behaviour is generally not apparent from the model representation itself.

This is a common difficulty in modern machine learning. Many flexible models obtain strong predictive performance precisely by learning complex nonlinear relationships and interactions, but these same properties can make the resulting models difficult to interpret. Such models are therefore often described as black-box models, motivating a large literature concerned with making complex model behaviour more transparent.

Since related terms such as interpretability, explainability, and explainable artificial intelligence are not used uniformly in this literature, the present thesis uses the term interpretability in a broad sense. Following Molnar (2022), interpretability is taken to concern the extent to which model behaviour and individual predictions can be made accessible to human users.

Interpretability can serve several purposes. Molnar (2022) distinguishes, among other things, between interpretability for model justification and for knowledge discovery. In the first case, interpretations help stakeholders assess whether the fitted model behaves reasonably and can be trusted for decision support. In the second case, they may reveal relationships in the data that were not known in advance. Interpretability can also support model debugging and validation by revealing implausible effects, unexpected interactions,

or patterns that suggest data-quality issues.

The importance of interpretability depends on the application. As discussed by Molnar (2022), explanations may be less necessary in low-stakes settings or for well-studied tasks where the consequences of individual errors are limited. This is clearly not the case for the credit-risk setting. Interpretability is therefore both a practical and regulatory concern: developers and validators must be able to assess whether the model has learned plausible risk relationships, while institutions must be able to document, challenge, monitor, and justify the model for its intended use.

Interpretability may be achieved in different ways. Molnar (2022) distinguishes between inherently interpretable models and post-hoc interpretability methods. Inherently interpretable models are models whose structure is understandable by design, usually because the model class has been restricted so that the fitted model remains sufficiently simple or structured to inspect directly. Examples include classical statistical models such as linear and logistic regression, as well as small decision trees.

However, inherent interpretability is not a binary property. Molnar (2022) discusses several levels at which a model may be interpretable. A model may be interpretable as a whole, meaning that the complete fitted model can be inspected directly. This is typically only realistic for very simple models. Alternatively, individual components may be interpretable even when the full model is too large to understand at once, as in a regression model with many coefficients. Finally, individual predictions may be interpretable if the reasoning behind a specific prediction can be followed, for example by tracing the path of an observation through a decision tree.

Gradient boosting models such as XGBoost are generally not inherently interpretable in this sense. Although each individual tree may be simple, the final prediction is obtained by combining the contributions of many sequentially fitted trees, making the ensemble as a whole difficult to inspect directly. This motivates the use of post-hoc interpretability methods, which are applied after a model has been fitted in order to explain, summarize, or visualize the behaviour of the prediction function. Many such methods are model-agnostic, meaning that they require only access to model inputs and predictions rather than the internal structure of the model.

Post-hoc methods answer different interpretability questions. Local methods aim to explain individual predictions, with LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) being common examples. Global methods instead summarize the behaviour of the fitted model more broadly. For example, partial dependence plots (Friedman, 2001) and accumulated local effects plots (Apley and Zhu, 2020) visualize how predictions change as selected features vary. Functional decomposition methods, introduced in the next section, may also be viewed as global interpretability methods, since they aim to represent the fitted prediction function through lower-dimensional components.

A further distinction is between feature-effect methods and feature-importance methods (Molnar, 2022). Feature-effect methods describe the relationship between selected input variables and the model prediction, while feature-importance methods rank variables according to their influence on the model.

One influential approach to feature attribution is based on the Shapley value. The Shapley value originates from cooperative game theory, where it was proposed by Shapley (1953) as a solution for distributing the total gains from cooperation among players in a coalition game. Its central idea is to allocate value to each player according to their average marginal contribution across possible coalitions.

Let N denote the set of players, with $n = |N|$, and let $v : 2^N \to \mathbb{R}$ be a characteristic

function assigning a value to each coalition. The Shapley value for player $i \in N$ is given by

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup \{i\}) - v(S) \right]. \tag{2.5}$$

It is therefore a weighted average of the marginal contribution of player i across all coalitions not containing i. The Shapley value is uniquely characterized by axioms such as efficiency, symmetry, the dummy-player property, and additivity, which makes it a natural allocation rule.

Lundberg and Lee (2017) apply this allocation principle to machine-learning interpretability through SHAP. Features are treated as players and the value function is based on the model prediction. For a given observation, SHAP values allocate the difference between the prediction and a baseline across input variables. They are therefore primarily local explanations, since they describe how a specific prediction is attributed to individual features. A practical challenge is that exact Shapley values require evaluating all feature subsets, leading to exponential complexity. Lundberg and Lee (2017) address this by proposing approximation methods and model-specific algorithms, including efficient implementations for tree-based models.

Although SHAP values are local explanations, they are often aggregated across observations to obtain global feature-importance summaries. A common measure is mean absolute SHAP importance,

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_j^{(i)}|, \tag{2.6}$$

where $\phi_j^{(i)}$ denotes the SHAP value of variable j for observation i. This ranks variables according to the typical magnitude of their local contributions across the data.

However, such importance summaries do not describe the shape of the fitted prediction function. A variable may receive high importance because it often contributes substantially to predictions, but this does not show whether its effect is increasing, decreasing, nonlinear, threshold-like, or mainly expressed through interactions. Thus, SHAP feature importance is useful for ranking variables, but it does not by itself provide the type of functional representation considered in this thesis.

For this reason, feature-attribution methods are often complemented by feature-effect methods such as partial dependence plots and accumulated local effects plots. Partial dependence plots describe how the average model prediction changes as selected features are varied, while the remaining features are averaged over (Friedman, 2001). They can therefore visualize marginal feature effects and, in the two-dimensional case, interaction patterns. However, when predictors are dependent, partial dependence plots may evaluate the model at rare or unrealistic feature combinations.

Accumulated local effects plots address this issue by summarizing local changes in the fitted prediction function over regions where observations are available (Apley and Zhu, 2020). ALE is also closely related to functional decomposition, since first-order ALE functions may be interpreted as estimated main-effect components, while higher-order ALE functions describe interactions.

The methods discussed above can provide useful summaries of a fitted model, but they do not by themselves make the model inherently interpretable. When an explanation is produced by an additional procedure applied to an otherwise complex prediction function, the explanation may itself be regarded as a secondary model of the original model. It must therefore be assessed not only in terms of whether it is understandable, but also in

terms of whether it is robust and faithful to the model being explained.

This issue has led to criticism of relying on post-hoc explanations in high-stakes applications. Rudin (2019) argues that, when decisions have serious consequences, one should not use black-box models and then explain them afterwards, but should instead use models that are interpretable by design whenever possible. A central concern is that a post-hoc explanation may be plausible or useful without being fully faithful to the model being explained. If the explanation is not faithful, it risks describing an approximation to the model rather than the model itself. If it is fully faithful, this raises the question of why the interpretable representation could not be used more directly.

This concern is also reflected in work on interpretable machine learning for regulated and high-risk applications. Sudjianto and Zhang (2021) argue that inherently interpretable machine-learning models become especially important in sectors related to health, safety, fundamental rights, and banking, where models may be subject to legal, regulatory, and model-risk-management requirements. They note that black-box models explained through model-agnostic methods may be more difficult to defend under regulatory scrutiny, since transparency and explainability are important for assessing conceptual soundness and effective model risk management.

In the present credit-risk setting, this issue is closely related to the regulatory expectations discussed in Section 2.1. The ECB's guide to internal models does not exclude the use of machine-learning techniques in internal models (European Central Bank, 2025). However, it emphasizes that increased model complexity must be justified and that institutions must maintain sufficient explainability for model development, validation, monitoring, and human judgement.

These expectations include both global and local interpretability. At the global level, institutions should be able to assess whether the relationship between risk drivers and model estimates is intuitive and economically plausible. At the local level, they should be able to understand the contribution of individual key drivers to specific estimates. This corresponds closely to the distinction introduced above between global interpretability, which concerns the behaviour of the fitted model as a whole, and local interpretability, which concerns individual predictions.

The ECB's expectations do not rule out the use of post-hoc explainability tools. Indeed, such tools may support several of the required tasks: global feature-effect methods can help assess the plausibility of risk-driver relationships, while local attribution methods can help explain individual estimates. However, the ECB also emphasizes that explainability techniques and tools should be defined and documented, including their evaluation criteria, weaknesses, limitations, and the explainability requirements of different stakeholders. Thus, when an explanation is produced by a separate procedure applied to an otherwise opaque model, the explanation itself becomes an additional object that must be justified.

This suggests that, in a regulatory credit-risk setting, interpretability is easier to defend when it is closely connected to the fitted prediction function itself. An inherently interpretable model, or a model whose fitted prediction function can be represented directly in an interpretable form, can more naturally support both global and local analysis because the relevant risk-driver relationships are part of the model representation rather than only external explanations.

To make this notion more concrete, Sudjianto and Zhang (2021) propose a qualitative template for assessing the inherent interpretability of machine-learning models. The template evaluates interpretability through structural characteristics such as additivity

or modularity, sparsity, linearity, smoothness, monotonicity, visualizability, projection, and segmentation. These criteria are useful in the present thesis because they provide a vocabulary for discussing whether the proposed functional-decomposition approach supports model understanding. They will therefore be returned to later when evaluating the extent to which the functional ANOVA representation makes the fitted XGBoost model interpretable.

The discussion above motivates the approach taken in this thesis. Rather than treating interpretability only as an external explanation applied after model fitting, the aim is to study whether the fitted XGBoost score function can itself be represented through components that support global understanding, local attribution, and visual inspection. The next section introduces functional ANOVA decomposition as the mathematical framework used to construct such a representation.

2.5 Functional ANOVA Decomposition

The previous section argued that interpretability in a regulatory credit-risk setting requires more than post-hoc summaries of model behaviour. In particular, an explanation should remain closely connected to the fitted prediction function and support both global assessment of risk-driver relationships and local understanding of individual predictions. This motivates studying whether the fitted prediction function itself can be represented in an interpretable form.

The approach considered in this thesis is based on functional decomposition, where a multivariate function is represented as a sum of lower-dimensional component functions. Let $F(\mathbf{x})$ be a function of $\mathbf{x} = (x_1, \ldots, x_p)$. A functional decomposition of F is a representation of the form

$$F(\mathbf{x}) = \sum_{u \subseteq \{1,\ldots,p\}} f_u(\mathbf{x}_u),$$

where each component f_u depends only on the subset of variables indexed by u. These components may then be interpreted as a baseline term, main effects, interaction effects, and, in principle, higher-order interactions.

For notational convenience, we assume that each subset u is associated with a single component function, since multiple functions depending on the same subset can always be combined. This representation is attractive from an interpretability perspective because it describes not only which variables are important, but also how individual variables and interactions contribute to the fitted prediction function. It is therefore similar in spirit to generalized additive models, but is not restricted to additive main effects. Instead, it can also represent interactions through components depending on multiple variables.

However, such a decomposition is not unique in general. Multiple distinct sets of component functions may yield exactly the same overall function while implying different interpretations of the roles of the individual variables. This immediately raises the question of how the component functions should be chosen.

To illustrate this ambiguity, consider the function

$$y = a + bx_1 + cx_2 + dx_1x_2. \tag{2.7}$$

At first glance, the decomposition appears straightforward, and one might argue that the representation is already directly provided by the equation itself. However, for arbitrary constants α and β, the same function can equivalently be written as (Lengerich

et al., 2020)

$$y = (a - d\alpha\beta) + (b + d\beta)x_1 + (c + d\alpha)x_2 + d(x_1 - \alpha)(x_2 - \beta). \tag{2.8}$$

While all such decompositions represent exactly the same function, the implied interpretations of the main and interaction effects differ substantially. For example, setting $\beta = -\frac{b}{d}$ removes the main effect of x_1, implying that the effect of x_1 acts entirely through interaction with x_2.

Thus, even in this simple example, the distinction between main effects and interactions depends on how the decomposition is chosen. One might object that the original algebraic form still provides a natural decomposition. Indeed, for linear models this corresponds to one possible rule for selecting the decomposition components. However, most machine-learning prediction functions do not possess such a simple analytical structure, and even simple functions may admit competing interpretations.

To illustrate this further, consider the function

$$y = \log(x_1 x_2), \qquad x_1, x_2 > 0.$$

Under one decomposition, this may be treated as a pure interaction effect with no main effects. On the other hand, rewriting the function as

$$y = \log(x_1) + \log(x_2)$$

instead yields a purely additive representation with no interaction effect at all.

These examples show that the distinction between main effects and interactions is not determined by algebraic form alone. A principled decomposition must therefore specify the criteria by which the component functions are chosen. Functional ANOVA provides one such framework.

The idea is closely related to the classical analysis of variance. In classical ANOVA, variation in a response is decomposed into contributions associated with different factors and their interactions. Functional ANOVA extends this idea from observed responses to functions, by decomposing a multivariate function into lower-dimensional components depending on subsets of its input variables. Early forms of such decompositions appear in the work of Hoeffding (1948). Later developments used spline and tensor-product spline spaces for functional ANOVA modelling (Stone, 1994), and studied projection estimation of regression functions onto functional ANOVA model spaces (Huang, 1998).

Functional ANOVA is also closely connected to sensitivity analysis, which studies how variation in a model output can be apportioned to variation in its inputs (Saltelli et al., 2008). In the present thesis, these inputs are the explanatory variables entering the fitted prediction function. This connection is useful because functional decomposition represents the fitted function through main effects and interactions, while sensitivity analysis provides a way to interpret these components in terms of how input variation is reflected in output variation. It also links functional ANOVA to interpretable machine learning, which Scholbeck et al. (2023) argue can be viewed as a form of sensitivity analysis applied to machine-learning models.

The discussion above motivates the need for low-order representations of complex functions. In practice, a full decomposition containing all possible interactions is rarely interpretable when the number of variables is large. The aim is therefore often to construct a representation that captures the most relevant functional behaviour using components of low dimensionality.

Hooker (2007) lists four desiderata for low-order representations of a function:

(a) **Comprehensibility**
(b) **Optimality**
(c) **Additivity**
(d) **Faithfulness to the underlying measure**

Comprehensibility requires the retained low-dimensional components to correspond to interpretable aspects of the function. Optimality requires the representation to capture as much relevant functional behaviour as possible without giving a misleading summary. Additivity requires truly additive functions to be recovered exactly as sums of their components. Faithfulness to the underlying measure requires the decomposition to respect the input distribution, rather than emphasizing unrealistic regions of the feature space.

A related principle is the preference for lower-order explanations when they are adequate. Yu et al. (2019) describe this as a form of reluctance toward interactions: all else equal, main effects should be preferred over interaction effects. This is consistent with the interpretability motivation above, since lower-order components are generally easier to visualize and understand. It also reflects the idea that interaction effects should represent behaviour that cannot be adequately explained by lower-order terms.

Functional ANOVA can now be understood as a particular rule for choosing the component functions in a functional decomposition. It does so by imposing centering and orthogonality conditions with respect to a chosen input measure.

Hooker (2004) gives the following definition of the functional ANOVA decomposition. Let $F(x) : \mathbb{R}^p \to \mathbb{R}$ be square integrable, with input vector $x = (x_1, \ldots, x_p)$. The functional ANOVA decomposition represents F as a sum of components of increasing dimensionality:

$$F(x) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{ij}(x_i, x_j) + \cdots \tag{2.9}$$

where f_0 is a constant term, $f_i(x_i)$ are main effects, and $f_{ij}(x_i, x_j)$ represent pairwise interaction effects and so on.

Let $[p] = \{1, \ldots, p\}$. Then, more generally, this decomposition can be written as

$$F(x) = \sum_{u \subseteq [p]} f_u(x_u), \tag{2.10}$$

where each component f_u depends only on the subvector x_u, $u \subset \{1, \ldots, p\}$.

The formal content of functional ANOVA lies in the additional conditions used to make this representation unique and interpretable. In the standard formulation, these conditions are centering, orthogonality, and the resulting variance decomposition.

Hooker (2007, p. 714) describe three desirable properties for the functional ANOVA decomposition. These are:

(i) **Zero Means:** $\int f_u(x_u) dx_u = 0$ for each $u \neq \varnothing$
(ii) **Orthogonality:** $\int f_u(x_u) f_v(x_v) dx = 0$ for $u \neq v$
(iii) **Variance Decomposition:** Let $\sigma^2(f) = \int f(x)^2 dx$ then

$$\sigma^2(F) = \sum_{u \subset [p]} \sigma_u^2(f_u) \tag{2.11}$$

The zero-mean condition ensures that each non-constant component is centered. Orthogonality ensures that different components do not contain overlapping variation. Together, these properties imply that the variability of the function can be decomposed into the variability contributed by the individual component functions.

The following decomposition, known as the "standard" functional ANOVA (Hooker, 2004, 2007) is defined recursively as:

$$f_u(x_u) = \int_{x_{-u}} \left(F(x) - \sum_{v \subset u} f_v(x_v) \right) dx_{-u}. \tag{2.12}$$

In this approach the functional ANOVA components are defined hierarchically: We subtract the lower order terms and project the resulting residual onto the subspace of functions depending only on x_u. This ensures a uniquely defined decomposition that satisfies all of the properties given above. Furthermore by defining it hierarchically we ensure that as much of the functional behaviour as possible is captured by the lower-order effects. The estimation procedure is also simple as it can be performed using simple Monte Carlo integration.

However, this particular decomposition comes with some caveats. Going back to Hooker (2007)'s four desiderata stated above, it is evident that the functional ANOVA satisfies the first three criteria. However, the same cannot be said of the fourth. Since the decomposition is taken under a product measure, it can be said to intrinsically assume that the underlying random variables are independent. Under strong dependence, the resulting decomposition can thus emphasize regions that have low probability mass (Hooker, 2007).

Hooker (2007, p. 715) thus proposed an alternative version of the functional ANOVA taken under a general measure $w(x)$. This alternative is known as generalized or weighted functional ANOVA and is defined as the projection of F onto a space of additive functions under the inner product.

$$< f, g >_w = \int_{\mathbb{R}^p} f(x)g(x)w(x)dx \tag{2.13}$$

All effects $\{f_u(x_u) | u \subset [p]\}$ are defined as the solution to

$$\{f_u(x_u) \mid u \subseteq [p]\} = \underset{\{g_u \in \mathcal{L}^2(\mathbb{R}^u)\}_{u \subseteq [p]}}{\arg\min} \int \left(\sum_{u \subseteq [p]} g_u(x_u) - F(x) \right)^2 w(x)dx \tag{2.14}$$

conditional on hierarchical orthogonality

$$\forall v \subset u,\ \forall g_v : \int f_u(x_u)g_v(x_v)w(x)dx = 0 \tag{2.15}$$

meaning that each component function f_u is orthogonal to the entire subspace of functions depending only on variables indexed by any subset $v \subset u$.

Furthermore by redefining the highest-order interaction term $f_{[p]}$ from (2.14) as the residual of F and all lower-order components

$$f_{[p]}(x) = F(x) - \sum_{u \subset [p]} f_u(x_u) \tag{2.16}$$

it is evident that (2.14) will always be minimized at 0. Hooker (2004) combines this fact with that the projection theorem (Luenberger, 1969) ensures that the residual term will be orthogonal to the subspace of lower-order additive functions, thereby satisfying (2.15). Any estimation procedure can thus be reformulated as solving for the lower-order components under the hierarchical orthogonality constraints, with the highest-order interaction term defined implicitly as the residual.

Although this constraint simplifies the potential estimation procedure somewhat, it is still not directly tractable, as it is formulated in terms of orthogonality with respect to all functions defined on lower-order subsets. In particular, it is not immediately clear how to verify or enforce this condition in practice, nor whether it uniquely determines the decomposition.

However, Hooker (2007, p. 716) introduces equivalent conditions through the following lemma:

Lemma 2.1. *The orthogonality conditions* (2.15) *are true over* $\mathcal{L}^2(\mathbb{R}^p)$ *if and only if the integral conditions*

$$\forall u \subseteq [p],\ \forall i \in u,\ \int f_u(x_u) w(x)\, dx_i\, dx_{-u} = 0 \tag{2.17}$$

hold.

Meaning that an equivalent requirement is that each component integrates to zero along each coordinate direction in u, with respect to the marginal distribution of w on x_u. The full proof is provided in Hooker (2007).

However, these conditions alone still do not guarantee uniqueness of the decomposition. For example, degenerate distributions can result in non-identifiability of the component functions. To ensure a unique solution, additional regularity conditions on w are thus required. Specifically, Hooker (2007) shows that under a grid closure of the support of w, the component functions are linearly independent and the decomposition is uniquely defined.

Before turning to the computational form used for tree models, it is useful to return to the sensitivity-analysis perspective introduced above. Functional ANOVA not only provides an interpretable representation of the fitted function, but also provides the basis for variance-based global importance measures such as Sobol indices and Shapley effects. As discussed earlier, mean absolute SHAP values summarize the typical magnitude of local feature attributions. Variance-based sensitivity analysis instead asks how much each variable contributes to the variability of the model output across the input distribution.

In the standard orthogonal functional ANOVA setting, the variance of the model output can be decomposed into contributions from the component functions. Let

$$V = \mathrm{Var}(F(X)), \qquad V_u = \mathrm{Var}(f_u(X_u)).$$

When the components are orthogonal,

$$V = \sum_{u \subseteq [p],\, u \neq \varnothing} V_u.$$

This leads to Sobol indices,

$$S_u = \frac{V_u}{V},$$

which measure the share of output variance attributable to the component indexed by u. For $u = \{j\}$, this is a main-effect contribution, while for $|u| > 1$, it is an interaction contribution (Saltelli et al., 2008).

Sobol indices therefore separate main-effect and interaction contributions, but they assign variance to subsets of variables rather than directly to individual variables. For example, a large interaction contribution $S_{\{j,k\}}$ shows that the joint behaviour of variables j and k is important, but does not determine how this contribution should be divided

between them. This creates an allocation problem when a single global importance value is desired for each variable.

One way to address this allocation problem is to apply the Shapley value to a variance-based set function. In the independent-input setting, Owen (2014) use variance contributions derived from the Sobol decomposition to construct such allocations. For dependent inputs, Song et al. (2016) define Shapley effects by applying the Shapley value to a cost function based on conditional variances. This cost function measures how much of the output variance is explained, or left unexplained, when different subsets of input variables are known. The resulting Shapley effect allocates the total output variance among the individual variables and therefore gives a single global sensitivity measure for each input variable.

Shapley effects should therefore be distinguished from the local SHAP values discussed earlier. SHAP values allocate the difference between an individual prediction and a baseline value, whereas Shapley effects allocate output variance across the input distribution. They are therefore global sensitivity measures rather than local prediction explanations.

Although generalized functional ANOVA addresses the issue of faithfulness to the underlying measure, it introduces other conceptual and computational difficulties. Apley and Zhu (2020) highlight several fundamental critiques of the method. In particular, while the hierarchical orthogonality constraints separate lower- and higher-dimensional effects, the same does not hold for lower-order effects when the input variables are dependent. As a result, if two variables are correlated, their lower-order effects may become entangled.

Even when using a decomposition that respects the joint distribution, the resulting main effects generally reflect both direct contributions as well as indirect contributions stemming from the dependence structure. Apley and Zhu (2020) argue that, in some contexts, one may instead prefer effects that capture the direct functional relationship between the fitted prediction function and each predictor. For example, in a model of the form $f(x) = x_1 + x_2$, the natural main effects would be $f_1(x_1) = x_1$ and $f_2(x_2) = x_2$, regardless of any dependence between x_1 and x_2. However, distribution-based decompositions will in general not recover these forms.

In addition to these conceptual concerns, Apley and Zhu (2020) also highlight challenges related to estimation. In the general case, the generalized functional ANOVA formulation of Hooker (2007) requires information about the joint distribution of the inputs, which is difficult to estimate in high-dimensional settings. Moreover, the components are not obtained recursively as in the standard functional ANOVA, but must instead be computed simultaneously by solving a constrained system of equations.

This computational difficulty is substantially reduced when F is piecewise constant on a finite partition of the input space (Lengerich et al., 2020). This is precisely the setting obtained for tree ensembles, where the split thresholds induce a finite set of bins for each feature and the fitted prediction function is constant on the resulting product partition. Let Ω_j denote the finite set of bins for feature j. For a piecewise-constant function F defined on this partition, the integral conditions in (2.17) reduce to finite weighted summation constraints:

$$\forall u \subseteq [p],\ \forall i \in u,\ \forall x_{u\setminus i},\quad \sum_{x_i \in \Omega_i} f_u(x_{u\setminus i}, x_i) \sum_{x_{-u}} w(x) = 0 \tag{2.18}$$

This finite representation makes the generalized functional ANOVA constraints much easier to enforce. Following Lengerich et al. (2020), each component f_u may be represented as a tensor over the bins corresponding to the variables in u. The condition in (2.18) then

requires each relevant tensor slice to have weighted mean zero.

The decomposition can therefore be obtained by purifying the component tensors. Starting from an initial decomposition, weighted slice means are iteratively subtracted from higher-order tensors and transferred to the corresponding lower-order components until the constraints in (2.18) are satisfied. For tree ensembles, this gives a practical way to obtain a functional ANOVA representation of the fitted prediction function. Importantly, the procedure reconstructs the fitted function exactly; it changes the representation of the model, not the fitted model itself. The implementation used in this thesis is described in Section 3.1.

Lengerich et al. (2020) apply this method to XGBoost models with maximum interaction depth two and show that the decomposition can recover ground-truth functional structure in simulated settings. They also evaluate the approach on real-world datasets and find that purification can separate main and interaction effects in ensemble models. At the same time, they note that purified effects may exhibit noisy local behaviour and that the resulting decomposition depends on the weighting distribution used during purification.

Building on this work, Yang et al. (2024) argue that depth-constrained tree ensembles combined with functional ANOVA decompositions may be interpreted as inherently interpretable generalized additive models with interactions. They emphasize that limiting tree depth is central for interpretability, since depth-two ensembles yield pairwise interaction components that are easier to visualize than higher-order effects. They also discuss additional strategies for improving interpretability, including monotonicity constraints, regularization, and pruning negligible effects after purification. At the same time, shallower trees may require larger ensembles to maintain predictive flexibility, reflecting a trade-off between interpretability and model flexibility.

Most directly related to the present thesis, Sudjianto (2025) argues that purified functional ANOVA decompositions provide a way to use flexible machine-learning models as structured and inherently interpretable models in high-stakes settings. The key point is that the fitted model is not only explained after training, but represented directly through main-effect and interaction components. Moreover, by leveraging results from Herren and Hahn (2022), local Shapley values can be obtained directly from the functional ANOVA representation by redistributing interaction components among the variables involved. This allows local explanations and global feature-importance summaries to be computed from the decomposed fitted model itself, rather than through a separate post-hoc explainer.

This line of work motivates the approach taken in the present thesis. By fitting a depth-constrained XGBoost model and decomposing the fitted score function into purified functional ANOVA components, the aim is to retain nonlinear predictive flexibility while obtaining a structured representation in terms of main effects and pairwise interactions.

3 Methodology

This chapter describes how the functional ANOVA representation is constructed and evaluated in practice. The chapter first presents the implementation used to decompose fitted XGBoost models into main-effect and pairwise-interaction components. It then introduces the evaluation and attribution summaries used to evaluate the functional ANOVA representation. Finally, the simulated and real-data studies used to assess the method are described.

3.1 Implementation of the functional ANOVA representation

3.1.1 Aggregation of the fitted XGBoost model

The fitted XGBoost model is first aggregated into intercept, one-way, and two-way components, referred to as the *raw* effects. The *base score* is extracted and added to the intercept, while the tree structure, stored in JSON format, is scanned to identify the features used by the model and their split points.

The split points are used to construct empty tensors: vectors for one-way effects and matrices for two-way effects. A feature with d splits generates a vector of length $d+1$. Similarly, the two-way effect between features with d and k splits forms a tensor of dimension $(d{+}1){\times}(k{+}1)$, or its transpose depending on the ordering of the variables. Each bin corresponds to an interval of the form $(a, b]$, consistent with the XGBoost splitting rule where observations less than or equal to the threshold are sent to the left child.

All trees are then iterated through, and their leaf values are added to the corresponding raw effects. A tree consisting only of a single leaf is added directly to the intercept, while a stump is added to the relevant main effect.

Figure 3 illustrates such a stump, where the leaves are aggregated into the main effect for x_1. Specifically, w_L is added to the bins corresponding to $(-\infty, s_1]$, while w_R is added to the bins corresponding to (s_1, ∞).

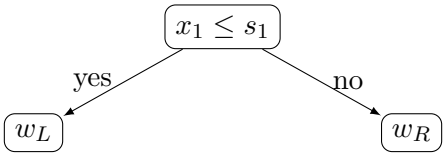


Figure 3: A stump (depth-1 tree) splitting on x_1.

Trees with two splits may contribute either to a main effect or to an interaction effect, depending on whether the second split uses the same feature or a different feature. To illustrate this, Figure 4 shows a depth-2 tree involving two features.

The left-hand side splits twice on the same feature and is therefore aggregated into the raw one-way effect for x_1. Specifically, w_{LL} is added to the bins corresponding to $(-\infty, s_2]$, while w_{LR} is added to the bins corresponding to $(s_2, s_1]$, with the remaining entries unchanged.

The right-hand side splits on two different features and is therefore aggregated into the raw interaction effect for x_1 and x_2. Specifically, w_{RL} is added to the bins corresponding to $(s_1, \infty) \times (-\infty, s_3]$, while w_{RR} is added to the bins corresponding to $(s_1, \infty) \times (s_3, \infty)$.

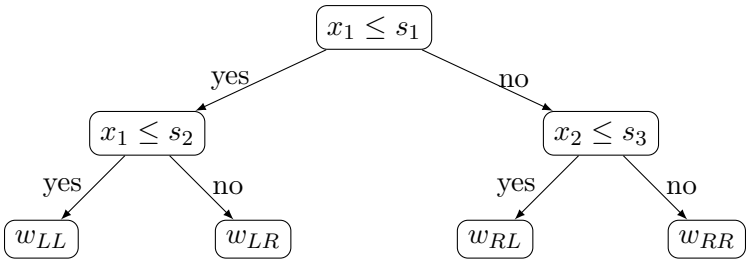


Figure 4: Depth-2 tree with repeated split on x_1 and interaction with x_2.

3.1.2 Purification of the raw effects

After aggregation, the fitted gradient-boosted tree model is represented by raw intercept, main-effect, and interaction tensors. These raw effects reconstruct the fitted model, but they do not generally satisfy the functional ANOVA constraints. The purpose of purification is therefore to transform the raw tensors into functional ANOVA components while preserving the fitted prediction function. The implementation is based on the mass-moving procedure of Lengerich et al. (2020), which applies to arbitrary piecewise constant functions.

As detailed in (2.18), the functional ANOVA constraints for piecewise constant functions are satisfied when all relevant tensor slices have weighted mean zero. The raw effects can therefore be viewed as a collection of tensors, where each tensor T_u represents the contribution associated with the variable subset u. For example, a one-way effect is represented by a vector over the bins of a single feature, while a two-way effect is represented by a matrix over the joint bins of two features.

The purification procedure works as follows. Let

$$m(T_u, i, x_{u\setminus i}) = \sum_{x_i \in \Omega_i} T_u(x_{u\setminus i}, x_i) \sum_{x_{-u}} w(x) \tag{3.1}$$

be the weighted mean of the slice of T_u obtained by varying x_i while keeping the remaining variables $x_{u\setminus i}$ fixed. In terms of the tensor representation, this corresponds to taking a slice of T_u along the dimension associated with x_i and computing the weighted average of its entries across the bins of x_i. In the present thesis, w is taken to be the empirical distribution of the observations over the bin discretization.

Because the model prediction is given by the sum of all tensor contributions, any quantity depending only on $x_{u\setminus i}$ can be transferred from T_u to the lower-order tensor $T_{u\setminus i}$ without changing the overall prediction. In particular, $m(T_u, i, x_{u\setminus i})$ is subtracted from the corresponding slice of T_u and added to $T_{u\setminus i}$. This operation ensures that the slice of T_u along dimension i has weighted mean zero while leaving the model prediction unchanged. Lengerich et al. (2020) refer to this procedure as *mass-moving*.

In practice, this means that for an aggregated tree ensemble we begin with the highest-order tensor, compute the weighted slice mean along one dimension, subtract this mean from the tensor, and move the removed mass to the corresponding lower-order tensor. This guarantees that all slices along that dimension satisfy the zero-mean condition. The same operation is then repeated for the remaining dimensions.

Performing the operation along one dimension may destroy the zero-mean property previously achieved along another dimension. However, by repeatedly cycling through the dimensions and successively removing the corresponding slice means, the tensors eventually converge to a representation in which all slices satisfy the required zero-mean conditions.

The algorithm for purifying a single tensor together with the overall purification procedure is given below.

Algorithm 1 Purify-Matrix (Lengerich et al., 2020)

Require: T, w, u, Ω ▷ Will purify T_u so that every slice has zero-mean according to weighting w

end ← False
while end ≠ True **do**
    end ← True
    for $i \in u$ **do**
        for $x_{u\setminus i} \in \Omega_{u\setminus i}$ **do**
            $m^0 \leftarrow m(T_u, i, x_{u\setminus i})$ ▷ (3.1)
            if $m^0 \neq 0$ **then** ▷ There is mass to move
                end ← False
                $T_u[x_{u\setminus i}, :] \leftarrow T_u[x_{u\setminus i}, :] - m^0$
                $T_{u\setminus i}[x_{u\setminus i}] \leftarrow T_{u\setminus i}[x_{u\setminus i}] + m^0$
            end if
        end for
    end for
end while
return T

Algorithm 2 Purify (Lengerich et al., 2020)

Require: T, w, Ω ▷ T is the set of tensors, Ω the values, w the weighting

order ← sort_descending($|T|$) ▷ Arrange all index sets u in decreasing size
for $u \in$ order **do**
    $T \leftarrow$ Purify-Matrix(T, w, u, Ω)
end for
return T

The convergence of this procedure is characterized by Lengerich et al. (2020).

Specifically, define a measure of the remaining "unpurified mass" of a tensor at iteration t. For a two-way interaction tensor $T^t_{a,b}$, the weighted row and column means are defined as

$$r^t_i = \sum_j w_{i,j} T^t_{a,b}[i,j], \qquad c^t_j = \sum_i w_{i,j} T^t_{a,b}[i,j]. \tag{3.2}$$

The total deviation from the zero-mean conditions can then be quantified by

$$M_t = \sum_{i,j} w_{i,j}(|r^t_i| + |c^t_j|). \tag{3.3}$$

The tensor is therefore purified precisely when $M_t = 0$.

Lengerich et al. (2020) provide the following convergence results.

Theorem 3.1. *For any $T_{a,b}, \Omega$, if $w_{i,j} = w_{i,j'} \forall i, j'$:*

$$M^t = 0 \quad \forall t \geq 2 \tag{3.4}$$

Theorem 3.2. *For any* $T_{a,b}, \Omega, w$*, for any* $\epsilon > 0$

$$M^t \leq \epsilon \quad \forall t \geq \tau(\epsilon) \tag{3.5}$$

where $\tau(\epsilon) = \log_2(\frac{M^0}{\epsilon})$

Thus, under equal weighting along each dimension, the algorithm converges after a single purification pass through each dimension. For more general non-degenerate weighting distributions, convergence is typically achieved after only a small number of iterations. Proofs are provided in Lengerich et al. (2020).

To illustrate the procedure, Figure 5 shows an example under uniform weighting, adapted from Lengerich et al. (2020). Consistent with the first convergence result, the zero-mean conditions are satisfied after one purification pass through each dimension.

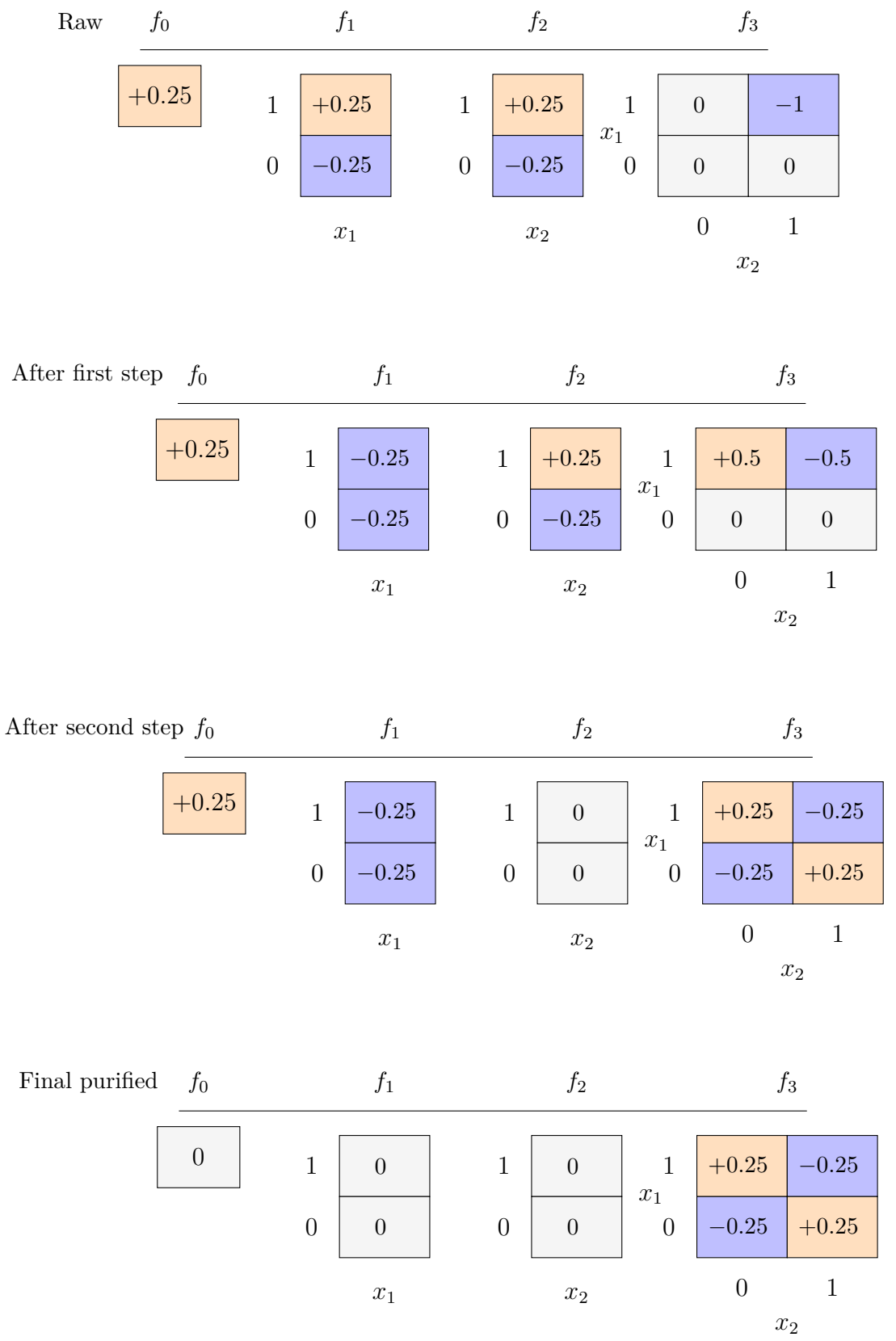


Figure 5: Illustration of the purification procedure under uniform weighting. From top to bottom: the raw tensor components, the state after purification along x_1, the state after purification along x_2, and the final purified representation. Positive values are shown in orange and negative values in blue.

3.2 Evaluation and attribution summaries

This section describes the attribution and stability summaries used to assess the purified functional ANOVA representation. The summaries considered are SHAP values, global feature-importance measures, and resampling-based uncertainty intervals.

3.2.1 Local attribution

First, SHAP values are used to measure local feature attribution. Following Herren and Hahn (2022), the SHAP value for feature X_j for a given observation x can be expressed as

$$\phi_j(x) = f_j(x_j) + \sum_{k \neq j} \frac{1}{2} f_{jk}(x_j, x_k) + \sum_{\substack{k<l \\ k,l \neq j}} \frac{1}{3} f_{jkl}(x_j, x_k, x_l) + \dots \tag{3.6}$$

Thus, each functional ANOVA component is distributed equally among the variables on which it depends. This provides a direct decomposition of the prediction into feature-level contributions. The formulation is attractive because it is obtained directly from the functional ANOVA decomposition and does not require an additional post-hoc approximation procedure.

3.2.2 Global feature importance

For global feature importance, two measures are considered: mean absolute SHAP values and pseudo-Shapley effects. Mean absolute SHAP importance is computed by averaging the absolute SHAP values over the training observations, as described in Section 2.4.

For the pseudo-Shapley effects, additional considerations arise in the presence of dependence. As discussed in Section 2.5, under independence the functional ANOVA components are orthogonal, so the variance of the model output can be decomposed into a sum of variance components. In that setting, Shapley effects can be computed directly from the variance decomposition.

Under dependence, this orthogonal variance decomposition no longer holds. Consider the functional ANOVA representation

$$F = f_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) + f_{12}(X_1, X_2) + f_{13}(X_1, X_3) + f_{23}(X_2, X_3).$$

Although hierarchical orthogonality still holds, the variance of the model output is now given by

$$\mathrm{Var}\,(F) = \sum_{u \in \mathcal{U}} \mathrm{Var}\,(f_u(X_u)) + 2 \sum_{\substack{u,v \in \mathcal{U} \\ u<v}} \mathrm{Cov}(f_u(X_u), f_v(X_v)),$$

where $\mathcal{U}$ denotes the set of all non-empty index sets corresponding to the functional ANOVA components. The covariance terms arise from dependence between the inputs and represent shared contributions to the variability of the model output. Consequently, the variance can no longer be uniquely attributed to individual components or variables using only the component variances.

Classical Shapley effects address this allocation problem by applying the Shapley value to a variance-based value function involving conditional expectations. However, computing such quantities is generally nontrivial, since it requires evaluating conditional expectations with respect to subsets of variables, which is particularly challenging under dependence. Motivated by this, we instead consider a direct allocation of the variance

expansion induced by the functional ANOVA representation. This should not be interpreted as a standard Shapley effect, but rather as a variance-allocation scheme inspired by the Shapley-value principle.

Let $\mathcal{U}$ denote the set of component index sets, and let j denote a variable. We define the pseudo-Shapley effect for variable j as

$$\Phi_j = \sum_{u \in \mathcal{U}: j \in u} \frac{1}{|u|} \mathrm{Var}\left(f_u(X_u)\right) + \sum_{\substack{u,v \in \mathcal{U} \\ u<v}} \left(\frac{\mathbf{1}\{j \in u\}}{|u|} + \frac{\mathbf{1}\{j \in v\}}{|v|} \right) \mathrm{Cov}(f_u(X_u), f_v(X_v)).$$

The allocation can be interpreted component-wise. Each variance term is distributed equally among the variables involved in the corresponding component. For a covariance term between components u and v, the total contribution to the variance expansion is $2\mathrm{Cov}(f_u(X_u), f_v(X_v))$. The allocation above assigns one covariance term to component u and one to component v, after which each component's share is distributed equally among the variables appearing in that component.

This yields intuitive allocation rules in several cases. A covariance between two main effects is split between the two corresponding variables. A covariance between a main effect and an interaction assigns one covariance contribution to the main-effect variable, while the interaction contribution is split equally among the variables in the interaction. A covariance between two interaction effects is handled analogously, with variables appearing in both interactions receiving contributions from both component sides.

Applying this allocation to all variables gives a decomposition of the total variance into feature-level contributions that accounts for both individual and shared effects. Since the construction is inspired by the Shapley value but is not derived from a value function over feature subsets, it does not in general coincide with classical Shapley effects. Instead, it provides a direct way to quantify how individual variables contribute to the variability of the model output in the presence of dependence.

Because covariance terms may be negative, a variable may also receive a negative pseudo-Shapley contribution. This can occur when the variance contributions associated with components containing the variable are outweighed by sufficiently negative covariance terms.

3.2.3 Resampling uncertainty

Finally, because the modelling framework is non-parametric, classical significance tests and confidence intervals are not directly applicable. A resampling-based procedure was therefore used to assess the stability of the functional ANOVA effects and the associated importance measures.

New training and test sets were repeatedly resampled. For each resample, the model was refitted and the raw effects were purified, yielding 50 decompositions in total. All bins from the resampled models were then collected, and the corresponding functional ANOVA effects were evaluated at the midpoint of each bin.

The median effect across the resampled models was plotted together with an empirical 95 percent quantile interval based on the 2.5 and 97.5 percent quantiles. Similarly, the pseudo-Shapley effects and mean absolute SHAP values were summarized using medians and empirical quantile-based error bars.

Each resampled model was estimated using the same hyperparameters as the original fitted model. This isolates the variability induced by the data sample, rather than

introducing additional variation from repeated hyperparameter tuning. Thus, the resampling procedure evaluates the stability of the purified functional ANOVA effects under a fixed modelling specification, rather than the robustness of the XGBoost tuning procedure itself.

3.3 Simulation Study

The simulation study evaluates the purification procedure in settings where the underlying functional structure is known. The first two simulations use continuous responses, first with independent inputs and then with dependent inputs, so that the theoretical functional ANOVA effects can be derived directly and compared with the recovered effects. The third simulation uses a binary response generated from a latent logistic score function, reflecting the fact that gradient boosting classification estimates a latent score rather than observing it directly. Finally, a credit-risk toy model is used to study how ratio-based covariates may affect the interaction structure recovered by the functional ANOVA decomposition.

3.3.1 Independent linear model

The first simulation considers the function

$$F = a + bX_1 + cX_2 + dX_1X_2, \tag{3.7}$$

where X_1 and X_2 are independent normally distributed variables with means μ_1, μ_2 and standard deviations σ_1, σ_2.

Since the inputs are independent, the theoretical functional ANOVA effects are obtained directly from the recursive definition in (2.12). This gives

$$\begin{aligned} f_0 &= \mathbb{E}[F] = a + b\mu_1 + c\mu_2 + d\mu_1\mu_2, \\ f_1(x_1) &= \mathbb{E}[F \mid X_1 = x_1] - f_0 = (b + d\mu_2)(x_1 - \mu_1), \\ f_2(x_2) &= \mathbb{E}[F \mid X_2 = x_2] - f_0 = (c + d\mu_1)(x_2 - \mu_2), \\ f_{12}(x_1, x_2) &= \mathbb{E}[F \mid X_1 = x_1, X_2 = x_2] - f_0 - f_1(x_1) - f_2(x_2) \\ &= d(x_1 - \mu_1)(x_2 - \mu_2). \end{aligned} \tag{3.8}$$

The specific data-generating function used in the simulation is

$$\begin{aligned} F &= 3 + X_1 + 4X_2 - X_1X_2, \\ X_1 &\sim N(2, 1), \quad X_2 \sim N(5, 1). \end{aligned}$$

This specification was chosen because it produces positive raw marginal effects but a negative interaction effect. Consequently, the functional ANOVA main effect for X_1 has the opposite direction from its raw marginal effect, since the average interaction contribution with X_2 dominates the positive raw coefficient of X_1. The example therefore illustrates that the functional ANOVA specification can lead to a substantially different interpretation of a variable's main effect than the raw algebraic form of the function.

Figures 6 and 7 show the theoretical raw and purified effects. The purified main effect for X_1 reverses direction relative to the raw effect, while the interaction effect retains the same overall structure but is centered around the feature means.

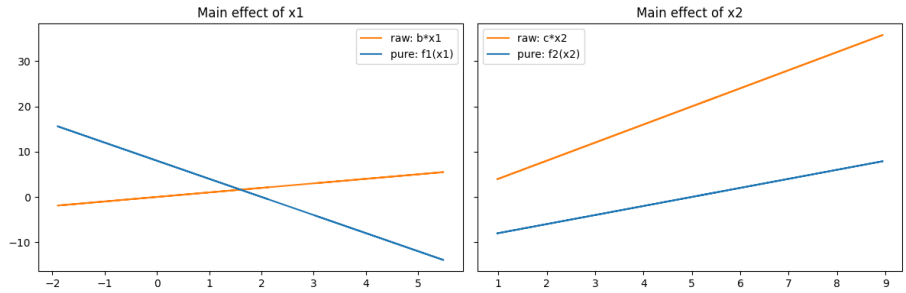


Figure 6: Theoretical raw (orange) and purified (blue) main effects for the independent linear model.

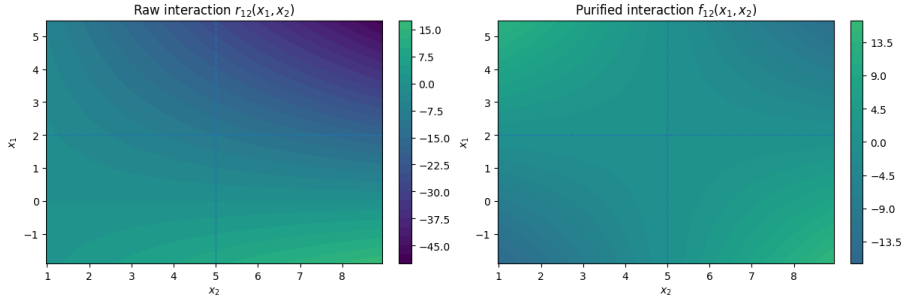


Figure 7: Theoretical raw (left) and purified (right) interaction effects for the independent linear model.

3.3.2 Dependent linear model

The second simulation introduces dependence between X_1 and X_2:

$$(X_1, X_2) \sim N_2(\mu, \Sigma),$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Since the inputs are now dependent, the independent-input recursion in (2.12) no longer applies. As discussed in Section 2.5, the generalized functional ANOVA components must instead satisfy the hierarchical orthogonality conditions in (2.15). Although closed-form expressions are rarely available in general, they can be derived for the present model.

Define

$$U = \frac{X_1 - \mu_1}{\sigma_1}, \qquad V = \frac{X_2 - \mu_2}{\sigma_2},$$

so that $(U, V) \sim N_2((0,0), \Sigma_U)$, where

$$\Sigma_U = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We can rewrite the function as

$$\begin{aligned} F &= a + b(\sigma_1 U + \mu_1) + c(\sigma_2 V + \mu_2) + d(\sigma_1 U + \mu_1)(\sigma_2 V + \mu_2) \\ &= a + (b + d\mu_2)\sigma_1 U + (c + d\mu_1)\sigma_2 V + d\sigma_1\sigma_2 UV + b\mu_1 + c\mu_2 + d\mu_1\mu_2 \\ &= a + pU + qV + kUV + b\mu_1 + c\mu_2 + d\mu_1\mu_2, \end{aligned}$$

where

$$p = (b + d\mu_2)\sigma_1, \qquad q = (c + d\mu_1)\sigma_2, \qquad k = d\sigma_1\sigma_2.$$

Since $f_0 = \mathbb{E}[F]$, we obtain

$$f_0 = a + b\mu_1 + c\mu_2 + d(\mu_1\mu_2 + \rho\sigma_1\sigma_2),$$

and therefore

$$F - f_0 = pU + qV + k(UV - \rho).$$

To construct the interaction component, we impose hierarchical orthogonality:

$$\mathbb{E}[f_{12}(X_1, X_2) \mid X_1] = 0, \qquad \mathbb{E}[f_{12}(X_1, X_2) \mid X_2] = 0.$$

Since conditioning on X_1 is equivalent to conditioning on U, and similarly for X_2 and V, the conditions may be imposed in terms of U and V. By symmetry, it suffices to solve for m using $\mathbb{E}[f_{12} \mid U] = 0$.

We first compute the conditional expectation of the interaction term. Since (U, V) is jointly normal, the conditional distribution of V given U is normal with

$$\mathbb{E}[V \mid U] = \rho U, \qquad \mathrm{Var}(V \mid U) = 1 - \rho^2$$

and thus

$$\mathbb{E}[k(UV - \rho) \mid U] = kU\mathbb{E}[V \mid U] - k\rho = k\rho U^2 - k\rho = k\rho(U^2 - 1).$$

Thus, the term $k(UV - \rho)$ does not satisfy the orthogonality condition and must therefore be orthogonalized by subtracting lower-order functions of U and V. We therefore define

$$f_{12} = k(UV - \rho) - m(U^2 - 1) - m(V^2 - 1),$$

and solve for m such that $\mathbb{E}[f_{12} \mid U] = 0$.

Using

$$\mathbb{E}[V^2 \mid U] = (1 - \rho^2) + (\rho U)^2,$$

we obtain

$$\begin{aligned}
\mathbb{E}[f_{12} \mid U] &= k\rho(U^2 - 1) - m(U^2 - 1) - m\Big((1 - \rho^2) + \rho^2 U^2 - 1\Big) \\
&= k\rho(U^2 - 1) - m(U^2 - 1) - m\rho^2(U^2 - 1) \\
&= \Big(k\rho - m(1 + \rho^2)\Big)(U^2 - 1).
\end{aligned}$$

Setting this equal to zero yields

$$m = \frac{k\rho}{1 + \rho^2}.$$

The canonical functional ANOVA components are therefore

$$\begin{aligned}
f_0 &= a + b\mu_1 + c\mu_2 + d(\mu_1\mu_2 + \rho\sigma_1\sigma_2), \\
f_1 &= pU + \frac{k\rho}{1 + \rho^2}(U^2 - 1), \\
f_2 &= qV + \frac{k\rho}{1 + \rho^2}(V^2 - 1), \\
f_{12} &= k(UV - \rho) - \frac{k\rho}{1 + \rho^2}(U^2 - 1) - \frac{k\rho}{1 + \rho^2}(V^2 - 1).
\end{aligned}$$

The correlation parameter was set to $\rho = 0.8$. Figures 8 and 9 show the resulting theoretical raw and functional ANOVA effects. The dependence structure causes the functional ANOVA main effects to become quadratic, even though the original function is linear in each marginal variable. The interaction surface is also rotated and stretched. Because of the dependence structure, observations are concentrated along the main diagonal, while the top-left and bottom-right regions contain little probability mass. Thus, although the theoretical interaction effect is defined over the full feature space, these regions are unlikely to be observed in practice.

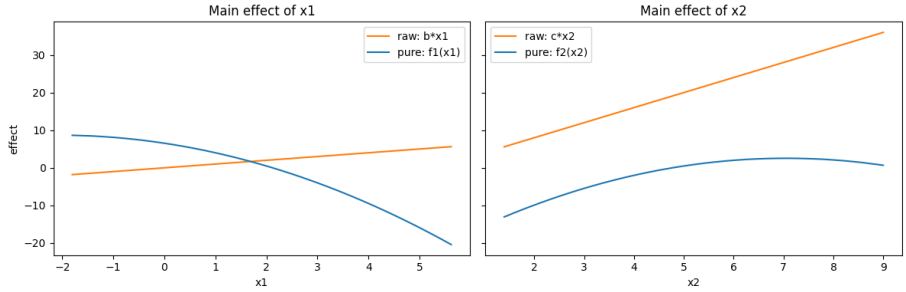


Figure 8: Theoretical raw (orange) and purified (blue) main effects for the dependent linear model.

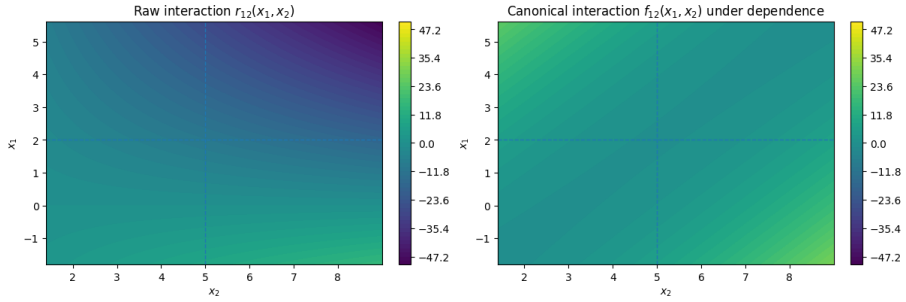


Figure 9: Theoretical raw (left) and purified (right) interaction effect for the dependent linear model.

3.3.3 Binary response with latent score

The third simulation uses the same type of latent functional form as the linear simulations, but observes only a binary response. Let

$$\eta(X_1, X_2) = a + bX_1 + cX_2 + dX_1X_2$$

denote the latent score function. The response is generated from this score using the logistic Bernoulli model described in Section 2.2.

This simulation is included because it is closer to the binary probability-estimation setting considered in the real-data application. In contrast to the regression simulations, the fitted model does not observe the latent score function directly, but only binary realizations generated from probabilities determined by that score. The functional ANOVA effects of interest are therefore still the effects of the latent score function η, but they must now be recovered indirectly from the binary response.

The parameter values were changed relative to the previous linear simulations. Using the original parameters together with the logistic link produced an extremely imbalanced response, with almost all observations belonging to one class. The parameters were therefore chosen to produce a less degenerate class distribution while retaining the same qualitative form of the latent model.

3.3.4 Credit-risk toy model

Finally, we consider a toy example designed to illustrate a potential tension between functional ANOVA interaction effects and ratio-based covariates in non-retail credit-risk modelling. As discussed in Section 2.1, financial ratios are commonly used as risk drivers instead of the underlying balance-sheet quantities. From a functional ANOVA perspective, this matters because ratios may already encode nonlinear relationships between the underlying variables. Consequently, interaction effects may appear differently depending on whether the model is expressed in terms of ratios or in terms of the original accounting quantities.

To explore this, we construct a stylised binary-response model with latent score

$$\eta_i = -5.5 - 8R_{Pi} + 4R_{Li},$$

where the binary response is generated from η_i using the logistic Bernoulli model described above. Here $R_P = P/A$ denotes profitability and $R_L = L/A$ denotes leverage, where P, L, and A represent profit, liabilities, and assets. The latent score is therefore additive in the two ratios. However, the mapping from the underlying quantities (P, L, A) to (R_P, R_L) is nonlinear, so interaction structure may reappear when the same relationship is expressed in terms of the original variables.

For the purpose of this toy example, the ratio variables and total assets are generated independently. This assumption is stylised and should not be interpreted as a realistic description of financial data. Rather, it creates a controlled setting in which the ratio variables are free of dependence, allowing the interaction structure induced by the choice of covariates to be studied separately from dependence in the input distribution.

The covariates are generated according to

$$\log A \sim N(10, 1), \quad R_L \sim \beta(2, 2), \quad R_P \sim N(0.03, 0.08),$$

and

$$P = AR_P, \quad L = AR_L.$$

The log-normal distribution for assets captures the skewness typically observed in firm size, while the beta distribution keeps leverage bounded between 0 and 1. The profitability ratio is modelled as approximately centered around a small positive value while allowing both positive and negative profitability. The distributions are not intended to be fully realistic, but to provide a controlled setting in which the relationship between ratios, firm size, and interaction effects can be examined.

3.3.5 Model training for the simulation study

For all simulation settings, the XGBoost trees were restricted to `max_depth`=2 when constructing the functional ANOVA decomposition. This restricts the representation to main effects and pairwise interactions, which are substantially easier to visualize and interpret than higher-order effects. It also avoids estimating higher-order interaction

components on sparse partitions of the feature space, which would likely produce irregular and difficult-to-interpret structures.

For simulations where the target function was directly observed, the models were trained using the default XGBoost hyperparameters apart from the depth restriction. For simulations where the target function had to be estimated indirectly from binary outcomes, hyperparameter tuning was performed using the same procedure as in the real-data study, described in Section 3.4.2.

3.4 Real-data study

In addition to the simulation study, we consider the Polish Companies Bankruptcy dataset from the UCI Machine Learning Repository (Tomczak, 2016). The dataset was constructed from financial information on Polish manufacturing companies obtained from the Emerging Markets Information Service (EMIS) database. It contains financial variables derived from annual financial statements together with a binary bankruptcy indicator.

The dataset is provided in five versions corresponding to different forecasting horizons, where the target indicates whether a firm goes bankrupt within one, two, three, four, or five years after the observation period. This study uses the 1-year horizon, which contains 7027 observations, of which 271 correspond to bankruptcies. The explanatory variables consist of 64 financial measures capturing profitability, liquidity, leverage, operational efficiency, and related aspects of firm condition, as defined in Tomczak (2016). Most variables are financial ratios, although some absolute quantities are also included.

The dataset does not contain explicit firm identifiers or time indices. Although the underlying information was collected from financial statements over multiple years, the publicly available dataset is provided in processed form and is therefore treated as a cross-sectional sample. The number of bankruptcies is also small compared with what would typically be expected in an IRB PD modelling setting for SME portfolios. This is a limitation, but the purpose of the present study is to evaluate the proposed functional ANOVA representation rather than to develop a fully specified regulatory PD model.

The main reason for using this dataset is that publicly available datasets suitable for non-retail credit-risk modelling are limited, and the included variables resemble the types of risk drivers commonly used in non-retail PD models. Unlike many larger retail credit-risk datasets, this dataset is based on financial statement variables and accounting ratios rather than behavioural or customer-level variables. This is important in the present thesis because financial ratios may have nonlinear effects and often share common numerators or denominators, producing strong and complicated dependence structures among the explanatory variables. The dataset therefore provides a useful setting for evaluating the proposed framework on variables closer to those encountered in non-retail credit-risk modelling.

Bankruptcy prediction is not identical to default prediction, but the problems are closely related through the modelling of severe financial distress. The 1-year forecasting horizon was chosen because default prediction is typically formulated over a one-year horizon, making this version of the dataset the closest to the PD modelling setting considered in the thesis.

The dataset was previously used by Zięba et al. (2016), who evaluated XGBoost for bankruptcy prediction. It has therefore already been studied in the context of the main modelling framework used here. However, the present thesis imposes additional interpretability constraints, so the aim is not to match the predictive performance of an

unconstrained XGBoost model.

3.4.1 Data cleaning and initial feature selection

The real data require pre-processing before modelling.

The first issue was the treatment of missing values. Although XGBoost can handle missing values natively by learning a default direction at each split, missing values were excluded in the present study.

There are two reasons for this choice. First, the functional ANOVA decomposition is defined with respect to a function over the feature space and an associated input distribution. If missing-value routing is learned implicitly by XGBoost, the effect of a feature depends not only on its numerical value but also on whether the feature is observed. This would introduce an additional modelling dimension related to data availability rather than the economic relationships of interest.

Second, missingness is important in a regulatory credit-risk context and would normally require explicit treatment and justification. EBA guidelines on PD and LGD estimation require institutions to identify and address data deficiencies, including missing data, and to account for remaining uncertainty through a margin of conservatism (2017). Since the objective of this thesis is not to compare missing-data strategies, missing values were excluded from the analysis.

The removal procedure was performed in two stages. First, features with more than 5 percent missing observations were removed. This removed two variables: sales (n) / sales (n-1) and (current assets - inventories) / long-term liabilities. After this, all remaining observations containing missing values were removed. This removed approximately 10 percent of the observations, leaving 6318 observations and 62 explanatory variables. The observed bankruptcy rate in the reduced dataset was 1.66 percent.

Many explanatory variables are ratios, so missing values may arise from economically meaningful configurations such as division by zero in the denominator. However, since the underlying accounting components are not available, such cases cannot be verified directly and would require explicit modelling treatment.

The second preprocessing issue was strong dependence and redundancy among explanatory variables. From a predictive perspective, XGBoost can often handle correlated variables because overlapping predictors may provide similar split gains. The generalized functional ANOVA framework also allows dependent inputs. However, the iterative purification procedure has favourable convergence properties only for non-degenerate distributions. This is problematic for ratio-based financial variables, many of which are either almost deterministically related or strongly connected through common accounting components. For example, variables such as total liabilities / total assets and total assets / total liabilities are deterministic transformations of one another, while other ratios may differ only through small accounting adjustments.

To identify such relationships, both ordinary Pearson correlations and log-based Pearson correlations were inspected. Ordinary Pearson correlation captures linear association on the original scale, while log-based correlations can reveal inverse or multiplicative relationships among positive ratio variables. The log-based correlations were computed by retaining strictly positive observations for each pair of variables and applying a logarithmic transformation before computing Pearson correlation.

Figure 10 illustrates the resulting correlation matrices for the explanatory variables.

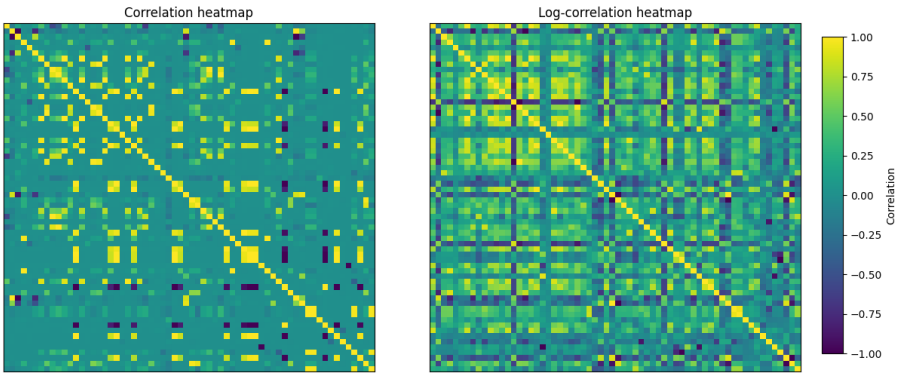


Figure 10: Ordinary Pearson correlation matrix (left) and log-based Pearson correlation matrix (right) for the explanatory variables after missing-value filtering. Feature names have been abbreviated for readability.

The figure shows several large clusters of highly correlated variables, with additional associations visible in the log-correlation matrix. In several near-degenerate variable combinations, the purification algorithm failed to converge even after 1000 iterations. A feature-reduction procedure was therefore introduced to remove highly redundant variables before fitting the final model.

For each pair of variables, similarity was defined as the maximum absolute value of the ordinary and log-based Pearson correlations. Variables with similarity above 0.9 were connected in a graph, and clusters were formed as connected components. For each cluster, the representative variable was chosen as the variable with the strongest absolute correlation with the response. This procedure is intended as a pragmatic reduction step rather than a full production-style variable selection exercise.

This procedure produced one large cluster of 23 variables primarily related to sales-based ratios, together with smaller clusters related to profitability, current assets and equity, gross profit and total liabilities, capital and fixed assets, and total assets. In total, 40 of the 62 explanatory variables were removed. The final retained feature set is shown in Table 1.

Table 1: Final retained feature set after correlation-based feature reduction.

Feature
net profit / total assets
total liabilities / total assets
working capital / total assets
(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation) * 365
retained earnings / total assets
sales / total assets
gross profit / short-term liabilities
gross profit (in 3 years) / total assets
(equity - share capital) / total assets
profit on operating activities / financial expenses
logarithm of total assets
operating expenses / total liabilities
total sales / total assets
constant capital / total assets
(current assets - inventory - receivables) / short-term liabilities
total liabilities / ((profit on operating activities + depreciation) * (12/365))
EBITDA (profit on operating activities - depreciation) / total assets
equity / fixed assets
working capital
(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
long-term liabilities / equity
sales / fixed assets

3.4.2 Model training for the real-data study

For the real-data application, the target function is not observed directly and must be estimated from binary outcomes. Hyperparameter tuning was therefore performed. The same tuning procedure was also used for the binary simulation settings.

The data were split into training and test sets using an 80/20 split with stratified sampling to preserve the class proportions. The tuning criterion was ROC AUC (see Hastie et al. (2009)), which aligns with the objective of ranking observations according to risk. Hyperparameter tuning was performed using randomized search with stratified 5-fold cross-validation on the training data. For each hyperparameter configuration, the model was trained on four folds and evaluated on the remaining fold, and the mean ROC AUC across folds was used as the performance measure. A total of 40 hyperparameter configurations were sampled, resulting in 200 fitted models. The best-performing configuration was then refitted on the full training set and evaluated on the test set.

Class imbalance was handled through the parameter `scale_pos_weight`. The baseline value was defined as

$$\texttt{scale\_pos\_weight} = \frac{N_{\text{neg}}}{N_{\text{pos}}},$$

where N_{neg} and N_{pos} denote the number of negative and positive observations in the

training data, respectively.

The hyperparameter search space was given by

$$
\begin{aligned}
\texttt{n\_estimators} &\in \{200, 400, 800, 1200, 1600\},\\
\texttt{learning\_rate} &\in \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1\},\\
\texttt{min\_child\_weight} &\in \{1, 2, 5, 10, 20, 40\},\\
\texttt{gamma} &\in \{0, 0.1, 0.3, 0.5, 1, 2, 5\},\\
\texttt{subsample} &\in \{0.6, 0.7, 0.8, 0.9, 1.0\},\\
\texttt{colsample\_bytree} &\in \{0.4, 0.6, 0.8, 1.0\},\\
\texttt{reg\_alpha} &\in \{0, 10^{-3}, 10^{-2}, 0.1, 1, 5\},\\
\texttt{reg\_lambda} &\in \{0.5, 1, 2, 5, 10, 20\},\\
\texttt{scale\_pos\_weight} &\in \left\{1.0,\ \frac{N_{\text{neg}}}{N_{\text{pos}}},\ 0.5 \cdot \frac{N_{\text{neg}}}{N_{\text{pos}}},\ 1.5 \cdot \frac{N_{\text{neg}}}{N_{\text{pos}}}\right\}.
\end{aligned}
$$

Since all hyperparameters were tuned simultaneously, certain parameter combinations may produce clearly unsuitable configurations, such as relatively high learning rates combined with a very large number of trees. In some cases, minor adjustments to the search space were therefore made by removing clearly unsuitable parameter regions for a given fit.

Hyperparameter tuning was not the primary focus of this thesis, and no extensive sequential tuning strategy was employed. The purpose was to obtain a reasonable fitted model for evaluating the functional ANOVA representation, rather than to fully optimize predictive performance. Whenever the model specification was modified, for example through feature selection or monotonicity constraints, the hyperparameters were re-tuned.

Monotonicity constraints were also used in selected model specifications to impose expected effect directions based on domain knowledge and to reduce noisy effect behaviour. These constraints act on the underlying XGBoost model rather than directly on the purified functional ANOVA components. Nevertheless, monotonicity was generally observed to be preserved in the purified main effects, since purification transfers systematic lower-order structure from interaction tensors to the corresponding main effects.

Monotonicity is not generally preserved in the purified interaction effects themselves. This is not necessarily problematic, since interaction effects act as corrections to the main effects rather than as standalone marginal relationships. A non-monotone interaction component may therefore still be consistent with a monotone overall effect once the corresponding main effect is included.

4 Results and Analysis

This chapter presents the results of the simulation and real-data studies. The simulation study is used to assess whether the functional ANOVA implementation recovers known structures under controlled conditions and to illustrate how dependence affects the interpretation of the components. The real-data study then applies the method to a bankruptcy-prediction dataset in order to evaluate whether the recovered components provide useful global and local insight into a fitted credit-risk model. The purpose of the analysis is not primarily to establish XGBoost as a superior predictive model, but to investigate whether the fitted model can be represented and interpreted in a useful way.

4.1 Simulation Study Results

4.1.1 Independent Linear Model

The first model evaluated is the linear model with independent features. As explained in Section 3.3, this example is deliberately somewhat counterintuitive: X_1 and X_2 separately have increasing raw marginal effects on the response, while together they have a negative interaction effect.

A sample of 10,000 independent observations was simulated, with normally distributed noise $\epsilon \sim N(0, 0.5^2)$ added to the response. The data were split into training and test sets as previously described, and an XGBoost model was fitted using the default parameters except for the depth-2 restriction. The fitted tree ensemble was then aggregated and purified using the procedure described in Section 3.1.

The recovered purified effects closely resemble the theoretical functional ANOVA effects, particularly in regions where the data density is high. Figure 11 compares the recovered and theoretical main effects. The main discrepancies are the visible jumps around the feature means and the local noise introduced by the piecewise-constant tree approximation.

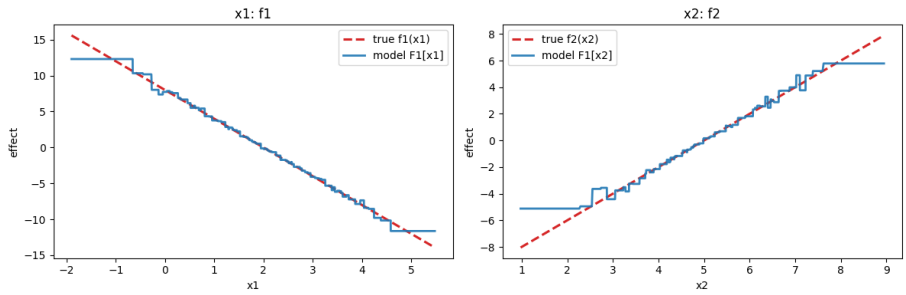


Figure 11: Comparison between the purified and theoretical main effects for the independent linear model.

The interaction effect is harder to compare directly as a surface. Figure 12 therefore shows the difference between the theoretical and purified interaction effects at each observation. The differences display the same plateau structure induced by the tree approximation, with larger discrepancies in boundary regions where data support is weaker.

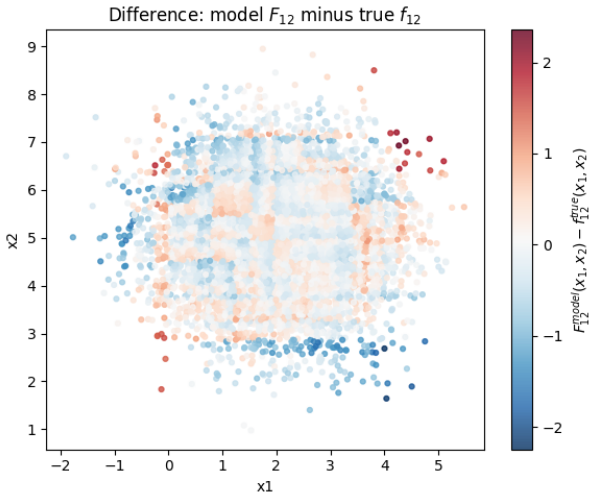


Figure 12: Difference between the purified and theoretical interaction effects for the independent linear model.

Table 2 summarizes the differences between the theoretical and recovered effects.

Table 2: Difference diagnostics between theoretical and purified effects.

Effect	Mean difference	RMSE difference	Max abs difference
f_1	0.024523	0.234772	3.290308
f_2	-0.010424	0.204892	2.926837
f_{12}	-0.013745	0.353013	2.356567

The mean differences are close to zero and small relative to the scale of the effects, suggesting little systematic under- or overestimation after purification. The RMSE values are moderate relative to the maximum deviations, indicating that the largest discrepancies are localized. The interaction effect has the largest RMSE, consistent with interaction effects being more difficult to estimate accurately.

Figure 13 shows the global SHAP importance.

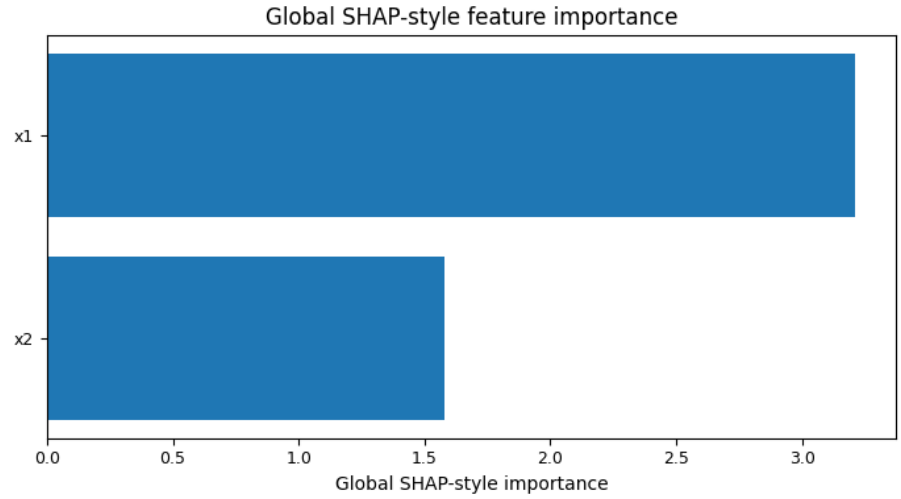


Figure 13: Global SHAP importance for the independent linear model.

The SHAP importance assigns greater importance to X_1 than to X_2, even though the original linear coefficients are $b = 1$ and $c = 4$. This reflects that the functional ANOVA main effects are not determined by the raw linear coefficients alone, but also by the average contribution of the interaction term under the input distribution. In this case,

$$f_1(x_1) = (b + d\mu_2)(x_1 - \mu_1), \qquad f_2(x_2) = (c + d\mu_1)(x_2 - \mu_2).$$

Substituting the parameter values gives

$$f_1(x_1) = (-4)(x_1 - 2), \qquad f_2(x_2) = 2(x_2 - 5).$$

Thus, the interaction contribution shifts the slope of the X_1 main effect from 1 to -4, making its average absolute contribution larger than that of X_2.

For normally distributed inputs,

$$\mathbb{E}[|X_i - \mu_i|] = \sigma_i \sqrt{\frac{2}{\pi}}.$$

Since $\sigma_1 = \sigma_2 = 1$, the expected absolute main-effect magnitudes are

$$\mathbb{E}[|f_1(X_1)|] = 4\sqrt{\frac{2}{\pi}}, \qquad \mathbb{E}[|f_2(X_2)|] = 2\sqrt{\frac{2}{\pi}}.$$

Using the local SHAP attribution rule discussed in Section 3.2, where half of the interaction effect is additionally assigned to each variable, the corresponding theoretical mean absolute SHAP values are approximately 3.19 and 1.60. These values closely match the recovered importance values.

Figure 14 shows the pseudo-Shapley effects. Since the inputs are independent and the functional ANOVA components are orthogonal, this allocation coincides with the usual variance-based allocation obtained from the Sobol decomposition.

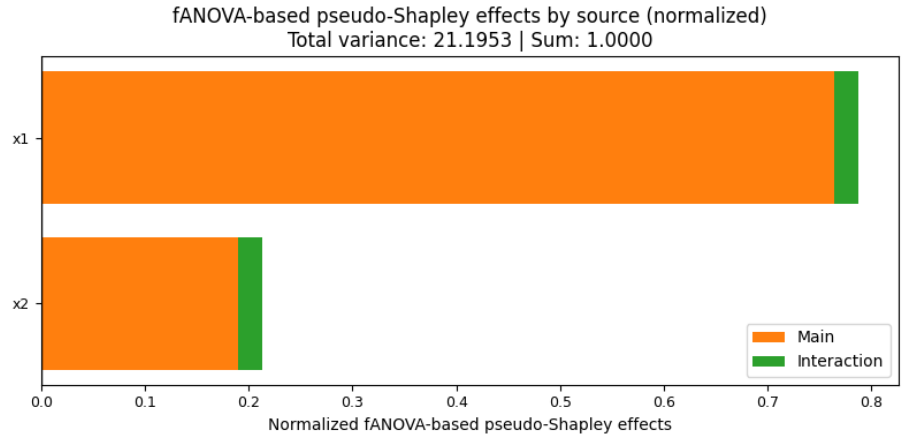


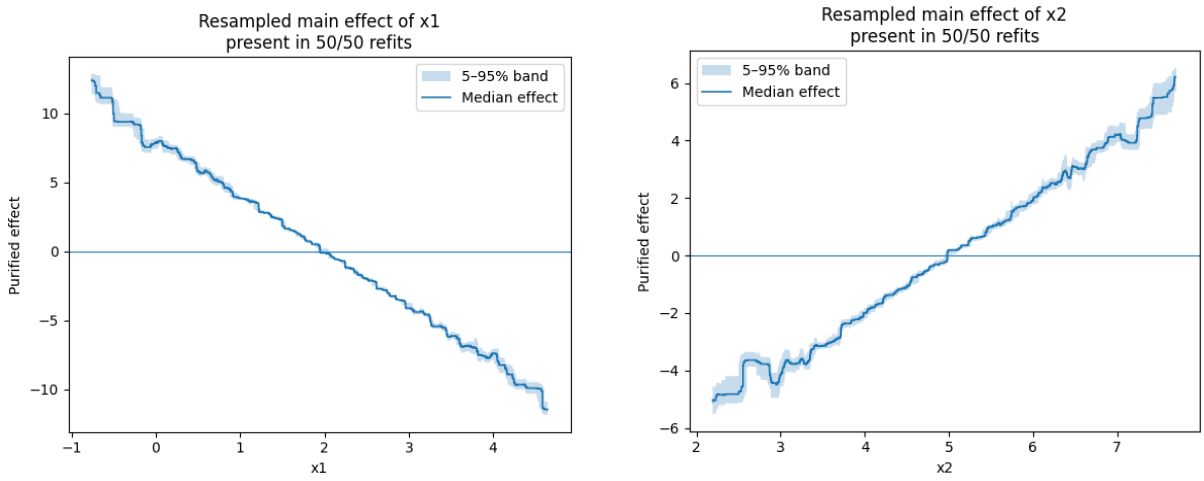
Figure 14: Pseudo-Shapley effects for the independent linear model.

The difference between the two variables is more pronounced for the variance-based measure, since variance contributions depend quadratically on effect magnitude. While the SHAP main-effect magnitudes differ by a factor of two, the corresponding variance contributions differ by a factor of four. The theoretical variance-allocation values are approximately 0.79 for X_1 and 0.21 for X_2, and the recovered pseudo-Shapley effects

are close to these values. This illustrates that the two importance measures answer different questions: mean absolute SHAP values measure the typical magnitude of local contributions, while pseudo-Shapley effects measure contribution to variability in the model output.

The resampling procedure was then applied to assess the stability of the recovered effects. The training and test sets were resampled 50 times, and the model was refitted for each resample. The median and quantile bands were computed using the procedure described in Section 3.2.

Figure 15 shows the resulting resampled main effects.



(a) X_1 (b) X_2

Figure 15: Resampled purified main effects for the independent linear model.

The uncertainty bands are tight in the central regions of the input distribution, where the resampled training sets contain similar observations. They widen in the tails, where the presence and exact location of extreme observations vary more between resamples.

In addition, localized regions of higher uncertainty appear around several split locations. This does not necessarily mean that the fitted models disagree about the overall effect. Rather, the resampled XGBoost models often place splits in similar parts of the feature space, but not at exactly the same thresholds. Since the recovered effects are piecewise constant and the resampling summary is evaluated over the collected bins from the different fitted models, small shifts in split location cause pre- and post-jump values to be mixed in narrow regions. This produces local uncertainty bands around common split regions even when the broader functional shape is stable.

The same phenomenon is visible for the interaction effect in Figure 16. Here the uncertainty is highest in regions where many splits are placed for both variables, particularly around the feature means. This reflects the fact that small changes in the two-dimensional partition can shift the location of plateaus in the recovered interaction surface.

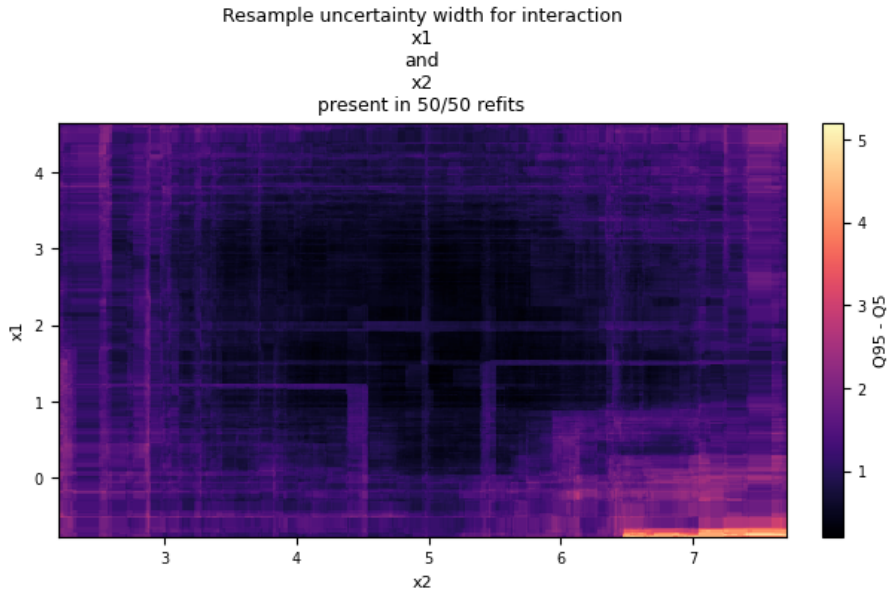


Figure 16: Resampled purified interaction effect for the independent linear model.

Overall, the independent linear simulation shows that the proposed approach can recover the theoretical functional ANOVA effects well when the target function is directly observed. The recovered effects closely match the theoretical effects in regions with sufficient data support. The remaining discrepancies are mainly caused by the piecewise-constant tree approximation, limited data support in boundary regions, and local instability in the exact placement of split thresholds across resampled fits.

4.1.2 Dependent Linear Model

We next consider the linear model with dependent features. A train/test split was performed and the model was fitted using the same default settings as in the independent case, apart from the depth restriction. The model obtained an RMSE of 0.55, slightly lower than in the independent setting. This likely reflects that the strong dependence structure concentrates the observations in a smaller effective region of the feature space, making the response surface easier to approximate with the available splits.

Figure 17 compares the recovered purified main effects with the theoretical generalized functional ANOVA effects.

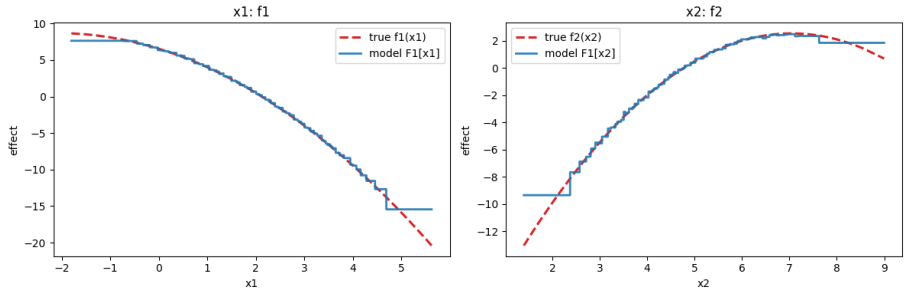


Figure 17: Comparison between the purified and theoretical main effects for the dependent linear model.

Despite the strong correlation between the inputs, the recovered main effects align

closely with the theoretical effects in the central region of the distribution. The largest deviations occur near the edges, where there are fewer observations and the tree approximation is less stable.

For the interaction effect, Figure 18 shows the pointwise difference between the recovered and theoretical interaction effects.

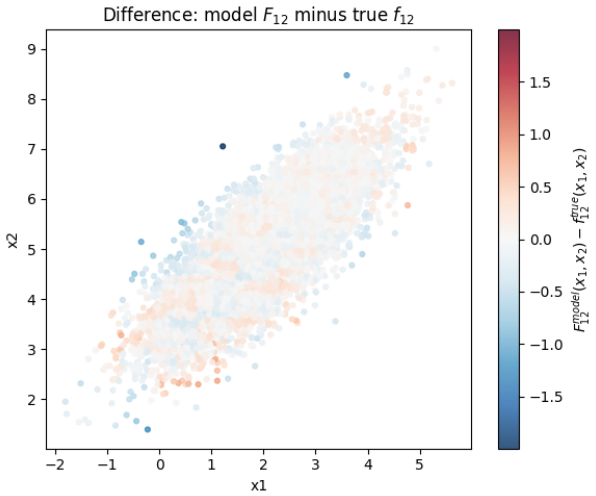


Figure 18: Difference between the purified and theoretical interaction effects for the dependent linear model.

The interaction differences are smaller than in the independent case. This is consistent with the dependence structure concentrating the observations in a denser region of the input space. With the same depth restriction, the model can therefore use its splits more efficiently in the region where observations are likely to occur, rather than approximating the interaction surface over the full rectangular feature space.

Table 3 summarizes the corresponding difference diagnostics.

Effect	Mean Difference	RMSE Difference	Max Absolute Difference
f_1	0.047434	0.174638	4.954616
f_2	-0.003583	0.131474	3.700384
f_{12}	0.006277	0.129659	1.995569

Table 3: Difference diagnostics between theoretical and recovered effects for the dependent linear model.

Compared with the independent case in Table 2, the RMSE values are smaller for all effects. The mean differences are also close to zero, indicating little systematic deviation from the theoretical decomposition. The larger maximum deviations for the main effects are mainly localized to edge regions, while the overall recovery is accurate in the high-density part of the input distribution.

Figure 19 shows the global SHAP feature importance.

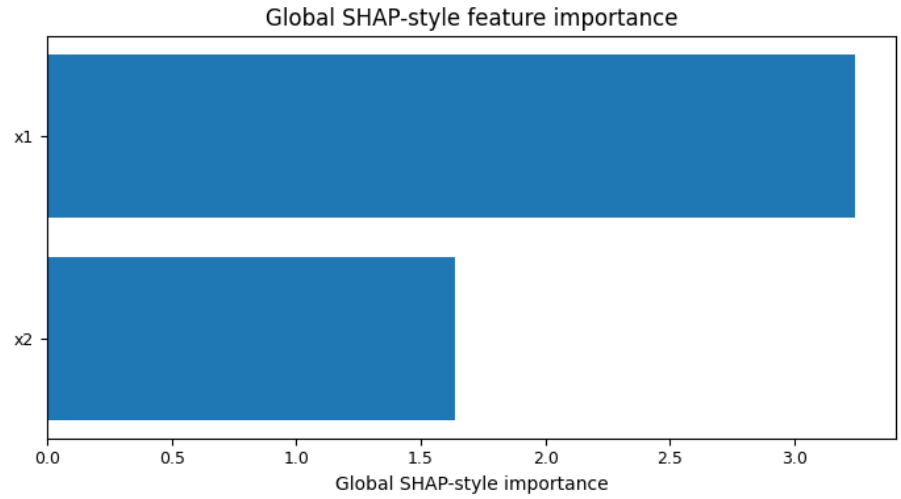


Figure 19: Global SHAP feature importance for the dependent linear model.

The SHAP-based importance is broadly similar to the independent case. Although dependence changes the theoretical functional ANOVA components, the average magnitudes of the local feature contributions remain relatively similar.

Larger differences appear for the pseudo-Shapley effects, shown in Figure 20.

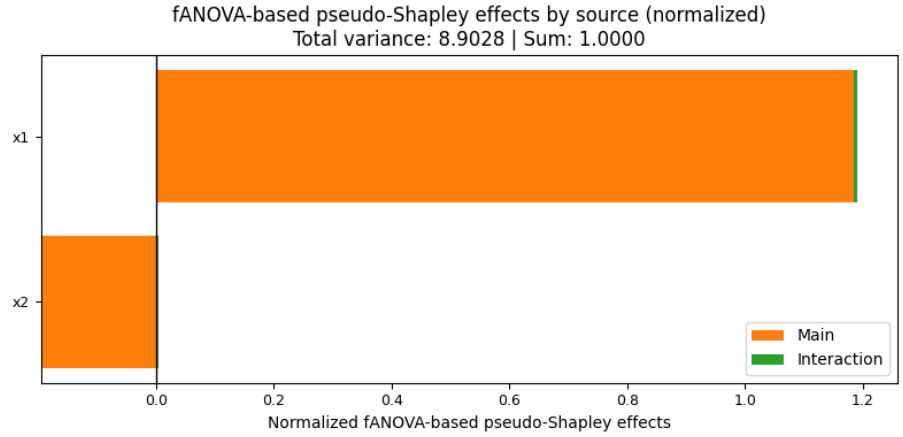


Figure 20: Pseudo-Shapley variance importance for the dependent linear model.

The contribution assigned to X_2 becomes negative under the pseudo-Shapley variance allocation. This occurs because the allocation includes covariance terms between functional ANOVA components. For the dependent linear model, the canonical main effects are

$$f_1 = pU + m(U^2 - 1), \qquad f_2 = qV + m(V^2 - 1), \qquad m = \frac{k\rho}{1 + \rho^2}.$$

Using standard moments of the jointly normal distribution gives

$$\mathrm{Var}\,(f_1) = p^2 + 2m^2, \qquad \mathrm{Var}\,(f_2) = q^2 + 2m^2, \qquad \mathrm{Cov}(f_1, f_2) = pq\rho + 2m^2\rho^2.$$

The derivation is provided in Appendix A.1.

In the present parameter setting, the effective slopes of the two main effects have opposite signs, while the variables themselves are strongly positively correlated. Consequently, the covariance between the main effects becomes strongly negative. Substituting

the parameter values gives

$$\mathrm{Var}\,(f_2) \approx 4.48, \qquad \mathrm{Cov}(f_1, f_2) \approx -6.10.$$

The negative covariance contribution is therefore larger in magnitude than the positive variance contribution of the second main effect itself. As a result, the variance allocated to X_2 under the pseudo-Shapley decomposition becomes negative.

This should not be interpreted as meaning that X_2 is unimportant. Rather, it indicates that the variability associated with X_2 is offset by covariance cancellation under the dependent input distribution. This illustrates an important distinction between the two importance summaries. The SHAP-based measure reflects the typical magnitude of local feature contributions, while the pseudo-Shapley effect describes a variance allocation that also accounts for covariance between components. Under dependence, these two notions of importance can therefore lead to substantially different conclusions.

The same resampling procedure was performed as in the independent case. The resulting uncertainty bands were tighter overall, reflecting that the strong dependence concentrates observations in a smaller and denser effective region of the input space. The split locations were therefore more stable across resamples, and the sparse corner regions contributed little to the resampling variability. The importance measures were also stable across resamples, with narrow uncertainty intervals. The corresponding resampling figures are provided in Appendix A.2.

Overall, the dependent linear model shows that the aggregation and purification procedure can recover the theoretical generalized functional ANOVA effects even under strong dependence between the inputs. The results also illustrate that dependence affects more than the shape of the component functions. Because the components are no longer mutually orthogonal, covariance terms enter the variance allocation and may substantially change the interpretation of global feature importance.

4.1.3 Binary response with latent score

The previous simulations show that the proposed approach can recover the theoretical functional ANOVA effects when the response surface is observed directly. However, as discussed in Section 3.3, in a binary classification setting the latent score function is not observed directly. The model instead observes binary realizations generated from probabilities determined by the latent score, making recovery of the underlying functional ANOVA effects substantially more difficult.

Four model specifications were evaluated. The first was a classification model using the default XGBoost hyperparameters together with the depth-2 restriction. The second was a tuned classification model, while the third additionally imposed monotonicity constraints. Technically, under dependence the theoretical main effects are quadratic rather than globally monotone. However, within the observed sample region the main effects are effectively monotone, since the opposite side of the curve has very low probability mass. The final specification was a regression XGBoost model fitted directly to the latent score η, corresponding to the setting considered in the previous simulations.

All classification models achieved broadly similar predictive performance. The default-parameter model performed slightly worse, with a test ROC AUC of approximately 0.93, while the tuned and monotone models achieved ROC AUC values around 0.94. The latent-score regression model is included as a reference for how well the functional structure can be recovered when the score itself is observed.

Figure 21 shows the recovered purified main effects for the different model specifications.

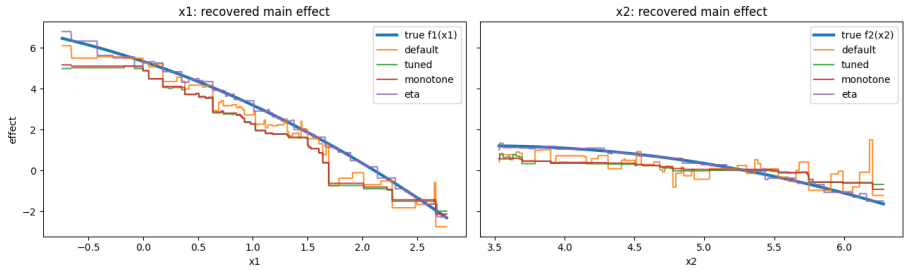


Figure 21: Recovered purified main effects for the latent binary-response models under different training specifications.

Compared with the regression case where the target function is observed directly, the binary-response setting is substantially more difficult. The default-parameter classification model displays considerable local variability, but recovers the overall shape of the X_1 effect with less systematic shrinkage. The tuned models produce smoother effects, making the functional shape easier to identify visually, but at the cost of stronger shrinkage towards zero. For X_2, the recovered effect is noisier overall, and tuning appears to provide a clearer improvement.

Figure 22 compares the recovered purified interaction effects evaluated at the observations with the theoretical interaction effect.

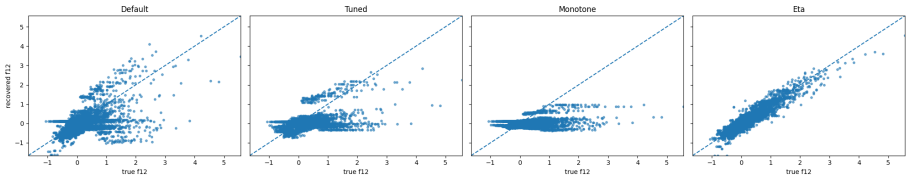


Figure 22: Recovered purified interaction effects against the theoretical interaction effect for the latent binary-response models.

The interaction effects are substantially harder to recover than the main effects. The default model retains more visible interaction structure, but with considerable dispersion around the theoretical effect. The tuned models reduce this variability and produce a more concentrated relationship, but also shrink the interaction more strongly towards zero. This reflects the same bias–variance tradeoff observed for the main effects, but amplified for the interaction effect.

The monotone model exhibits the strongest shrinkage of the interaction effect, despite using the same hyperparameter configuration as the tuned unconstrained model. This suggests that monotonicity constraints act as an additional form of regularization.

Table 4 reports quantitative recovery metrics for the purified functional ANOVA effects. The recovered effects are compared with the theoretical purified effects from the data-generating process. For each effect, the table reports the MAE, RMSE, relative mean squared error, correlation with the theoretical effect, and the reconstruction error for the constant component f_0.

Table 4: Recovery metrics for purified functional ANOVA effects.

Model	Effect	MAE	RMSE	RelMSE	Corr.	f_0 error
Default	f_1	0.888	1.464	0.204	0.905	0.770
	f_2	0.532	0.763	0.413	0.803	
	f_{12}	0.306	0.455	0.989	0.420	
Tuned	f_1	1.213	1.772	0.299	0.872	1.356
	f_2	0.608	0.847	0.509	0.896	
	f_{12}	0.300	0.436	0.909	0.364	
Monotone	f_1	1.149	1.705	0.277	0.885	1.306
	f_2	0.566	0.777	0.429	0.937	
	f_{12}	0.301	0.444	0.944	0.245	
Score regression	f_1	0.115	0.161	0.002	0.999	0.011
	f_2	0.075	0.110	0.009	0.996	
	f_{12}	0.122	0.170	0.137	0.931	

The recovery metrics are consistent with the visual results. When the latent score is observed directly, the theoretical functional ANOVA effects are recovered almost exactly. In the binary-response setting, recovery is substantially less accurate, especially for the interaction effect. The main effects retain relatively high correlations with their theoretical counterparts, while the interaction effect has much weaker correlation and relative mean squared error close to one.

The tuned and monotone models produce smoother estimates, but this comes with stronger shrinkage of the recovered effects. The default model is more variable, but retains more of the interaction structure. Additional simulations with 100,000 observations showed that the tuned model improves substantially with more data, producing both smoother functions and better overall recovery. The monotone model, however, continued to display stronger shrinkage of the interaction effect.

Overall, the binary-response simulation shows that recovering functional ANOVA effects from binary outcomes is considerably more difficult than recovering them from an observed response surface. The main effects can still be recovered to a useful degree, but interaction effects are much more sensitive to noise, regularization, and modelling constraints. This is an important limitation for the real-data setting, where the score function is not observed directly and the available number of default observations is limited.

4.1.4 Credit-risk toy model

The final simulation considers the credit-risk toy model, where the objective is to illustrate how interaction structure may already be embedded in ratio-based covariates. Two models were compared: one using the underlying accounting quantities (A, L, P) as predictors, and one using the ratios $R_L = L/A$ and $R_P = P/A$ directly.

The resulting test ROC AUC was approximately 0.74 for the underlying-quantity model and 0.76 for the ratio-based model. The latter is close to the maximum achievable performance in this setup, since the ROC AUC obtained using the true underlying score is approximately 0.76. Thus, both models capture most of the available predictive signal, although the ratio-based specification is closer to the data-generating structure.

The recovered interaction effects differ substantially between the two feature repre-

sentations. For the underlying-quantity model, weak but visible interaction structure remains between the underlying accounting quantities. For the ratio-based model, however, the purified interaction effect is largely dominated by noise and contains little stable structure. The corresponding interaction plots are provided in Appendix A.3.

This is consistent with the construction of the toy model. The latent score is additive in the ratios R_L and R_P, but the ratios themselves are nonlinear functions of the underlying quantities A, L, and P. Consequently, interaction structure is expected when the model is expressed in terms of the original accounting quantities, while much of this structure disappears when the economically meaningful ratios are used directly.

The toy model therefore highlights an important point for the real-data application. Weak recovered interaction effects should not necessarily be interpreted as evidence that no interaction exists in the underlying financial process. Rather, the chosen feature representation may already encode part of this structure, particularly when the predictors are accounting ratios constructed from common underlying quantities.

Taken together, the simulation studies illustrate both the strengths and limitations of the proposed decomposition approach. When the underlying response surface is observed directly, the proposed approach recovers the theoretical functional ANOVA components closely, even under strong dependence. When the response is latent and observed only through a binary outcome, recovery becomes substantially more difficult, especially for interaction effects. The toy model further shows that the apparent strength of recovered interactions depends on the chosen feature representation, since ratio variables may already encode nonlinear structure from the underlying accounting quantities.

These observations motivate the real-data analysis that follows. In the real-data setting, theoretical effects are no longer available for comparison. The focus therefore shifts from recovery of known functional components to the interpretability, stability, and economic plausibility of the decomposed effects obtained from the fitted credit-risk model.

4.2 Real Data Results

In the real-data application, the theoretical functional ANOVA effects are not available for comparison. The analysis therefore focuses on whether the fitted model can be represented through stable and economically interpretable low-order effects while retaining reasonable predictive performance. The real-data model was developed in two stages. First, an initial depth-2 XGBoost model was fitted using the retained feature set from Table 1 in Section 3.4. This model was used to assess predictive performance and to identify variables whose recovered effects were unstable, difficult to interpret, or of limited importance. A final reduced model was then fitted after removing such variables and imposing monotonicity constraints where needed.

4.2.1 Initial model and final specification

The initial model was fitted using the full retained feature set and the tuning procedure described in Section 3.4.2. For comparison, two additional models were fitted: an XGBoost model in which deeper trees were allowed, and a logistic regression model with tuned elastic-net regularization. When deeper trees were allowed, cross-validation again selected a depth-2 model, suggesting limited predictive benefit from higher-order interactions in this dataset.

Table 5 reports the test-set performance of the initial constrained XGBoost model and the logistic regression benchmark.

Table 5: Predictive performance on the test set for logistic regression and the initial constrained depth-2 XGBoost model.

Model	ROC AUC	Accuracy	Recall	Precision
Logistic regression	0.854	0.741	0.857	0.052
Depth-2 XGBoost	0.879	0.922	0.524	0.111

The comparison should be interpreted with care. The logistic regression coefficients did not fully converge, and no detailed logistic-regression specification search was performed. In a practical credit-risk modelling setting, substantially more attention would normally be given to multicollinearity, transformations, variable selection, and other specification choices. The comparison is therefore not intended to establish that the depth-2 XGBoost model is generally superior to logistic regression. Rather, it serves as a benchmark showing that the constrained XGBoost model achieves reasonable discriminatory performance in this dataset.

The constrained XGBoost model obtained a higher test ROC AUC than logistic regression, indicating stronger ranking ability between bankrupt and non-bankrupt firms. The threshold-dependent classification metrics are less central, since they were computed using the default probability threshold of 0.5 in a highly imbalanced dataset. The XGBoost model achieved higher precision for the bankruptcy class, while logistic regression achieved higher recall, reflecting a more aggressive classification behaviour with more predicted bankruptcies.

The initial constrained XGBoost model selected through cross-validation consisted of 400 depth-2 trees with learning rate 0.03. The selected hyperparameters also implied relatively strong regularization and imbalance correction, with `min_child_weight` $= 20$ and `scale_pos_weight` ≈ 30. Overall, the fitted model was therefore a conservative low-order nonlinear model.

Although the initial model achieved reasonable predictive performance, several recovered effects were difficult to interpret. Some variables had low importance and resampling intervals that largely covered zero, while others produced unstable or economically counterintuitive main effects. This motivated a second modelling step in which variables with limited importance or problematic effect shapes were removed. Variables that were important but displayed counterintuitive effects were first retained and examined further, since such behaviour may itself reveal limitations of the feature representation.

Monotonicity constraints were also considered for variables whose learned effects displayed implausible directions or unstable behaviour in central regions of the feature distribution. In practice, the only variable for which a monotonicity constraint was imposed in the final model was `working capital / total assets`, due to the large drop observed around 0.2 in the initial model. Other variables displayed smaller local irregularities or edge-region behaviour, but constraints were not imposed more broadly in order to avoid adding unnecessary modelling assumptions in the limited-data setting.

The final model included 11 features, shown in Table 6. Hyperparameters were retuned after the feature reduction and monotonicity constraint were introduced.

Table 6: Features included in the final XGBoost model.

Final model features
operating expenses / total liabilities
retained earnings / total assets
(equity - share capital) / total assets
constant capital / total assets
gross profit (in 3 years) / total assets
total liabilities / total assets
logarithm of total assets
net profit / total assets
gross profit / short-term liabilities
working capital / total assets
long-term liabilities / equity

Table 7 reports the test-set performance of the final reduced-feature models.

Table 7: Test-set predictive performance for the reduced-feature models. Precision and recall are reported for the bankruptcy class.

Model	**ROC AUC**	**Accuracy**	**Recall**	**Precision**
Logistic regression	0.829	0.725	0.857	0.050
Depth-2 XGBoost	0.866	0.876	0.667	0.090

Both models experienced a reduction in predictive performance compared with the initial specification, indicating that some removed variables contained predictive information. Nevertheless, the final XGBoost model retained stronger ranking ability than logistic regression, with a test ROC AUC of 0.866 compared with 0.829. The reduction in predictive performance is therefore moderate relative to the gain in interpretability from a smaller feature set and more stable effect structure.

Figure 23 illustrates this ranking behaviour by showing the density-normalized distributions of predicted probabilities for the two outcome classes in the test set. The bankrupt firms are assigned higher predicted probabilities on average, although the two distributions still overlap. Since the densities are normalized separately by class, the figure compares the shapes of the distributions rather than the class proportions.

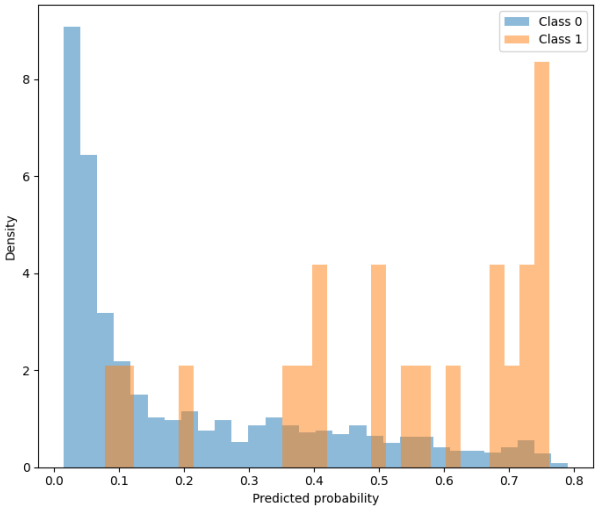


Figure 23: Density-normalized distributions of model-predicted probabilities in the test set.

The selected hyperparameters for the final model again implied a conservative and strongly regularized fit, with a low learning rate, substantial L2 regularization, and a high split penalty. The selected imbalance weighting was approximately half the empirical class imbalance ratio, suggesting that partial rather than full imbalance correction generalized best in cross-validation. As seen in the latent-link simulation, such regularization may involve a trade-off between stability and exact recovery of underlying functional effects. This should be kept in mind when interpreting the recovered real-data components.

4.2.2 Global feature importance

We first examine the global importance measures for the final reduced model. Figures 24 and 25 show the mean absolute SHAP importance and the pseudo-Shapley variance allocation, respectively.

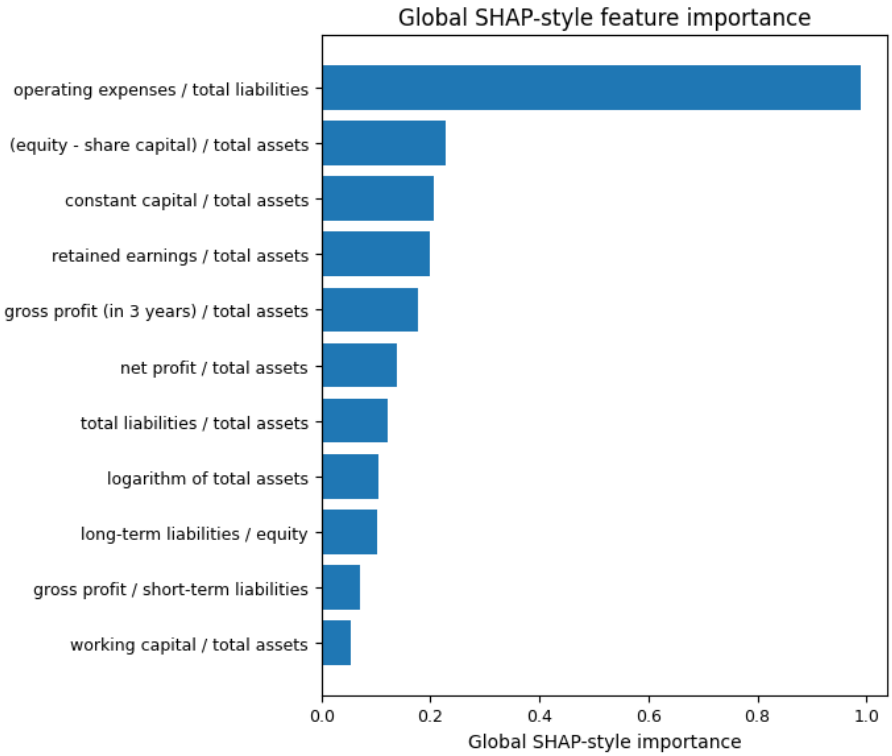


Figure 24: Global SHAP-based feature importance for the final constrained XGBoost model.

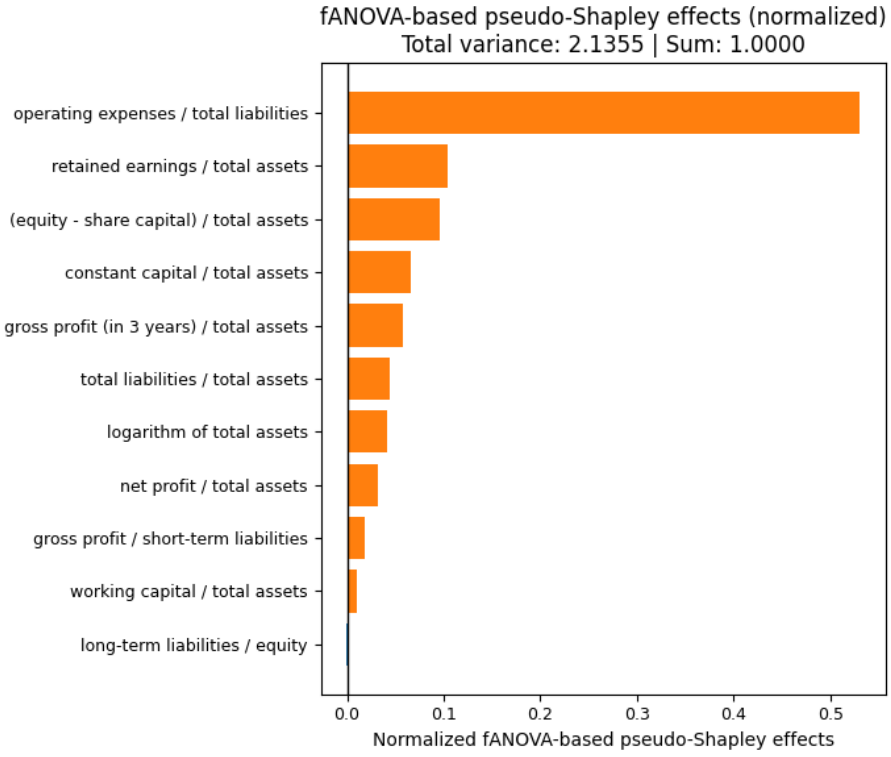


Figure 25: Pseudo-Shapley variance importance for the final constrained XGBoost model.

Both importance measures identify `operating expenses / total liabilities` as the dominant variable. This is plausible in a short-term bankruptcy prediction setting,

since the ratio can be interpreted as a measure of how large operating expenses are relative to the firm's liabilities. The final transition into bankruptcy is often associated with deteriorating liquidity or solvency conditions, making such a variable a natural source of short-term predictive information.

The two importance measures agree on the dominant role of this variable, but they differ in how importance is distributed among the remaining features. This reflects the distinction discussed in Section 3.2. Mean absolute SHAP importance measures the typical magnitude of local feature contributions, while the pseudo-Shapley effect allocates variance in the fitted model output across variables. A variable may therefore have non-negligible local contribution magnitudes while still contributing little to the overall output variance.

This is visible for `long-term liabilities / equity`, which receives some local attribution importance but has an approximately zero net pseudo-Shapley contribution. This should not be interpreted as meaning that the variable is unused by the model. Rather, positive and negative covariance contributions largely cancel under the variance-based allocation. Thus, as in the dependent simulation, the pseudo-Shapley effect reflects net contribution to output variability rather than the magnitude of local use in individual predictions.

Overall, the importance measures suggest that the predictive structure of the final model is concentrated in a small number of financial ratios. The remaining variables contribute more modestly, either through smaller local effects, covariance-adjusted variance contributions, or specific local corrections.

4.2.3 Main effects in the final model

The intercept and purified main effects for the final model are shown in Figure 26. Since the model contains 11 retained features, the purpose of the figure is not to inspect each effect in isolation, but to assess whether the recovered low-order representation provides a coherent global description of the fitted score function.

When interpreting such effects in an applied modelling setting, one would normally also consider the empirical distribution of the underlying feature values in detail. In the present thesis, however, the data set is primarily used to illustrate the proposed decomposition framework on a real data set with relevance for credit-risk modelling, rather than to perform a detailed empirical study of the firms themselves. Moreover, the available background information about the data is limited compared with what would be available in an actual modelling situation. The evaluation is therefore focused on whether the decomposition provides a coherent and interpretable representation of the fitted model. Selected examples showing how the estimated effects relate to the empirical feature distributions are provided in Appendix A.4.

Overall, several of the recovered main effects are broadly consistent with economic intuition. For example, `total liabilities / total assets` displays an increasing relationship with the fitted bankruptcy score, while `gross profit (in 3 years) / total assets` and `retained earnings / total assets` display decreasing relationships over important parts of the feature space. This suggests that the decomposed model captures several plausible risk-driver relationships.

Several effects also exhibit threshold-like or regime-switching behaviour. For `gross profit / short-term liabilities` and `net profit / total assets`, the effects are relatively flat over parts of the positive region but differ more clearly around zero. This is plausible for profitability-type ratios, since the distinction between negative and positive

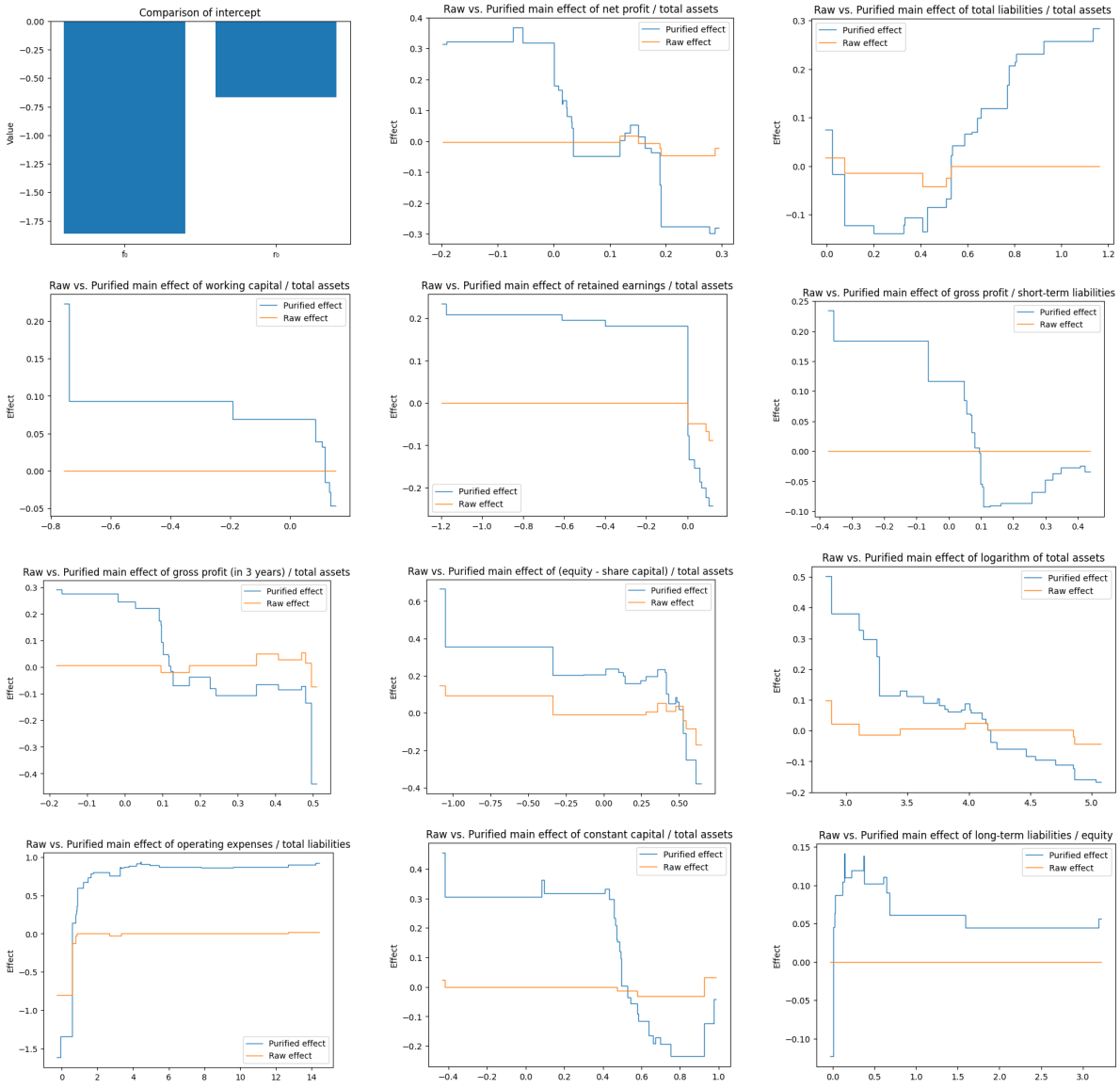


Figure 26: Intercept and purified main effects for the final depth-2 XGBoost model.

profitability may be more informative than small changes within the positive range. Similarly, `retained earnings / total assets` suggests a different relationship depending on whether retained earnings are negative or positive.

Other effects are less straightforward to interpret. For example, `constant capital / total assets` contains a relatively large plateau, and some variables display local jumps or edge-region behaviour that is difficult to assess economically. Such patterns may partly reflect the piecewise-constant nature of the fitted tree ensemble and the limited number of observations in some regions of the feature space. They may also reflect the fact that several financial ratios combine accounting quantities with different economic interpretations, making it difficult to determine whether the predictive signal comes from the ratio itself, its numerator, its denominator, or interactions with other variables.

The main effects therefore illustrate both the usefulness and the limitations of the representation. On the one hand, the functional ANOVA decomposition makes it possible to inspect the fitted score function through a small number of one-dimensional components. On the other hand, interpretability still depends on the economic meaning and stability of the input variables themselves. Ratios that are unstable, highly dependent on related variables, or difficult to interpret economically may still produce effects that are hard to justify, even when the model representation is transparent.

4.2.4 Interaction effects

The final model also contains purified pairwise interaction effects. Overall, these effects were small in magnitude compared with the main effects, suggesting that most of the fitted score function is explained by main effects. This is consistent with the depth-2 model specification and with the results from the toy model, where ratio-based predictors were shown to absorb part of the interaction structure between underlying accounting quantities.

Two representative interaction effects are shown in Figures 27 and 28. The first concerns `gross profit (in 3 years) / total assets` and `net profit / total assets`, while the second concerns `constant capital / total assets` and `net profit / total assets`.

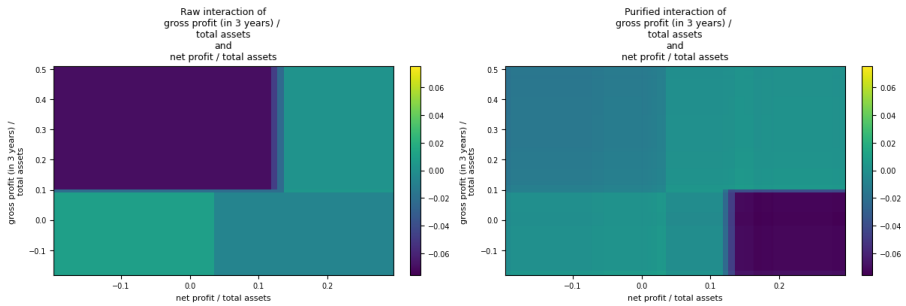


Figure 27: Raw and Purified interaction effect between `gross profit (in 3 years) / total assets` and `net profit / total assets`.

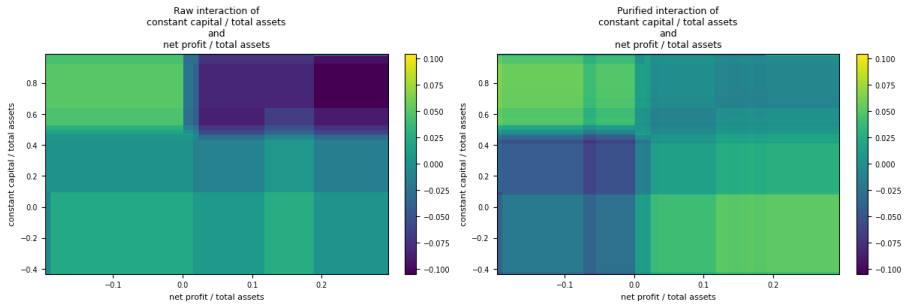


Figure 28: Raw and Purified interaction effect between `constant capital / total assets` and `net profit / total assets`.

As discussed previously, purified interaction effects should be interpreted as deviations from additive main-effect behaviour rather than as standalone risk relationships. The first interaction can be interpreted as suggesting that high net profit partly compensates for weak long-term gross profitability. The second interaction has a more saddle-like structure, where the effect of net profit depends on the level of constant capital and vice versa.

However, both interaction effects are modest in magnitude. Their economic interpretation should therefore be treated as secondary to the main effects. In this application, the interaction components appear to function mainly as local corrections to the additive structure rather than as dominant drivers of the fitted bankruptcy score.

4.2.5 Local attribution

In addition to global main and interaction effects, the functional ANOVA representation also supports local attribution. As described in Section 3.2, SHAP values can be computed directly from the functional ANOVA components by distributing each component equally among the variables on which it depends. This provides a feature-level decomposition of individual predictions without requiring a separate post-hoc explainer.

Figures 29, 30, and 31 show local SHAP decompositions for a true negative, a true positive, and a false positive prediction, respectively. The examples are based on the default classification threshold of 0.5.

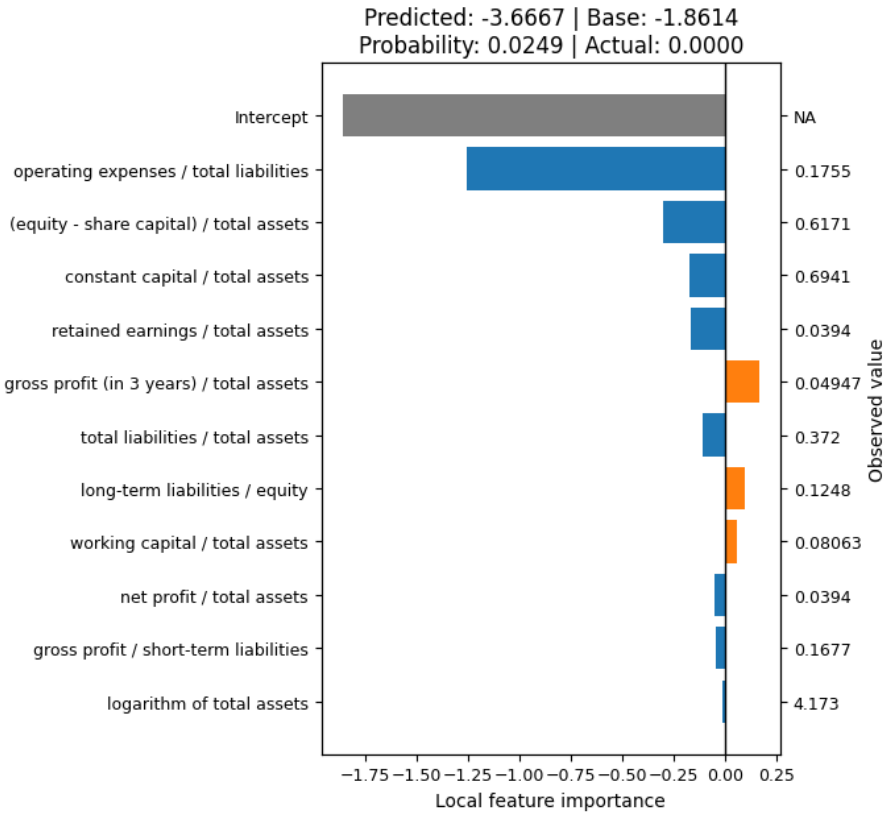


Figure 29: Local SHAP decomposition for a true negative prediction.

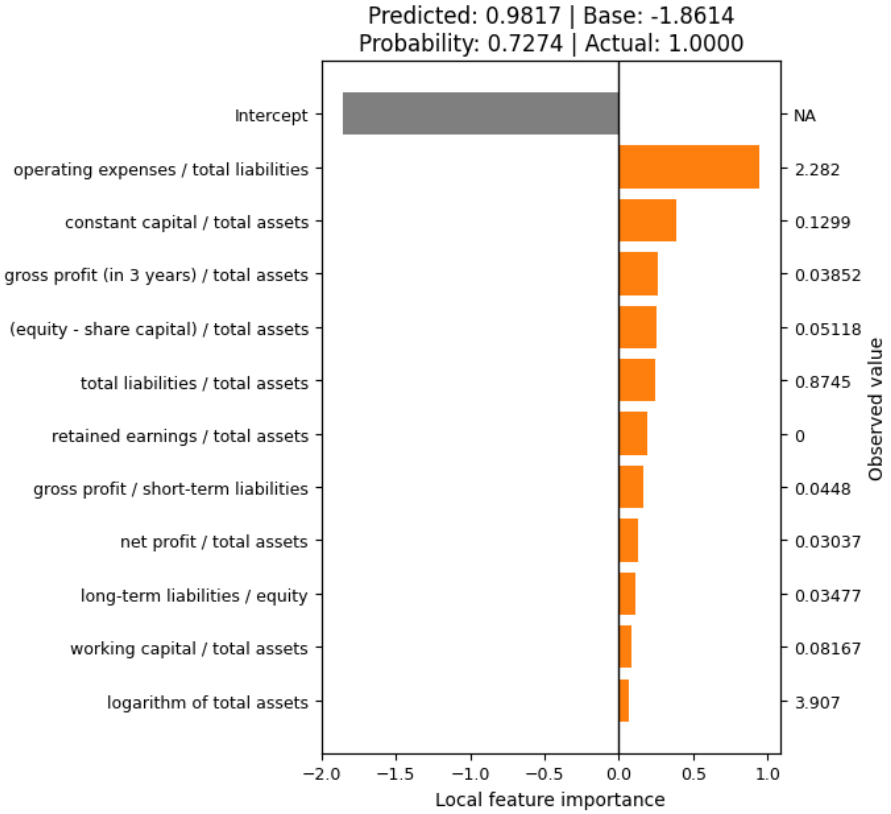


Figure 30: Local SHAP decomposition for a true positive prediction.

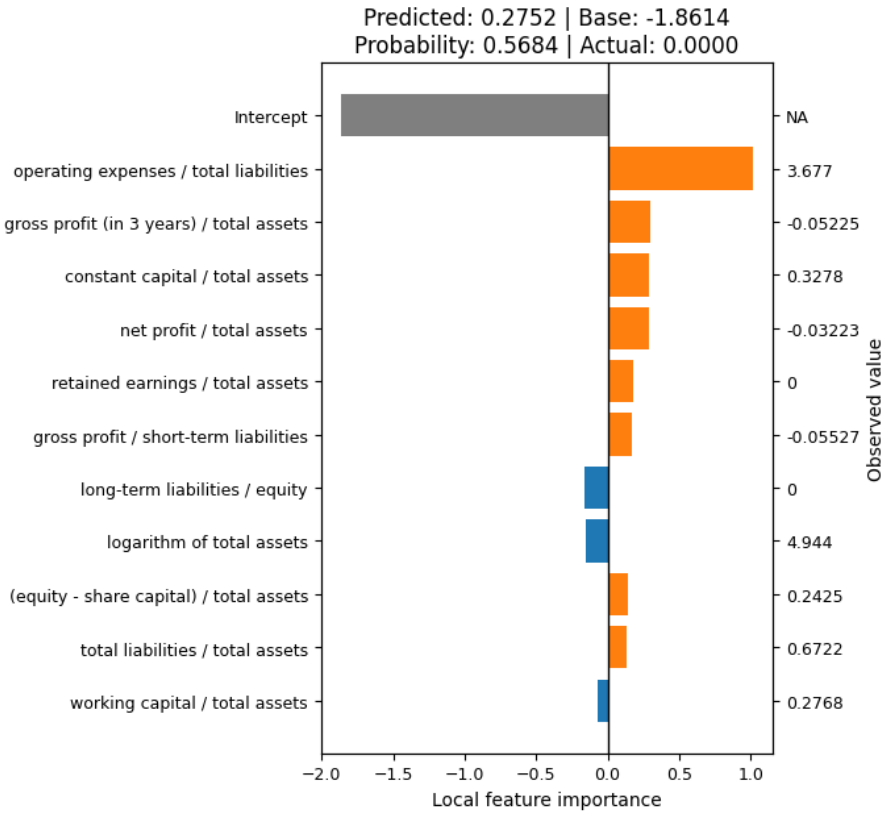


Figure 31: Local SHAP decomposition for a false positive prediction.

For the true negative prediction, most variables contribute towards a lower fitted bankruptcy score. The largest contribution comes from `operating expenses / total liabilities`, consistent with its dominant role in the global importance measures.

For the true positive prediction, `operating expenses / total liabilities` again contributes strongly, but the prediction is less dominated by a single variable. Instead, several variables jointly increase the fitted bankruptcy score. This illustrates how the local decomposition can identify whether a prediction is driven by one dominant signal or by the accumulation of several smaller contributions.

For the false positive prediction, the local decomposition helps explain why the model assigned high risk even though the firm did not go bankrupt. The observation has operating expenses almost four times larger than total liabilities, negative profitability, and low constant capital relative to total assets. These characteristics provide economically plausible reasons for the elevated fitted score, even though the classification is ultimately incorrect.

The local decompositions therefore illustrate how the fitted model can be inspected at the level of individual predictions. Because the attributions are computed directly from the decomposed functional ANOVA representation, they remain connected to the same main and interaction effects used in the global analysis.

4.2.6 Resampling stability

Finally, the resampling procedure described in Section 3.2 was applied to the final model in order to assess the stability of both the importance measures and the recovered effects. The same hyperparameters were used across resampled fits, so the intervals reflect variability induced by the sampled data rather than repeated hyperparameter tuning.

Figures 32 and 33 show the resampled medians and quantile intervals for the normalized mean absolute SHAP importance and pseudo-Shapley variance allocation. The importance values were normalized within each resampled model, since different resampled fits may have different overall prediction magnitudes and total output variance. Consequently, the reported medians are computed from normalized quantities and do not necessarily sum to one.

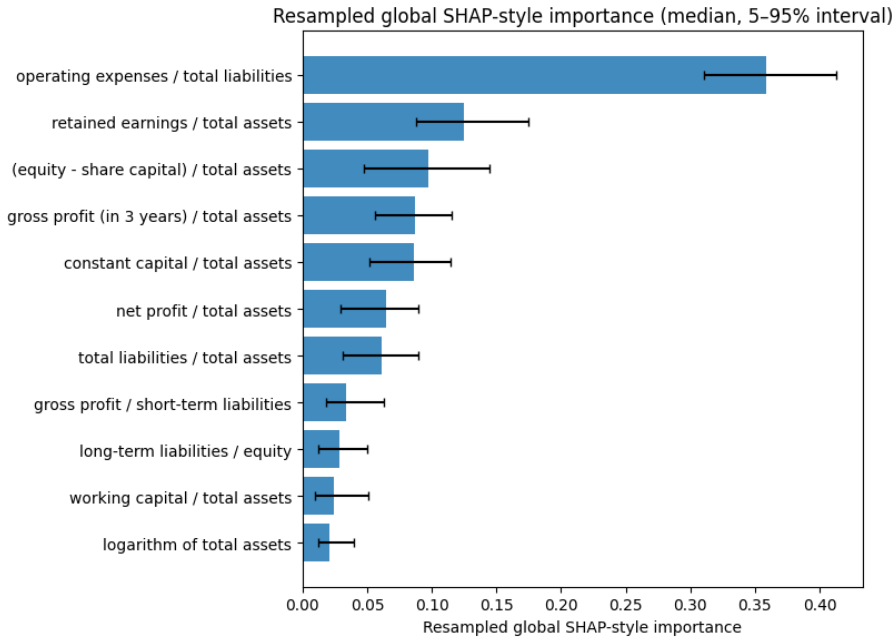


Figure 32: Resampling intervals for the normalized global SHAP-based feature importance.

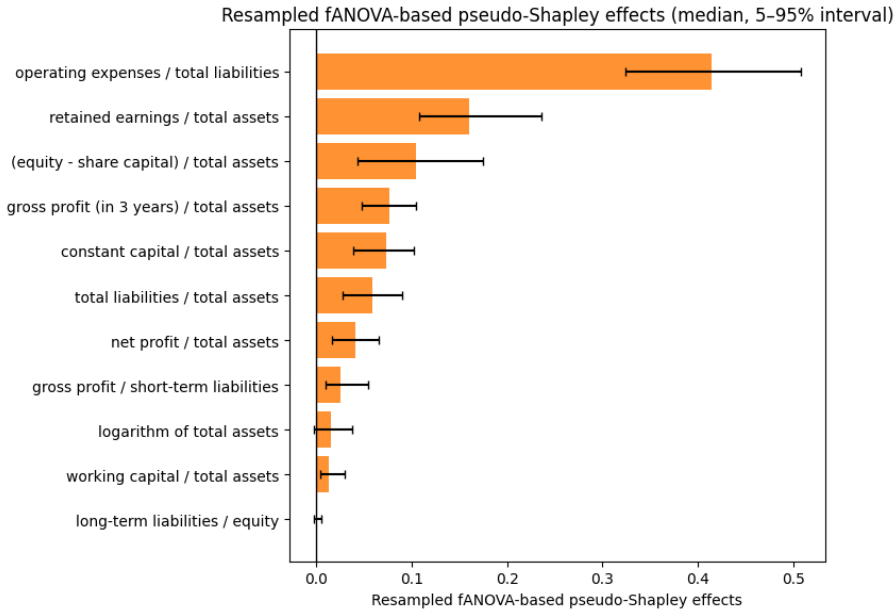


Figure 33: Resampling intervals for the normalized pseudo-Shapley effects.

The resampling intervals support the main conclusions from the fitted model. Both importance measures consistently identify `operating expenses / total liabilities`

as the dominant variable. The remaining variables show lower and more variable importance, but the overall ranking structure is still reasonably stable. The pseudo-Shapley intervals also show that some variables may have small or near-zero net variance contributions even when they have non-negligible SHAP importance, again reflecting the difference between local contribution magnitude and covariance-adjusted variance allocation.

Figure 34 shows the resampled medians and quantile intervals for the purified main effects.

Overall, the resampled main effects align well with the effects from the fitted final model, indicating that the dominant functional patterns are reasonably stable across resampled datasets. Several effects become smoother in the resampled medians, suggesting that some of the local irregularities in the fitted effects are not stable across samples.

For example, `gross profit (in 3 years) / total assets`, `gross profit / short-term liabilities`, and `retained earnings / total assets` display clearer threshold-like or regime-switching behaviour in the resampled summaries. In these cases, the resampling results suggest increased bankruptcy risk around negative profitability values, followed by a lower or flatter contribution once profitability becomes positive.

Some effects are less stable. For example, `logarithm of total assets` becomes flatter in the resampled median than in the fitted model, and its quantile interval overlaps zero over larger parts of the feature range. This suggests that the effect of firm size is weaker and less stable than the dominant financial-ratio effects. Similarly, `gross profit / short-term liabilities` and `long-term liabilities / equity` have intervals that overlap zero in parts of the positive range, indicating that their main contribution may come from specific regions rather than from a stable effect over the entire feature space.

Taken together, the resampling results indicate that the main qualitative structure of the final model is stable, especially for the most important variables. At the same time, weaker effects and local irregularities are less stable, which supports interpreting them cautiously. This is consistent with the simulation results, where main effects were more stable than interaction effects and small local features were sensitive to split placement and sample variation. Overall, the results support the usefulness of the proposed functional ANOVA framework, but also highlight several limitations. In the simulation studies, the aggregation and purification procedure recovered the theoretical functional ANOVA components closely when the response surface was observed directly, including under strong dependence between predictors. When the target function was latent and observed only through a binary response, recovery became substantially more difficult, especially for interaction effects.

The real-data application showed that the final depth-2 XGBoost model could be represented through a relatively small set of interpretable main effects and weaker pairwise interaction corrections. The dominant variables and several of the recovered main-effect shapes were economically plausible, and the most important effects remained stable under resampling. At the same time, some effects were difficult to interpret, particularly the interactions.

Taken together, the results suggest that the proposed framework can provide a useful compromise between nonlinear predictive modelling and functional interpretability in a credit-risk setting. However, the results also show that the interpretability of the decomposition depends not only on the modelling method, but also on the feature representation, the amount of available default information, and the stability of the recovered effects. These issues are discussed further in the following section.

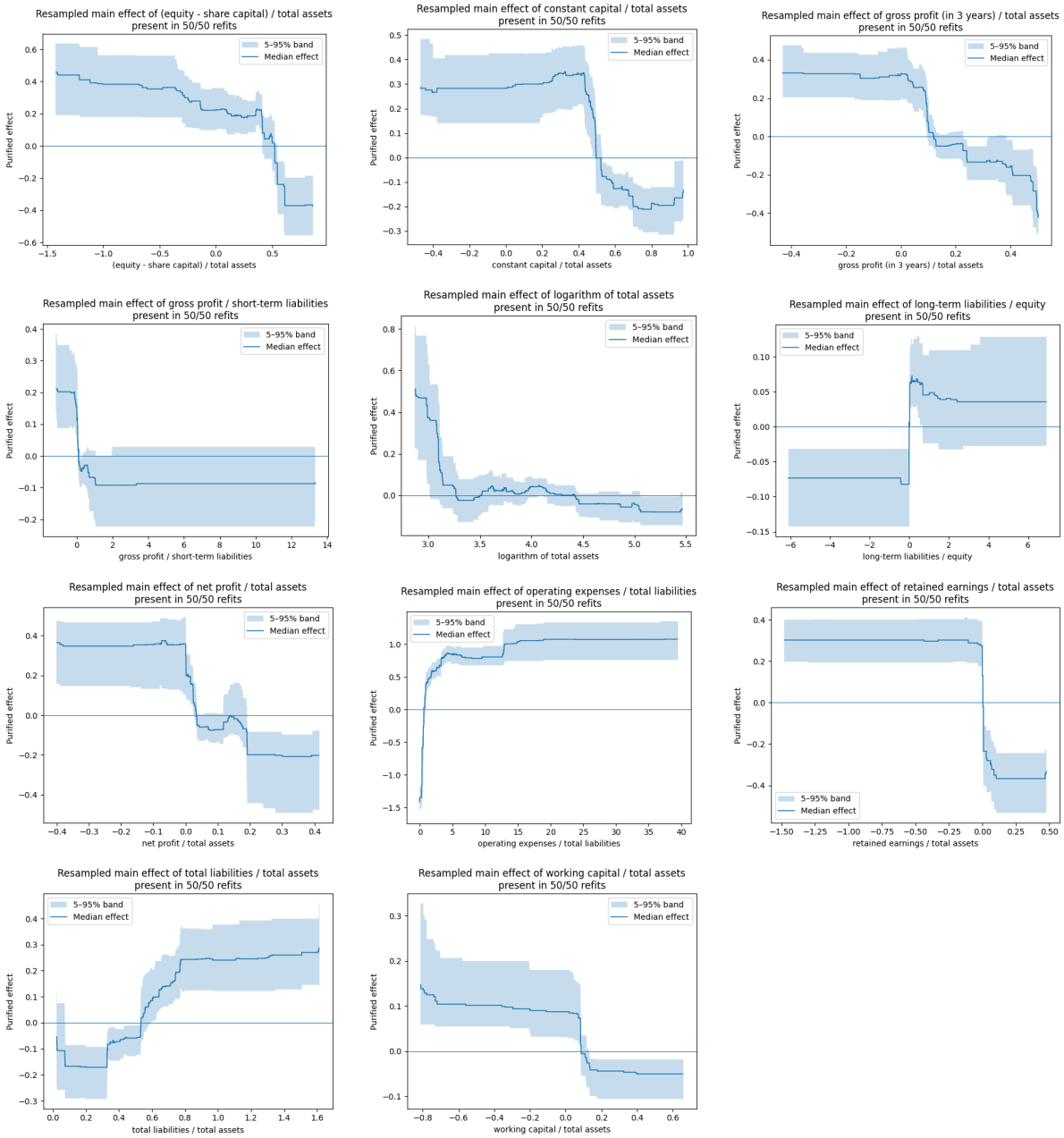


Figure 34: Resampled medians and quantile intervals for the purified main effects of the final depth-2 XGBoost model.

5 Discussion

5.1 Functional ANOVA as an interpretable model representation

We begin by considering whether the functional ANOVA representation of a depth-constrained gradient-boosted tree model may itself be regarded as inherently interpretable. As previously discussed, although the decomposition is computed after the XGBoost model has been fitted, it is not merely a post-hoc approximation but an exact reconstruction of the fitted score function. The original tree ensemble may therefore, in principle, be replaced by the functional ANOVA representation. In this sense, the interpretation is part of the model representation rather than an external explanation of it.

The central question is therefore not whether the decomposition is faithful to the fitted model, since faithfulness follows from the exact reconstruction. The relevant question is instead whether the resulting components are sufficiently simple, stable, and meaningful to support interpretation in a credit-risk setting.

This may be related to the levels of interpretability discussed in Section 2.4. At the level of the complete model, the representation was said to be interpretable only when the complete fitted model could be inspected directly. For the functional ANOVA representation, this could be said to be reasonably satisfied when the model is limited to a small collection of features, albeit not to the same extent as linear and logistic regression.

The representation more clearly supports component-level and local interpretability. Each main effect and selected pairwise interaction can be visualized directly, while individual predictions can be decomposed into baseline, main-effect, and interaction contributions. This provides both a global representation of the fitted score function and a local attribution mechanism for individual observations.

The proposed approach also aligns with several of the qualitative criteria for interpretable machine-learning models discussed by Sudjianto and Zhang (2021). It most clearly satisfies criteria related to additivity, visualizability, and projection. Additivity is built into the representation, since the fitted score function is written as a sum of lower-dimensional components. Visualizability is a major strength for main effects and selected pairwise interactions, which can be displayed directly as one- and two-dimensional functions. The functional ANOVA components are also defined through projection-type orthogonality conditions, with the empirical distribution determining how the fitted function is separated into lower-order effects.

Other criteria are only partially satisfied. Sparsity may be encouraged through feature selection, regularization, and pruning, but it is not guaranteed by the decomposition itself, even when motivated. Even when most interaction effects are small, a depth-two model may formally contain many nonzero pairwise components. If sparsity is regarded as paramount, then additional pruning or direct regularization of the functional ANOVA components, as discussed by Yang et al. (2024), may be needed.

Monotonicity is also not guaranteed by the decomposition itself. It may be encouraged through monotonicity constraints in the underlying XGBoost model, and the results suggest that such structure may often be reflected in the purified main effects.

By contrast, linearity and smoothness are generally not satisfied. Linearity is not the objective of the method: if the relevant relationships are well described by linear effects, then a linear model would generally be preferable. The benefit of the functional ANOVA representation is instead that it can expose nonlinearities and low-order interactions while

retaining a structured additive form. Smoothness is also limited, since tree-based models produce piecewise-constant component functions. Regularization and other constraints may reduce noisy local variation, but the recovered effects remain fundamentally non-smooth.

5.2 Interpretability relative to traditional credit-risk models

The next question is how the proposed representation compares with a traditional model such as logistic regression. In many respects, the two approaches are conceptually similar. Both produce additive decompositions of the prediction function, and both allow individual predictions to be expressed as sums of contributions from different risk drivers. The main difference is that logistic regression represents each effect through a single coefficient, while the functional ANOVA representation expresses effects as functions of the input variables.

In the empirical application, the constrained XGBoost model achieved stronger discriminatory performance than logistic regression, even under the depth-two restriction. However, this result should not be interpreted as a general claim that XGBoost is superior to logistic regression for credit-risk modelling. The trade-off between predictive performance and interpretability should not be understood too rigidly. Rudin (2019) argues that, in many high-stakes applications, the assumption that black-box machine-learning models are necessarily more accurate than interpretable models is misleading. According to this view, carefully designed interpretable models may often achieve competitive performance, particularly when substantial domain knowledge and feature engineering are used.

At the same time, extensive feature engineering affects what is meant by interpretability. A logistic regression model may remain mathematically simple, since each constructed variable enters through a single coefficient. However, if strong predictive performance requires many transformations, then part of the model complexity has been moved from the statistical model into the feature construction process. The fitted coefficients remain interpretable with respect to the engineered variables, but the relationship between the original risk drivers and the final prediction may be less clear.

The proposed framework thus provides a different form of interpretability. Rather than representing nonlinear structure through a large set of manually engineered variables, nonlinear main effects and pairwise corrections are learned by the model and exposed through the functional ANOVA components. These components are represented on the scale of the variables used for modelling, so if the model is fitted using economically meaningful accounting variables or financial ratios, the learned relationships can be inspected directly on that scale.

The advantage is therefore not that the proposed approach avoids modelling choices or domain-based feature construction. Such choices remain necessary in any credit-risk model. Rather, the advantage is that nonlinearities and interactions become directly inspectable as part of the fitted model representation itself, instead of being embedded indirectly in an extensively engineered feature space. In this sense, the framework may narrow the interpretability–performance trade-off in a different way than traditional feature-engineered interpretable models.

5.3 Practical usefulness and limitations of the recovered components

The empirical application should be understood primarily as an illustration of the proposed interpretability framework rather than as an attempt to construct a production-ready PD model. The objective was not to determine whether gradient-boosted tree models are generally suitable for credit-risk modelling from a purely predictive perspective, but to study whether a fitted gradient-boosted tree model can be represented in a form that supports meaningful interpretation. This distinction is important because interpretability cannot compensate for weak predictive modelling or insufficient data. The dataset used in this thesis is limited in size, contains relatively few bankruptcy observations, and consists mainly of financial-statement variables. The main question is therefore not whether the final model is optimal, but whether the decomposition provides useful insight into what the model has learned.

The recovered main effects are practically useful because they allow the learned relationship between each retained risk driver and the fitted score function to be inspected visually. This is more informative than a global importance score alone, since it shows how the contribution changes over the range of the variable. A central advantage is that the shape of this relationship does not need to be specified in advance. The model may learn nonlinearities, thresholds, saturation effects, or regime-switching behaviour directly from the data, after which these structures can be inspected through the purified main effects.

This also makes the decomposition potentially useful as a diagnostic tool during model development. Features whose recovered effects are economically implausible or unstable can be investigated, constrained, transformed, or removed. In the empirical application, this was precisely how the initial model was used: the recovered effects helped identify variables that were predictive but difficult to interpret, motivating the reduced final specification.

However, the purified components must be interpreted according to the logic of the functional ANOVA decomposition. In particular, a main effect should not be interpreted as an isolated ceteris paribus effect obtained by varying one variable while holding all others fixed. Instead, it represents a distribution-dependent contribution associated with that variable under the chosen input measure, after accounting for the hierarchical structure of the decomposition.

The interaction effects were less immediately useful than the main effects in the empirical application. Their interpretation as corrections to lower-order structure makes them more difficult to validate directly against economic intuition. Moreover, the features in the dataset mainly consist of financial ratios. Such ratios already relate one accounting quantity to another, for example by scaling liabilities, profits, or liquidity measures by firm size or by another balance-sheet quantity. As illustrated by the toy simulation, part of the interaction structure between underlying financial quantities may therefore already be embedded in the covariates themselves.

In a real SME modelling setting, one would likely also have access to other types of variables, such as behavioural information, payment history, sector indicators, or other qualitative characteristics. In such settings, stronger and more interpretable interaction effects may emerge. For example, high leverage together with a history of missed payments may jointly produce substantially greater risk than would be implied by either factor individually. Interactions involving qualitative variables may also be easier to visualize

and interpret, since the effect of a continuous risk driver can often be compared across a small number of categories. The observed weak interaction effects should therefore not be interpreted as evidence that interactions are generally unimportant in credit-risk modelling. Rather, they should be understood with regard to the features included in this particular application.

5.4 Dependence and importance measures

Another aspect that is conceptually more difficult is the interpretation of functional ANOVA components under dependence. As previously discussed, when the input variables are dependent, the main effects become entangled. This means that a lower-order component may contain structure originating from other variables whenever that structure is predictable under the input distribution.

A simple example is $F(x_1, x_2) = x_1 + x_2$. If X_1 and X_2 are perfectly correlated, then, on the support of the data,

$$F(X_1, X_2) = 2X_1 = 2X_2.$$

The main effect of either variable may therefore describe the same underlying movement in the fitted function. This is not necessarily wrong: distributionally, a one-unit increase in X_1 is associated with a one-unit increase in X_2, and hence with a two-unit increase in the function. However, it also means that the two main effects partly describe overlapping variation.

This is why interaction effects under dependence should be interpreted as corrections to the distribution-dependent lower-order structure. Even in a structurally additive function, an interaction component may be needed to avoid double-counting variation that has been absorbed by several main effects. With imperfect correlation, the same idea applies only partially: the overlap is no longer complete, but each main effect may still absorb some structure associated with the other variable.

This dependence-relative interpretation may actually be desirable in practice. If certain combinations of feature values are unlikely or essentially absent under the observed distribution, then understanding the model under those artificial combinations may be less relevant for practical model assessment. In applications, we are often interested in how the model behaves for observations that can realistically occur, rather than how it behaves under theoretical perturbations that are mathematically possible but unlikely in practice.

Thus, the entanglement of components under dependence is not necessarily a deficiency. Rather, it reflects that the decomposition is tied to the distribution on which the model is used. The recovered components should therefore be interpreted as distribution-dependent summaries of the fitted model, rather than as causal or purely structural effects. This makes them useful for assessing whether the model behaves plausibly for realistic combinations of risk drivers, even though it also means that the components cannot be interpreted as isolated economic mechanisms.

The importance summaries also reflect this issue. The SHAP values obtained from the functional ANOVA representation are attractive because they are simple to compute once the decomposition is available, while still providing both local explanations and global magnitude-based summaries. The pseudo-Shapley effects are also computationally convenient, since they allocate the variances and covariances of the recovered components directly. However, under dependence this allocation is not a proper Shapley-effect construction based on a variance value function. Its interpretation can therefore be difficult,

since positive and negative covariance contributions may cancel.

Consequently, a low pseudo-Shapley effect does not necessarily imply that a variable is unimportant. It may instead indicate that the variable contributes to the fitted score in a way that is offset by covariance with other components. This was seen both in the dependent simulation and in the real-data application. An important direction for future work is therefore to develop a proper variance-based Shapley-effect measure under dependence that retains the computational simplicity made possible by the explicit low-order representation of the fitted model.

5.5 Stability, tuning, and model quality

The interpretability of the recovered components is strongly affected by the modelling choices used to fit the underlying XGBoost model. Hyperparameters such as learning rate, regularization strength, subsampling, split penalties, and the number of trees influence not only predictive performance, but also the sparsity, smoothness, and stability of the decomposed effects. A model selected purely for discriminatory performance may therefore not be the model whose functional ANOVA representation is easiest to interpret.

This was visible in the simulation study. In the latent binary-response setting, tuning and monotonicity constraints produced smoother and more stable main effects, but also increased shrinkage, particularly for the interaction effect. The default model retained more of the interaction structure, but at the cost of greater local variability. This illustrates that effect recovery involves a trade-off between fidelity to the underlying functional structure, smoothness, and stability.

The same issue appeared in the real-data application. The final model was deliberately conservative, with a reduced feature set, depth-two trees, and relatively strong regularization. These choices improved the interpretability and stability of the recovered effects, but also reduced predictive performance relative to the initial full-feature model. This should not necessarily be viewed as a failure of the approach. Rather, it reflects the fact that interpretability-oriented modelling may require selecting a model that is not purely optimized for predictive performance.

A further practical requirement is stability. A component function may be visually interpretable in a single fitted model, but it is less useful for validation or economic interpretation if its shape changes substantially under reasonable perturbations of the data. The resampling analysis is therefore an important part of the interpretability assessment. It provides information about which effects are robust and which should be interpreted more cautiously.

In the empirical application, the dominant importance rankings and several main-effect shapes were reasonably stable across resampled datasets. However, weaker effects, local irregularities, and interaction components were less stable. This is consistent with the simulation results, where main effects were easier to recover than interactions and uncertainty was higher in sparse regions or near unstable split locations.

Stability should therefore be viewed as part of interpretability itself. Effects that are visually clear, economically plausible, and stable under resampling are more suitable for model validation than effects that appear only in a single fitted model. In applications where the decomposition is intended to support model understanding, hyperparameter selection should therefore consider both predictive performance and the behaviour of the recovered functional ANOVA components.

5.6 Regulatory implications and limitations

Finally, we consider the extent to which the proposed framework aligns with the regulatory expectations discussed earlier. Machine-learning methods are not ruled out in principle in regulatory credit-risk modelling, but their use requires that the resulting model can be understood, documented, validated, monitored, and explained. In this respect, the functional ANOVA representation is promising, since it provides explanations that are directly tied to the fitted score function rather than to a separate surrogate model.

The framework also addresses the distinction between global and local explanations in a natural way. The purified main and interaction effects provide a global representation of the learned risk-driver relationships, while individual predictions can be decomposed into baseline, main-effect, and interaction contributions. Thus, compared with a black-box model explained only through external post-hoc methods, the proposed approach provides a more integrated form of model transparency.

This is particularly relevant in a regulatory setting. Because the decomposition reconstructs the fitted score function, the explanations are not merely approximate summaries of an otherwise opaque model. Instead, the same representation can be used to inspect global risk-driver relationships, explain individual predictions, and assess whether the fitted model behaves plausibly across relevant regions of the input space. This makes the framework well suited to the type of model understanding, documentation, and validation expected in credit-risk modelling.

Several modelling and governance choices would nevertheless need to be handled carefully in practical applications. First, the decomposition is distribution-dependent. In this thesis, the empirical distribution is used as a proxy for the relevant input distribution, meaning that the recovered effects describe the model relative to the observed development sample. This is appropriate when the objective is to understand model behaviour for realistic observations, but it also means that the explanations may depend on the sample used for purification. The resampling procedure provides one way to assess this sensitivity, and in a larger application it would be natural to evaluate stability across validation samples, time periods, and portfolio segments.

Second, the hyperparameter-selection procedure would need to be justified not only in terms of predictive performance, but also in terms of interpretability. As discussed earlier, hyperparameters affect the sparsity, smoothness, and stability of the recovered functional ANOVA components. In the present framework, tuning therefore becomes part of the interpretability procedure itself. A model that maximizes discriminatory performance may not necessarily produce the most stable or meaningful component functions.

The depth restriction occupies a slightly different role. Since the interpretability framework is based on main effects and pairwise interactions, the depth-two restriction is not merely a hyperparameter to be freely optimized, but part of the intended model class. This restriction may reduce predictive flexibility, but it also limits the decomposition to components that can be visualized, validated, and communicated. Moving beyond pairwise interactions would likely be difficult to justify in many tabular credit-risk settings, since higher-order interaction effects are substantially harder to interpret and document.

The empirical application in this thesis is exploratory, but it demonstrates the type of analysis that would be useful in a more complete model-development process. In a production setting, the same framework could be combined with more extensive validation, robustness checks, missing-data treatment, monitoring over time, and calibration procedures. The contribution of the present thesis is therefore to show that a constrained XGBoost model can be represented and evaluated through functional ANOVA compo-

nents in a way that is both faithful to the fitted model and practically interpretable.

Overall, the proposed framework appears well aligned with the direction of regulatory expectations for explainable machine-learning models. It does not remove the need for careful model validation, documentation, and governance, but it provides a structured and promising way to inspect and explain a nonlinear fitted model. The remaining challenge is therefore not whether the method can provide explanations, but how the modelling and tuning procedure should be designed so that the resulting explanations are stable, interpretable, faithful to the fitted model, and generalizable beyond the development sample.

6 Conclusion and Further Work

The present thesis investigated whether a depth-constrained XGBoost model decomposed through functional ANOVA can provide an interpretable representation suitable for credit-risk modelling.

Overall, the results are promising. Gradient-boosted tree models have previously shown strong predictive performance in credit-risk applications, but their practical use in regulated settings is limited by interpretability concerns. The proposed framework addresses this by representing the fitted XGBoost score function through main effects and pairwise interaction effects. Since the functional ANOVA representation reconstructs the fitted score function, the resulting explanations are directly tied to the model itself rather than to a separate post-hoc approximation. This makes the framework well aligned with regulatory expectations that machine-learning models should support both global and local understanding.

The simulation study showed that the aggregation and purification procedure can recover theoretical functional ANOVA effects accurately when the response surface is observed directly. Recovery became more difficult in the binary-response setting, which is closer to the PD-modelling problem, but the recovered main effects still captured much of the global structure of the theoretical effects. The interaction effects were more sensitive to noise, regularization, and model constraints, indicating that they should generally be interpreted more cautiously.

The real-data application showed that the framework can produce meaningful and inspectable component functions in a credit-risk-style setting. Several recovered main effects displayed nonlinearities and threshold-like behaviour that were broadly consistent with economic intuition. The functional ANOVA representation also enabled local explanations to be obtained directly, allowing individual predictions to be analysed in terms of feature contributions. The recovered interaction effects were comparatively small and harder to interpret, which was argued to be partly related to the financial-ratio-based feature representation, where some interaction structure may already be embedded in the covariates.

Taken together, the results suggest that depth-constrained XGBoost combined with functional ANOVA decomposition provides a promising compromise between nonlinear predictive modelling and functional interpretability. The approach does not remove the need for careful model development, validation, and governance, but it provides a structured way to inspect, explain, and challenge a nonlinear fitted model in a credit-risk setting.

Several directions for future work remain.

First, the sparsity of the representation could be investigated further. In the empirical application, many pairwise interaction components were small, suggesting that the representation could potentially be simplified. One possible approach is the post-hoc regularization of the components proposed by Yang et al. (2024), where the original model predictions are regressed on the recovered component functions and weak components are shrunk or removed. Such an approach would generally sacrifice exact reconstruction of the original XGBoost model. However, if the regularized representation is treated as the final predictive model, this may be viewed as a deliberate trade-off between exact faithfulness to the original ensemble and practical interpretability.

Second, future work could study whether the recovered component functions should be smoothed. Smoothing may reduce local jumps and piecewise-constant artefacts, making

the effects easier to interpret and communicate. However, this would also weaken exact reconstruction and the exact functional ANOVA constraints. This raises an important trade-off between exact faithfulness and practical interpretability.

Finally, the framework should be evaluated on datasets that more closely resemble actual non-retail PD-modelling settings. In particular, it would be valuable to use richer datasets with more default observations, a broader set of risk drivers, and access to both financial ratios and the underlying accounting quantities from which they are constructed. This would make it possible to study how the chosen feature representation affects the recovered main and interaction effects. Additional behavioural variables, payment-history variables, sector indicators, and qualitative information could also provide a more informative setting for evaluating whether the proposed approach is useful in credit-risk modelling.

Bibliography

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

Raymond A. Anderson. *Credit Intelligence & Modelling: Many Paths Through the Forest of Credit Rating and Scoring*. Oxford University Press, Oxford, UK, 2021. URL https://academic.oup.com/book/38955.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020. doi: 10.1111/rssb.12377.

Helmi Ayari, Ramzi Guetari, and Naoufel Kraïem. Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review*, 59(1):13, November 2025. doi: 10.1007/s10462-025-11416-2. URL https://doi.org/10.1007/s10462-025-11416-2.

Basel Committee on Banking Supervision. Credit risk modelling: Current practices and applications. Technical report, Bank for International Settlements, Basel, April 1999.

William H. Beaver. Financial ratios as predictors of failure. *Journal of Accounting Research*, 4:71–111, 1966. doi: 10.2307/2490171. URL https://www.jstor.org/stable/2490171. Empirical Research in Accounting: Selected Studies.

Jodi L. Bellovary, Don E. Giacomino, and Michael D. Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, 33:1–42, 2007. URL https://www.jstor.org/stable/41948574. Winter.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213726.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984. ISBN 978-0534980538.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005. URL https://stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf. Working draft.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

European Banking Authority. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. Technical Report EBA/GL/2017/16, European Banking Authority, November 2017. URL https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1751636/2d9e7c9a-03f9-4b5e-88af-5f5a79682b02/Guidelines%20on%20PD%20and%20LGD%20estimation.pdf. In force; applicable to IRB institutions under the CRR.

European Banking Authority. Discussion paper on machine learning for IRB models. Technical Report EBA/DP/2021/04, European Banking Authority, 2021. URL https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf. Accessed: 2026-05-15.

European Central Bank. ECB guide to internal models. Supervisory guide, European Central Bank, July 2025. URL https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf. Consolidated version, July 2025.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997. doi: 10.1006/jcss.1997.1504.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–407, 2000. doi: 10.1214/aos/1016218223.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2 edition, 2009. ISBN 978-0-387-84857-0.

Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1990. ISBN 9780412343902.

Andrew Herren and P. Richard Hahn. Statistical aspects of SHAP: Functional ANOVA for model interpretation. *arXiv preprint arXiv:2208.09970*, 2022. doi: 10.48550/arXiv.2208.09970. URL https://arxiv.org/abs/2208.09970.

Wassily Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 575–580. ACM, 2004. doi: 10.1145/1014052.1014122.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Jianhua Z. Huang. Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1):242–272, 1998. URL https://www.jstor.org/stable/119985. Accessed: 2026-01-23.

Luisa Izzi, Gianluca Oricchio, and Laura Vitale. *Basel III Credit Rating Systems: An Applied Guide to Quantitative and Qualitative Models*. Palgrave Macmillan, Cham, 2019. ISBN 9783030203090.

Dohyeong Ki and Adityanand Guntuboyina. What functions does XGBoost learn? *arXiv*

preprint arXiv:2601.05444, 2026. doi: 10.48550/arXiv.2601.05444. URL https://arxiv.org/abs/2601.05444.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2402–2412. PMLR, 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969. ISBN 9780471553591.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4765–4774, 2017.

Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book/. Available online.

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

Art B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233. URL https://epubs.siam.org/doi/10.1137/130936233.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

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

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, Chichester, UK, 2008. ISBN 978-0470059975.

Robert E. Schapire. The strength of weak learnability. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, pages 28–33. IEEE, 1990.

Christian A. Scholbeck, Julia Moosbauer, Giuseppe Casalicchio, Hoshin Gupta, Bernd Bischl, and Christian Heumann. Position paper: Bridging the gap between machine learning and sensitivity analysis. *arXiv preprint arXiv:2312.13234*, 2023. doi: 10.48550/arXiv.2312.13234. URL https://arxiv.org/abs/2312.13234.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4:1060–1083, 2016. doi: 10.1137/15M1048070.

Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, 22(1):118–171, 1994. doi: 10.1214/aos/1176325361.

Agus Sudjianto. From classical ANOVA to gradient boosted trees: A functional ANOVA perspective on inherently interpretable models. Working Paper 5315091, SSRN, 2025.

Agus Sudjianto and Aijun Zhang. Designing inherently interpretable machine learning models. *arXiv preprint arXiv:2111.01743*, 2021. doi: 10.48550/arXiv.2111.01743. URL `https://arxiv.org/abs/2111.01743`.

Sebastian Tomczak. Polish companies bankruptcy data, 2016. URL `https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data`. Data donated April 10, 2016.

Zebin Yang, Agus Sudjianto, Xiaoming Li, and Aijun Zhang. Inherently interpretable tree ensemble learning. *arXiv preprint arXiv:2410.19098*, 2024. doi: 10.48550/arXiv.2410.19098. URL `https://arxiv.org/abs/2410.19098`.

Guo Yu, Jacob Bien, and Ryan Tibshirani. Reluctant interaction modeling. *arXiv preprint arXiv:1907.08414*, 2019. doi: 10.48550/arXiv.1907.08414. URL `https://arxiv.org/abs/1907.08414`.

Maciej Zięba, Sebastian K. Tomczak, and Jakub M. Tomczak. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58:93–101, 2016. doi: 10.1016/j.eswa.2016.04.001. URL `https://www.sciencedirect.com/science/article/pii/S0957417416301592`.

A Appendix

A.1 Dependent linear model variance allocation

This appendix gives the variance and covariance calculations used to interpret the pseudo-Shapley effects in the dependent linear simulation. Recall that, in the notation of Section 3.3, the canonical main effects are

$$f_1 = pU + m(U^2 - 1), \qquad f_2 = qV + m(V^2 - 1), \qquad m = \frac{k\rho}{1+\rho^2},$$

where (U, V) are standard normal variables with correlation ρ.

Since the functional ANOVA components have mean zero,

$$\mathrm{Var}\,(f_1) = \mathbb{E}[f_1^2], \qquad \mathrm{Var}\,(f_2) = \mathbb{E}[f_2^2], \qquad \mathrm{Cov}(f_1, f_2) = \mathbb{E}[f_1 f_2].$$

For the first main effect,

$$\begin{aligned}\mathbb{E}[f_1^2] &= \mathbb{E}\left[\Big(pU + m(U^2-1)\Big)^2\right] \\ &= p^2\mathbb{E}[U^2] + 2pm\mathbb{E}[U(U^2-1)] + m^2\mathbb{E}[(U^2-1)^2].\end{aligned}$$

By symmetry,

$$\mathbb{E}[U(U^2-1)] = \mathbb{E}[U^3] - \mathbb{E}[U] = 0,$$

and, using standard moments of the standard normal distribution,

$$\mathbb{E}[(U^2-1)^2] = \mathbb{E}[U^4] - 2\mathbb{E}[U^2] + 1 = 3 - 2 + 1 = 2.$$

Hence,

$$\mathrm{Var}\,(f_1) = p^2 + 2m^2.$$

Similarly,

$$\mathrm{Var}\,(f_2) = q^2 + 2m^2.$$

For the covariance between the main effects,

$$\begin{aligned}\mathbb{E}[f_1 f_2] &= \mathbb{E}\left[\Big(pU + m(U^2-1)\Big)\Big(qV + m(V^2-1)\Big)\right] \\ &= pq\mathbb{E}[UV] + pm\mathbb{E}[U(V^2-1)] + qm\mathbb{E}[V(U^2-1)] + m^2\mathbb{E}[(U^2-1)(V^2-1)].\end{aligned}$$

The mixed terms vanish by symmetry:

$$\mathbb{E}[U(V^2-1)] = 0, \qquad \mathbb{E}[V(U^2-1)] = 0.$$

Furthermore,

$$\mathbb{E}[UV] = \rho,$$

and, by Isserlis' theorem,

$$\mathbb{E}[U^2V^2] = 1 + 2\rho^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}[(U^2-1)(V^2-1)] &= \mathbb{E}[U^2V^2] - \mathbb{E}[U^2] - \mathbb{E}[V^2] + 1 \\ &= (1 + 2\rho^2) - 1 - 1 + 1 \\ &= 2\rho^2.\end{aligned}$$

Thus,

$$\mathrm{Cov}(f_1, f_2) = pq\rho + 2m^2\rho^2.$$

For completeness, we also derive the variance of the interaction component. Let

$$A = UV - \rho, \qquad B = U^2 - 1, \qquad C = V^2 - 1,$$

so that

$$f_{12} = kA - mB - mC.$$

Then

$$\begin{aligned}\mathrm{Var}\,(f_{12}) &= \mathbb{E}[(kA - mB - mC)^2] \\ &= k^2\mathbb{E}[A^2] + m^2\mathbb{E}[B^2] + m^2\mathbb{E}[C^2] - 2km\mathbb{E}[AB] - 2km\mathbb{E}[AC] + 2m^2\mathbb{E}[BC].\end{aligned}$$

Using standard Gaussian moments and Isserlis' theorem,

$$\mathbb{E}[A^2] = 1 + \rho^2, \qquad \mathbb{E}[B^2] = \mathbb{E}[C^2] = 2,$$
$$\mathbb{E}[BC] = 2\rho^2, \qquad \mathbb{E}[AB] = \mathbb{E}[AC] = 2\rho.$$

Substituting these expressions gives

$$\mathrm{Var}\,(f_{12}) = k^2(1 + \rho^2) + 4m^2 + 4m^2\rho^2 - 8km\rho.$$

Finally, substituting

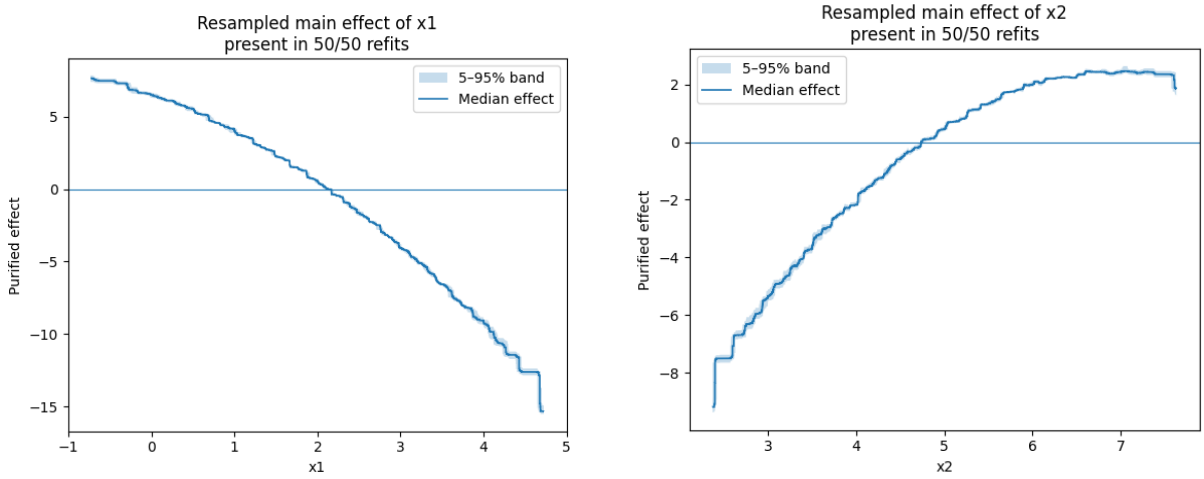
$$m = \frac{k\rho}{1 + \rho^2}$$

and simplifying yields

$$\mathrm{Var}\,(f_{12}) = k^2\frac{(1 - \rho^2)^2}{1 + \rho^2}.$$

Thus, as $|\rho|$ increases, the variance of the interaction component decreases. This reflects that an increasing proportion of the original interaction structure is absorbed into the lower-order effects in order to satisfy hierarchical orthogonality.

A.2 Dependent linear model resampling figures



(a) X_1 (b) X_2

Figure 35: Resampled uncertainty bands for the purified main effects in the dependent linear model.

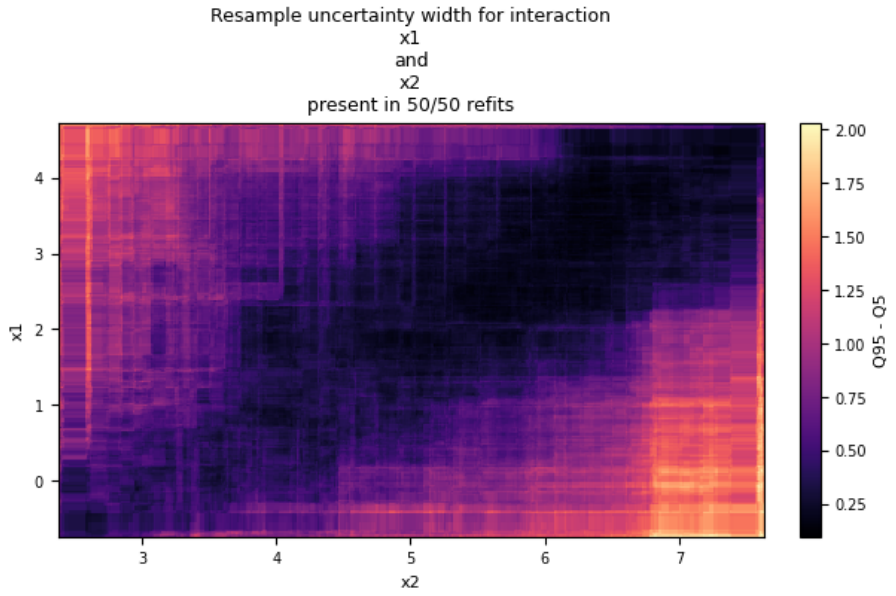


Figure 36: Resampled uncertainty bands for the purified interaction effect in the dependent linear model.

A.3 Credit-risk toy model interaction figures

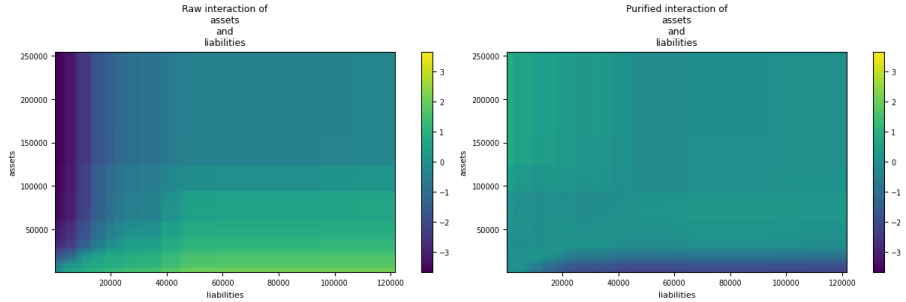


Figure 37: Recovered purified interaction effect between assets and liabilities in the component-based toy model.

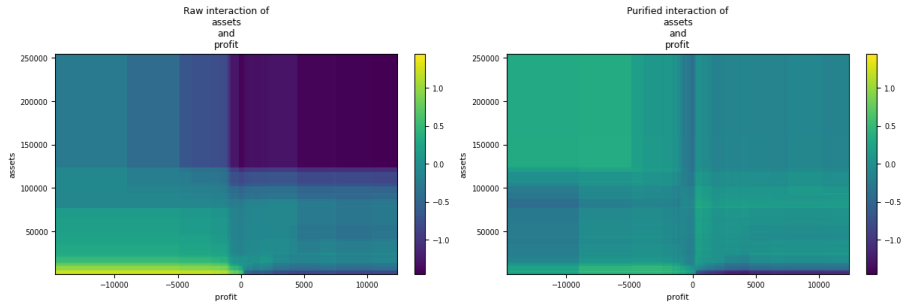


Figure 38: Recovered purified interaction effect between assets and profit in the component-based toy model.

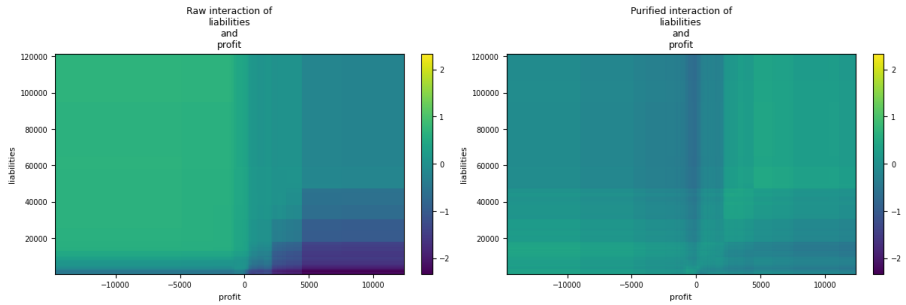


Figure 39: Recovered purified interaction effect between liabilities and profit in the component-based toy model.

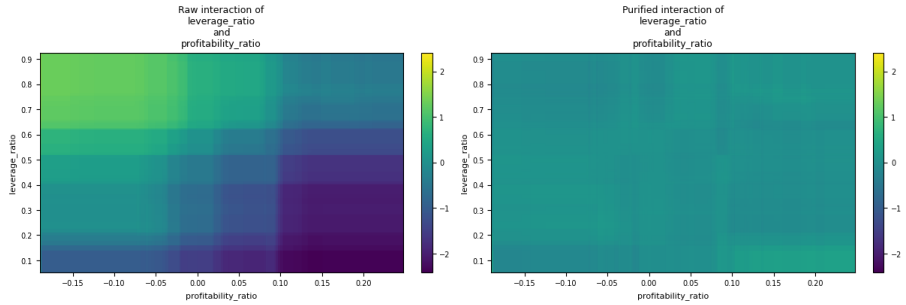


Figure 40: Recovered purified interaction effect in the ratio-based toy model.

A.4 Selected main effects with empirical feature distributions

This section presents two selected main effects together with the empirical distribution of the corresponding feature values in the training data. The x-axis is restricted to the range over which the estimated main effect changes. Observations outside this range fall in constant-effect regions and are reported separately in each figure.

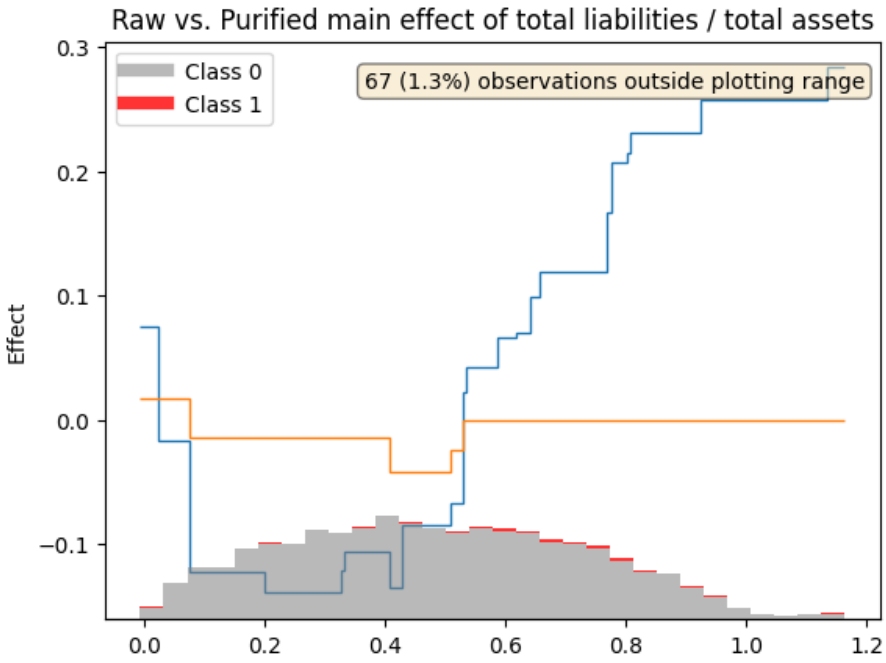


Figure 41: Raw and purified main effect for total liabilities divided by total assets, shown together with the empirical class distribution.

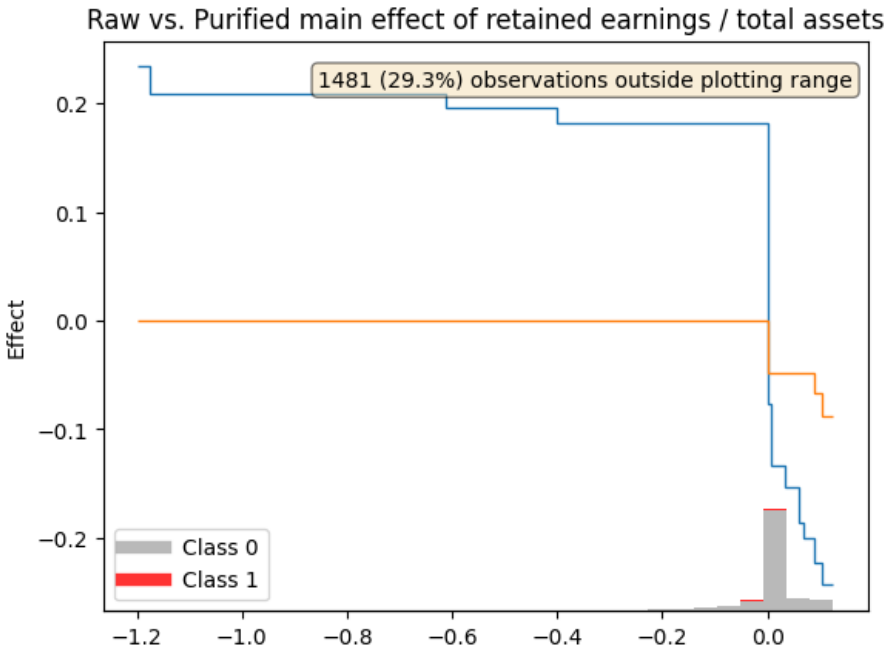


Figure 42: Raw and purified main effect for retained earnings divided by total assets, shown together with the empirical class distribution.

We observe that for the first main effect, total liabilities divided by total assets, nearly all observations lie within the range where the estimated main effect changes. The distribution of bankrupt firms also aligns with the structure of the estimated effect, as bankruptcies become more common toward larger values of the ratio. For the second main effect, retained earnings divided by total assets, the distribution is more concen-

trated around zero, while a larger share of observations lies outside the plotted range. Although most of these observations are located to the right, most of the corresponding bankruptcies are located to the left, which helps explain why the model continues to split in that region.