Are Logistic Models Really Interpretable?

Danial Dervovic[1] , **Freddy Lécué**[2] , **Nicolás Marchesotti**[3] and **Daniele Magazzeni**[3]
[1]JP Morgan AI Research, Edinburgh, UK
[2]JP Morgan AI Research, New York City, NY, USA
[3]JP Morgan AI Research, London, UK
{danial.dervovic, freddy.lecue, nicolas.p.marchesotti, daniele.magazzeni}@jpmchase.com

Abstract

The demand for open and trustworthy AI models points towards widespread publishing of model weights. Consumers of these model weights must be able to act accordingly with the information provided. That said, one of the simplest AI classification models, Logistic Regression (LR), has an unwieldy interpretation of its model weights, with greater difficulties when extending LR to generalised additive models. In this work, we show via a User Study that skilled participants are unable to reliably reproduce the action of small LR models given the trained parameters. As an antidote to this, we define Linearised Additive Models (LAMs), an optimal piecewise linear approximation that augments any trained additive model equipped with a sigmoid link function, requiring no retraining. We argue that LAMs are more interpretable than logistic models – survey participants are shown to solve model reasoning tasks with LAMs much more accurately than with LR given the same information. Furthermore, we show that LAMs do not suffer from large performance penalties in terms of ROC-AUC and calibration with respect to their logistic counterparts on a broad suite of public financial modelling data.

1 Introduction

In high-stakes domains such as finance and healthcare, there is renewed interest in *inherently interpretable* models [Lipton, 2018; Molnar, 2022; Gunning, 2019], where the model form is such that it admits useful explanations of its output without any post-processing. Calls for transparency of algorithms being used on the public are widespread [Veale *et al.*, 2018]. Within finance there is already increased regulatory scrutiny [OCC, 2021; Commission, 2021; Register, 2021] being introduced with regards to the usage of AI, which could extend in the future – within certain contexts – to require full algorithmic transparency, i.e. sharing model coefficients.

One of the prototypical inherently interpretable classification models often used as a baseline is Logistic Regression (LR) [Molnar, 2022]. Other additive models such as Generalised Additive Models (GAMs) and variants thereof [Lou *et al.*, 2013; Gkolemis *et al.*, 2023] generalise LR to have more flexibility [Hastie *et al.*, 2009] and are also considered to be interpretable. As with LR, such models entail evaluating a real-valued function of the input data in logit, or log-odds space, that is subsequently transformed into probability space via a non-linear logistic link function. We call this class of models *logistic models* – a rigorous definition is given in Definition 2.1. In some sense, logistic models can be thought of as *reasoning in logit space*, in that the model weights naturally find their interpretation in terms of log-odds rather than probabilities, the units of the eventual model output.

Logistic models are now ubiquitous within Explainable AI (XAI) [Lou *et al.*, 2013; Sudjianto and Zhang, 2021; Vaughan *et al.*, 2018] but the literature is scant on evaluating the interpretability of these models. Indeed, there is a small amount of evidence to the contrary [Harris, 2017; von Hippel, 2015], with no more thorough study to the authors' knowledge. To what extent is this family of models truly interpretable? The present work aims to (at least partially) answer this question. We show via a User Study that for at least one definition of interpretability, based on Human-Grounded Evaluation [Doshi-Velez and Kim, 2017], such models provide limited and misleading explanations. For contexts where a certain level of interpretability is required we propose a remedy, *Linearised Additive Models (LAM)*, that largely keeps the properties of any base logistic model the same, while dispensing with the non-linearities that cause confusion when using model weights as explanations.

Contributions. We outline the primary contributions below.

1. **Identification of interpretability limitations for LR.** A concrete motivating example demonstrating that model explanations provided in log-odds can be difficult for humans to interpret.
2. **Linearised Additive Models (LAM).** An efficient procedure to convert any trained logistic additive model that reasons in log odds to one that reasons directly about probabilities, without any retraining. For the special case of LR, LAM is rigorously proved to be the optimal approximation out of a large class of possible models.
3. **Empirical evaluation of performance preservation**. On a collection of public datasets from credit modelling, we establish that there is only a very small penalty in classification performance and a somewhat larger – but still small – penalty in calibration incurred for using

LAMs versus logistic models.

4. **User evaluation.** We conduct a user study with $N = 36$ participants, concluding via Human-Grounded Evaluation that LAMs are more interpretable than logistic models, as suggested by the motivating example. The measured outcomes of the user study are statistically significant.

Related Work. There is a vast literature on defining and evaluating performance of inherently interpretable models in general, a non-exhaustive group of which is [Dash *et al.*, 2018; Lou *et al.*, 2013; Yang *et al.*, 2021; Vaughan *et al.*, 2018; Kraus and Feuerriegel, 2019; De Bock and De Caigny, 2021; Vidal and Schiffer, 2020; Alaa and van der Schaar, 2019]. The evaluation of interpretability itself is still an open question, with detailed discussion of this issue in the references [Halliwell *et al.*, 2022; Chen *et al.*, 2022b; Narayanan *et al.*, 2018; Lipton, 2018]. There are several different approaches, with the current preferred (and most expensive) approach being User Studies, as this allows one to directly measure resulting outcomes from explanations [Doshi-Velez and Kim, 2017]. Indeed, measuring explanation quality via measured outcomes when humans are asked to simulate the action of an algorithm predates most of the AI interpretability literature, for instance works such as [Kulesza *et al.*, 2013]. [Rong *et al.*, 2022] provide a recent survey paper on User Studies for evaluating explanations and interpretability of AI models.

The closest works in the literature to this paper are [Abdul *et al.*, 2020; Poursabzi-Sangdeh *et al.*, 2021]. In [Abdul *et al.*, 2020], the authors measure interpretability of sparse linear models and GAMs via user studies that measure cognitive load on participants carrying out tasks with these models and their associated explanations. In the work by [Poursabzi-Sangdeh *et al.*, 2021], users are presented with the coefficients of 2 and 8-variable linear models. The quality of the explanations from each model is measured by how adept users are at simulating the action of the model. Crucially, both these works evaluate regression models and are not concerned with issues arising due to the non-linearity of the sigmoid transformation required for logistic classification models.

The present work highlights that experts are not immune from misinterpreting certain model explanations, as also observed in the works: [Kaur *et al.*, 2020; Bhatt *et al.*, 2020].

Low-degree polynomial approximations to the sigmoid are employed in Private Machine Learning, e.g. [Kim *et al.*, 2018a; Chen *et al.*, 2018; Kim *et al.*, 2018b], but to our knowledge the piecewise linear approximation in this work is unique.

Structure. In Section 2 we provide the motivating example and present the LAM definition and optimality results. Section 3 contains the performance evaluation of LAMs against their logistic counterparts and Section 4 details the User Study. We include most detail in the main text for the User Study, providing derivations, proofs and experimental details in the Supplementary Material (SM). We conclude with limitations and future work in Section 5.

Notation. This work considers binary classification, and we use $\{(\boldsymbol{x}^{(j)}, y^{(j)})\}_{j=1}^{M}$ to denote the data, where $\boldsymbol{x} \in \mathbb{R}^d$ is a vector of numerical features in a given dataset. The binary labels are indicators of some (usually bad) event such as a loan default: $y_i \in \{0, 1\}$. We assume data points are drawn i.i.d. For a point $\boldsymbol{x} \in \mathbb{R}^d$, we denote the i^{th} element of $\boldsymbol{x}$ by x_i. The i^{th} Euclidean basis vector is denoted by $\mathbf{e}_i$. The sigmoid function is defined as $\sigma(z) := (1 + e^{-z})^{-1}$ for $z \in \mathbb{R}$. We follow credit modelling terminology, where *risk* $\hat{y}(\boldsymbol{x}^{(j)})$ is a model's subjective probability in $[0, 1]$ for a data point $\boldsymbol{x}^{(j)}$ to be of positive class, that is $y^{(j)} = 1$. The set $[d] := \{1, \ldots, d\}$ for $d \in \mathbb{N}$.

2 Logistic and Linear Probability Modelling

GAMs [Hastie *et al.*, 2009; Molnar, 2022] are a widely-known and long standing class of models considered to be inherently interpretable. In this work we restrict attention to GAMs without feature interactions. We formalise the notion of additive models as understood in this paper in Definition 2.1.

Definition 2.1 (Logistic Additive Model)**.** Let $\boldsymbol{x} \in \mathbb{R}^d$. We call $\hat{y} : \mathbb{R}^d \to [0, 1]$ a *logistic additive model* if takes the form $\hat{y}(\boldsymbol{x}) = \sigma(f(\boldsymbol{x})) := \sigma(\beta_0 + \sum_{i=1}^{d} \beta_i f_i(x_i))$, where the *bias* $\beta_0 \in \mathbb{R}$ and for all $i \in [d]$, $\beta_i \in \mathbb{R}$ and the $f_i : \mathbb{R} \to \mathbb{R}$ are univariate shape functions.

We refer to logistic additive models as defined in Definition 2.1 simply as logistic models or additive models when it is clear from context. The simplest and most common additive model in wide usage is LR, where $f_i(x_i) = x_i$ for all $i \in [d]$ [Hastie *et al.*, 2009]. Another example would be an Explainable Boosting Machine of [Lou *et al.*, 2013] for classification, where the f_i are piecewise constant functions (when there are no feature interactions).

2.1 Logistic Modelling and Interpretability

Historically, linear probability modelling, i.e. linear regression on dichotomous variables, was used prior to the advent of efficient methods for fitting LR models; see [Aldrich and Nelson, 1984; Hastie *et al.*, 2009]. It is generally accepted that the application of linear regression to binary classification problems is unwise due to the propensity of the model returning probability estimates outside the $[0, 1]$ interval and sensitivity to outliers [Ng, 2011; Hastie *et al.*, 2009; Molnar, 2022]. In certain circumstances, these issues are not observed, with linear regression obtaining similar classification performance to LR [Hellevik, 2009; von Hippel, 2015]. LR models have superceded linear probability models.

The LR model coefficients are typically interpreted as follows [Molnar, 2022; Hastie *et al.*, 2009]: a unit change in variable x_i leads to a multiplicative increase in odds for the positive class of $\exp(\beta_i)$. However, as observed by [Harris, 2017; von Hippel, 2015] this interpretation can be unwieldy for experts and nigh-on impossible for non-experts to reason with when we are concerned with probabilities, which is how the model outputs are typically presented and thought about.

Motivating Example. Suppose there is a LR model $\hat{y}$ used to predict some negative outcome and coefficients are shared with downstream users of the model. Inputs with risk ≥ 0.5 are considered "high-risk" and "low-risk" otherwise. Referring to Figure 1, Alice has a predicted risk of $\hat{y}(\boldsymbol{x}^{(A)}) = 0.1$ and Bob has a predicted risk of $\hat{y}(\boldsymbol{x}^{(B)}) = 0.25$. Both Alice and Bob are interested in what happens to their risk if they increase

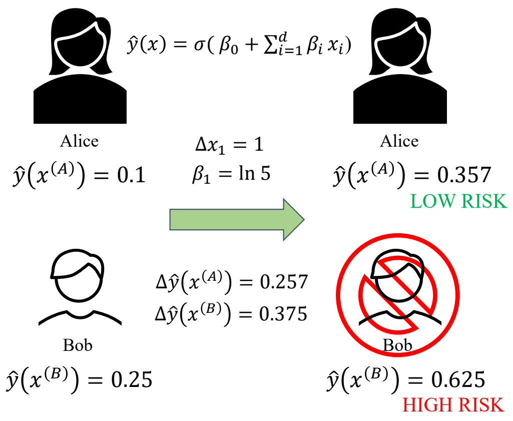
$\hat{y}(x) = \sigma(\beta_0 + \sum_{i=1}^{d} \beta_i x_i)$
Alice
$\hat{y}(x^{(A)}) = 0.1$
$\Delta x_1 = 1$
$\beta_1 = \ln 5$
Alice
$\hat{y}(x^{(A)}) = 0.357$
LOW RISK
Bob
$\Delta\hat{y}(x^{(A)}) = 0.257$
$\Delta\hat{y}(x^{(B)}) = 0.375$
Bob
$\hat{y}(x^{(B)}) = 0.25$
$\hat{y}(x^{(B)}) = 0.625$
HIGH RISK

Figure 1: The cost of misinterpration of model coefficients as explanations. Alice and Bob receive the same explanation, but incur a different change in model output, ultimately leading to different outcomes.

the value of feature x_i by one unit, all else equal. They are told the model coefficient $\beta_i = 1.61 \approx \ln 5$, a common form of transparent model explanation. First, both exponentiate β_i which gives 5. They now know that increasing feature x_i by increases their odds by a factor 5. The model outputs a risk score in units of probability, so they now have to compute what increasing their odds corresponds to in probabilities.

In this instance for Alice, $\text{odds}(A) = \hat{y}(\boldsymbol{x}^{(A)})/(1 - \hat{y}(\boldsymbol{x}^{(A)})) = 0.1/(1-0.1) \approx 0.111$. Alice then multiplies her odds by 5, yielding 0.556. Converting back to probabilities we have $\hat{y}(\boldsymbol{x}^{(A)} + \mathbf{e}_i) \approx 0.556/(1+0.556) \approx 0.357$. Subtracting her original risk score, we have that increasing x_i by one unit increases the risk by a probability of 0.257 and Alice remains low-risk. A similarly laborious computation gives an increase in risk for Bob to approximately 0.625. Bob would be considered high-risk under a unit increase in x_i. This was not obvious on first inspection before carrying out the computation explicitly.

The nonlinearity of odds as a function of probabilities (and vice versa) means that *users with different risk scores cannot attribute logistic regression model outputs to the model coefficients in the same way.* Moreover, the necessary computations are such that one cannot easily reason about the model's input-output relationship without a significant amount of practice. On the contrary, the coefficients β_i of a linear probability model admit the more direct interpretation of the increase in output model probability arising from a unit increase in x_i, regardless of the risk value of the user in question.

2.2 Linearised Additive Models (LAMs)

Given the preceding example, we wish to keep the interpretability characteristics of linear probability modelling while simultaneously sidestepping the issues arising from using a non-linear link function as in LR. To this end, we present the Linearised Additive Model (LAM), which is defined with respect to an already trained logistic model, $\sigma \circ f$. Informally, a LAM replaces the sigmoid link function with a clipping function and scales f by an affine transformation. Denote by $\Pi_{[0,1]}$ the projector from $\mathbb{R}$ onto the unit interval, that is $\Pi_{[0,1]}(z) = \max(0, \min(1, z))$.

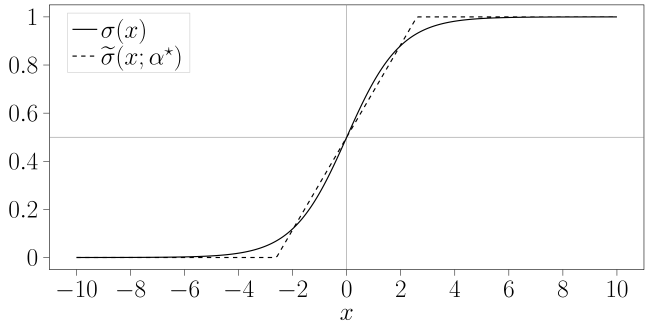

Figure 2: Optimal approximation $\widetilde{\sigma}(x;\alpha \approx 2.5996)$ to sigmoid function $\sigma(x)$. The parameter α corresponds the half the width along the x-axis of the middle line segment in the piecewise linear function.

Definition 2.2 (Linearised Additive Models (LAM))**.** Let $\boldsymbol{x} \in \mathbb{R}^d$ and let $\hat{y}(\boldsymbol{x}) = \sigma(f(\boldsymbol{x}))$ be an additive model per Definition 2.1. Moreover, set $\alpha^\star := \frac{80000}{30773} \approx 2.5996$ as a universal constant. Then, the *Linearised Additive Model*, $\hat{y}_{\text{LAM}}$, relative to f is given by

$$\hat{y}_{\text{LAM}}(\boldsymbol{x}) = \Pi_{[0,1]}\left(\frac{1}{2} + \frac{\beta_0}{2\alpha^\star} + \sum_{i=1}^{d} \frac{\beta_i}{2\alpha^\star} f_i(x_i)\right).$$

For brevity we refer to $\hat{y}_{\text{LAM}}$ as a *linearised* model and say that $\hat{y}_{\text{LAM}}$ is the *LAM induced by* $\hat{y}$.

The LAM as defined in Definition 2.2 is derived by considering the optimal 3-piece piecewise linear approximation (see Figure 2) to the sigmoid function in terms of squared error, using this function as a link function for $f(\boldsymbol{x})$ and invoking linearity. We choose this family of approximating functions as 3 pieces gives the simplest non-trivial approximation to the logistic sigmoid, while simulataneously allowing for a similar interpretation to a linear probability model. Squared error is chosen as it is a common metric for function approximation [Hastie *et al.*, 2009] for which we are able to derive a universal tractable approximator.

As an example, an LR model is written as $\hat{y}(\boldsymbol{x}) = \sigma(\sum_{i=0}^{d} \beta_i x_i)$, where we fix $x_0 = 1$ according to convention. Then, the linearised version will be $\hat{y}_{\text{LAM}}(\boldsymbol{x}) = \Pi_{[0,1]}(\frac{1}{2} + \sum_{i=0}^{d} \frac{\beta_i}{2\alpha^\star} x_i)$. In the case of linearised LR we can interpret the coefficients $\beta_i/2\alpha^\star$ as the contribution to model output in probability space for a unit increase in x_i, as with a linear probability model.

Remark 2.3. *To train a LAM, all that is required is to train the underlying logistic additive model, then apply Definition 2.2 using the trained coefficients* $\{\beta_i\}_{i=0}^{d}$.

Motivating Example Revisited. Under an LR model, Alice and Bob had to interpret a unit increase in feature x_i as having different and unintuitive effects on their respective risk scores. Using the induced LAM, a unit increase in x_i gives rise to *the same change in model output, regardless of the input*, i.e. $\frac{\beta_i}{2\alpha^\star} = \frac{1.61}{2 \times 2.5996} \approx 0.310$, modulo outputs greater than unity which will be clipped. Alice and Bob's respective inputs of 0.1 and 0.25 can be readily seen to change to 0.41 and 0.56. Note this gives the same qualitative result as the non-linearised model, namely Alice stays low-risk and Bob becomes high-risk. We surmise that this direct interpretation of the LAM coefficients incurs smaller cognitive overhead in answering questions of this type, in constrast to LR.

The following optimality result for the LAM approximation to LR lends theoretical support to our definition of LAMs.

Theorem 2.4 (LAM Optimality)**.** *Let* PL_3 *be the space of 3-piece piecewise linear functions of one variable and* $\mathcal{X} = \mathbb{R}^d$*. For any LR model* $\hat{y}(\boldsymbol{x}) = \sigma(f(\boldsymbol{x})) = \sigma(\beta_0 + \sum_{i=1}^d \beta_i x_i)$ *on* $\mathcal{X}$*, an approximator* $\widetilde{\sigma}(f(\boldsymbol{x}))$ *is defined for all* $\widetilde{\sigma} \in \mathrm{PL}_3$*. Then,* $\hat{y}_{\mathrm{LAM}}$ *is the squared-error optimal approximator for arbitrary* f*, that is,*

$$\hat{y}_{\mathrm{LAM}}(\boldsymbol{x}) = \widetilde{\sigma}(f(\boldsymbol{x}); \alpha^\star) \quad \textit{where}$$
$$\widetilde{\sigma}(\,\cdot\,; \alpha^\star) = \arg\min_{\widetilde{\sigma} \in \mathrm{PL}_3} \left\{ \int_{\mathcal{X}} (\widetilde{\sigma}(f(\boldsymbol{x})) - \sigma(f(\boldsymbol{x})))^2 \mathrm{d}\boldsymbol{x} \right\},$$

with $\widetilde{\sigma}(z; \alpha^\star) := \Pi_{[0,1]}(\frac{1}{2}(1 + \frac{z}{\alpha^\star}))$*,* $\alpha^\star \approx 2.5996$*.*

Proof (Sketch). One can show via symmetry arguments that the minimising approximator comes from a one-parameter function family, when $d = 1$ with $f(x) = x$. The error is a convex function of this parameter α, which is minimised at $\alpha^\star$. The argument can then be extended to d-dimensional, affine $f(\boldsymbol{x})$ by considering the error integral explicitly, from which the result follows. □

3 Performance Comparison

In this section we detail our experiments comparing the model performance of logistic additive models against their linearised (LAM) counterparts.

3.1 Experimental Setup

Models. We compare all models to XGBoost [Chen *et al.*, 2022a], with shorthand XGB, and XGB with monotone constraints imposed (MonoXGB). XGB classification performance serves as an effective upper-bound on the competing models. As state-of-the-art baseline logistic additive models we consider the Additive Risk Models (ARMs) of [Chen *et al.*, 2022a], since they are GAMs that explicitly incorporate monotone constraints that are often required in sensitive domains such as finance [Reserve, 2011]. These models come in 1-layer (ARM1) and 2-layer (ARM2) variants. Further baselines include NNLR (Non-negative LR [Chen *et al.*, 2022a]) with raw feature inputs as an effective lower bound on model performance. For an additive model with shorthand M, we denote its linearised version (in the sense of Definition 2.2) by LAM-M. We include the linearised models LAM-NNLR, LAM-ARM1 and LAM-ARM2 in our experiments. LAM-ARM2 has both the individual subscale models and global NNLR model linearised.

Datasets. In this work we are principally interested in the consumer credit domain. Bankruptcy prediction datasets are also included due to their similar problem structure and origin. We consider publically available datasets from the UCI repository [Kelly *et al.*], namely, the German Credit dataset [Hofmann, 1994], Australia credit approvals [Quinlan, 1987], Taiwanese bankruptcy [Liang *et al.*, 2020] prediction, Japanese credit screening [Sano, 1992] and the Polish companies bankruptcy [Tomczak, 2016] dataset. We consider also the FICO Home Equity Line of Credit dataset (HELOC) [FICO, 2018], Give Me Some Credit (GMSC) and Lending Club (LC) [Kaggle, 2019] datasets.

3.2 Performance Metrics

We are chiefly interested to what extent linearising relative to logistic models introduces degredation (if any) of both classification performance and calibration – the latter being of interest as we are modifying the probability estimates of a trained logistic model. For each metric the 10-fold stratified cross-validation score is computed for every (classifier, dataset) combination.

Classification Performance. To measure of classification performance, we use the area under the curve of the receiver operating characteristic [Bradley, 1997; Hanley and McNeil, 1983], denoted as AUC.

Calibration. We consider two widely-used numerical summary statistics for the calibration, *Expected Calibration Error (ECE)* and *Maximum Calibration Error, (MCE)*. Lower values of ECE and MCE correspond to better calibration of a particular model, with the idealised model having a value of zero for both.

Statistical Methodology. For the purposes of discussion here, consider a graph where for each classification algorithm $\mathcal{A}$ we draw a node. We draw an edge between any nodes corresponding to algorithm pairs $(\mathcal{A}, \mathcal{A}')$ such that the performance of $\mathcal{A}$ cannot be distinguished from the performance $\mathcal{A}'$ with significance $\alpha = 0.05$ according to a a Wilcoxon signed-rank test [Wilcoxon, 1945] conducted over the considered datasets. We display this graph, where the nodes $\mathcal{A}$ are arranged according to their average rank $R_{\mathcal{A}}$ in Figure 3. These are the Critical Difference (CD) diagrams of [Demšar, 2006] for AUC, ECE and MCE . Not only are we interested in whether an observed difference in cross validated score between two algorithms is statistically significant, but also the size of this difference. In Table 1 for the AUC metric. The quantity $\hat{\theta}_{\mathrm{HL}}$ is a robust point estimate[1] computed across the datasets.

Results for Classification Performance. We highlight several observations from the CD diagram for AUC (Figure 3). Unconstrained XGB is the strongest model in terms of AUC, as we may expect. The weakest performing model families are {NNLR, LAM-NNLR}, presumably due to their simplicity as compared with the other models. LAM-NNLR models are indistinguishable in AUC performance from their logistic counterparts, NNLR. Interestingly, LAM-ARM1 is connected to MonoXGB in the CD graph. MonoXGB models are considered state of the art for monotone constrained models on tabular data and LAM-ARM1 is more interpretable according to a number of criteria, yet here they display statistically indistinguishable classification performance. However, LAM-ARM1 can be distiguished with statistical significance from

[1] The Hodges-Lehmann estimator associated to Wilcoxon's signed rank test [Wilcox, 2022].

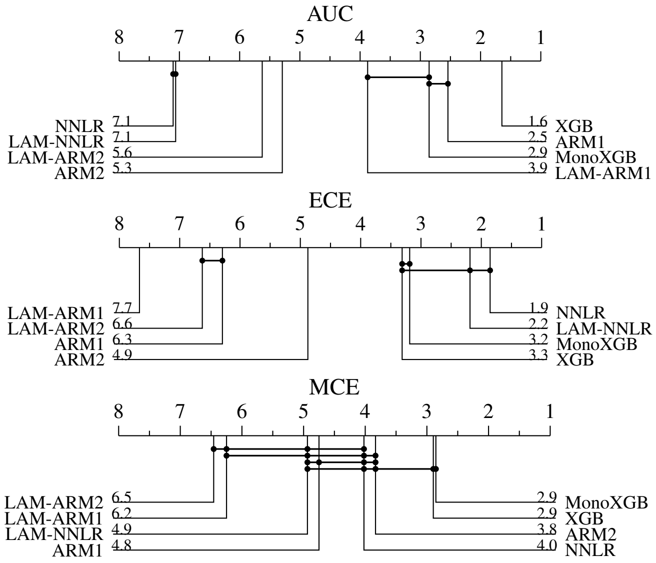

Figure 3: Critical Difference diagrams for AUC, ECE and MCE. The x-axis represents the mean rank averaged over all datasets, with each classifier's mean rank reported adjacent to its name (lower rank $\equiv$ better). Classifiers connected by an edge *cannot* be distinguished with significance $\alpha = 0.05$.

	NNLR	*LAM-NNLR*	*ARM1*	*ARM2*	*LAM-ARM1*	*LAM-ARM2*	*MonoXGB*	*XGB*
NNLR	—	0.000	**-0.327**	**-0.314**	**-0.319**	**-0.311**	**-0.320**	**-0.340**
LAM-NNLR	—	—	**-0.327**	**-0.314**	**-0.319**	**-0.311**	**-0.320**	**-0.340**
ARM1	—	—	—	**0.008**	**0.003**	**0.013**	0.000	**-0.011**
ARM2	—	—	—	—	**-0.005**	**0.003**	**-0.009**	**-0.017**
LAM-ARM1	—	—	—	—	—	**0.009**	-0.003	**-0.013**
LAM-ARM2	—	—	—	—	—	—	**-0.014**	**-0.020**
MonoXGB	—	—	—	—	—	—	—	**-0.002**
XGB	—	—	—	—	—	—	—	—

Table 1: Point estimate for difference in AUC scores between classifiers. Negative values mean column model is better than row model. Bold values indicate statistical significance.

its logistic counterpart, ARM1. Nonetheless they are in the same connected component of the CD graph indicating a very similar level of proficiency. The AUC performance of LAM-ARM2 can be separated from the performance of ARM2. In this instance we hypothesise that this is happening due to linearisation being applied in two layers as opposed to just one, allowing errors to accumulate. *In terms of absolute numerical difference in AUC performance, linearisation incurs a very small penalty*, as shown in Table 1. The AUC penalties (point estimate) for linearising NNLR, ARM1 and ARM2 are 0.000, 0.003 and 0.003 respectively across the datasets.

Results for Calibration. Inspecting the CD diagrams for the ECE and MCE calibration metrics in Figure 3, we see there is a penalty incurred on model calibration for linearising. Indeed for both metrics, only the linearisation of NNLR models gives a penalty in calibration that cannot be distinguished from the logistic model with statistical significance. However, we note the numerical difference across the datasets is small,being of order ∼0.005 for ECE and ∼0.03 for MCE. We believe this discrepancy in calibration is likely due to "model certainty" evinced by the linearised models, namely output values lying in $\{0, 1\}$. Notably, the LC datasets get a very large fraction of predictions with certainty (∼80%) using LAM-ARM1 and LAM-ARM2 models. The MCE ranks are all fairly close to one another, meaning all of the classifiers are more closely matched on this metric as compared with AUC and ECE.

4 User Survey

The motivating example in Section 2.1 suggests that linearised models will be easier to interpret and reason about by users as opposed to logistic models. To substantiate this, we conduct a user study where the aim is to ascertain which class of models is more interpretable: logistic models or LAMs. Our proxy for interpretability is how capable users are at carrying out basic reasoning tasks about the models' outputs, given the model coefficients. This falls within the paradigm of Human-Grounded Evaluation [Doshi-Velez and Kim, 2017], where empirically measured human performance on a simplified task serves as a proxy for explanation quality. Using human simulation of model outputs as a proxy for model interpretability is a similar strategy to that used in the work by [Poursabzi-Sangdeh *et al.*, 2021].

4.1 Setup

The study is structured as a questionnaire, wherein participants are shown a small LR model alongside its linearised counterpart and are asked to predict how the output of each model will change in both direction and magnitude from some initial (input, output) pairs. Participants are shown several instantiations, which we call scenarios. More formally, a *scenario* consists of the following elements:

- A tuple of model coefficients (A_0, A_1, A_2).
- A tuple of model coefficients (B_0, B_1, B_2).
- An input to the models $\boldsymbol{x} := (x_1, x_2)$
- The result of applying models A and B to $\boldsymbol{x}$, $\hat{y}_A(\boldsymbol{x}) \in [0, 1]$ and $\hat{y}_B(\boldsymbol{x}) \in [0, 1]$.
- A modified input to the models $\boldsymbol{x}' = \boldsymbol{x} + \delta \boldsymbol{e}_m$, where $m \in \{1, 2\}$. Both $\boldsymbol{x}'$ and $\delta \in \mathbb{R}$ are shown to participants. The feature m being modified is also highlighted.

Crucially, *participants are given no indication as to what kind of model the coefficients represent.* This choice was made so as to not prime participants with any expectation of the models' behaviour [Natesan *et al.*, 2016]. Nor are participants led to believe the models are related to one another. In fact, *for all scenarios Model A was the linearisation of Model B.* Model B is the LR model $\hat{y}_B(\boldsymbol{x}) = \sigma(B_0 + B_1 x_1 + B_2 x_2)$, with the Model A coefficients computed using Definition 2.2. Upon being shown a particular scenario a user is directed to carry out the following tasks:

Direction Task. For both models A and B give the change of direction in model output, that is, compute $\text{sign}(\hat{y}_A(\boldsymbol{x}') - \hat{y}_A(\boldsymbol{x}))$ and $\text{sign}(\hat{y}_B(\boldsymbol{x}') - \hat{y}_B(\boldsymbol{x}))$. The options given to participants are {"OUTPUT INCREASES", "OUTPUT DECREASES"}.

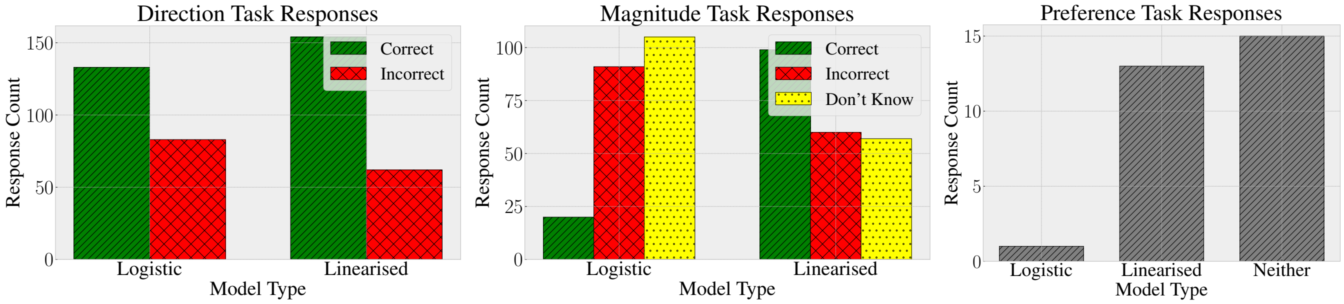

Figure 4: Summary of responses to User Survey comparing interpretability of LR models against LAM-LR. Users predict the change in direction of model output correctly at a slightly higher rate for linearised models (left). When users are asked about the magnitude of this change, users fare overwhelmingly better using LAM-LR as opposed to LR (center). When asked about which model they found easiest to use (right), the majority of users said neither model, with slightly fewer opting for LAM-LR and *only one* for LR.

Magnitude Task. For both models A and B give the magnitude of the change in model output, that is, compute $|\hat{y}_A(\boldsymbol{x}') - \hat{y}_A(\boldsymbol{x})|$ and $|\hat{y}_B(\boldsymbol{x}') - \hat{y}_B(\boldsymbol{x})|$. The options given to participants are $\{\approx 0.05, \approx 0.1, \approx 0.2, \approx 0.3,$ “DON’T KNOW”$\}$.

Respondents are shown six scenarios, not including three training scenarios shown at the beginning of the survey. The training scenarios show the same input $\boldsymbol{x}$ with progressively increasing (and decreasing) inputs $\boldsymbol{x} + \delta_u \mathbf{e}_m$ such that $|\delta_1| < |\delta_2| < \cdots$ along with the corresponding outputs, so that the user can learn the behaviour of each model class. The six test scenarios were designed to be balanced, in the sense that positive and negative changes were included as well as positive and negative values for the x_m, (B_0, B_1, B_2), so as to not skew the results. The scenarios were randomly shuffled three times and each permutation assigned to a different subgroup of participants at random. As a final task after all scenarios are presented to study participants, we query the following.

Preference Task. This is a question asked upon completion of the survey about which class of model was easiest to use. The choices are: $\{$“MODEL A”, “MODEL B”, “NEITHER”$\}$.

The survey was sent to 101 possible respondents via the SurveyMonkey platform [Inc., 2022], who are AI researchers and practitioners in the authors’ firm. Participation was anonymous and on a voluntary basis. There were 46 respondents in total, from which 36 successfully completed the training scenarios and gave answers to more than 35% of scenarios. This was the data that was analysed. Despite the small number of participants, the design of the experiment, providing multiple scenarios per-respondent, meant that we can still report results with statistical significance.

4.2 Results and Analysis

We summarise the findings of the User Survey in Figure 4. In the Direction Task we see that a slightly greater proportion of correct answers are given for the LAM as opposed to the logistic model. In the Magnitude Task we see that the logistic models received a response of “DON’T KNOW” most often, followed by an incorrect response and a small number of correct responses. The linearised counterparts of these models yielded mostly correct answers, followed by “DON’T KNOW”, then incorrect answers. In the Preference Task, the majority of users declared neither model easiest to use, followed closely by the LAM. *Only one respondent declared logistic models easiest to use*.

Statistical Analysis. The Direction and Magnitude Tasks are instances of a *clustered matched-pair binary data* experiment, with the matched pairs being the logistic model and corresponding LAM within each scenario, the binary data comprising a correct vs not correct response to an individual task on each scenario and each cluster corresponding to an individual respondent’s answers to multiple scenarios. We use the statistical test for non-inferiority in clustered matched-pair binary data of [Yang *et al.*, 2012], which accounts for within cluster correlated responses. In our setting this corresponds to an individual’s responses possibly being correlated with one another, but independent from the responses of other individuals. Suppose that $p_{\log}$ is the success probability of a respondent for a given task on a logistic model and p_{LAM} the corresponding success probability for the linearised version. The true difference between the two classes of model is $\delta = p_{\mathrm{LAM}} - p_{\log}$. Choose a small non-inferiority margin[2] $\delta_0 > 0$. Then the hypothesis test we are conducting is

$$\mathrm{H}_0 \; : \; p_{\mathrm{LAM}} - p_{\log} \leq \delta_0; \quad \text{vs} \quad \mathrm{H}_1 \; : \; p_{\mathrm{LAM}} - p_{\log} > \delta_0.$$

Yang *et al.*’s Z_{MO} test statistic asymptotically follows a normal distribution assuming the null hypothesis H_0. For the Direction Task we observe $Z_{\mathrm{MO}} = 2.822$ corresponding to a p-value of 0.0024, which is significant at the $\alpha = 0.05$ level. Moreover, the 95% confidence interval for the difference in success rates between LAMs and logistic models was $\delta \in (0.03, 0.16)$. This small value of the performance difference in the Direction Task is expected, as for both logistic models and LAMs we can easily inspect the signs of the model coeficients to get the directions, a strategy which many respondents guessed from the training scenarios. For the Magnitude Task we observe $Z_{\mathrm{MO}} = 4.55$ corresponding to a p-value of 2.72×10^{-6}, which is significant at the $\alpha = 0.05$ level. The 95% confidence interval for the difference in success rates between LAMs and logistic models was $\delta \in (0.23, 0.50)$, a significant gap. We attribute this gap to inherent non-interpretability of reasoning in log-odds space vs reasoning directly in probabilities.

[2] We choose $\delta_0 = 0.001$

For the Preference Task, the data correspond to matched pairs, each pair belonging to one participant and corresponding to a preference of logistic models vs LAM models, that is, possible responses are $\{A \prec B, A \succ B, A \sim B\}$, with Model A corresponding to LAM and Model B to logistic. As there is no numerical or ranked comparison, the usual appropriate statistical test is the Sign Test [Dixon and Mood, 1946]. Given there are many ties $A \sim B$, we use the Trinomial Test of [Bian *et al.*, 2011], specially developed for this regime. Let p_A denote the probability a randomly chosen participant prefers LAMs to logistic models and let p_B denote the converse. Moreover, p_0 is the probability neither is preferred. Then our null hypothesis is $\mathrm{H}_0 : p_A = p_B$ and alternative is $\mathrm{H}_1 : p_A > p_B$. Let N_A, N_B and N_0 be the random variables denoting the observed counts corresponding to Model A preferred, Model B preferred and neither respectively, with $N := N_A + N_B + N_0$. Assuming H_0, the test statistic $N_d = N_A - N_B$ has critical value at significance $\alpha = 0.05$ of $C_{0.05} = 6$, which is exceeded in our data with $n_d = 12$, corresponding to a p-value of 0.0009, which is statistically significant. This supports there being a user preference for LAMs over logistic models, although the impact of this is somewhat dampened by the comparatively large number of responses preferring neither.

Findings. We interpret these findings as follows: as measured by performance in basic reasoning tasks about model behaviour, *logistic models are less interpretable than their linearised counterparts*. This difference is more pronounced when it comes to reasoning about actual numerical model outputs in response to varying input variables as opposed to reasoning just about the general direction of change. Interestingly, in spite of the gulf in respondents' performance between the LAMs and logistic models this difference is not necessarily felt strongly by the respondents themselves, a large number of whom stated that neither class of models was easier to reason about. In this study, consumers of LAM explanations were correctly interpreting them *without even knowing what the underlying model was*, but were not generally confident in their interpretations. For logistic models, participants did not correctly interpret the explanations and did not declare them easy to use, in spite of their preexisting modelling expertise.

5 Conclusion

This work introduces techniques for improving the interpretability characteristics of existing models while incurring only very small penalties in classification performance. For an additive logistic model such as ARM, one can use our linearisation scheme (LAM) with the model and incur only a very small reduction in ROC-AUC and a small increase in calibration error. Via the User Study, we showed that when participants are required to simulate the output of a model, that is, predict its behaviour, their performance was far better with linearised models than their logistic counterparts.

Lessons Learned. In this work we closely examined a common tacit assumption within the XAI literature, that there is no reduction in interpretability of GAMs when moving from regression to classification via a non-linear link function (here, the logistic function). We showed that this assumption in fact does not hold in general, with the commonly used explanation method of sharing model weights or shape functions proving to be misleading to human respondents. Our LAM construction is shown to mostly overcome this issue while largely preserving an underlying model's behaviour. The implication here is that when providing model coefficients as explanations, it may be worth paying a small price in performance by linearising a trained logistic model to ensure explanations are correctly understood.

Limitations and Future Work. With this work we must take the following into consideration.

- **Model Certainty.** For a LAM $\hat{y}_{\mathrm{LAM}}$ induced by a logistic model $\hat{y}$, inputs $\boldsymbol{x}$ such that $\hat{y}(\boldsymbol{x}) \notin [0.07, 0.93]$ correspond to LAM outputs $\hat{y}_{\mathrm{LAM}}(\boldsymbol{x}) \in \{0, 1\}$. Risk scores of $\leq 7\%$ and $\geq 93\%$ correspond to confident predictions. LAMs effectively round these risk scores to certainty. If a difference in risk score between, say, 97% and 99.99% is important, then using a LAM may not be appropriate. Possible mitigations are: *i.* clip to the interval $[\epsilon, 1-\epsilon]$ for some small $\epsilon > 0$; or *ii.* increase α in the approximation to the sigmoid, thereby growing the set of model inputs with output in $(0, 1)$, while incurring a penalty in approximation.
- **Alternate Linearisation Schemes.** One could consider alternate methods of linearising as opposed to LAMs. As an example, Average Marginal Effects (AME) [Scholbeck *et al.*, 2024; Bartus, 2005] are local model explanations based on evaluating model prediction function derivatives. The AME values could potentially be interpreted as coefficients of a clipped linear model, although they would lack the theoretical guarantees afforded by LAM.
- **User Study Scope.** One could increase the scope of the User Study in terms of tasks and information provided to users, considering interpretation of the model coefficients integrated as part of a downstream task, as opposed to predicting model outputs being the tasks' focus. A potentially fruitful investigation is measuring the effect of linearisation on interpretability of general shape functions in GAMs, since LR models have linear shape functions.

6 Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[Abdul *et al.*, 2020] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, 2020.

[Alaa and van der Schaar, 2019] Ahmed M. Alaa and Mihaela van der Schaar. Demystifying Black-box Models with Symbolic Metamodels. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[Aldrich and Nelson, 1984] John Aldrich and Forrest Nelson. *Linear Probability, Logit, and Probit Models*. SagePub, 1984.

[Bartus, 2005] Tamás Bartus. Estimation of marginal effects using margeff. *The Stata Journal*, 5(3):309–329, 2005.

[Bhatt *et al.*, 2020] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA, 2020. Association for Computing Machinery.

[Bian *et al.*, 2011] Guorui Bian, Michael McAleer, and Wing-Keung Wong. A trinomial test for paired data when there are many ties. *Mathematics and Computers in Simulation*, 81(6):1153–1160, 2011.

[Bradley, 1997] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[Chen *et al.*, 2018] Hao Chen, Ran Gilad-Bachrach, Kyoohyung Han, Zhicong Huang, Amir Jalali, Kim Laine, and Kristin Lauter. Logistic regression over encrypted data from fully homomorphic encryption. *BMC Medical Genomics*, 11(4):81, 2018.

[Chen *et al.*, 2022a] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems*, 152:113647, 2022.

[Chen *et al.*, 2022b] Zixi Chen, Varshini Subhash, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. Does the explanation satisfy your needs?: A unified view of properties of explanations. *arXiv preprint arXiv:2211.05667*, 2022.

[Commission, 2021] European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts., 2021.

[Dash *et al.*, 2018] Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean Decision Rules via Column Generation. In *Advances in Neural Information Processing Systems (NIPS 2018)*, pages 4655–4665, 2018.

[De Bock and De Caigny, 2021] Koen W De Bock and Arno De Caigny. Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling. *Decision Support Systems*, page 113523, 2021.

[Demšar, 2006] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[Dixon and Mood, 1946] W. J. Dixon and A. M. Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[FICO, 2018] FICO. FICO xML Challenge found at community.fico.com/s/xml, 2018.

[Gkolemis *et al.*, 2023] Vasilis Gkolemis, Anargiros Tzerefos, Theodore Dalamagas, Eirini Ntoutsi, and Christos Diou. Regionally additive models: Explainable-by-design models minimizing feature interactions, 2023.

[Gunning, 2019] David Gunning. DARPA's Explainable Artificial Intelligence (XAI) Program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii, 2019.

[Halliwell *et al.*, 2022] Nicholas Halliwell, Fabien Gandon, Freddy Lecue, and Serena Villata. The need for empirical evaluation of explanation quality. In *AAAI: Explainable Agency in Artificial Intelligence Workshop*, 2022.

[Hanley and McNeil, 1983] J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[Harris, 2017] Naftali Harris. Logistic regression isn't interpretable, 2017.

[Hastie *et al.*, 2009] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[Hellevik, 2009] Ottar Hellevik. Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, 43(1):59–74, 2009.

[Hofmann, 1994] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

[Inc., 2022] Momentive Inc. Surveymonkey, 2022.

[Kaggle, 2019] Kaggle. Lending Club Data, 2019.

[Kaur *et al.*, 2020] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, 2020.

[Kelly *et al.*,] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. UCI machine learning repository.

[Kim *et al.*, 2018a] Andrey Kim, Yongsoo Song, Miran Kim, Keewoo Lee, and Jung Hee Cheon. Logistic regression model training based on the approximate homomorphic encryption. *BMC Medical Genomics*, 11(4):83, 2018.

[Kim *et al.*, 2018b] Miran Kim, Yongsoo Song, Shuang Wang, Yuhou Xia, and Xiaoqian Jiang. Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation. *JMIR Med Inform*, 6(2):e19, 2018.

[Kraus and Feuerriegel, 2019] Mathias Kraus and Stefan Feuerriegel. Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. *Decision Support Systems*, 125:113100, 2019.

[Kulesza *et al.*, 2013] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, 2013.

[Liang *et al.*, 2020] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Taiwanese Bankruptcy Prediction. UCI Machine Learning Repository, 2020.

[Lipton, 2018] Zachary C. Lipton. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *ACM Queue*, 16(3):31–57, 2018.

[Lou *et al.*, 2013] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, 2013.

[Molnar, 2022] Christoph Molnar. *Interpretable Machine Learning*. 2nd edition, 2022.

[Narayanan *et al.*, 2018] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[Natesan *et al.*, 2016] Divya Natesan, Morgan Walker, and Shannon Clark. Cognitive bias in usability testing. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, volume 5, pages 86–88. SAGE Publications Sage CA, 2016.

[Ng, 2011] Andrew Ng. Logistic Regression: Classification, 2011.

[OCC, 2021] OCC. Comptroller's Handbook on Model Risk Management (by US Office of the Comptroller of the Currency), 2021.

[Poursabzi-Sangdeh *et al.*, 2021] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[Quinlan, 1987] Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository, 1987.

[Register, 2021] US Federal Register. Request for information and comment on financial institutions' use of artificial intelligence, including machine learning (by US Agencies), 2021.

[Reserve, 2011] US Federal Reserve. SR 11-7/OCC11-12: Supervisory Guidance on Model Risk Management (by Federal Reserve Board and Office of the Comptroller of the Currency), 2011.

[Rong *et al.*, 2022] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: User studies for model explanations. *arXiv preprint arXiv:2210.11584*, 2022.

[Sano, 1992] Chiharu Sano. Japanese Credit Screening. UCI Machine Learning Repository, 1992.

[Scholbeck *et al.*, 2024] Christian A Scholbeck, Giuseppe Casalicchio, Christoph Molnar, Bernd Bischl, and Christian Heumann. Marginal effects for non-linear prediction functions. *Data Mining and Knowledge Discovery*, pages 1–46, 2024.

[Sudjianto and Zhang, 2021] Agus Sudjianto and Aijun Zhang. Designing Inherently Interpretable Machine Learning Models. *arXiv preprint arXiv:2111.01743*, 2021.

[Tomczak, 2016] Sebastian Tomczak. Polish Companies Bankruptcy. UCI Machine Learning Repository, 2016.

[Vaughan *et al.*, 2018] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N Nair. Explainable Neural Networks based on Additive Index Models. *arXiv preprint arXiv:1806.01933*, 2018.

[Veale *et al.*, 2018] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.

[Vidal and Schiffer, 2020] Thibaut Vidal and Maximilian Schiffer. Born-Again Tree Ensembles. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9743–9753, 2020.

[von Hippel, 2015] Paul von Hippel. Linear vs. Logistic Probability Models: Which is Better, and When?, 2015.

[Wilcox, 2022] Rand R. Wilcox. Chapter 3 - Estimating Measures of Location and Scale. In *Introduction to Robust Estimation and Hypothesis Testing*, pages 45–106. Academic Press, 5th edition, 2022.

[Wilcoxon, 1945] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[Yang *et al.*, 2012] Zhao Yang, Xuezheng Sun, and James W. Hardin. Testing non-inferiority for clustered matched-pair binary data in diagnostic medicine. *Computational Statistics & Data Analysis*, 56(5):1301–1320, 2012.

[Yang *et al.*, 2021] Zebin Yang, Aijun Zhang, and Agus Sudjianto. Enhancing Explainability of Neural Networks Through Architecture Constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2610–2621, 2021.