Data Science in Finance and Economics

DSFE, 4(1): 1–52.
DOI: 10.3934/DSFE.2024001
Received: 25 July 2023
Revised: 09 November 2023
Accepted: 11 December 2023
Published: 04 January 2024

http://www.aimspress.com/journal/dsfe

Research article

Benchmarking alternative interpretable machine learning models for corporate probability of default

Michael Jacobs, Jr*

Ph.D, CFA, Senior Vice-President, Lead Quantitative Analytics & Modeling Expert, Head – C & I 1st Line Model Development Validation & Quality Assurance, PNC Financial Services Group – Balance Sheet Analytics & Modeling/Model Development, 340 Madison Avenue, New York, N.Y. 10173, U.S.A

* **Correspondence:** Email: michael.jacobsjr@pnc.com.

Abstract: In this study we investigate alternative *interpretable machine learning* ("IML") models in the context of *probability of default* ("PD") modeling for the large corporate asset class. IML models have become increasingly prominent in highly regulated industries where there are concerns over the unintended consequences of deploying black box models that may be deemed conceptually unsound. In the context of banking and in wholesale portfolios, there are challenges around using models where the outcomes may not be explainable, both in terms of the business use case as well as meeting model validation standards. We compare various IML models (deep neural networks and explainable boosting machines), including standard approaches such as logistic regression, using a long and robust history of corporate borrowers. We find that there are material differences between the approaches in terms of dimensions such as model predictive performance and the importance or robustness of risk factors in driving outcomes, including conflicting conclusions depending upon the IML model and the benchmarking measure considered. These findings call into question the value of the modest pickup in performance with the IML models relative to a more traditional technique, especially if these models are to be applied in contexts that must meet supervisory and model validation standards.

Keywords: probability of default; credit risk; model validation, model risk; interpretable machine learning; deep neural networks

JEL Codes: G28, G17, E47, G33

1. Introduction and summary

Machine learning ("ML") models and algorithms have become predominant in several industries, notably including credit risk management. Following a long period of resistance from supervisors and model risk managers, there is at present a transition from academia to the credit risk practice, ranging from model development to various other applications in this domain. Since with this movement comes a new set of uncertainties and other difficulties (e.g., transparency around what drives model outcomes), the current focus of research is in the design of ML models that meet business requirements and supervisory expectations.

A view gaining acceptance is that there is no bright line between ML and traditional statistical models, considering that even standard regression models may be extremely complex and not readily interpretable (Breeden, 2021). Some examples include factor variables, spline approximations, interaction terms and numerous descriptive input variables. Therefore, it can be argued that what distinguishes ML from traditional statistical algorithms are optimization methodologies developed long ago in the context of fields apart from where we have applied standard econometric models. These include techniques such as bagging, boosting and random forests that are related to so-called *ensemble methods* (Clemen, 1989; Opitz and Maclin, 1999). We can gain further insight into these differences by considering the taxonomy proposed by Harrell (2018):

- ***Uncertainty:*** Statistical models specify a probability law governing the data generation process that induces the model uncertainty.
- ***Structure***: A parametric specification is typically imposed in statistical models, for example the linearity of the target variable or the parameter estimate with respect to the explanatory variables.
- ***Empiricism***: ML has a greater more feature, in which interactions are admitted of high order that are not prespecified, in contrast to statistical models that identify key parameters.

In the case of credit risk modeling, particularly for non-retail asset classes such as commercial and industrial, the datasets at hand are usually limited in depth. This is illustrated by the survey of credit scorecard models performed by Lessmann et al. (2015), who find that in about 90% of the studies reviewed there where less than 10K observations. This stands in contrast to other "big data" domains where there is more emphasis on extreme non-linearities, such as image processing (Krizhevsky et al., 2012) and natural language processing (Collobert and Weston, 2008). That said, ML has gained some traction in small dataset settings, for example through emphasizing concepts such as model robustness or the use of simplified interaction effects (Breeden, 2021).

While ML techniques are generating traction in the wider domain of credit risk, the gains have been relatively more limited in PD modeling, and even less so in wholesale as contrasted to retail asset classes. Non-wholesale PD applications have been in areas such as alternative banking channels (Abdulrahman et al., 2014), social media (Allen et al., 2020) or mobile phone use (Bjorkegren and Grissen, 2020). Indirect applications in PD modeling include ML algorithms that preprocess deposit histories (American Banking Association, 2018), that create input factors that are used in traditional methods such as logistic regression. In looking at applications of ML to risk beyond credit, we find methodologically kindred areas such as fraud detection (Zhou et al., 2018) or anti-money laundering (li et al., 2020).

The U.S. banking supervisors issued a request for information and comment on the use of ML (U.S. Banking Regulatory Agencies, 2021) where one of the top questions relates to lack of *explainability* (also termed *interpretability*) in some ML approaches and applications (e.g., fair

lending). Furthermore, a less transparent and explainable approach might result in difficulties in evaluating the conceptual soundness of a model, which is an important model risk management consideration outlined in SR11-7/OCC11-12 (U.S. Banking Regulatory Agencies, 2011). In August 2021, the Office of the Comptroller of the Currency ("OCC") released the model risk handbook (OCC, 2021) prescribing with respect to banks employing models that examiners are expected to determine if credit rating systems account for explainability and transparency as key considerations in managing the model risk of models deemed to be complex. While ML models that are not interpretable (or "black-box") are by construction not explainable nor transparent, there are methods such as *locally interpretable model-agnostic explanations* that are model-agnostic (Ribeiro et al., 2016). Similarly, methods such as *Shapley additive explanations* (Lundberg and Lee, 2017) avail us of explanations that are approximate, but as with the former technique this has reason for us to exercise caution (Molnar et al., 2020). Analogous to the no *free lunch theorem* in finance in the context of ML, it can be shown that there exists no model-agnostic and universally applicable notion of explainability. Extensive academic literature criticizes the uncritical applications of these methods (Kumar et al., 2020; Rudin, 2019, Slack et al., 2020). However, there is a class of *inherently interpretable* ML models with model-based explainability, where this means that the model is transparent and self-explanatory (Sudjianto and Zhang, 2021). Yang et al. (2021a) argue the inherent interpretability of models deemed complex should rely on constraints of practical nature, which the authors extend to a methodology for qualitative assessment of the interpretability of ML models.

Credit risk modeling in an ML context is typically assumed to involve large volumes of data for training purposes, which is more common in the common in the data rich domain of the retail asset classes, whereas in wholesale asset classes (e.g., corporate or commercial real estate) the situation is in fact similar albeit having less defaults as well as fewer risk factors of a standardized nature. While it is widely accepted that in the wholesale context one of the leading uses of ML is the standardization of risk factors that may be rather heterogeneous, in fact there exist several large datasets in this asset class, and there is a large body of published research that have applied ML in model building (Vahid and Ahmadi, 2016; Anagnostou et al., 2020). There is a distinction between bankruptcy and default, and since the former are public the focus of most research is on modeling that event (Odom and Sharda, 1990, Coats and Fant, 1993, Mckee, 2000, Min and Lee, 2005; Vassiliou, 2013), with methods tested covering all ML techniques. Models for lending to *small- and medium-sized enterprises* ("SMEs") fall in between consumer and commercial approaches, because the performance is more closely tied to a small group of owners. Although less data is available for SMEs there has been some research in applying ML in this domain (Li et al., 2016; Zhu et al., 2017), with findings commercialized by the novel fintech industry in this market leveraging ML methods and alternative data sources.

In this study we investigate alternative *interpretable machine learning* ("IML") models in the context of *probability of default* (PD) modeling using a dataset of corporate borrowers. IML models have become increasingly prominent in highly regulated industries where there are concerns over the unintended consequences of black box models that may be deemed conceptually unsound. In the context of banking and in wholesale portfolios, there are challenges around deploying models where the outcomes may not be explainable, both in terms of the business use case as well as meeting model validation standards. We compare various IML models, including standard approaches such as logistic regression, using a long and deep history of corporate borrowers sourced from Moody's studied in Jacobs (2022a, 2022b). This data consists of approximately 200K observations at a quarterly frequency comprised of North American based large corporate obligors having public credit ratings in the period

1990–2015. This dataset contains a large set of potential explanatory variables, including financial and macroeconomic risk factors.

In our comparison of various IML models (deep neural networks and explainable boosting machines), including standard approaches such as logistic regression, we find that there are material differences between the approaches in terms of dimensions such as model predictive performance and the importance or robustness of risk factors in driving outcomes, including conflicting conclusions depending upon the IML model and the interpretability or robustness measure considered. While we observe that the IML models all demonstrate some pickup in performance relative to a more traditional technique, the degree of this outperformance is modest, especially on an out-of-sample basis. We also observe in a comparison of interpretability measures across the IML models and a more traditional technique an overarching complete lack of consistency across both models and measures, which calls into question the value of this pickup in performance with the IML models, especially if these models are to be applied in contexts that must meet model validation standards.

This paper shall proceed according to the following outline. A review the literature is presented in Section 2. In Section 3 we present modeling methodology, where we introduce the alternative IML frameworks that we will compare empirically. In Section 5 we describe the modeling data. Section 6 encompasses the estimation results and benchmarking analysis. Section 7 we conclude, address implications for policy and propose related research for future work.

2. Review of the literature

In this review of the relevant literature, we will first cover some of the basic literature in credit risk modeling, and then proceed to ML research as they apply to this study.

The seminal study in PD modeling that introduced the industry standard PD scoring model was presented by Altman (1968) who applied the methodology of *multiple discriminant analysis* ("MDA"). While MDA has the advantage of computational convenience since it assumes a Gaussian error structure and linear model specification, we note that this edge is marginal at best given recent advances in computational capabilities. Mester (1997) notes the prevalence and growth of such models in U.S. banking. Altman and Narayanan (1997) find that spanning geographies and borrower types that such approaches are remarkably similar. A popular vendor PD scorecard model, very common in banking, is the *Moody's Analytics* (Dwyer et al., 2004; "MA") model for private firms, which is considered very adaptable.

Merton (1974) takes a rather different approach than credit scoring, modeling a levered firm's equity as a call option on the assets of the firm, where the strike price is equal to the debt owed, an approach based on option pricing theory. The PD is estimated iteratively in a method that extracts unobserved value and volatility of assets, using the amount of debt owed at some point in the future, in the process deriving the *distance-to-default* ("DTD") construct. DTD is understood as the quantity of standard deviations from asset value at a point in time from the value of the debt obligations, so that DTD is inversely related to the PD. There are strict assumptions associated with this construct that have been addressed in the subsequent literature. A popular vendor model based upon this approach is MA's *CreditEdge*™ ("CE") model meant for firms with publicly traded equity, which is calibrated to stock process and observed defaults to derive the *expected default frequency* ("EDF"). Since the EDF is based upon equity market data it is more volatile measures as compared to PD ratings derived from credit scoring models.

An alternative to the above is the *reduced form* approach that applies intensity modeling to derive a stochastic hazard rate (Jarrow and Turnbull, 1995; Duffie and Singleton, 1999; Bonds, 1999). This approach differs from the structural models by eschewing the underlying economics driving the default process through decomposing prices of defaultable debt as a means of estimating a random intensity process. This features the benefit of not being restricted by the assumptions at play in the structural approach but has the downside of potentially measuring risks apart from credit such as liquidity risk premia that complicate the construct. The *Kamakura Risk Manager*TM vendor model is in this class, which statistically implements the reduced form *Jarrow-Chava Model* (Chava and Jarrow, 2004) model. This version of the reduced form approach adjusts explicitly for the aforementioned liquidity risk premia, but at the cost of incorporating embedded optionality noise and other market distortions.

There is substantial ML literature apart from the domain of PD modeling, where subsequently applications in the latter area have been found. Clemen (1989) surveys the considerable literature regarding the combination of forecasts and concludes that forecast accuracy can be substantially improved through the combination of multiple individual forecasts. This author considers contributions from the forecasting, psychology, statistics, and management science literatures. Opitz and Maclin (1999) consider an ensemble of a set of individually trained classifiers whose predictions are combined when classifying novel instances, such as *bagging* and *boosting*, and evaluate these methods on 23 datasets using both neural networks and decision trees as the classification algorithm. The authors find that bagging is almost always more accurate than a single classifier but sometimes much less accurate than boosting, while in boosting the performance is dependent on the characteristics of the dataset being examined.

Turning to applications of ML techniques to non-wholesale PD modeling, fuzzy logic is applied in Abdulrahman et al. (2014) in the context of micro-finance in a developing country. The American Banking Association (2018; "ABA") announced a new credit score, called the UltraFICO score as part of a bid to ensure that creditworthiness is better reflected through credit scoring, drawing on consumer-contributed data (e.g., information from checking and savings accounts) that reflects responsible financial management. The ABA argues that the new score could improve credit access for the majority of Americans and is particularly relevant for those who fall in the grey area in terms of credit scores or fall just below a lender's score cut-off. Allen et al. (2020) consider social networks associated with the demand for and supply of consumer and small business loans originated on lending marketplaces, finding that loan demand increases substantially with past borrowing activities of geographically distant but socially connected areas, whereas borrower-area social proximity to deposits increases funding likelihood by 5.61% and improves ex-post loan performance. The authors argue that social networks improve capital allocation by increasing the awareness of alternative lending platforms and facilitates the transmission of less accessible information complementary to loan-specific data.

There is a deep stream of literature that considers applications of ML techniques to the prediction of corporate defaults or bankruptcies. Vahid and Ahmadi (2016) model a set of credit states (good, past due, overdue and doubtful) defined by the Iranian central bank to model solvency and insolvency rates. The model utilizes a hybrid a radial basis function neural network featuring a self-organizing map methodology that is found to perform better than methods such as a single- and four-step classifications as well as *support vector machines* ("SVM"). Anagnostou et al. (2020) propose a method to enhance credit portfolio models based on the model of Merton by incorporating contagion effects using Bayesian network methods while maintaining the convenient representation of factor models. A range of techniques are applied to learn the structure and parameters of financial networks

from real default swaps data and the impact on standard risk metrics is estimated in a stylized portfolio. Odom and Sharda (1990) develop a neural network prediction of bankruptcy using financial data from various companies, which is compared to a multivariate discriminant model, where results show some potential for outperformance in this context. Coats and Fant (1993) build a neural network to discriminate solvent from distressed obligors using financial ratio risk factors with an application an early distress identification system. Mckee (2000) builds a bankruptcy prediction model based upon rough sets theory with variables identified in prior recursive partitioning research that is shown to be was 93% accurate in predicting bankruptcies on a large sample of U.S. public companies, as well as 88% accuracy on a separate 100-company holdout sample, with for the same dataset a baseline recursive partitioning model shows only 65% accuracy. Min and Lee (2005) apply SVMs for the prediction of public large corporate bankruptcy, utilizing a grid-search methodology that features 5-fold cross-validation, to estimate optimized coefficient estimates parameterizing the kernel function of the SVM. The authors find that on the same dataset this SVM approach has superior performance to MDA, logistic regression and three-layer fully connected back-propagation neural network models. Vassiliou (2013) studies the credit rating agency methodologies though applying fuzzy set theory, demonstrating that a fuzzy economy is admissible if and only if there exists an equivalent martingale measure exists, and further constructs a forward probability measure under which a default-free security price once discounted is a martingale. The author proceeds to apply these finding in modeling the credit migration dynamics of a defaultable bond as an inhomogeneous semi-Markov process having fuzzy states, and analyzes the consequences of switching from a physical to a forward probability measure, as well as implements parameter estimation and calibration on a sample of traded corporate bonds.

There is also a growing literature that considers applying ML techniques to the prediction of SME defaults or bankruptcies. Li et al. (2016) introduce a PD modeling approach for SMEs that is hybrid model combining logistic regression and *artificial neural networks* ("ANN") applied to Finnish firms from the fiscal years 2004 to 2012. Their results suggest that the proposed hybrid model is more accurate than either the ANN or logistic regression models in isolation. Zhu et al. (2017) applies six ML methods (one decision tree method, three ensemble ML methods, a random subspace method and two *integrated ensemble ML* methods ("IEML")) to predict default for 57 SMEs listed on the Shenzhen and Shanghai Stock Exchanges during the period of 2012–2013. They find that IEML methods outperform either the decision tree or random subspace methods.

There is a body of supervisory guidance for the financial industry that is a motivation for considering IML techniques in credit risk modeling. The interagency guidance on the management of model risk (U.S. Banking Regulatory Agencies 2011) provides comprehensive guidance for banks that transcends model validation, addressing standards for model development, implementation and use. Furthermore, this guidance encompasses governance and control mechanisms such as board and senior management oversight, policies and procedures, controls and compliance, and an appropriate incentive and organizational structure. The bank examiners handbook issued by the U.S. Office of the Comptroller of the Currency (OCC 2021) provides guidance in performing consistent and high-quality model risk management examinations. This guidance presents the concepts and general principles of model risk management, informs and educates examiners about sound model risk management practices that should be assessed during an examination and provides information needed to plan and coordinate examinations on model risk. The interagency request for information (U.S. Banking Regulatory Agencies 2021) gathers information and comments on financial institutions' use of *artificial intelligence* ("AI") models, including ML models. The purpose of this is to understand

respondents' views on the use of AI by financial institutions in their provision of services to customers and for other business or operational purposes; including appropriate governance, risk management, and controls over AI; and any challenges in developing, adopting, and managing AI.

We conclude this review by surveying the relevant literature in the field of IML. Ribeiro et al. (2016) propose the *local interpretable model-agnostic explanations* ("LIME") technique. This approach forms predictions through discovering locally an interpretable in the neighborhood of the prediction. Lundberg and Lee (2017) propose *Shapley additive explanations* ("SHAP") as a means of interpreting predictions, assigning an importance value to each feature corresponding to a prediction. SHAP identifies a novel additive feature importance class of measures and derives a unique solution to this problem having properties considered attractive. It is shown that SHAP subsumes six known techniques that lack such good qualities with the added benefit of more rapid calibration and a more intuitive expression. In a point-of-view piece, Rudin (2019) contrasts the explainability of black box ML models versus ML models that are inherently interpretable. The author highlights the dangers of using explainable black boxes for important decision making, challenges inherent in interpretable ML and identifies some applications where IML models are competitive with black box ML models. Molnar et al. (2020) point out some downsides in of ML model interpretation including poor generalization, feature dependence or spurious causal interpretation. Kumar et al. (2020) apply the game theoretic construct of a cooperative game to feature importance computation in an ML setting, where influence is distributed between input factors, deriving a form of the unique SHAP values characterizing the game. The authors justify this method based upon mathematical properties deemed desirable as well as the applicability of this construct in model explainability. They further demonstrate the mathematical issues that arise in using Shapley values in the context of feature importance, and that while there exist mitigating solutions to these problems, the latter give rise additional model complexity like the necessity of applying causal reasoning. Slack et al. (2020) put forth the argument that post hoc explanation methods reliant upon perturbations of inputs, such as LIME or SHAP, can be shown to be unreliable. Instead of those methods they propose a new scaffolding technique that masks the biases associated with a single classifier through enabling an adversary, resulting in construction of arbitrary needed explanations. Sudjianto et al. (2020) apply local linear representation "unwrapping" of the black box of deep ReLU networks, using the concept of an *activation pattern* capable of disentangling a complex network into a set of *local linear models* (LLMs) that are equivalent to the underlying network. As part of this process the authors develop a user-friendly package for implementing this construct that includes feature importance metrics, model diagnostics and the provision of a pre-trained deep *rectified linear unit* ("ReLU") network that simplifies the original problem. The authors further propose visual interpretations and diagnostics, such as the *local linear profile plot*, and argue that these tools are an effective means for simplifying a network simplification. Finally, they implement these methods in simulation exercises, benchmark training and testing data, as well as a real world example mortgage credit risk modeling. Yang et al. (2021a) approach the explainability of a deep neural networks though employing various constraints on model architecture (projection pursuit with orthogonality constraints, sparse additive subnetworks and smooth function approximations), which gives rise to an *enhanced explainable neural network* ("ExNN"), and the authors argue that this construct strikes an improved balance amongst the interpretability and performance of the model. The authors further demonstrate sufficient conditions of identifiability of this ExNN model, and implement estimation of the model using a modified minibatch gradient descent method, featuring a backpropagation algorithm that computes the Cayley

transform and derivatives which preserving orthogonality of the projection. In a simulation exercise featuring six alternative scenarios the authors compare their method with a number of benchmarks (SVM, least absolute shrinkage and selection operator, extreme learning machine, random forest and multilayer perceptron), demonstrating that the ExNN model maintains the flexibility of favorable model performance while achieving augmented interpretability. Yang et al. (2021b) develop an ExNN model based on the *generalized additive models with structured interaction* ("GAMI-Net") model and argue that this construct provides good balance between prediction accuracy and model interpretability. The authors point out that GAMI-Net is a decomposed feedforward network having multiple additive subnetworks, where each subnetwork consist of multiple hidden layers, where this construct is architected for the capture of one main effect or one pairwise interaction. Three interpretability features are analyzed (heredity, sparsity and marginal clarity) and they develop an adaptive training algorithm, in which first main effects are trained and second pairwise interactions are optimized with respect to the residuals. Numerical experiments on both real-world datasets and synthetic functions show that their model has augmented interpretability and favorable performance as compared to explainable boosting machines as well as other traditional ML models.

3. Modeling methodology and econometric technique

In this section we will describe the IML techniques that we will benchmark in our application to corporate PD modeling. We first describe the *logistic regression modeling* ("LRM") technique, which widely understood in the literature and applied by practitioners, and serves as our base case model against which we compare the other IML models in this paper. Defining the classes $\{\omega_i\}_{i=1}^{2}$ of the classification problem, we can write the *log-odds* (which is also called the *logit function*):

$$ln\left(\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}\right) = \boldsymbol{\theta}^T\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{x} = (x_1, .., x_k) \in \mathbb{R}^k$ is a vector of risk factors having dimension k and $\boldsymbol{\theta} = (\theta_1, .., \theta_k) \in \mathbb{R}^k$ is a vector of coefficients. Note that we set $x_1 = 1$, which implies that the intercept is absorbed into $\boldsymbol{\theta}$. As $P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x}) = 1$, we get that:

$$P(\omega_1|\boldsymbol{x}) = \frac{1}{1+exp(-\boldsymbol{\theta}^T\boldsymbol{x})} = \sigma(-\boldsymbol{\theta}^T\boldsymbol{x}), \tag{2}$$

where the $\sigma(-\boldsymbol{\theta}^T\boldsymbol{x})$ is the *logistic sigmoid* (also called the *sigmoid link*) function. This function has properties mathematically equivalent to a cumulative distribution function, as it has range in $(0,1)$ with the real line as the domain, and we may interpret as a PD conditioned upon $\boldsymbol{\theta}^T\boldsymbol{x}$, where the latter can be viewed as a score since it is positively related to the risk of default.

A standard method to estimate the vector of parameters $\boldsymbol{\theta}$ is *maximum likelihood estimation* ("MLE"). In this construct we specify set of training samples defined by features $\{\boldsymbol{x}_n\}_{n=1}^{N}$ and a classification target $\{y_n\}_{n=1}^{N}$, with $y_n \in \{0,1\}$, and specify the likelihood function as:

$$P(y_1, \ldots, y_N|\boldsymbol{\theta}) = \prod_{n=1}^{N}\left(\sigma(-\boldsymbol{\theta}^T\boldsymbol{x}_n)\right)^{y_n}\left(1-\sigma(-\boldsymbol{\theta}^T\boldsymbol{x}_n)\right)^{1-y_n}. \tag{3}$$

Typically will have as our objective the negative *log-likelihood function* (also called the *cross-entropy error*), which has advantages in computation, a transformation of (3) that is monotonic and increasing:

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} y_n\, ln\left(\sigma(-\boldsymbol{\theta}^T\boldsymbol{x}_n)\right) + (1-y_n)\, ln\left(1-\sigma(-\boldsymbol{\theta}^T\boldsymbol{x}_n)\right). \tag{4}$$

We minimize (4) with respect to $\boldsymbol{\theta}$ by means of iterative methods, examples being Newton's scheme or steepest descent. Importantly, this model has properties of computational convenience and in most cases the estimation is stable. The reason for this is that $\sigma(-\boldsymbol{\theta}^T \boldsymbol{x}_n) \in (0,1)$, which implies that the covariance matrix $\mathbf{R}$ is positive definite, from which it follows in turn that the Hessian matrix $\nabla^2 L(\boldsymbol{\theta})$ is positive definite. This means that $L(\boldsymbol{\theta})$ is *convex* and the optimization will have a unique minimum. A potential issue arises if the development data is *linearly separable,* a situation in which for any point on the hyperplane $\widehat{\boldsymbol{\theta}}_{MLE}^T \boldsymbol{x} = 0$, so in solving the problem the separation of the classes occurs *perfectly*. As a result, every point receives an assignment probability of identically one as $\boldsymbol{\sigma}\left(\widehat{\boldsymbol{\theta}}_{MLE}^T \boldsymbol{x}\right) = \frac{1}{2}$, and the algorithm makes the estimates infinite ($\widehat{\boldsymbol{\theta}}_{MLE}^T \to \infty$). In terms of geometry, rather than an s-curve the link function evolves toward a step shape. In this setting we are dealing with overfitting to the development data. Treatments for this problem include k-fold cross-validation, else regularization terms residing within a penalty function that restricts the size of the coefficient estimates, examples of the latter being the LASSO technique that features a linear penalty $C(\boldsymbol{\theta}|\lambda) = \lambda|\boldsymbol{\theta}|$ having cost parameter λ.

We now will describe the ReLU-DNN model, among the most popular of DNNs, as the the ReLU activation function in the hidden layers is known to have better properties than DNNs using the sigmoid activation function as discussed above for the LRM. As detailed by Sudjianto et al. (2020), this model exhibits superior performance, rich expressivity, universal approximation ability, double descent risk curves under over-parameterization, as well as fast and scalable training algorithms by stochastic gradient descent. $\mathcal{N}$ is the symbol that denotes a feedforward ReLU network and we define an input layer $x \in \Omega \subseteq \mathbb{R}^d$ and L hidden layers with neuron sizes $\left[n_1, \ldots, n_L\right]$. In the case of each hidden neuron $\boldsymbol{u}_{\mathrm{i}}^{(1)}$, denote its leftward input as $z_{\mathrm{i}}^{(1)}$ and the rightward output as $\chi_{\mathrm{i}}^{(1)}$, then by the ReLU activation

$$\chi_{\mathrm{i}}^{(1)} = \max\left\{0, z_{\mathrm{i}}^{(1)}\right\}, \quad i = 1, \ldots, n_l \text{ and } l = 1, \ldots, L. \tag{5}$$

From layer $l-1$ to l, given the weight matrix $\boldsymbol{W}^{(i-1)}$ of size $n_i \times n_i$ and the bias vector $\boldsymbol{b}^{(i-1)}$ of size $n_i \times 1$, the layer $l-1$ output $\boldsymbol{\chi}^{(1-1)}$ leads to the layer-l input $z^{(1)}$ by the affine transformation

$$z^{(1)} = \boldsymbol{W}^{(i-1)} \boldsymbol{\chi}^{(1-1)} + \boldsymbol{b}^{(\mathrm{i}-1)}, \quad \text{for } l = 1, \ldots, L, \tag{6}$$

where the layer zero corresponds to the input layer with $\boldsymbol{\chi}^{(0)} = \boldsymbol{x}$, $\boldsymbol{W}^{(0)}$ of size $n_1 \times d$ and $\boldsymbol{b}^{(0)}$ of size $n_1 \times 1$ with respect to an output layer (i.e., layer $L+1$). At L layer we have the output $\chi_{\mathrm{i}}^{(\mathrm{L})}$ which predicts the final target y according to a *generalized linear model* ("GLM"), which in the case of classification can be the sigmoid link function as defined previously for the LRM:

$$\mathbb{E}(y) = \sigma\left(\boldsymbol{w}^{(L)} \boldsymbol{\chi}^{(\mathrm{L})} + b^L\right) = \sigma(\eta(\boldsymbol{x})), \tag{7}$$

where a univariate formulation with a binary response $\boldsymbol{W}^{(L)} \equiv \boldsymbol{w}^{(\mathrm{L})}$ is a row vector and $\boldsymbol{b}^{(L)} \equiv b^{(\mathrm{L})}$ is a scalar. For convenience, let denote the set of network parameters by $\theta = \{\boldsymbol{W}^l, \boldsymbol{b}^l: \ l = 0, 1, \ldots, L\}$. In the case where y is "one-hot encoded" (i.e., categorical indicator that take the values zero or one) the *softmax link* may be utilized, and this construct may be readily extended. In the case of every hidden neuron there are two states in ReLU activation: active when $z \geq 0$ ("on") or inactive when $z < 0$. We analyze these states pertaining to all of the hidden neurons in the ReLU network by leveraging the concept of an *activation pattern* (Sudjianto et al., 2020), which means that if we have a network $\mathcal{N}$ having L hidden layers with sizes of neurons $\left[n_1, \ldots, n_L\right]$, the activation pattern is a binary vector

$$\boldsymbol{P} = \{\boldsymbol{P}^{(1)}; \ldots; \boldsymbol{P}^{(L)}\} \in \{0,1\}^{\sum_{l=1}^{L} n_l}, \tag{8}$$

which indicates for each hidden neuron state (i.e., on vs. off). We term the component $\mathbf{P}^{(l)}$ for $l = 1, \ldots, L$ a *layered pattern*. In the case where there is at least one $P^{(l)} = 0$ for some l we call the activation pattern $\mathbf{P}$ *trivial*. $\sum_{l=1}^{L} n_l$, the count of the network's hidden neurons, is called the activation pattern's *length*. With respect to the input $x \in \Omega$, through the feedforward neural network with fixed parameters, it would determine the values of $\left\{z^{(l)}, \chi^{(l)}\right\}$ by (6) and (5) in sequence pertaining to $l = 1, \ldots, L$. $P^{(l)}(x)$ governing how every $\boldsymbol{z}^{(l)}$ determines the layer l neurons' states, which is equivalent to a layered pattern. This implies that an instance of input $\boldsymbol{x}$ is paired with a set activation pattern

$$P(x) = \left[P^{(1)}(x); \ldots; P^{(L)}(x)\right], \tag{9}$$

and we note that such may or may not be trivial. Furthermore, the *activation region* associated with every distinct pattern is given by a *convex polytope*

$$\mathfrak{R} = \{x \in \Omega: P(x) = P\}, \tag{10}$$

having closed-form boundaries which we discuss in more detail later. The concept of the activation region is critical in grasping the way in which hidden layers are used to partition the entire space Ω sub-spaces that are into disjoint. Therefore, in the general case of ReLU_DNNs in general, where $P = \left[P^{(1)}; \ldots; P^{(L)}\right] \in \mathscr{P}_{\text{expr}}(\mathcal{N})$ represents the set of distinct activation patterns, each may be paired with a *convex activation region* $\mathscr{R}^P \subset \Omega$. The latter obeys a set of inequality constraints that is derived from layer-wise affine-transformed variables of the following form:

$$(-1)^{P(l)} \odot z(l) \leq 0, \quad l = 1, \ldots, L \tag{11}$$

where $\odot$ is the Hadamard product. In the case of any pair of distinct activation patterns their associated activation regions are disjoint. This implies that the ReLU_DNN represents a partition of the underlying input space into a finite set of convex sub-spaces

$$\Omega = \bigcup_{P \in \mathscr{P}_{\text{expr}}(\mathcal{N})} \mathscr{R}^P. \tag{12}$$

Note that $\mathscr{P}_{\text{expr}}(\mathcal{N})$ is not yet determined and continues to be an ongoing research area in the so-called in the study of DNN *expressivity* that is dictated by the upper and lower bounds of the aggregate count of distinct patterns (Sudjianto et al., 2020). The latter authors note the cardinality of $\mathscr{P}_{\text{expr}}(\mathcal{N})$ has an upper bound of $O\left(k^{dL}\right)$ for a ReLU-DNN having L hidden layers with width of k. Any ReLU-DNN performs a partition of the input space into multiple sub-regions determined by a particular activation pattern. The multi-layer propagation process transforms the original input sequentially according to (6) and constrained by the ReLU activation as defined in (5). At the end of the process linear prediction takes the form

$$\eta(\boldsymbol{x}) = w^{(L)} \chi^{(L)} + b^{(L)}, \tag{13}$$

up to a $\sigma(\bullet)$, a pre-specified link function. Then the final features $\chi^{(L)}$ are explicitly derived with respect to each sub-region, giving rise to closed form local linear models. If the activation pattern is trivial (i.e., at least one layer features all neurons that are all inactive) the output with respect to this layer is all zero. In the associated trivial region the prediction is constant value in each region as the final output is only influences by layer-wise bias terms and not by $\boldsymbol{\chi}$, the original input variables.

In the case where the activation pattern, $\boldsymbol{P}=\left[\mathbf{P}^{(1)};\ldots;\mathbf{P}^{(L)}\right]$, is non-trivial we define a diagonal matrix $\boldsymbol{D}^{(l)}$ corresponding to each layer, where diagonal feature identical $(0,1)$ values as in the case $\boldsymbol{P}^{(l)}$,

$$\boldsymbol{D}^{(l)}=\text{diag}(\boldsymbol{P}^{(l)}), \qquad \text{for } l=1,\ldots,L, \tag{14}$$

so that the output in each layer after activation by the ReLU (5) may be expressed according to the chain rule vector notation,

$$\boldsymbol{\chi}^{(l)}=\max\left\{0,z^{(l)}\right\}=D^{(l)}z^{(l)}=\boldsymbol{D}^{(l)}(\mathbf{W}^{(l-1)}\boldsymbol{\chi}^{(l-1)}+\mathbf{b}^{(l-1)}) \quad \text{for } l=1,\ldots,L. \tag{15}$$

Thus, we may derive the final features $\boldsymbol{\chi}^{(L)}$ recursively as follows,

$$\begin{aligned}
\boldsymbol{\chi}^{(\mathrm{L})} &= \mathbf{D}^{(\mathrm{L})}(\mathbf{W}^{(\mathrm{L}-1)}\boldsymbol{\chi}^{(\mathrm{L}-1)}+\mathbf{b}^{(\mathrm{L}-1)}) \\
&= \mathbf{D}^{(\mathrm{L})}\mathbf{W}^{(\mathrm{L}-1)}(\mathbf{D}^{(\mathrm{L}-1)}\mathbf{W}^{(\mathrm{L}-2)}\boldsymbol{\chi}^{(\mathrm{L}-2)}+\mathbf{b}^{(\mathrm{L}-2)})+\mathbf{D}^{(\mathrm{L})}\mathbf{b}^{(\mathrm{L}-2)} \\
&= \ldots \\
&= \mathbf{D}^{(\mathrm{L})}\mathbf{W}^{(\mathrm{L}-1)}\ldots\mathbf{D}^{(0)}\mathbf{W}^{(0)}\mathbf{x}+\sum_{\mathrm{l}=1}^{\mathrm{L}-1}\mathbf{D}^{(\mathrm{L})}\mathbf{W}^{(\mathrm{L}-1)}\ldots\mathbf{D}^{(\mathrm{l}+1)}\mathbf{W}^{(\mathrm{l})}\mathbf{D}^{(\mathrm{l})}\mathbf{b}^{(\mathrm{l}-1)}+\mathbf{D}^{(\mathrm{L})}\mathbf{b}^{(\mathrm{L}-1)} \\
&= \prod_{\mathrm{h}=1}^{\mathrm{L}}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{W}^{(\mathrm{L}-\mathrm{h})}\mathbf{x}+\sum_{\mathrm{l}=1}^{\mathrm{L}-1}\prod_{\mathrm{h}=1}^{\mathrm{L}-1}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{W}^{(\mathrm{L}-\mathrm{h})}\mathbf{D}^{(\mathrm{l})}\mathbf{b}^{(\mathrm{l}-1)}+\mathbf{D}^{(\mathrm{L})}\mathbf{b}^{(\mathrm{L}-1)}.
\end{aligned} \tag{16}$$

It follows that the linear prediction (13) in the output layer has an explicit expression given by

$$\eta(\boldsymbol{x})=\prod_{\mathrm{h}=1}^{\mathrm{L}}\mathbf{W}^{(\mathrm{L}-\mathrm{h})}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{W}^{(0)}\mathbf{x}+\sum_{\mathrm{l}=1}^{\mathrm{L}-1}\prod_{\mathrm{h}=1}^{\mathrm{L}+1-h}\mathbf{W}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{b}^{(\mathrm{l}-1)}+\mathbf{b}^{(\mathrm{L})}, \tag{17}$$

in which $\mathbf{W}^{(\mathrm{L})}\equiv\mathbf{w}^{(\mathrm{L})}$ for notational convenience. This leads to the result that characterizes a ReLU-DNN network and any of its expressible activation pattern $\boldsymbol{P}$ on the activation region $\mathscr{R}^{\boldsymbol{P}}$ in terms of an LLM:

$$\eta(\boldsymbol{x})=\tilde{\mathbf{w}}^{P}\boldsymbol{x}+\tilde{b}^{P}, \quad \forall\boldsymbol{x}\in\mathscr{R}^{P} \tag{18}$$

with the following closed-form parameters

$$\tilde{\mathbf{w}}^{P}=\prod_{\mathrm{h}=1}^{\mathrm{L}}\mathbf{W}^{(\mathrm{L}-\mathrm{h})}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{W}^{(0)}\mathbf{x}, \quad \tilde{\mathbf{b}}^{\mathrm{P}}=\sum_{\mathrm{l}=1}^{\mathrm{L}-1}\prod_{\mathrm{h}=1}^{\mathrm{L}+1-h}\mathbf{W}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{D}^{(\mathrm{L}+1-\mathrm{h})}\mathbf{b}^{(\mathrm{l}-1)}+\mathbf{b}^{(\mathrm{L})}. \tag{19}$$

Through combinations of the region partitioning in (12) and the LLMs in (19) this gives rise to the *local linear representation* with respect to ReLU-DNNs. This means that a ReLU-DNN has an equivalent representation as a finite set of LLMs, and each of these functions is defined only on one of these disjoint convex sub-regions. We can see from this that the activation pattern is a central concept in this local linear representation. It only remains to determine the activation patterns for purposes of interpretability of the network, where we refer the reader to Sudjianto et al. (2020) for the development an effective unwrapper for pre-trained ReLU-DNNs that we utilize in this paper.

The 2nd IML model under consideration is the GAMI-Net. In this model we formulate a complex functional relationship through building up from lower-order representations. The latter include nonlinear main effects as well as pairwise interactions. We denote S_1, S_2 as a set of active main effects and its respective pairwise interactions. Then the GAMI-Net may be written as follows:

$$g\left(\mathrm{E}\left(y|x\right)\right)=\mu+\sum_{j\in S_1}h_j\left(x_j\right)+\sum_{(j,k)\in S_2}f_{j,k}\left(x_j,x_k\right). \tag{20}$$

In this set-up each of the main effects and the pairwise interactions are assumed to be quantities having zero mean,

$$\begin{aligned} \int h_j\left(x_j\right) dF\left(x_j\right) = 0, \ \forall j \in S_1, \\ \int f_{j,k}\left(x_j, x_k\right) dF\left(x_j, x_k\right) = 0, \ \ \forall (j,k) \in S_2, \end{aligned} \tag{21}$$

where in (21) $dF(x_j)$ and $dF(x_j, x_k)$ are the respective distribution functions. With a view to ensuring identifiability we enforce the near orthogonality of each pairwise interaction term $dF(x_j, x_k)$ with respect to their corresponding parent main effects $h_j(x_j)$ and $h_k(x_k)$. Note that this requirement is subject to a *marginal clarity constraint* that we will discuss below.

The architecture of the GAMI-Net model features a main effect and pairwise interaction sub-model. Every of the main effect $h_j(x_j)$ in (20) consists of a sub-network having a single input node, multiple hidden layers and then a single output node. The pairwise interaction $f_{j,k}(x_j, x_k)$ in (20) are each represented by a sub-network having two input nodes. These networks are all combined linearly with an additional bias node to capture the intercept μ, which produces the ultimate output. In particular, the sub-network of the main effect is projected upon a 1-dimensional curve, while the sub-network of interaction terms is calibrated to a 2-dimensional surface. In the approximation of an arbitrary surface of curve, a single-hidden-layer feedforward neural network, having a sufficiently numerous quantity of hidden nodes, may be used. In this process we leverage the techniques of modern deep learning in order to facilitate the utilization of multiple hidden layers, which results augmented model performance. If the network is properly configured, this construct has enough flexibility to capture any functional form, including the admission of categorical variables that with one-hot encoding may be preprocessed. We may simplify this to many bias nodes with subnetworks used for fitting the main effects of categorical variables in which each node represents an intercept effect with respect to a corresponding categorical variable. Finally we note that *sparsity*, *heredity* and *marginal clarity* constraints are imposed upon the GAMI-Net model development. Sparsity and heredity constraints are meant to facilitate the interpretability of the estimated model, while the purpose of the marginal clarity constraint is to ensure that the main effects and their corresponding child pairwise interactions are uniquely identifiable.

The final model that we consider is the *explainable boosting machine* ("EBM") introduced by Lou et al. (2013), an IML model designed at the time to have accuracy comparable to widely used ML methods such as random forests or boosted trees, which we will describe at a high level and refer the reader to the reference for mathematical details. The EBM is tree-based and a cyclic gradient boosting GAM where the interaction detection is automatic. While this model takes more computational overhead to estimate as compared to other similar ML models, the EBM has the desirable features of extreme compactness and rapid time to produce predictions in execution. The EBM is a GAM of the same form as the GAMI-Net model, but there are some differences in that the constraints imposed above are not imposed in this setting. In this model the GAM link function is adapted to alternative objectives, such as the regression or classification problems, and when introduced it was recognized as an improvement over traditional ML models. Firstly, the EBM learns the functional form for each input variable through standard ML algorithms, for example techniques like gradient boosting or bagging. In the boosting process restrictions are imposed to train features one by one (a "round-robin"), in which the learning rate is very low, implying that the order of features is irrelevant. The algorithm then cycles across the features in order to mitigate collinearity effects while learning the optimal corresponding functional form. This process provides visibility into the manner in which feature

contributes to the prediction of the model. Second of all, the EBM automatically detects and then includes the pairwise interaction terms having the same form, a factor that further augments model accuracy while at the same time enhancing interpretability. Finally, implementation of the EBM is in a parallel manner with respect to the senses of multi-core as well as multi-machine architectures. Since the EBM is additive in input factors, every feature contributes to the prediction modularly, which facilitates comprehension of the relative individual contributions to the prediction. In making individual predictions each functions as a lookup table with respect to each feature, which producing in the process a quantification of the feature's contribution. Such term contributions are then summed and entered into the link function in order to produce the final prediction. This additivity means that term contributions may be sorted and optically represented in order to demonstrate features having the greatest influence on any particular prediction. In order to enforce additivity of the individual terms there is a training penalty imposed upon the EBM, which augments training time as compared to similar techniques. But since the prediction process is a matter of lookups and arithmetic operations inside the GAMs of the feature, EBMs are amongst the most rapid ML models for prediction at the point of model implementation or time of production.

4. Description of modeling data

In this section we describe the data used in developing the models benchmarked in this paper.

The primary source of data is Moody's *Default Risk Service*™ ("DRS") history of credit ratings, a comprehensive source for rating migrations, default and recovery rates spanning regions and industrial sectors. We obtain standardized financial statement line items and market data from the *Compustat™* database. This includes the industry classification system identifiers *Global Industry Classification Standards* ("GICS") and *North American Industry Classification System* ("NAICS"). This data is available starting in 1979 and spans several economic cycles. This database also provides indicators of company default to supplement DRS - namely types of bankruptcy, liquidation as well as default rating grades from the *National Recognized Rating Agency's* ("NSROs") - which are all considered standard industry definitions of default. We further supplement the latter default types in DRS and Compustat™ by a list of company defaults provided by New Generation Research, which is known in the industry as *Bankruptcydata.com*. In the remainder of this paper we will refer to this base dataset as the "Moody's population" or "Moody's rated obligors".

We then apply a series of filters to this Moody's base dataset in order to obtain a population representative of a North American segment of large *commercial & industrial* ("C&I") corporate borrowers having agency ratings. We accomplish this with a combination of NAICS and GICS industry codes, regional indicators and a floor for historical annual net sales amounts. Non-C&I obligors flagged for non-inclusion are defined by the NAICS codes indicating financial firms, commercial real estate, real estate investment trusts, public sector, government, dealer finance and not-for-profit (see Figure 1 below). We perform a similar filter with respect to the GICS classification for education, financials and real estate (see Figure 2 below). Regional filters are applied to choose only obligors based in the U.S. and Canada and only obligors having a maximum historical annual net sales of a minimum $1B. There are further exclusions for cases of missing or invalid GICS or NAICS codes. We consider only observations after the 1st quarter of 1991, for the rationale that market and accounting regulations were rather different prior to 1990's, as well as that the macroeconomic data used in this study are only available starting in 1990.

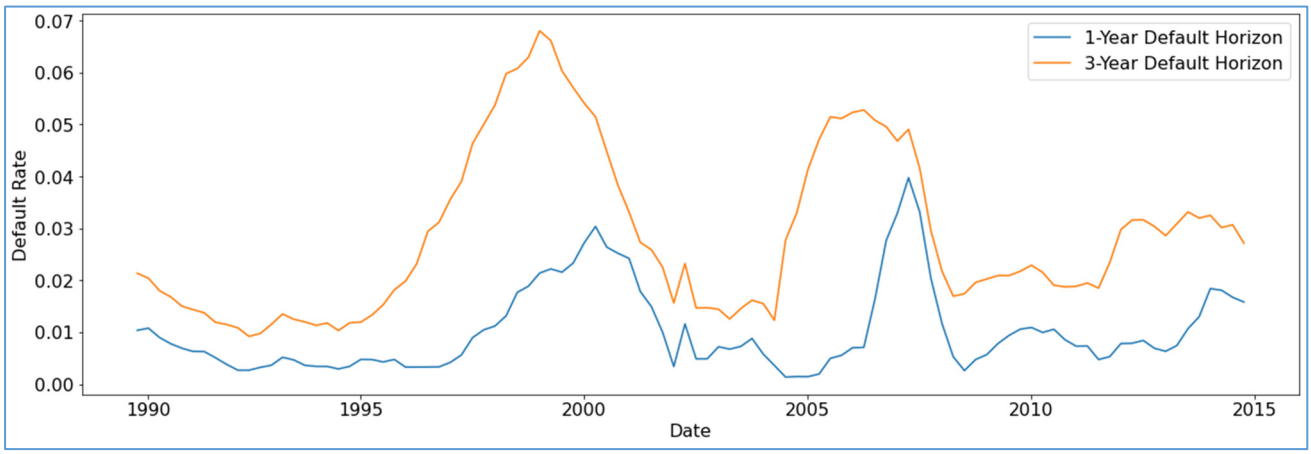


Figure 1. Moody's obligors one-and three-year default rates.

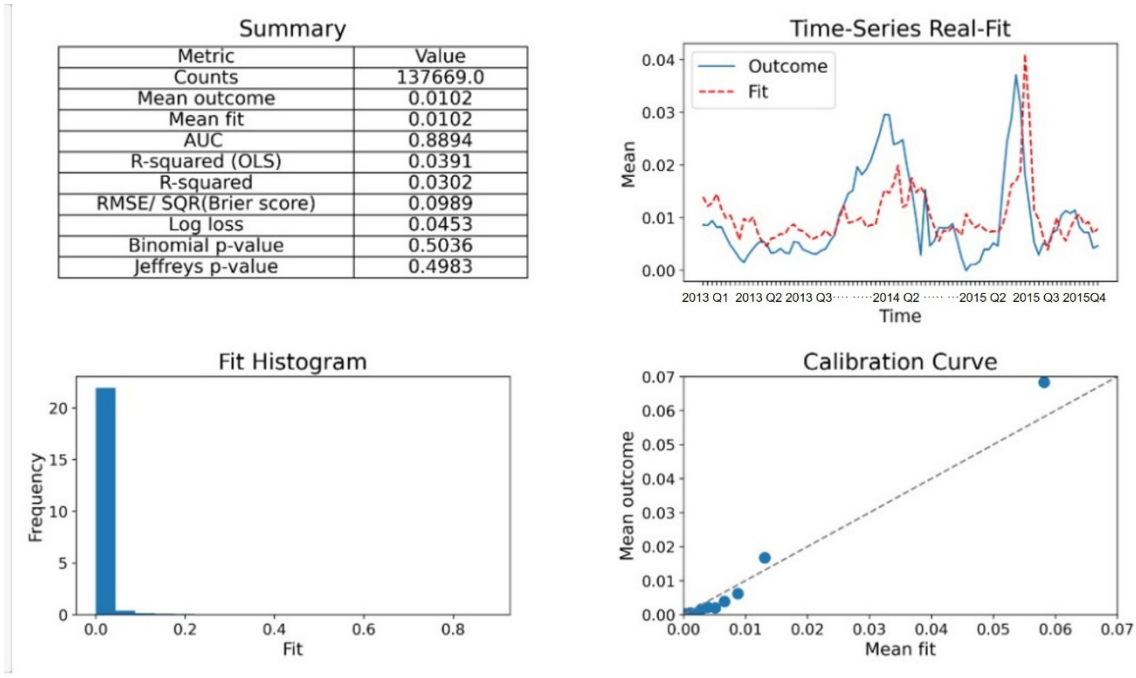


Figure 2. In-Sample performance measures for the estimated logistic regression model.

There are also filters applied that are based upon the default markings. Records considered to be too near to an event of are excluded, an industry standard treatment, where the rationale is the records of a company in such time window do not reflect information about a future default but rather it is more likely existing problems are reflected in such information. We further limit ourselves to data that are in a range of 6–18 (as opposed to 1–12) months prior to default, as this time frame is typically more reflective of period between when financial statements are issued and when credit ratings are refreshed (i.e., this process usually takes up to about six months depending on the time needed to receive complete raw financials, input processed financial ratios and then finalize the ratings). Another point is that while in general we do not consider obligors' financial statements after the default date, there are some cases in which an obligor may exit a default state or "cure" (e.g., emerge from bankruptcy), in which case only the statements between the default date and cure date are excluded.

Table 1. Segment composition by GICS industry classifications for all vs. defaulted Moody's obligors.

GICS Segment	Defaulted Obligors	All Obligors
Consumer Discretionary	30.9%	19.6%
Consumer Staples	6.4%	8.4%
Energy	5.9%	7.6%
Healthcare Equipment and Services	2.9%	2.9%
Industrials	15.1%	31.6%
Materials	11.3%	10.5%
Pharmaceuticals and Biotechnology	0.2%	2.7%
Software and IT Services	1.8%	2.5%
Technology Hardware and Communications	11.3%	4.3%
Utilities	5.6%	7.6%

The time period covered for the model development data extends to the 4th quarter of 2015. In Table 1 above we show the comparison of the modeling population according to the GICS industry sectors. In the case of each sector the defaulted obligors columns represents the proportion of the defaulted obligors comprising the sector in entire population. The observations are concentrated in Consumer Discretionary (20%), Industrials (17%), Technology Hardware and Communications (12%) and Energy except E&P (11%) sectors. We show a similar industry composition according to the NAICS classification system below in Table 2.

Table 2. Segment composition by NAICS industry classifications for all vs. defaulted Moody's obligors.

NAICS Segment	Defaulted Obligors	All Obligors
Agriculture, Forestry, Hunting and Fishing	0.4%	0.2%
Accommodation and Food Services	2.9%	2.3%
Waste Management and Remediation Services	2.1%	2.4%
Arts, Entertainment and Recreation	1.0%	0.7%
Construction	2.5%	1.7%
Educational Services	0.2%	0.1%
Healthcare and Social Assistance	1.6%	1.6%
Information Services	12.1%	11.5%
Management Compensation Enterprises	0.1%	0.1%
Manufacturing	34.4%	37.7%
Mining, Oil and Gas	8.6%	6.8%
Other Services (excluding Public Administration)	0.6%	0.4%
Professional, Scientific and Technological Services	2.5%	2.3%
Real Estate, Rentals and Leasing	1.6%	0.9%
Retail Trade	12.4%	9.6%
Transportation and Warehousing	7.0%	5.4%
Utilities	5.4%	8.3%
Wholesale Trade	2.7%	7.0%

The model development dataset is comprised of financial ratios and default indicators most recently available from in DRS^{TM}, Compustat^{TM} and bankruptcydata.com, and as a result we consider this data to be timely and of favorable quality to support the development of robust models. As the time period for model development of 1Q91–4Q15 spans two economic downturns (i.e., a complete business cycle), the length of the dataset is another factor that supports the good data quality. Related to the latter point, we plot in Figure 1 below annual one- and three- year default rates in the model development dataset, and we can see that trends in the default rates are intuitive as they peak during

the downturn periods. In developing the models in this paper we elect to use the 1-year default rate are the target variable, for reasons to be discussed in more detail in the following section.

In Table 3 below shows summary statistics for the variables appearing in our final models. The final model were chosen based upon an exhaustive search algorithm along with 5-fold cross-validation[1]. Observation counts are around 157K and the one-year default rate is about 1%. Below are the variable categories and variable names of the explanatory variables that appear in the final models[2]:

- ***Size*** Change in Total Assets ("CTA")
- ***Leverage*** Total Liabilities to Total Assets Ratio ("TLTAR")
- ***Coverage*** Cash Use Ratio ("CUR")
- ***Efficiency*** Net Accounts Receivables Days Ratio ("NARDR")
- ***Liquidity*** Net Quick Ratio ("NQR")
- ***Profitability*** Before Tax Profit Margin ("BTPM")
- ***Macroeconomic*** S&P 500 Equity Price Index Quarterly Average Annual Change ("SP500EP"), Consumer Confidence Index ("CCI")

Table 3. Summary statistics of default indicators, financial and macroeconomic explanatory variables appearing in the final models.

Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Default Indicator	157,353	0.01	0.10	0.00	0.00	0.00	0.00	1.00
CTA		0.14	0.35	−0.40	−0.01	0.06	0.17	3.21
TLTAR		0.60	0.23	0.12	0.45	0.59	0.71	1.53
CUR		1.90	2.84	−22.43	1.41	2.06	2.65	19.00
NARDR		130.25	101.44	11.26	68.98	106.74	159.43	754.09
NQR		0.34	1.07	−0.85	−0.28	0.06	0.59	6.11
BTPM		5.94	21.00	−146.67	1.85	7.09	12.85	48.70
SP500EPI		1.91	6.09	−27.33	−0.19	2.19	5.68	12.81
CCI		2.34	21.58	−60.97	−7.02	4.89	15.35	73.21

Missing rates and the *area under the receiver operating characteristic curve* ("AUC") statistics (measuring the power of the variables to distinguish default from non-default on a univariate basis) for the explanatory variables appearing in the final models are summarized below in Table 4[3]. Across risk factors the univariate AUCs lie in a range of around 0.6 to 0.8, which indicates a strong capability to rank order default risk amongst these variables on a univariate basis. As the rate of rate of missing

[1] Clarifying our model selection criteria and process, we balance multiple criteria, both in terms of statistical performance as well as some qualitative considerations. Firstly, all models have to exhibit the stability of factor selection (where the signs on coefficient estimates are constrained to be economically intuitive) and statistical significance in k-fold cross validation sub-sample estimation. However, this is constrained by the requirement that we have only a single financial factor chosen from each category. Then the models that meet these criteria are evaluated according to statistical performance metrics such as AIC and AUC, as well as other considerations such as rating mobility and relative factor weights.

[2] All candidate explanatory variables are Winsorized at either the 10th, 5th or 1st percentile levels at either tail of the sample distribution, in order to mitigate the influence of outliers or contamination in data, according to a customized algorithm that analyzes the gaps between these percentiles and caps / floors where these are maximal.

[3] The plots are omitted for the sake of brevity and are available upon request.

observations lies in a range of 5 and 10% across risk factors, this indicates good data quality that supports development of robust models.

Table 4. Missing Rates and AUCs for Financial and Macroeconomic Explanatory Variables Appearing in the Final Models.

Category	Explanatory Variables	AUC	Missing Rate
Size	CTA	0.726	8.52%
Leverage	TLTAR	0.843	4.65%
Coverage	CUR	0.788	7.94%
Efficiency	NARDR	0.615	8.17%
Liquidity	NQR	0.653	7.71%
Profitability	BTPM	0.827	2.40%
Macroeconomic	SP500EPI	0.603	0.00%
	CCI	0.607	0.00%

5. Estimation results & model benchmarking

Table 5. Estimation results for the logistic regression model.

Explanatory Variable	Parameter Estimate	P-Value	Factor Importance	AIC	AUC	HL P-Value	Mobility Index
CTA	−0.4837	0.0000	0.0455	7,231.00	0.9312	0.5945	0.7184
TLTAR	2.6170	0.0104	0.1091				
CUR	−0.0428	0.0000	0.1545				
NARDR	0.0005	0.0000	0.2273				
NQR	−0.4673	0.0000	0.0909				
BTPM	−0.0161	0.0000	0.2736				
SP500EPI	−0.0189	0.0000	0.0759				
CCI	−0.0099	0.0000	0.0232				

We shall first describe the LRM model estimation results shown below in Table 5. The signs of the coefficient estimates are seen to be consistent with economic intuition and the levels of statistical significance indicate that the parameters are estimated very precisely. The AUC statistics indicate that the model has a strong ability to rank order default risk, in line with favorable performance by industry standards. Regarding a measure of predictive accuracy the p-value of 0.60 in the *Hosmer-Lemeshow* ("HL") tests shows that the model fits the data well. The AIC corroborates the latter as it shows that the model has a favorable fit to the data. The singular value decomposition ("SVD") mobility measure has a value of 0.72, demonstrating the expectation that PD implied ratings from this model are very sensitive to the state of the macroeconomy, as should be the case for a such a class of PD model, as discussed below. The relative sizes of the *factor importance* ("FI") measures, measuring the percent of the model score in totality attributed to a risk factor, are also consistent with a model that is showing the expected characteristics of this type of PD model. The so-called rating philosophy reflected here is of a *point-in-time* ("PIT") model, which predicts default within a relatively short horizon and used in early warning systems, and the factors that are expected to have more importance should span the dimensions of borrower profitability, liquidity or profitability. This contrasts to so-called *through-the-cycle* ("TTC") models, which predict default over a relatively long horizon and are suitable for credit underwriting, and which tend to place more importance on dimensions such as capital structure, size or debt service coverage (Jacobs, 2022b). Accordingly, there is less weights on risk factors more critical to credit underwriting (i.e., factors in the categories Size/Scale, Leverage/Capital Structure and

Debt Service Coverage) that would get higher weight in a TTC model, whereas this trend is reversed and there is greater emphasis on factors considered more critical to early warning or credit portfolio management (i.e., Liquidity, Profitability and Efficiency) in this PIT model. Note that we elect to choose a one- versus three- year prediction horizon, as we wish to construct a PIT model, as we believe this construct to make the most meaningful comparison to ML models, as PIT models tend to have stronger performance than TTC models in terms of predictive accuracy (Jacobs, 2022b).

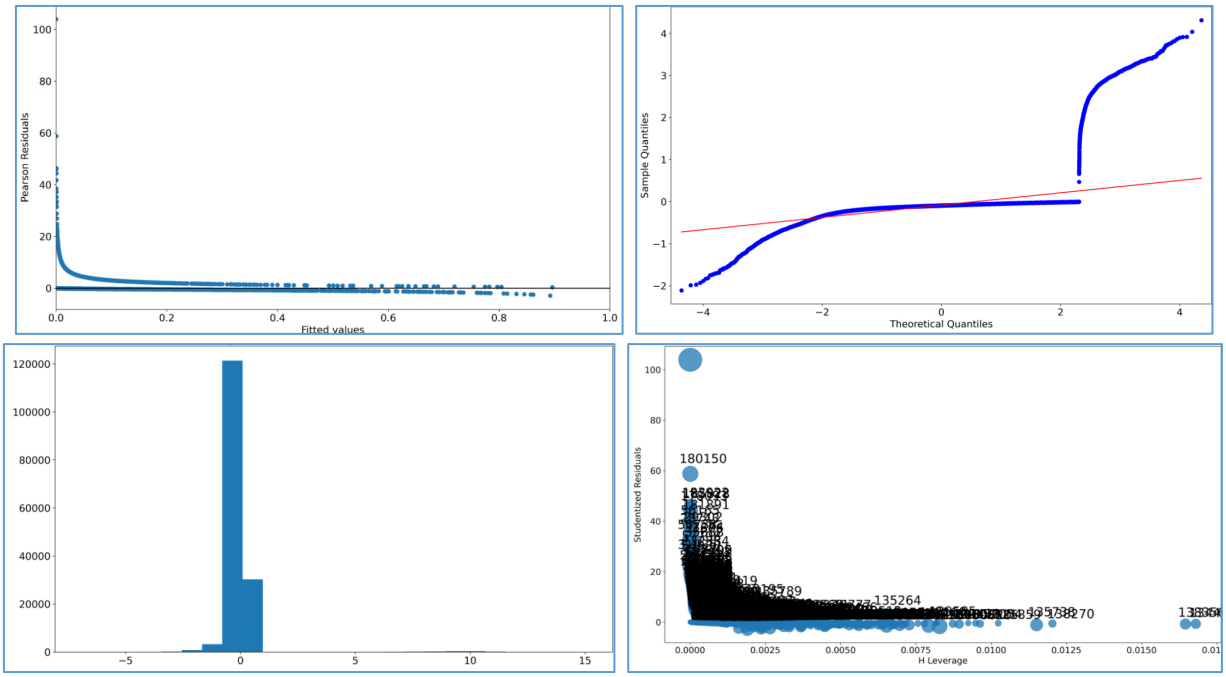


Figure 3. In-Sample residual diagnostic for the estimated logistic regression model.

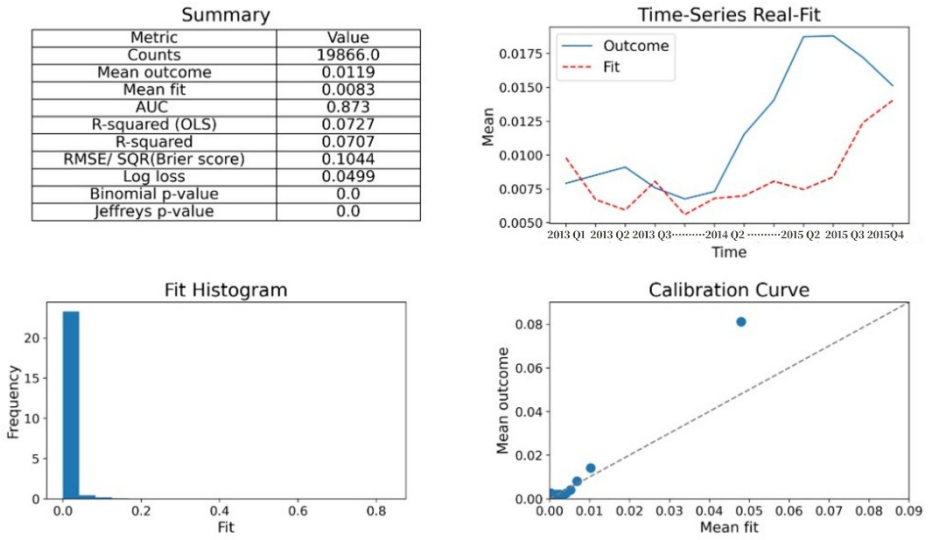


Figure 4. Out-of-Sample performance measures for the estimated logistic regression model.

In the above figures we present additional in-sample and out-of-sample performance statistics and diagnostic plots for the final LRM model. We observe that in-sample that these optical diagnostics (e.g., time series and calibration plots; fit histograms) and additional goodness-of-fit statistics (e.g., binomial and Jeffrey's p-values; OLS r-squareds) confirm the previously discussed results, in that the

final model shows favorable fit to the data. However, other optical residual diagnostic measures (e.g., residual vs. fitted values, quantile-quantile plots, residual histograms and leverage plots) show that the model has some issues in predictive accuracy or model specification. Finally, the out-of-sample analysis shows that the final model performs adequately.

Table 6. ML model benchmarking PD modeling analysis – comparison of AUC and accuracy performance measures.

		Logistic Regression Model	ReLU Activation Deep Neural Network	Generalized Additive Model with Structured Interactions	Explainable Boosting Machine Tree
Training Sample	AUC	0.9312	0.9508	0.9527	0.9870
	Accuracy	0.9935	0.9941	0.9943	0.9960
Testing Sample	AUC	0.9958	0.9677	0.9668	0.9506
	Accuracy	0.9958	0.9958	0.9955	0.9965

We now present the benchmarking analysis of alternative IML models (ReLU-DNN, GAMI-Net and EBM) and the LRM baseline model. This analysis is developed using the Python PiML Toolbox (Sudjianto et al., 2023). In Table 6 above we show the comparison of the AUC and Accuracy (defined as the number of true positives and negatives over the number of observations) discriminatory power performance and measures. We can see that the IML models all demonstrate some pickup in AUC performance, albeit the degree of improvement is modest, especially on an out-of-sample basis, ranging in 2–5% (1–2%) in testing (training) samples. ReLU-DNN (EBM) shows the greatest increase (decrease) out-of-sample of 1.4% (0.4%), while EBM (ReLU-DNN) shows the greatest (least) increase in-sample of 5.6% (1.9%). This suggests that, on this basis, EBM (ReLU-DNN) is most (least) prone to overfitting. On the other hand, according to the Accuracy measure EBM performs best on both an in- and out-of-sample basis, whereas in the training (testing) sample LRM (GAMI-Net) performs worst, which leads to a rather different conclusion than AUC. That said, AUC is a more preferred measure in PD classification applications, so we have more confidence in the conclusions based upon AUC. We depict this analysis graphically in Figures 5 through 8 in the ROC and Precision-Recall plots below.

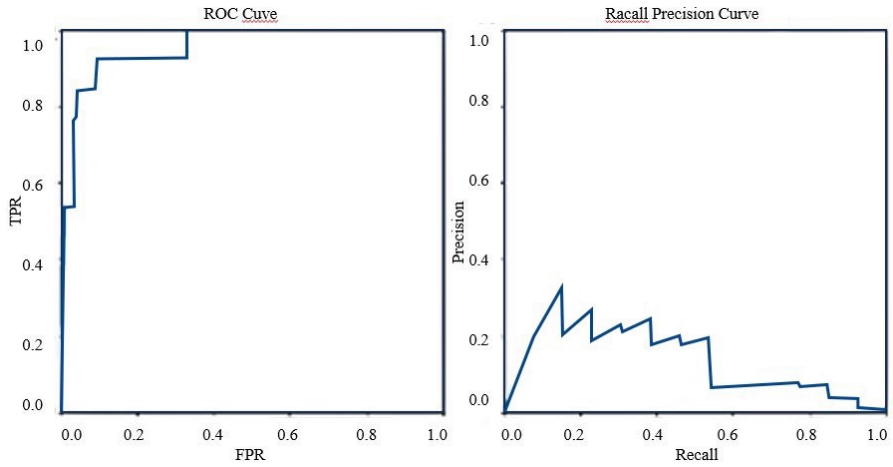


Figure 5. ML model benchmarking PD modeling analysis – comparison of the ROC and Precision-Recall plots for the LRM model.

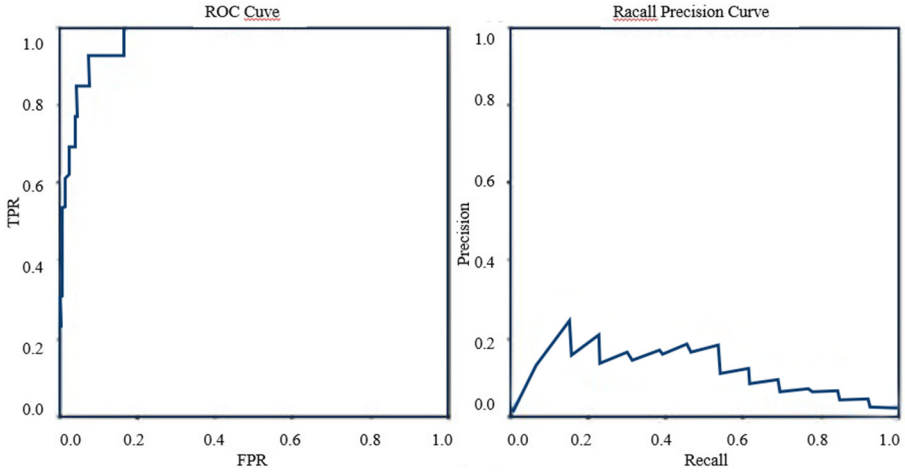


Figure 6. ML model benchmarking PD modeling analysis – comparison of the ROC and Precision-Recall plots for the ReLU-DNN model.

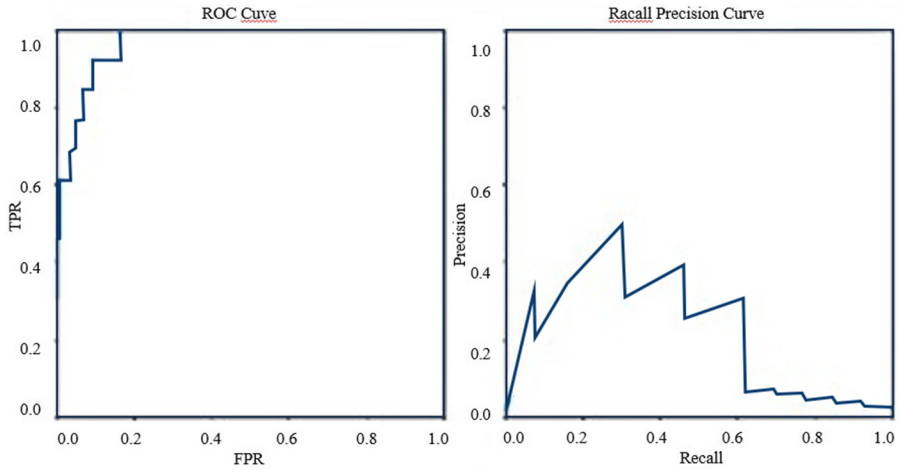


Figure 7. ML model benchmarking PD modeling analysis – comparison of the ROC and Precision-Recall plots for the GAMI-Net model.

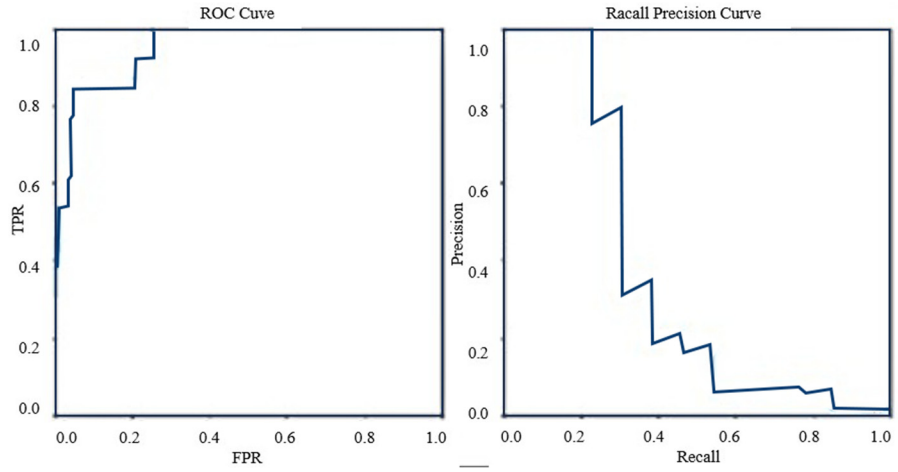


Figure 8. ML model benchmarking PD modeling analysis – comparison of the ROC and precision-recall plots for the EBM model.

We now consider the concept of FI as discussed in the LRM model estimation results previously in more detail. The FI numeric value is interpreted as a score where a higher value indicates more importance of the factor. The concept is similar to that of the magnitude of a regression coefficient in OLS, or a sensitivity measure in other kinds of models, but note that there are various concepts of FI that provide slightly different information. FI is a commonly used as a tool for ML model interpretability, as from the FI scores it is in principle possible to explain why a ML model makes particular predictions and how we can manipulate features to change its predictions. We consider the following types of FIs:

- ***Range FI*** ("RANGE-FI") is commonly used in the industry in LRM applications and serves as our baseline, which is used in the previous estimation and shown in Table 5. The RANGE-FI is a post hoc explanation defined for each risk factors as the coefficient estimate multiplied by the range of the risk factors, scaled by the sum of the ranges of all the risk factors: $RANGE - FI(x_i) \triangleq \frac{\beta_i \times [max(x_i) - min(x_i)]}{\sum_{j=1}^{p} [max(x_j) - min(x_j)]}$, where β_i is the i^{th} coefficient estimate corresponding to each risk factor $x_i : i =, 1, ..., p.$ While this gives us a sense of how influential a factor is as part of fitting the model, this measure is sample dependent as it only gives a particular view of importance, and furthermore does not measure sensitivity to the factor or how the factor contributes to some measure of model fit.
- ***Permutation FI*** ("PERM-FI") post hoc local explanations measure the influence of individual risk factors on the model prediction that calculates the increase in a loss measure (in this context of PD modeling the AUC) when they are permuted. When a factor value is randomly shuffled, the relationship between the feature and the target is broken, and the resulting drop in model performance indicates the feature's significance. However, as different models can have very different FI rankings, this measure only reveals the importance of each feature to that specific model.
- ***Shapley FI*** ("SHAP-FI") additive post hoc local explanations are an ML tool that can explain the output of any model by computing the contribution of each feature to the final prediction based upon coalition game theory. SHAP-FI measures indicate how to justly allocate a "payout" (i.e., a prediction of the model) among the coalition of features as an average marginal contribution of a

feature value across all the possible combinations of features. SHAP-FI possesses several attractive properties, such as local accuracy, missingness and consistency.

- ***Local Interpretable Model Agnostic FI*** ("LIME-FI") post hoc local explanations are a model agnostic explanation tool. The procedure underlying this measure involves creating a surrogate interpretable model, such as a Lasso or decision tree, to explain how the original model makes predictions for a given input sample. The algorithm creates a simulated dataset by randomly perturbing the input sample with noise, evaluates the output of the original model on these perturbed samples and then fits an interpretable model (Lasso in our case) on this simulated data, with weights assigned to each perturbed sample based on its proximity to the original input sample.
- ***Global FI*** ("GLOB-FI") is an inherent explanation that measures the global relative importance of each feature calculated by measuring the variance of the marginal effect $Var[\beta_i \times x_i] : i = ,1, \ldots, p.$ on the training dataset. In the case of categorical features, we aggregate the marginal effects of all of its dummy variables and then calculate the variance. Therefore, the GLOB-FI provides a measure of how much the feature contributes to the overall variability in the model's predictions. In order to interpret the relative importance of each feature as a proportion of the total importance across all features, we normalize the feature importance so that their sum equals 1: $GLOB - FI(x_i) \triangleq \frac{Var[\beta_i \times x_i]}{\sum_{j=1}^{p} Var[\beta_i \times x_i]}$.

We show a comparison of FI measures across the three IML models and the LRM below in Table 7. While the overarching observation is the complete lack of consistency across both models and FI measures, we note that overall the SHAP-FI and LIME-FI measures are least inconsistent across models. We only show RANGE-FI for the LRM as that is our baseline and not computed for the IML models in the PiML package that we use. As we noted previously in the LRM estimation results the top three factors under RANGE-FI are NARDR, BTPM and CUR, which we deem to be a conceptually sound outcome for this type of PIT PD model where the factors that are expected to have more importance should span the dimension reflecting more short term default risk (borrower profitability, liquidity or cash management), in contrast to TTC PD models more suitable for credit underwriting, where the latter tend to place greater importance on longer term dimensions of default risk (i.e., capital structure, size or debt service coverage.) In the case of the other IML models or other FI measures such factors do not consistently show up as having the most importance, and the rank ordering are rather different depending on the combination of model and FI measure.

First focusing on the LRM model, in the case of PERM-FI the most important risk factors are the macroeconomic factors CCI and MDY500EP at 1st and 2nd place, respectively, with NARDR at 3rd and the only risk factor that overlaps with the top three factors in the RANGE-FI ranking. That said, it can be argued that it makes sense for macroeconomic factors to carry more weight in PIT models, but the fact that such factors are ranked near the bottom in RANGE-FI is hard to justify. In the case of SHAP-FI the top three are TLTAR, NQR and CTA in descending order – we see here that two of the factors are more suitable for PIT models, and while NQR is a PIT factor, it does not make the top ranking according to any of the other FI measures for the LRM. However, in LIME-FI the 2nd and 3rd ranked factors are NARDR and CUR, respectively, which are also ranked among the top three under RANGE-FI, but the 1st ranked factor under this measure is TLTR. Finally, for the LRM we get a completely different picture with GLOB-FI where NQR, TLTAR and CTA take the 1st, 2nd and 3rd places, respectively.

Table 7. ML model benchmarking PD modeling analysis – comparison of factor importance measures.

		CTA	TLTAR	CUR	NARDR	NQR	BTPM	SP500EP	CCI
LRM	Range	0.0455	0.1091	0.1545	0.2273	0.0909	0.2736	0.0759	0.0232
	Permutation	5.66E-06	5.10E-06	7.21E-06	1.26E-05	6.94E-06	6.64E-06	5.51E-05	4.81E-04
	Shapley	0.5332	0.5414	0.0166	0.1421	1.4116	0.0542	0.0677	0.4136
	LIME	−0.0533	0.0025	0.0005	0.0012	−0.1178	−0.0125	−0.0007	−0.0137
	Global	0.1621	0.1069	0.0003	0.0069	0.6857	0.0018	0.0008	0.0356
ReLU-DNN	Permutation	9.11E-06	4.90E-06	8.16E-06	3.21E-04	2.41E-04	8.90E-06	9.00E-05	8.00E-05
	Shapley	0.0011	0.0048	0.0001	0.0013	0.0026	0.0009	0.0005	0.0015
	LIME	−0.0457	0.0297	−0.0006	0.0162	−0.0468	−0.0364	0.0038	−0.0018
	Global	0.3618	0.1584	0.0017	0.0235	0.3295	0.1181	0.0028	0.0043
GAMI-Net	Permutation	1.13E-05	2.41E-04	3.21E-04	1.60E-04	8.00E-05	3.53E-04	3.98E-05	4.01E-04
	Shapley	9.50E-04	1.59E-03	6.30E-04	4.80E-04	1.57E-03	1.10E-03	1.50E-04	6.50E-04
	LIME	−0.0497	0.0154	−0.0168	0.0079	−0.0234	0.0008	−0.0519	−0.0049
	Global	0.0949	0.0876	0.2061	0.0165	0.2783	0.0920	0.0084	0.0269
EBM	Permutation	1.44E-03	1.85E-03	1.30E-03	8.41E-04	1.61E-03	1.20E-03	8.00E-05	4.81E-04
	Shapley	6.60E-04	7.50E-04	4.80E-04	2.70E-04	1.23E-03	5.20E-04	2.20E-04	2.10E-04
	LIME	−0.0231	0.0136	−0.0132	0.00476	−0.0154	−0.0265	0.0007	−0.0033
	Global	0.2033	0.1823	0.1273	0.0350	0.2783	0.1170	0.0168	0.0400

Key: (*across each row in the table*)

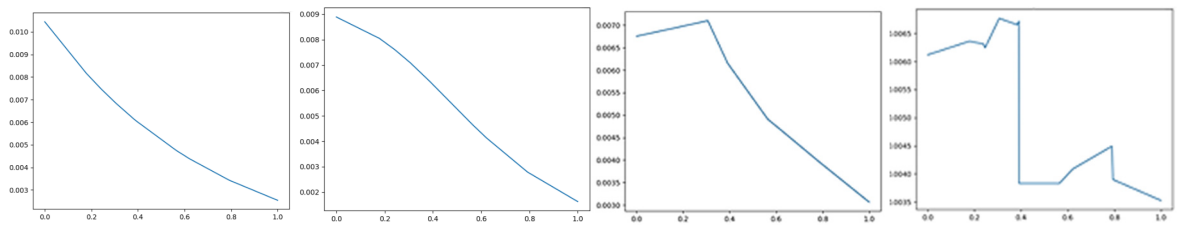
Top			Bottom		
1	2	3	3	2	1

We will conclude our discussion of FI by going over the comparison for the ReLU-DNN model, for the sake of completeness and to not belabor the point about complete inconsistency, as the comparisons for the GAMI-Net and EBM models are similarly incoherent. PERM-FI does feature two PIT risk factors as 1st and 2nd, NARDR and NQR, respectively; with the macroeconomic factors SP500EP and CCI coming in at 3rd and 4th, respectively. In contrast, the top risk factor under SHAP-FI is the TTC risk factor TLTAR, with NQR coming in at 2nd and CCI at 3rd. LIME-FI also ranks TLAR at the top, but now we observe NARDR and SP500EP in the 2nd and 3rd slots, respectively. Finally, GLOB-FI ranks the PIT risk factors NQR and CUR at 1st and 2nd, respectively, with the TTC risk factor CTA coming in at 3rd.

We next consider another mode of interpretability, *partial dependence plots* ("PDP"), a model-agnostic tool that helps visualize the relationship between a subset of risk factors and the predicted dependent variable. This allows us to determine whether this relationship is linear, monotonic or something more complex. If we have a set of risk factors X and a fitted model $\hat{f}$ (which in this context of binary classification for PD modeling this would be the log-odds), then we form a partition into a subset of interest X_S and its complement X_C, and define the PDP function as $PDP_S(x_s) = E_{X_C}\left[\hat{f}(x_s, X_C)\right] = \int \hat{f}(x_s, X_C)p(X_C)\,dX_C$. This integral is approximated by the summation $\frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_s, x_C^{(i)}\right)$, where $x_C^{(i)}$ is the complementary set of risk factors in the i^{th} training sample. We

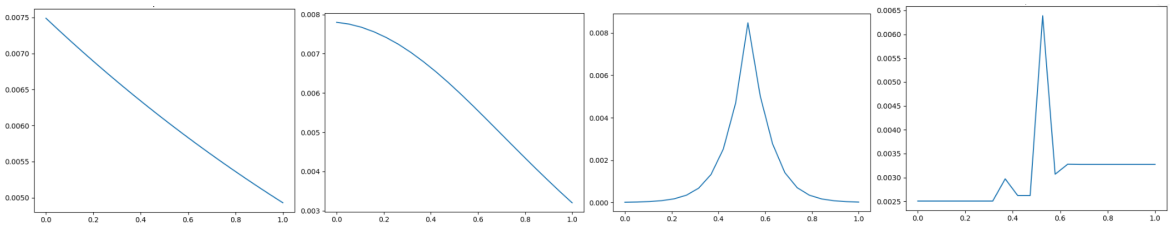
note here that PDPs have several limitations, the most critical being the assumption of independence amongst risk factors. If the risk factors are highly correlated then the outcomes may be inaccurate, as extrapolation of the dependent variable at predictor values required, that the result may lie far beyond the training data's multivariate envelope. However, we have tested for collinearity and have not detected an egregious problem in this regard, so we do not deem this assumption to be a fatal flaw. Also, PDPs may result in inconsistencies between global and local explanations, as PDPs provide an average view of risk factors' influence on the predictions, and local effects specific to certain subsets of data may differ from global ones.

The PDPs are shown below in Figures 9 through 16 for each of the risk factors. While generally speaking the three IML models considered show relationships to default risk that are as expected, as compared to the LRM model there are several instances of non-monotonicities or non-sensical results. In regard to the latter observation, the ReLU-DNN model has results that are most intuitive, and EBM tends to display patterns that are most aberrant, while GAMI-Net is intermediate in this respect. Considering the CCI risk factor in Figure 9, the LRM and ReLU-DNN models show declines as expected, and very similar linear patterns, while the GAMI-Net model shows an increase at low levels of the factor and then linear decrease, while the EBM model has a similar early increase to the latter followed by a discontinuous drop to a flat-line pattern. In the case of the CUR risk factor shown in Figure 10, the LRM and ReLU-DNN models both show similar linear declines, whereas the GAMI-Net model exhibits an inexplicable tent shape, and the EBM model shows a rather odd appearing blip in the middle range of the factor. The situation is rather different for the NQR risk factor in Figure 11 as we observe an agreement across all models of exponential decays. The situation is again unique for the NARDR risk factor in Figure 12 as in contrast to the mildly convex and monotone increase of the LRM, for the other 3 IML models we observe all variations of a hump shape: smooth for the ReLU-DNN model, tent shaped for the GAMI-Net model and a highly irregular kind of step function for the EBM model. Looking at the BTPM risk factor in Figure 13, again we see an agreement between the LRM and ReLU-DNN models in nearly linear decline, whereas both the GAMI.



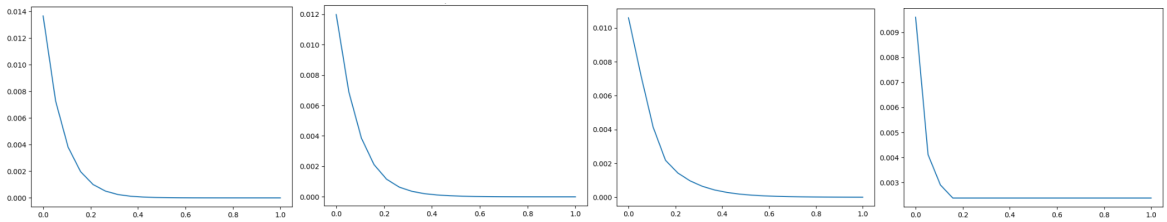
LRM ReLU-DNN GAMI-Net EBM

Figure 9. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the Consumer Confidence Index risk factor.



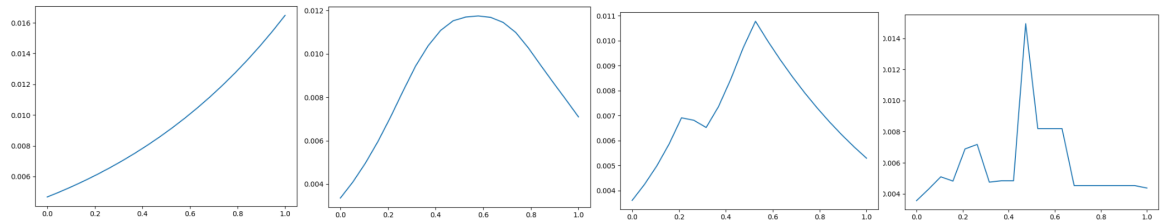
LRM ReLU-DNN GAMI-Net EBM

Figure 10. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the Cash Use Ratio risk factor.



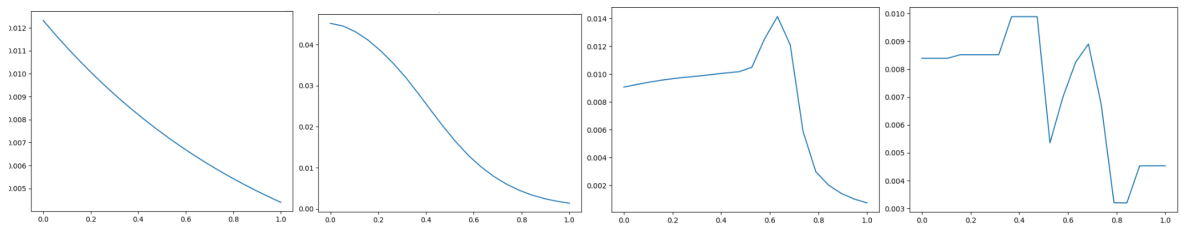
LRM ReLU-DNN GAMI-Net EBM

Figure 11. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the Net Quick Ratio risk factor.



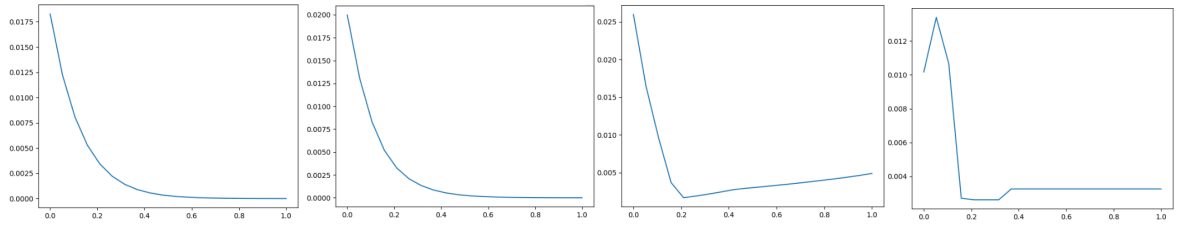
LRM ReLU-DNN GAMI-Net EBM

Figure 12. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the Net Account Receivables Days risk factor.



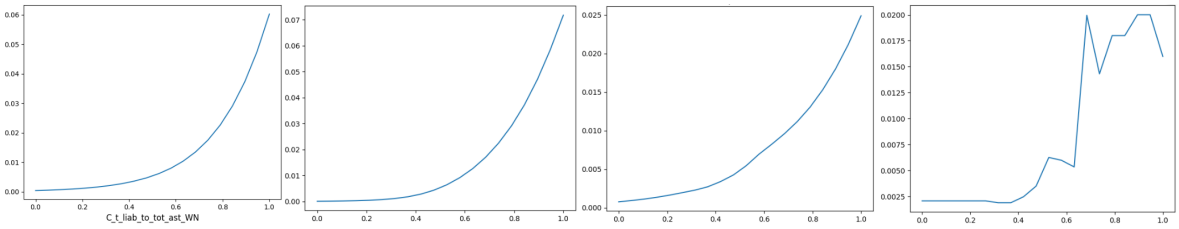
LRM ReLU-DNN GAMI-Net EBM

Figure 13. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the before tax profit margin risk factor.



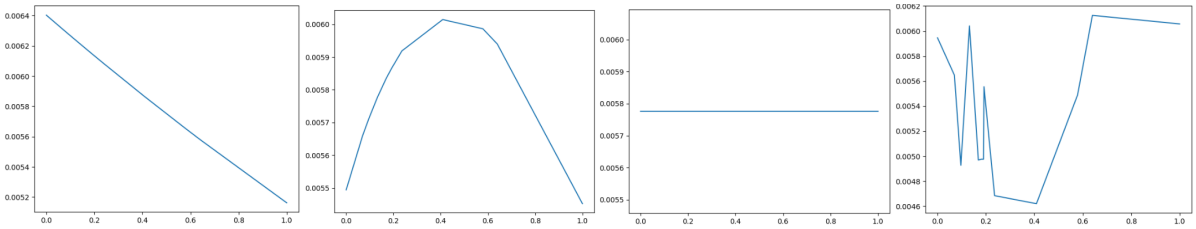
LRM ReLU-DNN GAMI-Net EBM

Figure 14. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the Change in Total Assets risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 15. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the total liabilities to total assets ratio risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 16. ML model benchmarking PD modeling analysis – comparison of partial dependence plots for the S&P 500 Equity Price Index risk factor.

GAMI-Net and EBM models show overall albeit non-monotonic decline, with again the pattern in the EBM model appearing very jagged. The CTA risk factor is shown in Figure 14, and across models there is agreement in version of an intuitive exponential decline. Similarly for the TLTAR risk factor in Figure 15 there is a consistent version of an intuitive exponential increase across models, where again EBM shows the most irregular pattern. Finally, for the SP500EP risk factor in Figure 16 the patterns are the most divergent: the LRM shows a linear decline, the ReLU-DNN model a hump shape, the GAMI-Net model a flat line and the EBM model a pattern that is difficult to succinctly describe.

Given the limitations of the PDPs noted, we consider an alternative that is robust to these assumptions, the *accumulated local effects* ("ALE") plot explainability construct. ALE resolve the issues with PDP calculating differences in predictions instead of averages based on the conditional distribution of the features, showing how model predictions change in a small window of the factor

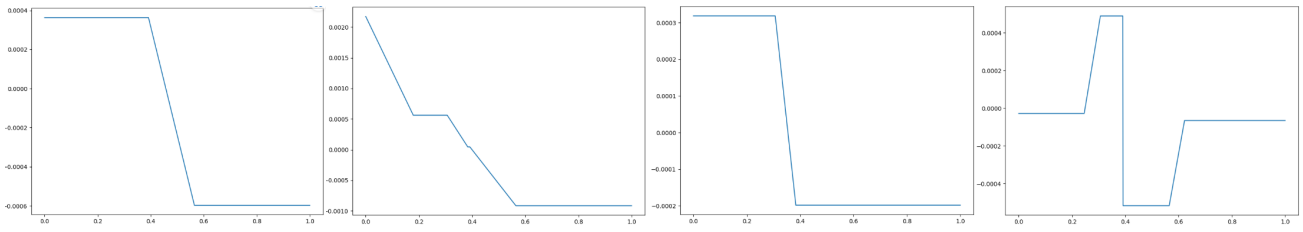
around a defined grid value for data instances in that window. ALE plots average the changes in the predictions and accumulates them over a grid, mathematically represented as

$$\hat{f}_{A,ALE}(x_S) = \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}\left[\hat{f}^S(X_s, X_C|X_S = x_S)\right] dz_S$$
$$= \int_{z_{0,S}}^{x_S} \left(\left(\int_{X_C} \hat{f}^S(z_S, X_C|X_S = x_S)\, d\, P(X_C|X_S = x_S) \right) \right) dz_S$$

where the change is at point $z_{0,S}$is defined as the cross partial derivative $\hat{f}^S(x_s, x_c) = \frac{\partial \hat{f}^S(x_c, x_C)}{\partial x_s}$. This is computed by replacing $z_{0,S}$ with a grid of intervals over which we compute the changes in the prediction, so instead of directly averaging the predictions, the ALE method calculates the prediction differences conditional on the factor set and integrates the derivative over them to estimate the effect.

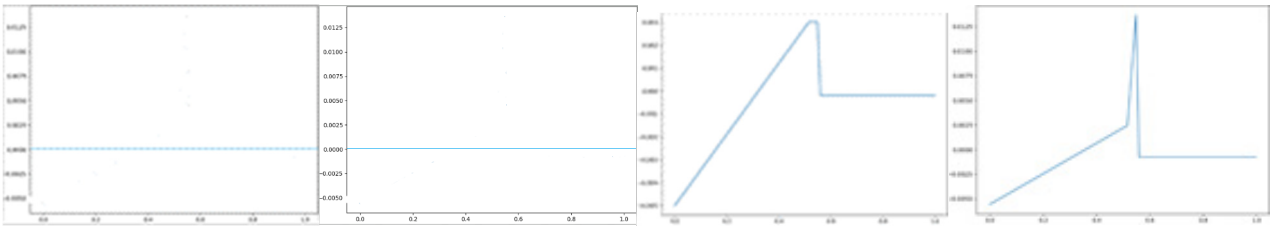
The ALE plots are shown below in Figures 17 through 24 for each of the risk factors. The first and rather obvious conclusion is that the shapes of the ALE plots differ radically from that of the PDP plots, as all of these exhibit extreme non-continuities of variations in step-functions, for example including in many cases "hockey-stick" or L- (reverse L-) shapes. This is by design as the y-axes are not interpreted as average PD estimates at some level of a factor but rather the marginal change in the log-odds for incremental changes in the factor at various levels; that said, the jumps are often hard to rationalize. Furthermore, as with the PDP plots, we are seeing inconsistencies across models and marginal changes that are in counter-intuitive directions. First considering the CCI risk factor in Figure 17, in the LRM there in linear decrease in the 0.4–0.6 region, with flat lines out of this range. In the ReLU-DNN model we get a rotated hockey-stick that approximates linear decrease below 0.6, punctuated by flatness in the 0.2–0.3 region, and flat elsewhere. The GAMI-Net model shows a linear decline in a very narrow range of about 0.38 to 0.40, and is flat elsewhere. Finally for the CCI, the EBM model shows a completely erratic pattern, stepwise increases at high and low levels of the factor in the respective narrow ranges of about 0.20–0.25 and 0.55–0.60, with an almost straight drop in the middle at about 0.40. The CUR risk factor is shown in Figure 18, where in the case of both the LRM and GAMI-Net models the lines are flat at zero, as the ALE scores are identically zero throughout the range of the factor, whereas in the case of the ReLU-DNN model there is linear increase until about 0.60 after which there is a step drop to a flat line at zero, while for the EBM model we see a similar pattern in the rage of up to about 0.50 and then a similar drop to zero. Considering the NQR risk factor in Figure 19 there are steep nearly L-shaped drops to a level of near zero in the narrow lows ranges of around 0.05–0.10 in the LRM, GAMI-Net and EBM models; while the ALE score is zero throughout in the ReLU-DNN model. Turning to the NARDR risk factor in Figure 20, for the LRM the effect is zero until a linear increase at about near 0.40, while for the ReLU-DNN model there is a stepwise increase until about 0.40 followed by linear decline, with a similar pattern in the GAMI-Net model albeit with a single step before around 0.40, and finally for the EBM a single step between flat lines occurring at around 0.05. In the case of the BTPM risk factor in Figure 21 in three models we have overall decrease, a linear increase up to a flat line immediately (after a nearly vertical drop) at around 0.60 for the LRM (both the GAMI-Net and EBM models), and only for the LRM do we get an intuitive overall linear decrease throughout most of the range until at around 0.65 where it flat-lines. In all models for the CTA risk factor shown in Figure 22 we get agreement of L-shaped overall declines, steep linear decline in the low range below around 1.50 for the LRM, ReLU-DNN and GAMI-Net models; while the EBM model shows an almost stepwise decline at about the value 0f 1.80 to a flat line at a little below zero. Similarly for the TLTAR risk factor in Figure 23 there is a consistency of a

hockey-stick overall increase, with the linear rise all starting at just below the level of 0.60. Finally, for the SP500EP risk factor in Figure 24 the patterns are rather divergent, with zero ALE scores throughout in both LRM and GAMI-Net models, whereas we observe an overall hump shape (decline) of a stepwise increase (an erratic decrease) in the lower range up to around 0.40, with linear ascent to a peak at about 0.60 followed by linear decline (to a flat line in the ReLU-DNN (EBM) models.



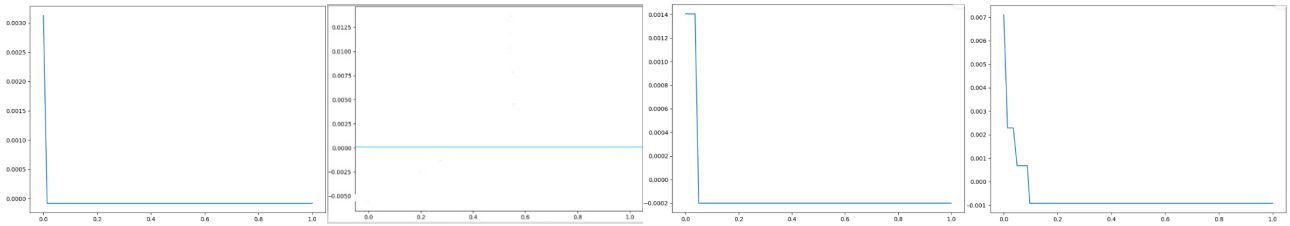
LRM ReLU-DNN GAMI-Net EBM

Figure 17. ML Model Benchmarking PD Modeling Analysis – Comparison of Accumulated Local Effects Plots for the Consumer Confidence Index Risk Factor.



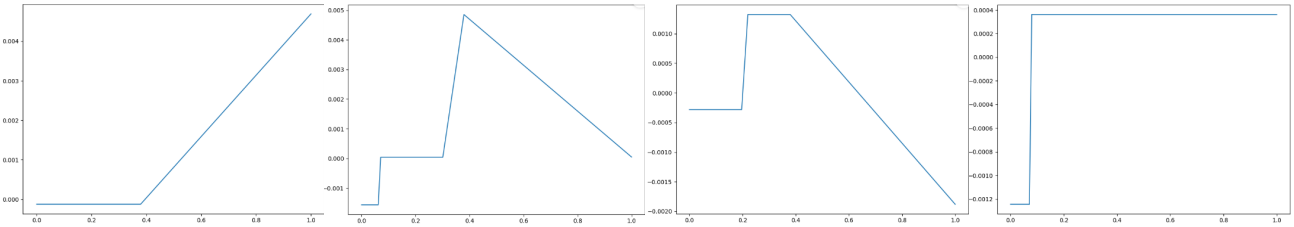
LRM ReLU-DNN GAMI-Net EBM

Figure 18. ML model benchmarking PD modeling analysis – comparison of Accumulated Local Effects Plots for the Cash Use Ratio risk factor.



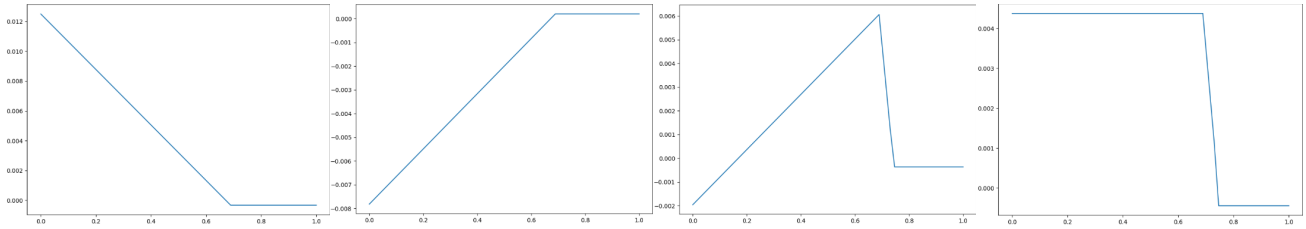
LRM ReLU-DNN GAMI-Net EBM

Figure 19. ML model benchmarking PD modeling analysis – comparison of Accumulated Local Effects Plots for the Net Quick Ratio risk factor.



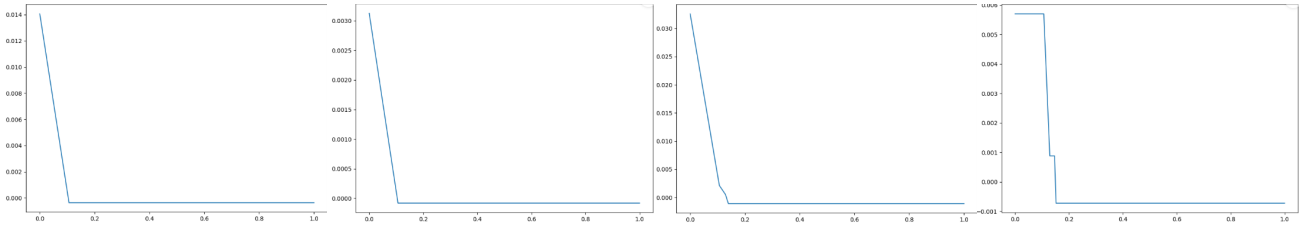
LRM ReLU-DNN GAMI-Net EBM

Figure 20. ML model benchmarking PD modeling analysis – comparison of accumulated local effects plots for the net account receivables days risk factor.



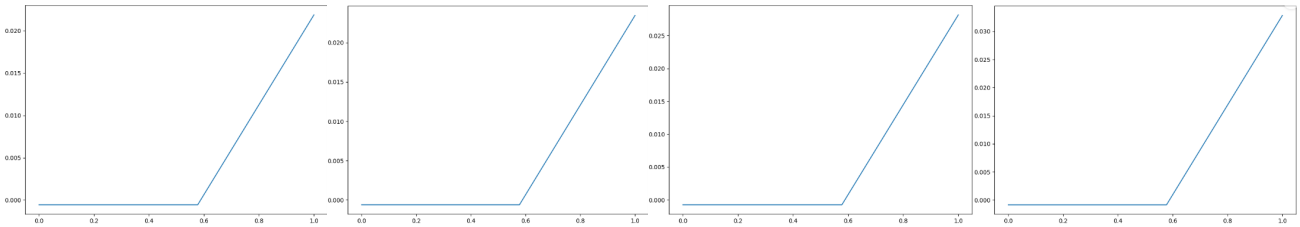
LRM ReLU-DNN GAMI-Net EBM

Figure 21. ML model benchmarking PD modeling analysis – comparison of accumulated local effects plots for the before tax profit margin risk factor.



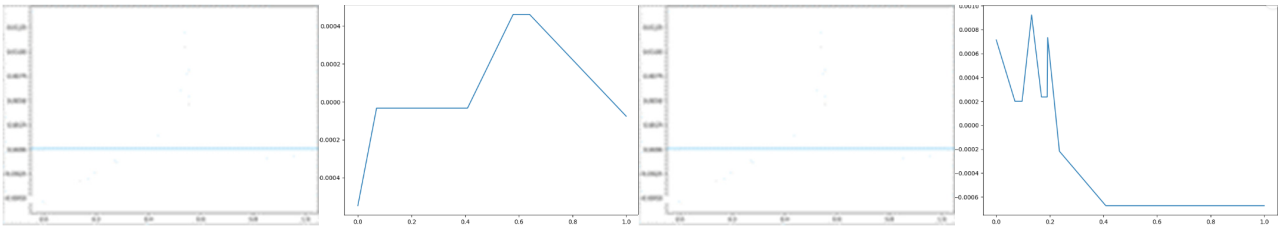
LRM ReLU-DNN GAMI-Net EBM

Figure 22. ML model benchmarking PD modeling analysis – comparison of Accumulated Local Effects Plots for the Change in Total Assets risk factor.



LRM ReLU-DNN GAMI-Net EBM

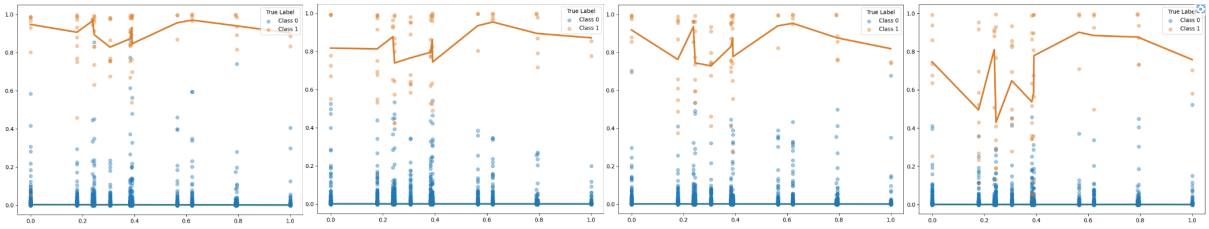
Figure 23. ML model benchmarking PD modeling analysis – comparison of Accumulated Local Effects Plots for the total liabilities to total assets ratio risk factor



LRM ReLU-DNN GAMI-Net EBM

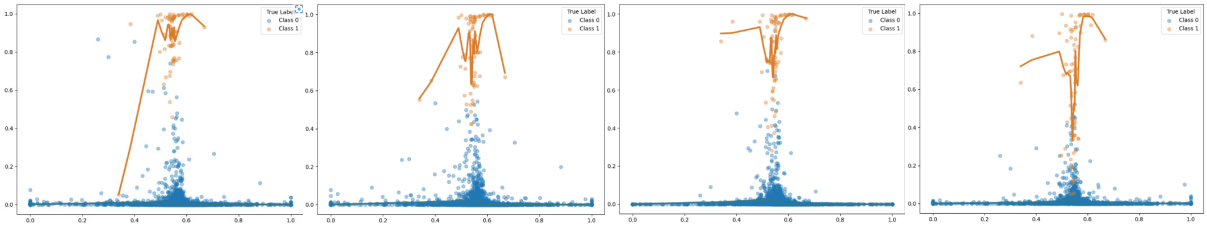
Figure 24. ML model benchmarking PD modeling analysis – comparison of Accumulated Local Effects Plots for the S&P 500 Equity Price Index risk factor.

We will now describe *prediction residuals* between the default indicator dependent variable and the predicted PD estimates against the risk factor explanatory variables. Through an examination of these plots we can analyze patterns and trends in the residuals and assess how they vary with changes in the selected risk factors. As the response variable is binary, we plot the absolute residuals separately for classes of 0 (1) indicating non-default (default), and in addition we include smoothing curves for each class estimated using a *locally weighted scatterplot smoothing* ("LOWESS") estimator. The accuracy plots are shown below in Figures 25 through 32 for each risk factor and model. In general, we observe that, as expected in the case of a relatively low-default setting (or said in the language of statistics, an *unbalanced panel* where the target rate is much lower than the non-incidence rate), the errors are much larger (much smaller) and closer to one for default (non-default). We also note that there are some differences in the patterns across models and risk factors that are in some cases hard to rationalize. First considering the CCI risk factor in Figure 25, in three of the models (LRM, ReLU-DNN and GAMI-Net) we see a very similar pattern of a saw-tooth curve across most of the range of the factor very near one (a flat line very near zero) for default (non-default), whereas this pattern differs in EMB in that the trend line for default rather is lower than one in the lower range of the factor. The residual plot patterns are rather similar across all models for the CUR risk factor shown in Figure 26, where the errors all have a non-monotonic V-shaped spike for the default class in the middle range of around 0.50 to 0.70. The residual plot patterns are also rather similar across all models for the NQR risk factor shown in Figure 27, where the errors all have a non-monotonic jagged series of V-shaped spikes that resolve into a flat line for the default class in the low range of around below 0.30. The commonality across models for the NARD risk factor in Figure 28 is that the residual trend line for the default class spans the entire range, but there are some differences in the patterns, with all showing an early high and jagged pattern, followed by almost linear declines (V-shapes) at higher levels of range for both LRM and ReLU-DNN (GAMI-Net and EBM). The BTPM is shown in Figure 29, and for three of the models (LRM, ReLU-DNN and GAMI-Net) we observe a similar overall increase in the default class trend like of the residual across the entire range from about a level of 0.80 to very near 1.00, punctuated by a V-shaped dip in the middle of the range, whereas EBM differs that this V plunges to a much lower level and at the higher part of the range there is another steep drop to a low level. The residual plots for CTA risk factor are shown in Figure 30 where we observe divergent patterns vis a vis the other factors, in three cases (ReLU-DNN, GAMI-Net and EBM) overall non-monotonic lack of increase or decrease at low levels of the factor, whereas for the LRM there is an overall linear decline over most of the range except for the low region as well as a slight linear trend in the non-default class. In the case of the TLTAR risk factor shown in Figure 31, unlike the other cases in all case the default class residual trend is on the whole decreasing monotonically over most of the range (with LRM closest to linear, ReLU-DNN and GAMI-Net having V- shapes in the middle range and EBM showing a pronounced U-Shape in the middle range), as well as a mild linear increase to still a low level in the non-default trend over all the range. Finally, we consider the SP500EPI risk factor in Figure 32 and observe across all models that there is a non-monotonic lack of change over the entire range in the trend line for the default class, where in the case of EBM a reverse V-shaped dip is the most pronounced as compared to the other models.



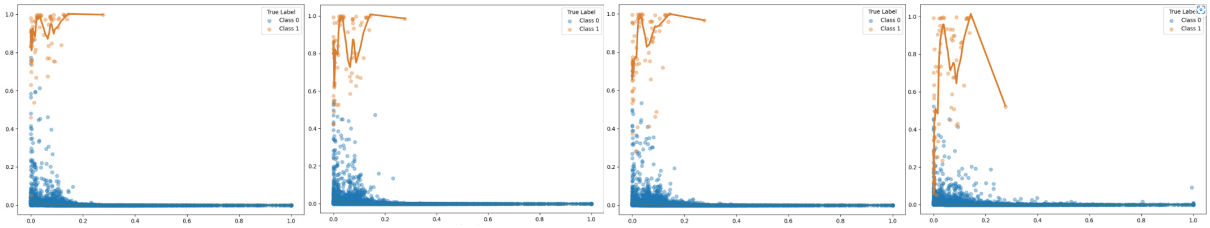
LRM ReLU-DNN GAMI-Net EBM

Figure 25. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the consumer confidence index risk factor.



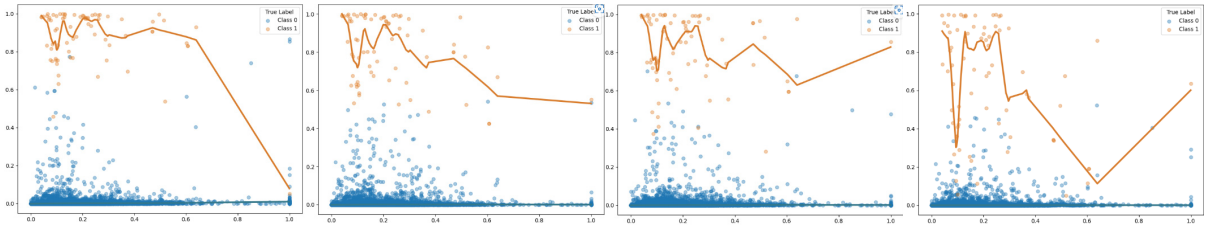
LRM ReLU-DNN GAMI-Net EBM

Figure 26. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the Cash Use Ratio risk factor.



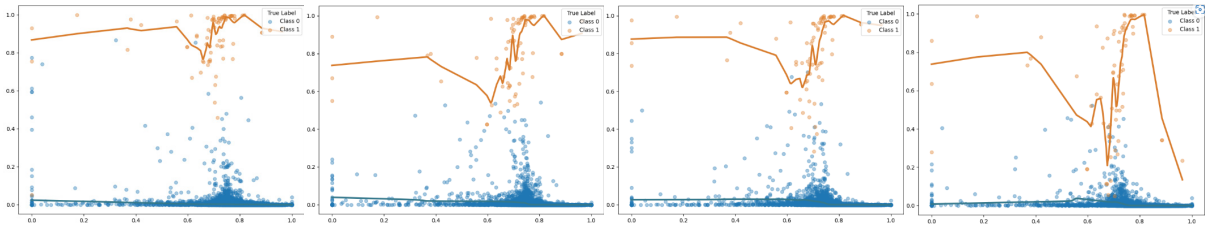
LRM ReLU-DNN GAMI-Net EBM

Figure 27. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the Net Quick Ratio risk factor.



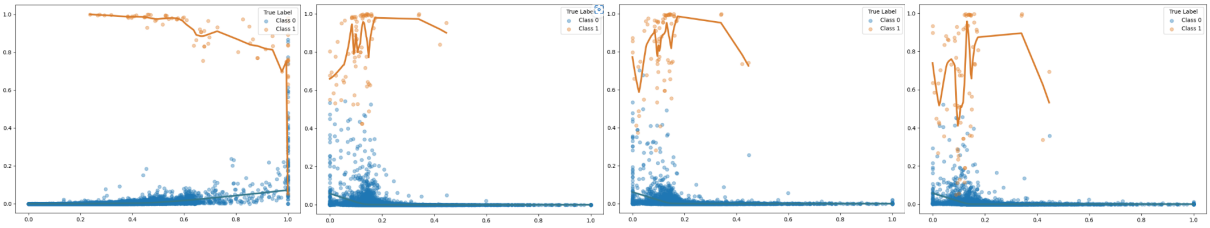
LRM ReLU-DNN GAMI-Net EBM

Figure 28. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the Net Account Receivables Days risk factor.



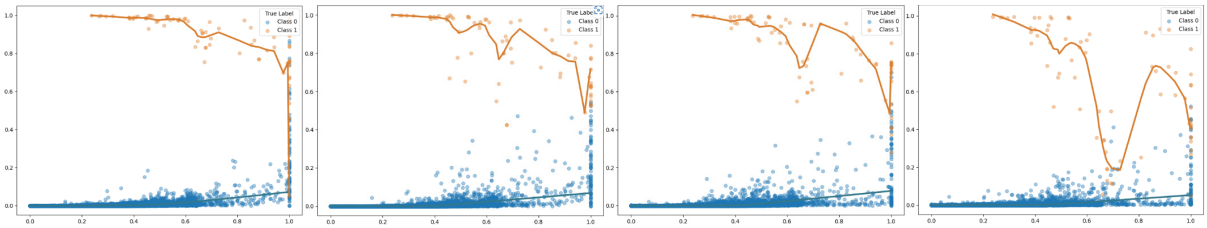
LRM ReLU-DNN GAMI-Net EBM

Figure 29. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the before tax profit margin risk factor.



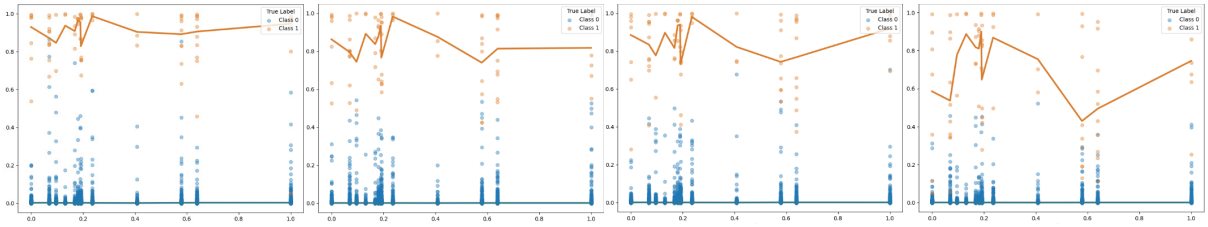
LRM ReLU-DNN GAMI-Net EBM

Figure 30. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the Change in Total Assets risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 31. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the total liabilities to total assets ratio risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 32. ML model benchmarking PD modeling analysis – comparison of prediction accuracy plots for the S&P 500 Equity Price Index risk factor.

We now turn to an analysis of performance in terms of *robustness.* The performance of a model can be adversely affected when it encounters noisy data or experiences distribution shifts. Such data drift or shift may occur due to unexpected changes, which can alter the underlying patterns and relationships between the input and target variables. The robustness test assesses model performance through subjecting it to small changes in the covariate space by perturbing a covariate x with changes Δx, calculating the model's output $\hat{f}(x+\Delta x)$ on the perturbed data and then evaluating the performance metric of $Score\left(y, \hat{f}(x+\Delta x)\right)$ according to the AUC. This process is iterated multiple times (we choose ten) for all the test samples, with performance metrics recorded for each repetition. It is important to note that the assumption is made that the response remains unchanged throughout. In the case of numerical features as in this setting, there are two perturbation options. *Raw perturbation* directly adds i.i.d. Gaussian noise $N\left(0, \lambda^2 \operatorname{Var}(x)\right)$ to x, where λ is the *perturbation size*. However, this method may not be suitable when the data is skewed and has a long tailed distribution (as is the case with our risk factors), in which case the calculation of the standard deviation may become unstable, and it is relatively hard to choose a suitable perturbation size. Therefore, we choose to perform *quantile perturbation* in order to resolve this issue, which is implemented as follows. First, the feature is converted to the quantile space, and then a uniform noise $U(-0.50\lambda, 0.50\lambda)$ is added to perturb the quantiles, where λ also represents the perturbation size. Finally, we transform the quantiles back to the original space. Note that we may perform this analysis on all the covariates or on subsets of the covariates.

In Figure 33 below we show the robustness results where we perturb all the risk factors, where we note that in the key that "GLM_5" is the LRM model, and for each perturbation with sizes ranging from 0.10 to 0.40 in increments of 0.10 we show the box plots of the AUC for each model. We observe that as expected AUC deteriorates as we increase the perturbation size, and the ReLU-DNN and LRM models appear to hold up best as the variance of the AUC measure appears most stable with increasing size, while the GAMI-Net and EBM models show worse performance.

In Figures 34 through 41 below we present the robustness plots for each risk factor separately, from which we conclude that for a given variable there is some material variation in robustness across models for some risk factors, and in general while the dispersion of the AUC consistently becomes greater with the shock the means of the distributions do not get materially worse. Also, in general the LRM, ReLU-DNN and GAMI-Net models are most consistent in patterns across risk factors, while the EBM is least consistent in this regard and in some cases shows counterintuitive patterns.

In the case of the CCI risk factor shown in Figure 34, the degradation is greater and similar in the LRM and ReLU-DNN models with more downside, while the GAMI-Net and EBM models appear to be more similar, with less degradation as well as more asymmetry. In the case of the CUR risk factor shown in Figure 35, the degradation is greater and similar in the LRM and GAMI-Net models with more downside, while the ReLU-DNN model appears to show less degradation as well as asymmetry, and the EBM model actually shows an improvement at the greatest perturbation shock. The NQR risk factor is shown in Figure 36, and now we see that the EBM model is very resilient to shocks, whereas the LRM has the greatest drop in mean performance (although the variance of the distributions are rather stable), while the ReLU-DNN and GAMI-Net models are intermediate in robustness (although the variance of the later distribution is less stable as compared to the former). The NARD risk factor shown in Figure 37 has the mean AUC getting better or no worse with shock size, but in terms of increasing variability performance is worse in the LRM and GAMI-Net models, as compared to the

ReLU-DNN and EBM models with the latter actually becoming less dispersed at the largest perturbation size.

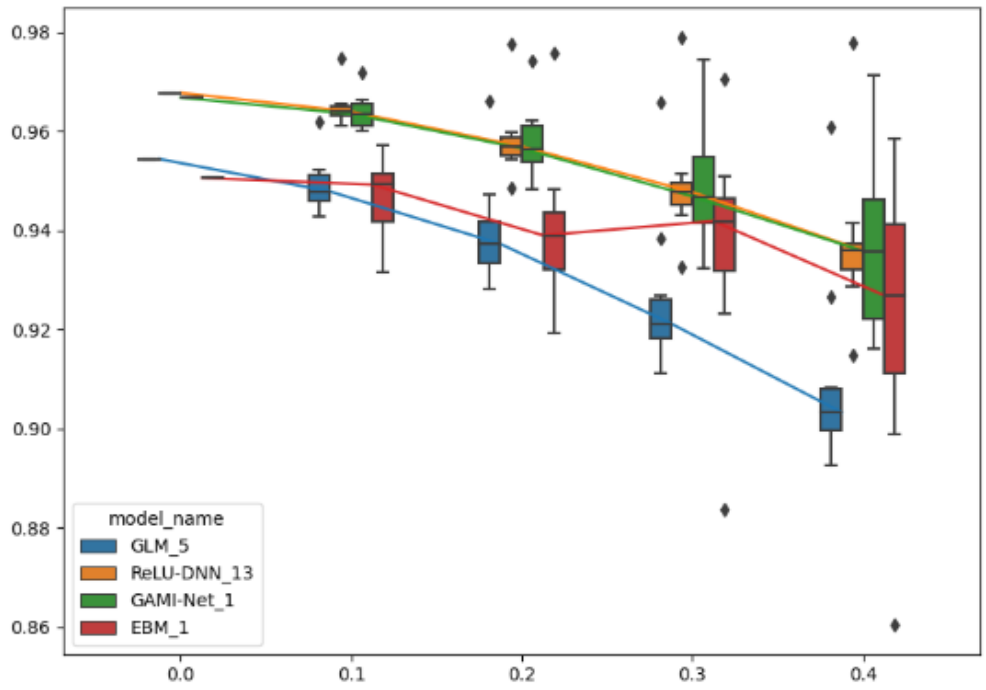


Figure 33. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for all risk factors.

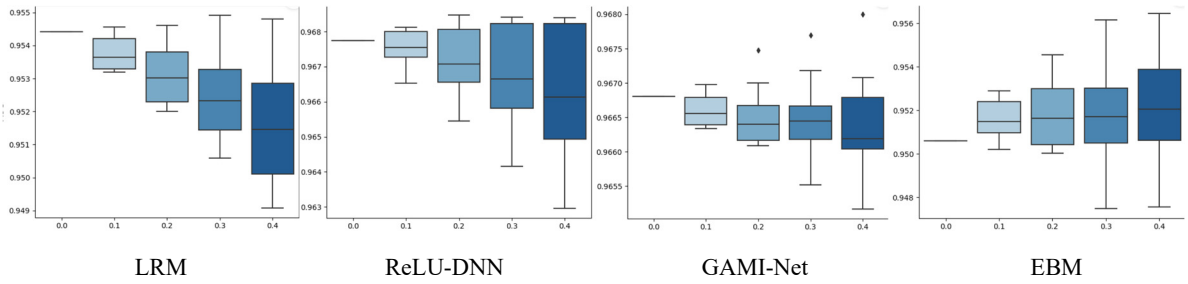


Figure 34. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the Consumer Confidence Index risk factor.

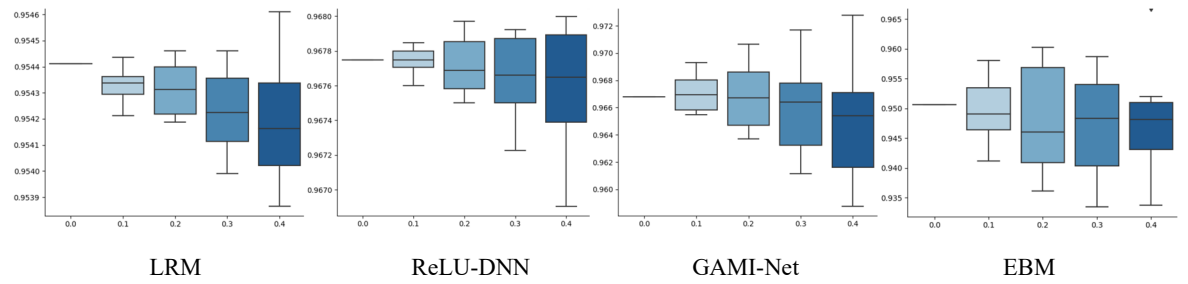


Figure 35. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the Cash Use Ratio risk factor.

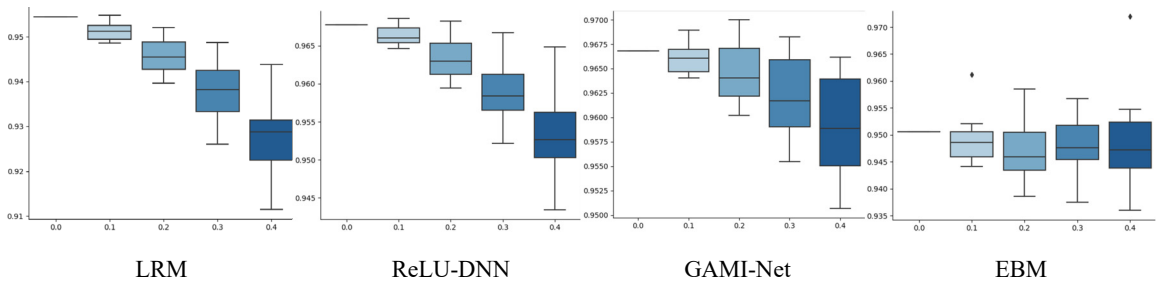


Figure 36. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the Net Quick Ratio risk factor.

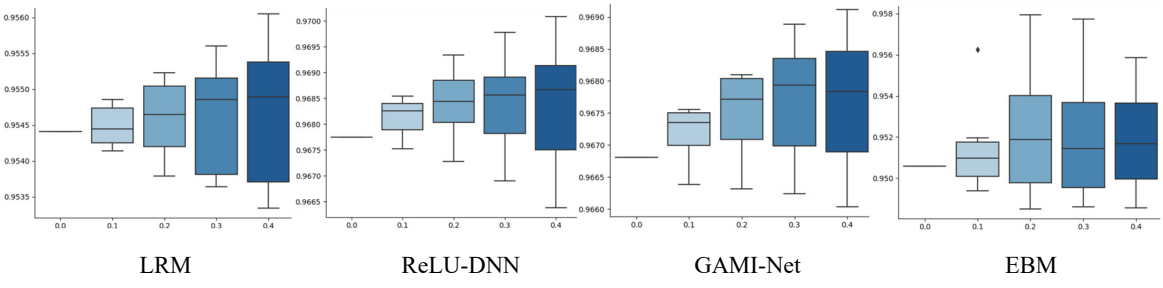


Figure 37. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the Net Account Receivables Days risk factor.

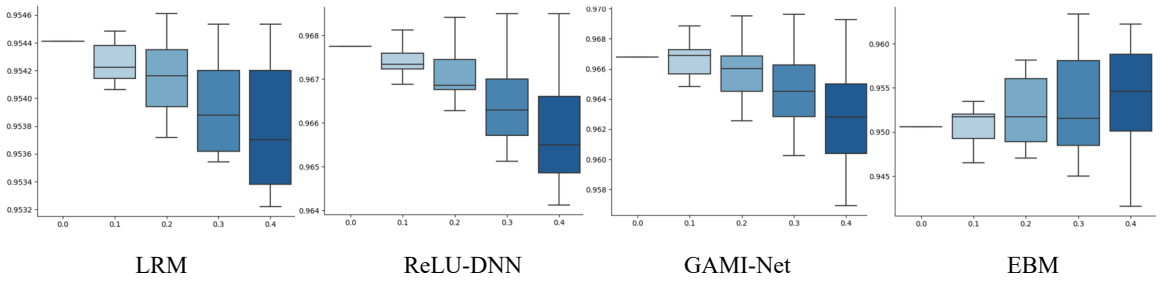


Figure 38. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the before tax profit margin risk factor.

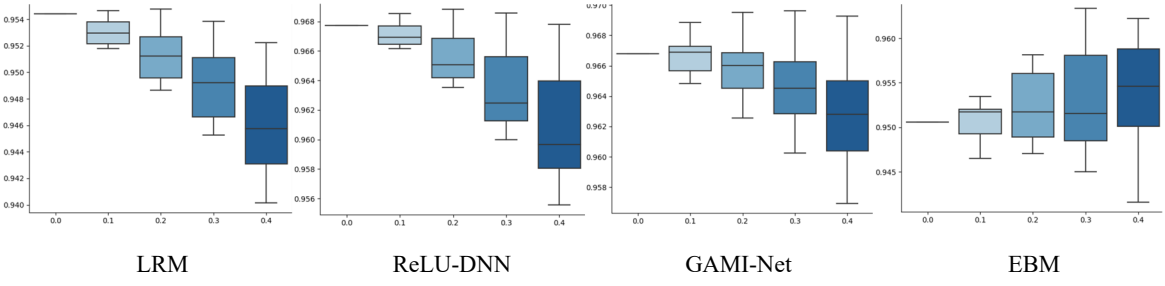


Figure 39. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the Change in Total Assets risk factor.

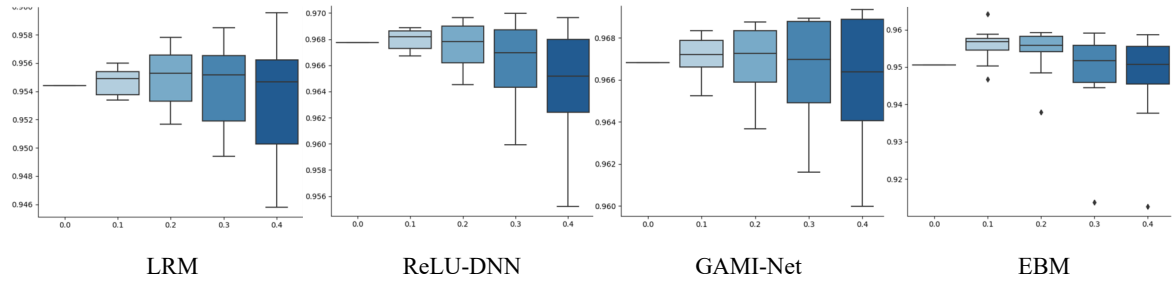


Figure 40. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the total liabilities to total assets ratio risk factor.

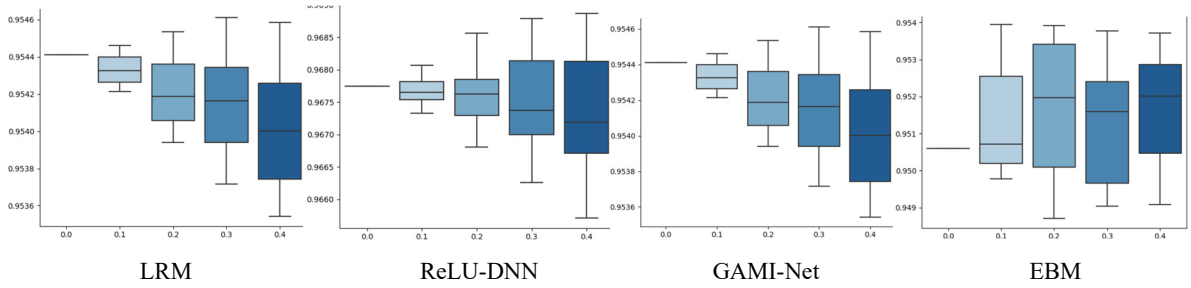


Figure 41. ML model benchmarking PD modeling analysis – comparison of Robustness Plots for the S&P 500 Equity Price Index risk factor.

The BTPM risk factor is shown in Figure 38, where the LRM, ReLU-DNN and GAMI-Net models have similar patterns of declining (increasing) mean (dispersion) of the AUC distribution, in order of improving overall performance; while the EBM model has a non-decreasing mean AUC and widening dispersion until the largest shock. In the case of the CTA risk factor shown in Figure 39, the patterns of decreasing mean and increasing dispersion with shock size are similar for the LRM, ReLU-DNN and EBM models; whereas again we see an odd pattern in the EBM model, a stable or even improving mean coupled with rising dispersion as the shock increases. We see a somewhat similar pattern as the former case for the TLTAR risk factor shown in Figure 40, the patterns of decreasing mean and increasing dispersion with shock size are similar for the LRM, ReLU-DNN and EBM models; but this time we see a slightly different weird pattern in the EBM model, a stable or even improving mean (but this time for intermediate values of the shock and not the maximal value) coupled with rising dispersion as the shock increases. Finally, for the SP500EPI risk factor shown in Figure 41, yet again the pattern across models similar to the last two cases, in terms of the similarity of the first three models in the same way, as well as an anomalous pattern in the EBM model of a non-monotonically increase overall in the mean now paired with almost no increase in the dispersion.

We will now discuss weak region overfitting analysis for each risk factor. In this technique, we evaluate the performance metric of choice (in this case the accuracy gap – "ACC") for each sample (testing and training) as *pseudo-responses*, and then we use a decision tree (or tree ensemble) technique to segment the range of the variable of interest into sub-regions. In each sub-region we then identify those where the ACC exceeds a pre-specified threshold according to a *minimum sample condition*. In the overfit plots we identify these weak regions as *positive bars* as we take the negative of the ACC deviations.

The results of the weak region overfit analysis are shown below in Figures 42 through 49. In general, there is a moderate degree of agreement amongst models as to which regions are weak, with a material amount of disagreement. In the case of the CCI risk factor shown in Figure 42, all models show the regions around 0.10–0.20 and 0.30–0.40 to be most weak, with EBM showing the least deviation in ACC among them; as well as the regions around 0.80–0.85 and 0.95–1.00, although the degree of weakness here is less. The CUR risk factor is shown in Figure 43, where we see identified weak regions of around 0.25–0.30 in the LRM, 0.30–0.35 in all models except the LRM and 0.50–0.55 in all models. In Figure 44 for the NQR risk factor the only region identified as weak is around 0.00–0.05 for all models, in the order of severity: the LRM, ReLU-DNN, GAMI-Net and EBM model. In the case of the NARD risk factor shown in Figure 45, the LRM and ReLU-DNN models find the regions around 0.60–0.65 to be weak, whereas this is the case and to a greater degree of severity for only the LRM around 0.85–0.90. In the case of the BTPM risk factor shown in Figure 46 we observe some regions of agreement as well as disagreement across models in terms of weakness, with all models identifying narrow regions around just below 0.20, just above 0.40 and just above 0.60 as such; while only the LRM identifies a narrow region around just above 0.30 similarly. Considering the CTA risk factor in Figure 47 we observe complete agreement across models of regions identified as weak, around below 0.10, 0.30–0.35 and 0.40–0.45. The story is the same as the latter for the TLTAR risk factor shown in Figure 48, where almost all models to varying degrees shows the region of around greater than 0.85 as weak, with the exception of the EBM model at around greater than 0.95. Finally, we see another case of almost near agreement amongst models for the SP500EPI risk factor in Figure 49, where almost all models identifies the region around 0.40–0.70 as weak, with the exception of the EBM model for the subset of that around 0.55–0.60.

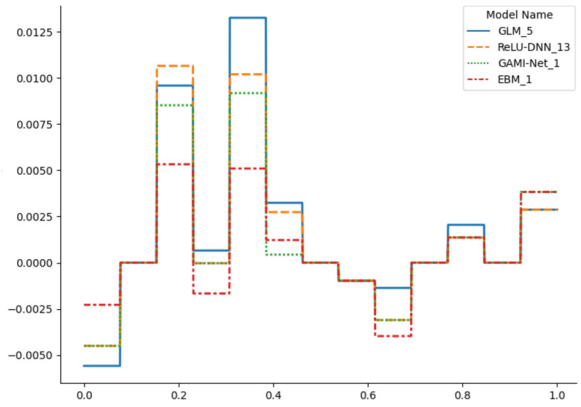


Figure 42. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the Consumer Confidence Index risk factor.

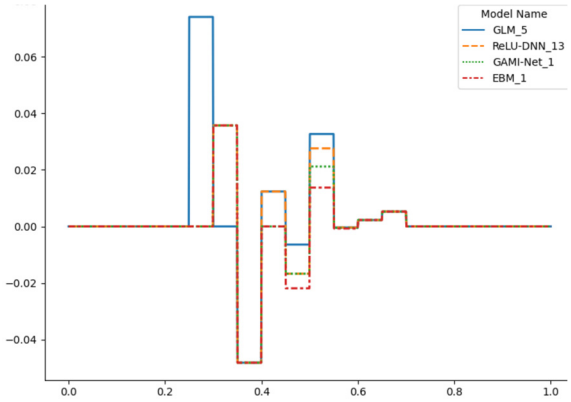


Figure 43. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the Cash Use Ratio risk factor.

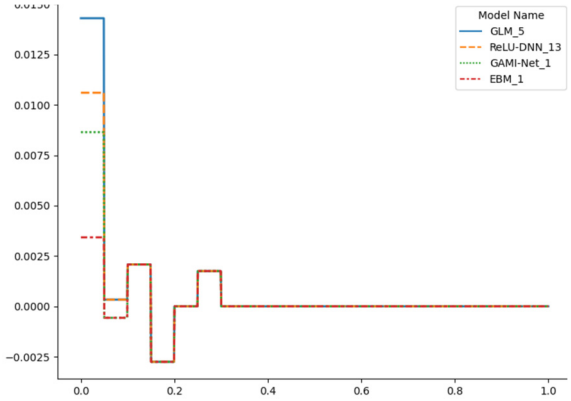


Figure 44. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the Net Quick Ratio risk factor.

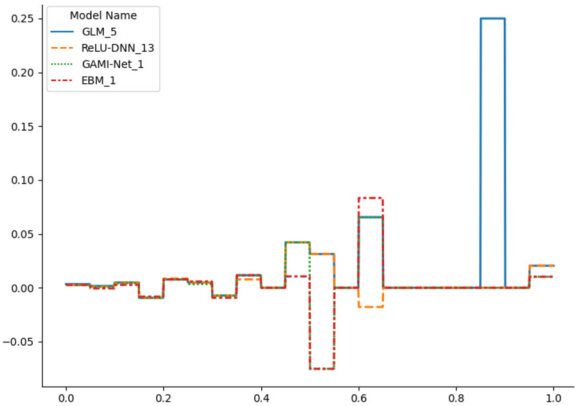


Figure 45. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the Net Account Receivables Days risk factor.

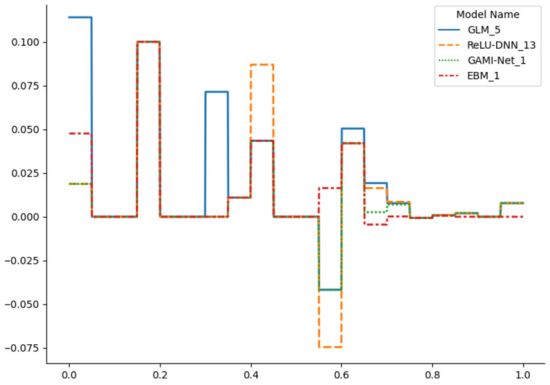


Figure 46. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the before tax profit margin risk factor.

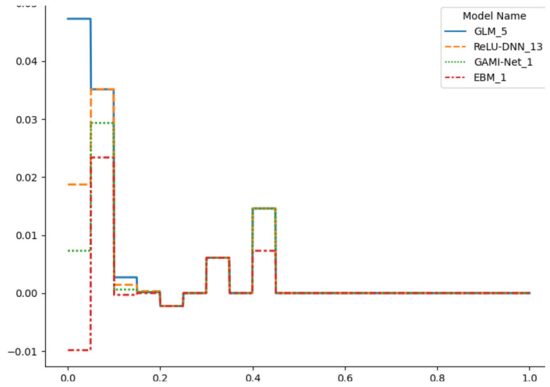


Figure 47. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the Change in Total Assets risk factor.

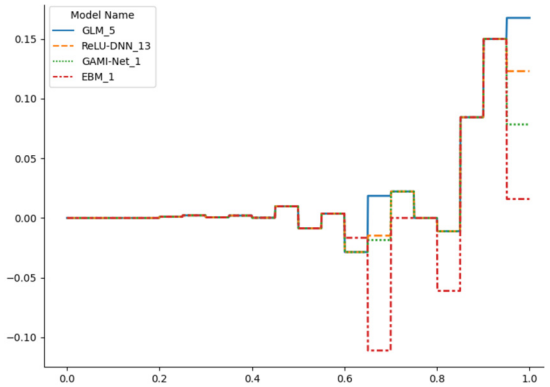


Figure 48. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the total liabilities to total assets ratio risk factor.

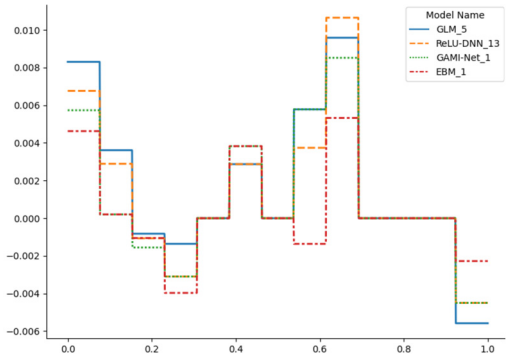


Figure 49. ML model benchmarking PD modeling analysis – comparison of weak region overfitting analysis plots of the ACC deviation for the S&P 500 Equity Price Index risk factor.

We now will discuss *resiliency distance analysis* (or *weak region over-fitting analysis*), where we assess the performance of the worst samples in a model across various out-of-distribution scenarios, according to the distributional distance between the worst test sample and the full test sample that is calculated for each risk factor that are then ranked by distance. We choose the PSI distributional distance measure and set the worst sample proportion to 10%. The resilience distribution analysis plot is shown below in Figure 50. We observe that there is general agreement across models in which risk factors have the greatest shifts in distribution, with all models identifying the TLTAR, NQR, BTPF and CTA features as being top four in terms of shift.

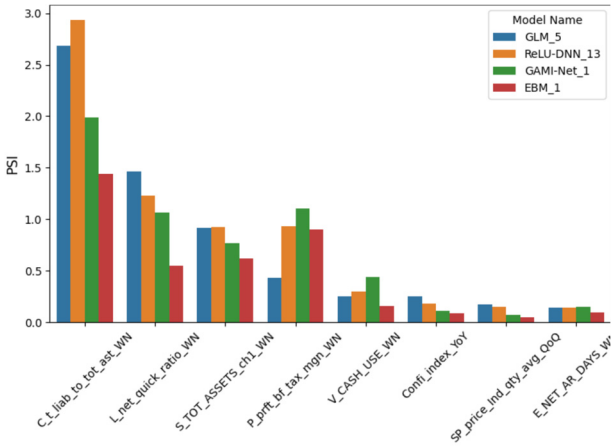


Figure 50. ML model benchmarking PD modeling analysis – comparison of resiliency distance analysis plots of the 10% worst samples for all risk factors.

We may also perform the resilience analysis in terms of a performance metric, such as AUC, for each risk factor in the plots showing *resiliency performance*. In this analysis we demonstrate the model's performance in the "worst-sample" scenario, varying the worst sample ratios ranging from 0.10 to 1.0, where a ratio of 1.0 implies that all test samples are treated as worst samples, while a ratio of 0.10 signifies that only 10% of the test samples are considered as worst samples. In the plot, the red dotted line represents the model's performance on the entire test sample. If the curve depicted in the plot is monotonic this indicates that as the worst sample ratio increases, the model's performance tends

to decrease, a visualization that allows for a clear understanding of how the model's performance varies when different proportions of worst samples are considered.

The resiliency performance plots are shown below in Figures 51 through 58, and all the plots are monotonic and show the expected degradation in performance as a higher proportion of samples are considered to be worst, with the rank ordering of the AUC generally similar across worse sample proportions. However, at the very lower proportions we do see some divergence in performance according to the model for some of the risk factors. While all these differences are generally minor there is a lack of consistency in the patterns across risk factors. In the case of the CCI risk factor shown in Figure 52, at the low proportions we observe the GAMI-Net model to be outperforming the ReLU-DNN model, while the LRM and EBM models converge. In the case of the CUR risk factor shown in Figure 53 the LRM and EBM models converge to the downside at the low proportions, while the GAMI-Net model performs best in this region relative to the ReLU-DNN model. In the case of the NQR risk factor shown in Figure 54 all models converge at the lower proportions except for the EBM model that diverges to the downside. All models converge at the lowest proportions for the NARD risk factor shown in Figure 55. The LRM and EBM models converge, while the ReLU-DNN model diverges to the downside from the GAMI-Net model in this region, for the BTPM risk factor shown in Figure 56. In Figure 57 we can see for the CTA risk factor that all models diverge at the lowest proportions, in decreasing levels of performance the GAMI-Net, ReLU-DNN, LRM and EBM models. There is a similar pattern to the latter observed for the TLTAR risk factor as seen in Figure 58 at the lower proportions, with the EBM model also performing the worst and the ReLU-DNN model the best, while there is intermediate performance for the LRM and GAMI-Net models with the former crossing the latter to the upside. Finally, in the case of the SP500EP risk factor shown in Figure 59, all models converge at the lowest level of the proportion, with EBM diverging to the downside.

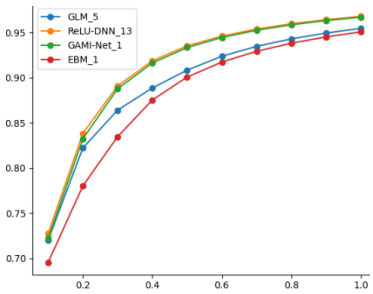


Figure 51. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the Consumer Confidence Index risk factor.

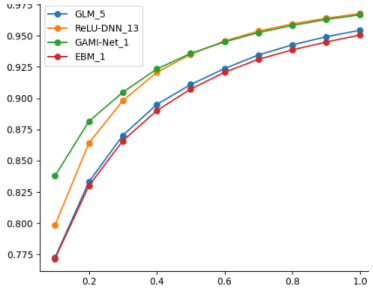


Figure 52. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the Cash Use Ratio risk factor.

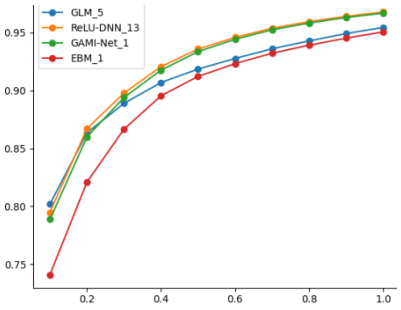


Figure 53. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the Net Quick Ratio risk factor.

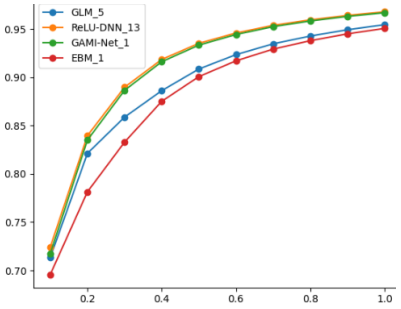


Figure 54. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the Net Account Receivables Days risk factor.

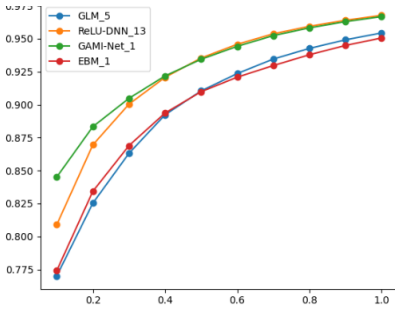


Figure 55. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the before tax profit margin risk factor.

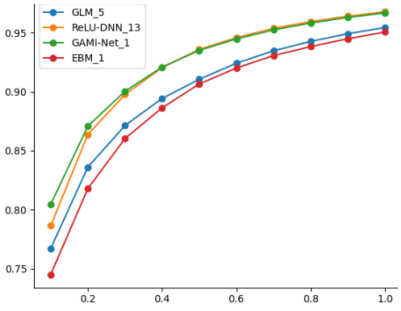


Figure 56. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the Change in Total Assets risk factor.

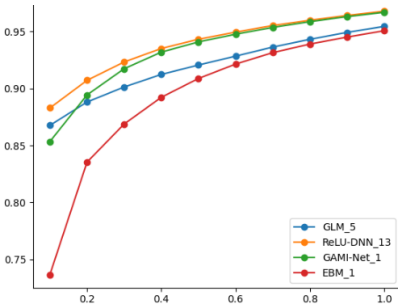


Figure 57. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the total liabilities to total assets ratio risk factor.

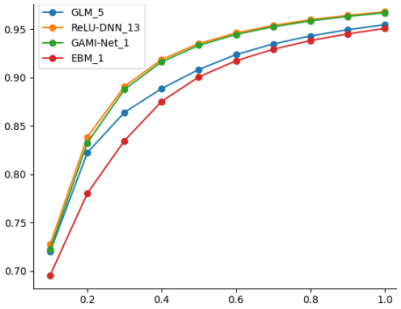


Figure 58. ML model benchmarking PD modeling analysis – comparison of resiliency performance plots for the S&P 500 Equity Price Index risk factor.

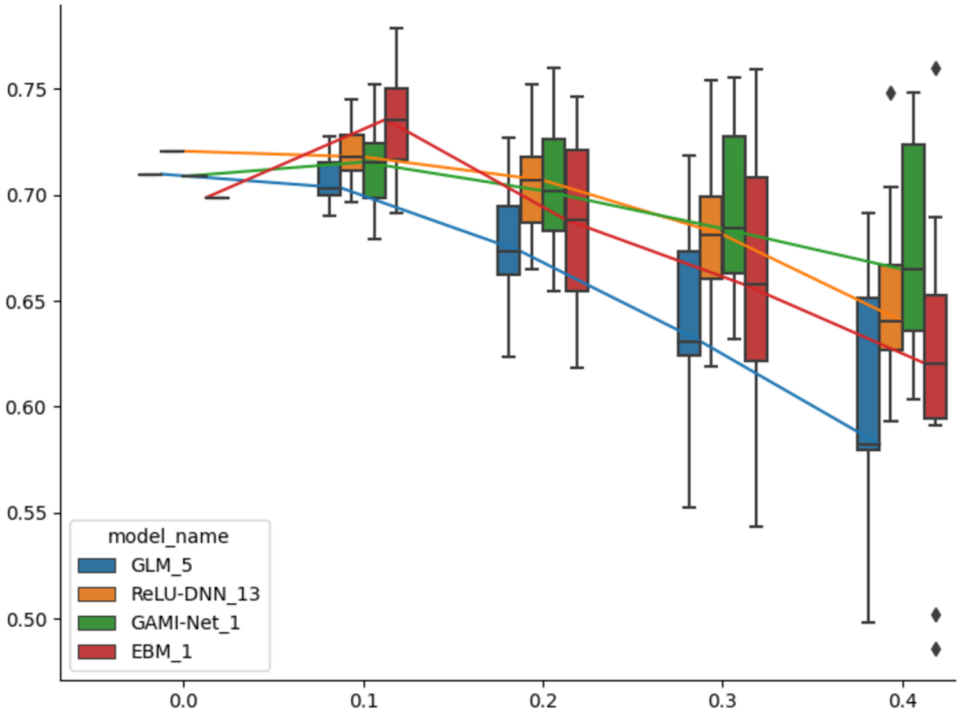


Figure 59. ML model benchmarking PD modeling analysis – comparison of robustness plots of the worst samples for all risk factors.

In addition to the robustness test on all test samples, we also test model robustness on so-called *worst-performing* test samples. We achieve this by identifying a top percentage of test samples with

the largest absolute residuals. These samples are considered the worst-performing because they have the highest discrepancies between the predicted values and the actual values, and then we apply the perturbations to these samples. This analysis is shown below in Figure 59 for all risk factors simultaneously in terms of ACC and in Figures 60 through 67 for each risk factor individually in terms of AUC.

In terms of the decline in mean ACC for the simultaneous analysis shown below in Figure 59 across all risk factors, where we choose to vary the worse sample proportion in increments 10% from 10% to 40%, the most robust model is the GAMI-Net. However, in these terms and also accounting for the dispersion of the distribution the best performing model is the ReLU-DNN, with the other two models intermediate in these regards, conclusions which differ slightly from the previously discussed robustness analysis.

This worst sample robustness analysis can also be conducted individually for each risk factor, where we show the distribution shift density, where we interpret density to the positive side as iterations where there is degradation in a performance metric. We show these densities below in Figures 60 through 67 for a 10% proportion of worst samples and the AUC performance metric. We observe in all cases that the densities vary in shape across risk factors, but for a given risk factor they have very similar in shape across models. In the case of the CCI risk factor in shown in Figure 60, we observe that the distributions are rather dispersed and for the 100% worst sample density a very multi-modal shape, with mean decrease in AUC ranging near 15% across models. In the case of the CUR risk factor shown in Figure 61, we observe that the distributions are rather concentrated in around the range of about 0.40 to 0.60, and for the 100% worst sample density a very similar shape, with mean decrease in AUC ranging around 13–15% across models. In the case of the NQR risk factor in shown in Figure 62, we observe that the distributions are rather concentrated around the range of about 0.00 to 0.30 and very right skewed, and for the 100% worst sample the density has a very similar shape, with mean decrease in AUC ranging around about 11–18% across models. In the case of the NARD risk factor in shown in Figure 63, we observe that the distributions are rather concentrated in around about 0.60 to 0.90 and very left skewed, and for the 100% worst sample density there is very similar shape, with mean decrease in AUC ranging around 11–18% across models.

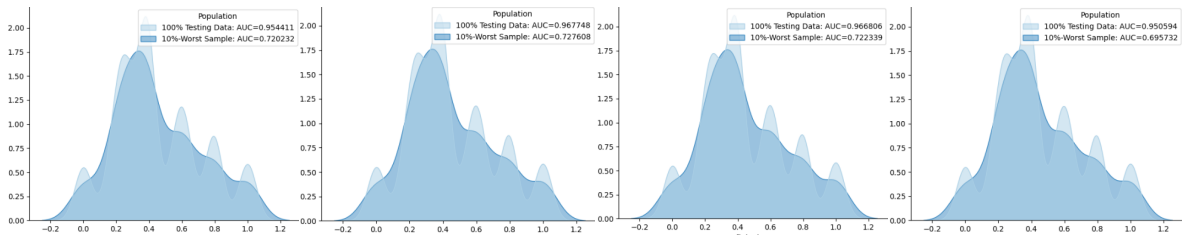


Figure 60. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the Consumer Confidence Index risk factor.

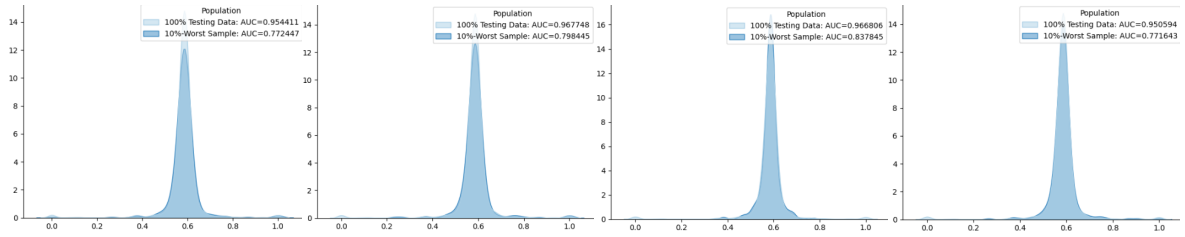


Figure 61. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the Cash Use Ratio risk factor.

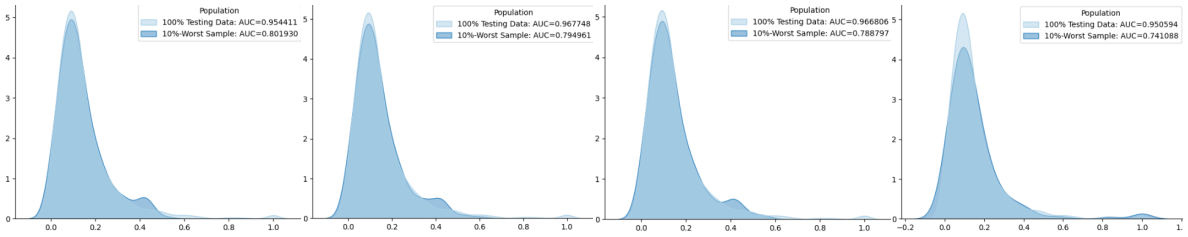


Figure 62. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the Net Quick Ratio risk factor.

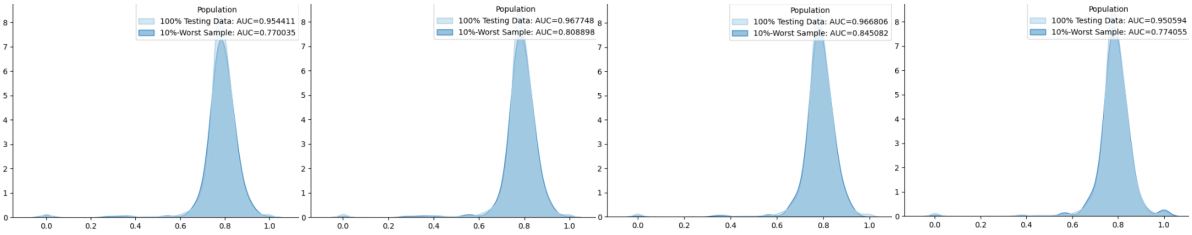


Figure 63. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the Net Account Receivables Days risk factor.

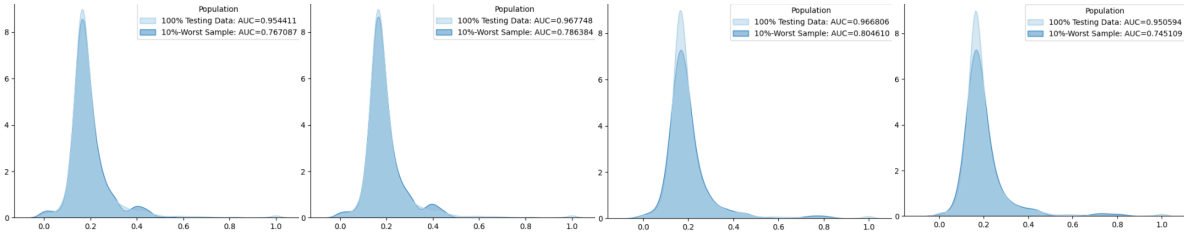
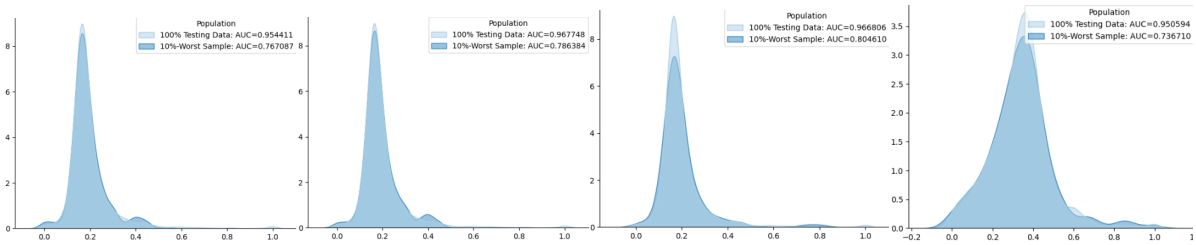
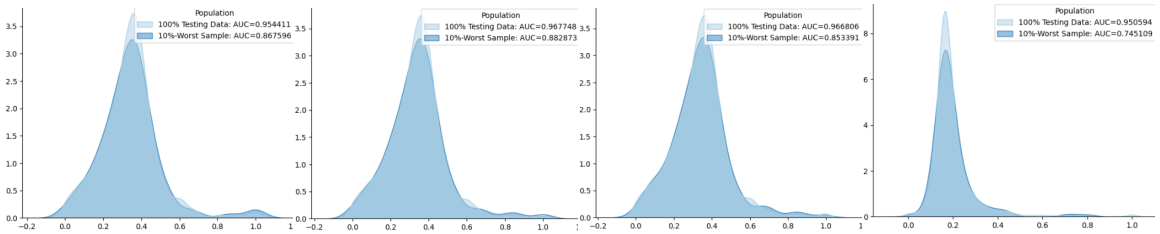


Figure 64. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the before tax profit margin risk factor.



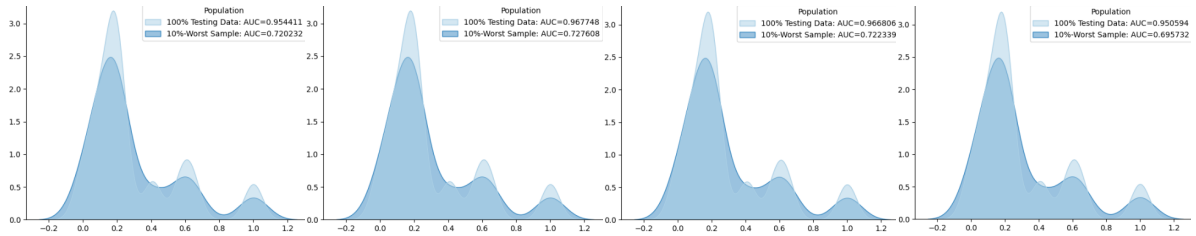
LRM ReLU-DNN GAMI-Net EBM

Figure 65. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the Change in Total Assets risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 66. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the total liabilities to total assets ratio risk factor.



LRM ReLU-DNN GAMI-Net EBM

Figure 67. ML model benchmarking PD modeling analysis – comparison of Robustness Plots of the 10% worst samples for the S&P 500 Equity Price Index risk factor.

In the case of the BTPM risk factor in shown in Figure 64, we observe that the distributions are rather concentrated in around the range of about 0.00 to 0.40 and very right skewed, and for the 100% worst sample density a very similar shape albeit with a lower mode, with mean decrease in AUC ranging around about 17–20% across models. In the case of the CTA risk factor in shown in Figure 65, we observe that the distributions are rather concentrated in around the range of about 0.00 to 0.40 and very right skewed, and for the 100% worst sample density a very similar shape albeit with a lower mode, with mean decrease in AUC ranging of around about 17–20% across models. In the case of the TLTAR risk factor in shown in Figure 66, we observe that the distributions are rather concentrated in around the range of about 0.00 to 0.60 and very right skewed, and for the 100% worst sample density a very similar shape albeit with a lower mode, with mean decrease in AUC ranging around about 9–21% across models. Finally, in the case of the SP500EPI risk factor in shown in Figure 67, we observe

that the distributions are rather dispersed with multiple modes in around the ranges of about 0.00 to 0.40, 0.50 to 0.70 and 0.90 to 1.10. They are also very right skewed, and for the 100% worst sample density there is a very similar shape albeit with a lower mode, with the mean decrease in AUC ranging around about 23–25% across models.

6. Conclusions and directions for future research

This study has investigated alternative IML models in the context of PD modeling applied to the large corporate asset class. This is motivated by the fact that IML models have become increasingly prominent in highly regulated industries where there are concerns over the unintended consequences of black box models that may be deemed conceptually unsound. In the context of banking and in wholesale portfolios, we have noted the challenges around deploying models where the outcomes may not be explainable, both in terms of meeting business use cases as well as in satisfying model validation standards. We compared various IML model, including standard approaches such as logistic regression, using a history of corporate borrowers sourced from Moody's, a dataset used in Jacobs (2022a, 2022b).

In our comparison of various IML models (GAMI-Net, ReLU-DNN and EBM), including standard approaches such as LRM, we found that there are material differences between the approaches in terms of dimensions such as model predictive performance the importance of risk factors in driving outcomes. While we observed that the IML models all demonstrate some pickup in AUC performance relative to the LRM, the degree of this outperformance is modest, especially on an out-of-sample basis: the ReLU-DNN (EBM) models showed the greatest increase (decrease) out-of-sample, while the EBM (ReLU-DNN) models showed the greatest (least) increase in-sample. We also observed in a comparison of FI measures across the three IML models and the LRM an overarching complete lack of consistency across both models and FI measures, but we noted that overall the SHAP-FI and LIME-FI measures were least inconsistent across models. This observation called into question the value of this pickup in performance with the IML models, especially if these models are to be applied in contexts that must meet model validation standards.

In the analysis of PDPs for each of the risk factors we found that while generally speaking the three IML models considered showed relationships to default risk that are as expected, as compared to the LRM model there are several instances of non-monotonicities or non-sensical results. In regard to the latter observation, the ReLU-DNN model had results that are most intuitive, and the EBM model tended to display patterns that are most aberrant, while the GAMI-Net model was intermediate in this respect. Given the limitations of the PDPs that we have noted, we also considered the ALE alternative that is robust to certain restrictive assumptions underlying the PDP construct. We noted the primary and rather obvious conclusion that the shapes of the ALE plots differed radically from that of the PDP plots, as all of these exhibited extreme non-continuities of variations in step-functions, for example having included in many cases "hockey-stick" or L- (reverse L-) shapes. Furthermore, as with the PDP plots we observed inconsistencies across models and marginal changes that in counterintuitive directions.

We analyzed prediction residuals between the default indicator dependent variable and the predicted PD estimates against the risk factor explanatory variables. In general we observed that, as expected this the case of a relatively low-default setting, the errors are much larger (much smaller) and closer to one for default (non-default), although there were some differences in the patterns across models and risk factors that were in some cases hard to rationalize.

In the robustness analysis where we perturbed all the risk factors, we observed that as expected AUC deteriorates as we increase the perturbation size, and that the ReLU-DNN and LRM models held up best as the variance of the AUC measure appeared most stable with increasing perturbation size, while the GAMI-Net and EBM models showed worse performance in this regard. In the robustness analysis for each risk factor separately we concluded that for a given variable there is some material variation in robustness across models for some risk factors, and in general that while the dispersion of the AUC consistently became greater with the shock, the means of the distributions did not get materially worse. Also, in general we observed that the LRM, ReLU-DNN and GAMI-Net models were most consistent in patterns across risk factors, while the EBM was least consistent in this regard and in some case has counterintuitive patterns.

In the resilience distribution analysis, where we assessed the performance of the worst samples in a model across various out-of-distribution scenarios according to the PSI distributional distance between the worst test sample and the full test sample calculated for each risk factor, we observed that there is general agreement across models in which risk factors have the greatest shifts in distribution. In the related resiliency performance analysis in terms the AUC performance metric, where we demonstrated the model's performance in the "worst-sample" scenario, all the plots were observed to be monotonic and to show the expected degradation in performance as a lower proportion of samples are considered to be worst, with the rank ordering of the AUC generally similar across worse sample proportions. However, at the very lower proportions we did see some divergence in performance according to the model for some of the risk factors, and while all these differences were seen to be generally minor there was an observed lack of consistency in the patterns across risk factors.

We also tested a variation of the above, the model robustness on so-called worst-performing test samples, achieved by identifying a top percentage of test samples with the largest absolute residuals between the predicted values and the actual values, and then applying the perturbations to these samples. In terms of the decline in mean ACC for the simultaneous analysis for all risk factors, the most robust model was the GAMI-Net. However, in these terms and also accounting for the dispersion of the distribution the best performing model was found to be the ReLU-DNN, with the other two models intermediate in this regards, conclusions which differed slightly from the previously discussed robustness analysis.

While there have been significant advances in IML that have opened the possibility for ML models to be more used in domains of heightened supervisory scrutiny, such as in credit risk modeling, these findings suggest that the industry may have a long way to go before this aspiration is fully realized. Furthermore, as can be seen in looking at the deep and rapidly growing literature on IML models, there are still many theoretical and technical questions that are yet unanswered, such as the interpretability measures and IML model variants that best achieve the promised objective of IML Given these considerations, depending upon the application of the PD model, we counsel practitioners to proceed cautiously in putting IML models into production in a champion capacity until these issues are resolved. That said, in spite of these limitations, we see scope for applying IML models in development or validation capacities as challengers or benchmarks as part of model testing and evaluation.

Given the wide relevance and scope of the topics addressed in this study, there is no shortage of fruitful avenues along which we could extend this research. Some proposals include but are not limited to:

- Asset classes beyond the large corporate segments, such as small business, real estate or even retail;

- applications to stress testing of credit risk portfolios[4];
- the consideration of industry specificity in model specification;
- different modeling methodologies, such as ratings migration or hazard rate models; and,
- datasets in jurisdictions apart from the U.S., else pooled data encompassing different countries with a consideration of geographical effects.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The author declare no conflict of interest. The views expressed herein are those of the author do not necessarily represent an official position of PNC Financial Services Group.

References

Abdulrahman UFI, Panford JK, Hayfron-Acquah JB (2014) Fuzzy logic approach to credit scoring for micro finances in Ghana: a case study of KWIQPUS money lending. *Int J Comput Appl* 94: 11–18. https://doi.org/10.5120/16362-5772

Allen L, Peng L, Shan Y (2020) Social networks and credit allocation on fintech lending platforms. working paper, social science research network. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537714.

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Financ* 23: 589–609. Available from: https://pdfs.semanticscholar.org/cab5/059bfc5bf4b70b106434e0cb665f3183fd4a.pdf.

Altman EI, Narayanan P (1997) An international survey of business failure classification models. in financial markets, institutions and instruments. New York: New York University Salomon Center, 6. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0416.00010.

American Banking Association (2018) New credit score unveiled drawing on bank account data. *ABA Bank J*, October 22.

Anagnostou I, Kandhai D, Sanchez Rivero J, et al. (2020) Contagious defaults in a credit portfolio: a Bayesian network approach. *J Credit Risk* 16: 1–26. https://dx.doi.org/10.2139/ssrn.3446615

Bjorkegren D, Grissen D (2020) Behavior revealed in mobile phone usage predicts credit repayment. *World Bank Econ Rev* 34: 618–634. https://doi.org/10.1093/wber/lhz006

Bonds D (1999) Modeling term structures of defaultable bonds. *Rev Financ Stud* 12: 687–720. Available from: https://academic.oup.com/rfs/article-abstract/12/4/687/1578719?redirectedFrom=fulltext.

[4] Refer to Jacobs Jr et al. (2015), Jacobs Jr. (2020a), Jacobs Jr. (2020b) and Jacobs Jr. (2022) for studies that address model validation and model risk quantification methodologies. These studies include supervisory applications such as the comprehensive capital analysis and review ("CCAR") and current expected credit loss ("CECL") framework; and further feature alternative credit risk model specifications (including machine learning model), macroeconomic scenario generation techniques, as well as the quantification and aggregation of model risk (including the principle of relative entropy as studied in this paper.)

Breeden J (2021) A survey of machine learning in credit risk. *J Risk Model Validat* 17: 1–62.

Chava S, Jarrow RA (2004) Bankruptcy prediction with industry effects. *Rev Financ* 8: 537–69. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=287474.

Clemen R (1989) Combining forecasts: a review and annotated bibliography. *Int J Forecast* 5: 559–583 Available from: https://people.duke.edu/~clemen/bio/Published%20Papers/13.CombiningReview-Clemen-IJOF-89.pdf.

Coats PK, Fant LF (1993) Recognizing financial distress patterns using a neural network tool. *Financ Manage* 22: 142–155. Available from: https://ideas.repec.org/a/fma/fmanag/coats93.html.

Collobert R, Weston J (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1: 160–167. Association for Computing Machinery, New York. https://doi.org/10.1145/1390156.1390177

Duffie D, Singleton KJ (1999) Simulating correlated defaults. Paper presented at the Bank of England Conference on Credit Risk Modeling and Regulatory Implications Working Paper, Stanford University. Available from: https://kenneths.people.stanford.edu/sites/g/files/sbiybj3396/f/duffiesingleton1999.pdf.

Dwyer DW, Kogacil AE, Stein RM (2004) Moody's KMV RiskCalc™ v2.1 Model. Moody's Analytics. Available from: https://www.moodys.com/sites/products/productattachments/riskcalc%202.1%20whitepaper.pdf.

Harrell FE Jr (2018) Road Map for Choosing Between Statistical Modeling and Machine Learning, *Stat Think.* Available from: http://www.fharrell.com/post/stat-ml.

Jacobs Jr M (2022a) Quantification of model risk with an application to probability of default estimation and stress testing for a large corporate portfolio. *J Risk Model Validat* 15: 1–39. https://doi.org/10.21314/JRMV.2022.023

Jacobs Jr M (2022b) Validation of corporate probability of default models considering alternative use cases and the quantification of model risk. *Data Sci Financ Econ* 2: 17–53. https://doi.org/10.3934/DSFE.2022002

Jarrow RA, Turnbull SM (1995) Pricing derivatives on financial securities subject to credit risk. *J Fnanc* 50: 53–85. https://doi.org/10.1111/j.1540-6261.1995.tb05167.x

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1: 1097–1105. Available from: https://cse.iitk.ac.in/users/cs365/2013/hw2/krizhevsky-hinton-12-imagenet-convolutional-NN-deep.pdf.

Kumar IE, Venkatasubramanian S, Scheidegger C, et al. (2020) Problems with Shapley-Value-Based Explanations as Feature Importance Measures. In *International Conference on Machine Learning Research*: 5491–5500. Available from: https://proceedings.mlr.press/v119/kumar20e.html.

Lessmann S, Baesens B, Seow HV, et al. (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Europ J Oper Res* 247: 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li K, Niskanen J, Kolehmainen M, et al. (2016) Financial innovation: Credit default hybrid model for SME lending. *Expert Syst Appl* 61: 343–355. https://doi.org/10.1016/j.eswa.2016.05.029

Li X, Liu S, Li Z, et al. (2020) Flowscope: spotting money laundering based on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 4731–4738. https://doi.org/10.1609/aaai.v34i04.5906

Lou Y, Caruana R, Gehrke J, et al. (2013) Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 623–631. https://dl.acm.org/doi/abs/10.1145/2487575.2487579

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. Available from: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Mckee TE (2000) Developing a bankruptcy prediction model via rough sets theory. *Intell Syst Account Financ Manage* 9: 159–173. https://doi.org/fv22ks

Merton RC (1974) On the pricing of corporate debt: The risk structure of interest rates. *J Fnanc* 29: 449–470.

Mester LJ (1997) What's the point of credit scoring? Federal Reserve Bank of Philadelphia. *Bus Rev* 3: 3–16. Available from: https://fraser.stlouisfed.org/files/docs/historical/frbphi/businessreview/frbphil_rev_199709.pdf.

Min JH, Lee YC (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Syst Appl* 28: 603–614 https://doi.org/10.1016/j.eswa.2004.12.008

Molnar C, König G, Herbinger J, et al. (2020) Pitfalls to avoid when interpreting machine learning models. Working Paper, University of Vienna. Available from: http://eprints.cs.univie.ac.at/6427/.

Odom MD, Sharda R (1990) A neural network model for bankruptcy prediction. In *Joint Conference on Neural Networks*, 163–168. IEEE Press, Piscataway, NJ. Available from: https://ieeexplore.ieee.org/abstract/document/5726669.

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. *J Artif Intell Res* 11: 169–198.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135–1144. Available from: https://dl.acm.org/doi/abs/10.1145/2939672.2939778.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x

Slack D, Hilgard S, Jia E, et al. (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. https://dl.acm.org/doi/abs/10.1145/3375627.3375830

Sudjianto A, Knauth W, Rahul S, et al. (2020) Unwrapping the black box of deep ReLU networks: Interpretability, diagnostics, and simplification. *arXiv preprin*, Cornell University. https://arxiv.org/abs/2011.04041

Sudjianto A, Zhang A (2021) Designing inherently interpretable machine learning models. In *Proceedings of ACM ICAIF 2021 Workshop on Explainable AI in Finance*. ACM, New York. https://arxiv.org/abs/2111.01743

Sudjianto A, Zhang A, Yang Z, et al. (2023) PiML toolbox for interpretable machine learning model development and validation. *arXiv preprint arXiv*. https://doi.org/10.48550/arXiv.2305.04214

U.S. Banking Regulatory Agencies (2011) The U.S. Office of the Comptroller of the Currency and the Board of Governors of Federal Reserve System. SR 11-7/OCC11-12: Supervisory Guidance on Model Risk Management. Washington, D.C. Available from: https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

U.S. Banking Regulatory Agencies (2021) The U.S. Office of Comptroller of the Currency, the Board of Governors of the Federal Reserve System, the Federal Deposit Insurance Corporation, the Consumer Financial Protection Bureau, and the National Credit Union Administration. Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning. Washington, D.C. Available from: https://www.federalregister.gov/documents/2021/03/31/2021-06607/requestfor-information-and-comment-on-financial-institutions-use-of-artificialintelligence.

U.S. Office of the Comptroller of the Currency (2021) Comptroller's Handbook on Model Risk Management. Washington, D.C. Available from: https://www.occ.treas.gov/publicationsand-resources/publications/comptrollers-handbook/files/model-riskmanagement/index-model-risk-management.html.

Vahid PR, Ahmadi A (2016) Modeling corporate customers' credit risk considering the ensemble approaches in multiclass classification: evidence from Iranian corporate credits. *J Credit Risk* 12: 71–95.

Vassiliou PC (2013) Fuzzy semi-Markov migration process in credit risk. *Fuzzy Sets and Syst* 223: 39–58. https://doi.org/10.1016/j.fss.2013.02.016

Yang Z, Zhang A, Sudjianto A (2021a) Enhancing explainability of neural networks through architecture constraints. I*EEE T Neur Net Learn Syst* 32: 2610–2621. https://doi.org/10.1109/TNNLS.2020.3007259

Yang Z, Zhang A, Sudjianto A (2021) GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recogn* 120: 108192. https://doi.org/10.1016/j.patcog.2021.10819

Zhu Y, Xie C, Wang G J, et al. (2017) Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Comput Appl* 28: 41–50. https://doi.org/10.1007/s00521-016-2304-x

AIMS Press