Foundation Models for Credit Risk Prediction: A Game Changer?

Bart Baesens [1], Andreas Goethals [1], Stefan Lessmann [2], Simon De Vos [1], Cristián Bravo [3], David Martens [4], Victor Medina-Olivares [5], Christophe Mues [6], Maria Óskarsdóttir [7], Seppe vanden Broucke [8,1], Tim Verdonck [9,10], and Wouter Verbeke [1]

[1]Faculty of Economics and Business, KU Leuven, Belgium
[2]School of Business and Economics, Humboldt University of Berlin, Germany
[3]Department of Statistical and Actuarial Sciences, Western University, Canada
[4]Department of Engineering Management, University of Antwerp, Belgium
[5]Business School, University of Edinburgh, United Kingdom
[6]Business School, University of Southampton, United Kingdom
[7]School of Mathematical Sciences, University of Southampton, United Kingdom
[8]Department of Business Informatics and Operations Management, Ghent University, Belgium
[9]Department of Mathematics, University of Antwerp, Belgium
[10]Department of Mathematics, KU Leuven, Belgium

Abstract

Predictive models play a pivotal role in credit risk management, guiding critical decisions through accurate estimation of default probabilities and losses. Their performance influences the profitability of lending operations and the stability of the financial system. The relevance of predictive modeling has generated extensive interest in credit risk research, with numerous studies introducing new modeling techniques. Large-scale benchmarking studies complemented these endeavors by periodically consolidating the state-of-the-art and systematically uncovering the merits and demerits of methodological advancements. Today, quasi-standards such as a gradient-boosting model for prediction, often paired with a SHAP explainer, have emerged for specific contexts. However, the continuous improvement of risk models and modeling practices remains a top priority.

Concurrently, the rapid advancements in AI, most notably through large language models, have disrupted predictive modeling paradigms in many fields. Foundation models, pretrained on extensive datasets from diverse domains, have demonstrated remarkable performance by leveraging prior knowledge acquired during pretraining. While prevalent in natural language processing and computer vision, foundation models specifically designed for tabular data have only recently emerged and may hold great potential for credit risk management. We conjecture that pretraining on out-of-domain data is particularly beneficial in small-data settings, such as SME lending or managing specialized corporate portfolios. More technically, foundation models may help address longstanding credit-scoring challenges, including low default portfolios or class imbalance. However, the actual value of foundation models for credit risk prediction remains an open question.

This paper focuses on recently proposed foundation models for tabular data. We benchmark these models against a broad set of competitors, including established and advanced machine learning techniques in two core prediction tasks: PD and LGD modeling. Our evaluation extends on the design of previous benchmarking studies, encompassing various datasets, performance indicators, and experimental conditions to clarify pretraining benefits across relevant risk modeling challenges. We find that tabular foundation models generally perform best across datasets and tasks in comparison with state-of-the-art alternatives. Moreover, we find that tabular foundation models offer significant improvement in predictive performance when the size of the dataset in terms of number of observations grows smaller. These results are remarkable, since the tabular foundation models are tested out-of-the-box, without applying any hyperparameter tuning, ensuring ease of use and mitigating the computational costs associated with employing foundation models.

Keywords Credit Risk Modeling · Foundation Models · Probability of Default · Loss Given Default

1 Introduction

Foundation models are large-scale machine learning systems that have been trained on vast and heterogeneous corpora and, as such, acquire representations and inductive biases that transfer to a wide range of downstream tasks (possibly after fine-tuning through further training on application-specific data) [Schneider et al., 2024]. Foundation models have transformed domains such as natural language processing (NLP) and computer vision. In NLP, large language models such as GPT-4 have redefined the state of the art in tasks such as next-word prediction and text generation [Bubeck et al., 2023]; in computer vision, models such as CLIP [Radford et al., 2021] and diffusion-based generators have enabled zero-shot classification—classification of previously unseen classes or tasks without task-specific training—and high-fidelity synthesis that were previously out of reach [Awais et al., 2025]. These advances exemplify a shift away from traditional model-training approaches towards developing and using general-purpose models with broad applicability and, potentially, improved task performance.

Analogous developments have recently reached tabular data—the workhorse format of credit risk analytics. Tabular learning differs markedly from unstructured domains due to heterogeneous feature types, disparate scales, and dataset-specific semantics that have historically impeded knowledge transfer. Tree-based ensembles such as gradient boosting machines (GBMs) have long dominated tabular benchmarks, consistently outperforming deep learning alternatives in extensive comparative studies [Lessmann et al., 2015, Gunnarsson et al., 2021, Shwartz-Ziv and Armon, 2022a]. Prior-Data Fitted Networks (PFNs), however, extend the foundation model idea to the tabular setting by pretraining on large numbers of synthetic datasets drawn from priors over data-generating processes [Müller et al., 2021, Hollmann et al., 2025a]. At inference, a PFN processes an entire dataset, including labeled and unlabeled instances, in a single forward pass and outputs predictive distributions without further training or tuning.

This paper examines whether the foundation model paradigm benefits credit risk modeling. Credit risk management is critically dependent on accurate predictions of risk parameters, notably the probability of default (PD) and the loss given default (LGD). Improvements in predictive accuracy directly affect capital calculation, allocation, and portfolio steering under the Basel framework, provisioning under IFRS 9, and risk-adjusted pricing. If PFNs can deliver competitive accuracy without task-specific training, they could simplify model development, monitoring, and maintenance, reduce operational costs, and improve time-to-model, especially in small-data contexts.

PFNs promise advantages that are particularly relevant for credit risk modeling practice. First, as zero-shot forecasters, they avoid repeated hyperparameter tuning and retraining across tasks and portfolios. This property is attractive for portfolios with limited data availability, such as low-default portfolios, specialized corporate segments, or new credit products. Second, PFNs could enable a consistent modeling approach across risk parameters (PD, LGD, and potentially exposure at default, EAD), thereby benefiting governance by simplifying validation, monitoring, and model risk management. From a supervisory perspective, cross-parameter consistency within financial institutions and methodological uniformity across institutions may reduce the burden of model review. Third, by combining broad synthetic pretraining with in-context learning on institutional data, PFNs may temper historical biases embedded in local datasets and thereby support fairer, more inclusive decision making [Óskarsdóttir et al., 2019, Kozodoi et al., 2022, Moldovan, 2023].

Despite this potential, there are reasons for caution. Credit risk modeling presents specific challenges—severe class imbalance, heterogeneous borrower populations, economic non-stationarities, and regulatory constraints—that may not be fully addressed by current tabular foundation models (TFMs). Moreover, a systematic evaluation of PFNs for credit risk prediction has been lacking. While anecdotal evidence suggests strong performance on small datasets, it remains unclear whether PFNs can consistently match tuned baselines across diverse credit portfolios and targets.

We address this gap with an extensive empirical benchmark. We compare PFNs against classical and advanced machine learning methods on a collection of PD and LGD datasets that vary in size, dimensionality, and label characteristics. Our evaluation protocol follows the benchmarking tradition in credit scoring, uses cross-validation, and assesses a comprehensive set of performance metrics. We further conduct statistical tests to assess the significance of algorithm comparisons across datasets. As a by-product, our benchmark provides the first credit-risk assessment, to the best of our knowledge, of several recent non-PFN deep learning algorithms for tabular data.

The remainder is organized as follows. Section 2 reviews credit risk modeling and recent advances in deep learning for tabular data. Section 3 introduces PFNs and their Bayesian interpretation. Section 4 details the experimental setup in terms of datasets, preprocessing steps, the evaluated learners, and the evaluation protocol. Section 5 presents empirical results and discusses practical implications. Section 6 concludes.

2 Literature review

This paper contributes to the empirical credit scoring literature by providing original results concerning the predictive performance of TFMs in two crucial prediction applications, PD and LGD modeling. In doing so, the paper draws on a large body of prior work on credit risk prediction and extends it by introducing a new family of predictive methods. Beyond credit risk and, more generally, financial applications, the paper also draws on prior work on deep learning for tabular data, which paved the way for the recently introduced PFNs, which are central to this study. Below, we review prior work in these two fields and elaborate on our contributions.

2.1 Credit risk modeling

Credit risk refers to the potential loss arising from a counterparty's failure to meet contractual obligations [Baesens et al., 2016]. In retail and wholesale banking, risk models quantify key risk parameters used for pricing, capital, provisioning, and portfolio management: the PD, which estimates the likelihood of a borrower defaulting over a fixed horizon, and the LGD, which measures the proportion of exposure not recovered if default occurs. In both PD and LGD modeling, predictive accuracy is key to the efficiency and effectiveness of credit risk management [Baesens et al., 2003, Gunnarsson et al., 2021, Loterman et al., 2012].

Better predictions directly translate into improved capital efficiency (e.g., when estimating expected loss under the Basel Accord's IRB approach), more precise and timely provisioning under IFRS 9 through improved estimation of lifetime expected credit losses, reduced impairments, and better pricing decisions [Bastos and Matos, 2022]. Regulators likewise emphasize the discriminatory power and calibration of scorecards, for example through backtesting, monitoring stability, and validation requirements.

Aiming to unlock financial rewards from more accurate predictions, the community routinely assesses the potential of new prediction methods and proposes methodological adjustments tailored to characteristic modeling challenges. For PD, these challenges include class imbalance [Brown and Mues, 2012, Marqués et al., 2013, Wang et al., 2025, Engelmann and Lessmann, 2021], reject inference [Banasik and Crook, 2007, Kozodoi et al., 2025], economic cycle dependency [Bellotti and Crook, 2012, Djeundje et al., 2025, Distaso et al., 2025, Dirick et al., 2019], or non-linear interactions in borrower behavior [Thomas, 2000, Van Gestel et al., 2005]. Compared with PD prediction, fewer studies have considered LGD modeling [Yao et al., 2015, Cheng et al., 2025, Hurlin et al., 2018a], which may be due to the limited public availability of LGD datasets. Public datasets that are frequently used in LGD modeling research include the *LendingClub* and *Freddie Mac* datasets [Calabrese and Zanin, 2022, Zhou et al., 2018], which provide sufficient information to approximate LGD. Given the characteristic bi-modal shape of real-world LGD distributions, arising, for example, from mixed portfolios comprising both secured and unsecured loans, many studies recommend multi-stage models [Tomarchio and Punzo, 2019, Starosta, 2021, Papouskova and Hajek, 2019]. Dynamic models that explicitly recognize capital inflows and outflows throughout the workout period [Thomas et al., 2016] are another popular research topic in the LGD literature, often involving survival modeling [e.g., Joubert et al., 2021].

Given the prevalence of machine learning and advanced predictive methods in the credit scoring literature, several benchmarking studies have periodically consolidated the state of the art by applying an arsenal of statistical and machine learning methods across multiple datasets, performance metrics, and experimental conditions. Prominent examples include Baesens et al. [2003], Lessmann et al. [2015], Gunnarsson et al. [2021], Jiang et al. [2023] and Loterman et al. [2012], Bellotti et al. [2021], for PD and LGD prediction, respectively. Closely connected to comparing the predictive performance of alternative methods is the question of how to properly assess predictions. Several attempts have been made to connect the accuracy of model predictions to profitability and other measures of business impact [Verbraken et al., 2014, Martin et al., 2025, Garrido et al., 2018]. While quantification of business impact is more developed in PD modeling, some studies have pursued similar goals in the context of LGD [Hurlin et al., 2018b].

Conditional on the availability of suitable indicators of a risk model's business impact, the predictive models themselves can be optimized for cost-minimization [Bahnsen et al., 2015] or profit maximization [Finlay, 2010, Serrano-Cinca and Gutiérrez-Nieto, 2016, Xu et al., 2025, Alfonso-Sánchez et al., 2024]. Beyond the actual risk model, other stages in the modeling pipeline can be optimized for cost/profit objectives. For example, Maldonado et al. [2017], Kozodoi et al. [2019] introduce techniques for profit-oriented feature selection. Also targeting the input data of predictive models, recent work aims to extract information from borrower networks that standard methods relying on the IID assumption would fail to capture. Using methods from geospatial econometrics, related studies have, for example, established correlations between default events across geographical neighborhoods [Medina-Olivares et al., 2025, Calabrese and Crook, 2020]. More generally, graph-based algorithms, such as graph neural networks (GNNs), facilitate the processing of comprehensive dependency structures, providing modelers with considerable freedom in how to design an influence network.

Óskarsdóttir and Bravo [2021] pioneered the use of multilayer network analysis in credit-risk modeling, showing that borrower connections can provide substantial predictive gains. This line of work has since developed towards graph-based and graph neural network approaches for credit-risk prediction [Zandi et al., 2025, Shi et al., 2024]. The pursuit to incorporate auxiliary information in risk models beyond what classical scorecards embody has also inspired scholars to systematically explore novel sources of data, including social media postings [De Cnudde et al., 2018], psychometric data [Djeundje et al., 2021], and, perhaps most prominently, textual data [Wu et al., 2025, Stevenson et al., 2021, Kriebel and Stitz, 2022]. A distinctive advantage of such alternative data is that it may facilitate accurate risk assessment and, by extension, lending to *thin-book clients* whose limited credit history may otherwise exclude them from accessing financial services [Óskarsdóttir et al., 2019].

Facing increasing concern related to the omnipresence of advanced, AI-based, decision models in lending—and beyond—much recent work concentrates on explainable AI (XAI) techniques to ensure that the mechanisms underlying such models are well-understood [De Bock et al., 2024]. Model explainability is a regulatory imperative and prerequisite for achieving higher-level objectives, including robustness, safety, and fairness [Barredo Arrieta et al., 2020]. Recent work on XAI for credit risk examines the interplay between ML-based scorecards and XAI tools in general [e.g., Bastos and Matos, 2022], the robustness of explanations [Ballegeer et al., 2025], or the moderating effect of longstanding challenges, such as class imbalance [Chen et al., 2024]. Benefiting from its strong theoretical foundations in cooperative game theory, the SHAP framework for additive feature attribution has gained considerable popularity in credit risk and may be considered a quasi-standard [du Toit et al., 2023, Borgonovo et al., 2024]. Unlike SHAP, which decomposes a model's prediction, be it right or wrong, Hué et al. [2025] introduces an interesting challenger approach that inherits the same theoretical underpinnings but decomposes a scorecard's predictive performance. The corresponding insights might be even more meaningful to decision-makers as, on one hand, performance statistics are easier to interpret than predictions, and, on the other hand, knowing which features drive performance provides directly actionable insights to improve the model. Apart from techniques to explain otherwise opaque machine learning models, several studies devise new, intrinsically interpretable learning algorithms and demonstrate that these often avoid sacrificing (much) predictive accuracy. Representative examples include tree-based approaches [Carrizosa et al., 2025, De Caigny et al., 2018] , generalized additive models [Kraus et al., 2024], and interpretable deep learning techniques [Medina-Olivares et al., 2024, Zografopoulos et al., 2025].

Regulatory frameworks governing the use of AI in consumer-facing applications, such as the EU AI Act, the EU GDPR, and recent revisions to the Basel capital accord, require financial institutions to demonstrate that their risk models are not only explainable but also free from discriminatory bias against historically disadvantaged groups. Seminal papers provide evidence that scorecards exhibit such biases and that machine-learning-based approaches are especially vulnerable [Fu et al., 2021, Fuster et al., 2022]. Focusing on retail lending decisions, Kozodoi et al. [2022] systematically analyze available statistical indicators of algorithmic fairness and algorithms to mitigate disparate impact in the machine learning pipeline for their suitability in credit scoring. Hurlin et al. [2025] introduce a powerful tool to test a given scorecard for algorithmic bias, enabling lenders to verify compliance (or the lack thereof) with regulatory requirements.

In summary, it is clear that machine learning and AI play a pivotal role in credit risk modeling. They continuously supply novel technologies that aid lenders in their quest for more effective, efficient, and ultimately profitable decision models that are also compliant with regulation.

This paper contributes to the empirical credit risk modeling literature. Although we focus on established classification and regression settings for PD and LGD modeling, respectively, our goal extends beyond assessing yet another prediction method. Instead, we evaluate a *novel paradigm* for risk prediction: the use of tabular data foundation models for zero-shot forecasting. Adopting a zero-shot forecasting method to support lending operations entails a fundamental shift in established risk modeling practices, moving from task-specific model training and calibration to the use of pretrained foundation models and in-context learning. Given these consequences and the prevalence of predictive modeling in credit risk, we consider a holistic evaluation of the new paradigm in canonical risk modeling tasks imperative.

2.2 Deep learning for tabular data

The study follows prior work and established industry practices of approaching PD and LGD modeling as classification and regression problems, respectively. Consequently, we use supervised machine learning algorithms for labeled tabular data [Baesens et al., 2016].

Deep learning (DL) has revolutionized various data modalities and represents the uncontested state-of-the-art in NLP and computer vision, to name only a few [LeCun et al., 2015]. Given that natural language is a specific type of sequential data, it is not surprising that DL has also become a *go-to method* for time series forecasting [Benidis et al.,

2022]. With respect to cross-sectional tabular data, however, classic ML methods, particularly gradient-boosted trees (GBM), may be considered the de facto standard. Benchmarking experiments have shown the superiority of GBM over DL-based approaches in that the former often provide more accurate predictions at lower computational costs and/or with less need for hyperparameter tuning [Shwartz-Ziv and Armon, 2022b, Grinsztajn et al., 2022]. Shmuel et al. [2025] contextualize this finding by elaborating on the specific conditions (e.g., dataset characteristics) when DL excels. Considering PD modeling, however, Gunnarsson et al. [2021] argue that the potential of DL lies in unlocking non-standard data sources (e.g., text) to augment risk models rather than serving as a vehicle for model estimation. For model estimation, Gunnarsson et al. [2021] find that GBM-type approaches excel, which echoes earlier benchmarking results.

We argue that the Gunnarsson et al. [2021] study represents the credit risk modeling space well, in particular for tabular data. DL has provided excellent results in use cases involving semi-structured and unstructured data modalities [Bravo et al., 2026], such as text data or networks [e.g., Stevenson et al., 2021, Zandi et al., 2025], and settings that employ time-varying covariates [e.g., Korangi et al., 2023, Medina-Olivares et al., 2024], whereas GBM-type algorithms, often paired with a post-hoc XAI approach, remain the standard solution when the goal is to develop highly accurate risk scorecards.[1]

The unmatched success of DL in unstructured data settings stands in sharp contrast to its moderate performance on tabular data, which has inspired much recent work aimed at closing the performance gap with GBM-type approaches. Given that the transformer architecture was instrumental to the success of DL in unstructured data settings and that it relies on the attention mechanism, unlocking the power of attention for tabular data was a natural next step. The work by Gorishniy et al. [2021], introducing FT-Transformer, may be seen as a seminal paper paving the way for more advanced DL-based tabular data models. The authors demonstrate effective ways to utilize attention when working with tabular data, which is essential because it enables models to flexibly capture heterogeneous feature interactions common in tabular domains. Unlike convolutional or recurrent layers, attention can dynamically weight dependencies across diverse variables, reducing the need for manual feature engineering. Building on FT-Transformer, subsequent work has refined attention variants for tabular data, confirming its role as a central design principle in closing the performance gap with GBMs [Ye et al., 2024].

Jiang et al. [2025] provide a comprehensive survey of DL models for tabular data and their evolution. The authors identify three main evolutionary stages of deep tabular learning. The first stage includes approaches similar to FT-Transformer, which are trained and evaluated within a single distribution and focus on feature-level encoding, sample-level interactions, and objective-driven regularization [e.g., Arik and Pfister, 2021, Huang et al., 2020]. These approaches introduced innovations such as feature tokenization, attention-based feature selection, and inter-sample retrieval, but their generalization remained limited to the dataset at hand.

The second stage, transferable models, extended this paradigm by pre-training on one or multiple source datasets and fine-tuning on downstream tasks. Such models seek to overcome heterogeneity in feature and label spaces, often leveraging self-supervised objectives or language-model-inspired pretraining to enable cross-dataset knowledge transfer [e.g., Yoon et al., 2020, Ucar et al., 2021, Somepalli et al., 2021]. For example, TabLLM serializes tabular data by integrating feature names into text and combining them with task descriptions, which enables approaching tabular prediction tasks as language generation problems and, thus, the use of large language models in a few-shot mode [Hegselmann et al., 2023].

The most recent, third stage is marked by foundation models for tabular data, which provide zero-shot capabilities across diverse domains without the need for fine-tuning. Within this category, Prior-Data Fitted Networks (PFNs) stand out as a pioneering approach that harnesses synthetic pretraining and in-context learning with transformers to deliver robust generalization across a wide range of tabular problems [Hollmann et al., 2023]. PFNs approximate Bayesian inference by pretraining on synthetic datasets generated from, for example, structured causal models, thereby endowing the transformer with a learned prior that enables strong in-context learning capabilities [Müller et al., 2022]. Introducing TabPFNv2, Hollmann et al. [2025b] is a landmark paper demonstrating the generality, predictive capability, and scalability of a PFN-based zero-shot learner. The authors conduct comprehensive empirical comparisons to demonstrate that their approach matches the performance of GBM algorithms across a large set of regression and classification problems. Given that in-context learning requires a PFN to process an entire dataset in a single forward pass, the scalability of TabPFNv2 remains limited to moderately sized tables (e.g., 10,000 observations). Addressing this bottleneck is the subject of current research [Qu et al., 2025, Grinsztajn et al., 2026].

The three stages illustrate how the field of deep tabular learning has shifted from approaches that adopt the classic ML paradigm of training task-specific models to increasingly versatile architectures, which mimic the generality of

[1] We do not mean to imply that GBM-type approaches, or any ML-based approach, can be considered an industry standard. Linear models are the only approach that can potentially claim this role.

contemporary AI models. It is this generality, achieved through in-context learning and pretraining on large out-of-domain data collections, which we consider a *paradigm shift* with potentially far-reaching implications for credit risk modeling. This prospect, as well as the cautionary remarks of Klein and Hoffart [2025], who argue that the impressive results of PFNs in laboratory environments may give a false impression of how easy it is to use corresponding methods in practice, motivate the focal study, in which we undertake a comprehensive empirical evaluation of PFNs for credit risk modeling.

We acknowledge that our work does not advance deep tabular data learning methodology. Instead, we aim to contribute original empirical insights into how that field's latest innovations compare with established domain standards in credit risk modeling. Such evidence is valuable because the evaluation of PFNs has not moved beyond the standard benchmarking datasets routinely used in machine learning. These dataset collections are comprehensive and aim to be general by combining data from various fields. While clearly effective for demonstrating the potential of a new approach, decision-makers may appreciate targeted results that reflect the characteristics of their domain. Thus, our study provides a critical test of whether PFNs can move from promising benchmarks to delivering real impact in credit risk modeling.

3 Prior-data fitted networks (PFNs)

PFNs are pretrained models that approximate the Bayesian posterior predictive distribution at inference time [Müller et al., 2022]. While applicable to various data types, TabPFN, a specific type of PFN, has shown great potential as a foundation model for tabular prediction [Hollmann et al., 2023]. The core idea is to encode prior beliefs about likely relationships between inputs and outputs by defining a prior over data-generating processes, which are functions that map features to targets. Next, TabPFN repeatedly samples functions from this prior to generate synthetic datasets, which are subsequently used as training examples. Using these synthetic datasets as training points, a permutation-invariant transformer architecture is trained to map observed training pairs and features to posterior predictive distributions over the targets. Through this training, the model learns to approximate the posterior predictive distribution that is implied by the defined prior, conditioned on the observed data. Conceptually, TabPFN amortizes Bayesian inference: the computationally expensive inference is learned once during pretraining, so that inference on a new dataset reduces to a single forward pass.

Formally, consider a tabular dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ with features $x_i \in \mathbb{R}^d$ and targets y_i (binary for PD, continuous and bounded for LGD). TabPFN receives as input both labeled *context* examples and unlabeled *query* examples and outputs predictive distributions $p(y \mid x, \mathcal{D}_{\text{context}})$ over the target. During pretraining, synthetic tasks are generated by sampling from a prior over joint distributions $p(x, y)$, including mechanisms that induce missingness, correlations, and noise features, thereby ensuring that TabPFN encounters a diverse range of realistic tabular structures. The model is trained to minimize a proper scoring rule, e.g., log-loss for classification or a suitable continuous target loss for regression, over held-out query examples from each synthetic dataset.

A Bayesian interpretation arises because TabPFN aims to learn a mapping that approximates the posterior predictive distribution $p(y \mid x, \mathcal{D}_{\text{context}})$ that is implied by the training prior over data-generating processes [Müller et al., 2022]. In the limit of infinite synthetic data and capacity, TabPFN converges to Bayes-optimal predictions for tasks drawn from the prior. In practice, expressivity, prior misspecification, and optimization constraints introduce approximation errors, causing deviations from the true Bayesian posterior distribution. Nevertheless, empirical results indicate strong performance, especially on small datasets where traditional learners are prone to overfitting or require heavy regularization [Hollmann et al., 2025a].

Two properties are particularly salient for credit risk modeling. First, TabPFN provides *zero-shot* predictions on new datasets; it does not require any task-specific training or hyperparameter tuning, as inference reuses the pretrained network. Second, TabPFN can process tabular datasets with varying numbers of features and examples using cell embeddings and masked attention, allowing the same pretrained network to operate across heterogeneous structures. This flexibility is essential for lenders, as they must model risk exposure across heterogeneous credit portfolios, drawing on datasets that vary substantially in size, dimensionality, and distributional characteristics.

4 Experimental design

4.1 Data

We evaluate PFNs for PD and LGD estimation, corresponding to classification and regression tasks, respectively. Our experiments use both public and proprietary datasets spanning retail and small-business portfolios. Tables 1 and 2

summarize the key characteristics of the PD and LGD datasets, respectively, including sample size, dimensionality, minority class proportion, and data origin.

Our datasets vary substantially in the number of observations, the number of features, and the distribution of the target variable. For PD, the default rate differs substantially across portfolios, ranging from approximately 6.7% to 40.0% (mean $\approx$ 22.1%, median $\approx$ 22.4%). For the LGD datasets, the target variable exhibits a typical bimodal, zero-one-inflated distribution across most datasets, as reported in Loterman et al. [2012]. The bimodality of the LGD distribution arises due to, on the one hand, the high risk of losing the entire outstanding amount in the event of default for unsecured exposures, resulting in an LGD value equal to or close to 100%, while for secured loans, on the other hand, the full outstanding amount at time of default is often recovered, resulting in an LGD value equal to or close to 0%. Figure 1 visualizes the distribution of the target variable in the LGD datasets.

Table 1: Characteristics of the PD datasets. Minority % represents the positive class rate (default).

ID	Dataset name	Observations	Variables	Minority %	Source
PD1	GMSC (Give Me Some Credit)	150,000	10	6.7%	Kaggle
PD2	Taiwan Credit Card	30,000	23	22.1%	UCI
PD3	Vehicle Loan	233,154	35	21.7%	Kaggle (LTFS)
PD4	LendingClub (PD)	9,578	13	16.0%	LendingClub
PD5	Myhom	7,000	8	40.0%	Kaggle
PD6	Hackerearth	532,428	35	23.6%	Hackerearth
PD7	Cobranded	80,000	47	24.6%	Kaggle
PD8	German Credit	1,000	20	30.0%	UCI
PD9	Bank Status	100,000	16	22.6%	Proprietary
PD10	Thomas	1,225	14	26.4%	Thomas et al. [2002]
PD11	Loan Default	105,471	759	9.3%	Proprietary
PD12	Home Credit	307,511	120	8.1%	Kaggle
PD13	HMEQ	5,960	12	19.9%	SAS
PD14	MicroFinance	158,700	2,986	37.8%	Kozodoi et al. [2025]

Table 2: Characteristics of LGD datasets.

ID	Dataset name	Observations	Variables	Source
LGD1	HELOC	57,931	8	FICO
LGD2	Loss2	4,637	52	–
LGD3	AXA	2,545	2	AXA
LGD4	Base Model	762	202	–
LGD5	Base Modelisation	594	256	–
LGD6	Freddie Mac (LGD)	16,002	20	Freddie Mac
LGD7	LendingClub (LGD)	5,627	17	LendingClub

4.2 Data preprocessing

We apply a consistent preprocessing pipeline across all datasets to ensure comparability, implemented in TALENT [Liu et al., 2024], a unified benchmarking framework for tabular learning that standardizes data handling, method integration, and evaluation across a broad set of algorithms. All preprocessing parameters are learned exclusively from the training data of each cross-validation fold (cf. infra) and applied to the validation and test partitions of the corresponding cross-validation iteration, thereby avoiding data leakage. We impute missing values in numerical features with their median, and introduce a novel category level for missing values in categorical features. We subsequently encode these as integer indices for methods that support categorical inputs natively, or via one-hot encoding for methods that require real-valued inputs; the encoding policy is specified and enforced per method within TALENT. We apply normalization in a method-specific manner: all numerical variables are standardized to zero mean and unit variance, except for TFMs (TabPFN, TabPFNv2, TabPFN-Real, MITRA, TabICL), which operate on raw, non-normalized inputs.

4.3 Benchmarking protocol

We compare a diverse panel of competitive learners, ranging from classical statistical methods to state-of-the-art deep tabular architectures. For PD, we compare 29 different methods, including foundation models (TabPFN, TabPFNv2,

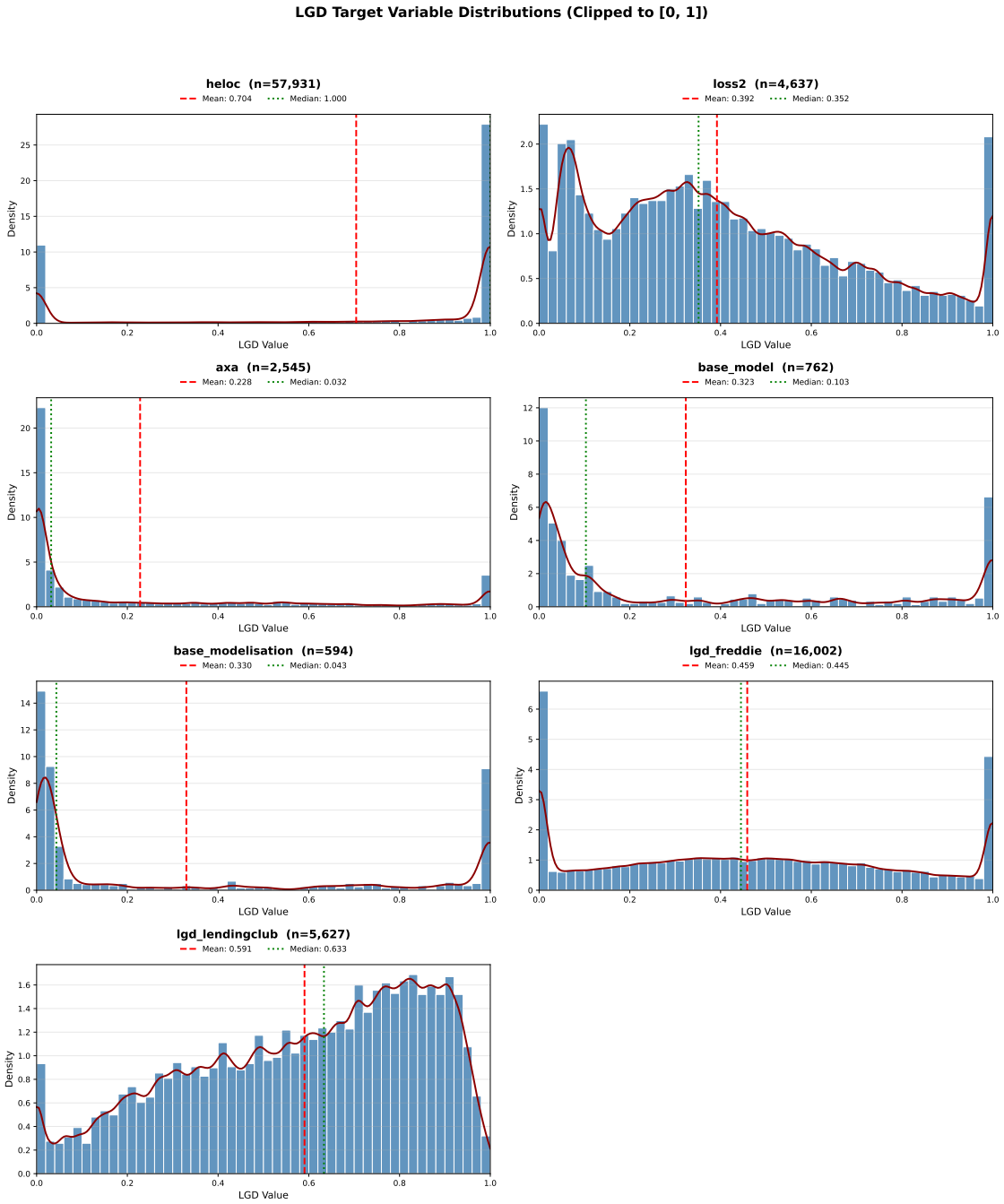

Figure 1: Distribution of the target variable in the LGD datasets.

TabPFN-Real, MITRA, and TabICL), several classical methods, particularly tree-based ensembles (XGBoost, LightGBM, CatBoost), and a wide array of deep learning models, including transformers (e.g., FT-Transformer, ExcelFormer) and specialized architectures like TabNet and TabM.

For LGD, the benchmark comprises 22 methods, including foundation models that support regression at the time of evaluation (TabPFNv2), classical regression baselines, gradient-boosting approaches, and advanced deep learning architectures such as ModernNCA, ResNet, and PTARL. Table 3 summarizes the full set of methods considered across both tasks.

Table 3: Overview of benchmarked methods for PD (classification) and LGD (regression).

Method Name	**PD**	**LGD**
Foundation Models		
TabPFN	✓	
TabPFNv2	✓	✓
TabPFN-Real	✓	
MITRA	✓	
TabICL	✓	
Tree Boosting		
CatBoost	✓	✓
LightGBM	✓	✓
XGBoost	✓	✓
Deep Tabular (Transformers)		
FT-Transformer (ftt)	✓	✓
AutoInt	✓	✓
ExcelFormer	✓	✓
AMFormer	✓	
T2G-Former	✓	✓
PTARL		✓
Deep Tabular (MLP & Specialized)		
MLP	✓	✓
ResNet		✓
SNN	✓	
RealMLP	✓	✓
MLP-PLR	✓	✓
DANets	✓	
SwitchTab	✓	
TabNet	✓	✓
DCN2	✓	✓
TabM	✓	✓
TANGOS	✓	✓
ModernNCA		✓
Classical ML		
Logistic Regression	✓	
Linear Regression		✓
K-Nearest Neighbors (KNN)	✓	✓
Random Forest	✓	✓
Support Vector Machine (SVM/SVR)	✓	✓
Naive Bayes	✓	
Nearest Class Mean (NCM)	✓	

Performance is evaluated on the held-out test partition within a five-fold cross-validation scheme. We tune hyperparameters on a validation split comprising 20% of the training fold by maximizing the area under the ROC curve (AUC) for PD models, and by minimizing the mean squared error (MSE) for LGD models. These objectives are used to select hyperparameters within each fold. Hyperparameter optimization is performed independently in each fold using the Optuna framework [Akiba et al., 2019] with up to 20 trials.

For PD modeling, we consider a diverse set of metrics that evaluate discrimination, calibration, and decision performance, including AUC, Gini, KS (Kolmogorov-Smirnov), Brier score, LogLoss, and Average Precision (PR-AUC). We also report threshold-based metrics, including Accuracy, Balanced Accuracy, F1-score, Precision, Recall, and the Matthews Correlation Coefficient (MCC). To compute these, we convert model-estimated PD probabilities to labels using an optimal threshold determined by maximizing the F1-score on the validation set.

For LGD prediction, we measure forecast accuracy in terms of R^2, MSE, RMSE, and MAE, as well as MedAE, Max Absolute Error, Pearson and Spearman correlation coefficients, and Explained Variance. All LGD predictions are

clipped to the $[0, 1]$ interval before metric calculation to ensure they remain within the logical bounds of loss given default.[2]

4.4 Statistical analysis

Principles for comparing learners across datasets in a statistically sound way have been established by Demšar [2006]. Recent machine learning benchmarks advocate modifications, involving a pairwise comparison of learners' performance using a Student's t-test (e.g., comparing AUCs over datasets or cross-validation folds) followed by Holm's step-down adjustment to control the family-wise error rate at $\alpha = 0.05$ in multiple comparisons [Ye et al., 2024, Liu et al., 2024]. We adopt a similar approach, yet we replace the t-test with the Wilcoxon signed-rank test [Wilcoxon, 1945], a non-parametric alternative that does not assume normality of performance differences and accounts for both the direction and magnitude of differences across datasets.

We further summarize the likelihood that a given method "wins" a dataset using the probability of achieving maximum accuracy (PAMA), computed as the relative frequency of a learner to achieve the top performance (in a given metric) across all cross-validation splits and datasets [Fernández-Delgado et al., 2014]. Whereas the classic statistical tests assess whether performance differences between learners are significant, PAMA highlights the practical dominance frequency, thereby complementing the analysis of differences in learners' performance and their statistical significance.

5 Results and discussion

The empirical results comprise estimates of predictive performance across learning algorithms, datasets, and cross-validation folds, using various evaluation metrics.

5.1 PD modeling

Figure 2 reports the average performance of classification methods in terms of AUC, across the five folds of all PD datasets. A first observation is that TabICL, one of the foundation models considered here, achieves the best performance overall. Although the observed performance differences are small in absolute terms, it is notable that—without any training or hyperparameter optimization—a TFM outperforms widely credited boosting-based ensembles, which may be considered a de facto standard in credit risk modeling.

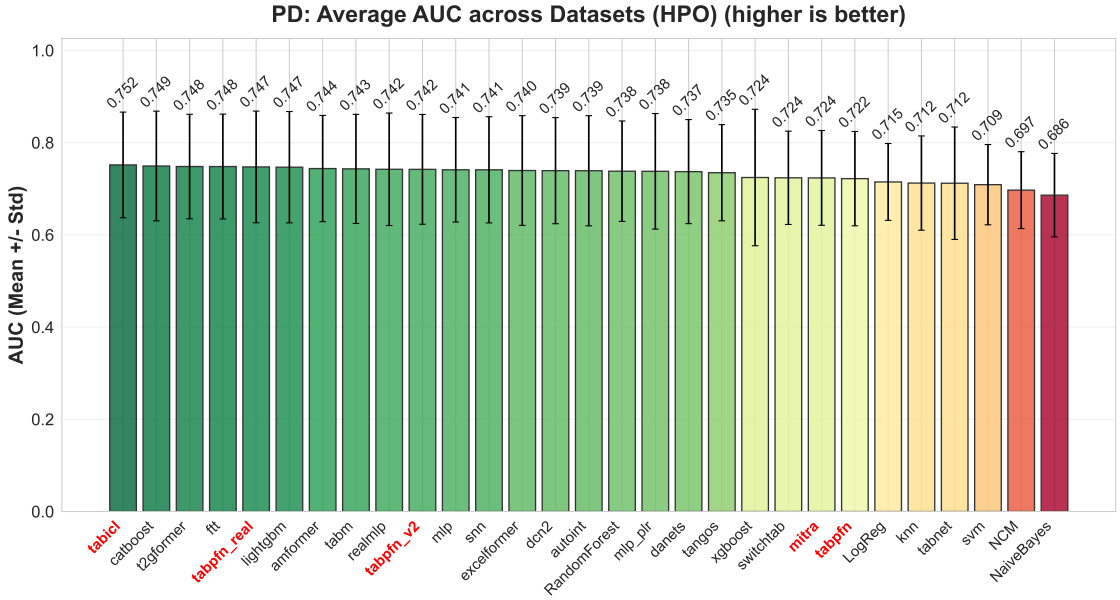

Figure 2: Average AUC across PD datasets.

[2]Only one of our datasets exhibits raw LGD values outside the zero-one range.

A second observation from Figure 2 concerns the small magnitude of performance differences for a set of well-performing learning algorithms, including GBM-type approaches, PFNs, and Random Forests. We observe a similar trend for other metrics of classification performance, as reported in Table 4.

A third observation concerns the performance differences among TFMs. Whereas TabICL is best in class, TabPFN-Real also ranks among the stronger methods. TabPFNv2 performs in the middle of the distribution, while MITRA and TabPFN rank among the weaker TFMs for default prediction.

Table 4 summarizes the full set of metrics utilized in the PD classification benchmark.

Table 4: Average performance of all methods on the PD (classification) benchmark across 14 datasets and 5 folds under HPO conditions. Arrows indicate optimization direction (↓ lower is better; all others higher is better). The best value in each column is shown in **bold**.

Method	Discrimination				Calibration		Classification (threshold-based)					
	AUC	Gini	KS	AP	Brier ↓	LogLoss ↓	Acc	Bal-Acc	F1	Precision	Recall	MCC
Foundation models												
TabICL	**0.7517**	**0.5035**	**0.3987**	0.4933	0.1619	0.5081	0.7125	0.6734	0.5047	0.4368	0.6547	0.3265
TabPFNv2	0.7421	0.4843	0.3867	0.4891	0.1320	**0.4119**	0.6970	0.6685	0.5025	0.4364	0.6666	0.3201
TabPFN-Real	0.7474	0.4948	0.3972	**0.4949**	0.1612	0.5067	0.6975	0.6732	**0.5092**	0.4426	0.6744	**0.3296**
MITRA	0.7236	0.4471	0.3466	0.4563	0.1391	0.4329	0.6449	0.6449	0.4702	0.3944	0.6801	0.2729
TabPFN	0.7220	0.4440	0.3433	0.4498	0.1407	0.4378	0.6778	0.6491	0.4716	0.3926	0.6510	0.2758
Gradient boosting												
CatBoost	0.7494	0.4988	0.3982	0.4923	0.1331	0.4146	**0.7129**	**0.6746**	0.5074	0.4425	0.6540	0.3288
LightGBM	0.7468	0.4936	0.3956	0.4918	0.1318	0.4187	0.6919	0.6714	0.5064	0.4351	0.6826	0.3216
XGBoost	0.7244	0.4489	0.3595	0.4781	0.1343	0.4319	0.6416	0.6597	0.5010	0.4311	**0.7230**	0.3051
Classical ML												
Logistic Reg.	0.7148	0.4296	0.3330	0.4294	0.1428	0.4432	0.6700	0.6429	0.4599	0.3760	0.6475	0.2609
Random Forest	0.7381	0.4763	0.3726	0.4765	0.1358	0.4226	0.6895	0.6588	0.4868	0.4160	0.6567	0.2976
Naive Bayes	0.6861	0.3723	0.2956	0.3974	0.3069	2.2809	0.6124	0.6173	0.4317	0.3387	0.6768	0.2124
KNN	0.7124	0.4248	0.3370	0.4438	0.1385	0.5808	0.6573	0.6462	0.4753	0.3948	0.6825	0.2749
SVM	0.7089	0.4178	0.3268	0.4239	0.1434	0.4453	0.6601	0.6387	0.4557	0.3712	0.6508	0.2530
NCM	0.6971	0.3941	0.3111	0.4094	0.2186	0.6258	0.6471	0.6321	0.4498	0.3620	0.6618	0.2390
Deep learning												
MLP	0.7412	0.4824	0.3829	0.4816	0.1332	0.4402	0.6928	0.6682	0.5002	0.4334	0.6686	0.3191
MLP-PLR	0.7379	0.4758	0.3844	0.4865	0.1344	0.4688	0.6971	0.6679	0.4995	0.4347	0.6553	0.3181
SNN	0.7411	0.4823	0.3829	0.4817	0.1370	0.4550	0.7072	0.6675	0.4970	0.4320	0.6427	0.3154
RealMLP	0.7423	0.4846	0.3869	0.4884	0.1656	0.5167	0.7085	0.6708	0.5033	**0.4461**	0.6490	0.3259
TabNet	0.7120	0.4241	0.3363	0.4483	0.1392	0.4342	0.6617	0.6431	0.4728	0.4057	0.6544	0.2716
SwitchTab	0.7238	0.4476	0.3501	0.4540	0.1390	0.4324	0.6799	0.6532	0.4772	0.4035	0.6561	0.2862
FTT	0.7482	0.4964	0.3942	0.4895	**0.1314**	0.4129	0.7000	0.6734	0.5041	0.4342	0.6701	0.3248
T2G-Former	0.7483	0.4965	0.3930	0.4894	0.1318	0.4159	0.7073	0.6737	0.5056	0.4375	0.6615	0.3274
DCN2	0.7393	0.4786	0.3814	0.4775	0.1338	0.4425	0.6963	0.6673	0.4996	0.4308	0.6641	0.3149
ExcelFormer	0.7396	0.4792	0.3809	0.4813	0.1333	0.4162	0.6945	0.6646	0.4960	0.4294	0.6578	0.3114
AutoInt	0.7391	0.4782	0.3791	0.4816	0.3200	5.3695	0.8112	0.6092	0.3228	0.5500	0.2715	0.2723
DANets	0.7372	0.4745	0.3739	0.4778	0.1354	0.4428	0.6921	0.6648	0.4935	0.4328	0.6557	0.3127
TANGOS	0.7349	0.4697	0.3681	0.4737	0.1362	0.4389	0.6847	0.6602	0.4863	0.4184	0.6611	0.3024
AMFormer	0.7439	0.4877	0.3879	0.4885	0.1330	0.4243	0.7063	0.6706	0.4992	0.4359	0.6510	0.3206

5.2 LGD modeling

Next, we examine the performance of TFMs for LGD prediction. Figure 3 reports the corresponding results in terms of the average R^2 measure across datasets. Results for (R)MSE and MAE are similar and reported in Table 5.

A first observation from Figure 3 is that, as for PD modeling, the best performance is, on average, achieved by a TFM, i.e., TabPFNv2. LGD modeling is a challenging task, which prior work often associates with the peculiar shape of the loss distributions, as observed in Figure 1. Therefore, it is notable to find that also for LGD prediction, despite the challenge involved in this task, TFMs outperform the state-of-the-art methods in the field. This confirms the occurrence of a paradigm shift, both for classification and regression, as highlighted in the previous section.

TabPFNv2 supports regression by discretizing the continuous target into a piecewise-constant probability distribution and training the model to predict this distribution, effectively framing regression as an ordinal classification problem. This approach can be considered conceptually suitable for LGD distributions, as it provides the required flexibility to fit widely varying and non-normal distributions, and is confirmed by our experiments to perform well on the LGD data employed here.

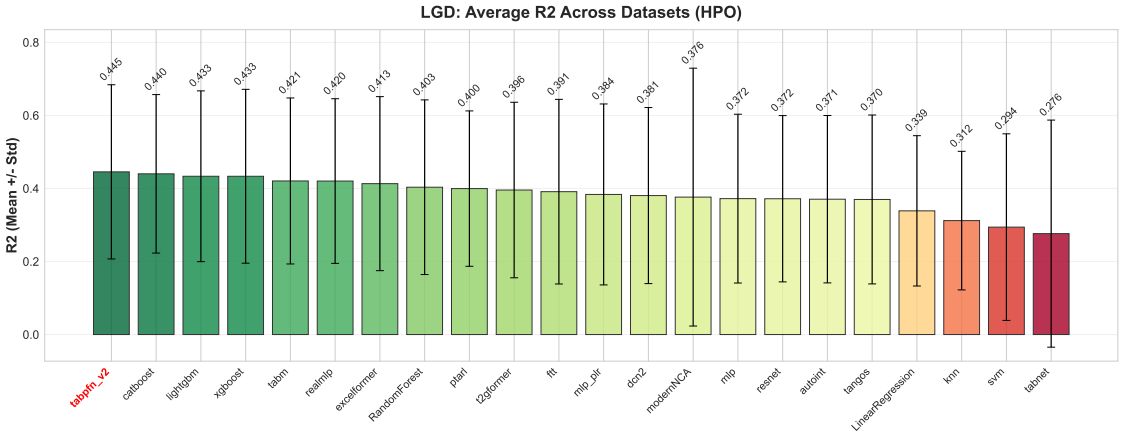

Figure 3: Average R^2 across LGD datasets.

A second observation is that, as for PD estimation, several methods are close in average performance, although more variability is observed compared to the PD results.

A third observation is that our experiments confirm the strong performance of boosting ensemble methods, across PD and LGD modeling tasks, as reported in the literature.

The full set of metrics utilized in the LGD regression benchmark is summarized in Table 5.

5.3 Statistical analysis of PFN performance

To contextualize performance differences across learners and datasets, we employ three complementary statistical analyses, following standard practice in the benchmarking literature [Gunnarsson et al., 2021, Lessmann et al., 2015, Demšar, 2006]. We begin with the PAMA analysis to identify learners that achieve top performance most frequently, then use the Friedman test to formally establish whether performance differences are globally non-random, and finally use pairwise Wilcoxon signed-rank tests with Win/Loss ratios to probe the magnitude and significance of the pairwise differences. All analyses are based on the AUC for PD models and R^2 for LGD models.

A natural first question that emerges is the following: across all datasets, which learner most often yields the best performance? The Probability of Achieving Maximal Accuracy (PAMA) analysis [Fernández-Delgado et al., 2014] provides an answer to this question by computing, across all learners, the fraction of fold-level observations in which a learner achieves the highest score. This provides an intuitive, rank-based summary of competitiveness that is robust to outlying datasets and does not require us to make any assumptions about the distribution of performance differences.

Figures 4 and 5 show the PAMA results for the PD and LGD benchmarks, respectively. For PD, the performance of TabICL stands out. It achieves the highest AUC in 25.7% of fold-level observations (18 out of 70), followed by CatBoost at 15.7%. Notably, foundation models as a group achieve the best performance in 44.3% of all folds, indicating their collective competitiveness, even when no single model dominates. For LGD regression, TabPFNv2 is the only foundation model included. It achieves the highest R^2 in 45.7% of folds (16 out of 35), placing it well ahead of all competitors on this benchmark too.

The PAMA analysis is descriptive and does not assess whether observed performance differences could have arisen by chance. To establish this formally, we apply the Friedman test with Iman–Davenport correction [Friedman, 1940, Iman and Davenport, 1980, Demšar, 2006], a non-parametric test that assesses whether the overall ranking of learners across datasets deviates significantly from what would be expected under the null hypothesis that all methods perform equally. For the PD benchmark ($N = 14$ datasets, $k = 29$ methods), $F_F = 7.01$ ($p = 1.11 \times 10^{-16}$); for the LGD benchmark ($N = 7$ datasets, $k = 22$ methods), $F_F = 3.84$ ($p = 1.12 \times 10^{-6}$). Both tests strongly reject the null hypothesis, confirming that the performance differences observed in the PAMA analysis are not attributable to chance and that post-hoc pairwise comparisons are warranted.

Having established that global differences are statistically significant, we turn to pairwise comparisons to identify which specific pairs of learners differ statistically significantly in terms of performance. For each pair, we test the null hypothesis that the median performance difference is zero using the Wilcoxon signed-rank test [Wilcoxon, 1945], which accounts for both the direction and magnitude of differences across datasets. p-values are adjusted for multiple

Table 5: Average performance of all methods on the LGD (regression) benchmark across 7 datasets and 5 folds under HPO conditions. Arrows indicate optimisation direction (↓ lower is better; all others higher is better). The best value in each column is shown in **bold**.

Method	Error / Fit						Correlation & Variance		
	R^2	MSE ↓	RMSE ↓	MAE ↓	MedAE ↓	MaxErr ↓	Pearson	Spearman	Expl. Var
Foundation model									
TabPFNv2	**0.4455**	0.0704	**0.2522**	0.1889	0.1410	0.8890	**0.6596**	**0.6094**	**0.4586**
Gradient boosting									
CatBoost	0.4401	**0.0702**	0.2540	0.1885	0.1369	0.9052	0.6499	0.6023	0.4435
LightGBM	0.4334	0.0706	0.2545	0.1879	0.1376	0.9076	0.6434	0.5969	0.4407
XGBoost	0.4334	0.0704	0.2540	0.1911	0.1442	0.8998	0.6393	0.5982	0.4380
Classical ML									
Linear Reg.	0.3387	0.0832	0.2777	0.2139	0.1725	0.9366	0.5673	0.5359	0.3410
Random Forest	0.4035	0.0748	0.2619	0.1962	0.1423	0.8925	0.6178	0.5710	0.4100
KNN	0.3120	0.0839	0.2815	0.2216	0.1778	**0.8488**	0.5583	0.5145	0.3185
SVM	0.2942	0.0885	0.2856	0.2117	0.1515	0.9238	0.5548	0.5302	0.3187
Deep learning									
MLP	0.3722	0.0798	0.2700	0.1898	0.1200	0.9261	0.6096	0.5704	0.3814
MLP-PLR	0.3836	0.0782	0.2669	0.1787	0.1030	0.9365	0.6231	0.5806	0.3999
RealMLP	0.4203	0.0730	0.2588	0.1844	0.1268	0.9212	0.6369	0.5958	0.4261
TabM	0.4206	0.0727	0.2581	0.1852	0.1290	0.9265	0.6384	0.5914	0.4255
TabNet	0.2763	0.0942	0.2902	0.2278	0.1772	0.9363	0.4799	0.4543	0.2839
FTT	0.3912	0.0774	0.2649	0.1801	0.1068	0.9362	0.6314	0.5907	0.4057
T2G-Former	0.3957	0.0768	0.2644	0.1811	0.1089	0.9419	0.6354	0.5876	0.4106
DCN2	0.3805	0.0793	0.2685	0.1945	0.1409	0.9388	0.6049	0.5657	0.3861
AutoInt	0.3706	0.0800	0.2701	0.1909	0.1234	0.9326	0.6105	0.5679	0.3882
ExcelFormer	0.4132	0.0739	0.2597	0.1843	0.1270	0.9361	0.6234	0.5780	0.4212
ResNet	0.3719	0.0800	0.2704	0.1917	0.1270	0.9414	0.6095	0.5617	0.3807
TANGOS	0.3699	0.0798	0.2705	0.1898	0.1195	0.9288	0.6047	0.5652	0.3810
Specialised / other									
ModernNCA	0.3763	0.0789	0.2616	**0.1718**	**0.0968**	0.9356	0.6202	0.5884	0.3904
PTARL	0.3997	0.0763	0.2645	0.1986	0.1497	0.9070	0.6193	0.5808	0.4027

comparisons using Holm's step-down method [Holm, 1979, Garcia and Herrera, 2008], applied over all 406 PD and 231 LGD pairs, respectively. Figures 6 and 7 display the results where cell text reports the Win/Loss (W/L) ratio, and a trailing asterisk (*) marks pairs with a statistically significant difference ($p \leq 0.05$, Holm-corrected).

For the PD benchmark, 22 out of 406 pairwise comparisons are statistically significant after Holm correction. For LGD, none of the 231 pairwise comparisons reach significance. The limited number of significant pairwise differences is not surprising and reflects two structural features of the benchmarking setup. First, the number of datasets constrains statistical power: the Wilcoxon signed-rank test requires sufficient paired observations to reliably detect non-zero median differences, and with $N = 14$ PD datasets and only $N = 7$ LGD datasets, this power is inherently limited. For LGD in particular, the minimum achievable two-sided p-value for a single comparison is $2/2^7 \approx 0.016$, and after Holm correction for 231 simultaneous tests, this threshold rises substantially, making it virtually impossible to find a difference in performance between any individual pair to be statistically significant. Second, the relatively small magnitude of performance differences among the top methods — which can be clearly observed in Figures 2 and 3 — means that even where directional evidence is consistent, differences may not be large enough to cross the significance threshold after multiple-comparison correction. These results, therefore, do not contradict the omnibus Friedman test, which has substantially greater power to detect global heterogeneity than individual pairwise tests have to localize it.

Taken together, the three analyses paint a consistent and favorable picture for TFMs. The PAMA analysis shows that foundation models collectively achieve the highest performance in 44.3% of PD folds, with TabICL leading individually, and that TabPFNv2 alone wins 45.7% of LGD folds. The Friedman test confirms that these differences in method rankings are not attributable to chance. Overall, these results provide consistent support for the conclusion that TFMs are competitive challengers to established methods in credit risk prediction.

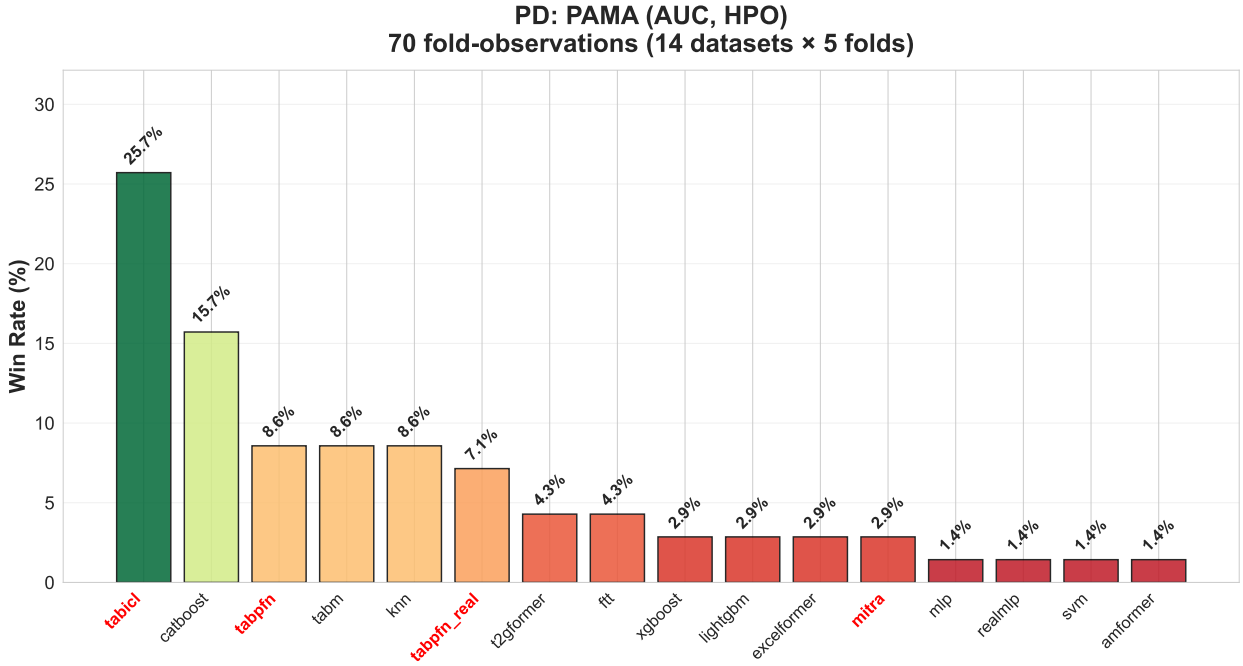

Figure 4: Probability of maximal AUC (PAMA) analysis for PD datasets.

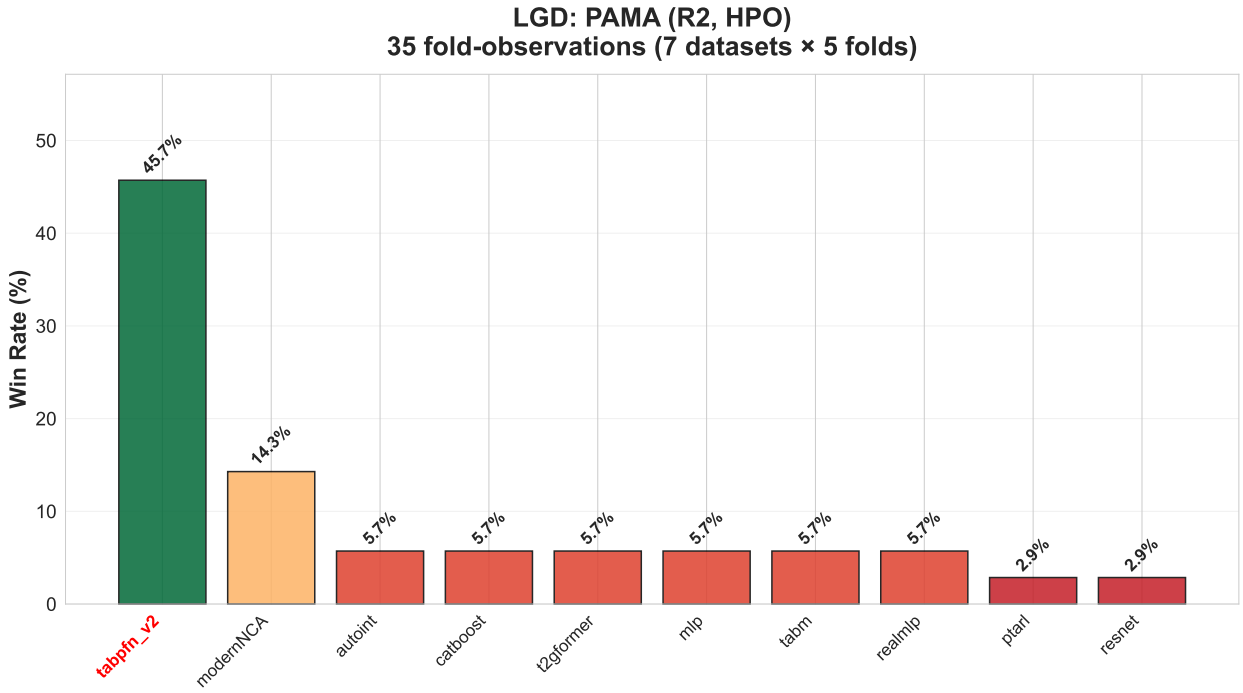

Figure 5: Probability of maximal R^2 (PAMA) analysis for LGD datasets.

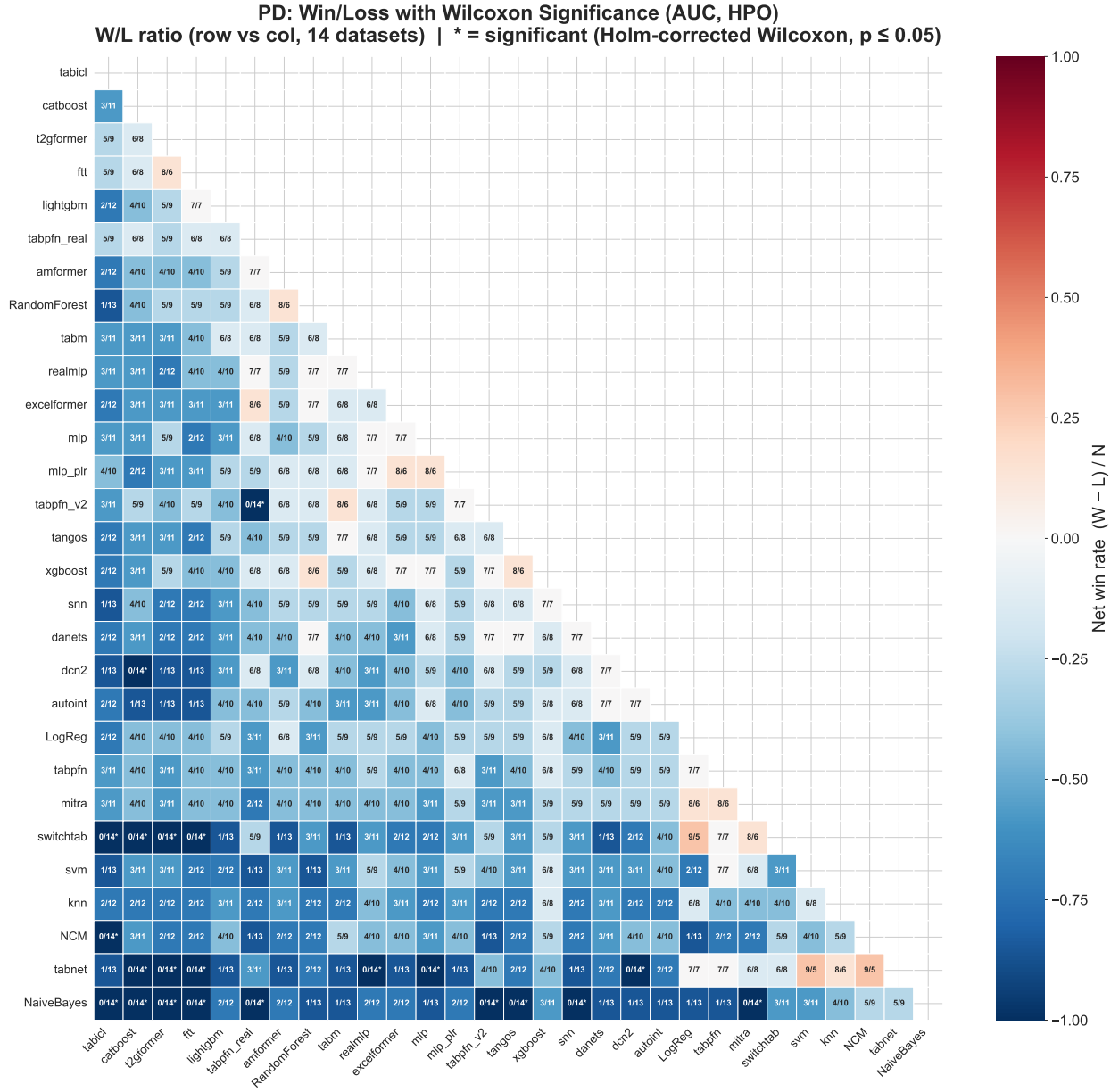

Figure 6: Win/Loss ratio matrix for the PD benchmark ($N = 14$ datasets). Cell text gives the W/L count from the row method's perspective; cell color encodes the net win rate $(W - L)/N$ - blue: row method wins more often, red: row method loses more often. An asterisk (*) denotes a statistically significant difference (Holm-corrected Wilcoxon, $p \leq 0.05$).

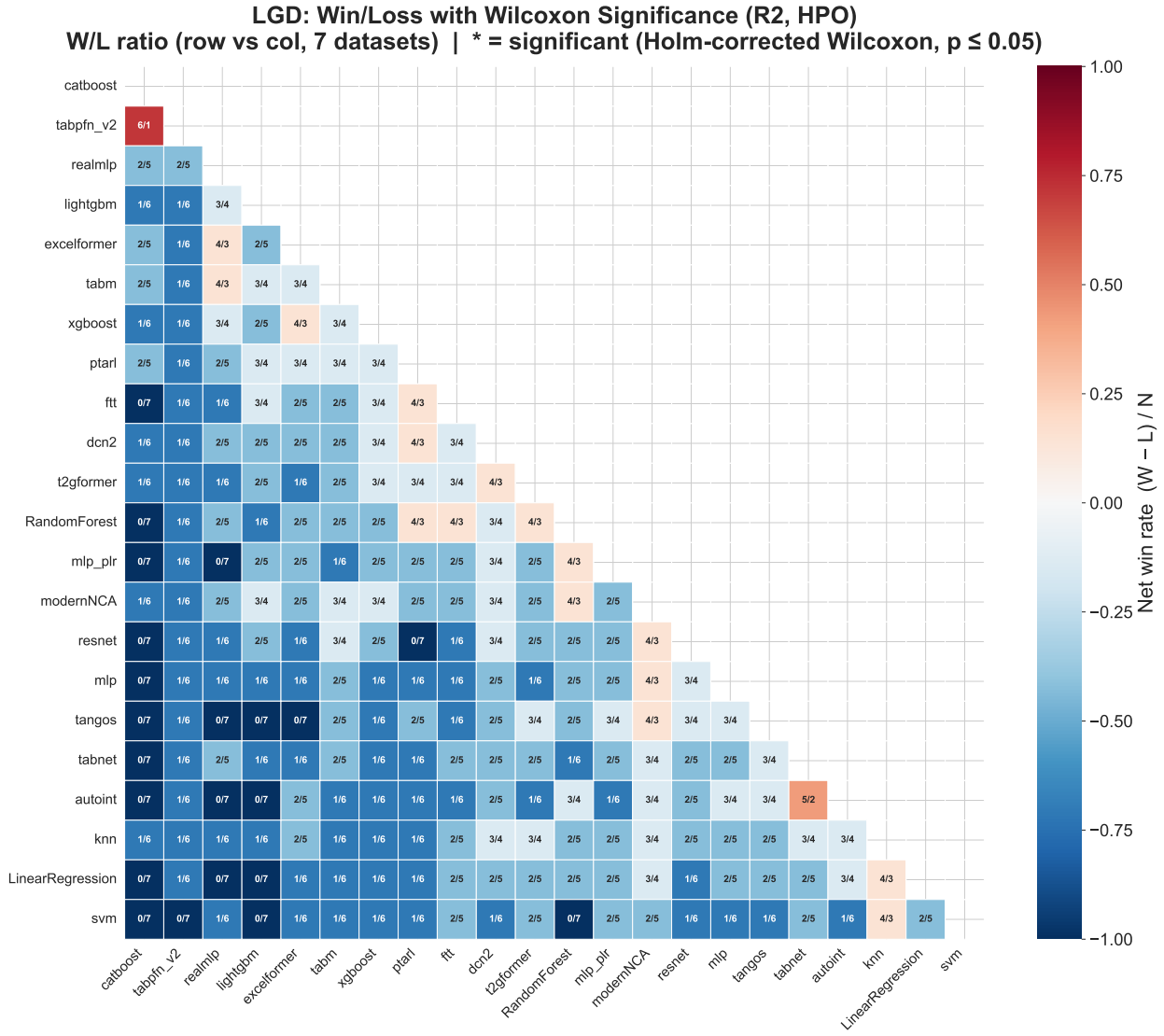

Figure 7: Win/Loss ratio matrix for the LGD benchmark ($N = 7$ datasets). Cell text gives the W/L count from the row method's perspective; cell color encodes the net win rate $(W - L)/N$ - blue: row method wins more often, red: row method loses more often. An asterisk (*) denotes a statistically significant difference (Holm-corrected Wilcoxon, $p \leq 0.05$).

5.4 Dataset size

A particular feature of TFMs that may be appealing towards credit risk modeling is their potential use for small portfolios, resulting in small datasets.

To assess the ability of TFMs for handling small datasets, we next analyze the relation between learner performance, more specifically, the rank of a given method, and dataset size (in number of observations) using Spearman's rank correlation coefficient. Figures 8 and 9 report the corresponding results for PD and LGD modeling, respectively. The positive correlation observed for TFMs indicates that they rank higher (i.e., show relatively worse performance compared to other methods) on larger datasets. In other words, for smaller datasets, TFMs perform relatively better. We observe this tendency for PD and LGD modeling, supporting our hypothesis that TFMs are particularly appealing for small portfolios, and, more generally, small data settings.

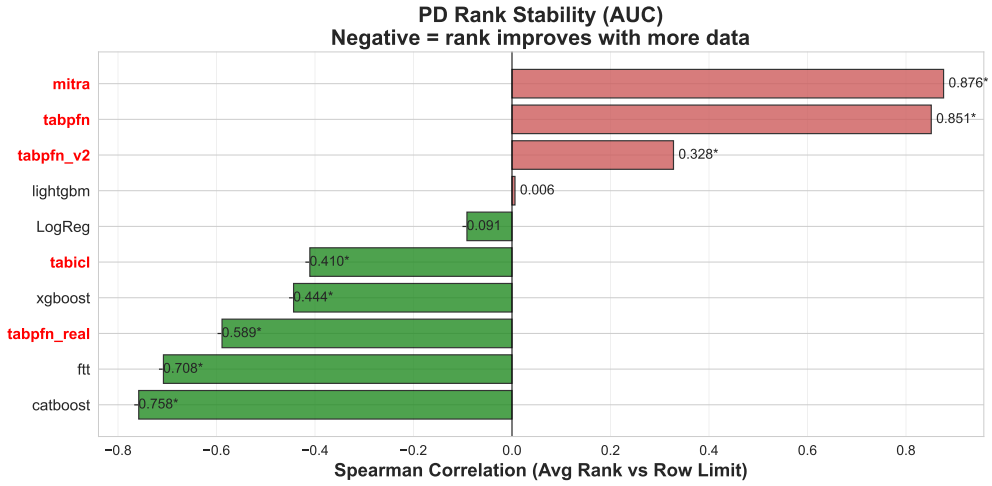

Figure 8: Spearman rank correlation between a method's performance rank and PD dataset size (number of observations) across the 14 PD datasets. A positive correlation indicates that the method's rank number increases (i.e., relative performance deteriorates) as dataset size grows, whereas a negative correlation indicates stronger relative performance on larger datasets. Ranks are based on AUC and computed per dataset across all 29 methods. TFMs exhibit positive correlations, consistent with the hypothesis that they provide greater relative benefit in small-data settings.

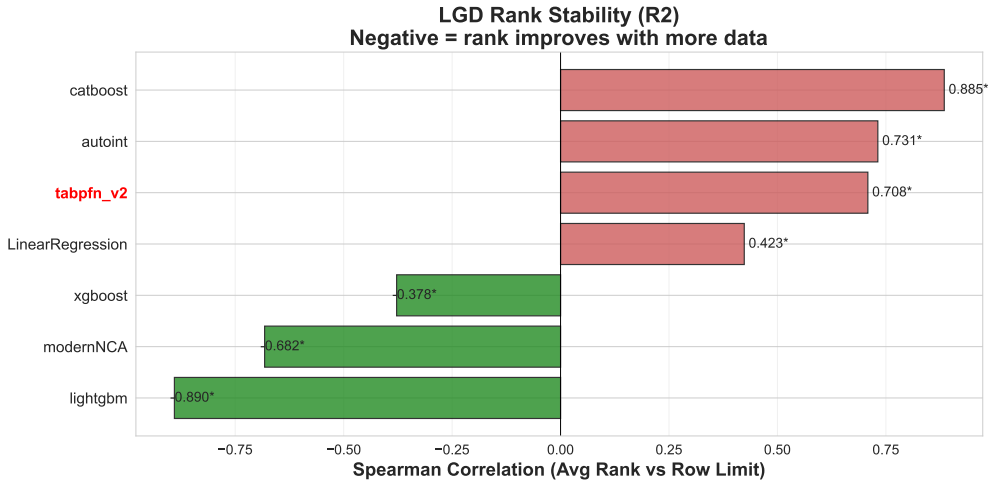

Figure 9: Spearman rank correlation between a method's performance rank and LGD dataset size (number of observations) across the 7 LGD datasets. Ranks are based on R^2 and computed per dataset across all 22 methods. As with PD, the positive correlation observed for TabPFNv2 indicates relatively stronger performance on smaller datasets.

A second experiment to assess the impact of dataset size on the performance of TFMs evaluates the evolution of average performance as the number of randomly sampled observations increases (from 500 to 15,000) for datasets that include more than 15,000 observations. Figures 10 and 11 report the results for PD and LGD modeling, respectively, and include the so-called learning curves for the TFMs and a selected set of baseline methods, i.e., Logistic Regression, CatBoost, FTT, LightGBM, and XGBoost. A general—unsurprising—tendency is that performance increases with dataset size across methods. The performance of TFMs is substantially above the performance of the baseline methods for smaller numbers of observations, and the difference in performance decreases as the size of the dataset increases. For LGD modeling, similar results are found, as shown in Figure 11. Whereas the baseline method, XGBoost, performs substantially worse than TabPFNv2 for small dataset sizes, it eventually overtakes TabPFNv2 when dataset sizes exceed 8,000 observations. It is unclear why TabPFNv2's performance declines at this point. Further research may focus on confirming and explaining this behavior.

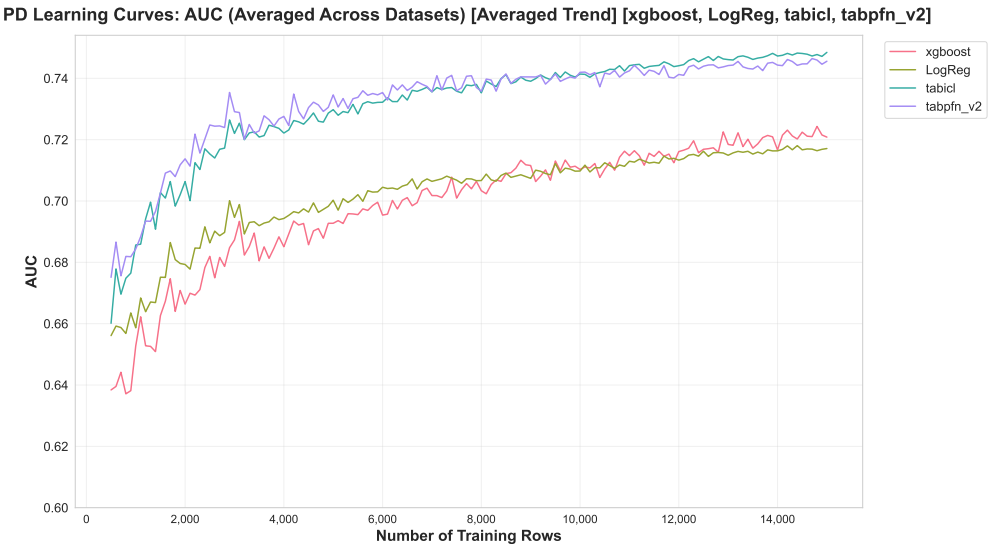

Figure 10: PD learning curves: average AUC across PD datasets with more than 15,000 observations as the number of randomly sampled training observations increases from 500 to 15,000. Curves are shown for TabICL, TabPFN, Logistic Regression, and XGBoost.

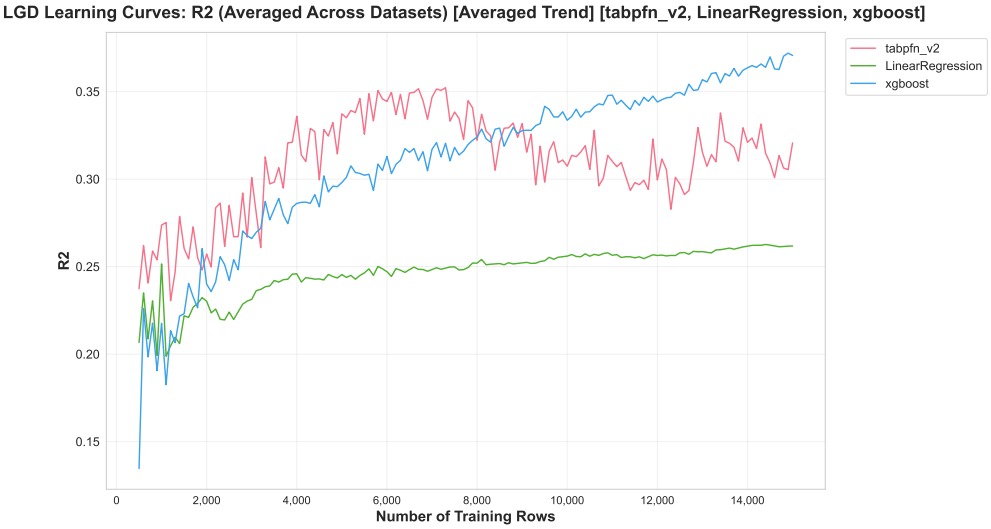

Figure 11: LGD learning curves: average R^2 across LGD datasets with more than 15,000 observations as the number of randomly sampled training observations increases from 500 to 15,000. Curves are shown for TabPFNv2, Linear Regression, and XGBoost.

6 Conclusions

This paper provides an extensive empirical evaluation of TFMs for credit risk prediction, benchmarking them against a broad panel of classical, tree-based, and deep learning methods across 14 PD and 7 LGD datasets. The results consistently favor foundation models. For PD classification, TabICL achieves the highest average rank and leads the PAMA analysis with the highest AUC in 25.7% of fold-level observations, while foundation models as a group attain top performance in 44.3% of all folds. For LGD regression, TabPFNv2, the only foundation model evaluated, achieves the highest R^2 in 45.7% of folds, placing it ahead of all tuned competitors, including GBMs. Statistical analysis via the Friedman test confirms that these differences in learner rankings are not attributable to chance, and Win/Loss comparisons show that top-ranked methods, including foundation models, significantly outperform weaker baselines.

A noteworthy finding concerns the relationship between dataset size and relative performance. TFMs show a consistent advantage on smaller datasets, with their relative performance deteriorating as dataset size grows. This finding directly supports the use case motivating much of the interest in foundation models for credit risk: small-data settings such as SME lending, low-default portfolios, and specialized corporate segments where conventional methods struggle due to limited training signals. The learning curve analysis confirms this pattern, showing that TabPFNv2 leads clearly at small sample sizes before tuned GBMs gradually close the gap as data availability increases.

From a practical perspective, PFNs offer several advantages. They eliminate the need for task-specific hyperparameter tuning and retraining, improving time-to-model and reducing operational costs. Their generality supports consistent modeling across risk parameters, which can simplify governance and may ease supervisory review. Potential benefits from a customer perspective include more stable, consistent decisions across portfolios and institutions and, with appropriate safeguards, opportunities to reduce historical biases by leveraging priors learned from broad synthetic data. Importantly, PFNs can also complement established risk modeling practices, facilitating risk prediction in early stages of the product lifecycle where available data is very limited. A later migration to, for example, a logit- or GBM-based score once a sufficient amount of (repayment) data has been gathered is straightforward, potentially reducing dependency on external credit rating agencies for new product scoring.

Important avenues for future work remain. We plan to include established LGD econometric baselines (e.g., beta regression and two-part models), and evaluate profit-based measures such as EMPC more broadly. We also aim to analyze PFN behavior under specific credit risk challenges, including low-default portfolios, reject inference framed as missing data, distribution shifts, and fairness constraints. Finally, interpretability remains essential: we will compare feature attributions derived from PFN and GBM pipelines using SHAP and alternatives such as XPER [Hué et al., 2025].

References

Johannes Schneider, Christian Meske, and Pauline Kuss. Foundation models: A new paradigm for artificial intelligence. *Business & Information Systems Engineering*, 66(2):221–231, 2024.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European journal of operational research*, 247(1):124–136, 2015.

Björn Rafn Gunnarsson, Seppe Vanden Broucke, Bart Baesens, María Óskarsdóttir, and Wilfried Lemahieu. Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305, 2021.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90, 2022a.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637 (8045):319–326, 2025a.

María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, and Bart Baesens. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74:26–39, 2019.

Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, 2022.

Darie Moldovan. Algorithmic decision making methods for fair credit scoring. *IEEE Access*, 11:59729–59743, 2023.

Bart Baesens, Daniel Roesch, and Harald Scheule. *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons, 2016.

Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

Gert Loterman, Iain Brown, David Martens, Christophe Mues, and Bart Baesens. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28(1):161–170, 2012.

João A. Bastos and Sara M. Matos. Explainable models of credit losses. *European Journal of Operational Research*, 301(1):386–394, 2022. ISSN 0377-2217. doi: 10.1016/j.ejor.2021.11.009.

Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, 39(3):3446–3453, 2012.

Ana Isabel Marqués, Vicente García, and José Salvador Sánchez. On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7):1060–1070, 2013.

Zhongyi Wang, Yuhang Tian, Sihan Li, and Jin Xiao. A secure cross-silo collaborative method for imbalanced credit scoring. *European Journal of Operational Research*, 326(2):357–373, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.04.020.

Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021. doi: 10.1016/j.eswa.2021.114582.

John Banasik and Jonathan Crook. Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3):1582–1594, 2007.

Nikita Kozodoi, Stefan Lessmann, Morteza Alamgir, Luis Moreira-Matias, and Konstantinos Papakonstantinou. Fighting sampling bias: A framework for training and evaluating credit scoring models. *European Journal of Operational Research*, 324(2):616–628, 2025.

Tony Bellotti and Jonathan Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1):171–182, 2012.

Viani B Djeundje, Jonathan Crook, and Galina Andreeva. The devil in the details: Dynamic prediction of loan portfolio profitability with macroeconomic drivers through multi-state modelling. *European Journal of Operational Research*, 2025.

Walter Distaso, Francesco Roccazzella, and Frédéric Vrins. Business cycle and realized losses in the consumer credit industry. *European Journal of Operational Research*, 323(3):1024–1039, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2024.12.026.

Lore Dirick, Tony Bellotti, Gerda Claeskens, and Bart Baesens. Macro-economic factors in credit risk calculations: Including time-varying covariates in mixture cure models. *Journal of Business and Economic Statistics*, 37(1): 40–53, 2019. ISSN 0735-0015. doi: 10.1080/07350015.2016.1260471.

Lyn C Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172, 2000.

Tony Van Gestel, Bart Baesens, Peter Van Dijcke, Johan Suykens, Joao Garcia, and Thomas Alderweireld. Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *Journal of credit Risk*, 1(4), 2005.

Xiao Yao, Jonathan Crook, and Galina Andreeva. Support vector regression for loss given default modelling. *European Journal of Operational Research*, 240(2):528–538, 2015. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2014.06.043.

Hui Cheng, Cuiqing Jiang, Zhao Wang, and Xiaoya Ni. Multi-view locally weighted regression for loss given default forecasting. *International Journal of Forecasting*, 41(1):290–306, 2025. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2024.05.006.

Christophe Hurlin, Jérémy Leymarie, and Antoine Patin. Loss functions for loss given default model comparison. *European Journal of Operational Research*, 268(1):348–360, 2018a. ISSN 0377-2217. doi: 10.1016/j.ejor.2018.01.020.

Raffaella Calabrese and Luca Zanin. Modelling spatial dependence for loss given default in peer-to-peer lending. *Expert Systems with Applications*, 192:116295, 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.116295.

Guangyou Zhou, Yijia Zhang, and Sumei Luo. P2p network lending, loss given default and credit risks. *Sustainability*, 10(4):1010, 2018. ISSN 2071-1050.

S. D. Tomarchio and A. Punzo. Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(4):1247–1266, 2019. doi: 10.1111/rssa.12466.

Wojciech Starosta. Loss given default decomposition using mixture distributions of in-default events. *European Journal of Operational Research*, 292(3):1187–1199, 2021. ISSN 0377-2217. doi: 10.1016/j.ejor.2020.11.034.

Monika Papouskova and Petr Hajek. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118:33–45, 2019. ISSN 0167-9236. doi: 10.1016/j.dss.2019.01.002.

Lyn C. Thomas, Anna Matuszyk, Mee Chi So, Christophe Mues, and Angela Moore. Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research*, 249(2):476–486, 2016. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2015.09.013.

Morne Joubert, Tanja Verster, Helgard Raubenheimer, and Willem D. Schutte. Adapting the default weighted survival analysis modelling approach to model ifrs 9 lgd. *Risks*, 9(6):103, 2021. ISSN 2227-9091. doi: 10.3390/risks9060103.

Cuiqing Jiang, Wang Lu, Zhao Wang, and Yong Ding. Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. *Expert Systems with Applications*, 213:118878, 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.118878.

Anthony Bellotti, Damiano Brigo, Paolo Gambetti, and Frédéric Vrins. Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, 37(1):428–444, 2021. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2020.06.009.

Thomas Verbraken, Cristián Bravo, Richard Weber, and Bart Baesens. Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238 (2):505–513, 2014. ISSN 0377-2217. doi: 10.1016/j.ejor.2014.04.001.

John Martin, Mali Abdollahian, Sona Taheri, and David Akman. A novel financial performance metric to minimize misclassification costs in model selection. *Annals of Operations Research*, 2025. ISSN 1572-9338. doi: 10.1007/s10479-025-06514-x.

Franco Garrido, Wouter Verbeke, and Cristián Bravo. A robust profit measure for binary classification model evaluation. *Expert Systems with Applications*, 92:154–160, 2018. ISSN 0957-4174. doi: 10.1016/j.eswa.2017.09.045.

Christophe Hurlin, Jérémy Leymarie, and Antoine Patin. Loss functions for loss given default model comparison. *European Journal of Operational Research*, 268(1):348–360, 2018b. ISSN 0377-2217. doi: 10.1016/j.ejor.2018.01.020.

Alejandro Correa Bahnsen, Djamila Aouada, and Björn Ottersten. Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19):6609–6619, 2015. ISSN 0957-4174. doi: 10.1016/j.eswa.2015.04.042.

Steven Finlay. Credit scoring for profitability objectives. *European Journal of Operational Research*, 202(2):528–537, 2010.

Carlos Serrano-Cinca and Begoña Gutiérrez-Nieto. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems*, 89:113–122, 2016. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/j.dss.2016.06.014.

Yong Xu, Gang Kou, and Daji Ergu. Profit-based uncertainty estimation with application to credit scoring. *European Journal of Operational Research*, 325(2):303–316, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.03.007.

Sherly Alfonso-Sánchez, Jesús Solano, Alejandro Correa-Bahnsen, Kristina P. Sendova, and Cristián Bravo. Optimizing credit limit adjustments under adversarial goals using reinforcement learning. *European Journal of Operational Research*, 315(2):802–817, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2023.12.025.

Sebastián Maldonado, Cristián Bravo, Julio López, and Juan Pérez. Integrated framework for profit-based feature selection and svm classification in credit scoring. *Decision Support Systems*, 104:113–121, 2017. ISSN 0167-9236. doi: 10.1016/j.dss.2017.10.007.

Nikita Kozodoi, Stefan Lessmann, Konstantinos Papakonstantinou, Yiannis Gatsoulis, and Bart Baesens. A multi-objective approach for profit-driven feature selection in credit scoring. *Decision Support Systems*, 120:106–117, 2019. doi: 10.1016/j.dss.2019.03.011.

Victor Medina-Olivares, Finn Lindgren, Raffaella Calabrese, and Jonathan Crook. Joint model for longitudinal and spatio-temporal survival data. *European Journal of Operational Research*, 327(3):892–904, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.07.060.

Raffaella Calabrese and Jonathan Crook. Spatial contagion in mortgage defaults: A spatial dynamic survival model with time and space varying coefficients. *European Journal of Operational Research*, 287(2):749–761, 2020. ISSN 0377-2217. doi: 10.1016/j.ejor.2020.04.031.

María Óskarsdóttir and Cristián Bravo. Multilayer network analysis for improved credit risk prediction. *Omega*, 105: 102520, 2021. ISSN 0305-0483. doi: 10.1016/j.omega.2021.102520.

Sahab Zandi, Kamesh Korangi, María Óskarsdóttir, Christophe Mues, and Cristián Bravo. Attention-based dynamic multilayer graph neural networks for loan default prediction. *European Journal of Operational Research*, 321(2): 586–599, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2024.09.025.

Yong Shi, Yi Qu, Zhensong Chen, Yunlong Mi, and Yunong Wang. Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation. *European Journal of Operational Research*, 315(2):786–801, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2023.12.028.

Sofie De Cnudde, Julie Moeyersoms, Marija Stankova, Ellen Tobback, Vinayak Javaly, and David Martens. What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance. *Journal of the Operational Research Society*, page 1–11, 2018. ISSN 0160-5682. doi: 10.1080/01605682.2018.1434402.

Viani B. Djeundje, Jonathan Crook, Raffaella Calabrese, and Mona Hamid. Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163:113766, 2021. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.113766.

Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. Unleashing the power of text for credit default prediction: Comparing human-written and generative ai-refined texts. *European Journal of Operational Research*, 326(3): 691–706, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.04.032.

Matthew Stevenson, Christophe Mues, and Cristián Bravo. The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, 295(2):758–771, 2021. ISSN 0377-2217. doi: 10.1016/j.ejor.2021.03.008.

Johannes Kriebel and Lennart Stitz. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, 302(1):309–323, 2022. ISSN 0377-2217. doi: 10.1016/j.ejor.2021.12.024.

Koen W. De Bock, Kristof Coussement, Arno De Caigny, Roman Słowiński, Bart Baesens, Robert N. Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, Sebastián Maldonado, David Martens, María Óskarsdóttir, Carla Vairetti, Wouter Verbeke, and Richard Weber. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2): 249–272, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2023.09.026.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

Matteo Ballegeer, Matthias Bogaert, and Dries F. Benoit. Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *European Journal of Operational Research*, 326(3):630–640, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.05.039.

Yujia Chen, Raffaella Calabrese, and Belen Martin-Barragan. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1):357–372, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2023.06.036.

Hendrik Andries du Toit, Willem DaniÃ, Helgard Raubenheimer, et al. Shapley values as an interpretability technique in credit scoring. *Journal of Risk Model Validation*, 2023.

Emanuele Borgonovo, Elmar Plischke, and Giovanni Rabitti. The many shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *European Journal of Operational Research*, 318(3):911–926, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2024.06.023.

Sullivan Hué, Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. Measuring the driving forces of predictive performance: Application to credit scoring. *ArXiv preprint*, arXiv:2212.05866v4, 2025. doi: 10.48550/arXiv.2212.05866.

Emilio Carrizosa, Kseniia Kurishchenko, and Dolores Romero Morales. On enhancing the explainability and fairness of tree ensembles. *European Journal of Operational Research*, 323(2):599–608, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2025.01.008.

Arno De Caigny, Kristof Coussement, and Koen W. De Bock. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269 (2):760–772, 2018. ISSN 0377-2217. doi: 10.1016/j.ejor.2018.02.009.

Mathias Kraus, Daniel Tschernutter, Sven Weinzierl, and Patrick Zschech. Interpretable generalized additive neural networks. *European Journal of Operational Research*, 317(2):303–316, 2024. ISSN 0377-2217. doi: 10.1016/j.ejor.2023.06.032.

Victor Medina-Olivares, Stefan Lessmann, and Nadja Klein. The deep promotion time cure model. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18848 – 18858, 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2024.3398559.

Lazaros Zografopoulos, Maria Chiara Iannino, Ioannis Psaradellis, and Georgios Sermpinis. Industry return prediction via interpretable deep learning. *European Journal of Operational Research*, 321(1):257–268, 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2024.08.032.

Runshan Fu, Yan Huang, and Param Vir Singh. Crowds, lending, machine, and bias. *Information Systems Research*, 2021. ISSN 1047-7047. doi: 10.1287/isre.2020.0990.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022. ISSN 0022-1082. doi: 10.1111/jofi.13090.

Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. The fairness of credit scoring models. *Management Science*, 2025. doi: 10.1287/mnsc.2022.03888.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, François-Xavier Aubet, Laurent Callot, and Tim Januschowski. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):Article 121, 2022. ISSN 0360-0300. doi: 10.1145/3533382.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90, 2022b. doi: 10.1016/j.inffus.2021.11.011.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, 2022.

Assaf Shmuel, Oren Glickman, and Teddy Lazebnik. A comprehensive benchmark of machine and deep learning models on structured data for regression and classification. *Neurocomputing*, 655:131337, 2025. ISSN 0925-2312. doi: 10.1016/j.neucom.2025.131337.

Cristian Bravo, Sebastian Maldonado, and Maria Oskarsdottir. *Deep Learning in Banking: Integrating Artificial Intelligence for Next-Generation Financial Services*. John Wiley & Sons, 2026.

Kamesh Korangi, Christophe Mues, and Cristián Bravo. A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, 308(1):306–320, 2023. ISSN 0377-2217. doi: 10.1016/j.ejor.2022.10.032.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning on tabular data. *arXiv preprint arXiv:2407.00956*, 2024.

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qile Zhou, and Han-Jia Ye. Representation learning for tabular data: A comprehensive survey. *ArXiv preprint*, arXiv.2504.16109, 2025. doi: 10.48550/arXiv.2504.16109.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. In *arXiv preprint arXiv:2012.06678*, 2020. URL https://arxiv.org/abs/2012.06678.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. VIME: extending the success of self- and semi-supervised learning to tabular domain. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/7d97667a3e056acab9aaf653807b4a03-Abstract.html.

Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18853–18865, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9c8661befae6dbcd08304dbf4dcaf0db-Abstract.html.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: improved neural networks for tabular data via row attention and contrastive pre-training. *CoRR*, abs/2106.01342, 2021. URL https://arxiv.org/abs/2106.01342.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. Tabllm: Few-shot classification of tabular data with large language models. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR, 2023. URL https://proceedings.mlr.press/v206/hegselmann23a.html.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637 (8045):319–326, 2025b. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint*, 2502.05564, 2025. doi: 10.48550/arXiv.2502.05564.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *ArXiv preprint*, arXiv:2511.08667v2, 2026. doi: 10.48550/arXiv.2511.08667. URL https://arxiv.org/abs/2511.08667.

Tassilo Klein and Johannes Hoffart. Foundation models for tabular data within systemic contexts need grounding. *ArXiv preprint*, arXiv:2505.19825, 2025. doi: 10.48550/arXiv.2505.19825.

Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. *Credit Scoring and its Applications*. Siam, Philadelphia, 2002.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox. *arXiv preprint arXiv:2407.04057*, 2024.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2623–2631. ACM, 2019. doi: 10.1145/3292500.3330701. URL 10.1145/3292500.3330701.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014. URL http://jmlr.org/papers/v15/delgado14a.html.

Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The annals of mathematical statistics*, 11(1):86–92, 1940.

Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Salvador Garcia and Francisco Herrera. An extension on" statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9(12), 2008.