

AUC Is Not Enough: A Utility-Aware Comparison of XGBoost and Multi-Task FT Variants for Proxy-Prescriptive Credit Modeling

Edinaldo Vicente de Oliveira, Anderson Tenório Sergio

¹CESAR School
Recife – PE – Brasil

evo@cesar.school, ats@cesar.school

Abstract. *Credit-scoring models are often ranked by AUC, yet lending decisions depend on calibrated default probabilities, offer rules, and economic consequences. Using the Home Credit Default Risk benchmark, we compare XGBoost with three multi-task neural baselines: MLP, FT-flat, and source-aware FT-block, under a calibrated proxy-prescriptive workflow. XGBoost achieves the highest AUC (0.7833), with a significant paired-DeLong gap over FT-block (0.7751), but the ranking inverts under proxy utility: FT-block leads in all economic scenarios, about 12% above XGBoost in the base case. FT-flat does not reproduce the inversion, and post-hoc diagnostics localise the effect mainly in the credit/offer channel.*

Keywords: credit scoring; tabular deep learning; multi-task learning; probability calibration; utility-aware evaluation.

1. Introduction

Credit-scoring models are routinely benchmarked by discriminative metrics, and the area under the ROC curve (AUC) is the de facto standard. Discrimination, however, is only one ingredient of a lending decision. Operational pipelines act on *calibrated* default probabilities, apply economically interpretable offer rules, and are ultimately judged by the monetary consequences of approvals and rejections. Hand [Hand 2009] showed that AUC implicitly averages misclassification costs in a way that varies across classifiers, making it incoherent for cost-aware comparison and motivating decision-aware alternatives. Profit-driven evaluation [Verbraken et al. 2013, Verbraken et al. 2014], cost-sensitive learning [Elkan 2001], and calibration requirements under Basel-style governance [Basel Committee on Banking Supervision 2006] all point in the same direction: a model that ranks applicants well may still be a poor basis for the decision actually taken.

This raises a concrete question for credit modelling that is rarely tested head-on: *does the model with the best AUC also deliver the best decision quality under a calibrated, prescriptive criterion?* If discriminative and decision-aware rankings can diverge, selecting models by AUC alone may be actively misleading in deployment.

We study this on the *Home Credit Default Risk* benchmark and address two research questions. **RQ1:** does the model ordering induced by AUC coincide with the ordering induced by a calibrated proxy utility? **RQ2:** is any such inversion robust to economic scenarios, approval-threshold sensitivity, and sampling uncertainty?

Our setting is *proxy-prescriptive*: we apply a fixed, transparent decision rule (approval plus offer sizing) on top of calibrated model outputs and score its monetary consequence under explicit economic scenarios. The comparison spans four base families detailed in Section 5: XGBoost, a multi-task perceptron (MLP-MTL), a flat Feature-Tokenizer Transformer (FT-Transformer, hereafter FT-flat) [Gorishniy et al. 2021], and a source-aware block-aware variant (FT-block). Two guardrails are fixed in advance: the target claim concerns *proxy* utility rather than realised profit, and fairness outputs are diagnostic only.

This paper makes three contributions. First, it documents a reproducible data-and-audit workflow for credit modelling, including leakage checks and monetary-periodicity and split-disjointness audits. Second, it evaluates the base models under a common protocol, with validation-based calibration and a paired DeLong test on the full test split. Third, it shows that the AUC-induced ranking can *invert* under the proxy-utility rule, supports the inversion with threshold-sensitivity and paired-bootstrap analyses, and localises the effect, by a post-hoc channel decomposition, in the credit/offer channel rather than in risk ranking alone.

The study is deliberately bounded: an i.i.d. split, selective labels retained, and utility treated as a transparent heuristic proxy. Section 7 states these boundaries in full.

2. Related Work

2.1. Tree-Based Models and Credit-Scoring Benchmarks

Since its introduction by Chen and Guestrin [Chen and Guestrin 2016], XGBoost has become a dominant baseline for tabular credit modelling, combining strong predictive performance, engineering maturity and operational familiarity. The most influential empirical benchmark for credit scoring is the large-scale comparison of Lessmann et al. [Lessmann et al. 2015], which evaluates 41 classifiers across eight credit data sets and finds that ensembles consistently outperform single classifiers. The survey of Dastile et al. [Dastile et al. 2020] extends this picture to deep learning, which is increasingly competitive but raises interpretability concerns. Grinsztajn et al. [Grinsztajn et al. 2022] give structural reasons why tree-based methods remain strong on heterogeneous tabular data, while McElfresh et al. [McElfresh et al. 2023] show that neural models close or overturn the gap only under specific conditions. This motivates XGBoost as a strong industrial baseline, though not as exhaustive coverage of tree-based methods such as LightGBM [Ke et al. 2017] and CatBoost [Prokhorenkova et al. 2018].

2.2. Deep Learning for Tabular Data

The FT-Transformer adapts attention to tabular inputs by tokenising features and learning interactions through self-attention [Gorishniy et al. 2021]. Early benchmarks favoured boosted trees [Shwartz-Ziv and Armon 2022], but later work showed that model family and tuning budget matter greatly [McElfresh et al. 2023].

2.3. Multi-Task Learning

Lending needs correlated outputs—risk, offer size and affordability. Multi-Task Learning (MTL) exploits this correlation: jointly learning related tasks regularises the

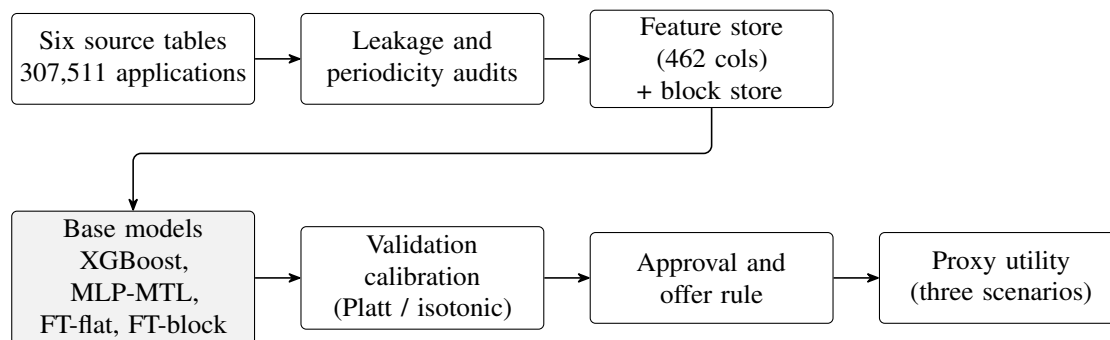


Figure 1. Proxy-prescriptive evaluation pipeline.

shared representation and typically improves generalisation when the tasks are related [Caruana 1997, Ruder 2017, Kendall et al. 2018], which makes it a natural framework for the credit pipeline studied here.

2.4. Calibration and Selective Labels

Post-hoc calibration is operationally important under Basel-style requirements [Basel Committee on Banking Supervision 2006]. Platt scaling and isotonic regression are standard tools [Platt 1999, Zadrozny and Elkan 2002]; Guo et al. [Guo et al. 2017] characterise miscalibration of modern neural networks, and Kull et al. [Kull et al. 2017] propose beta calibration for asymmetric scores. Lending data also suffer from *selective labels*: defaults are observed only for historically approved applications [Lakkaraju et al. 2017, Coston et al. 2021], limiting causal claims about counterfactual decision quality.

2.5. Profit Scoring, Fairness and Governance

Discriminative metrics alone do not capture decision value. Verbraken et al. [Verbraken et al. 2013] introduced the Expected Maximum Profit measure, later applied to consumer credit [Verbraken et al. 2014]; Elkan [Elkan 2001] formalises cost-sensitive learning, and Thomas et al. [Thomas et al. 2017] consolidate decision-aware methodology. Kozodoi et al. [Kozodoi et al. 2022] study the fairness–profit trade-off, while governance work emphasises reproducibility, explainability and the bounded scope of fairness claims [Pineau et al. 2021].

2.6. Positioning

Our contribution is neither a new architecture nor a state-of-the-art claim. We place an established tabular baseline and a small family of multi-task neural variants under a *single* calibrated, decision-aware protocol and ask whether discriminative and proxy-utility rankings agree. The novelty is the controlled demonstration that they need not, with a falsifiable negative control (FT-flat) and a post-hoc localisation of the effect in the offer channel.

3. Problem Formulation

By *proxy-prescriptive* we mean that evaluation applies a fixed, transparent decision rule—approval plus offer sizing—to calibrated model outputs and scores the resulting monetary consequence. The workflow has two stages (Figure 1). The first prepares and audits the

data, aggregating six source tables into a flat feature store and a parallel source-aware block store that preserves the origin of each predictor. The second is decision-aware: the base models are trained under a common protocol, their risk scores calibrated on validation data, and a fixed approval-and-offer policy maps the calibrated outputs to a monetary proxy utility across three economic scenarios.

3.1. Task Definition

Let $\mathcal{D} = \{(x_i, y_i, c_i, a_i)\}_{i=1}^N$, where x_i is the applicant feature vector, $y_i \in \{0, 1\}$ indicates default, $c_i = \log(1 + \text{AMT_CREDIT}_i)$ is the (log) contracted credit amount, and $a_i = \log(1 + \text{AMT_ANNUITY}_i)$ is the (log) contractual annuity, treated as a proxy of instalment size rather than an observed payment series. The neural models are trained in a multi-task setting with three heads: a *risk* head, a *credit* head, and an auxiliary *annuity* head. The proxy utility uses the calibrated risk score and the credit head; the annuity head acts as an auxiliary regularisation signal.

3.2. Masked Regression under Selective Labels

Because outcomes are observed only for historically approved applicants, regression losses are evaluated under a good-payer mask. With $m_i = \mathbb{I}[y_i = 0]$,

$$\mathcal{L}_c = m_i \text{Huber}_\delta(\hat{c}_i - c_i), \quad \mathcal{L}_a = m_i \text{Huber}_\delta(\hat{a}_i - a_i), \quad (1)$$

averaged over the mini-batch, where the Huber loss of a residual r with threshold δ is

$$\text{Huber}_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta. \end{cases} \quad (2)$$

This is operationally coherent for the available labels but does not solve the causal-identification problem induced by selective labels: it introduces a train/inference shift for the credit head, which sees only good payers at training time. We return to this in Section 7.

3.3. Calibration and Proxy Policy

Raw default scores are calibrated on the validation set by selecting the better of Platt scaling and isotonic regression by validation Brier score, $\text{PD}_{\text{cal},i} = g(\text{PD}_{\text{raw},i})$, $g \in \{\text{Platt}, \text{isotonic}\}$. The approval threshold and scenario parameters are fixed a priori, not tuned to favour any model. We use $\text{PD}_{\text{cal},i} \leq 0.15$, an operational choice that yields an approval rate near the historical portfolio level ($\approx 85\%$). Approval is $\text{approved}_i = \mathbb{I}[\text{PD}_{\text{cal},i} \leq 0.15]$, and the proxy offer rule is

$$\hat{o}_i = \text{clip}\left[(\exp(\hat{c}_i) - 1) \cdot (1 - \text{PD}_{\text{cal},i}), \ell, u\right], \quad (3)$$

where ℓ and u are the 1st and 99th percentiles of training good-payer credit values; clipping is applied to the offer, after scaling by $(1 - \text{PD}_{\text{cal},i})$. For scenario s , the proxy utility is

$$u_i^{(s)} = \hat{o}_i \cdot (r_s - \text{PD}_{\text{cal},i} \cdot \text{LGD}_s \cdot \kappa_s) \cdot \text{approved}_i, \quad (4)$$

where r_s is the scenario return rate, LGD_s the loss given default, and κ_s a scenario constant scaling the expected-loss severity term, capturing scenario-specific capital and severity assumptions not absorbed by LGD_s . The rule is applied only at evaluation time, on top of standard losses, which distinguishes it from cost-sensitive training [Elkan 2001] and places the comparison within profit-driven evaluation [Verbraken et al. 2013, Verbraken et al. 2014].

3.4. Evaluation Criteria

Two families of criteria are compared throughout. *Discriminative* criteria measure ranking quality independently of any decision: AUC is the area under the ROC curve, and PR-AUC the area under the precision–recall curve, which is more informative under the 8.07% class imbalance of this benchmark. *Calibration* criteria measure whether predicted probabilities match observed frequencies: the Brier score is the mean squared error of $\text{PD}_{\text{cal},i}$ against y_i , and the expected calibration error (ECE) is the average absolute gap between predicted and empirical default rates across probability bins. The *decision-aware* criterion is the proxy utility of Eq. (4), aggregated over the test split. The central question of this paper is whether the model ordering induced by the first family coincides with the ordering induced by the third.

3.5. Training Objective

The neural models minimise

$$\mathcal{L}_{\text{total}} = \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c + \lambda_a \mathcal{L}_a, \quad (5)$$

where $\mathcal{L}_r$ is binary cross-entropy and $\mathcal{L}_c, \mathcal{L}_a$ are the masked Huber losses of Eq. (1). We probe both uncertainty-weighting [Kendall et al. 2018] and fixed weights.

4. Data, EDA and Audits

4.1. Dataset, Exploratory Analysis and Feature Store

We use the *Home Credit Default Risk* benchmark, a public application-scoring dataset. We use its application table and five of its six history tables, summarised in Table 1; POS_CASH_balance was not aggregated. The label is the binary TARGET of application_train, and the task is strongly imbalanced—the default rate is 8.07% ($\approx 11.4:1$ negative-to-positive), stable across seeds and splits. After aggregation and joins the flat feature store has shape (307,511, 462): the six blocks of Table 1 contribute 455 features, five cross-source coverage flags indicate which histories exist for each applicant, and the identifier and label complete the count; training bundles therefore use 444 numerical and 16 categorical predictors per seed. A parallel source-aware *block store* groups features by origin for FT-block. Coverage varies widely across history-dependent sources—from 94.8% for instalment history down to 28.3% for revolving credit—reflecting applicants without the corresponding records; the block-aware design lets FT-block condition on this structure rather than on imputed values alone. Missingness concentrates in a minority of derived columns (previous-application interest-rate aggregates exceed 98% missing) and is handled by training-set-fitted imputation.

4.2. Leakage and Structural Audits

Before modelling we apply leakage controls: anteriority filters in bureau (DAYS_CREDIT ≤ 0) and previous_application (DAYS_DECISION ≤ 0), merge checks by SK_ID_CURR, and exclusion of known post-outcome columns. Imputers, scalers and encoders are fitted on the training set only; no leakage was detected. A monetary-periodicity audit verifies the semantic convention linking AMT_INCOME_TOTAL, AMT_ANNUITY and AMT_CREDIT; under the selected convention 99.7% of evaluable rows satisfy a debt-service ratio ≤ 0.6 and 99.996% an implied term of 6–360 months.

Table 1. Source tables: size, content, features contributed to the block store, and coverage, defined as the fraction of the 307,511 applicants for whom the source contains at least one record. Coverage is measured per source table independently, so child tables need not nest within their parent.

Source table	Content	Rows	#feat.	Coverage
application_train	applications, label	307,511	120	100.0%
bureau	external credits	1,716,428	65	85.7%
bureau_balance	their monthly balances	27,299,925	16	30.0%
previous_application	prior applications	1,670,214	100	94.6%
installments_payments	repayment history	13,605,401	43	94.8%
credit_card_balance	revolving credit	3,840,312	111	28.3%

4.3. Split Design

The public benchmark exposes no reliable application timestamp, so we use a stratified i.i.d. protocol with three seeds (2026, 2027, 2028) and a 64/16/20 train/validation/test split by TARGET: 196,806 training, 49,202 validation and 61,503 test rows per seed. A persisted disjointness audit confirms zero overlap across partitions. The absence of a temporal split is an explicit limitation (Section 7, L1).

5. Experimental Protocol

5.1. Model Families

XGBoost. One classifier for risk plus two independent regressors, following modular credit-modelling practice. **MLP-MTL.** A shared-body perceptron with task-specific heads, giving a non-attentional multi-task comparator. **FT-flat.** A flat FT-Transformer kept as a methodological control for whether any Transformer reproduces the inversion. **FT-block.** A source-aware FT-Transformer, the primary neural family here; it tokenises features with source embeddings, so attention is informed by the block structure of Table 1.

5.2. Architectures and Training

Table 2 lists the selected hyperparameters. For the neural models we use learning rate 2×10^{-4} , weight decay 10^{-3} , gradient clipping 0.5, batch size 1024, and at most 50 epochs with patience 8. Both FT variants use mean pooling over feature tokens as their readout. This departs from the [CLS]-token aggregation of the original FT-Transformer [Gorishniy et al. 2021]; we treat it as a fixed architectural choice applied identically to FT-flat and FT-block, so it cannot by itself explain any difference between them.

Table 2. Base-model hyperparameters (selected by probe on seed 2026).

Model	depth/layers	lr/heads	other
XGBoost	depth=6, est.=1200	lr=0.03	sub=0.8, col=0.8
MLP-MTL	[384, 192, 96]	—	drop=0.1
FT-flat	layers=3, $d = 128$	heads=8	mean pool, drop=0.1
FT-block	layers=2, $d = 128$	heads=4	mean pool, src-emb, drop=0.1

5.3. Calibration and Statistical Testing

All models are calibrated on the validation split by choosing between Platt and isotonic calibration via validation Brier score. Only the risk head is calibrated; the credit head is

used uncalibrated, a point we revisit when interpreting the mechanism (Section 6) and in Section 7. Pairwise AUC comparisons use the paired DeLong test [Sun and Xu 2014], computed on the calibrated probabilities over the full test split ($n = 61,503$ per seed). Two no-retrain analyses strengthen inference on the central metric: a threshold-sensitivity sweep that re-applies the canonical policy for approval thresholds from 0.05 to 0.30, and a paired bootstrap [Efron and Tibshirani 1993] with $N_{\text{boot}}=2000$ replications and independent RNGs per scope, estimating the uncertainty of the FT-block – XGBoost utility difference at the canonical threshold. The three economic scenarios are conservative $(r, \text{LGD}, \kappa) = (0.06, 0.40, 0.020)$, base $(0.10, 0.45, 0.015)$, and optimistic $(0.14, 0.35, 0.010)$.

6. Results

6.1. Base Benchmark

Table 3 separates predictive from decision-aware ranking. XGBoost is the best base model on AUC and PR-AUC. FT-block is not competitive with XGBoost on those discriminative metrics, yet it leads on proxy utility. FT-flat is a negative control: despite a higher AUC than FT-block, it has the lowest utility of the four families.

Table 3. Base benchmark on the test split (mean over 3 seeds; base-scenario utility in billions of the benchmark’s nominal currency units, denoted R\$ by convention).

Model	AUC	PR-AUC	Brier	ECE	Approval	Utility _{base}
XGBoost	0.7833±0.0005	0.2705±0.0038	0.0660	0.0025	85.7±1.8%	3.052±0.054
MLP	0.7709±0.0006	0.2458±0.0026	0.0672	0.0022	83.9±1.5%	2.879±0.574
FT-flat	0.7791±0.0006	0.2617±0.0021	0.0665	0.0024	84.7±1.1%	1.918±0.799
FT-block	0.7751±0.0017	0.2541±0.0031	0.0668	0.0024	84.3±0.9%	3.430±0.085

6.2. AUC Comparisons

The paired DeLong tests (Table 4) are unambiguous: XGBoost wins on discrimination and FT-block does not match it on AUC. The contribution is not a hidden superiority on the same metric; it is that the AUC-induced and utility-induced rankings differ. No multiple-testing correction is applied across the comparisons in Table 4, and the reported significance should be read accordingly.

Table 4. Key paired DeLong comparisons (range over 3 seeds; $n = 61,503$ per seed).

Pair	ΔAUC range	p -value range	Interpretation
XGBoost vs. FT-block	+0.0066 to +0.0106	1.7×10^{-11} – 4.3×10^{-5}	XGBoost higher
XGBoost vs. MLP	+0.0115 to +0.0136	$< 10^{-13}$ – 6.4×10^{-12}	XGBoost higher
FT-flat vs. FT-block	+0.0027 to +0.0057	4.9×10^{-6} – 2.5×10^{-2}	FT-flat higher

6.3. The Utility Inversion and Its Mechanism

The central result is in Table 5: FT-block yields the highest base-model utility in all three scenarios. In the base scenario it exceeds XGBoost by about 12.4%. The inversion is specific to FT-block—FT-flat does not reproduce it despite a higher AUC. A post-hoc

Table 5. Proxy utility across scenarios (mean $\pm$ std, 3 seeds; R\$ billion).

Model	Conservative	Base	Optimistic
XGBoost	1.826 $\pm$ 0.032	3.052 $\pm$ 0.054	4.281 $\pm$ 0.077
MLP	1.722 $\pm$ 0.344	2.879 $\pm$ 0.574	4.039 $\pm$ 0.805
FT-flat	1.147 $\pm$ 0.478	1.918 $\pm$ 0.799	2.691 $\pm$ 1.121
FT-block	2.052$\pm$0.051	3.430$\pm$0.085	4.812$\pm$0.120

decomposition localises the effect in the credit/offer channel rather than in risk ranking. Auditing the credit head against the true (log) credit target on the test split, FT-block attains a mean predicted/true credit ratio of 1.14 and a log-credit MAE of 0.132, whereas the block-equivalent FT-flat *under*-predicts offer scale (ratio 0.60, MAE 0.510). In a counterfactual swap, replacing the flat credit channel with the block credit channel closes most of the utility gap, while replacing only the risk channel closes little. Because the offer rule (Eq. 3) multiplies utility by $\hat{o}_i$, and the credit head is *uncalibrated* (Section 5), part of the gap is attributable to offer-scale calibration. The honest reading is therefore sharper than “a Transformer wins”: it is the *joint* calibration of risk ranking and offer-scale prediction that governs proxy utility. This is a post-hoc diagnostic, not a pre-registered hypothesis (Section 7, L8, L12).

6.4. Robustness

FT-block outperforms XGBoost at every evaluated approval threshold (Table 6), with a relative gain between +4% and +14%. The paired bootstrap (Table 7) keeps the 95% confidence interval of the FT-block – XGBoost utility difference strictly above zero in every seed. Averaged over seeds the difference is +0.226, +0.378 and +0.530 B in the conservative, base and optimistic scenarios; the per-seed intervals, not a sum of per-seed differences, are the inferential quantity.

Table 6. Approval-threshold sensitivity, base scenario (mean across seeds, no retraining). Gain: FT-block vs. XGBoost.

Threshold	XGBoost (B)	FT-block (B)	Δ (B)	Rel. gain
0.05	2.022	2.104	+0.082	+4.0%
0.10	2.723	3.027	+0.305	+11.2%
0.15	3.052	3.430	+0.378	+12.4%
0.20	3.194	3.650	+0.456	+14.3%
0.25	3.269	3.727	+0.459	+14.0%
0.30	3.327	3.774	+0.447	+13.5%

Table 7. Paired bootstrap of the FT-block – XGBoost utility difference at threshold 0.15 ($N_{\text{boot}}=2000$; 95% CI).

Scope	Scenario	Δ (B)	95% CI (B)
Seed 2026	Base	+0.260	[0.252, 0.268]
Seed 2027	Base	+0.434	[0.426, 0.442]
Seed 2028	Base	+0.440	[0.431, 0.448]

6.5. Distributional Disclosure

Because utility-oriented selection changes who is approved and how much is offered, a recommendation to adopt it carries an obligation to disclose the distributional effect. We therefore report one bounded diagnostic. We apply no formal fairness criterion—no demographic parity, equalised odds, or calibration-within-groups—and no fairness constraint enters training or model selection. We partition the test set by `CODE_GENDER`, commonly treated as a protected attribute in the credit-scoring fairness literature [Kozodoi et al. 2022] and recorded directly as a category in a benchmark that holds no race, ethnicity or nationality, and recompute the approval rate and mean calibrated PD within each subgroup. Age is the other protected attribute observable here, via `DAYS_BIRTH`, but it is continuous and would require an arbitrary banding choice; we leave it to future work. For FT-block, approval is 88.1% for women against 77.1% for men, with mean calibrated PD of 0.069 and 0.103 respectively: under a single global threshold, the subgroup with lower predicted default receives more approvals. These are unadjusted associations, not causal effects, and must not be read as compliance certification; formal fairness criteria, additional protected attributes and mitigation strategies are left to future work [Kozodoi et al. 2022].

7. Limitations

We state the boundaries of the study explicitly. **(L1) No temporal validation.** All time-related fields in the benchmark (e.g. `DAYS_BIRTH`, `DAYS_CREDIT`, `DAYS_DECISION`) are measured relative to each application’s own date, so no chronological ordering *between* applications is recoverable and no out-of-time split is definable on this data. The absence of temporal validation is thus a property of the benchmark, not a protocol choice, but it still limits operational conclusions under drift. **(L2) Selective labels.** Defaults are observed only under the historical approval policy [Lakkaraju et al. 2017]; the good-payer mask is operationally coherent but does not solve causal identification. **(L3) Proxy utility.** Equation (4) is a transparent heuristic, not an institution’s profit-and-capital model. **(L4) Small seed count.** With three seeds, inter-seed variability is estimated imprecisely, and initialisation dominates sampling as a source of variability: the per-seed difference ranges from +0.260 to +0.440 B, about eleven times the mean width of the within-seed bootstrap intervals (≈ 0.016 B). Table 7 should therefore be read as uncertainty conditional on a seed, not of the effect itself. Treating the seed as the unit of analysis, the three per-seed differences give a mean of +0.378 B with a 95% *t*-interval of [0.124, 0.632] B (*df*=2), still strictly above zero. **(L5) Single dataset.** All evidence comes from one benchmark; external validity is open. **(L6) Diagnostic-only governance.** Fairness artefacts are descriptive and insufficient for normative assessment. **(L7) Variant-specific inversion.** The inversion is supported by FT-block, not by FT-flat; it is not a generic “Transformers beat trees” claim. **(L8) Post-hoc mechanism.** The credit-channel decomposition was conducted to explain the finding, not as a pre-registered hypothesis. **(L9) Auxiliary annuity head.** The third head uses `AMT_ANNUITY` as a contractual-annuity proxy, not observed repayment behaviour. **(L10) Non-exhaustive tabular baselines.** XGBoost is a strong industrial baseline, but LightGBM, CatBoost, TabNet [Arik and Pfister 2021] and SAINT [Somepalli et al. 2021] were not evaluated; we make no exhaustive state-of-the-art claim. TabPFN is excluded on technical grounds rather than by omission: the cited version [Hollmann et al. 2023] targets problems of order 10^3 samples and 10^2 features, far below the $307,511 \times 462$ setting used here. **(L11) Utility functional form.** Proxy

utility depends on the offer rule, scenario parameters and credit-head scale. Under the chosen κ_s the expected-loss term is at most 2% of the revenue term, so the three scenarios differ essentially by a scale factor on r_s and scenario variation is a weak robustness test: the FT-block over XGBoost utility ratio is 1.124 in all three. Robustness over alternative functional forms, such as a concave offer rule or a capital-adjusted criterion, is left to future work. **(L12) Uncalibrated credit head.** Only the risk head is calibrated, so part of the FT-block advantage reflects offer-scale calibration rather than risk discrimination (Section 6). Re-scaling the credit head to a unit predicted/true ratio would quantify how much of the gap that accounts for; it is left to future work. **(L13) Confounded comparison axes.** The comparison does not disentangle model family from the shared multi-task representation (XGBoost uses one classifier plus independent regressors, whereas the neural models share a backbone across three heads), nor does it ablate whether FT-block’s advantage stems from source-aware tokenisation, its smaller capacity acting as a regulariser, or both. Both ablations are left to future work.

8. Conclusion

On the *Home Credit Default Risk* benchmark, XGBoost leads on discrimination (AUC 0.7833 ± 0.0005 against 0.7751 ± 0.0017 , a gap significant under paired DeLong in all three seeds), yet FT-block leads on proxy utility in all three scenarios, exceeding XGBoost by about 12.4% in the base scenario. The ordering holds across every approval threshold tested, and the paired bootstrap keeps the gap above zero in every seed.

The central contribution is that the AUC-induced ranking can invert under a calibrated proxy-utility rule, with the effect localised by a post-hoc decomposition in the credit/offer channel rather than in risk ranking alone.

The result is deliberately bounded. It is not a generic “Transformers beat trees” claim, since FT-flat does not reproduce the inversion; not an institutional profit claim, since utility is a proxy; and not a fairness certification, since governance outputs are diagnostic only. The practical implication is direct: credit-model evaluation should not stop at AUC. In proxy-prescriptive settings, calibrated probabilities, offer scale and utility-aware comparison can reveal a qualitatively different ordering of models.

Future work should test whether the FT-block advantage survives institution-specific utility functions, out-of-time validation, multiple datasets, and LightGBM/CatBoost comparisons; and should quantify the offer-scale contribution by re-scaling the credit head.

References

- Arik, S. Ö. and Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687.
- Basel Committee on Banking Supervision (2006). Basel II: International convergence of capital measurement and capital standards. Technical report, Bank for International Settlements.
- Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Coston, A., Rambachan, A., and Chouldechova, A. (2021). Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 2144–2155.
- Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34. OpenReview version updated in 2023.
- Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330.
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123.
- Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30.
- Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491.
- Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094.
- Kull, M., de Menezes e Silva Filho, T., and Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631.

- Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284.
- Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.
- McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., and White, C. (2023). When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.
- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Larochelle, H. (2021). Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164):1–20.
- Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv:1706.05098.
- Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.
- Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., and Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv:2106.01342.
- Sun, X. and Xu, W. (2014). Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393.
- Thomas, L. C., Crook, J. N., and Edelman, D. B. (2017). *Credit Scoring and Its Applications*. SIAM, 2nd edition.
- Verbraken, T., Bravo, C., Weber, R., and Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2):505–513.
- Verbraken, T., Verbeke, W., and Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):961–973.
- Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.