

A Comparative Explainable Machine Learning Framework for Predicting QRIS Adoption Using Synthetic MSME Profiles

Erni Febrina Harahap^a, S Salfadri^{b,*}, Jhon Rinaldo^b, Yoga Perdana^c, Arciana Damayanti^c

^aUniversitas Bung Hatta, Padang, Indonesia

^bUniversitas Eka Sakti, Padang, Indonesia

^cUniversitas Pendidikan Indonesia, Bandung, Indonesia

Abstract

Quick Response Code Indonesian Standard (QRIS) has become the national interoperable payment infrastructure for Indonesian micro, small, and medium enterprises (MSMEs), yet detailed public microdata linking MSME characteristics to adoption status remain unavailable, which limits the development of predictive analytics for merchant onboarding. This study develops and evaluates a comparative explainable machine learning framework for predicting potential QRIS adoption using theory-informed synthetic MSME profiles generated through Monte Carlo simulation. A transparent data-generating process combining linear, quadratic, interaction, and threshold components produced 2,000 profiles under a moderate digital environment and 2,000 additional profiles under digitally enabling and digitally constrained scenarios. Four algorithmic families were compared: Logistic Regression, Support Vector Machine with radial basis function kernel (SVM-RBF), CatBoost, and the Explainable Boosting Machine (EBM). Evaluation integrated discrimination, probability calibration, distribution-shift testing, permutation feature importance, Accumulated Local Effects, intrinsic EBM feature functions, counterfactual explanations, and robustness analysis over sample size, label noise, class imbalance, coefficient perturbation, and random seeds. On the independent test set, SVM-RBF achieved the highest balanced accuracy (0.767) and the lowest Brier score (0.173), while Logistic Regression attained the highest ROC-AUC (0.821); the intrinsically interpretable EBM remained within approximately one percentage point of the best model on every metric. Feature-importance rankings were highly consistent across the four model families (Spearman $\rho \geq 0.91$), with facilitating conditions, digital financial literacy, and performance expectancy dominating the simulated predictions. Because all profiles are synthetic, the reported probabilities and feature effects are computational outcomes under explicitly stated simulation assumptions rather than empirical estimates or causal evidence. The framework demonstrates that a transparent, calibrated, and robustness-audited prediction pipeline for QRIS adoption is computationally feasible and provides a reproducible template for future validation with real MSME data.

Keywords: QRIS, MSMEs, Explainable Artificial Intelligence, Synthetic Data, CatBoost

Received: 17 June 2026

Revised: 17 October 2026

Published: 31 October 2026

1. Introduction

Digital payments have become a central instrument of operational transformation for micro, small, and medium enterprises (MSMEs) in emerging economies. In Indonesia, the Quick Response Code Indonesian Standard (QRIS), introduced by Bank Indonesia in 2019 as the national QR payment standard, allows a single merchant-presented QR code to accept payments from any licensed payment provider, thereby lowering acceptance costs and simplifying reconciliation for small merchants (Hamzah Muchtar et al., 2024; Indonesia, 2019). Despite this standardization, MSMEs differ considerably in their readiness to adopt digital payment technology because they operate under heterogeneous technological, financial, infrastructural, and market condition (Juniarini et al., 2020; Suryani et al., 2017; Tambunan, 2021). Constructs from technology-acceptance research, including performance expectancy, effort expectancy, social influence, facilitating conditions, trust, perceived security, perceived risk, digital financial literacy, customer demand, and perceived transaction cost, offer a theoretically grounded vocabulary for describing this heterogeneity (Davis, 1989; Venkatesh Morris, M. G., Davis, G. B., & Davis, F. D., 2003; Venkatesh et al., 2012). In

* Corresponding author.

E-mail address: salfadrisemsi123@gmail.com

the present study these constructs are not empirical findings; they are theoretical inputs to a controlled simulation framework.

Conventional technology-adoption research typically applies the Technology Acceptance Model, the Unified Theory of Acceptance and Use of Technology, regression analysis, or structural equation modeling (Davis, 1989; Venkatesh Morris, M. G., Davis, G. B., & Davis, F. D, 2003). These inferential approaches are valuable for hypothesis testing and parameter estimation, but they emphasize population-average relationships and frequently impose linear and additive functional forms. They are therefore less suited to individual-level prediction, and they may miss threshold effects and interactions among business and behavioral characteristics. Machine learning complements inferential modeling by prioritizing out-of-sample predictive accuracy and flexible pattern recognition (Breiman, 2001; Hastie et al., 2009). However, high-performing learners such as gradient-boosted ensembles and nonlinear support vector machines are difficult to inspect directly, which raises legitimate concerns for decision support in financial-inclusion settings (Guidotti et al., 2018; Rudin, 2019). This tension motivates comparative explainable machine learning, in which candidate models are evaluated jointly on predictive performance, probability calibration, interpretability, and robustness rather than on accuracy alone.

A practical obstacle for predictive QRIS research is the absence of a sufficiently detailed public dataset that jointly records MSME characteristics, technology-acceptance constructs, infrastructural conditions, and adoption status (Agustiana et al., 2025; Muhammad Hilmansyah Lukman et al., 2025). Following an established tradition of simulation-based methodology development and synthetic-data research (Jordon, 2022; Rubin, 1993) this study constructs a controlled computational environment: synthetic MSME profiles are generated from a fully documented, theory-informed Monte Carlo process, and the adoption target is produced by a transparent latent-score mechanism specified before any model is trained. The purpose of the synthetic environment is to test the feasibility, comparative behavior, interpretability, and robustness of the modeling pipeline prior to empirical validation. A mandatory boundary therefore applies throughout the article: the synthetic profiles do not estimate actual QRIS adoption prevalence, do not validate causal determinants, and do not represent the population of Indonesian MSMEs. All results describe the behavior of models under explicitly stated simulation assumptions.

Six gaps motivate the study. First, QRIS adoption research is predominantly inferential rather than predictive. Second, machine-learning studies of payment adoption commonly report accuracy while neglecting probability calibration, even though calibrated probabilities are essential for segmentation and prioritization (Niculescu-Mizil & Caruana, 2005). Third, explainability is often reduced to a single feature-importance ranking without analysis of nonlinear feature effects (Apley & Zhu, 2020; Molnar, 2022). Fourth, local explanations are usually descriptive rather than actionable, whereas counterfactual formulations can translate predictions into minimum-change recommendations (Mothilal et al., 2020; Wachter et al., 2018). Fifth, synthetic-data studies risk circular validation when a model merely recovers the generation rules; this risk must be surfaced and managed through noise injection, distribution shift, and coefficient perturbation. Sixth, few comparative designs explicitly test the stability of explanations across algorithms, random seeds, scenarios, and noise levels (Fisher et al., 2019).

The study addresses six research questions. RQ1: How accurately can different machine-learning algorithms classify potential QRIS adopters from synthetic MSME profiles? RQ2: Which algorithm provides the best balance between predictive performance, probability calibration, and interpretability? RQ3: How consistent are global feature-importance patterns across Logistic Regression, SVM-RBF, CatBoost, and EBM? RQ4: How do selected MSME characteristics generate nonlinear and interaction effects on simulated adoption predictions? RQ5: What minimum actionable changes are required to shift a synthetic profile from predicted non-adoption to adoption? RQ6: How robust are performance and explanations under changes in sample size, class imbalance, label noise, coefficient assumptions, and digital-environment scenarios?

The contributions are methodological (an integrated comparison of discrimination, calibration, explainability, explanation stability, and counterfactual analysis), computational (a reproducible generator for theory-informed synthetic MSME profiles with a frozen data-generating process), theoretical (a predictive operationalization of adoption constructs including nonlinearities, thresholds, and interactions, without any claim of empirical validation), practical (a prototype workflow that may later support MSME digital-readiness segmentation after validation with real data), and reproducibility-oriented (full disclosure of distributions, coefficients, seeds, and hyperparameter search spaces). The study is deliberately distinct from prior simulation work that treated QRIS adoption as an input scenario for transaction-efficiency assessment; here, adoption is the classification target and the object of explanation.

2. Methods

2.1 Research Design

The research uses a quantitative computational design (Harahap et al., 2023) that combines theory-informed synthetic data generation, supervised binary classification, comparative machine learning, explainable artificial intelligence, probability calibration, scenario analysis, and sensitivity testing. Figure 1 summarizes the framework. Theoretical constructs feed a Monte Carlo profile generator; a transparent latent-score mechanism produces the simulated adoption target; four model families are trained and compared; and an explainability and robustness layer audits every stage.

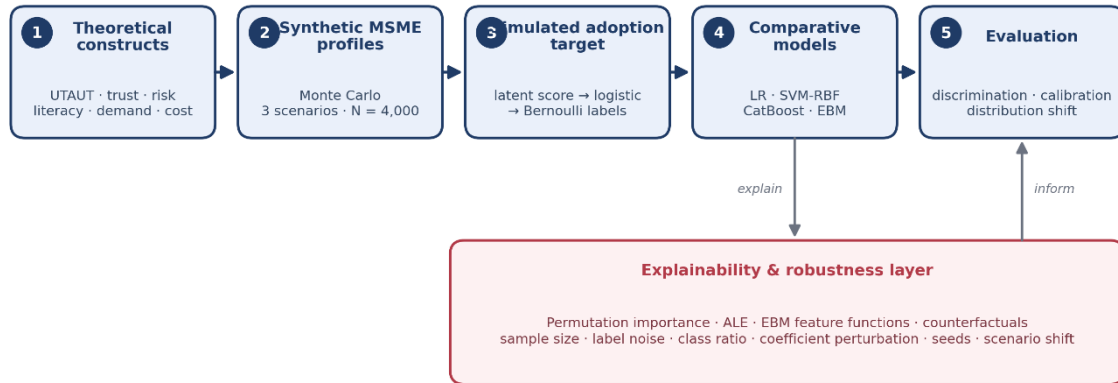


Fig 1. Conceptual research framework and computational workflow.

2.2 Variables

The binary target Y_i equals 1 for a simulated potential QRIS adopter and 0 for a simulated potential non-adopter. The wording “simulated adoption status” is used consistently; “actual adoption status” is avoided because no real validation data are used. Eighteen predictors span four dimensions: technology-acceptance constructs, capability and market constructs, infrastructure, and business or owner characteristics (Table 1). Variables that directly reveal the target, such as the number of QRIS transactions or QRIS-generated revenue, are excluded from the predictor set to prevent leakage.

Table 1. Variables, measurement scales, and synthetic generation rules.

Code	Variable	Scale	Generation rule
PE	Performance expectancy	Likert 1–5	Discretized latent MVN construct
EE	Effort expectancy	Likert 1–5	Discretized latent MVN construct
SI	Social influence	Likert 1–5	Discretized latent MVN construct
FC	Facilitating conditions	Likert 1–5	Discretized latent MVN construct
TR	Trust	Likert 1–5	Discretized latent MVN construct
SEC	Perceived security	Likert 1–5	Discretized latent MVN construct
PR	Perceived risk	Likert 1–5	Discretized latent MVN construct
DFL	Digital financial literacy	Likert 1–5	Discretized latent MVN construct
CD	Customer demand	Likert 1–5	Discretized latent MVN construct
TC	Perceived transaction cost	Likert 1–5	Discretized latent MVN construct
IR	Internet reliability	Likert 1–5	Discretized latent MVN construct
OwnerAge	Owner age	18–65 years	Truncated Normal(40, 10, 18, 65)
Education	Education level	Ordinal 1–5	Categorical distribution
BusinessAge	Business age	Years	Shifted Poisson(6), age-consistent
RevenueCat	Monthly revenue category	Ordinal 1–4	Size-linked categorical rule
Employees	Number of employees	1–25	Zero-truncated Poisson(3)
Urban	Urban location	Binary	Bernoulli, scenario-specific
Smartphone	Smartphone availability	Binary	Bernoulli, literacy-linked

2.3 Synthetic Profile Generation

Latent correlated constructs are generated from a multivariate normal distribution,

$$Z \sim N(\mu, \Sigma) \quad (1)$$

where the covariance matrix Σ encodes theoretically plausible relations, including a positive correlation between trust and perceived security ($r = 0.50$), between digital financial literacy and facilitating conditions ($r = 0.40$), and between facilitating conditions and internet reliability ($r = 0.45$), together with a negative correlation between perceived risk and trust ($r = -0.40$). Latent values are discretized into five-point Likert categories using fixed thresholds at -1.5 , -0.5 , 0.5 , and 1.5 . Scenario-specific mean shifts of ± 0.3 to ± 0.7 standard deviations on literacy, infrastructure, trust, demand, risk, and cost constructs define the digitally enabling and digitally constrained environments; the moderate environment uses zero shifts. Business variables follow bounded distributions consistent with Table 1, and monthly revenue is deliberately represented as an ordinal category so that no unwarranted currency-level precision is implied.

2.4 Data-Generating Process for the Simulated Target

The simulated adoption target is produced by a latent adoption score that combines linear, quadratic, interaction, threshold, and stochastic components. All coefficients were frozen before any model training and are varied only during sensitivity testing. The linear base score is

$$A_i = \beta_0 + \sum_{j=1}^{13} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad (2)$$

with positive weights on performance expectancy (0.55), digital financial literacy (0.60), facilitating conditions (0.50), customer demand (0.50), trust (0.40), smartphone availability (0.45), and related enablers, and negative weights on perceived risk (-0.55) and perceived transaction cost (-0.35); Likert predictors are centered at 3 and $\sigma = 0.90$. The augmented score adds nonlinear structure,

$$A_i^* = A_i + \gamma_1 \text{sgn}(d_i) d_i^2 + \gamma_2 (FC - 3)(CD - 3) + \gamma_3 (TR - 3)(SEC - 3) + h(FC_i) \quad (3)$$

where $d_i = DFL_i - 3$, $\gamma_1 = 0.10$, $\gamma_2 = 0.15$, $\gamma_3 = 0.12$, and the facilitating-conditions threshold function is

$$h(FC) = \{0, FC \leq 3; \delta(FC - 3), FC > 3\} \quad (4)$$

with $\delta = 0.30$. The adoption probability and binary label follow the logistic-Bernoulli mechanism,

$$P_i = \frac{1}{1 + \exp(-A_i^*)}, \quad Y_i \sim \text{Bernoulli}(P_i) \quad (5)$$

The intercept $\beta_0 = -0.35$ yields a near-balanced moderate scenario; the same frozen mechanism applied to shifted construct distributions produces higher simulated prevalence in the enabling scenario and pronounced imbalance in the constrained scenario. No coefficient was tuned to favor any particular learning algorithm.

2.5 Simulation Scenarios and Data-Quality Checks

Table 2. Simulation scenario specifications and realized class distributions.

Scenario	Main conditions	Profiles	Simulated adopter share
Moderate	Zero construct shifts; balanced environment	2,000	54.2%
Digitally enabling	+0.3 to +0.6 SD on enablers; $-0.4/-0.5$ SD on risk and cost	1,000	77.0%
Digitally constrained	-0.3 to -0.7 SD on enablers; $+0.5/+0.6$ SD on risk and cost	1,000	25.7%

The three scenarios are controlled simulation environments and are not labeled as empirical descriptions of Indonesian provinces or cities. Data-quality auditing covered range validation of every variable, logical consistency between owner age and business age, inspection of category proportions and distribution shapes, correlation-matrix verification against the intended Σ , duplicate-profile screening, class-distribution inspection per scenario, and an explicit target-leakage audit confirming that no usage-derived variable enters the predictor set.

2.6 Predictive Models

Four models representing distinct algorithmic families are compared. Logistic Regression provides a transparent linear baseline that estimates adoption probability through the logit link (Hastie et al., 2009):

$$P(Y = 1 | x) = \frac{1}{1 + \exp[-(\beta_0 + \sum_j \beta_j x_j)]} \quad (6)$$

SVM-RBF acts as a nonlinear margin-based comparator (Cortes & Vapnik, 1995), with kernel

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (7)$$

CatBoost is the high-performance gradient-boosting model for mixed tabular data [8], producing an additive ensemble of M oblivious trees:

$$\hat{y}(x) = \sum_{m=1}^M \eta f_m(x) \quad (8)$$

The Explainable Boosting Machine is an intrinsically interpretable generalized additive model with learned nonlinear feature functions and selected pairwise interactions (Caruana et al., 2015; Lou et al., 2013; Nori et al., 2019):

$$g(E[Y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{j < k} f_{jk}(x_j, x_k) \quad (9)$$

2.7 Training, Validation, and Hyperparameter Optimization

The moderate-scenario dataset was split into 1400 training, 300 validation, and 300 test profiles (70:15:15) with stratification on the target. Stratified five-fold cross-validation was applied within the training partition for model comparison; hyperparameters were selected by randomized search with three-fold internal cross-validation on ROC-AUC (Pedregosa, 2011); final models were refitted on the combined training and validation data; and the untouched test partition was reserved for the final evaluation. All preprocessing, including standardization for the linear and kernel models, was learned from training folds only. Table 3 reports the search spaces and the selected configurations.

Table 3. Hyperparameter search spaces and selected configurations.

Model	Search space (randomized search, 3-fold CV, ROC-AUC)	Selected configuration
Logistic Regression	$C \in \text{logspace}(10^{-3}, 10^2)$; penalty = L2; class weight $\in \{\text{none}, \text{balanced}\}$	$C = 0.0015$; L2; no class weight
SVM-RBF	$C \in \text{logspace}(10^{-2}, 10^2)$; $\gamma \in \text{logspace}(10^{-3}, 10^0)$; class weight $\in \{\text{none}, \text{balanced}\}$	$C = 5.46$; $\gamma = 0.001$; balanced
CatBoost	depth $\in \{4, 5, 6\}$; learning rate $\in \{0.03-0.12\}$; iterations $\in \{300-600\}$; L2 $\in \{1-9\}$	depth = 5; lr = 0.03; 300 iterations; L2 = 1
EBM	learning rate $\in \{0.01-0.05\}$; max bins $\in \{32, 64\}$; interactions $\in \{5-10\}$	lr = 0.02; 64 bins; 5 interactions

2.8 Performance and Calibration Metrics

Discrimination and classification quality are summarized by accuracy, balanced accuracy, precision, recall, and the F1 score,

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

together with ROC-AUC for threshold-free discrimination and PR-AUC, which is more informative when the adopter class is less frequent (Saito & Rehmsmeier, 2015). Probability quality is assessed with reliability diagrams and the Brier score (Brier, 1950; Niculescu-Mizil & Caruana, 2005)

$$BS = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2 \quad (11)$$

Model selection follows a multi-criteria logic: a model is preferred when it offers a defensible balance of discrimination, calibration, stability, and interpretability rather than the single highest accuracy.

2.9 Explainability Methods

Global importance is computed with permutation feature importance on the independent test set (Fisher et al., 2019; Molnar, 2022) defined for feature j as the decrease in ROC-AUC after permuting that feature:

$$PFI_j = \text{Score}(\text{original}) - \text{Score}(\text{feature } j \text{ permuted}) \quad (12)$$

Feature-effect shapes are examined with Accumulated Local Effects, which remain valid under correlated predictors by accumulating local finite differences of the prediction function across the ordered support of each feature (Apley & Zhu, 2020). For an ordinal feature with grid points $z_0 < z_1 < \dots < z_k$, the uncentered ALE at z_k is

$$ALE(z_k) = \sum_{k=1}^K E[f(z_k, x_{-j}) - f(z_{k-1}, x_{-j})] \quad (13)$$

with expectation taken over profiles in the corresponding local interval, followed by mean-centering. Intrinsic explanations are read directly from the fitted EBM feature functions and interaction terms. Actionable local

explanations use counterfactual search (Apley & Zhu, 2020; Mothilal et al., 2020): for a profile x , the counterfactual x^{cf} solves

$$x^{cf} = \arg \min [\lambda L(f(x'), y_{target}) + d(x, x') + \Omega(x, x')] \quad (14)$$

implemented as a greedy one-step coordinate search over an actionability-constrained feature set. Immutable or difficult-to-change variables (owner age, education, business age, revenue category, number of employees, and location) are frozen; only literacy, facilitating conditions, risk, trust, security, demand, cost, internet reliability, and smartphone availability may move, in unit Likert steps within their admissible ranges. Counterfactuals are assessed on validity, proximity, and sparsity, and are reported strictly as model-based recommendations under the simulation.

2.10 Robustness, Sensitivity, and Explanation Stability

Robustness analysis regenerates data and refits every model across five independent seeds for each configuration: sample sizes $N \in \{500; 1,000; 2,000; 4,000\}$; label-noise levels of 0%, 5%, 10%, and 15% created by random label flips; class-prevalence settings obtained by shifting the generator intercept (realized adopter shares of approximately 53%, 45%, 37%, and 26%); and multiplicative perturbations of all generator coefficients by $\pm 10\%$ and $\pm 20\%$. Distribution-shift evaluation trains on the moderate environment and tests on the enabling and constrained environments without refitting (Moreno-Torres et al., 2012). Explanation stability is quantified by the Spearman rank correlation between feature-importance rankings,

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)} \quad (15)$$

computed across model pairs and across repeated CatBoost fits with different seeds. All experiments were implemented in Python 3.12 with scikit-learn 1.8, CatBoost 1.2, and InterpretML 0.7 [8], [11], [21]; every random seed, distribution parameter, coefficient, and search space is disclosed to support full reproducibility.

3. Result and Discussion

3.1 Synthetic Dataset

The final corpus contains 4,000 synthetic profiles across three scenarios. All Likert variables remained within the 1–5 range, owner age within 18–65 years, and no duplicate profiles or leakage variables were detected. Table 4 summarizes selected descriptive statistics by scenario, and Figure 2 shows the realized class distributions. Under the frozen generator, the simulated adopter share is 54.2% in the moderate environment, rises to 77.0% under enabling conditions, and falls to 25.7% under constrained conditions; these values are outcomes of the simulation parameters, not observed adoption rates. The empirical correlation matrix (Figure 3) reproduces the intended construct structure, including the trust–security, literacy–infrastructure, and risk–trust relations, with no unintended near-perfect dependencies.

Table 4. Descriptive statistics, mean (SD), for selected predictors by scenario.

Variable	Moderate	Enabling	Constrained
PE	2.98 (1.01)	3.02 (1.00)	3.00 (1.01)
FC	3.03 (1.03)	3.55 (0.97)	2.34 (0.97)
TR	2.98 (1.02)	3.45 (1.00)	2.63 (0.99)
PR	2.99 (0.99)	2.53 (0.99)	3.54 (0.97)
DFL	3.02 (1.03)	3.56 (0.97)	2.43 (0.99)
CD	3.00 (1.00)	3.40 (1.01)	2.56 (0.99)
IR	3.04 (1.03)	3.48 (0.99)	2.31 (0.98)
OwnerAge	40.39 (9.60)	39.46 (10.15)	39.79 (9.92)
Employees	3.06 (1.65)	3.08 (1.68)	3.10 (1.64)

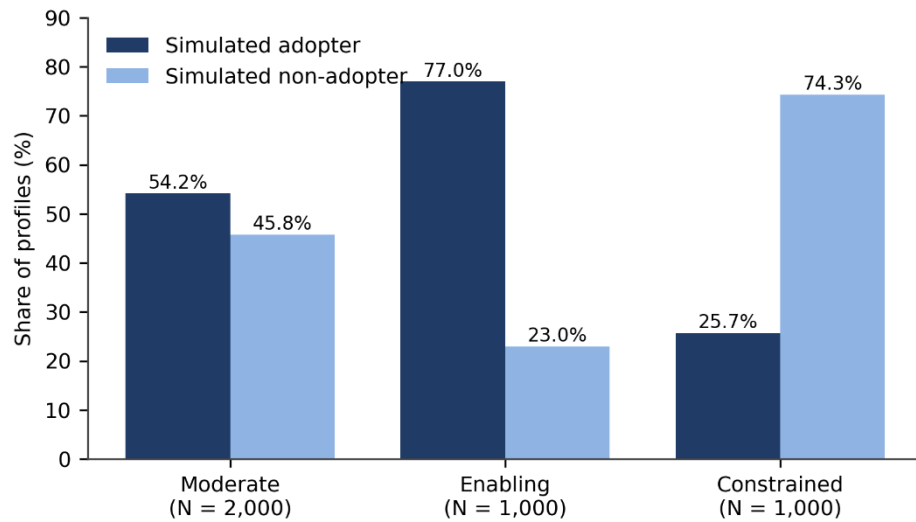


Fig 2. Simulated class distribution by scenario.

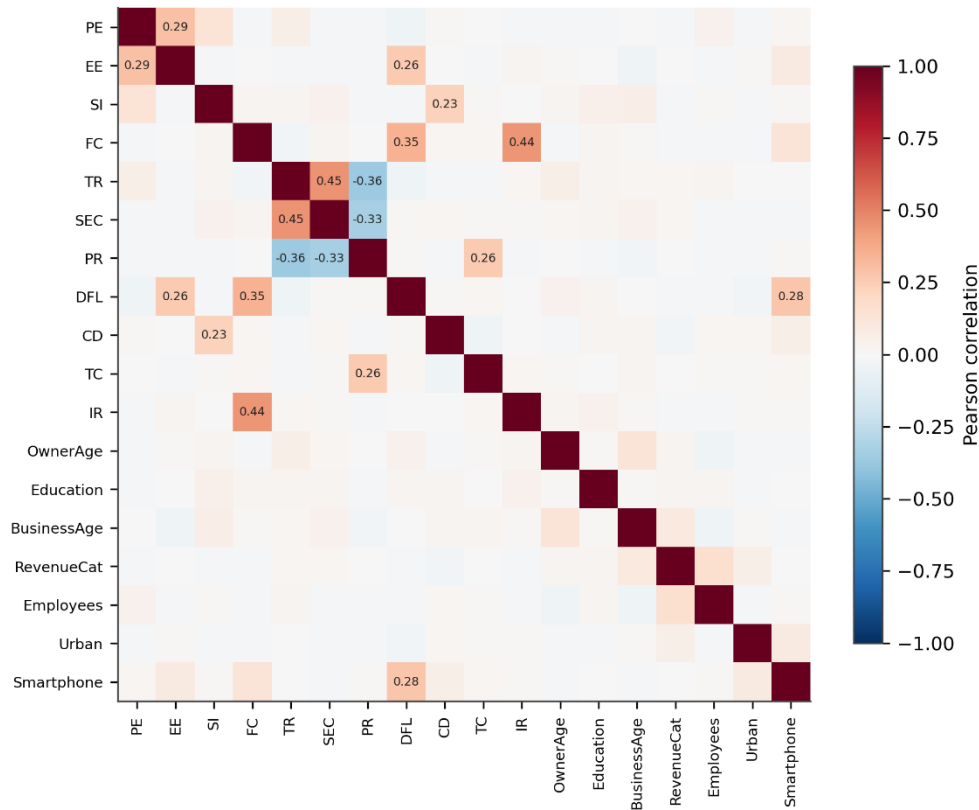


Fig 3. Empirical correlation matrix of the moderate-scenario predictors.

3.2 Cross-Validation Performance

Table 5 reports stratified five-fold cross-validation results on the training partition, expressed as mean $\pm$ standard deviation across folds. The four models are statistically close: cross-validated ROC-AUC spans 0.846 to 0.855, and balanced accuracy spans 0.753 to 0.760, with overlapping fold-level dispersion for every pair. This proximity is an expected and informative property of the simulated environment: the generator contains strong additive structure with moderate nonlinearity, so flexible learners cannot dramatically outperform a well-regularized linear model. Model rankings are therefore not inferred from any single fold or seed.

Table 5. Five-fold cross-validation performance on the training partition (mean $\pm$ SD).

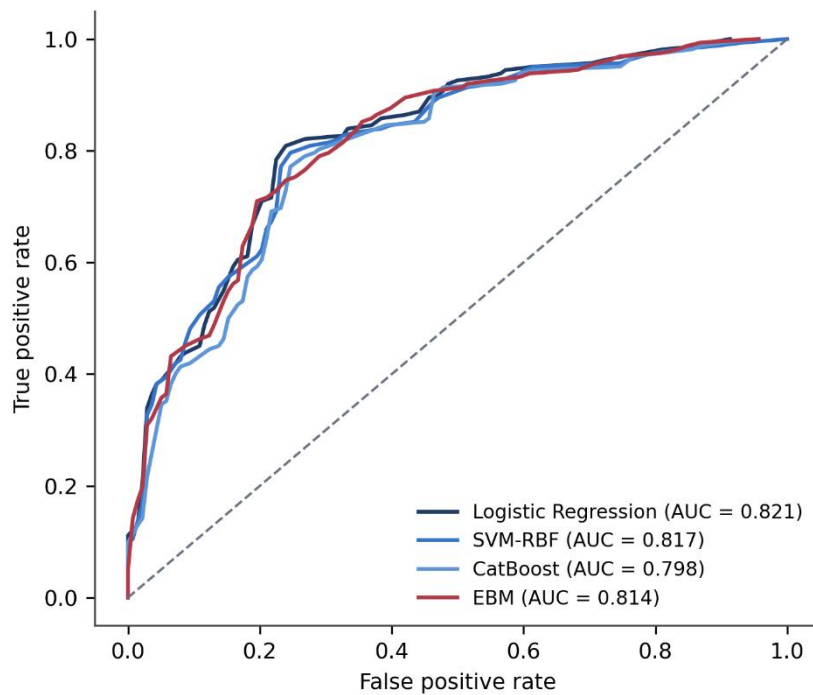
Model	Balanced acc.	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.754 $\pm$ 0.022	0.741 $\pm$ 0.020	0.868 $\pm$ 0.024	0.799 $\pm$ 0.019	0.855 $\pm$ 0.022
SVM-RBF	0.760 $\pm$ 0.032	0.770 $\pm$ 0.035	0.809 $\pm$ 0.029	0.788 $\pm$ 0.027	0.855 $\pm$ 0.019
CatBoost	0.753 $\pm$ 0.027	0.764 $\pm$ 0.024	0.798 $\pm$ 0.039	0.780 $\pm$ 0.027	0.847 $\pm$ 0.017
EBM	0.760 $\pm$ 0.027	0.770 $\pm$ 0.030	0.808 $\pm$ 0.022	0.788 $\pm$ 0.021	0.846 $\pm$ 0.019

3.3 Independent Test-Set Performance

Table 6 presents the final evaluation on the untouched test partition, and Figures 4 and 5 display the ROC and precision–recall curves. SVM-RBF attains the highest balanced accuracy (0.767) and F1 (0.792), Logistic Regression the highest ROC-AUC (0.821) and PR-AUC (0.835), while CatBoost and EBM remain within roughly one to two points of the leaders on every metric. Confusion matrices (Figure 6) show broadly symmetric error patterns, with the balanced-class-weighted SVM trading a small amount of precision for improved recall on the adopter class.

Table 6. Independent test-set performance (N = 300).

Model	Balanced acc.	Precision	Recall	F1	ROC-AUC	PR-AUC
Logistic Regression	0.739	0.731	0.840	0.782	0.821	0.835
SVM-RBF	0.767	0.775	0.809	0.792	0.817	0.831
CatBoost	0.761	0.776	0.790	0.783	0.798	0.812
EBM	0.747	0.757	0.790	0.773	0.814	0.828

**Fig 4.** ROC curves on the independent test set.

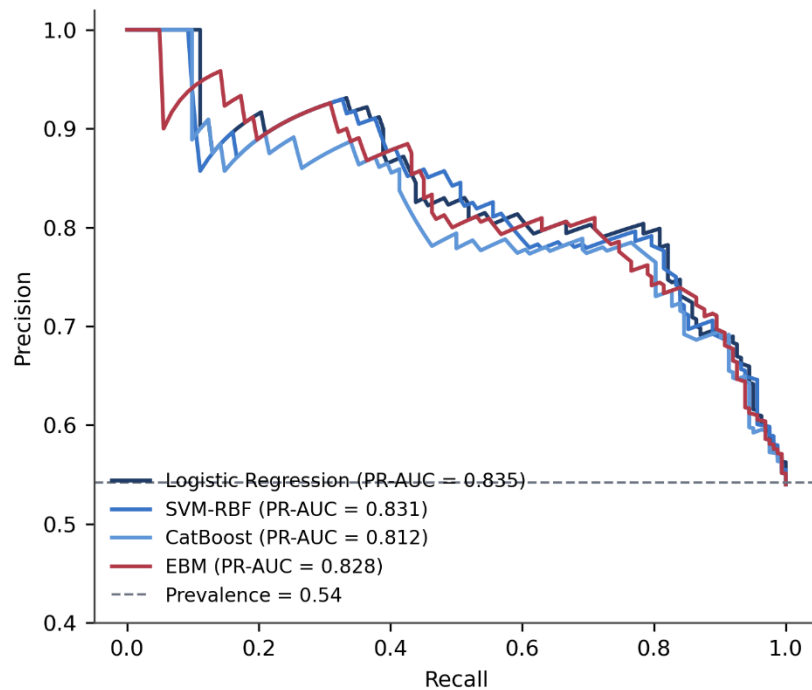


Fig 5. Precision–recall curves on the independent test set.

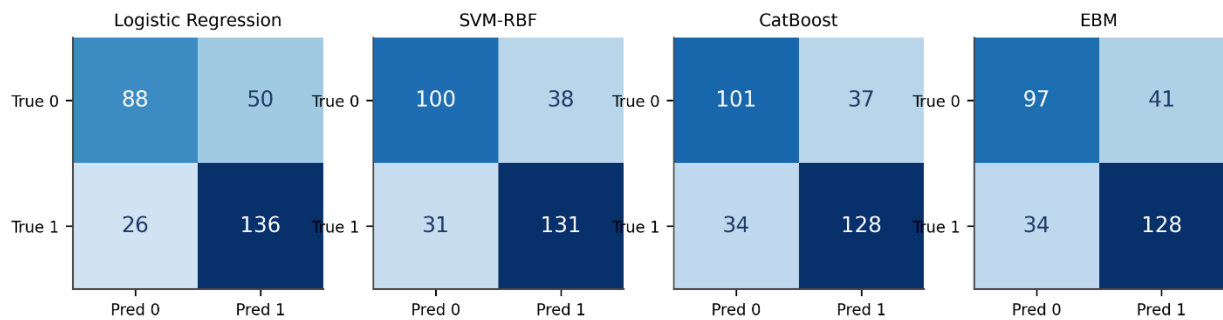


Fig 6. Confusion matrices on the independent test set.

3.4 Probability Calibration

Figure 7 and Table 7 assess probability quality. SVM-RBF with Platt-style probability estimates achieves the lowest Brier score (0.173), followed closely by EBM (0.176); Logistic Regression, despite its strong discrimination, produces the highest Brier score (0.191) because its heavily regularized configuration compresses predicted probabilities toward the center. The result illustrates a central methodological point: a model with high ROC-AUC may still deliver comparatively weak probability estimates, and calibration must be examined explicitly whenever predicted probabilities will be used for segmentation or prioritization.

Table 7. Probability-calibration summary on the test set.

Model	Brier score	Reliability-curve pattern
Logistic Regression	0.191	Under-confident; probabilities compressed toward 0.5
SVM-RBF	0.173	Closest to the diagonal across bins
CatBoost	0.182	Mildly miscalibrated in central bins
EBM	0.176	Near-diagonal with slight central deviation

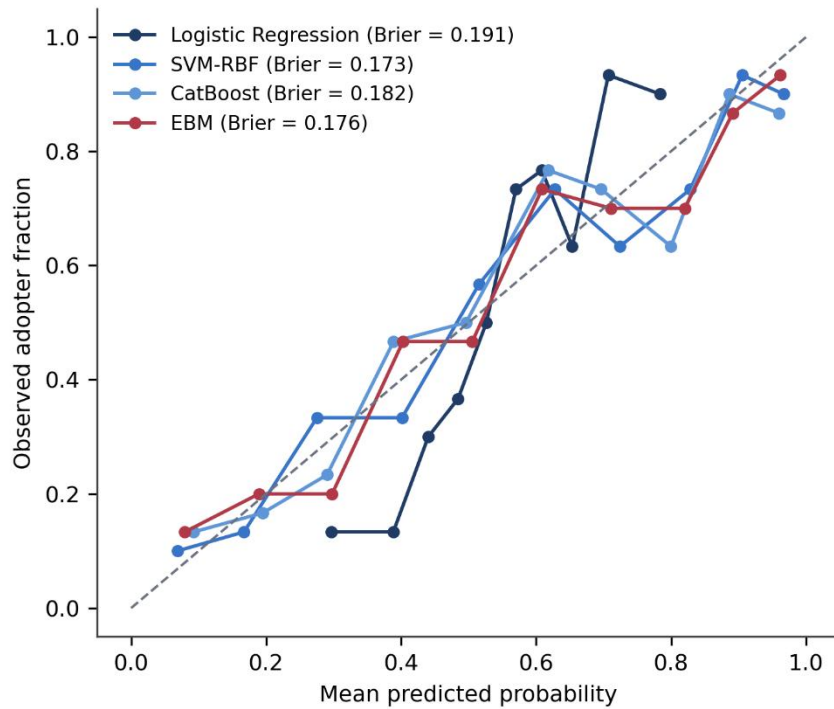


Fig 7. Reliability diagrams (quantile-binned) on the independent test set.

3.5 Performance under Distribution Shift

Table 8 reports zero-refit transfer of the moderately trained models to the enabling and constrained environments. Ranking discrimination is essentially preserved: ROC-AUC remains between 0.825 and 0.853 across all model–scenario pairs, indicating that the learned score orders shifted profiles nearly as well as in-distribution profiles. Balanced accuracy at the fixed 0.5 threshold, however, degrades by roughly six to nine points, because prevalence shifts move the optimal operating threshold. Brier scores improve mechanically in the enabling scenario, where the class distribution is more extreme. The practical implication is that threshold recalibration, rather than full model retraining, is the first remedy when the deployment environment shifts.

Table 8. Zero-refit performance under distribution shift (trained on the moderate environment).

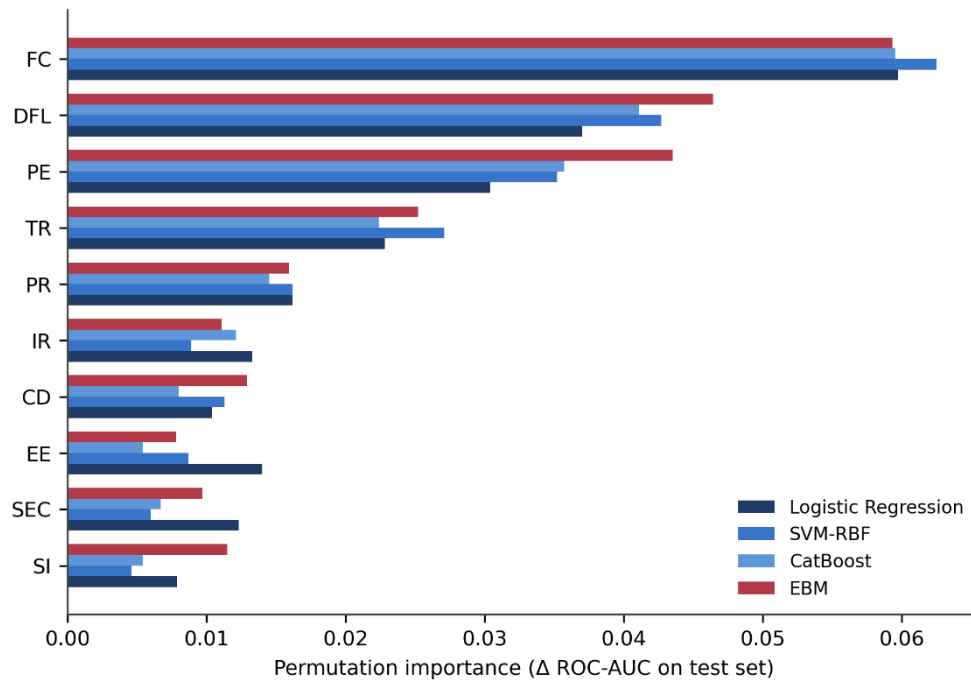
Model	Enab. bal. acc.	Enab. ROC-AUC	Constr. bal. acc.	Constr. ROC-AUC	Constr. PR-AUC
Logistic Regression	0.652	0.851	0.699	0.836	0.651
SVM-RBF	0.677	0.853	0.680	0.833	0.643
CatBoost	0.690	0.848	0.686	0.825	0.631
EBM	0.678	0.846	0.686	0.827	0.634

3.6 Global Feature Importance and Cross-Model Agreement

Figure 8 and Table 9 report permutation feature importance computed on the test set with ROC-AUC as the scoring basis. All four model families agree on the leading predictors of the simulated outcome: facilitating conditions ($\Delta\text{AUC} \approx 0.063$), digital financial literacy, performance expectancy, trust, and perceived risk. Demographic and firmographic variables (owner age, education, employees, revenue category) contribute negligibly, consistent with their near-zero weights in the frozen generator. Cross-model rank agreement is very high (Spearman ρ between 0.91 and 0.98; Table 10), and CatBoost rankings across five refit seeds correlate at $\rho = 0.94 \pm 0.03$. Correct wording is maintained deliberately: facilitating conditions consistently contributed to the simulated model predictions across the four algorithms; no claim is made that they are primary determinants of QRIS adoption among Indonesian MSMEs.

Table 9. Top-eight permutation feature-importance values (Δ ROC-AUC) by model.

Feature	Logistic Regression	SVM-RBF	CatBoost	EBM
FC	0.060	0.063	0.059	0.059
DFL	0.037	0.043	0.041	0.046
PE	0.030	0.035	0.036	0.043
TR	0.023	0.027	0.022	0.025
PR	0.016	0.016	0.015	0.016
CD	0.010	0.011	0.008	0.013
IR	0.013	0.009	0.012	0.011
EE	0.014	0.009	0.005	0.008

**Fig 8.** Permutation feature importance across the four model families.**Table 10.** Explanation-stability results (Spearman rank correlations of importance rankings).

Comparison	Spearman ρ
Logistic Regression vs SVM-RBF	0.973
Logistic Regression vs CatBoost	0.955
Logistic Regression vs EBM	0.913
SVM-RBF vs CatBoost	0.967
SVM-RBF vs EBM	0.946
CatBoost vs EBM	0.977
CatBoost across 5 refit seeds (mean $\pm$ SD)	0.941 $\pm$ 0.025

3.7 Nonlinear and Interaction Effects

Figure 9 shows Accumulated Local Effects for four theoretically salient predictors, and Figure 10 shows the corresponding intrinsic EBM feature functions. Three patterns emerge consistently. First, digital financial literacy exhibits an accelerating positive effect: the ALE gain from level 4 to level 5 exceeds the gain from level 1 to level 2, mirroring the signed-quadratic term of the generator. Second, facilitating conditions display the built-in threshold: effects are flat to mildly positive up to level 3 and rise steeply beyond it, and the EBM recovers the same kink without being told it exists. Third, perceived risk shows a monotone negative effect that steepens at high risk levels. The linear-model ALE curves are, by construction, straighter; the agreement of CatBoost, EBM, and SVM-RBF on curvature demonstrates that the nonlinear structure is learned from data rather than assumed. EBM interaction terms rank

facilitating conditions $\times$ customer demand and trust $\times$ perceived security among the strongest pairs, matching the generator's interaction design.

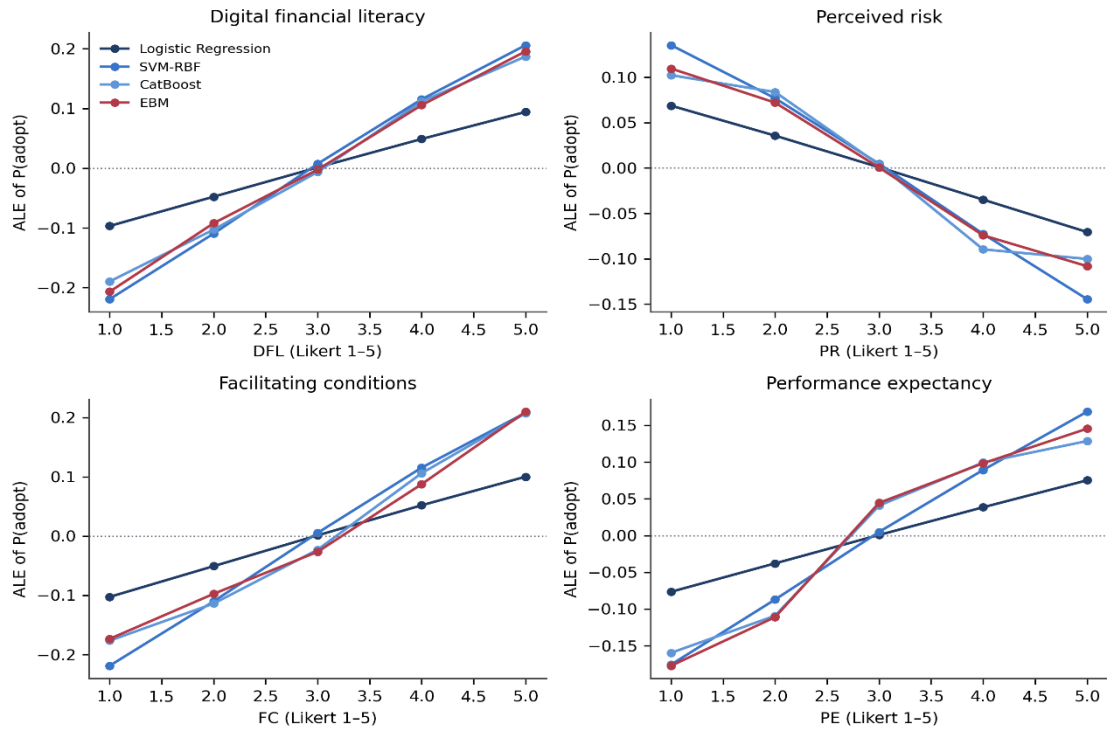


Fig 9. Accumulated Local Effects for digital financial literacy, perceived risk, facilitating conditions, and performance expectancy.

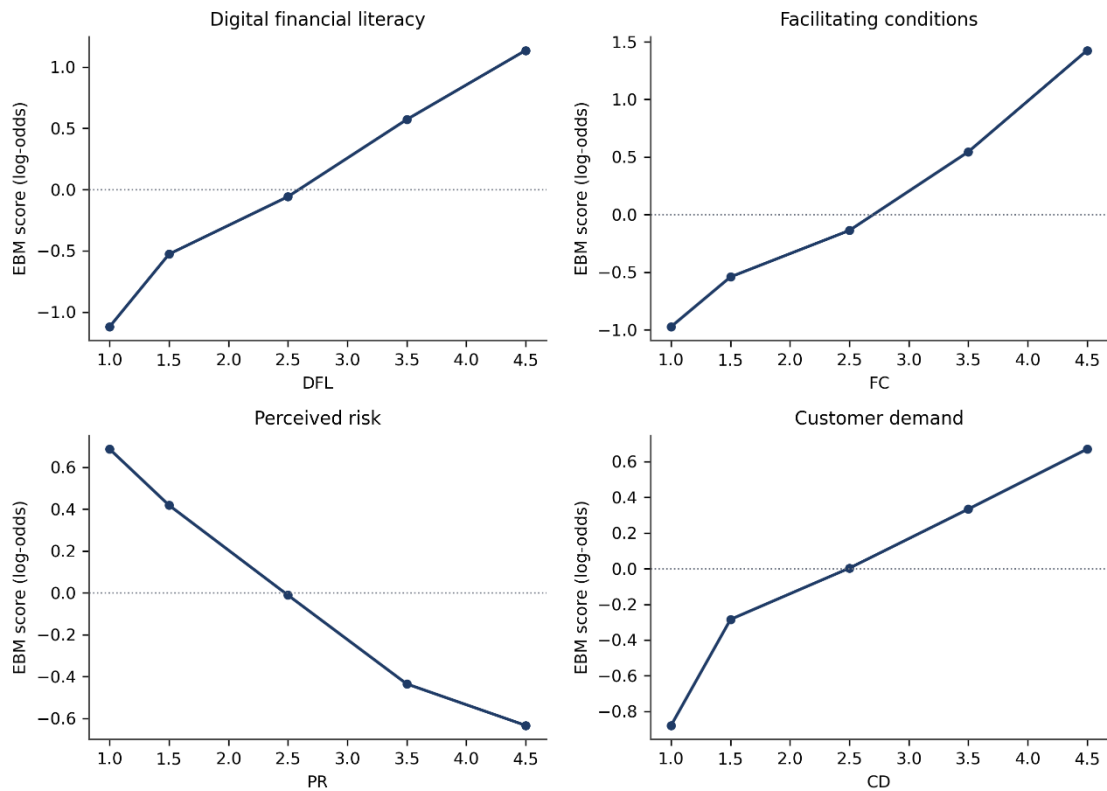


Fig 10. Intrinsic EBM feature functions (log-odds scale) for four leading predictors.

3.8 Local and Counterfactual Explanations

Table 11 presents counterfactual analyses for representative test profiles under the EBM reference model with the actionability constraints of Section 2.9. A borderline synthetic non-adopter (predicted probability 0.436) crosses the decision boundary with a single unit improvement in digital financial literacy, while a clearer non-adopter (0.295) requires a two-level literacy improvement. By contrast, profiles deep in either class are counterfactually stable: the highest-probability adopter retains a predicted probability of 0.909 even after four adverse unit changes, and the lowest-probability profile reaches only 0.131 after four favorable changes. This asymmetry is itself informative, indicating that sparse unit-step interventions matter most near the decision boundary. All counterfactual changes are model-based recommendations under the simulation and must not be read as experimentally validated causal interventions.

Table 11. Counterfactual explanations under the EBM reference model (actionable features only).

Profile	Initial $\hat{p}$	Final $\hat{p}$	Greedy unit changes applied
Borderline non-adopter	0.436	0.575	DFL: 3 $\rightarrow$ 4
Clear non-adopter	0.295	0.602	DFL: 3 $\rightarrow$ 4; DFL: 3 $\rightarrow$ 5
Uncertain case	0.501	0.336	FC: 4 $\rightarrow$ 3
High-probability adopter (reverse search)	0.993	0.909	FC: 5 $\rightarrow$ 4; FC: 5 $\rightarrow$ 3; Smartphone: 1 $\rightarrow$ 0; DFL: 4 $\rightarrow$ 3
Low-probability profile	0.012	0.131	DFL: 3 $\rightarrow$ 4; DFL: 3 $\rightarrow$ 5; Smartphone: 0 $\rightarrow$ 1; PR: 4 $\rightarrow$ 3

3.9 Robustness and Sensitivity Analysis

Figure 11 summarizes the robustness experiments (five independent seeds per configuration; mean $\pm$ SD). Panel (a) shows conventional learning-curve behavior: ROC-AUC improves from approximately 0.82–0.84 at $N = 500$ to 0.83–0.84 at $N = 4,000$, with shrinking seed variance and no model overtaking another decisively. Panel (b) shows graceful degradation under label noise: fifteen percent random flips reduce ROC-AUC by roughly 0.10 for the most affected model, with the four families degrading in near-parallel. Panel (c) shows that balanced accuracy declines moderately as the simulated adopter share falls from 53% to 26%, with the class-weighted SVM most resilient. Coefficient perturbation of $\pm 20\%$ changes mean ROC-AUC by at most 0.042 for any model, and the performance ordering of the four families is preserved in every perturbed configuration, indicating that the comparative conclusions are not artifacts of one specific coefficient vector.

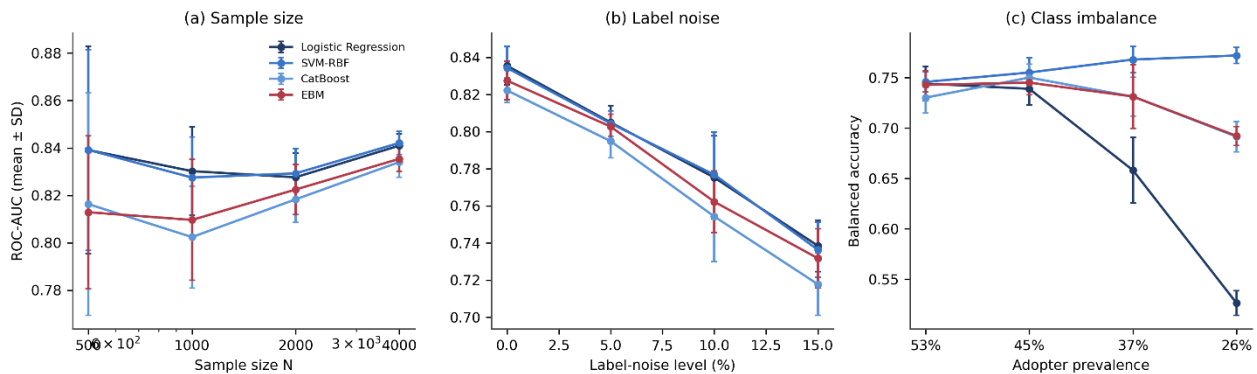


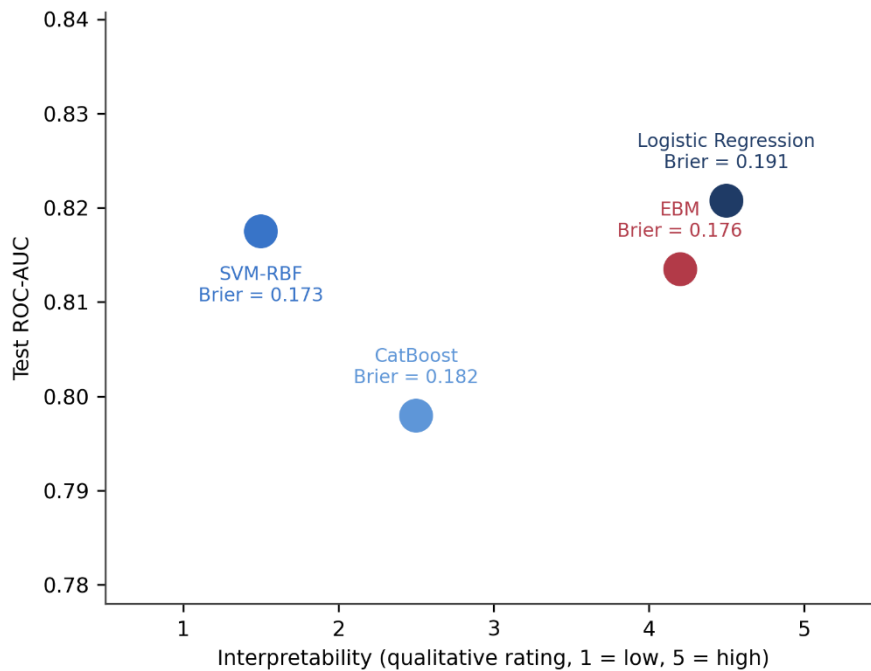
Fig 11. Robustness of ROC-AUC and balanced accuracy to (a) sample size, (b) label noise, and (c) class imbalance (mean $\pm$ SD over five seeds).

3.10 Accuracy–Interpretability Trade-Off

Table 12 and Figure 12 integrate the evidence. The maximum spread in test ROC-AUC across the four families is 0.023, which is smaller than typical fold-to-fold variation, whereas the interpretability gap between the glassbox models and SVM-RBF is categorical. Under the multi-criteria selection logic of Section 2.8, the EBM offers the strongest overall profile in this simulated environment: near-best discrimination, second-best calibration, intrinsic global and local transparency, and high explanation stability. SVM-RBF is preferable when calibrated probabilities at a fixed threshold are the dominant requirement, and Logistic Regression remains a fully defensible baseline when simplicity is paramount.

Table 12. Integrated accuracy–interpretability assessment.

Model	Test ROC-AUC	Brier	Global interpretability	Local interpretability	Cost
Logistic Regression	0.821	0.191	High (coefficients)	Moderate	Low
SVM-RBF	0.817	0.173	Low (post-hoc only)	Post-hoc / CF	Moderate
CatBoost	0.798	0.182	Moderate (PFI, ALE)	Post-hoc / CF	Moderate
EBM	0.814	0.176	High (feature functions)	High	Moderate

**Fig 12.** Accuracy–interpretability positioning of the four model families.

The nonlinear models did not decisively outperform the linear baseline in discrimination, and this null-like comparative result is substantively meaningful rather than disappointing. The frozen generator allocates most of its signal to additive main effects, adds moderate curvature and two interactions, and injects substantial latent noise ($\sigma = 0.90$) plus Bernoulli labeling noise; the Bayes-optimal ROC-AUC is therefore bounded well below one, and all four families approach that bound. The finding echoes the broader literature on tabular prediction: when the underlying signal is largely additive and the sample is moderate, well-regularized simple models are highly competitive (Breiman, 2001; Hastie et al., 2009). Where the families genuinely separate is calibration and transparency, which is precisely why the study's evaluation design integrates the Brier score, reliability diagrams, and interpretability audits instead of relying on discrimination alone (Niculescu-Mizil & Caruana, 2005).

Higher model complexity did not translate into higher practical value in this environment. The EBM matched the boosted and kernel models within noise while exposing its full functional form; its feature functions recovered the generator's threshold and curvature without access to the generation rules, providing a direct, verifiable audit of what the model learned (Caruana et al., 2015; Lou et al., 2013). This supports the position that for tabular decision-support problems, intrinsically interpretable models should be serious candidates rather than fallbacks, and that the common assumption of a steep accuracy–interpretability trade-off frequently fails on structured data (Rudin, 2019). Model selection is thus best framed as a multi-dimensional decision across discrimination, calibration, stability, transparency, and computational cost.

Every interpretive statement in this study is tied to the data-generating assumptions. The dominance of facilitating conditions, digital financial literacy, and performance expectancy in permutation importance reflects their generator weights amplified by the threshold and quadratic terms (Harahap et al., 2026); the recovered ALE shapes demonstrate that model-agnostic explanation methods can faithfully reconstruct known nonlinear mechanisms under correlated predictors (Apley & Zhu, 2020). The counterfactual asymmetry, in which boundary-adjacent profiles flip with one or two unit changes while extreme profiles are stable, illustrates a property that would matter in a validated deployment:

intervention effort is most productive for near-boundary merchants, and sparse, plausible, actionability-constrained recommendations are obtainable from glassbox models without auxiliary approximation (Mothilal et al., 2020; Wachter et al., 2018). None of these observations establishes real-world causal efficacy; they establish that the explanatory machinery behaves coherently when ground truth is known, which is the appropriate standard of evidence for a synthetic proof of concept.

Methodologically, the study contributes an integrated evaluation template that couples discrimination and calibration metrics with explanation-stability quantification, counterfactual auditing, and a five-axis robustness battery (sample size, noise, imbalance, coefficient perturbation, distribution shift), executed over repeated seeds and reported as mean $\pm$ dispersion. The circular-validation risk inherent to synthetic targets is managed explicitly: label noise prevents trivial rule recovery, coefficient perturbation shows that conclusions survive generator misspecification, and distribution shift tests transfer beyond the training distribution (Moreno-Torres et al., 2012). Theoretically, the study shows how UTAUT-family constructs can be operationalized inside a predictive, individually resolved framework that accommodates thresholds and interactions which population-average structural models would smooth away (Venkatesh Morris, M. G., Davis, G. B., & Davis, F. D, 2003; Venkatesh et al., 2012); this is offered as a complementary representation, not a validation of any acceptance theory.

After validation with real MSME data, the framework could support digital-readiness mapping, prioritization of literacy programs, targeted QRIS onboarding assistance, and calibrated segmentation of prospective merchants, with the EBM supplying regulator-auditable explanations for each score. The distribution-shift results carry a concrete operational lesson: when a scoring model trained in one environment is applied to a structurally different region, threshold recalibration should precede model retraining, because ranking quality transfers better than the decision threshold. These implications are stated as potential applications conditional on empirical validation, not as policy recommendations derived from synthetic results.

All profiles and adoption outcomes are synthetic; feature distributions, correlations, and target rules embody researcher assumptions; and the models are, in part, rediscovering structure the authors created, so absolute performance figures carry no empirical meaning. Results do not estimate actual adoption prevalence, establish no causal relationships, and may be sensitive to data-generating choices beyond the $\pm 20\%$ perturbations examined. Explainability methods explain the fitted predictive models, not actual MSME behavior. The greedy counterfactual search returns sparse but not provably minimal changes, robustness used five seeds per configuration rather than an exhaustive design, and external validation on real observations remains outstanding.

4. Conclusion

This study developed and evaluated a comparative explainable machine learning framework for predicting simulated QRIS adoption from theory-informed synthetic MSME profiles. Across 4,000 profiles and three digital-environment scenarios, the four model families proved statistically close in discrimination (test ROC-AUC 0.798–0.821) while separating clearly on calibration and transparency: SVM-RBF delivered the best probability quality (Brier 0.173) and balanced accuracy (0.767), and the Explainable Boosting Machine provided the strongest overall balance of near-best performance, second-best calibration, intrinsic interpretability, and high explanation stability. Global importance rankings were consistent across all four algorithms (Spearman $\rho \geq 0.91$) and across refit seeds, with facilitating conditions, digital financial literacy, and performance expectancy dominating the simulated predictions; ALE and EBM feature functions recovered the generator's threshold and curvature; and constrained counterfactual search produced sparse, actionable unit-step recommendations for boundary-adjacent profiles. Performance degraded gracefully under label noise, class imbalance, smaller samples, $\pm 20\%$ coefficient perturbation, and zero-refit distribution shift, with model ordering preserved throughout. The findings demonstrate the computational feasibility of a transparent, calibrated, and robustness-audited QRIS-adoption prediction pipeline and contribute a fully reproducible synthetic-data experiment; they do not estimate adoption prevalence, validate determinants, or establish causal effects in the Indonesian MSME population. Empirical validation with real merchant data is the essential next step before any operational use.

References

Agustiana, A., Mutiara Putri, T., Amallia Nurvianti, R., Dwi Utami, L., Soliha, I., syahwildan, M., Studi Manajemen, P., Ekonomi dan Bisnis, F., Pelita Bangsa, U., & Bekasi, K. (2025). Analisis Kritis QRIS dan GPN dalam

- Memperkuat Kedaulatan Ekonomi Digital Indonesia (Tinjauan Laporan USTR 2025). *Ikraith-Ekonomika*, 8(2), 708–720. <https://journals.upi-yai.ac.id/index.php/IKRAITH-EKONOMIKA/article/view/4968>
- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B*, 82(4), 1059–1086.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission BT - Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730).
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42.
- Hamzah Muchtar, E., Trianto, B., Maulana, I., Alim, M. N., Marasabessy, R. H., Hidayat, W., Junaedi, E., & Masrizal. (2024). Quick response code Indonesia standard (QRIS) E-payment adoption: customers perspective. *Cogent Business & Management*, 11(1). <https://doi.org/10.1080/23311975.2024.2316044>
- Harahap, M. A. K., Harmain, U., Purwoko, A., & Nasution, A. A. (2026). Urban Green Design in Public Infrastructure Planning: A Structured Literature Review for Smart, Livable, and Sustainable Competitive Cities. *International Journal of Sustainable Development and Planning*, 21(4). <https://doi.org/10.18280/ijstdp.210436>
- Harahap, M. A. K., Tarigan, I. M., Sipayung, S. P. B., Nasution, A. A., & Syaifuddin, M. (2023). A Biblio-Systematic Analysis of Development Planning for Better Environment: A Case Study of 4 Countries. *Civil Engineering and Architecture*, 11(3), 1209–1219. <https://doi.org/10.13189/cea.2023.110307>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. <https://doi.org/10.1007/978-0-387-84858-7>
- Indonesia, B. (2019). *Peraturan Anggota Dewan Gubernur No. 21/18/PADG/2019 tentang Implementasi Standar Nasional Quick Response Code untuk Pembayaran*. Bank Indonesia.
- Jordon, J. et al. (2022). Synthetic data — what, why and how? *ArXiv Preprint ArXiv:2205.03257*.
- Juniariani, N. M. R., Sanjaya, I. K. P. W., & Mariyatni, N. P. S. (2020). Pengelolaan Keuangan, Strategi Pemasaran, Dan Jiwa Kewirausahaan Untuk Mencapai Keunggulan Bersaing. *BISMA: Jurnal Bisnis Dan Manajemen*, 14(2), 125–132. <https://doi.org/10.19184/bisma.v14i2.17706>
- Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). *Accurate intelligible models with pairwise interactions BT - Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623–631).
- Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.).
- Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1), 521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>
- Mothilal, R. K., Sharma, A., & Tan, C. (2020). *Explaining machine learning classifiers through diverse counterfactual explanations BT - Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (pp. 607–617).
- Muhammad Hilmansyah Lukman, Dafin Firmansyah, Denada Nandini, Mochammad Zahid Khairu Ghani, Zelda Saleha Fermansyah, & Syti Sarah Maesaroh. (2025). Mekanisme Adopsi QRIS di Retail: Studi Empiris atas Fleksibilitas, Efisiensi Waktu, dan Protektivitas dalam Meningkatkan Transaksi Digital. *Jurnal Riset Rumpun Ilmu Ekonomi*,

4(2), 36–48. <https://doi.org/10.55606/jurrie.v4i2.5386>

- Niculescu-Mizil, A., & Caruana, R. (2005). *Predicting good probabilities with supervised learning BT - Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632).
- Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *ArXiv*.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, 9(2), 461–468.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3).
- Suryani, S., Ramadhan, D. S., Ekonomi, F., & Riau, U. I. (2017). Analisis Literasi Keuangan pada Pelaku Usaha Mikro di Kota Pekanbaru. *COSTING: Journal of Economic, Business and Accounting*, 1(1), 12–22. <https://doi.org/10.31539/COSTING.V1I1.67>
- Tambunan, T. T. H. (2021). EXPORT OF INDONESIAN MSEs AND THE ROLE OF PARTNERSHIP. *Journal of Developing Economies*, 6(2), 235. <https://doi.org/10.20473/jde.v6i2.28747>
- Venkatesh Morris, M. G., Davis, G. B., & Davis, F. D, V. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157–178.
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2), 841–887.