“Explainable AI for delinquency risk monitoring in U.S. auto lease securitizations”

AUTHORS	Zhanna Dryha Oleksandr Levchenko Larysa Sergiienko Liubomyr Kochubei Oleksiy Domashenko Kateryna Shcheglova

ARTICLE INFO	Zhanna Dryha, Oleksandr Levchenko, Larysa Sergiienko, Liubomyr Kochubei, Oleksiy Domashenko and Kateryna Shcheglova (2026). Explainable AI for delinquency risk monitoring in U.S. auto lease securitizations. *Investment Management and Financial Innovations*, *23*(3), 215–234. doi:10.21511/imfi.23(3).2026.16
DOI	http://dx.doi.org/10.21511/imfi.23(3).2026.16
RELEASED ON	Thursday, 13 August 2026
RECEIVED ON	Thursday, 02 July 2026
ACCEPTED ON	Tuesday, 04 August 2026

JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
29	4	16

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

Type of the article: Research Article

Received on: 2nd of July, 2026
Accepted on: 4th of August, 2026
Published on: 13th of August, 2026

Zhanna Dryha, Doctor of Economics, Professor, Department of National Security, Public Administration and Administration, Zhytomyr Polytechnic State University, Ukraine.

Oleksandr Levchenko, Ph.D. in Economics, Expert in Financial Regulation and Prudential Supervision, Certified Bank Auditor, Ukraine. (Corresponding author)

Larysa Sergiienko, D.Sc. in Public Administration, Associate Professor, Dean of the Faculty of National Security, Law and International Relations, Zhytomyr Polytechnic State University, Ukraine.

Liubomyr Kochubei, Master of Law, Master of Public Administration, Attorney, Chairman of the IT Lawyers Bar Association, Ukraine.

Oleksiy Domashenko, MBA, EA, Registered Investment Advisor, Accounting and Tax Expert, USA.

Kateryna Shcheglova, MSc, Strategic Marketing, London Metropolitan University, CJBS Research Associate, Certified AI Specialist (CAIP by Oxethica), Certified ISO 42001, AI certified by Saïd Business School, University of Oxford, CISI AI Ethics, Regulatory Innovations and AI Expert, UK.

Conflict of interest statement:
Author(s) reported no conflict of interest

Zhanna Dryha (Ukraine), **Oleksandr Levchenko** (Ukraine), **Larysa Sergiienko** (Ukraine), **Liubomyr Kochubei** (Ukraine), **Oleksiy Domashenko** (USA), **Kateryna Shcheglova** (UK)

EXPLAINABLE AI FOR DELINQUENCY RISK MONITORING IN U.S. AUTO LEASE SECURITIZATIONS

Abstract

This study evaluates whether explainable machine-learning models can provide reliable, operationally interpretable short-horizon delinquency monitoring within U.S. auto lease securitization panels. Six public SEC ABS-EE trust-family panels were harmonized at the contract-month level. The primary outcome is one-month-ahead incident escalation to 30 or more days past due among contracts below 30 days past due at the feature month. Fully tuned penalized logistic regression (M1), unconstrained gradient boosting (M2), and governance-constrained gradient boosting (M3/X-LEASE) were assessed through chronological development, calibration, locked out-of-time testing, contract-held-out validation, six leave-one-issuer-out experiments, and later temporal evaluation. The locked test comprised 198,301 observations and 768 events. M2 produced the strongest discrimination (AUC-ROC 0.8151; PR-AUC 0.0269), while M3 exceeded the logistic benchmark by PR-AUC but did not outperform M2. At an exact 1% review capacity, each model generated 1,984 alerts; M2 detected 104 events, compared with 61 for M3, so the hypothesized recall advantage of X-LEASE was not supported. Full-sample TreeSHAP analysis showed stable feature rankings across tested issuers, later periods, and contract-level resampling. M3 relied more heavily on credit score and issuer controls and had a more concentrated explanation structure, but this does not establish superior auditability or fairness. The results support human-reviewed early-warning monitoring within the tested securitization panels, not lifetime default prediction, automated adverse decisions, or universal transferability.

Keywords explainable artificial intelligence, rare-event prediction, auto lease securitization, delinquency monitoring, probability calibration, model governance, SHAP

JEL Classification C53, C55, G23, G28

INTRODUCTION

Machine-learning systems increasingly support risk monitoring in financial institutions, but their value depends on whether predictive signals remain usable under the operational and governance conditions in which decisions are reviewed. Financial early-warning applications must identify uncommon adverse transitions, preserve temporal separation between development and evaluation, produce probabilities that can be assessed for calibration, and provide sufficient documentation for challenge by risk managers, auditors, and supervisors. These requirements become especially demanding when data are assembled from heterogeneous public reports rather than from a single internally standardized portfolio.

Auto leasing provides a distinctive setting for this problem. Near-term payment behavior reflects borrower credit quality and payment burden, but also contract age, remaining term, vehicle value, residual-value structure, underwriting indicators, and servicing or reporting conventions. Delinquency escalation is rare, so conventional accuracy

measures can conceal weak event detection. At the same time, issuer-specific reporting patterns may become predictive even when they are not portable economic risk drivers. An operational model must therefore be evaluated not only by rank discrimination but also by calibration, alert burden, transportability, and the stability of its explanations.

The scientific problem is the limited evidence on whether rare, near-term delinquency escalation can be predicted consistently across heterogeneous public auto lease securitization panels while maintaining calibrated, reviewable, and temporally stable model behavior. Addressing that problem requires a narrow outcome definition, leakage-safe longitudinal construction, fully disclosed benchmarks, matched operating points, held-out-contract and held-out-issuer validation, and empirical assessment of explanation stability rather than an assumption that explainability is automatically reliable.

1. LITERATURE REVIEW

Research on credit-risk prediction has established that model comparisons depend on the outcome, sample construction, class imbalance, and evaluation metric. Large benchmarking studies show that ensemble methods can improve discrimination over traditional scorecards, but gains are data- and design-dependent rather than universal (Brown & Mues, 2012; Lessmann et al., 2015). Leasing studies provide more specific evidence: Hernes et al. (2021) apply deep learning to repayment prediction, while Kozina et al. (2023) compare machine-learning approaches for lease-contract default. These studies confirm predictive structure in leasing portfolios, yet repayment, default, delinquency transition, servicing persistence, loss severity, and origination scoring are economically distinct tasks. A model designed for one-month incident delinquency monitoring cannot be interpreted as a lifetime default or loss model without changing the estimand.

Rare-event evaluation requires particular care. ROC-AUC describes ranking across the full range of false-positive rates, whereas precision-recall analysis is more sensitive to event prevalence and the concentration of true positives among alerts (Saito & Rehmsmeier, 2015). Low prevalence also makes threshold-based results sensitive to review capacity. An apparently high recall can be obtained simply by flagging more contracts. Consequently, performance claims should compare models at the same alert volume or under an explicitly defined cost structure. Risk thresholds are not universally optimal; they depend on the consequences of false positives and false negatives as well as on the resources available for intervention (Wynants et al., 2019). This distinction is central for financial monitoring, where a model score supports a review queue rather than directly determining an adverse action.

Probability quality is separate from ranking. Calibration concerns the agreement between predicted probabilities and observed event frequencies, and it can deteriorate even when discrimination remains stable (Van Calster et al., 2019). Brier scores, calibration intercepts and slopes, observed-to-expected ratios, and calibration curves therefore complement AUC and PR-AUC. Validation design is equally important. Random cross-validation can overstate performance in longitudinal portfolios when the same contracts or reporting regimes appear on both sides of a split. Chronological out-of-time testing answers whether a frozen model transfers to later observations, strict contract-held-out validation addresses unseen contracts, and leave-one-issuer-out validation addresses transportability to a reporting entity excluded from preprocessing, tuning, refitting, and calibration.

Explainable AI introduces another layer of methodological choice. SHAP provides additive attributions for a model output and supports global and local analysis (Lundberg & Lee, 2017), but explanations remain properties of the fitted model rather than causal effects. Dependence among features can change Shapley-based attribution, and feature-importance summaries can be misleading when they are treated as intrinsic properties of the data-generating process (Aas et al., 2021; Kumar et al., 2020). In high-stakes settings, post hoc explanation should not replace consideration of simpler or constrained models (Rudin, 2019). Financial

http://dx.doi.org/10.21511/imfi.23(3).2026.16

research has therefore moved toward joint assessment of performance, explanation availability, and governance documentation (Bücker et al., 2022; Bussmann et al., 2021; Černevičienė & Kabašinskas, 2024). Nevertheless, many studies report a single global ranking without testing its stability across issuers, time, sample size, or resampling iterations.

Model governance literature similarly distinguishes technical evidence from institutional controls. The revised interagency U.S. guidance emphasizes risk-based model development, validation, use, documentation, and governance rather than a single prescriptive control set (Board of Governors of the Federal Reserve System, 2026). National Institute of Standards and Technology (NIST, 2023), the European Banking Authority (EBA, 2023), the Financial Stability Board (FSB, 2017; 2024), and BIS studies (Crisanto et al., 2024; Pérez-Cruz et al., 2025) stress accountability, monitoring, and context-sensitive explainability. Financial-sector surveys and implementation frameworks emphasize cross-functional stakeholder involvement, reproducibility, compliance review, and structured documentation, while complementary research highlights stakeholder-specific explanation needs and standardized model reporting (HKIMR, 2021; Kuiper et al., 2021; Mitchell et al., 2019). The EU Artificial Intelligence Act further reinforces risk-management and human-oversight expectations for high-risk systems (European Parliament and Council, 2024). These sources do not imply that an explanation method proves auditability or fairness. Auditability involves documentation, lineage, validation, change control, and accountable review; fairness requires relevant protected-attribute evidence and normative criteria. Issuer or manufacturer variables may indicate portfolio or reporting heterogeneity, but their use can only be assessed as proxy dependence when demographic attributes are unavailable.

Three gaps follow from this literature. First, public contract-month evidence for U.S. auto lease securitizations remains limited, particularly for a narrowly defined incident-delinquency transition. Second, leasing studies rarely combine chronological testing with matched alert-volume evaluation, probability calibration, contract-held-out validation, and true leave-one-issuer-out experiments. Third, explainability evidence is seldom evaluated for sample-size, issuer, temporal, and bootstrap stability while being separated from causal, fairness, and auditability claims. The objective of this study is to evaluate whether explainable machine-learning models can provide reliable and operationally interpretable short-horizon delinquency-risk monitoring in U.S. auto lease securitization panels under temporal and issuer heterogeneity.

Before the revised locked-test evaluation, the following hypotheses were formulated.

H1a: *On the locked out-of-time test, unconstrained gradient boosting achieves a higher PR-AUC than the fully tuned penalized logistic-regression benchmark; support requires the paired 95% contract-cluster bootstrap confidence interval for the difference to lie above zero.*

H1b: *On the same test, X-LEASE achieves a higher PR-AUC than the fully tuned logistic benchmark under the same criterion.*

H2: *At an exact fixed review capacity of the top 1% of locked-test observations, X-LEASE achieves higher recall than unconstrained gradient boosting; support requires the paired 95% confidence interval for the recall difference to lie above zero.*

H3: *The frozen X-LEASE model remains informative in every later temporal pair, defined by a lower 95% confidence bound for AUC-ROC above 0.50 and a lower bound for PR-AUC minus period-specific prevalence above zero.*

2. METHODS

The empirical design separates data construction, model development, calibration, locked evaluation, robustness analysis, and explanation assessment. All transformations and model-development decisions were restricted to their designated pre-test windows, while the additional robustness analyses were specified before execution.

2.1. Data, outcome, sample construction, and descriptive analysis

The data were obtained from the U.S. Securities and Exchange Commission's EDGAR system through Form ABS-EE filing pages and related Exhibit 102 asset-data submissions; available Exhibit 103 asset-related documents were retained as supporting documentation (SEC, n.d.). Six auto lease securitization trust families were analyzed: BMW Vehicle Lease Trust 2024-1, Ford Credit Auto Lease Trust 2024-B, GM Financial Automobile Leasing Trust 2025-1, Mercedes-Benz Auto Lease Trust 2025-A, Nissan Auto Lease Trust 2025-A, and World Omni Automobile Lease Securitization Trust 2023-A. Five trust panels cover April 2025 through March 2026, while World Omni covers April through November 2025. A 68-row source register identifies every trust-month filing by trust, CIK, reporting period, filing date, and accession number. Detailed source coverage, reporting periods, and the corresponding grouped issuer-level panel files are reported in Table A1 (Appendix A). Monthly asset-level records were organized by reporting period, harmonized within issuer, and grouped into compressed issuer-level combined panels before contract-month alignment. The resulting harmonized source panel contains 2,740,860 contract-month rows. Canonical longitudinal keys were used to align the same contract across monthly files, including a padding-insensitive Ford key.

The primary analytical population contains contracts that were active or observable and below 30 days past due (DPD) at feature month t, had an exact record in the immediately following calendar month, and had an observed delinquency status at t + 1. The binary outcome equals one when delinquency status at t + 1 is at least 30 DPD. Charge-off, liquidation, zero-balance, termination, and ready-made distress flags were excluded from both the primary target and predictors. The outcome therefore represents one-month-ahead incident escalation to 30 + DPD, not lifetime default, expected loss, origination scoring, liquidation, charge-off, or a universal measure of leasing risk. The reconstruction audit identified 1,065 contracts with internal monthly gaps; no synthetic months were created, and no outcome was bridged across a gap.

Feature months April-July 2025 formed the development window; August was used for hyperparameter selection; April-August formed the final-refit window; September was reserved for calibration and operating-rule selection; and October-to-November formed the locked six-issuer test. Four later five-issuer feature/outcome pairs, November-December 2025 through February-March 2026, were evaluated without refitting. Descriptive statistics include central tendency, dispersion, quartiles, range, and missingness for credit score, payment-to-income ratio (PTI), loan-to-value ratio (LTV), residual value, vehicle value, contract age, original term, remaining term, and related fields. The complete analytical-sample chronology and detailed descriptive-statistics and missingness summaries are provided in Tables A2 and B1 (Appendices A and B), respectively. Issuer-level distributions and missingness are supplied in the reproducibility package.

2.2. Models, baselines, and X-LEASE constraints

Three model families were compared. M1 is a fully tuned penalized logistic-regression benchmark covering L1, L2, and elastic-net candidates. M2 is unconstrained gradient boosting, following the gradient-boosting framework of Friedman (2001) and implemented with XGBoost (Chen & Guestrin, 2016). M3, labeled X-LEASE, uses the same admissible Track A feature architecture but applies stronger complexity and regularization constraints; it is a constrained specification, not a new algorithm. The selected M1 was elastic net (C = 1; l1 ratio = 0.9). M2 used depth 7, learning rate 0.10, minimum child weight 20, column subsampling 0.70, no L1 penalty, L2 penalty 10, and 596 boosting rounds. M3 used depth 3, learning rate 0.05, minimum child weight 100, row and column subsampling of 0.80, L1 penalty 5, L2 penalty 10, and 368 rounds.

Identifiers, future fields, outcome-construction variables, and ready-made distress indicators were prohibited. Missingness and categorical handling were fit on development or final-refit data only, as appropriate. Simple comparisons included a constant calibration-prevalence forecast, a credit-score-only logistic model, and a parsimonious FICO/PTI/LTV logistic model. Track C restricted the feature

http://dx.doi.org/10.21511/imfi.23(3).2026.16

set to origination and other fundamentals while retaining the selected model settings; it was treated as a robustness analysis. The incomplete servicing-persistence Track B was removed rather than reported without a complete analytical pipeline.

2.3. Validation, calibration, operating rules, and uncertainty

Candidate hyperparameters were selected using August PR-AUC, with deterministic tie-breaking based on the Brier score, AUC-ROC, complexity, and candidate identifier. Final models were then refit on April-August observations. Identity, Platt, and isotonic calibration were compared only on September data; isotonic calibration produced the lowest calibration-sample Brier score for all three models. Frozen probability cutoffs corresponding to a 1% calibration-sample alert budget were 0.039000 for M1, 0.051948 for M2, and 0.035066 for M3. The primary operational comparison was not based on those transferred cutoffs but on an exact, matched top-1% rank rule, with 0.5%, 2%, and 5% capacities reported as sensitivity analyses. The 1% capacity is a methodological benchmark, not an observed institutional staffing limit.

Evaluation comprised AUC-ROC, PR-AUC, PR-AUC relative to prevalence, Brier score, Brier Skill Score (BSS), log loss, calibration intercept and slope, observed-to-expected ratio, and complete confusion matrices. Operational measures included alert rate, precision, recall, false positives per detected event, number needed to review, missed events, and normalized costs for hypothetical false-negative to false-positive cost ratios. Strict contract-held-out validation used a deterministic issuer-stratified hash assignment that placed approximately 20% of contracts entirely outside fitting, tuning, preprocessing, and calibration. Six zero-shot leave-one-issuer-out (LOIO) experiments excluded each issuer from all source stages and removed issuer indicators. Paired contract-cluster bootstrap intervals used the same resampled contracts across models, with issuer stratification where applicable: 2,000 replicates for the main locked test and pooled temporal evaluation, and 1,000 for individual temporal, issuer, contract-held-out, and LOIO analyses where feasible. Detailed model specifications and calibration settings are presented in Table C1 of Appendix C, while locked-test probability-performance measures and exact matched top-1% confusion matrices are reported in Tables C2 and C3, respectively.

2.4. Explainability, stability, proxy risk, and reproducibility

Exact TreeSHAP contributions were computed for the complete 198,301-row locked-test matrix for M2 and M3. Contributions decompose the raw booster margin before external isotonic calibration; calibrated probability and alert status were reported separately. Expanded features were mapped to raw variables and semantic groups. Deterministic nested samples of 3,000, 6,000, 12,000, and 24,000 observations were compared with full-sample rankings. Stability was evaluated across issuers, four later temporal pairs, and 1,000 issuer-stratified contract-bootstrap replicates. Twelve local M3 cases – three true positives, three false positives, three false negatives, and three true negatives – were selected by predeclared boundary, median-score, and extreme-score rules, with deterministic hash tie-breaking, and the same rows were explained by M2.

Issuer, manufacturer, and reporting-related variables were evaluated through global attribution and fixed-hyperparameter ablations. Because SEC ABS-EE data do not include protected demographic attributes, these analyses measure proxy dependence, not individual or group fairness. SHAP values describe model behavior and are not interpreted causally. Auditability was not assigned an ad hoc score; the narrower claims concern documented, reproducible explanation outputs and audit-oriented review support. Historical session-specific binaries became unavailable after a runtime reset, so final robustness and explainability artifacts were functionally regenerated from frozen specifications, feature orders, seeds, and matrices. Regenerated predictions were reconciled with preserved outputs and passed integrity checks, but byte-identical equivalence to unavailable binaries is not claimed. Artificial intelligence tools were used to analyze, systematize, and synthesize the results of processing large datasets and to support editorial revision. Following the use of these tools, the authors reviewed the outputs, critically assessed their accuracy and interpretation, and assume full responsibility for the content, reliability, and conclusions presented in the study.

3. RESULTS

Results are presented in four stages: sample characteristics and locked-test performance; calibration and operational burden; robustness across contracts, issuers, and time; and explainability stability with proxy-risk diagnostics.

3.1. Dataset characteristics and primary locked-test performance

Table 1 shows the analytical chronology and predictor distributions. The locked test contained 198,301 eligible contract-month observations and 768 events, a prevalence of 0.3873%. Credit quality was generally high, and predictor distributions were broadly comparable across development, calibration, and test samples, although PTI included extreme values and World Omni had substantially higher delinquency-status missingness before eligibility filtering. The event rate rose to 0.5403% in the September calibration sample before returning to 0.3873% in the locked test, reinforcing the need to evaluate probability calibration under temporal shift.

Tables 2 and 3 summarize the selected specifications and primary performance, respectively. M2 produced the highest locked-test AUC-ROC (0.8151) and PR-AUC (0.0269), equivalent to 6.94 times event prevalence. M3 ranked second by AUC-ROC and PR-AUC and exceeded the fully tuned logistic benchmark. Paired contract-cluster bootstrap contrasts supported *H1a*: the M2-M1 PR-AUC difference was 0.01209 (95% CI [0.00741, 0.01878]). *H1b* was also supported: the M3-M1 difference was 0.00189 (95% CI [0.00037, 0.00429]). The evidence did not support M3 superiority over M2.

Table 1. Sample composition and selected descriptive statistics

Source: Authors' calculations based on SEC ABS-EE Exhibit 102 data.

Panel A. Analytical sample composition				
Sample	**Observations**	**Unique contracts**	**Events**	**Event rate (%)**
Development	864,663	222,643	3,502	0.4050
Tuning	206,573	206,573	706	0.3418
Calibration	202,682	202,682	1,095	0.5403
Locked test	198,301	198,301	768	0.3873

Panel B. Selected predictor distributions and missingness			
Variable	**Development**	**Calibration**	**Locked test**
Credit score	784.0 [722.0-835.0]; 0.8% miss	785.0 [723.0-835.0]; 0.8% miss	785.0 [723.0-835.0]; 0.7% miss
Payment-to-income ratio	0.061 [0.039-0.091]; 1.4% miss	0.061 [0.039-0.090]; 1.3% miss	0.061 [0.039-0.090]; 1.3% miss
Recomputed LTV ratio	0.909 [0.847-0.972]; 0.0% miss	0.909 [0.848-0.971]; 0.0% miss	0.909 [0.848-0.972]; 0.0% miss
Residual value (USD)	27,956 [21,853-37,042]; 0.0% miss	28,060 [21,945-37,139]; 0.0% miss	28,075 [21,965-37,151]; 0.0% miss
Vehicle value (USD)	48,285 [35,950-61,752]; 0.0% miss	48,555 [35,975-62,090]; 0.0% miss	48,645 [35,975-62,120]; 0.0% miss
Contract age (months)	14.0 [10.0-21.0]; 0.0% miss	17.0 [13.0-23.0]; 0.0% miss	18.0 [14.0-24.0]; 0.0% miss
Original term (months)	36.0 [36.0-36.0]; 0.0% miss	36.0 [36.0-36.0]; 0.0% miss	36.0 [36.0-36.0]; 0.0% miss
Remaining term (months)	21.0 [14.0-26.0]; 0.0% miss	18.0 [12.0-23.0]; 0.0% miss	17.0 [11.0-22.0]; 0.0% miss
Acquisition cost (USD)	43,328 [32,542-55,415]; 0.0% miss	43,537 [32,691-55,707]; 0.0% miss	43,596 [32,734-55,779]; 0.0% miss
Residual-value ratio	0.590 [0.540-0.640]; 0.0% miss	0.590 [0.539-0.640]; 0.0% miss	0.590 [0.538-0.640]; 0.0% miss

Note. Panel A reports analytical-sample counts. Panel B reports median [Q1-Q3] and missing percentage for the development, calibration, and locked-test samples. PTI and ratio variables are expressed as proportions.

http://dx.doi.org/10.21511/imfi.23(3).2026.16

Table 2. Model specifications and implemented X-LEASE constraints

Source: Authors' model-development records.

Dimension	M1	M2	M3 / X-LEASE
Model family	Penalized logistic regression (elastic net)	Unconstrained gradient boosting	Governance-constrained gradient boosting
Selected specification	C = 1.0; l1_ratio = 0.9; 100 rounds	depth = 7; eta = 0.1; min_child_weight = 20.0; 596 rounds	depth = 3; eta = 0.05; min_child_weight = 100.0; 368 rounds
Subsample / column sample	Not applicable	1.0 / 0.7	0.8 / 0.8
L1 / L2 regularization	Elastic-net penalty	0.0 / 10.0	5.0 / 10.0
Feature policy	Track A admissible features; future fields and identifiers excluded	Track A admissible features; future fields and identifiers excluded	Same admissibility plus stronger complexity and regularization constraints
Calibration	Isotonic, selected on September calibration sample	Isotonic, selected on September calibration sample	Isotonic, selected on September calibration sample
Primary operating comparison	Exact top 1% alert capacity	Exact top 1% alert capacity	Exact top 1% alert capacity

Note: M3/X-LEASE is a constrained gradient-boosting specification. Human review, override, escalation, and suspension are post-model governance controls rather than algorithmic parameters.

Table 3. Locked-test discrimination, calibration, and probability performance

Source: Authors' calculations using the locked October-November 2025 test and paired contract-cluster bootstrap.

Model	AUC-ROC (95% CI)	PR-AUC (95% CI)	PR-AUC / prevalence	Brier	BSS	Calibration intercept	Calibration slope
M1	0.8015 [0.7887, 0.8167]	0.0145 [0.0132, 0.0175]	3.75	0.003845	0.0040	−0.915	0.873
M2	0.8151 [0.8000, 0.8284]	0.0269 [0.0218, 0.0347]	6.94	0.003828	0.0083	−0.851	0.885
M3	0.8083 [0.7942, 0.8219]	0.0168 [0.0145, 0.0208]	4.35	0.003844	0.0041	−0.876	0.882

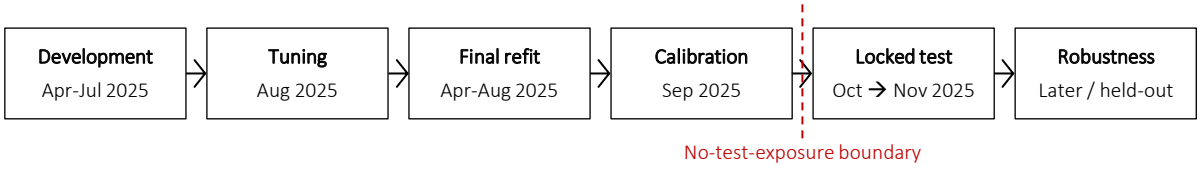
Note: PR-AUC denotes average precision; BSS is the Brier Skill Score relative to the frozen September calibration-prevalence forecast. Confidence intervals are conditional evaluation-sample intervals for frozen models.

3.2. Calibration, matched operating points, and alert burden

All three models used isotonic calibration selected on September data. Locked-test mean probabilities exceeded the observed event rate, producing observed-to-expected ratios near 0.72 and negative calibration intercepts, although BSS remained positive. Figure 2 shows that M2 dominated much of the precision-recall region and that all models exhibited some probability overprediction in the locked period. These diagnostics illustrate why rank performance and probability calibration must be reported separately.

At the exact top-1% capacity, every model generated 1,984 alerts. M2 detected 104 events, compared with 61 for M3 and 49 for M1. M2 there-

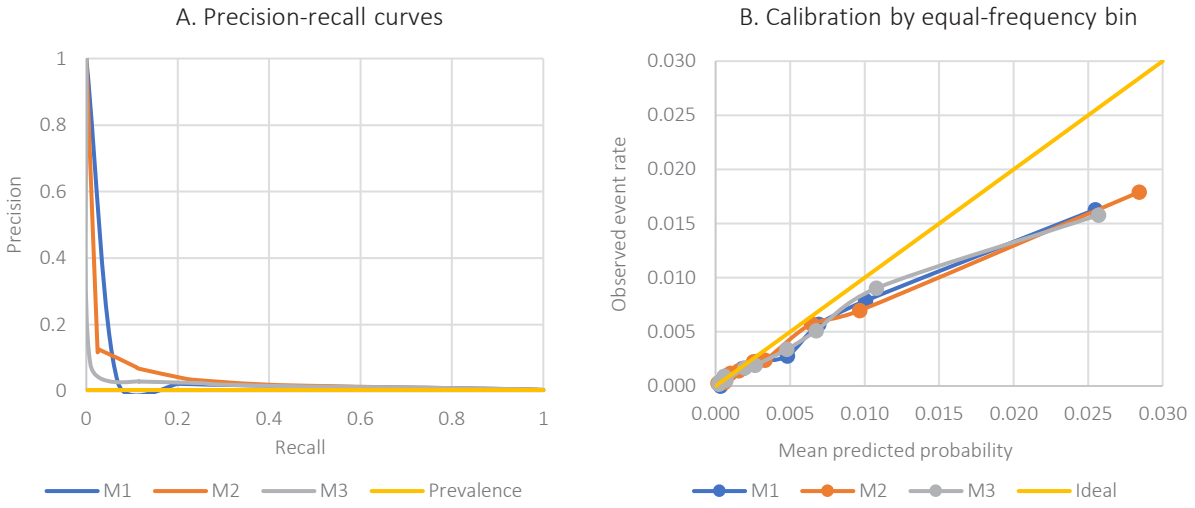
Source: Authors' design.



Note: The dashed line marks the no-test-exposure boundary. Hyperparameters were selected in August, final models were refit on April-August data, and calibration and operating rules were selected in September before the locked test was opened.

Figure 1. Temporal model-development and evaluation workflow

Source: Authors' calculations based on frozen locked-test predictions.



Note. Panel A compares precision-recall curves with the locked-test prevalence. Panel B uses equal-frequency calibration bins; the diagonal denotes perfect calibration.

Figure 2. Locked-test precision-recall and calibration curves

fore achieved recall of 0.1354 and required 19.1 reviews per detected event; M3 achieved recall of 0.0794 and required 32.5 reviews. The paired M3-M2 recall difference was −0.05556 (95% CI [−0.08184, −0.03157]), so *H2* was not supported. Transferred frozen probability cutoffs produced unequal alert rates of 1.47%, 1.27%, and 1.61% for M1, M2, and M3, respectively. The earlier appearance of an M3 recall advantage was therefore not robust to a matched review burden.

Normalized cost analysis reached the same qualitative conclusion across most false-negative to false-positive cost ratios at the matched 1% capacity: M2's larger number of detected events offset its false-alert burden. Because institution-specific review and intervention costs were unavailable, these values are scenario diagnostics rather than realized monetary benefits.

3.3. Contract-held-out, issuer-level, LOIO, and temporal robustness

The primary chronological test contained contracts observed in earlier months, so it addressed future monitoring rather than unseen-contract generalization. In the strict contract-held-out evaluation (39,703 observations; 145 events), M1 and M3 produced AUC-ROC values around 0.81, while M2 produced the highest PR-AUC (0.0150). These results did not reproduce M2's full-sample discrimination advantage uniformly, which indicates that contract novelty changes the comparative model profile.

Issuer-level performance varied with prevalence and sample size. Nissan supplied the largest event count, whereas Mercedes-Benz had only 38 locked-test events and World Omni 63. Six

Table 4. Exact matched top-1% confusion matrices and operational alert burden

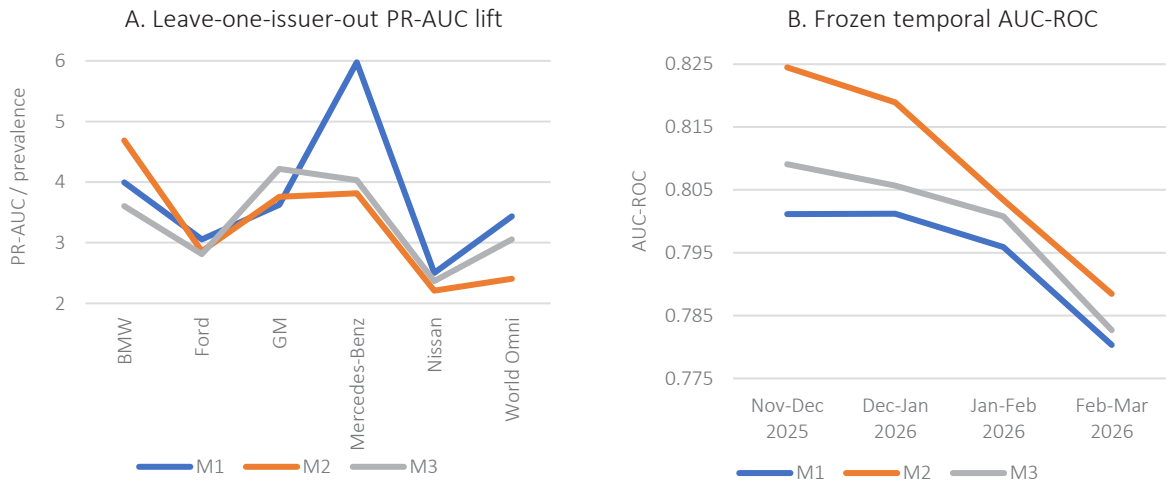
Source: Authors' calculations based on frozen locked-test rankings.

Model	Alerts	TP / FP / TN / FN	Precision	Recall	F1	NNR
M1	1,984	49 / 1,935 / 195,598 / 719	0.0247	0.0638	0.0356	40.5
M2	1,984	104 / 1,880 / 195,653 / 664	0.0524	0.1354	0.0756	19.1
M3	1,984	61 / 1,923 / 195,610 / 707	0.0307	0.0794	0.0443	32.5

Note: Each model flags exactly 1,984 observations. NNR is the number needed to review, equal to alerts divided by true positives.

http://dx.doi.org/10.21511/imfi.23(3).2026.16

Source: Authors' calculations based on strict LOIO and frozen temporal evaluations.



Note: Panel A reports PR-AUC relative to each held-out issuer's prevalence. Panel B reports AUC-ROC for four later five-issuer periods. Low-event issuer results require cautious interpretation.

Figure 3. Held-out-issuer and temporal robustness

zero-shot LOIO experiments further showed that transportability is issuer-dependent. No single model dominated all held-out issuers, and calibration was unstable in some low-event folds. These experiments support reporting issuer-specific uncertainty rather than treating pooled performance as uniform. Figure 3 summarizes LOIO PR-AUC lift and later temporal AUC-ROC.

Across all four later temporal pairs, M2 had the highest point AUC-ROC and PR-AUC. M3 nevertheless remained informative under the fixed H3 criteria: in every pair, the lower AUC confidence bound exceeded 0.50, and the lower bound for PR-AUC minus prevalence exceeded zero. H3 was therefore supported. Track C fundamentals-only results preserved the same broad ordering, with locked-test PR-AUC values of 0.0146, 0.0253, and 0.0160 for M1, M2, and M3, respectively. Track C is a robustness check; the incomplete Track B was removed. The complete contract-held-out, LOIO, temporal, and fundamentals-only robustness summary is provided in Table D1 of Appendix D.

3.4. Explainability stability and proxy dependence

Full-sample TreeSHAP analysis identified credit score as the dominant raw feature in both tree models. For M2, credit score accounted for 33.87% of normalized mean absolute SHAP, followed by issuer controls (8.05%), contract age (6.58%), remaining term (6.45%), and LTV (6.13%). For M3, credit score accounted for 47.48%, issuer controls for 14.77%, LTV for 6.45%, PTI for 5.01%, and remaining term for 3.53%. Figure 4 shows the corresponding full-test rankings.

The models shared a similar but non-identical explanation structure. Raw-feature rank correlation

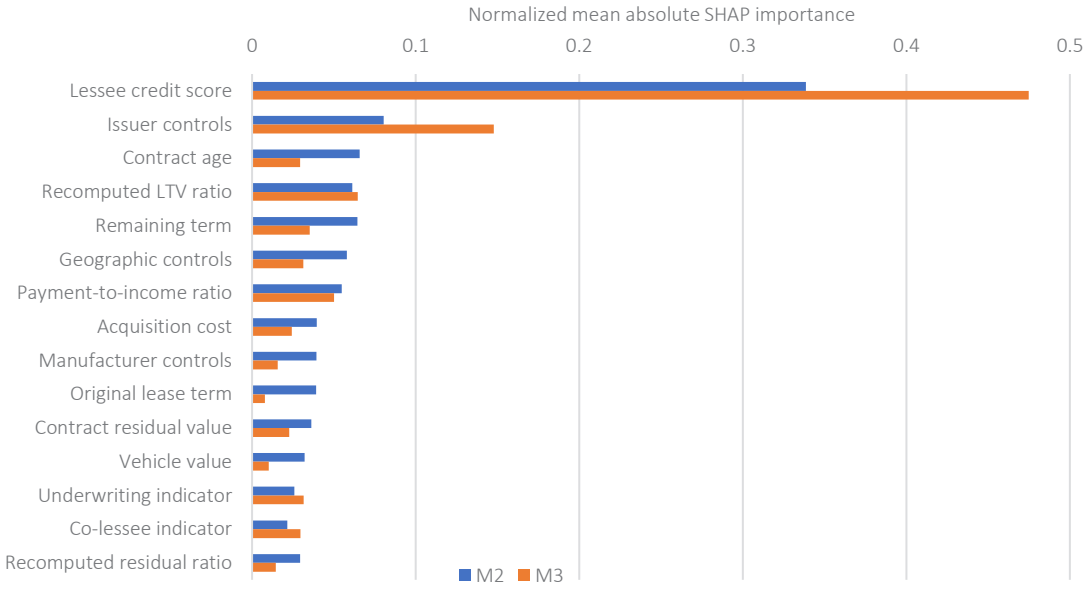
Table 5. Contract-held-out, LOIO, temporal, and fundamentals-only robustness summary

Source: Authors' calculations.

Model	Contract-held-out AUC / PR-AUC	LOIO AUC range	LOIO PR-AUC range	Temporal AUC range	Temporal PR-AUC range	Track C AUC / PR-AUC
M1	0.8075 / 0.0133	0.7508-0.8080	0.0081-0.0377	0.7803-0.8012	0.0139-0.0192	0.8027 / 0.0148
M2	0.7797 / 0.0150	0.7294-0.7752	0.0057-0.0264	0.7885-0.8245	0.0214-0.0289	0.8142 / 0.0253
M3	0.8062 / 0.0138	0.7374-0.8075	0.0060-0.0335	0.7827-0.8091	0.0145-0.0214	0.8076 / 0.0158

Note: Ranges summarize six strict LOIO folds and four frozen temporal pairs. Full issuer- and period-level calibration, uncertainty, and confusion matrices are provided in the reproducibility package.

Source: Authors' calculations using exact TreeSHAP for the reconstructed frozen M2 and M3 specifications.



Note: TreeSHAP decomposes the raw booster margin before external isotonic calibration. Values describe model behavior and are not causal effects.

Figure 4. Full locked-test normalized raw-feature SHAP importance

was 0.8911, top-10 Jaccard overlap was 0.6667, top-20 overlap was 0.9048, and normalized-importance cosine similarity was 0.9666. M3 was more concentrated: its top 10 raw features accounted for 91.86% of importance, compared with 84.17% for M2, and its effective number of important features was 3.84 versus 6.72. Concentration may simplify review, but it is not a quantitative auditability measure.

The deterministic 12,000-row visualization sample reproduced the full-test top-10 rankings for both models, with Spearman correlations above 0.9999 and cosine similarities above 0.99998. Issuer-specific rankings were generally close to the pooled ranking, although World Omni showed the largest shifts. Temporal rank correlations relative to the locked test exceeded 0.998 for both models, and 1,000 contract-bootstrap replicates yielded a median top-10 Jaccard overlap of 1.00. The evidence supports explanation stability over the tested issuers and short later periods, but not universal stability or broad superiority of M3 over M2. Detailed explanation-concentration, temporal-stability, bootstrap-stability, and proxy-dependence diagnostics are summarized in Table D2 of Appendix D.

Issuer controls represented 8.05% of M2 importance and 14.77% of M3 importance; manufacturer controls represented 3.94% and 1.57%, respectively. Fixed-hyperparameter ablations showed material but not uniformly directional changes in performance. These findings establish model dependence on portfolio and reporting controls, not prohibited discrimination. Protected demographic attributes are absent, so individual or group fairness cannot be evaluated. Twelve local cases were selected deterministically and confirmed that credit quality, exposure, term, and issuer-related contributions can move the model margin in either direction. The selected cases and their selection rules are reported in Table E1 (Appendix E); the corresponding waterfall and model-disagreement figures are available in the associated Zenodo reproducibility package (Dryha et al., 2026). Local explanations support review of model behavior but do not justify automated adverse decisions.

4. DISCUSSION

The revised evidence separates three questions that were previously conflated: which model ranks rare delinquency transitions most effectively, which model detects more events at a comparable review burden, and which model offers a more concentrated or stable explanation structure. M2 was

http://dx.doi.org/10.21511/imfi.23(3).2026.16

Table 6. Global SHAP importance and explanation-stability summary

Source: Authors' full locked-test TreeSHAP and stability calculations.

Panel A. Global raw-feature SHAP importance			
Model	**Rank**	**Raw feature**	**Normalized importance (%)**
M2	1	Credit score	33.87
M2	2	Issuer controls	8.05
M2	3	Contract age	6.58
M2	4	Remaining term	6.45
M2	5	Recomputed LTV ratio	6.13
M2	6	Geographic controls	5.79
M2	7	Payment-to-income ratio	5.47
M2	8	Acquisition cost	3.96
M3	1	Credit score	47.48
M3	2	Issuer controls	14.77
M3	3	Recomputed LTV ratio	6.45
M3	4	Payment-to-income ratio	5.01
M3	5	Remaining term	3.53
M3	6	Underwriting indicator	3.14
M3	7	Geographic controls	3.13
M3	8	Co-lessee indicator	2.95

Panel B. Explanation-stability and proxy-dependence diagnostics	
Diagnostic	**Result**
M2-M3 raw-feature Spearman	0.8911
M2-M3 top-10 / top-20 Jaccard	0.6667 / 0.9048
12,000-row top-10 Jaccard vs full test	1.00 for M2 and M3
Temporal Spearman vs locked test	0.9985-1.0000
Bootstrap median top-10 Jaccard	1.00 for M2 and M3
Issuer-control importance	8.05% (M2); 14.77% (M3)

Note: Panel A reports normalized mean absolute SHAP importance at the raw-feature level for the eight highest-ranked features of each model. Panel B reports cross-model similarity, sample-size, temporal, bootstrap, and issuer-control diagnostics. All stability metrics are conditional on the reconstructed frozen specifications. SHAP values describe model behavior and should not be interpreted as causal effects.

strongest on the first two questions. Its locked-test PR-AUC was materially higher than that of M1 and M3, and it detected 104 events at the exact 1% capacity, compared with 61 for M3. This finding is consistent with credit-scoring benchmarks in which gradient boosting performs well on nonlinear, imbalanced tabular data (Brown & Mues, 2012; Lessmann et al., 2015), but the contract-held-out and LOIO variation also confirms that no algorithm should be presumed dominant outside the tested population.

The non-support of *H2* is substantively important rather than merely negative. The earlier M3 recall advantage was calculated under different frozen probability cutoffs and therefore different alert volumes. Once review capacity was held constant, M2 ranked more events above the operational boundary. This illustrates the limitation of interpreting recall without its associated workload. Saito and Rehmsmeier (2015) emphasize prevalence-sensitive evaluation, and Wynants et al. (2019) show that threshold choice depends on the decision context. In portfolio monitoring, a matched alert budget is a more direct comparison of screening productivity than model-specific thresholds that generate unequal queues.

Calibration results provide a second qualification. Isotonic calibration improved September fit, but all models overpredicted the lower locked-test event rate. Positive BSS values indicate modest improvement over the frozen prevalence forecast, yet negative calibration intercepts and observed-to-expected ratios near 0.72 show that temporal prevalence shifts matter. This is consistent with the broader argument that discrimination does not guarantee probability reliability (Van Calster et al., 2019). A practical deployment would require continued calibration monitoring and potentially

governed recalibration under a new validation cycle, rather than ad hoc threshold changes after observing outcomes.

M3 remains scientifically useful even though it did not outperform M2. Its stronger regularization, shallower trees, and feature-admissibility controls produced a more concentrated explanation distribution, and it exceeded the fully tuned logistic benchmark by PR-AUC. The result is best interpreted as a constrained alternative rather than a superior model. Bücker et al. (2022) and Bussmann et al. (2021) distinguish multiple dimensions of transparency and reviewability; the present evidence supports documented model-behavior explanations and reproducible review, but not a single auditability score. The explanation-stability analysis also cautions against equating concentration with quality: M3 placed greater weight on credit score and issuer controls, and broader stability diagnostics did not establish that its explanations are uniformly more stable than M2's.

Issuer dependence is especially relevant for the article's scope. Pooled results can conceal differences in prevalence, missingness, reporting conventions, and portfolio composition. The LOIO experiments and issuer-specific SHAP analyses show that transportability and explanation structure are generally coherent but not identical. World Omni produced the largest explanation shifts, while Mercedes-Benz and World Omni had low event counts for some validation diagnostics. Accordingly, the evidence supports the exact held-out-issuer experiments, not a universal claim that the model transfers to every future issuer or securitization structure.

The results also clarify the proper governance boundary. The models can support human-reviewed early-warning queues, probability monitoring, model challenge, and investigation of feature contributions. The implemented controls – feature admissibility, prohibition of identifiers and future fields, frozen development chronology, documented calibration, matched-capacity evaluation, and reproducible explanations – are empirical features of the workflow. The Invariance-Based AI Risk Governance through Distributed Risk Consensus Methodology developed by Kateryna Shcheglova provides a conceptual governance extension for interpreting frozen model probabilities and TreeSHAP explanations as evidence for structured supervisory review. Its sentinel review, distributed risk-consensus, and human governance layers are informed by established model-risk and AI-governance principles (Board of Governors of the Federal Reserve System, 2026; NIST, 2023) and are summarized in Table F1 of Appendix F. These conceptual arrangements were not operationally tested and should not be presented as a validated control system, an automated decision process, or evidence of fairness, regulatory compliance, or quantitative auditability.

Several limitations bound the findings. The event is rare and observed over a one-month horizon in six public securitization panels. Some LOIO folds have few events, and World Omni differs in missingness and temporal availability. Public ABS-EE reporting may not capture all servicing, behavioral, geographic, or institutional variables available to an originator. Protected demographic attributes are absent, so fairness cannot be evaluated. Cost ratios are hypothetical, bootstrap intervals condition on frozen models, and SHAP describes model behavior rather than causal risk. Finally, historical session-specific binaries were unavailable after a runtime reset; final analyses were functionally regenerated from frozen specifications and reconciled, but byte-identical reproduction is not claimed. Future work should extend the issuer and vintage set, examine longer outcomes, study governed recalibration, and evaluate the conceptual governance architecture in an operational institution.

CONCLUSION

This study evaluated whether explainable machine-learning models can support reliable and operationally interpretable short-horizon delinquency monitoring in U.S. auto lease securitization panels. Unconstrained gradient boosting delivered the strongest locked-test discrimination and de-

http://dx.doi.org/10.21511/imfi.23(3).2026.16

tected the most events at the exact matched 1% review capacity. X-LEASE exceeded the fully tuned logistic benchmark by PR-AUC, but its hypothesized recall advantage over unconstrained boosting was not supported. Global and local TreeSHAP results were stable across the tested issuers, later periods, sample sizes, and contract-level resampling, while also showing material dependence on issuer controls. These findings support a distinction between predictive strength and governance-oriented reviewability: a constrained model can provide concentrated and reproducible explanations without being the empirically dominant predictor. The models are suitable only as human-reviewed early-warning tools for the tested short-horizon outcome; they do not establish lifetime default risk, causal effects, fairness, or automated decision authority. Further research should test additional issuers and vintages, longer delinquency and loss outcomes, governed recalibration, and the authors' conceptual governance extension in operational settings.

DATA AVAILABILITY

The underlying asset-level data were obtained from the U.S. Securities and Exchange Commission's EDGAR system through Form ABS-EE filings and related Exhibit 102 asset-data submissions. The specific filings used in the study are identified by trust, CIK, reporting period, filing date, and accession number in the supplementary source manifest. Monthly trust-level records were organized by reporting period, harmonized within issuer, and grouped into compressed issuer-level panels for subsequent contract-month alignment and analysis. The grouped analytical panels, verified source manifest, data and feature definitions, de-identified frozen predictions, statistical outputs, explainability results, retained scripts, environment records, and checksums for the deposited artifacts are available in the associated Zenodo dataset (Dryha et al., 2026): https://doi.org/10.5281/zenodo.21744987. Raw contract identifiers are not distributed.

DECLARATION OF AI-ASSISTED TECHNOLOGIES

During the preparation of this manuscript, the authors used artificial intelligence tools to analyze, systematize, and synthesize the results of processing large datasets and to support editorial revision. Following the use of these tools, the authors reviewed the outputs, critically assessed their accuracy and interpretation, and assume full responsibility for the content, reliability, and conclusions presented in the publication.

AUTHOR CONTRIBUTIONS

Conceptualization: Zhanna Dryha, Oleksandr Levchenko, Liubomyr Kochubei, Kateryna Shcheglova.
Data curation: Oleksandr Levchenko, Oleksiy Domashenko.
Formal analysis: Oleksandr Levchenko, Oleksiy Domashenko.
Investigation: Zhanna Dryha, Oleksandr Levchenko, Larysa Sergiienko, Liubomyr Kochubei, Oleksiy Domashenko, Kateryna Shcheglova.
Methodology: Zhanna Dryha, Oleksandr Levchenko, Oleksiy Domashenko, Kateryna Shcheglova.
Project administration: Oleksandr Levchenko.
Software: Oleksandr Levchenko.
Supervision: Zhanna Dryha, Larysa Sergiienko.
Validation: Zhanna Dryha, Larysa Sergiienko, Liubomyr Kochubei, Oleksiy Domashenko.
Visualization: Oleksandr Levchenko.
Writing – original draft: Zhanna Dryha, Oleksandr Levchenko.
Writing – review & editing: Zhanna Dryha, Oleksandr Levchenko, Larysa Sergiienko, Liubomyr Kochubei, Oleksiy Domashenko, Kateryna Shcheglova.

ACKNOWLEDGMENTS

The authors acknowledge the public availability of SEC EDGAR Form ABS-EE, Exhibit 102 asset-level data. No individuals or institutions outside the author team provided paid analytical, editorial, or funding support for this manuscript.

REFERENCES

1. Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence, 298,* 103502. https://doi.org/10.1016/j.artint.2021.103502
2. Board of Governors of the Federal Reserve System. (2026, April 17). *Revised guidance on model risk management* (SR 26-2). Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, & Federal Deposit Insurance Corporation. Retrieved from https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm
3. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033
4. Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society, 73*(1), 70-90. https://doi.org/10.1080/01605682.2021.1922098
5. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0
6. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*, Article 216. https://doi.org/10.1007/s10462-024-10854-8
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
8. Crisanto, J. C., Leuterio, C. B., Prenio, J., & Yong, J. (2024). *Regulating AI in the financial sector: Recent developments and main challenges* (FSI Insights No. 63). Bank for International Settlements. Retrieved from https://www.bis.org/fsi/publ/insights63.htm
9. Dryha, Z., Levchenko, O., Sergiienko, L., Kochubei, L., Domashenko, O., & Shcheglova, K. (2026). *Reproducibility package for "Explainable AI for Delinquency Risk Monitoring in U.S. Auto Lease Securitizations"* (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.21744987
10. European Banking Authority (EBA). (2023). *Follow-up report on the use of machine learning for internal ratings-based models* (EBA/REP/2023/28). Retrieved from https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf
11. European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence*. Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
12. Financial Stability Board (FSB). (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications.* Retrieved from https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service/
13. Financial Stability Board (FSB). (2024). *The financial stability implications of artificial intelligence.* Retrieved from https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/
14. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
15. Hernes, M., Kozierkiewicz, A., Maleszka, M., Rot, A., Kozina, A., Matenczuk, K., Janus, J., & Wróbel, E. (2021). Deep learning for repayment prediction in leasing companies. *European Research Studies Journal, 24*(2), 1134-1148. https://doi.org/10.35808/ersj/2178
16. Hong Kong Institute for Monetary and Financial Research (HKIMR). (2021). *Artificial intelligence and big data in the financial services industry: A regional perspective and strategies for talent development* (HKIMR Applied Research Report No. 2/2021). Retrieved from https://www.aof.org.hk/docs/default-source/hkimr/applied-research-report/aibdrep.pdf
17. Kozina, A., Kuźmiński, Ł., Nadolny, M., Miałkowska, K., Tutak, P., Janus, J., Płotnicki, F., Walaszczyk, E., Rot, A., Dziembek, D., & Król, R. (2023). The default of leasing contracts prediction using machine learning. *Procedia Computer Science, 225*, 424-433. https://doi.org/10.1016/j.procs.2023.10.027
18. Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2021).

http://dx.doi.org/10.21511/imfi.23(3).2026.16

Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. arXiv. https://doi.org/10.48550/arXiv.2111.02244

19. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 5491-5500). PMLR. https://doi.org/10.48550/arXiv.2002.11097

20. Lessmann, S., Baesens, B., Seow, H.-V., & 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. https://doi.org/10.1016/j.ejor.2015.05.030

21. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874

22. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

23. National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

24. Pérez-Cruz, F., Prenio, J., Restoy, F., & Yong, J. (2025). *Managing explanations: How regulators can address AI explainability* (FSI Occasional Papers No. 24). Bank for International Settlements. Retrieved from https://www.bis.org/fsi/fsipapers24.htm

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

26. 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), e0118432. https://doi.org/10.1371/journal.pone.0118432

27. U.S. Securities and Exchange Commission (SEC). (n.d.). *EDGAR full text search*. Retrieved May 27, 2026, from https://www.sec.gov/edgar/search/

28. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, Article 230. https://doi.org/10.1186/s12916-019-1466-7

29. Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., & Van Calster, B. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine, 17,* Article 192. https://doi.org/10.1186/s12916-019-1425-3

APPENDIX A. Data sources, accession status, and analytical populations

The analysis uses 68 monthly issuer-period source records obtained through the U.S. Securities and Exchange Commission's EDGAR system from Form ABS-EE filings and related Exhibit 102 asset-data submissions (SEC, n.d.). The accompanying machine-readable source manifest identifies every record by trust, CIK, reporting period, filing date, and accession number. Monthly records were organized by reporting period, harmonized within issuer, and grouped into compressed issuer-level combined panels before contract-month alignment. The grouped panels used for subsequent processing are listed below.

Table A1. SEC source coverage and grouped analytical panels

Source: Authors' compilation based on SEC EDGAR Form ABS-EE filings (SEC, n.d.), the verified 68-row accession manifest, and the grouped analytical panels.

Issuer / trust family	Coverage	SEC source records	Accession records documented	Grouped analytical panel
BMW BMW Vehicle Lease Trust 2024-1	2025-04 to 2026-03	12 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	12	BMW_VLT_2024-1_EX-102_12_months_combined_panel_public.csv.zip
Ford Ford Credit Auto Lease Trust 2024-B	2025-04 to 2026-03	12 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	12	Ford_Credit_ALT_2024-B_EX-102_12_months_combined_panel_public.csv.zip
GM GM Financial Automobile Leasing Trust 2025-1	2025-04 to 2026-03	12 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	12	GM_Financial_ALT_2025-1_EX-102_12_months_combined_panel_public.csv.zip
Mercedes-Benz Mercedes-Benz Auto Lease Trust 2025-A	2025-04 to 2026-03	12 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	12	Mercedes-Benz_ALT_2025-A_EX-102_12_months_combined_panel_public.csv.zip
Nissan Nissan Auto Lease Trust 2025-A	2025-04 to 2026-03	12 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	12	Nissan_ALT_2025-A_EX-102_12_months_combined_panel_public.csv.zip
World Omni World Omni Automobile Lease Securitization Trust 2023-A	2025-04 to 2025-11	8 monthly Form ABS-EE / EX-102 records; related EX-103 documentation where available	8	World_Omni_ALST_2023-A_EX-102_8_months_combined_panel_public.csv.zip

Note: The source register contains 68 monthly issuer-period records. Grouped files are derived analytical panels rather than original SEC XML submissions.

Table A2. Analytical sample chronology

Source: Authors' calculations based on the predefined analytical chronology.

Sample	Obs.	Contracts	Events	Event rate (%)
Development	864,663	222,643	3,502	0.4050
Tuning	206,573	206,573	706	0.3418
Calibration	202,682	202,682	1,095	0.5403
Locked test	198,301	198,301	768	0.3873

Note: Event rates are percentages. The primary outcome is exact adjacent-month incident escalation to 30+ DPD among contracts below 30 DPD at t.

World Omni is unavailable after November 2025. Internal monthly gaps were not filled or bridged.

http://dx.doi.org/10.21511/imfi.23(3).2026.16

APPENDIX B. Data dictionary, descriptive statistics, and leakage controls

The full 217-column data dictionary, 25-predictor Track A feature registry, and prohibited-future-field list are supplied as CSV files. The table below summarizes the variables explicitly requested by the editor.

Table B1. Descriptive statistics and missingness summary

Source: Authors' calculations from the harmonized analytical samples.

Sample	Variable	Summary
Development	Credit score	mean 775; SD 73.3; med. 784; IQR 722-835; range 303-900; miss. 0.8%
Development	Payment-to-income ratio	mean 0.0758; SD 1.67; med. 0.0615; IQR 0.039-0.0907; range 0-669; miss. 1.4%
Development	Recomputed LTV ratio	mean 0.911; SD 0.102; med. 0.909; IQR 0.847-0.972; range 0.333-1.63; miss. 0.0%
Development	Residual-value ratio	mean 0.591; SD 0.0757; med. 0.59; IQR 0.54-0.64; range 0.146-0.895; miss. 0.0%
Development	Vehicle value (USD)	mean 52,500; SD 22,700; med. 48,300; IQR 36,000–61,800; range 21,500–250,000; miss. 0.0%
Development	Residual value (USD)	mean 30,400; SD 11,900; med. 28,000; IQR 21,900–37,000; range 8,570–129,000; miss. 0.0%
Development	Contract age (months)	mean 15.9; SD 7.64; med. 14; IQR 10-21; range 3-53; miss. 0.0%
Development	Original term (months)	mean 36.3; SD 4.96; med. 36; IQR 36-36; range 13-60; miss. 0.0%
Development	Remaining term (months)	mean 20.3; SD 8.62; med. 21; IQR 14-26; range -4-57; miss. 0.0%
Development	Acquisition cost (USD)	mean 47,900; SD 22,200; med. 43,300; IQR 32,500–55,400; range 13,500–300,000; miss. 0.0%
Calibration	Credit score	mean 775; SD 73.2; med. 785; IQR 723-835; range 303-900; miss. 0.8%
Calibration	Payment-to-income ratio	mean 0.0759; SD 1.72; med. 0.0611; IQR 0.0388-0.0903; range 0-669; miss. 1.3%
Calibration	Recomputed LTV ratio	mean 0.911; SD 0.101; med. 0.909; IQR 0.848-0.971; range 0.333-1.63; miss. 0.0%
Calibration	Residual-value ratio	mean 0.59; SD 0.0759; med. 0.59; IQR 0.539-0.64; range 0.2-0.895; miss. 0.0%
Calibration	Vehicle value (USD)	mean 52,800; SD 22,800; med. 48,600; IQR 36,000–62,100; range 21,500–250,000; miss. 0.0%
Calibration	Residual value (USD)	mean 30,500; SD 11,900; med. 28,100; IQR 21,900–37,100; range 8,570–129,000; miss. 0.0%
Calibration	Contract age (months)	mean 18.8; SD 7; med. 17; IQR 13-23; range 8-51; miss. 0.0%
Calibration	Original term (months)	mean 36.4; SD 4.88; med. 36; IQR 36-36; range 13-60; miss. 0.0%
Calibration	Remaining term (months)	mean 17.4; SD 8.04; med. 18; IQR 12-23; range -3-52; miss. 0.0%
Calibration	Acquisition cost (USD)	mean 48,100; SD 22,300; med. 43,500; IQR 32,700–55,700; range 15,100–300,000; miss. 0.0%
Locked test	Credit score	mean 775; SD 73.1; med. 785; IQR 723-835; range 303-900; miss. 0.7%
Locked test	Payment-to-income ratio	mean 0.0759; SD 1.74; med. 0.061; IQR 0.0388-0.0901; range 0-669; miss. 1.3%
Locked test	Recomputed LTV ratio	mean 0.911; SD 0.101; med. 0.909; IQR 0.848-0.972; range 0.333-1.63; miss. 0.0%
Locked test	Residual-value ratio	mean 0.59; SD 0.0759; med. 0.59; IQR 0.538-0.64; range 0.2-0.895; miss. 0.0%
Locked test	Vehicle value (USD)	mean 52,800; SD 22,800; med. 48,600; IQR 36,000–62,100; range 21,500–250,000; miss. 0.0%
Locked test	Residual value (USD)	mean 30,600; SD 11,900; med. 28,100; IQR 22,000–37,200; range 8,570–129,000; miss. 0.0%
Locked test	Contract age (months)	mean 19.6; SD 6.87; med. 18; IQR 14-24; range 9-52; miss. 0.0%
Locked test	Original term (months)	mean 36.4; SD 4.85; med. 36; IQR 36-36; range 13-60; miss. 0.0%
Locked test	Remaining term (months)	mean 16.6; SD 7.9; med. 17; IQR 11-22; range -4-51; miss. 0.0%
Locked test	Acquisition cost (USD)	mean 48,200; SD 22,300; med. 43,600; IQR 32,700–55,800; range 15,100–300,000; miss. 0.0%

Note: Continuous-variable summaries report the statistics available in the verified descriptive outputs; missingness is shown where applicable.

Hard exclusions include all t+1 and outcome-construction fields, identifiers, row hashes, source names, ready-made distress/delinquency/charge-off/liquidation flags, and future lifecycle variables. Missingness treatment and categorical vocabularies were fitted only in authorized pre-test windows.

APPENDIX C. Model development, calibration, operating rules, and uncertainty

Table C1. Model specifications and selected settings

Source: Authors' model-development records.

Dimension	M1	M2	M3 / X-LEASE
Model family	Penalized logistic regression (elastic net)	Unconstrained gradient boosting	Governance-constrained gradient boosting
Selected specification	C = 1.0; l1_ratio = 0.9; 100 rounds	depth = 7; eta = 0.1; min_child_weight = 20.0; 596 rounds	depth = 3; eta = 0.05; min_child_weight = 100.0; 368 rounds
Subsample / column sample	Not applicable	1.0 / 0.7	0.8 / 0.8
L1 / L2 regularization	Elastic-net penalty	0.0 / 10.0	5.0 / 10.0
Feature policy	Track A admissible features; future fields and identifiers excluded	Track A admissible features; future fields and identifiers excluded	Same admissibility plus stronger complexity and regularization constraints
Calibration	Isotonic, selected on September calibration sample	Isotonic, selected on September calibration sample	Isotonic, selected on September calibration sample
Primary operating comparison	Exact top 1% alert capacity	Exact top 1% alert capacity	Exact top 1% alert capacity

Note: M3/X-LEASE is a constrained gradient-boosting specification. Conceptual human-governance controls are not model hyperparameters.

Calibration candidates were identity, Platt, and isotonic; isotonic was selected for M1-M3 by September Brier score. Frozen cutoffs corresponding to a 1% calibration-sample alert budget were 0.039000, 0.051948, and 0.035066. The primary operational comparison uses an exact top-1% rank rule, with deterministic row-ID tie-breaking.

Table C2. Locked-test discrimination, calibration, and probability performance

Source: Authors' calculations using the locked October-November 2025 test.

Model	AUC-ROC	PR-AUC	PR-AUC/ prevalence	Brier	BSS	Calibration intercept	Calibration slope
M1	0.8015	0.0145	3.75	0.003845	0.0040	-0.915	0.873
M2	0.8151	0.0269	6.94	0.003828	0.0083	-0.851	0.885
M3	0.8083	0.0168	4.35	0.003844	0.0041	-0.876	0.882

Note: PR-AUC denotes average precision; BSS denotes Brier Skill Score relative to the frozen September reference forecast.

Table C3. Exact matched top-1% confusion matrices

Source: Authors' calculations based on exact matched top-1% locked-test rankings.

Model	Alerts	TP	FP	TN	FN	Precision	Recall	F1	NNR
M1	1,984	49	1,935	195,598	719	0.0247	0.0638	0.0356	40.5
M2	1,984	104	1,880	195,653	664	0.0524	0.1354	0.0756	19.1
M3	1,984	61	1,923	195,610	707	0.0307	0.0794	0.0443	32.5

Note: All three models flag exactly 1,984 locked-test observations. NNR is the number needed to review.

Primary locked-test and pooled-temporal uncertainty used 2,000 paired issuer-stratified contract-cluster replicates; individual temporal, issuer, contract-held-out, and LOIO analyses used 1,000 where feasible. These are conditional evaluation-sample intervals for frozen models.

http://dx.doi.org/10.21511/imfi.23(3).2026.16

APPENDIX D. Robustness, explainability, and proxy dependence

Table D1. Robustness summary

Source: Authors' calculations from contract-held-out, LOIO, temporal, and Track C analyses.

Model	Contract-held-out AUC / PR-AUC	LOIO AUC range	LOIO PR-AUC range	Temporal AUC range	Temporal PR-AUC range	Track C AUC / PR-AUC
M1	0.8075 / 0.0133	0.7508-0.8080	0.0081-0.0377	0.7803-0.8012	0.0139-0.0192	0.8027 / 0.0148
M2	0.7797 / 0.0150	0.7294-0.7752	0.0057-0.0264	0.7885-0.8245	0.0214-0.0289	0.8142 / 0.0253
M3	0.8062 / 0.0138	0.7374-0.8075	0.0060-0.0335	0.7827-0.8091	0.0145-0.0214	0.8076 / 0.0158

Note: Ranges summarize six strict LOIO folds and four frozen temporal pairs. Low-event folds require cautious interpretation.

H1a and *H1b* were supported; H2 was not supported; H3 was supported. Full locked-test TreeSHAP was computed for M2 and M3. The 12,000-row visualization sample reproduced full-test top-10 rankings (Jaccard = 1.00; Spearman > 0.9999). Raw-feature similarity between M2 and M3 was Spearman 0.8911 and cosine 0.9666; M3 had a more concentrated explanation distribution.

Table D2. Explainability and stability diagnostics

Source: Authors' full locked-test TreeSHAP and stability calculations.

Evidence	M2	M3
Issuer-control SHAP share	8.05%	14.77%
Manufacturer-control SHAP share	3.94%	1.57%
Top-10 raw-feature importance share	84.17%	91.86%
Effective number of important features	6.72	3.84
Temporal rank stability	Spearman ≥ 0.9985	Spearman ≥ 0.9985
Bootstrap median top-10 Jaccard	1.00	1.00

Note: SHAP values describe model behavior before external isotonic calibration and are not causal effects.

Issuer/manufacturer ablations assess proxy dependence, not fairness. Protected demographic attributes are unavailable. SHAP explains raw booster margins before external isotonic calibration and is not causal.

APPENDIX E. Local explanations, conceptual governance extension, and reproducibility lineage

Twelve local cases were selected deterministically under the M3 exact top-1% rule: three observations in each of the TP, FP, FN, and TN categories using boundary, median-score, and extreme-score rules. The same rows were explained by M2 and M3. Four median-score waterfall figures and six model-disagreement figures, together with their underlying source CSV files, are available in the associated Zenodo reproducibility package (Dryha et al., 2026) and are not reproduced in this combined manuscript file.

Table E1. Deterministically selected local cases

Source: Authors' deterministic local-case selection records.

case_id	confusion_cell	selection_rule	observed_outcome	M3_calibrated_probability	M3_top1_alert
TP_BOUNDARY	TP	boundary	1	0.036161	1
TP_MEDIAN_SCORE	TP	median_score	1	0.036161	1
TP_EXTREME_SCORE	TP	extreme_score	1	0.285714	1
FP_BOUNDARY	FP	boundary	0	0.036161	1
FP_MEDIAN_SCORE	FP	median_score	0	0.036161	1
FP_EXTREME_SCORE	FP	extreme_score	0	0.285714	1
FN_BOUNDARY	FN	boundary	1	0.036161	0
FN_MEDIAN_SCORE	FN	median_score	1	0.010816	0
FN_EXTREME_SCORE	FN	extreme_score	1	0.000202	0

Table E1 (cont.). Deterministically selected local cases

case_id	confusion_cell	selection_rule	observed_ outcome	M3_calibrated_ probability	M3_top1_alert
TN_BOUNDARY	TN	boundary	0	0.036161	0
TN_MEDIAN_SCORE	TN	median_score	0	0.002500	0
TN_EXTREME_SCORE	TN	extreme_score	0	0.000000	0

Note: Cases were selected under deterministic boundary, median-score, and extreme-score rules; raw contract identifiers are not shown. Calibrated probabilities are rounded to six decimal places; full-precision values are retained in the associated Zenodo reproducibility package.

APPENDIX F. Conceptual Invariance-Based AI Risk Governance through Distributed Risk Consensus Methodology for frozen X-LEASE outputs

This section integrates the Invariance-Based AI Risk Governance through Distributed Risk Consensus Methodology developed by Kateryna Shcheglova as a conceptual governance extension for the frozen X-LEASE outputs. It does not modify the empirical design, model specifications, analytical population, outcomes, feature sets, calibration, thresholds, evaluation results, tables, or figures. Instead, it interprets frozen model probabilities, operating flags, and TreeSHAP explanations as supervisory evidence within a layered review architecture. The methodology was not separately tested as an empirical or operational-control workflow in the locked test, robustness analyses, or temporal evaluation. The empirically implemented controls – feature admissibility, frozen chronology, calibration records, validation procedures, and reproducible explanation outputs – remain components of the analytical workflow and are not reclassified as conceptual governance layers.

Table F1. Layered governance interpretation of frozen X-LEASE outputs

Source: Conceptual governance methodology developed by co-author Shcheglova and integrated in this study with the frozen X-LEASE and TreeSHAP evidence.

Layer	Status	Role in the governance framework	Governance function
XAI output layer	Empirically generated model-output evidence	Frozen X-LEASE probability, operating flag, and TreeSHAP explanation	Produces documented early-warning evidence for accountable human review.
Sentinel review layer	Conceptual	Reviews the signal from model-risk, leasing-risk, data-quality, conduct, and operational-resilience perspectives	Tests whether the signal remains reliable and within predefined governance conditions.
Distributed risk-consensus layer	Conceptual	Reviews disagreement, instability, issuer-specificity, and proximity to governance boundaries	Determines whether escalation, additional validation, enhanced monitoring, or no immediate action is appropriate.
Human governance layer	Conceptual institutional control	Defines invariants, escalation rules, override rights, and restriction or suspension conditions	Preserves accountability and human authority over final decisions.

Note: The methodology is conceptual and was not separately tested or validated as an operational-control workflow. The XAI output layer reports model-behavior evidence generated by the frozen model; the sentinel review, distributed risk-consensus, and human governance layers are proposed governance arrangements. They do not constitute an empirical or quantitative assessment of auditability, establish causal effects or fairness, guarantee regulatory compliance, or authorize automated decisions. LIME was not implemented and is not claimed.

Historical session-specific binaries were unavailable after a runtime reset; models were functionally reconstructed from frozen specifications and reconciled, but byte-identical equivalence is not claimed. The public Zenodo deposit includes the verified 68-row SEC source manifest, six grouped issuer-level analytical panels, data and feature definitions, manifests, hyperparameters, calibration and threshold records, de-identified predictions, performance and bootstrap outputs, SHAP summaries, the local waterfall and model-disagreement figures and their source data, retained scripts, environment records, and checksums. The deposited materials are available in the associated Zenodo dataset (Dryha et al., 2026), Version 1.0.0: https://doi.org/10.5281/zenodo.21744987

http://dx.doi.org/10.21511/imfi.23(3).2026.16