WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Analysis of Cost Sensitive Ensemble Classifiers for Reducing False Negative Predictions in Consumer Loan Defaults

Abdulrahman Adebola Iyaniwura

Independent Researcher/Senior Machine Learning Engineer, USA

ABSTRACT :

Consumer lending requires reliable default prediction to balance credit access, portfolio stability and financial losses. Conventional classifiers can overlook borrowers who subsequently default, particularly when default observations are uncommon and evaluation prioritises overall accuracy. This proposed study will compare cost-sensitive ensemble classifiers for reducing false-negative predictions in consumer loan default assessment. Default will constitute the positive class, making missed defaults the primary error of interest. Random forests, gradient boosting and adaptive boosting will be evaluated using consistent predictors, temporal partitions and development procedures. Misclassification costs, class weighting and decision thresholds will be tuned within training data, while independent evaluation will preserve observed default prevalence. Performance assessment will examine default recall, precision, precision–recall discrimination, calibration and expected lending costs alongside false-positive rates and approval implications. Sensitivity analyses will vary cost assumptions, default prevalence and threshold choices, with subgroup assessment examining uneven error burdens. Comparisons will distinguish improvements in detecting defaults from indiscriminate rejection of applicants. The study will identify trade-offs between missed-default reduction, predictive reliability and lending feasibility, without assuming that higher recall alone establishes economic value or equitable credit decisions.

Keywords: Consumer loan default; Cost-sensitive learning; Ensemble classifiers; False-negative reduction; Credit risk assessment; Decision thresholds.

1. INTRODUCTION AND RESEARCH OBJECTIVES

1.1 Consumer Lending and the Consequences of Missed Defaults

Default prediction supports consumer lending by informing credit assessment, portfolio monitoring and allocation of risk-bearing capacity. Credit-scoring research therefore evaluates models as decision-support tools rather than isolated classifiers [1]. With default defined as the positive class, a false negative occurs when a loan predicted not to default subsequently meets the specified default criterion.

Missed defaults can expose lenders to losses, collection expenditure and reduced portfolio resilience. However, false positives also matter: classifying a performing applicant as a future defaulter may produce unnecessary rejection, foregone revenue and restricted credit access. Costs can vary across exposures rather than remain constant within classes [2].

Default prevalence differs across products, borrower populations and economic periods. Where outcomes are imbalanced, high overall accuracy can coexist with poor default detection. Precision–recall evaluation helps reveal performance obscured by aggregate measures [3], motivating assessment of competing errors together.

1.2 Limitations of Conventional Classification and Ensemble Evaluation

Conventional training objectives often reward predictive fit without representing asymmetric lending consequences. Similarly, a probability threshold of 0.5 need not minimise application-specific loss. Cost-sensitive decision theory links classification choices to misclassification penalties and estimated probabilities [4].

Cost-sensitive learning changes the training emphasis; ensemble construction combines learners; threshold adjustment changes how scores become decisions. These mechanisms should be evaluated separately. Random forests, for example, aggregate randomised trees, but their ensemble structure alone does not specify lending costs [5].

Comparisons must therefore include standard models with tuned thresholds alongside cost-sensitive alternatives. Otherwise, apparent algorithmic gains may merely reflect more restrictive classification. Probability calibration requires separate examination because accurate ranking does not ensure reliable probabilities [6]. Temporal testing and documentation are also important in credit-model validation [7]. Finally, performance must be interpreted against the chosen costs rather than one universally preferred statistic [8].

1.3 Aim, Research Questions and Contribution

This study will compare cost-sensitive ensemble classifiers for reducing missed consumer loan defaults while accounting for false positives, calibration and economic consequences. It will ask which approaches provide the strongest trade-offs between default recall, expected cost and operational feasibility, and whether improvements persist across vintages, cost assumptions and borrower groups.

The contribution will be a controlled comparison using consistent predictors, development procedures and independent evaluation. Existing credit-scoring benchmarks demonstrate the importance of algorithm comparison [1]. Standard ensembles, threshold-tuned alternatives and cost-sensitive variants will therefore receive matched evaluation rather than presumed superiority.

Conclusions will distinguish better default detection from indiscriminate rejection. The next section will establish the prediction target and economic assumptions before data acquisition, feature engineering and modelling are specified.

2. PREDICTION TARGET, COST FRAMEWORK AND COMPARATIVE DESIGN

2.1 Default Definition, Prediction Moment and Performance Horizon

The unit of analysis will be an originated consumer loan, with prediction anchored immediately before the lending decision using information then available. A twelve-month horizon will define the primary outcome; suitability will be confirmed against dataset coverage. Default events will map documented charge-off, delinquency or restructuring codes to a prespecified definition, rather than equating every adverse status with default. Basel distinguishes unlikeliness to pay from material past-due triggers and permits specified qualifications [7].

\[y_i=\mathbb{1}\left(D_i\leq H\right) \tag{1} \]

Equation 1 defines the binary label. Here, \(D_i\) denotes elapsed time to qualifying default, \(H\) the twelve-month horizon, and the indicator returns one when default occurs within it. Default-free loans require observation through the horizon or documented full repayment terminating exposure. Immature loans and unresolved follow-up will not automatically receive negative labels. Distressed restructuring and charge-offs will follow mapping rules. Operational labels will be distinguished from regulatory compliance, and cohort exclusions will be reported transparently and reproducibly.

2.2 Misclassification Costs and Exposure Heterogeneity

False-negative costs represent consequences assigned to missed defaults; false-positive costs represent consequences of incorrectly flagging performing loans. These penalties will be specified before model comparison. A simplified matrix will initially assign zero cost to correct classifications, while acknowledging that actual lending includes funding costs, income and servicing expenditure. Performance comparisons depend on cost assumptions rather than a universally preferred metric [8].

$$C_{\mathrm{FN},i} = \mathrm{EAD}_i \, \mathrm{LGD}_i$$

Equation 2 expresses an exposure-based false-negative proxy. Multiplying exposure at default, measured in currency, by loss given default, expressed as a proportion, produces estimated loss severity in currency units. Loss severity varies with outstanding balances, security, recoveries and collection costs; Basel distinguishes exposure and economic loss concepts [7].

The product will not establish the incremental benefit of rejecting a loan or imply that all detected losses are avoidable. False-positive opportunity costs will be estimated separately using defensible assumptions about foregone margin and alternative decisions. Scenario analyses will vary uncertain inputs. Future realised exposures and recoveries may inform retrospective evaluation but will remain excluded from origination-time predictors to prevent leakage [13]. Estimated costs will retain their documented timing and provenance.

2.3 Experimental Comparisons and Acceptance Criteria

Three conditions will distinguish ensemble structure, training emphasis and decision thresholds: standard ensembles at a conventional threshold; standard ensembles with development-tuned thresholds; and cost-sensitive ensembles with development-tuned thresholds. Random forests illustrate that combining learners does not itself specify economic penalties [5]. Penalised logistic regression will provide an interpretable comparator, with elastic-net regularisation supporting shrinkage and variable selection [11].

The primary comparison will assess whether cost-sensitive training improves expected cost beyond threshold adjustment alone. Prespecified scenarios will vary cost ratios and exposure assumptions. Acceptance criteria will consider default recall, false-positive rates, calibration and operational feasibility; improved ranking will not substitute for reliable probabilities [6].

Comparisons will use identical evaluation loans and, where feasible, matched approval or false-positive rates. Tuning will remain within development data because shared selection and assessment can inflate performance estimates [12]. No universal performance cutoff will be presumed. These definitions will guide data acquisition and cohort construction, preserving the distinction between observed repayment outcomes and counterfactual lending decisions while documenting unresolved assumptions before empirical analysis or implementation proceeds.

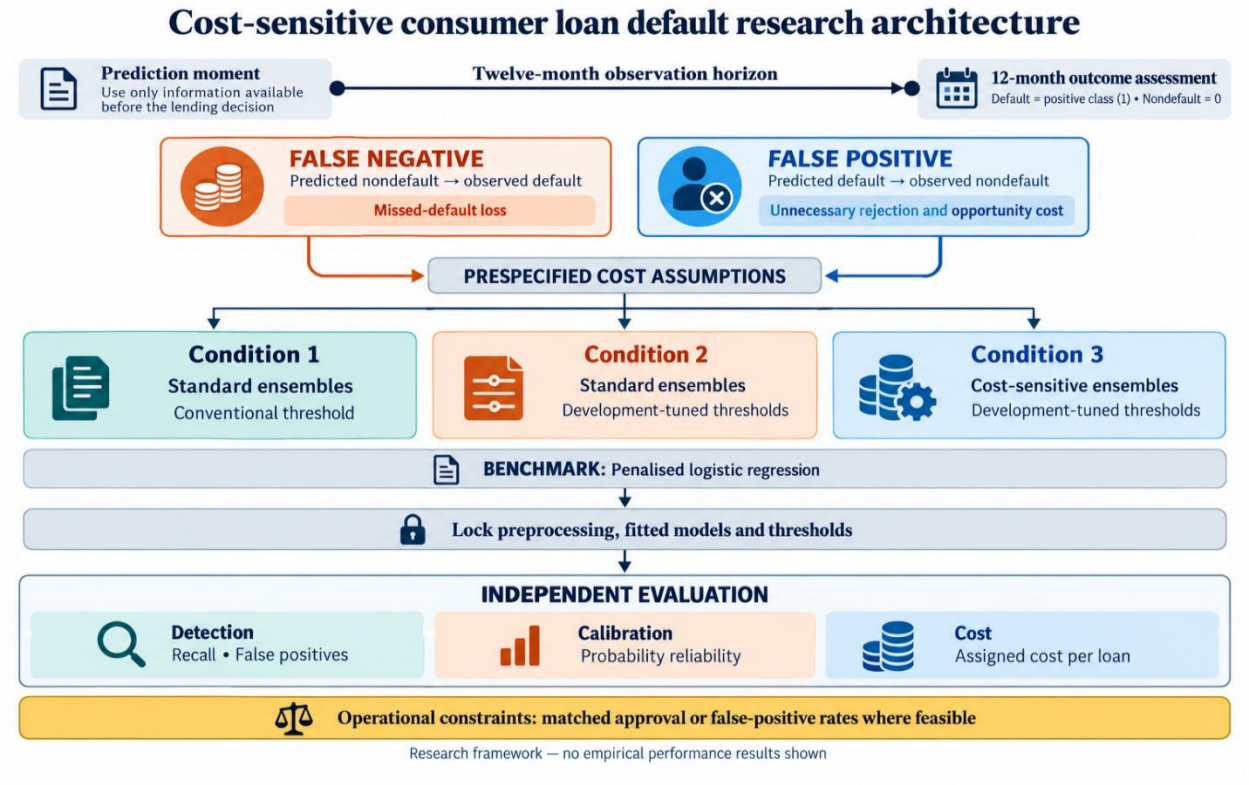


Figure 1. Research architecture linking default definition, cost assumptions and controlled model comparisons.

3. DATA ACQUISITION AND FEATURE ENGINEERING

3.1 Data Source, Acquisition and Cohort Construction

The proposed source is Freddie Mac's Single-Family Loan-Level Standard Dataset, which provides origination and monthly performance records for acquired mortgages [14]. This choice narrows consumer lending to U.S. fixed-rate mortgages; findings will not automatically generalise to unsecured loans. Proposed origination vintages are 2012–2019, with records retained only when the twelve-month outcome is observable or exposure terminates through documented repayment.

Access requires registration and acceptance of terms [14]. Extraction date, release identifier, performance cutoff and downloaded file checksums will be recorded when acquisition occurs. Release documentation will establish observation coverage rather than an assumed download date [15].

Loan identifiers will link origination and performance files. Duplicate loan-month records, invalid sequences and exclusions will be logged. Borrower linkage will be attempted only if valid identifiers exist; otherwise, repeat-borrower separation cannot be guaranteed. These are funded mortgages, not all applications, so rejected applicants' outcomes remain unobserved. Public identifiers will not support reidentification. Institutional review requirements and data-use arrangements will be documented before analysis, without assuming that public access removes governance obligations or restrictions.

3.2 Data Quality, Missingness and Leakage Audit

Missingness will be assessed by origination vintage, available product characteristics and outcome. Data checks will examine currency units, percentage coding, impossible ratios, duplicate records and dates inconsistent with loan progression. Valid extremes will be distinguished from recording errors; values will not be removed merely because they are unusual.

Monthly delinquency, collections, recoveries, disposition proceeds and settlement information will support outcome or cost assessment where appropriate, but will not enter origination predictor matrices. The source separates origination and subsequent performance information [14]. Availability at origination will not automatically prove availability before the lending decision.

Every candidate predictor will receive a timing assessment. Fields reflecting later lender decisions will be excluded or analysed separately. Imputation, encoding and transformations will be fitted within training partitions, addressing leakage that can inflate apparent predictive performance [13]. Processing rules and unresolved quality concerns will remain documented.

3.3 Financial and Behavioural Feature Construction

Candidate features will include available credit scores, debt-to-income ratios, original loan-to-value measures and loan characteristics. Income, requested amount, revolving utilisation and credit-history measures will enter only if a verified source supplies them. Original funded balances will not be relabelled as requested amounts. Affordability ratios will require compatible numerator and denominator definitions.

Currency units, percentage scales and missing-value codes will be harmonised. Zero denominators will produce flagged undefined ratios rather than arbitrary values. Categorical encoding will preserve documented meanings, and skew-reducing transformations will be selected within development data.

$$z_{ij} = \frac{x_{ij} - \mu_{j,\text{train}}}{s_{j,\text{train}}}$$

Equation 3 subtracts the training mean from raw feature value and divides by training standard deviation. Here, x_{ij} represents feature j for loan i; $\mu_{j,\text{train}}$ and $s_{j,\text{train}}$ are training-derived location and dispersion, while z_{ij} is dimensionless. Constant features require removal or separate handling. Parameters will be re-estimated inside training folds to prevent leakage [13]. Scaling supports penalised regression because penalties depend on coefficient scale [11], but is generally unnecessary for tree split ordering. Feature construction will retain provenance and decision-time availability constraints.

3.4 Mean Deviation, Distribution Checks and Feature Selection

$$\text{MAD}_j = \frac{1}{n}\sum_{i=1}^{n} \left| x_{ij} - \bar{x}_j \right|$$

Equation 4 defines mean absolute deviation around the arithmetic mean, not median absolute deviation. For feature j, n is the number of valid observations, x_{ij} an observed value and $\bar{x}_j$ its arithmetic mean. Absolute deviations prevent positive and negative differences from cancelling; dividing their sum by n gives average distance in the feature's original units.

Training and later-vintage distributions will be compared using means, standard deviations, mean absolute deviations, missingness and quantiles. These summaries describe variation rather than establish greater default risk. Outcome-conditioned comparisons will be interpreted cautiously because changes can reflect borrower composition or economic circumstances.

Feature screening will assess redundant information, unstable categories and correlated measurements. Selection criteria will be estimated within development folds and evaluated without reusing assessment outcomes, since model-selection procedures can otherwise produce optimistic estimates [12]. Later-vintage evaluation will not trigger retrospective feature revision. Engineered features, transformations and exclusion rules will be frozen with the model. This controlled dataset will support subsequent training, tuning and independent testing while preserving reproducibility and documented limits to generalisation.

Table 1. Data dictionary, engineered features and leakage safeguards.

Domain	**Source variable**	**Engineered definition**	**Unit/coding**	**Availability at prediction**	**Missing-data treatment**	**Leakage concern**
Creditworthiness	Available origination credit score	Documented score; flexible transformation where justified	Score units	Verify collection timing	Training-fitted imputation and indicator where justified	Updated scores must not replace application-time information
Debt burden	Reported debt-to-income ratio	Harmonised ratio using source definitions	Percentage	Verify timing and underwriting provenance	Preserve unknown codes separately before imputation	Do not reconstruct from later repayment records
Collateral leverage	Original loan-to-value ratio	Harmonised original leverage measure	Percentage	Verify appraisal and decision timing	Training-fitted treatment	Later property valuations are excluded
Loan amount	Original unpaid principal balance	Funded balance, explicitly distinguished from requested amount	Currency	May reflect the completed lending decision	Document missing or invalid balances	Exclude from a strict pre-decision model unless timing is supported
Loan characteristics	Purpose, occupancy, property type and term	Source-defined categories or numeric term	Categorical; months	Field-specific timing audit	Training-fitted encoding	Final contractual terms may follow the prediction moment
Income and utilisation	Conditional supplementary source	Ratios only where verified underlying variables exist	Currency; percentage	Must precede prediction	No fabricated reconstruction	Not assumed available in the selected dataset
Distribution summaries	Eligible numeric features	Mean, standard deviation, mean absolute deviation and	Original feature units	Training summaries fitted	Report valid observation counts	Test summaries must not drive model

Domain	Source variable	Engineered definition	Unit/coding	Availability at prediction	Missing-data treatment	Leakage concern
		quantiles		before evaluation		revision
Default outcome	Monthly performance records	Qualifying event within twelve months	Binary	Observed after prediction	Unresolved horizons remain unclassified	Outcome information excluded from predictors
Loss evaluation	Exposure, recoveries and expenses	Documented retrospective loss proxy	Currency	Generally post-outcome	Scenario analysis where incomplete	Evaluation use only; excluded from origination predictors
Record linkage	Loan identifier; borrower identifier if available	Origination–performance linkage and grouping	Identifier	Administrative use	Investigate unresolved linkage	Identifiers are not predictive features; loan IDs do not establish borrower identity

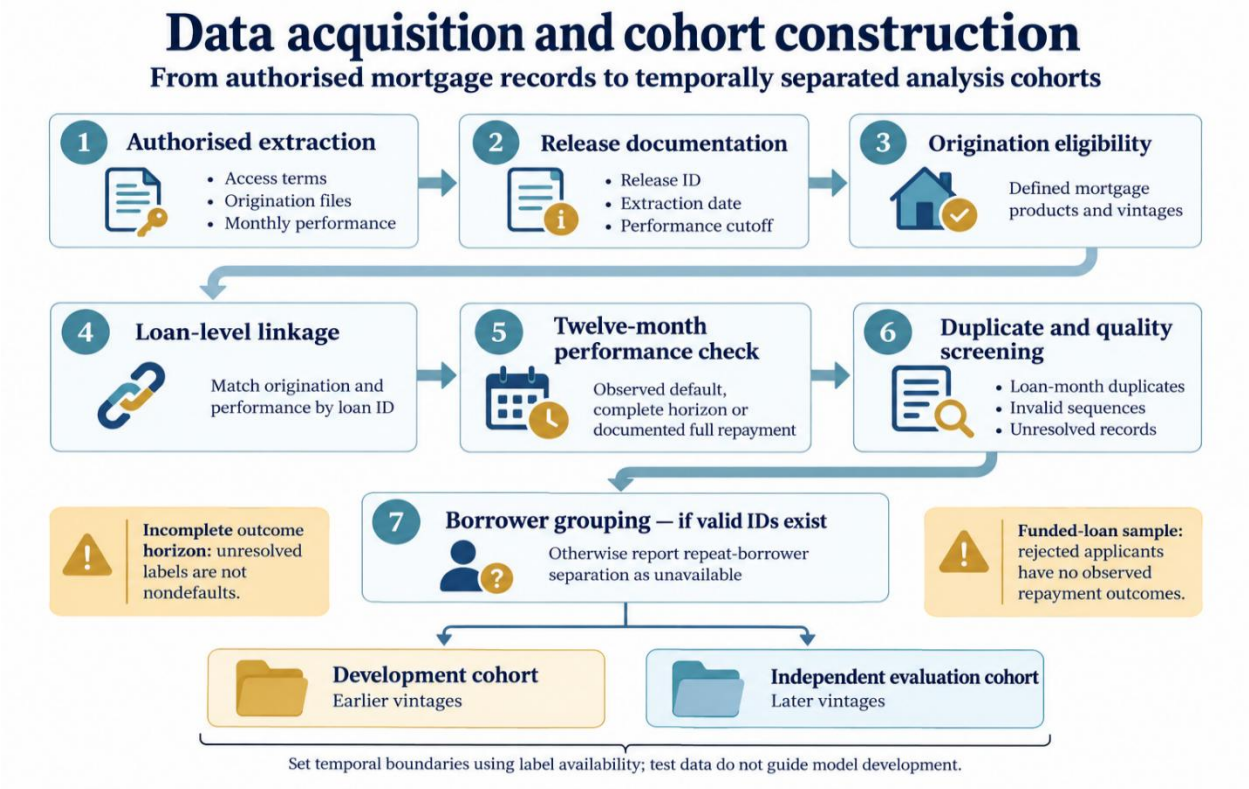


Figure 2. Data-acquisition and cohort-construction pipeline.

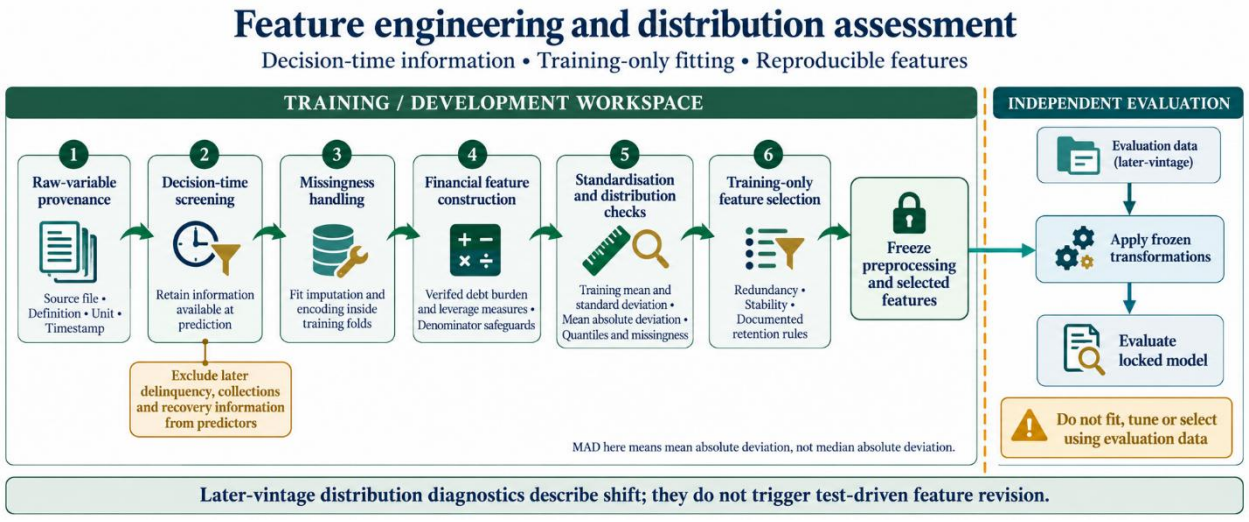


Figure 3. Feature-engineering and distribution-assessment workflow.

4. TRAINING, HYPERPARAMETER TUNING AND DECISION OPTIMISATION

4.1 Temporal Splitting and Borrower-Level Separation

Earlier origination vintages will support development, while later eligible vintages will provide independent temporal testing. Exact boundaries will be prespecified after examining observation coverage and event counts, without inspecting test performance. Within development data, chronological windows will separate fitting, validation, calibration and threshold selection.

Each simulated deployment date will permit training only on outcomes already observable by that date. A twelve-month label horizon therefore requires sufficient separation between training originations and subsequent prediction periods. Using future outcomes or transformations would introduce leakage even when origination dates appear correctly ordered [13].

Loan identifiers will prevent duplicate loans crossing partitions. Where valid borrower identifiers exist, an unseen-borrower evaluation will exclude cross-partition borrower overlap. Repeat-customer evaluation will be reported separately when relevant. Without borrower linkage, the study will acknowledge that unseen-borrower performance cannot be guaranteed.

Partition sizes will reflect default counts, seasonal coverage and feasibility of probability calibration. Calibration and threshold periods will retain representative outcome frequencies. The final test sample will remain untouched until predictors, hyperparameters, calibration method and decision rules are locked. Any later revision will require a new independent evaluation.

4.2 Candidate Ensembles and Cost-Sensitive Training

Random forests, gradient boosting and adaptive boosting will be compared with penalised logistic regression. Standard and cost-sensitive variants will share predictors and evaluation loans. Gradient boosting will support weighted probabilistic loss where implemented [16]; AdaBoost will use weighted weak-learner fitting and sample-weight updates [17].

$$\mathcal{L}(\theta) = -\frac{1}{\sum_i w_i}\sum_i w_i \left[y_i \log p_\theta(x_i) + (1-y_i)\log \quad \left(1-p_\theta(x_i)\right)\right] + \lambda\Omega(\theta)$$

Equation 5 defines weighted cross-entropy for compatible probabilistic learners. Here, θ denotes fitted parameters, $p_\theta(x_i)$ the estimated default probability, and \(w_i\) a nonnegative training weight. The denominator normalises total weight; $\Omega(\theta)$ penalises complexity, with strength controlled by λ.

Class weights will assign emphasis within outcome classes. Exposure-based weights will vary across loans using information available during development. Neither will be confused with post-training threshold adjustment. Multiplying these weights without justification could double-count costs [13].

Equation 5 will not be imposed as a universal ensemble objective. Random forests will incorporate weights into supported sampling and impurity calculations; AdaBoost will retain its algorithm-specific updating mechanism. Implementation details will document these differences rather than describe all methods as equivalent.

Weighted objectives can shift estimated class probabilities away from population probabilities. Scores will therefore undergo calibration assessment before economic thresholding. Post-origination exposure information will remain excluded from predictor inputs, and its training-cost role will require temporal justification [13].

4.3 Ensemble Construction and Complexity Controls

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} \alpha_m h_m(x)$$

Equation 6 represents a boosted score assembled through sequential additions to an initial function. Here, $F_0(x)$ is the starting score, $h_m(x)$ the learner added at stage m, M the number of stages, α_m its coefficient, and η learning-rate shrinkage. Repeatedly adding each scaled contribution produces the stated sum [16].

For logistic-loss boosting, the implementation-specific inverse link will convert the score into a default probability. AdaBoost margins will not automatically be interpreted as calibrated probabilities; their scale and conversion require separate documentation [17].

This additive representation does not describe random forests, which aggregate randomised trees rather than fit sequential corrections. Complexity controls will include depth, minimum leaf size, feature subsampling and ensemble size. Smaller learning rates will be evaluated jointly with stage counts. Model capacity will reflect available default events, not merely the number of candidate features or computational resources available for fitting. All settings will be recorded with software versions.

4.4 Hyperparameter Search and Internal Validation

Hyperparameters will be selected using rolling-origin validation within development data. Where feasible, nested temporal assessment will separate search from performance estimation. Preprocessing will be refitted in each training window rather than shared across folds, preventing information leakage [13].

Search ranges will cover tree depth, minimum leaf size, feature subsampling and ensemble size. Boosting will additionally vary learning rate, iterations and supported regularisation. Penalised regression will tune penalty strength and mixing parameters. Cost ratios and class weights will be selected within development procedures, not inferred from test outcomes.

A random-search budget will provide comparable tuning opportunities; random search offers an established alternative to exhaustive grids [18]. Early stopping will use development validation windows only. Seeds, versions, runtime, hardware and unsuccessful configurations will be retained.

Selection will minimise a prespecified cost criterion subject to constraints on false positives, approval implications and calibration. Cost inputs will represent stated scenarios rather than arbitrary preferences chosen to favour one algorithm. Final settings will be frozen before independent testing and any subsequent operational or economic interpretation [14].

4.5 Calibration and Threshold Selection

$$t^* = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}} + C_{\mathrm{FN}}}$$

Equation 7 follows from comparing conditional losses. For estimated default probability p, predicting default incurs false-positive loss $C_{\mathrm{FP}}(1-p)$, while predicting nondefault incurs false-negative loss $C_{\mathrm{FN}}p$. Choosing default when the first is no larger than the second and rearranging yields the stated threshold [4].

This derivation assumes calibrated probabilities, fixed nonnegative misclassification costs, positive total cost and zero costs for correct classifications. It also requires a defined relationship between prediction and lending action. Increasing missed-default cost lowers the threshold, but may increase unnecessary rejection.

Case-specific costs produce case-specific thresholds rather than one portfolio-wide cutoff. A complete profit model incorporating interest, funding, recoveries and capacity constraints may require another rule. Calibration will precede threshold selection using development observations; insufficient information will favour simpler calibration procedures. Models and thresholds will then be locked before testing whether fewer missed defaults translate into reliable, economically defensible performance under the prespecified assumptions and operational constraints in practice.

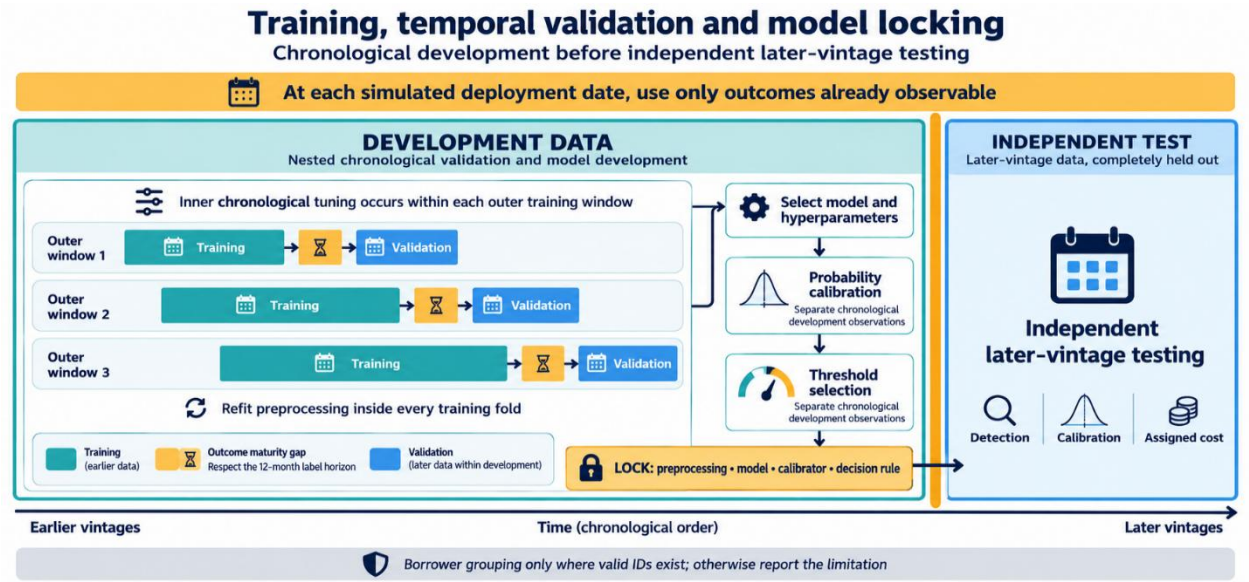


Figure 4. Training, temporal validation, tuning and model-locking architecture.

5. INDEPENDENT TESTING, BENCHMARKING AND STANDARDS COMPARISON

5.1 Classification, Calibration and Operational Metrics

Independent testing will report false-negative counts and rates, default recall, specificity and precision on identical eligible loans. Recall will use observed defaults as its denominator; precision will use predicted defaults, and specificity will use observed nondefaults. Fewer missed defaults will not imply lower underlying default prevalence.

ROC-AUC and PR-AUC will describe ranking, with precision–recall findings interpreted against test prevalence [22]. Calibration curves, intercept, slope and Brier score will assess probability reliability and overall error. These dimensions are complementary rather than interchangeable indicators of model quality [21].

Threshold-specific results will include simulated approval implications under an explicit prediction-to-action mapping. Because the source contains funded mortgages, these will represent retrospective policy simulations, not observed outcomes for previously rejected applicants. Runtime, training requirements and inference latency will be measured using documented hardware and batch sizes.

Subgroup reporting will include denominators, outcome counts and uncertainty. Comparable aggregate accuracy will not establish reliable predictions for every group. The evaluation will retain unsuccessful models and identify cases where sparse information prevents conclusions or stable comparison, rather than treating unavailable estimates as evidence of acceptable performance [24].

5.2 Cost Evaluation and Paired Benchmarking

$$\overline{C}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[C_{\mathrm{FN},i}\, y_i \mathbb{1}(p_i < t) + C_{\mathrm{FP},i}(1 - y_i)\mathbb{1}(p_i \geq t)\right]$$

Equation 8 averages penalties over the test sample. The first indicator selects defaulted loans scored below threshold t, while the second selects performing loans scored at or above it. Each selected case contributes its assigned false-negative or false-positive cost. Dividing by n produces currency per evaluated loan.

Changing t changes both error sets and therefore the average score. The expression assumes zero correct-classification costs and specified penalties; it does not calculate realised portfolio profit or establish that rejected defaults represent fully avoidable losses.

Cost-sensitive models will be compared with standard and threshold-tuned benchmarks using paired predictions. Uncertainty will preserve loan dependence and temporal structure where identifiable. Statistical comparison procedures must match the evaluation design rather than treat repeated measurements as independent datasets [23].

A predict-default-for-all comparator will expose trivial recall gains achieved through universal rejection. Recall will also be compared at matched false-positive or simulated approval rates. This will distinguish improved discrimination from a more restrictive decision policy, while scenario analyses will show dependence on uncertain loss and opportunity-cost assumptions.

5.3 Comparison With Standards and Governance Expectations

Standards comparison will examine default mapping, documentation, validation, stability and oversight. Basel requirements for relevant internal rating systems include out-of-time and out-of-sample testing, but their applicability depends on institutional and supervisory context [7]. Research using a public mortgage dataset will not establish regulatory compliance.

NIST AI RMF provides a voluntary framework addressing validity, reliability, transparency and harmful-bias management [24]. These principles will guide assessment of evidence and governance arrangements rather than supply numerical performance targets.

The manuscript will distinguish binding institutional requirements, voluntary frameworks, empirical model benchmarks and study-specific acceptance criteria. Neither Basel nor NIST will be presented as prescribing a universal acceptable recall or ROC-AUC. Operational tolerances will instead be justified through the intended use, cost assumptions and consequences.

Table 2 will map each evaluation dimension to evidence, comparator and limitations. Gaps in independent validation, subgroup coverage or monitoring will remain visible. A model meeting the study's performance criteria will therefore require institutional review before deployment, including consideration of data rights, operational capacity and the consequences of automated decisions.

Table 2. Benchmarking metrics and standards-based evaluation matrix.

Evaluation dimension	Metric/evidence	Comparator or framework	Study acceptance basis	Interpretation limitation
Default identification	Event mapping, horizon and label completeness	Applicable default definition; Basel CRE36 where relevant [7]	Reproducible mapping and adequate observation	Operational labels do not automatically satisfy regulatory definitions
Missed-default detection	False-negative count, rate and recall	Standard and threshold-tuned models	Improvement under prespecified operational constraints	Recall can increase through indiscriminate rejection
Ranking performance	ROC-AUC and PR-AUC	Identical test loans and benchmark models	Paired differences with uncertainty	Ranking does not establish calibration or economic benefit
Probability	Calibration curve, intercept,	Independent observed	Prespecified, application-	Good overall calibration can

Evaluation dimension	Metric/evidence	Comparator or framework	Study acceptance basis	Interpretation limitation
reliability	slope and Brier score	outcomes [21]	appropriate tolerances	conceal subgroup errors
Classification cost	Average cost per loan; cost differences	Equation 8 and stated cost scenarios	Lower cost without unacceptable competing consequences	Assigned penalties are not realised portfolio profit
Temporal reliability	Later-vintage performance and drift assessment	Out-of-time validation expectations [7]	Acceptable stability across relevant periods	Historical stability does not guarantee future performance
Statistical comparison	Paired differences and dependence-aware uncertainty	Evaluation-design-appropriate methods [23]	Consistent evidence beyond isolated favourable estimates	Related windows are not independent datasets
Subgroup reliability	Group-specific errors, calibration and information counts	NIST AI RMF principles [24]	Documented limitations and justified subgroup assessment	No single metric proves fairness; demographic data may be unavailable
Operational feasibility	Runtime, latency, resource use and review capacity	Intended lending workflow	Prespecified resource and service constraints	Laboratory timing may differ from production performance
Governance readiness	Provenance, documentation, oversight and monitoring	Applicable requirements and voluntary frameworks	Reviewable evidence and assigned responsibilities	Research performance does not establish deployment approval

5.4 Robustness, Cost Scenarios and Subgroup Testing

Robustness analyses will vary false-negative and false-positive cost ratios, exposure assumptions and default prevalence. Prevalence scenarios will be labelled hypothetical and distinguished from observed test results. Distribution shifts can involve predictor distributions, class proportions or relationships between predictors and outcomes [25].

Vintage and available product differences will be evaluated without treating related temporal samples as independent evidence. Alternative missing-data procedures and calibration choices will test dependence on development assumptions. Subgroup false negatives and false positives will be examined where legally available, suitable information permits.

Matched approval-rate comparisons will assess whether conclusions survive equivalent decision restrictiveness. Sparse outcomes, unavailable borrower linkage and uncertain costs will limit interpretation. Economic consequences and robustness, rather than isolated metric improvements, will guide subsequent discussion of comparative findings and their limits before any practical implementation decision.

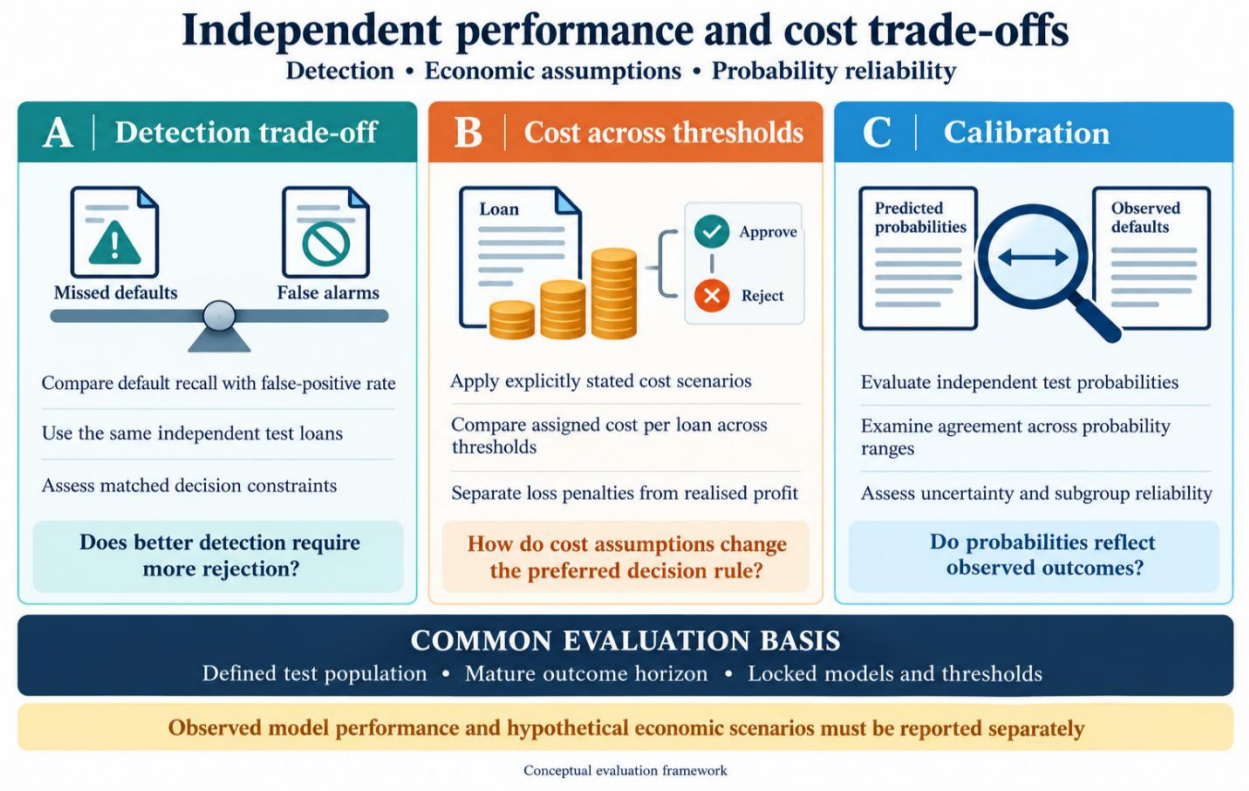


Figure 5. Independent performance and cost trade-offs.

6. RESULTS INTERPRETATION, LIMITATIONS AND DEPLOYMENT IMPLICATIONS

6.1 Comparative Results and Mean Performance Differences

Comparative interpretation will begin with the acquired cohort's composition, exclusions, origination coverage and observed twelve-month test prevalence. No numerical findings are available, so model superiority, default counts and economic gains remain unestablished. Table 3 will present outputs after independent testing rather than illustrative values presented as observations.

For every model and condition, reporting will include missed defaults, false-negative rate, recall, precision, calibration and average assigned cost. All comparisons will use the same eligible test loans. Apparent improvements will be examined against the threshold-tuned baseline to determine whether cost-sensitive training adds information beyond a changed operating point.

Paired differences across temporal validation windows will be summarised using their mean, standard deviation and, where informative, mean absolute deviation. Confidence intervals will account for the identifiable dependence structure. Related windows will not be treated as independent replications merely to obtain narrow uncertainty intervals.

Consistent gains across vintages would support reproducibility within the observed setting, while isolated favourable windows would require qualification. Null findings, failed configurations and weaker models will remain visible. Selection on funded loans limits the population represented by these comparisons, because repayment outcomes for rejected applicants are unavailable [28]. Consequently, even strong retrospective performance would establish neither benefits for all applicants nor improved lending decisions. Conclusions will distinguish empirical observations from model assumptions and untested deployment consequences throughout the interpretation process.

Table 3. Independent model-performance, cost and operational benchmarking results.

Model/condition	Test N	False-negative rate	Recall	Precision	PR-AUC	Brier score	Cost/loan ($)	Simulated approval rate	Illustrative training runtime (s)	Uncertainty
Penalised logistic regression: conventional threshold	10,000	0.70	0.30	0.600	0.48	0.071	71.00	95%	4	Not estimated
Penalised logistic regression: tuned threshold	10,000	0.30	0.70	0.350	0.48	0.071	36.50	80%	4	Not estimated
Random forest: conventional threshold	10,000	0.60	0.40	0.667	0.55	0.066	61.00	94%	85	Not estimated
Random forest: tuned threshold	10,000	0.25	0.75	0.375	0.55	0.066	31.25	80%	85	Not estimated
Cost-sensitive random forest: tuned threshold	10,000	0.20	0.80	0.400	0.58	0.065	26.00	80%	92	Not estimated
Gradient boosting: conventional threshold	10,000	0.55	0.45	0.750	0.61	0.060	55.75	94%	48	Not estimated
Gradient boosting: tuned threshold	10,000	0.20	0.80	0.400	0.61	0.060	26.00	80%	48	Not estimated
Cost-sensitive gradient boosting: tuned threshold	10,000	0.15	0.85	0.425	0.64	0.059	20.75	80%	54	Not estimated
AdaBoost: conventional threshold	10,000	0.65	0.35	0.500	0.50	0.074	66.75	93%	32	Not estimated
AdaBoost: tuned threshold	10,000	0.30	0.70	0.350	0.50	0.074	36.50	80%	32	Not estimated
Cost-sensitive AdaBoost: tuned threshold	10,000	0.22	0.78	0.390	0.54	0.070	28.10	80%	37	Not estimated
Predict default for every loan	10,000	0.00	1.00	0.100	Not applicable	Not applicable	45.00	0%	Not applicable	Not estimated

6.2 Explaining the Sources of Improvement

Sources of improvement will be examined through the three experimental conditions. Comparing standard models at conventional and tuned thresholds will isolate changes attributable to decision rules. Comparing threshold-tuned standard and cost-sensitive models under matched constraints will assess whether training changes provide additional value.

Better ranking, improved probability calibration and stricter rejection represent different mechanisms. A model may detect more defaults simply by flagging more loans; matched false-positive or simulated approval rates will therefore remain central to interpretation. Economic comparisons will retain identical cost assumptions across models.

Variable patterns will be examined for stability across vintages, algorithms and resamples. Correlated financial measures may redistribute importance,

making individual rankings sensitive to the explanation method and reference population [31]. Explanations will describe predictive behaviour rather than effects of changing borrower characteristics. Prediction and causal inference require different questions and assumptions [30].

Borrower-level consequences will include potential unnecessary rejection and missed assessment opportunities. Portfolio interpretation will consider concentration, loss severity and operational capacity. These implications will remain hypotheses until tested within a defined lending workflow rather than inferred from classification improvements alone or rankings.

6.3 Methodological and Economic Limitations

Funded-loan data condition analysis on earlier lending and acquisition decisions. Rejected applicants' repayment outcomes remain unobserved, and reject-inference procedures cannot be assumed to recover them without strong assumptions [28]. Freddie Mac mortgages also represent a narrower setting than consumer credit generally.

Label maturity, early repayment, censoring and inconsistent status coding may alter observed defaults. Research distinguishing whether from when borrowers default highlights the relevance of event timing [29]. A fixed twelve-month outcome cannot describe the complete lifetime repayment process.

Cost estimates depend on exposure, recoveries, collection expenses and foregone margin. Retrospective loss severity does not equal the causal benefit of rejection, particularly when alternative actions affect repayment. Observed loans do not reveal outcomes under every counterfactual approval policy [30].

Macroeconomic change, prevalence shifts and changing servicing practices may reduce calibration and alter error rates. Correlated predictors can destabilise explanations even when aggregate discrimination changes little [31]. Missing borrower linkage may prevent separation of repeat customers.

Finally, small test samples or few defaults can leave important comparisons unresolved. Strong retrospective results would not establish production reliability, improved borrower outcomes or regulatory compliance. These limitations will constrain generalisation and require independent validation, operational assessment and institution-specific review before any deployment decision proceeds.

6.4 Responsible Deployment and Future Research

Future research should validate models using independent lenders, additional products and longer temporal coverage. Mortgage findings should not be transferred to unsecured lending without evidence. Event-time approaches could complement fixed-horizon classification where default timing and incomplete follow-up are central [29].

A monitored shadow evaluation should precede automated decisions, allowing comparison with existing underwriting without immediately changing approvals. Review triggers should cover deteriorating calibration, rising missed-default rates, changing input distributions and unexpected subgroup errors. Recalibration or retraining will require controlled validation and renewed review.

Operational assessment should examine review capacity, latency, explanation usefulness and consequences for applicants. Stakeholders should define acceptable trade-offs and escalation procedures. Prospective evaluation must determine whether decision support improves loss management and credit access beyond existing practice, rather than merely changing classifications. The comparative contribution will therefore depend on reproducible evidence and explicit limits, establishing the conditions needed for responsible use before deployment claims can be justified.

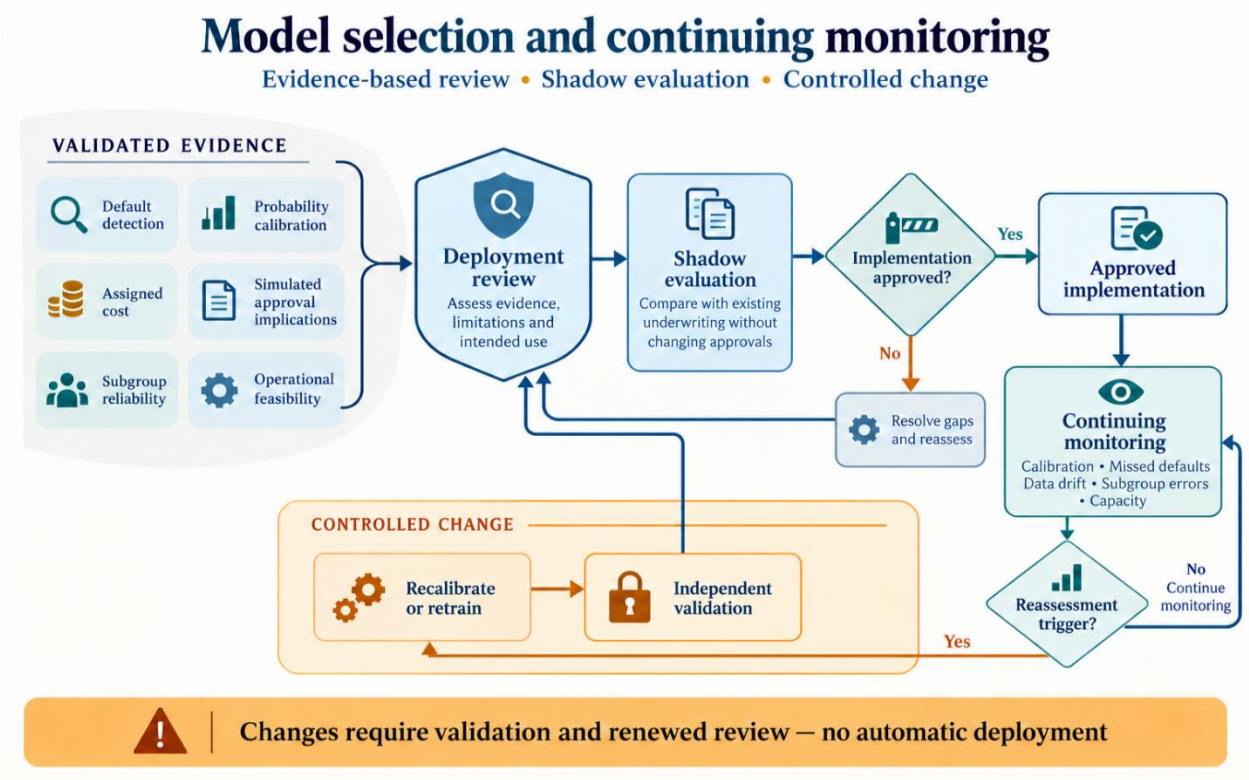


Figure 6. Model-selection and continuing-monitoring framework.

7. CONCLUSION

7.1 Comparative Contribution and Principal Findings

Missed defaults remain an important credit assessment problem because accurate classifiers can overlook loans that subsequently generate losses. However, improving default recall is insufficient when the same result can be achieved by rejecting more applicants. The proposed comparison therefore separates ensemble construction, cost-sensitive training and threshold adjustment, evaluating their contributions under consistent data, cost assumptions and operational constraints.

No completed analysis has established which approach performs best. The populated example table illustrates reporting conventions and cannot support empirical conclusions. Actual findings must determine whether cost-sensitive training improves performance beyond standard models with tuned thresholds, particularly at comparable simulated approval or false-positive rates.

Any demonstrated gains should be interpreted alongside calibration, precision, assigned cost and uncertainty across temporal windows. A simpler model may remain preferable when differences are small or operational demands outweigh benefits. The contribution lies in a reproducible comparative evaluation that identifies context-dependent trade-offs, rather than asserting universal superiority for an ensemble or presenting higher recall as economic success.

7.2 Conditions for Responsible Application

Responsible application requires mature outcome labels, documented exclusions and independent testing that reflects the intended lending setting. Cost assumptions must distinguish assigned misclassification penalties from realised profit and clarify the relationship between predictions and decisions. Calibration, temporal reliability and subgroup errors require assessment, with sparse evidence identified rather than interpreted as acceptable performance.

Deployment should follow external validation and monitored evaluation against existing underwriting. Review procedures must address drift, recalibration, operational capacity, explanation needs and potential borrower consequences. Funded-loan selection and unobserved rejected-applicant outcomes limit conclusions about wider credit access and counterfactual approval policies.

Predictive improvement alone does not demonstrate lower realised losses, fairer decisions or better borrower outcomes. These benefits require prospective evidence within a defined workflow. Disciplined evaluation can nevertheless support reliable and responsible credit decisions when uncertainty, competing costs and implementation consequences remain visible throughout institutional review.

REFERENCE

1. Aruleba I, Sun Y. Effective credit risk prediction using ensemble classifiers with model explanation. IEEE Access. 2024 Aug 16;12:115015-25.
2. IDOWU GA, OWOLABI JE. ENSEMBLE-BASED FRAUD DETECTION IN NIGERIAN BANKING: A SYNTHETIC DATA BENCHMARK AND COST-SENSITIVE ANALYSIS. 2026 Mar 9;6(1):107-29.
3. Trinh LT. A comparative analysis of consumer credit risk models in Peer-to-Peer Lending. Journal of Economics, Finance and Administrative Science. 2024 Oct 11;29(58):346-65.
4. Singh A. Peer-to-peer loan default prophecy in fintech: A comparative analysis of the predictive performance of machine learning models. Corporate Governance. 2024 Dec;6(2):1-2.
5. Petrides G, Verbeke W. Cost-sensitive ensemble learning: a unifying framework. Data Mining and Knowledge Discovery. 2022 Jan;36(1):1-28.
6. Ong P, Lee CN. A Comparative Study of Credit Default Risk Modelling with Ensemble Machine Learning Methods. InComputational Intelligence for Business Analytics and Decision Systems 2026 May 14 (pp. 239-249). Cham: Springer Nature Switzerland.
7. Chukwunweike J. Coordinating PLC-based load shedding and variable-speed drives to prevent transformer overloading during industrial production peaks. Int J Electr Data Commun. 2025;6(2 Pt B):118-135. doi:10.22271/27083969.2025.v6.i2b.109.
8. Opayemi OO. *Credit Risk Default Prediction: A comparative analysis of Machine Learning Techniques* (Doctoral dissertation, Dublin, National College of Ireland).
9. Akuji II, Olanrewaju AB, Ahmed TA, Alabi AJ, Adeyemi IB. A comparative study of classical, bagging, and hybrid methods for optimizing loan default prediction. Computer Science and Information Technologies. 2026 Jul 1;7(2):179-95.
10. Ogan RK. Machine-learning-based API failure prediction and root-cause analytics for improving reliability of enterprise banking integration architectures at scale. Int J Comput Artif Intell. 2021;2(2):142-153. DOI:10.33545/27076571.2021.v2.i2a.378.
11. Ekinci A, Sen S. Forecasting bank failure in the US: A cost-sensitive approach. Computational Economics. 2024 Dec;64(6):3161-79.
12. Cao J, Lu H, Wang W, Wang J. A loan default discrimination model using cost-sensitive support vector machine improved by PSO.

Information Technology and Management. 2013 Sep;14(3):193-204.

13. Alebiosu O. Engineering stakeholder alignment frameworks for resolving land rights, environmental compliance, and project delivery conflicts. Int J Appl Res. 2024;10(12):401-412.
14. Kim J, Choi K, Kim G, Suh Y. Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost. Expert Systems with Applications. 2012 Mar 1;39(4):4013-9.
15. Akinjole A, Shobayo O, Popoola J, Okoyeigbo O, Ogunleye B. Ensemble-based machine learning algorithm for loan default risk prediction. Mathematics. 2024 Oct 31;12(21):3423.
16. Bhandari, Ambika Baniya. 2026. "Bridging the Public Health–Cloud Security Divide: Workforce Implications of the HIPAA Security Rule Modernization for Public Health Informatics". *Current Journal of Applied Science and Technology* 45 (8):33-44. https://doi.org/10.9734/cjast/2026/v45i84730.
17. Taha AH, Abu-Naser SS. Predicting Loan Defaulters: A Comprehensive Analysis and Comparative Study of Machine Learning Algorithms Using a Large-Scale Loan Default Dataset. InProceedings of Eighth International Conference on Information System Design and Intelligent Applications 2025 Jan 3 (pp. 15-28). Singapore: Springer Nature Singapore.
18. Ramos SM. *Cost-Sensitive Machine Learning for Peer-To-Peer Credit Scoring: Balancing Predictive Accuracy, Risk and Regulatory Compliance* (Master's thesis, Universidade NOVA de Lisboa (Portugal)).
19. Bhandari, Ambika Baniya. 2026. "Quality Control Frameworks for Electronic Health Record Data Workflows in Resource-Constrained Settings: A Structured Review With Implications for Global Health Information Management". *Current Journal of Applied Science and Technology* 45 (8):1-16. https://doi.org/10.9734/cjast/2026/v45i84728.
20. Neema S, Soibam B. The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. Journal of Management Science and Business Intelligence. 2017 Aug;2(2):36-41.
21. Papouskova M, Hajek P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decision support systems. 2019 Mar 1;118:33-45.
22. Iyanuoluwa Oluwafemi. (2024). ENVIRONMENTAL MONITORING STRATEGIES INTEGRATING MOLECULAR DETECTION TECHNOLOGIES TO ENHANCE AGRICULTURAL BIOSECURITY AND SUSTAINABLE CROP PRODUCTION SYSTEMS. International Journal Of Engineering Technology Research & Management (IJETRM), 10(12), 757–772. https://doi.org/10.5281/zenodo.21152414
23. Li Z, Zhang J, Yao X, Kou G. How to identify early defaults in online lending: A cost-sensitive multi-layer learning framework. Knowledge-Based Systems. 2021 Jun 7;221:106963.
24. Aleksandrova Y, Armianova M. Evaluation of cost-sensitive machine learning methods for default credit prediction. In2022 International Conference Automatics and Informatics (ICAI) 2022 Oct 6 (pp. 89-94). IEEE.
25. Bhetwal S, Bastakoti R, Acharya N, Gupta GK, Bhandari AB. Benchmarking Local LLMs for Natural-Language-to-SQL Querying in Biopharmaceutical Manufacturing: An Empirical Benchmark on Consumer-Grade Hardware. arXiv preprint arXiv:2606.01338. 2026 May 31.
26. Chen N, Ribeiro B, Chen A. Comparative study of classifier ensembles for cost-sensitive credit risk assessment. Intelligent Data Analysis. 2015 Jan;19(1):127-44.
27. Gaffari MA. Explainable machine learning for predicting depression risk using sleep patterns, self-esteem, and psychosocial factors. International Journal of Computer Applications Technology and Research. 2026;15(11):26–37. doi:10.7753/IJCATR1511.1003.
28. Xiao J, Li S, Tian Y, Huang J, Jiang X, Wang S. Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring. Scientific Reports. 2025 Feb 18;15(1):6000.
29. Luo Z. An Adaptive Ensemble Learning Approach with Cost-Sensitive Optimization for Credit Default Prediction. InProceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy 2026 Mar 13 (pp. 523-531).
30. Sandra Darkwa. Transformer-Driven Strategic Communication Analytics for Identifying Synthetic Media Manipulation and Coordinated Narrative Amplification Networks Digitally. Int J Literacy Educ 2024;4(2):333-346. DOI: 10.22271/27891607.2024.v4.i2d.434
31. Yotsawat W, Wattuya P, Srivihok A. A novel method for credit scoring based on cost-sensitive neural network ensemble. Ieee Access. 2021 May 25;9:78521-37.