Benchmarking Machine Learning Algorithms for Credit Scoring: A Comprehensive Empirical Evaluation

Author First Author Last[1*]

[1*]Department, Institution, City, India.

Corresponding author(s). E-mail(s): author@institution.ac.in;

Abstract

Credit scoring is a critical function in retail lending, requiring lenders to estimate the probability that a borrower will default on a loan obligation. The academic literature has produced a large number of proposed classification algorithms for scorecard development, yet systematic empirical evidence on the comparative predictive performance of modern methods—including advanced ensemble learners and gradient boosting variants—remains limited for emerging market contexts. This paper presents a large-scale benchmarking study of *[N]* classification algorithms evaluated across *[K]* real-world credit scoring datasets. We assess classifier performance using multiple accuracy indicators spanning discriminatory ability, probability calibration, and categorical prediction correctness, including the Area Under the ROC Curve (AUC), the Brier Score (BS), the H-measure, the Kolmogorov–Smirnov statistic (KS), the Partial Gini index (PG), and the Percentage Correctly Classified (PCC). Statistical hypothesis testing via the Friedman test and pairwise post-hoc comparisons are employed to draw rigorous conclusions. Our results indicate that *[key finding 1, e.g., heterogeneous ensemble methods consistently outperform individual classifiers]*. *[Key finding 2, e.g., logistic regression remains a competitive baseline]*. We also examine the financial implications of deploying more accurate scorecards and find that *[financial result]*. This study provides updated guidance for practitioners and researchers in credit risk modelling regarding the choice of classification algorithm and performance evaluation methodology.

Keywords: Credit scoring, Benchmarking, Ensemble classifiers, Machine learning, Probability of default, AUC, Model comparison

JEL Classification: G21 , C44 , C53

1 Introduction

Credit scoring is concerned with developing empirical models to support decision-making in the retail credit business [1, 2]. Lenders employ predictive scorecards—statistical models that estimate the probability of default (PD) for a loan applicant—to decide whether to grant credit and under what terms. The accuracy of these scorecards has direct financial consequences: a more accurate scorecard translates into fewer bad loans approved and fewer creditworthy applicants rejected, thereby improving the lender's profitability while potentially widening access to credit.

The foundational benchmarking study by [3] provided a systematic comparison of classification algorithms for credit scoring. That study demonstrated that several advanced methods, including neural networks and support vector machines, offered modest but meaningful improvements over logistic regression, which remains the industry standard. More recently, [4] updated this comparison by including a broader set of 41 classifiers—including modern ensemble methods such as random forests, boosting algorithms, and selective heterogeneous ensembles—and found that heterogeneous ensemble strategies occupy the top performance ranks across multiple accuracy measures and datasets.

Despite these advances, several gaps motivate a further update and extension of this literature. *First*, the machine learning landscape has evolved substantially since 2015. Methods such as gradient boosting machines (e.g., XGBoost [5] and LightGBM [6]), deep neural network architectures, and attention-based models have emerged as state-of-the-art tools in applied prediction problems. Their credit-scoring performance has not been systematically assessed in a large-scale, multi-dataset benchmarking framework. *Second*, the majority of prior studies employ data from developed-economy lenders, predominantly from Europe and the United States. Whether classifier rankings generalise to *[context, e.g., Indian retail credit / emerging market lenders]* is an open empirical question. *Third*, evaluation methodology continues to evolve. In particular, the H-measure [7] and the Partial Gini index offer theoretically motivated alternatives to the AUC that have seen limited adoption in credit scoring research.

This paper addresses these gaps through the following contributions:

1. We conduct a large-scale benchmark of *[N]* classification methods—spanning individual classifiers, homogeneous ensembles, heterogeneous ensembles, and deep learning approaches—across *[K]* credit scoring datasets. *[Describe dataset sources briefly.]*
2. We employ six performance indicators, including the H-measure and the Partial Gini index, to provide a multi-dimensional assessment of scorecard quality.
3. We apply a statistically rigorous testing framework based on the Friedman test and the Rom procedure for family-wise error rate control in pairwise comparisons [8, 9].
4. We assess the financial value of more accurate classifiers by estimating expected misclassification cost reductions relative to a logistic regression baseline.

The remainder of this paper is organised as follows. Section 2 reviews prior benchmarking literature and motivates the current study. Section 3 describes the classification algorithms included in the comparison. Section 4 presents the experimental design, including the datasets, performance indicators, and statistical testing procedure. Section 5 reports empirical results. Section 6 discusses findings, and Section 7 concludes.

2 Literature Review

2.1 Benchmarking in retail credit scoring

The systematic comparison of classification algorithms for scorecard development has a substantial history. Table 1 summarises key empirical classifier comparison studies in retail credit scoring. *[Populate with studies most relevant to your scope.]* Early work established logistic regression (LR) and discriminant analysis as reference methods, against which neural networks and decision trees were compared [10, 11]. [3] provided the most comprehensive early benchmark, comparing 17 classifiers across eight datasets and finding that relatively simple methods such as logistic regression performed competitively.

A growing body of literature subsequently examined ensemble classifiers in credit scoring. [12] and [13] found evidence that ensemble methods such as random forests and boosting improve upon individual classifiers. [4] confirmed and extended these findings at scale, concluding that heterogeneous ensemble selection algorithms—particularly the Hill-Climbing Ensemble Selection with Bagging (HCES-Bag) [14]—achieved the strongest overall performance. Logistic regression remained a surprisingly competitive individual classifier, significantly outperforming artificial neural networks in their benchmark.

More recent work has incorporated gradient boosting variants. *[Cite 2–3 recent papers on XGBoost/LightGBM in credit scoring.] [Cite relevant deep learning papers.]* However, no study to date has combined this full breadth of methods in a single, rigorous benchmarking framework with multiple accuracy measures and formal statistical testing across *[your specific dataset domain]*.

2.2 Performance measurement in credit scoring

Performance indicators in credit scoring split broadly into three categories: measures of discriminatory ability (e.g., AUC, KS, Gini/PG), measures of the accuracy of probability estimates (e.g., Brier Score), and measures of categorical prediction accuracy (e.g., PCC, misclassification rate) [4, 7].

The AUC is by far the most widely used indicator. [7] identified a theoretical limitation of the AUC in that it implicitly assigns different, and potentially inconsistent, misclassification cost weights for different classifiers, making cross-classifier AUC comparisons potentially misleading. The H-measure was proposed as a remedy. However, empirical evidence suggests that AUC and H-measure rankings are highly correlated [4], implying that either measure supports broadly similar conclusions in practice.

The Brier Score (BS) captures calibration—the degree to which predicted probabilities agree with observed outcome frequencies—a dimension of performance distinct from discrimination. [4] showed that the BS provides complementary information to ranking measures, and recommend its inclusion in multi-measure evaluations. Similarly, the Partial Gini (PG) index focuses performance assessment on the most important segment of the score distribution for credit applications—the region near the acceptance threshold.

We adopt all six indicators used by [4] to ensure comparability with that study and to provide a multi-dimensional performance profile for each classifier.

2.3 Statistical testing in classifier comparisons

A persistent methodological limitation in empirical classifier comparisons is the neglect of formal statistical testing. Without hypothesis tests, observed rank differences may reflect sampling noise rather than genuine differences in classifier performance [9]. Parametric tests (e.g., paired t-tests) are widely used but assume normality of test statistics that is rarely met in classifier comparisons. [9] recommends non-parametric alternatives: the Friedman test to verify that at least two classifiers differ significantly, followed by pairwise Wilcoxon signed-rank tests or the Nemenyi test with appropriate multiple-comparison correction.

[4] adopted the Friedman test with pairwise comparisons and Rom p-value adjustment [8]—a uniformly more powerful procedure than Bonferroni or Holm correction. We follow the same testing framework here.

3 Classification Algorithms

We compare *[N]* classification algorithms drawn from three broad families: individual classifiers, homogeneous ensemble classifiers, and heterogeneous ensemble classifiers. The full list is provided in Table 2. Below we briefly characterise each family; algorithmic details and hyperparameter settings are documented in Appendix A.

3.1 Individual classifiers

Individual classifiers develop a single classification model from training data. We include both statistical methods that model $p(+|\mathbf{x})$ directly (e.g., logistic regression, LDA) and probabilistic methods that estimate class-conditional densities and apply Bayes' rule (e.g., naive Bayes). Semi-parametric kernel methods (SVM) and tree-based methods (CART) are also included. Following [4], we treat logistic regression (LR) as the industry-standard baseline, given its widespread adoption in retail credit.

[Briefly describe the specific individual classifiers included in your study, e.g., Logistic Regression (LR), Regularised Logistic Regression (LR-R), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Decision Trees (CART), k-Nearest Neighbour (k-NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), and any additional methods.]

3.2 Homogeneous ensemble classifiers

Homogeneous ensembles combine predictions of multiple base models developed using the same algorithm. Bagging [15] derives independent base models from bootstrap samples; boosting [16] builds base models sequentially to correct errors of the current ensemble. We include the following homogeneous ensembles: *[e.g., Bagged Decision Trees, Random Forest (RF), Rotation Forest, AdaBoost, XGBoost, LightGBM, Stochastic Gradient Boosting (SGB).]*

3.3 Heterogeneous ensemble classifiers

Heterogeneous ensembles combine base models built by *different* algorithms. Simple combination rules (unweighted averaging, weighted averaging, stacking) are included alongside selective ensemble strategies that search the model library for the optimal subset of base models [14, 17]. We consider both static selection approaches (which identify the model subset once) and, as benchmarks, two dynamic approaches. *[List the specific heterogeneous ensembles included, e.g., Simple Average (AvgS), Weighted Average (AvgW), Stacking, HCES-Bag, Top-T Ensemble, etc.]*

4 Experimental Setup

4.1 Credit scoring datasets

The empirical evaluation is based on *[K]* retail credit scoring datasets. Table 3 summarises the key characteristics of each dataset, including the number of observations, independent variables, prior default rate, and cross-validation scheme.

[Describe each dataset briefly. For example:]

- **Dataset 1:** *[N observations, p variables. Description of source and context.]*
- **Dataset 2:** *[N observations, p variables. Description of source and context.]*
- **Continue for all datasets.**

The datasets include covariates drawn from application forms (e.g., income, employment status, loan amount) and credit bureau information (e.g., credit history, existing obligations). A binary response variable indicates loan default ($y = 1$) or non-default ($y = 0$). *[Adjust coding convention as appropriate.]*

We note that class imbalance is prevalent in credit scoring data: the fraction of defaults (the minority class) typically ranges from *[XX%]* to *[XX%]* across our datasets (see Table 3). Consistent with [4], we do not apply resampling (e.g., SMOTE) prior to model training, so as to assess the *raw* relative performance differences across classifiers on a common ground. Calibration of probability predictions is applied using Platt scaling [18] for classifiers that do not natively produce well-calibrated probabilities.

4.2 Performance indicators

We assess classifier performance using six indicators that together provide a multidimensional view of scorecard quality:

1. **Area Under the ROC Curve (AUC):** Measures overall discriminatory ability. Equals the probability that a randomly selected positive instance receives a higher predicted score than a randomly selected negative instance.

2. **H-measure (H):** A coherent alternative to the AUC that specifies a common cost distribution across classifiers [7]. We follow [4] and use a Beta(2,2) distribution to specify the relative severity of false positives and false negatives.
3. **Partial Gini index (PG):** Concentrates performance assessment on the high-score region of the score distribution ($p(+|\mathbf{x}) \leq b$, with $b = 0.4$), capturing classifier accuracy for the most likely defaulters—the segment most relevant for credit decisions.
4. **Kolmogorov–Smirnov statistic (KS):** The maximum difference between the cumulative score distributions of positive and negative classes, widely used by practitioners in credit risk.
5. **Brier Score (BS):** The mean-squared error between predicted probabilities $p(+|\mathbf{x})$ and observed outcomes. Assesses calibration of probability predictions; lower values are better.
6. **Percentage Correctly Classified (PCC):** Proportion of observations correctly classified using an optimal threshold τ, set such that the fraction of cases classified as positive equals the prior default rate.

The first four indicators are "the higher the better"; for BS, we rank classifiers on $1 - \text{BS}$ to maintain a consistent direction across all six measures.

4.3 Data preprocessing and partitioning

Missing values are imputed using mean/mode replacement for numeric/nominal attributes respectively. We construct two versions of each dataset: one retaining mixed variable types and one in which nominal variables are converted to numeric scores using weight-of-evidence (WoE) coding [19]. This accommodates classifiers that require purely numeric inputs (e.g., SVMs, ANNs) while allowing tree-based and Bayes classifiers to operate on their preferred input representations. We use the version that yields the best validation performance for each classifier.

Data partitioning follows a stratified $N \times 2$-fold cross-validation scheme [20], where N is set according to dataset size (see Table 3). Specifically, each dataset is randomly split into two equal halves; the first half is used for training and the second for testing, and the procedure is then reversed. This is repeated N times with different random splits. For classifiers with hyperparameters, an inner five-fold cross-validation on the training partition is used to select the best hyperparameter configuration before evaluating on the held-out test partition.

We develop multiple classification models per algorithm by varying hyperparameter settings (see Appendix A for the full grid). Prior to comparing classifiers, we select the best-performing hyperparameter configuration per classifier per performance measure on the validation data, following [4].

4.4 Statistical testing

We employ the non-parametric Friedman test to verify that classifier ranks are not all equal [9]. Given a significant Friedman result, we conduct pairwise comparisons of each classifier against a control classifier (the best-performing classifier per performance measure) using the Rom procedure to control the family-wise error rate [8]. For selected classifier comparisons, we also report a full pairwise comparison matrix (Table 5) using the z-statistic [9]:

$$z = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6N}}} \tag{1}$$

where R_i and R_j are the average ranks of classifiers i and j, k is the total number of classifiers, and N is the number of datasets. p-values are obtained by referring z-statistics to the standard normal distribution, followed by Bergmann–Hommel correction [21] for the full comparison matrix.

5 Empirical Results

5.1 Benchmarking results

Table 4 presents the average classifier ranks across all *[K]* datasets for each of the six performance measures, along with the grand average rank (AvgR) and high-score position. *[Insert completed Table 4 once results are available.]*

The average ranks in Table 4 serve as the basis for statistical comparisons. We employ the Friedman test to verify that classifier performances are not all equivalent; the test statistic χ^2 and associated p-value are reported in the last row of Table 4. Given the highly significant Friedman result ($p < .001$ for all six performance measures), we proceed with pairwise comparisons against the control classifier.

[Example result sentence: "Heterogeneous ensemble methods occupy the top X ranks overall. The best-performing classifier by AvgR is CLASSIFIER, followed by CLASSIFIER and CLASSIFIER. Logistic regression ranks XXth on average, significantly outperforming X of the N classifiers."]

[Note which classifiers perform significantly worse than the best classifier (indicated by underscored p-values in Table 4), and discuss notable patterns by classifier family.]

5.2 Comparison of selected classifiers

To complement the broad ranking results, Table 5 reports a full pairwise comparison of selected classifiers: logistic regression (LR, the industry benchmark), *[CLASSIFIER 1]* (best individual classifier), *[CLASSIFIER 2]* (best homogeneous ensemble), and *[CLASSIFIER 3]* (best heterogeneous ensemble). *[Populate Table 5 with AvgR values and pairwise adjusted p-values.]*

[Example narrative: "Based on the Friedman χ^2 statistic of VALUE, we reject the null hypothesis that average ranks are equal ($p < .001$). LR predicts significantly less accurately than the best classifier across all six measures. CLASSIFIER predicts significantly more accurately than all other classifiers. The empirical evidence does not provide sufficient support to conclude that CLASSIFIER A and CLASSIFIER B differ significantly ($p = VALUE$)."]

5.3 Financial implications of scorecard accuracy

To assess the business value of more accurate classifiers, we estimate the misclassification costs of a scorecard following [22]. The cost function is:

$$C(s) = C(+|-) \cdot \text{FPR} + C(-|+) \cdot \text{FNR} \tag{2}$$

where $C(+|-)$ is the opportunity cost of denying credit to a good applicant, $C(-|+)$ is the cost of granting credit to a bad applicant (approximated by EAD × LGD less interest received), and FPR (false positive rate) and FNR (false negative rate) are computed at the Bayes-optimal threshold. We normalise costs relative to LR and consider 25 cost ratios $C(+|-) : C(-|+) = 1 : 2, \ldots, 1 : 50$.

Figure 1 presents the expected percentage cost reduction of *[CLASSIFIER 1]*, *[CLASSIFIER 2]*, and *[CLASSIFIER 3]* relative to LR across the range of cost settings. *[Insert Figure 2 once results are available.]*

[Example narrative: "The results in Figure 2 show that CLASSIFIER achieves the largest average cost reduction of X% over LR across cost settings. These improvements are economically meaningful given the scale of credit portfolios. The best classifier loses its advantage relative to LR when misclassification costs of bad risks are high, because EXPLANATION."]

6 Discussion

Our benchmarking results *[summarise key finding, e.g., confirm that heterogeneous ensemble methods outperform individual classifiers and that logistic regression remains competitive among individual methods]*. This is broadly consistent with [4] and provides evidence that these patterns extend to *[your data context, e.g., Indian retail credit datasets]*.

[Discuss any notable deviations from prior benchmarks. For example, do gradient boosting methods (XGBoost, LightGBM) rank more strongly than in earlier studies? Do deep learning approaches perform as expected?]

Several practical implications follow from these findings. First, *[practical implication 1, e.g., practitioners using only LR may be forgoing measurable accuracy gains, particularly when ensemble methods are available in standard modelling platforms]*. Second, *[practical implication 2, e.g., the strong performance of RF and gradient boosting methods suggests that tree-based ensembles represent a robust alternative to LR that does not require the same degree of variable transformation and manual feature engineering]*. Third, the multi-measure evaluation reveals that *[insight from comparing across measures, e.g., classifiers that rank highly on the AUC also tend to rank highly on the H-measure and KS, suggesting that the choice among these ranking measures has limited practical consequence]*.

The financial simulation (Section 5.3) underscores an important caveat: the mapping from statistical accuracy to business value is not one-to-one. *[Discuss the financial findings further.]* This finding reinforces the recommendation by [4] to include both accuracy measures and financial simulations when assessing the value of alternative scorecards.

Several limitations of this study warrant acknowledgement. *[List limitations, e.g., (i) datasets are drawn from a specific context and may not generalise to other lending environments; (ii) we do not consider dynamic scorecard recalibration; (iii) the treatment of missing data relies on simple imputation; (iv) the cost simulation assumes known and fixed cost ratios.]* Future research should address these limitations and examine the explainability of ensemble scorecards in the context of regulatory requirements (e.g., the requirement to provide reasons for adverse credit decisions).

7 Conclusion

This paper presents a comprehensive benchmarking study of *[N]* classification algorithms for retail credit scoring, evaluated across *[K]* datasets using six performance measures and rigorous non-parametric statistical tests. The principal findings are:

1. *[Key finding 1, e.g., "Heterogeneous ensemble classifiers consistently achieve the highest predictive accuracy, occupying X of the top Y ranks across performance measures."]*
2. *[Key finding 2, e.g., "Logistic regression remains the strongest individual classifier and significantly outperforms X of the N classifiers, including several methods regarded as state-of-the-art."]*
3. *[Key finding 3, e.g., "Gradient boosting methods, including XGBoost and LightGBM, rank among the top homogeneous ensemble classifiers, representing a meaningful advance on earlier benchmarks."]*
4. *[Key finding 4 — financial implication summary.]*
5. The Brier Score and Partial Gini index provide complementary information beyond what is captured by the AUC or KS, and should routinely form part of credit scorecard evaluations.

From a practical standpoint, we recommend that lenders consider ensemble methods—particularly random forests and gradient boosting—as alternatives to logistic regression when developing PD scorecards, subject to interpretability and regulatory constraints. We recommend RF as a minimum benchmark against which newly proposed classifiers should be compared.

A key direction for future research is to develop explainability frameworks for high-performing ensemble classifiers that satisfy the interpretability requirements of financial regulators. The integration of SHAP values and other post-hoc explanation methods with the benchmarking approach employed here represents a natural extension of this work.

Acknowledgements. *[Acknowledgements.]*

Declarations

- **Funding:** *[State funding source or "Not applicable".]*
- **Conflict of interest:** The author(s) declare no conflict of interest.
- **Ethics approval:** Not applicable.
- **Consent for publication:** Not applicable.

- **Data availability:** *[Describe data availability or provide links to public datasets used.]*
- **Code availability:** *[Describe code availability, e.g., "Code is available upon request" or provide a repository link.]*
- **Author contribution:** *[Describe each author's contribution.]*

Table 1 Analysis of classifier comparisons in retail credit scoring — literature summary

Retail credit scoring study	Data		Classifiers		Evaluation	
(chronological order)	No. datasets	Obs./vars	No. classifiers	ENS	AUC	ST
[3]	8	1895/21	17	×	×	P
[4]	8	30188/24	41	×	×	F/P
(add rows)	–	–	–	–	–	–
Present study	–	–	–	×	×	F/P

Note: ENS = Ensemble classifiers included; AUC = AUC reported; ST = Statistical testing employed (P = pairwise comparison, F = Friedman test, F/P = both). Complete this table with all reviewed studies.

Figure placeholder: Expected percentage reduction in misclassification costs compared to LR across cost ratio settings $C(+|-):C(-|+) = 1:2$ to $1:50$, for selected classifiers (RF, XGB, HCES-Bag). Y-axis: cost reduction (%); X-axis: $C(-|+)$ setting.

Fig. 1 Expected percentage reduction in misclassification costs compared to logistic regression (LR) across different cost settings $C(-|+)$, assuming $C(+|-) = 1$ and a Bayes-optimal classification threshold.

Table 2 Classification algorithms included in the benchmarking study

Family	Classification algorithm	Acronym	Models
Individual classifier	Logistic regression	LR	1
	Regularised logistic regression	LR-R	–
	Linear discriminant analysis	LDA	1
	Naive Bayes	NB	1
	Decision tree (CART)	CART	–
	k-Nearest neighbour	k-NN	–
	Artificial neural network	ANN	–
	Support vector machine (RBF)	SVM	–
	(additional methods)	–	–
Homogeneous ensemble	Bagged decision trees	Bag	–
	Random forest	RF	–
	Rotation forest	RotFor	–
	AdaBoost	Boost	–
	XGBoost	XGB	–
	LightGBM	LGBM	–
	(additional methods)	–	–
Heterogeneous ensemble	Simple average ensemble	AvgS	1
	Weighted average ensemble	AvgW	1
	Stacking	Stack	–
	HCES with bootstrap sampling	HCES-Bag	–
	Top-T ensemble	Top-T	–
	(additional methods)	–	–

Note: Dashes indicate that the number of models depends on hyperparameter grid size; see Appendix A.

Table 3 Summary of credit scoring datasets

Name	Cases	Variables	Default rate	$N \times 2$ CV	Source
DS1	–	–	–	–	–
DS2	–	–	–	–	–
DS3	–	–	–	–	–
(add rows)	–	–	–	–	–

Note: Default rate is the fraction of positive (default) cases. N×2 CV denotes the number of repetitions of two-fold cross-validation. Replace placeholder rows with actual dataset details.

Appendix A Hyperparameter Settings

Table A1 documents the hyperparameter configurations explored for each classification algorithm. For each algorithm, we evaluate all combinations of the listed settings and select the best-performing configuration on the inner validation fold prior to final test evaluation.

Table 4 Average classifier ranks across datasets for different performance measures

		AUC		PCC		BS		H		AvgR	High score
Classifier family	Classifier	Rank	(p)	Rank	(p)	Rank	(p)	Rank	(p)		
Individual	LR	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	LR-R	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	ANN	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	SVM	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	(others)	–	(–)	–	(–)	–	(–)	–	(–)	–	–
Homogeneous	RF	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	XGB	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	LGBM	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	(others)	–	(–)	–	(–)	–	(–)	–	(–)	–	–
Heterogeneous	AvgS	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	AvgW	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	HCES-Bag	–	(–)	–	(–)	–	(–)	–	(–)	–	–
	(others)	–	(–)	–	(–)	–	(–)	–	(–)	–	–
Friedman χ^2		–	(–)	–	(–)	–	(–)	–	(–)		

Note: Bold face indicates the best classifier (lowest average rank) per performance measure. Values in brackets are adjusted p-values of pairwise comparison to the best classifier (Rom procedure). Underscored values indicate $p < .05$. Replace dashes with actual results.

Table 5 Full pairwise comparison of selected classifiers

Classifier	AvgR	LR	RF	XGB	HCES-Bag
LR	–	—			
RF	–	(.p)	—		
XGB	–	(.p)	(.p)	—	
HCES-Bag	–	(.p)	(.p)	(.p)	—
Friedman χ^2	–	(.p)			

Note: AvgR = average rank across all datasets and performance measures. Cell values are adjusted p-values (Bergmann–Hommel procedure). Replace placeholders with actual computed values.

Table A1 Hyperparameter settings for classification algorithms

Algorithm	Hyperparameter	Values explored
LR-R	Regularisation (λ)	(fill in)
k-NN	Number of neighbours (k)	(fill in)
ANN	Hidden layers, nodes, learning rate	(fill in)
SVM (RBF)	C, γ	(fill in)
CART	Max depth, min leaf size	(fill in)
RF	Number of trees, features per split	(fill in)
XGBoost	Learning rate, depth, subsample	(fill in)
LightGBM	Learning rate, num leaves, subsample	(fill in)
(others)	(fill in)	(fill in)

Note: Add implementation platform details, e.g., Python scikit-learn version, R caret/mlr3, etc.

Appendix B Individual Dataset Results

Tables A.2 through *[A.K+1]* present raw performance estimates (not ranks) for each classifier on each individual dataset for all six performance measures. These complement the aggregated ranks in Table 4 and are provided for readers who wish to examine dataset-level results.

[Insert raw-results tables here once empirical runs are complete.]

References

[1] Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A **160**(3), 523–541 (1997)

[2] Crook, J.N., Edelman, D.B., Thomas, L.C.: Recent developments in consumer credit risk assessment. European Journal of Operational Research **183**(3), 1447–1465 (2007)

[3] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society **54**(6), 627–635 (2003) https://doi.org/10.1057/palgrave.jors.2601545

[4] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research **247**(1), 124–136 (2015) https://doi.org/10.1016/j.ejor.2015.05.030

[5] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)

[7] Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning **77**, 103–123 (2009) https://doi.org/10.1007/s10994-009-5119-5

[8] García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining. Information Sciences **180**(14), 2044–2064 (2010)

[9] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)

[10] Desai, V.S., Crook, J.N., Overstreet, G.A.: A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research **95**, 24–37 (1996)

[11] West, D.: Neural network credit scoring models. Computers & Operations Research **27**, 1131–1152 (2000)

[12] Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A.: Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications **40**, 5125–5131 (2013)

[13] Wang, G., Ma, J., Huang, L., Xu, K.: Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Systems **26**, 61–68 (2012)

[14] Caruana, R., Munson, A., Niculescu-Mizil, A.: Getting the most out of ensemble selection. In: Proceedings of the 6th International Conference on Data Mining, pp. 828–833. IEEE Computer Society, Hong Kong, China (2006)

[15] Breiman, L.: Bagging predictors. Machine Learning **24**, 123–140 (1996)

[16] Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, pp. 148–156. Morgan Kaufmann, Bari, Italy (1996)

[17] Partalas, I., Tsoumakas, G., Vlahavas, I.: An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Machine Learning **81**, 257–282 (2010)

[18] Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 61–74. Cambridge: MIT Press, ??? (2000)

[19] Thomas, L.C., Edelman, D.B., Crook, J.N.: Credit Scoring and Its Applications. SIAM, Philadelphia (2002)

[20] Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10**(7), 1895–1923 (1998)

[21] García, S., Herrera, F.: An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research **9**, 2677–2694 (2008)

[22] Viaene, S., Dedene, G.: Cost-sensitive learning and decision making revisited. European Journal of Operational Research **166**, 212–220 (2004)