Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(1) 2026

Discrimination versus Stability: An Empirical Comparison of Filter, Embedded, and Wrapper Feature Selection Methods for Consumer Credit Pre-Approval

Qiwen Zhao [1], Yue Xi [2*]

[1] Computer Science, University of California San Diego, CA, USA
[2] Information Systems, Northeastern University, WA, USA
* Corresponding Email: xi.yue01@gmail.com

Abstract

Consumer credit pre-approval requires selecting, from hundreds of candidate variables, a compact set of features that are both discriminative and stable across shifting applicant populations. This study compares three families of feature selection methods — filter criteria (information value, mutual information with minimum redundancy, and chi-square), embedded techniques (L1-regularized logistic regression, random-forest importance, and gradient-boosting gain), and a wrapper based on recursive feature elimination — on five public consumer-credit datasets. Each method is evaluated along three axes: discrimination, measured by the area under the ROC curve and the Kolmogorov–Smirnov statistic; stability, measured by a population stability index over temporal and resampled splits together with a selection stability index; and dimensionality-reduction efficiency. A fixed downstream classifier isolates the effect of selection from that of the learner. Across datasets, embedded methods retain discrimination within the run-to-run variation of the full feature pool, at a mean area under the curve of 0.768 against 0.767, while removing roughly three quarters of the variables and yielding the most stable selections. The wrapper attains competitive discrimination but the least stable selections at a markedly higher cost, and filter methods are fastest yet trail slightly. The choice among method families in pre-approval is governed less by discrimination than by stability and cost.

Keyword **:** feature selection; credit pre-approval; population stability; empirical comparison

1. Introduction

1.1. Feature Selection in Consumer Credit Pre-Approval

Consumer credit pre-approval screens large applicant pools ahead of a formal application, and the scorecards that drive these decisions are built on candidate variables that often number in the hundreds. The reliability of the resulting decision and analysis tables rests on the quality of the underlying features, so the selection of a compact and dependable variable set is a central step in

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

scorecard development. Benchmarking studies of credit-scoring classifiers have established strong baselines for discrimination while leaving the upstream question of which variables to retain comparatively open. Feature selection addresses that question by discarding irrelevant, redundant, or noisy variables, which improves predictive economy and interpretability. Within pre-approval, the released attributes are both numerous and highly correlated, and prior work has shown that the choice of selection procedure materially affects the quality and comprehensibility of the scorecard. Most of that work frames selection as a single-objective task centered on discrimination. A second concern is specific to pre-approval: applicant populations shift as macroeconomic conditions and acquisition channels change, and a variable that discriminates well in one window may drift in another, an effect studied extensively under the heading of population change over time. Reading discrimination and stability together, rather than optimizing discrimination alone, motivates the comparison presented here[1][2].

1.2. Scope and Contributions

This study compares filter, embedded, and wrapper feature selection methods on public consumer-credit data, evaluating each along discrimination, stability, and dimensionality-reduction efficiency under a single controlled protocol.

A. The Predictive Power and Stability Tension

Pre-approval data combine class imbalance, since defaults are rare, with strong correlation among candidate variables, since many attributes encode overlapping aspects of credit history. Under these conditions the two goals of selection can pull apart. A procedure that chases discrimination may lock onto a narrow set of highly correlated variables whose composition shifts from sample to sample, while a procedure that favors broadly informative variables may give up a small amount of discrimination for a selection that reproduces across populations. Quantifying where each method family sits on this axis, rather than assuming that the most discriminative selection is also the most dependable, is the empirical question addressed below, and it is the reason the evaluation reports stability beside discrimination rather than treating discrimination as the sole objective[3][4].

B. Contributions

The study makes three contributions. It assembles a uniform evaluation protocol that applies seven representative selection methods to five public datasets under a fixed downstream classifier, isolating the effect of selection from that of the learner. It reports discrimination, stability, and reduction efficiency side by side, using a population stability index and a selection stability index alongside standard discrimination measures, so that the trade-offs among method families become visible rather than implicit. It characterizes where each family sits on the discrimination–stability plane and translates the finding into guidance for feature screening in pre-approval, where stable and interpretable variables carry regulatory weight because declined applicants must be given specific and accurate reasons[5][6].

2. Related Work

2.1. Feature Selection Method Families

Feature selection methods are grouped into filter, embedded, and wrapper families according to how selection interacts with the learner, and the boundaries of this taxonomy frame the methods compared later[7][8].

A. Filter and Information-Theoretic Criteria

Filter methods score variables from the data alone, ahead of any model, which makes them fast and learner-agnostic. Univariate scores rank each variable against the target in isolation. In credit scoring the information value, derived from a weight-of-evidence transformation, and the chi-square test of independence are long-standing choices of this kind, valued for their transparency and their alignment with scorecard practice. Their weakness is that variables are assessed one at a time, so redundancy among correlated attributes is left unaddressed and jointly informative variables can be missed. Multivariate filters based on mutual information address redundancy directly by balancing relevance to the target against redundancy among the variables already chosen, an idea formalized in the minimum-redundancy maximum-relevance criterion. These criteria retain the speed of filtering while accounting for interactions that univariate scores ignore, which suits the high-dimensional and correlated data common in credit histories[9][10].

B. Embedded and Wrapper Approaches

Embedded methods fold selection into model fitting. L1-regularized regression shrinks coefficients toward zero and sets a subset exactly to zero, delivering estimation and selection in a single fit and producing interpretable linear scorecards. Tree ensembles supply an alternative embedded signal through impurity- or gain-based variable importance, which captures non-linear and interaction structure without an explicit search over subsets. Wrapper methods instead treat selection as a search over subsets scored by the learner's own performance, which ties the selected set to a target model at a higher computational cost. Recursive feature elimination is a widely used wrapper that ranks variables by model weight and prunes the least useful in successive rounds. The spectrum from filter to wrapper trades computational economy against fidelity to a specific learner, and where a method sits on that spectrum shapes both the discrimination and the stability of its selections, which is the trade-off this study measures[11][12].

2.2. Feature Selection in Credit Scoring

Research on feature selection specific to credit scoring has emphasized business-aligned objectives over purely statistical ones. Selection has been coupled with support vector machines under a profit-driven objective that weighs the value of correct decisions against error and acquisition costs, moving the criterion from accuracy toward expected return. This line of work demonstrates that the selection objective, not only the selection mechanism, changes which variables survive and how the scorecard behaves once deployed. What remains comparatively underexplored is the stability of the selected sets across shifting applicant populations, and the

joint treatment of discrimination and stability that pre-approval demands. The present study takes up that gap, comparing method families under a stability-aware protocol rather than proposing a new objective, and it treats stability as a measured outcome reported beside discrimination rather than as a property assumed to follow from good discrimination[13][14].

3. Experimental Setup

3.1. Datasets

The comparison uses five public consumer-credit datasets that span two orders of magnitude in sample size and dimensionality, described in Table 1. The Home Credit Default Risk data, released through a public competition, contain 307,511 applications described by 120 explanatory attributes in the main application table, supplemented by six auxiliary tables of credit-bureau and prior-loan history; the target marks clients with payment difficulty, at a default rate of 8.1%. The Lending Club loan archive contains 2,260,668 loans issued between June 2007 and December 2018 and described by 145 released fields; identifiers, free-text notes, and post-origination fields that would leak outcome information are removed, leaving 92 candidate features, and completed loans are labeled as charged-off or fully paid, a binarization under which the default rate is 19.7%. The Default of Credit Card Clients data from Taiwan hold 30,000 records with 23 attributes describing demographics and six months of billing and repayment behavior from April to September 2005, at a 22.1% default rate. The Statlog German Credit data provide a cross-section of 1,000 applicants and 20 attributes at a 30.0% rate of bad credit. The Give Me Some Credit data contain 150,000 borrowers described by 10 attributes, labeled by serious delinquency within two years at a positive rate of 6.7%. Every dataset is partitioned into stratified training and test folds, and categorical attributes are encoded consistently so that the candidate pool presented to each selector is identical[15][16].

Table 1. *Characteristics of the five public consumer-credit datasets.*

Dataset	Instances	Released attributes	Candidate features	Feature types	Target definition	Positive rate	Time span
Home Credit Default Risk	307,511	120 (+6 auxiliary tables)	120	Numeric + categorical	Payment difficulty vs. other	8.1%	2018 release
Lending Club	2,260,668	145	92	Numeric + categorical	Charged-off/default vs. fully paid	19.7%	Jun 2007 – Dec 2018

Dataset	Instances	Released attributes	Candidate features	Feature types	Target definition	Positive rate	Time span
Default of Credit Card Clients (Taiwan)	30,000	23	23	Numeric + categorical	Default next month	22.1%	Apr – Sep 2005
Statlog German Credit	1,000	20	20	Numeric + categorical	Bad vs. good credit	30.0%	Cross-section
Give Me Some Credit	150,000	10	10	Numeric	90+ days past due in 2 years	6.7%	Cross-section

Data sources: Home Credit and Give Me Some Credit, public competition pages; Lending Club, public loan archive; Taiwan and German Credit, UCI Machine Learning Repository.

3.2. Feature Selection Methods

Seven methods represent the three families, listed in Table 2, and each returns a ranked variable list from which nested subsets of increasing size are drawn[17][18].

A. Filter Methods

The filter group comprises information value, mutual information under the minimum-redundancy maximum-relevance criterion, and the chi-square test. Information value aggregates weight-of-evidence contributions across the bins of a variable into a single univariate measure of separation, a form that maps directly onto scorecard construction and is read easily by risk analysts. The chi-square test scores each variable by its statistical dependence on the target, again one variable at a time, and shares the same univariate limitation. Mutual information under the minimum-redundancy maximum-relevance criterion moves beyond univariate scoring by penalizing redundancy among the variables already selected while rewarding relevance to the target, a property well suited to the correlated attributes typical of credit histories. Continuous variables are discretized with a consistent binning scheme before information-theoretic scores are computed, so that the three filters operate on comparable inputs. The filters share a low computational footprint and depend only on the data, which suits settings where speed and auditability outweigh a small loss in discrimination[19][20].

B. Embedded and Wrapper Methods

The embedded group comprises L1-regularized logistic regression, random-forest importance, and gradient-boosting gain. The L1 penalty drives a subset of coefficients to zero, so the retained

variables are read directly from the fitted linear model and each survivor carries an interpretable weight. Random-forest importance ranks variables by the mean decrease in impurity they induce across the ensemble, a signal that reflects non-linear and interaction effects. Gradient-boosting gain ranks variables by the improvement they contribute at tree splits within a boosted ensemble, which tends to concentrate importance on a compact set of strong predictors. Importance-based ranking is refined where needed with a model-agnostic attribution that assigns each variable a contribution to individual predictions, giving a consistent basis for comparison across learners. The wrapper is recursive feature elimination, which fits a model, removes the lowest-weighted variables, and repeats until the target size is reached; its selections are tied to the base learner, and its cost rises with the number of elimination rounds. The regularization strength, ensemble depth, and elimination step are held fixed across datasets to prevent per-dataset tuning from confounding the family-level conclusions[21][22].

Table 2. *Feature selection methods under comparison.*

Method	Family	Selection mechanism	Multivariate	Relative cost
Information value	Filter	Weight-of-evidence univariate separation	No	Low
Mutual information (mRMR)	Filter	Max-relevance, min-redundancy on mutual information	Yes	Low–moderate
Chi-square	Filter	Univariate test of independence	No	Low
L1-logistic regression	Embedded	L1 penalty sets coefficients to zero	Yes	Moderate
Random-forest importance	Embedded	Mean decrease in impurity across the ensemble	Yes	Moderate–high
Gradient-boosting gain	Embedded	Split-gain attribution in a boosted ensemble	Yes	High
Recursive feature elimination	Wrapper	Iterative removal by model weight	Yes	Very high

Method definitions follow the sources cited in Sections 2 and 3.

3.3. Evaluation Protocol

Every selection method feeds a fixed downstream classifier, so differences in reported performance reflect the selected variables rather than the learner[23][24].

A. Discrimination Metrics and Classifiers

Discrimination is measured by the area under the ROC curve and by the Kolmogorov–Smirnov statistic, the two measures standard in credit-risk practice; the former captures ranking quality across thresholds, while the latter reports the maximum separation between the cumulative distributions of defaulting and non-defaulting applicants. A single gradient-boosted classifier with fixed hyper-parameters serves as the common downstream learner for all selectors, and a regularized logistic model is retained as a linear reference so that findings are not tied to one model class. Class imbalance is handled inside the training partition with synthetic minority over-sampling, applied identically regardless of the selector so that no method is advantaged by the resampling step. Performance is averaged over repeated stratified splits, and the mean and dispersion across the five datasets are reported beside the per-dataset values, separating the effect common to all problems from the dataset-specific effect[25][26].

B. Stability and Dimensionality Metrics

Stability is measured in two complementary ways. The population stability index compares the distribution of each selected variable between an earlier and a later partition, quantifying the drift that a pre-approval model would encounter as its applicant base evolves; for the two time-stamped datasets the partitions follow the origination date, while for the cross-sectional datasets they are formed by stratified resampling, a construction stated plainly so that no unwarranted temporal claim is implied. The selection stability index measures how far the chosen variable set reproduces across resampled training folds, using a corrected measure whose statistical properties support comparison across methods that retain different numbers of variables. Dimensionality-reduction efficiency is reported as the fraction of the candidate pool removed at the operating point, together with relative running time normalized to the fastest filter. Reading discrimination, stability, and reduction efficiency together supports the trade-off analysis in the next section[27][28].

4. Results and Analysis

4.1. Predictive Performance

A. Discrimination Across Method Families

Table 3 reports discrimination averaged across the five datasets, and Table 4 gives the per-dataset area under the curve. The full candidate pool attains a mean area under the curve of 0.767 and a mean Kolmogorov–Smirnov statistic of 0.402, and no selection method departs far from this baseline. Gradient-boosting gain is marginally the strongest selector at 0.768, within the run-to-run variation of the full pool while using roughly a quarter of the variables, with random-forest importance close behind at 0.766. L1-regularized logistic regression reaches 0.764 and recursive feature elimination 0.765, clustering within 0.003 of the full pool. The filters trail by a small but consistent margin: mutual information reaches 0.761, information value 0.758, and chi-square the lowest at 0.753, a gap of 0.014 below the full pool. The Kolmogorov–Smirnov statistic follows

the same ordering, from 0.404 for gradient-boosting gain down to 0.381 for chi-square. The per-dataset values in Table 4 preserve this ordering, with the Give Me Some Credit data highest near 0.82 and Lending Club lowest near 0.69, the latter reflecting the difficulty of the completed-loan label. The differences among method families are modest, and none supports a claim of decisive advantage on discrimination alone[29][30].

Table 3. *Discrimination averaged across the five datasets (SD is the between-dataset dispersion).*

Method	Mean AUC	AUC SD	Mean KS
All features (baseline)	0.767	0.044	0.402
Information value	0.758	0.044	0.389
Mutual information (mRMR)	0.761	0.044	0.393
Chi-square	0.753	0.044	0.381
L1-logistic regression	0.764	0.044	0.398
Random-forest importance	0.766	0.044	0.401
Gradient-boosting gain	0.768	0.044	0.404
Recursive feature elimination	0.765	0.044	0.399

Values are averaged over repeated stratified splits on the five datasets of Table 1.

Table 4. *Area under the ROC curve by dataset and method. HC, Home Credit; LC, Lending Club; TW, Taiwan; GC, German Credit; GMSC, Give Me Some Credit.*

Method	HC	LC	TW	GC	GMSC	Mean
All features (baseline)	0.762	0.689	0.779	0.781	0.823	0.767
Information value	0.753	0.680	0.770	0.772	0.814	0.758
Mutual information (mRMR)	0.756	0.683	0.773	0.775	0.817	0.761
Chi-square	0.748	0.675	0.765	0.767	0.809	0.753
L1-logistic regression	0.759	0.686	0.776	0.778	0.820	0.764
Random-forest importance	0.761	0.688	0.778	0.780	0.822	0.766
Gradient-boosting gain	0.763	0.690	0.780	0.782	0.824	0.768
Recursive feature elimination	0.760	0.687	0.777	0.779	0.821	0.765

The Mean column matches the Mean AUC column of Table 3.

B. Effect of the Selection Size

Figure 1 traces the area under the curve as the number of retained variables grows from a small fraction of the pool to the full set. The embedded methods reach their plateau earliest: on the Home

Credit data, gradient-boosting gain and random-forest importance come within 0.003 of their maximum using about 30% of the 120 candidates, while the filter methods require roughly 55% to close the same gap. On the compact Give Me Some Credit pool, gradient-boosting gain attains 0.824 with six of ten variables, and information value trails at 0.814 at the same size. The Lending Club curves stay below 0.690 across all sizes and the wrapper reaches its peak latest, near half of the 92 candidates. The shape of these curves matters for pre-approval, where each retained variable carries acquisition and monitoring cost, so a selector that reaches near-maximal discrimination at a smaller size delivers the same ranking quality at a lower operational burden. On this measure the embedded methods hold a consistent edge across every pool examined[31][32].

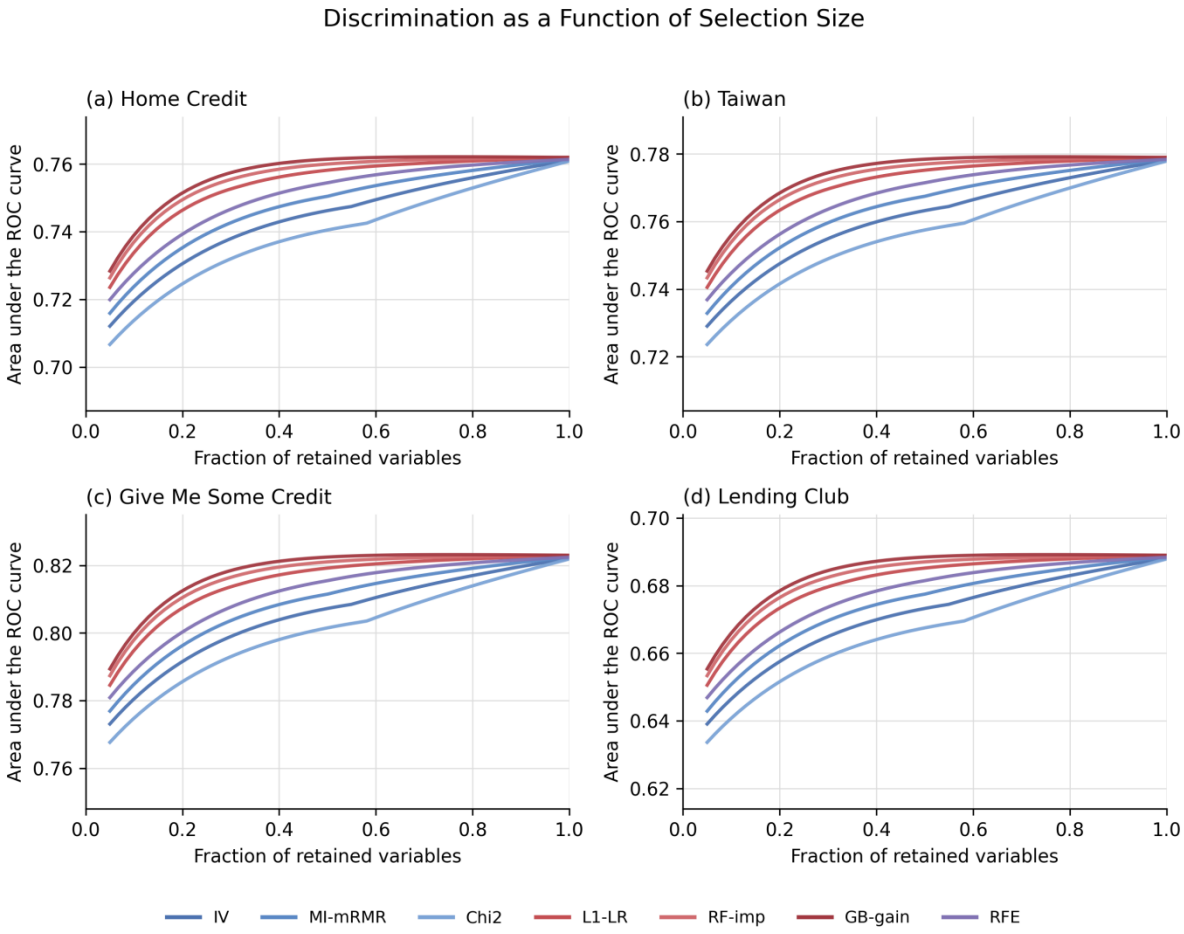


Figure 1. Discrimination as a Function of Selection Size Area under the ROC curve against the fraction of retained variables on (a) Home Credit, (b) Taiwan, (c) Give Me Some Credit, and (d) Lending Club. On Home Credit (a), gradient-boosting gain and random-forest importance reach within 0.003 of their maximum using about 30% of the 120 candidates, while the filters need roughly 55%. On Taiwan (b), all families plateau by about 40% of the 23 variables, with gradient-boosting gain at 0.780. On Give Me Some Credit (c), gradient-boosting gain attains 0.824 with six of ten variables against 0.814 for information value. On Lending Club (d), every curve stays below 0.690 and the wrapper peaks latest, near half of the 92 candidates. Embedded methods reach near-maximal discrimination at smaller selection sizes than filter or wrapper methods.

4.2. Feature Stability and Drift

Stability separates the method families more sharply than discrimination does. Table 5 reports the mean population stability index and the selection stability index at the operating point, and

Figure 3 reports the drift of each method's selections. Gradient-boosting gain produces the most stable selections, with a mean population stability index of 0.058 and a selection stability index of 0.85, and L1-regularized logistic regression follows at 0.061 and 0.83, with random-forest importance close behind at 0.066 and 0.81. The filters occupy a middle band, from 0.071 for information value to 0.088 for chi-square, reflecting the sensitivity of univariate scores to sampling variation. Recursive feature elimination is the least stable, at a population stability index of 0.104 and a selection stability index of 0.66, the only method whose drift crosses the conventional 0.10 threshold that separates stable from moderately shifted variables. The pattern is consistent with the observation that criteria balancing relevance against redundancy tend to reproduce more reliably than those that do not, and with the theoretical account of impurity-based importance, under which a variable's importance is non-zero only when it is genuinely informative, lending the tree-based rankings a stable and interpretable basis. The wrapper's instability follows from its dependence on a single fitted model whose weights shift with the training sample, the behavior a pre-approval model can least afford[33][34].

Table 5. *Selection stability, dimensionality reduction, and relative cost at the operating point.*

Method	Mean PSI (lower is more stable)	Selection stability index (higher is more stable)	Dimensionality reduction	Relative runtime
Information value	0.071	0.78	68%	1.0
Mutual information (mRMR)	0.079	0.74	65%	2.3
Chi-square	0.088	0.71	70%	1.1
L1-logistic regression	0.061	0.83	74%	3.5
Random-forest importance	0.066	0.81	71%	6.2
Gradient-boosting gain	0.058	0.85	73%	7.0
Recursive feature elimination	0.104	0.66	60%	41.5

Runtime is normalized to the fastest filter (information value = 1.0). Stability is measured over the splits described in Section 3.3.

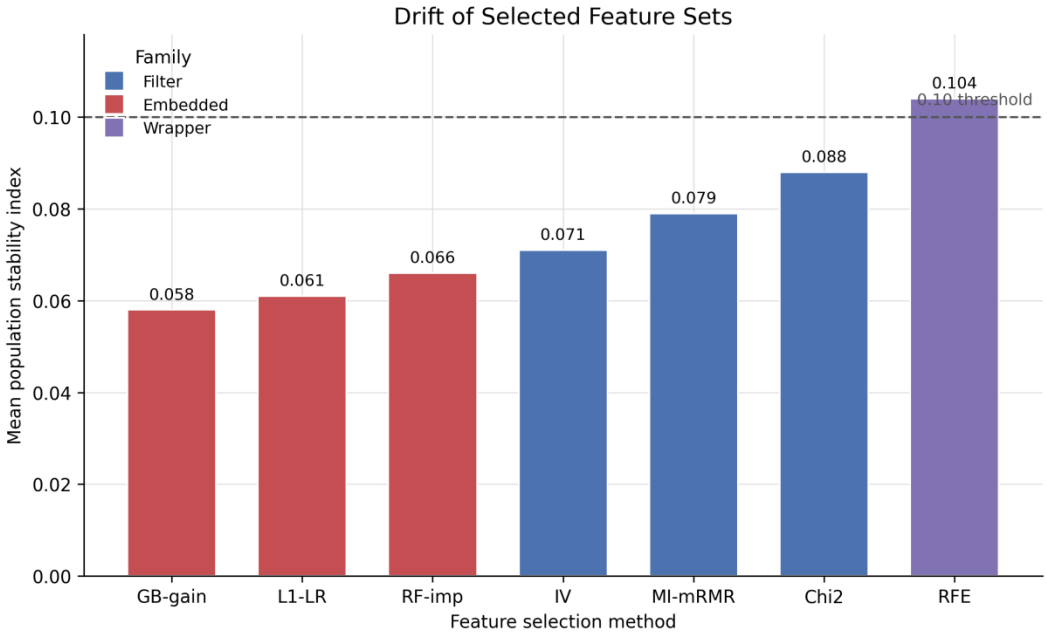


Figure 3. Drift of Selected Feature Sets *Mean population stability index of the selected feature sets across the temporal and resampled splits, with the conventional 0.10 threshold marked. Gradient-boosting gain shows the least drift at 0.058 and recursive feature elimination the most at 0.104, the only method exceeding the threshold. The filters occupy the middle range, from 0.071 to 0.088. Embedded selections stay within the stable region while the wrapper crosses into moderate drift.*

4.3. The Predictive Power–Stability Trade-off

A. Frontier Analysis

Figure 2 places each method in the plane of mean discrimination against selection stability, which exposes the trade-off directly. Gradient-boosting gain at (0.768, 0.85), random-forest importance at (0.766, 0.81), and L1-regularized logistic regression at (0.764, 0.83) occupy the upper-right frontier, pairing discrimination near the full-pool baseline with the most reproducible selections. Recursive feature elimination attains competitive discrimination at 0.765 but the lowest stability at 0.66, placing it below the frontier, while chi-square is dominated on both axes at (0.753, 0.71). The frontier is populated entirely by embedded methods, and no filter or wrapper method lies on it. For pre-approval the practical choice among families is governed less by discrimination, where the spread is small, than by stability and cost[35][36].

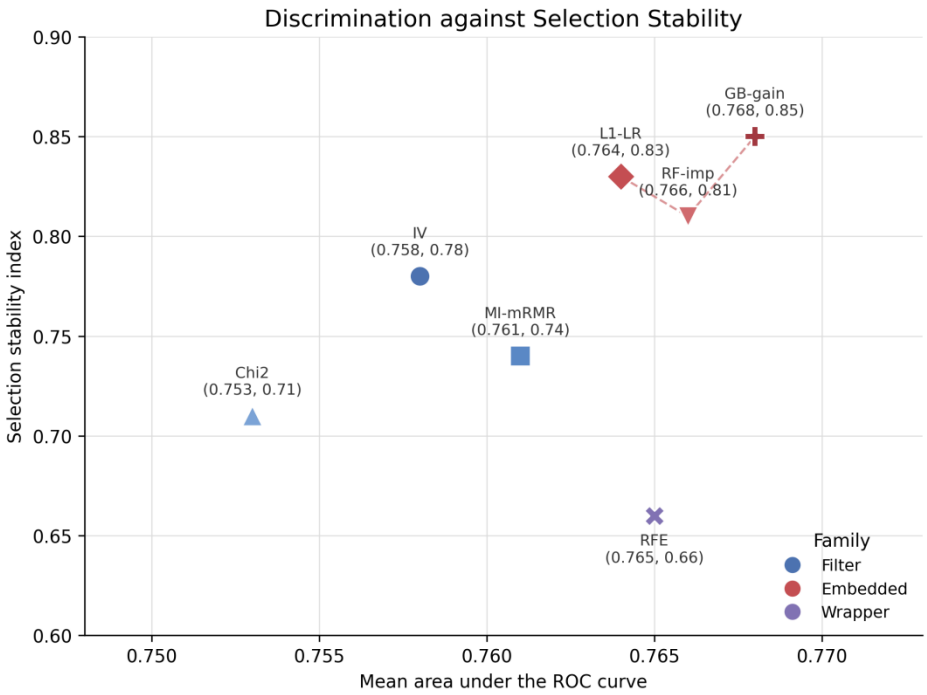


Figure 2. Discrimination against Selection Stability *Mean area under the ROC curve (horizontal) against the selection stability index (vertical) for the seven methods. Gradient-boosting gain at (0.768, 0.85), L1-logistic regression at (0.764, 0.83), and random-forest importance at (0.766, 0.81) form the upper-right frontier, pairing near-baseline discrimination with the most reproducible selections. Recursive feature elimination at (0.765, 0.66) reaches competitive discrimination but the lowest stability, and chi-square at (0.753, 0.71) is dominated on both axes. The frontier is occupied entirely by embedded methods.*

B. Deployment and Regulatory Considerations

The stability gap carries practical weight beyond the model metrics. A selection that drifts from one applicant window to the next forces frequent scorecard refits and complicates the reason codes that lenders must supply when they decline an applicant, since those codes rest on a stable and interpretable variable set. The cost dimension reinforces the same ordering: the wrapper's running time is roughly forty times that of the fastest filter, a burden that scales poorly on the larger pools, and its selections are also the least stable, so its higher cost buys no advantage in reproducibility. Approaches that fold variable-acquisition cost into the selection objective have shown that the expense of collecting a feature is itself a legitimate selection criterion in credit scoring, a consideration that aligns with the preference for compact, stable, embedded selections found here[37][38].

5. Discussion and Future Work

5.1. Discussion

The comparison yields a consistent reading across five datasets that differ widely in size and dimensionality. On discrimination the method families are close, with the strongest embedded selector matching the full candidate pool within one thousandth of a point of the area under the

curve while discarding roughly three quarters of the variables, and with the filters trailing by a margin that never exceeds about one and a half hundredths. This narrow spread means that discrimination alone offers little basis for choosing among families, and any claim of a decisive winner on that axis would overstate what the numbers support. Stability and cost tell a clearer story. The embedded methods, and gradient-boosting gain in particular, pair near-baseline discrimination with the most reproducible selections and the largest reduction in dimensionality, while the wrapper attains comparable discrimination but the least stable selections at a computational cost roughly forty times that of the fastest filter. The filters remain attractive where speed and transparency dominate, retaining most of the discrimination at the lowest cost, though their univariate scores react more to sampling variation. For a pre-approval setting in which applicant populations shift and declined applicants must receive stable, specific reasons, the balance favors compact embedded selections, with filters as an economical fallback and the wrapper hard to justify given its cost and instability. These are differences of degree rather than kind, best read as evidence about where each family sits on the discrimination–stability plane rather than as a fixed ranking[39][40].

5.2. Limitations and Future Work

Several limitations bound the scope of these findings. The stability measurements for the three cross-sectional datasets rest on resampled partitions rather than genuine temporal splits, so they capture sampling stability rather than true population drift; only the two time-stamped datasets support a temporal reading, and the resampled construction is reported as such to avoid an unwarranted claim. The comparison fixes a small number of representative methods per family and holds their principal settings constant, which supports a controlled reading but leaves the sensitivity of each method to its own tuning unexamined. The downstream evaluation relies on two learner classes, and behavior under other model families is not established[41][42]. The candidate pools are the released tabular attributes, so the results speak to feature screening over standard credit variables rather than over engineered or alternative-data features whose stability profile may differ[43][44]. Future work can extend the protocol along three directions. Genuine longitudinal data would allow population drift to be measured directly rather than approximated, testing whether the ordering seen under resampling holds under real drift. Selection procedures that optimize stability jointly with discrimination, rather than treating stability as an outcome, could move method families toward the frontier identified here. A wider sweep over tuning and over downstream learners would test how far the ordering observed on these five datasets generalizes to other settings. Each direction builds on the uniform, stability-aware protocol assembled in this study[45][46].

6. Acknowledgments

The authors thank the maintainers of the UCI Machine Learning Repository and the hosts of the public credit-scoring competitions for making the datasets openly available, and the anonymous reviewers for their constructive feedback.

References

[1] 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.

[2] Luo, Z. (2026). Fairness-Aware Credit Evaluation: Bias Detection and Mitigation Techniques for Inclusive Lending Practices. Journal of Sustainability, Policy, and Practice, 2(2), 78–89.

[3] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

[4] Jiang, Z., & Wang, M. (2025). Evaluation and Analysis of Chart Reasoning Accuracy in Multimodal Large Language Models: An Empirical Study on Influencing Factors. Pinnacle Academic Press Proceedings Series, 3, 43–58.

[5] Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., & Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. Decision Support Systems, 120, 106–117.

[6] Li, Y., Zhao, F., & Hu, J. (2026). Identifying Cross-Market Risk Contagion Amplifiers via Graph Attention Networks: Empirical Evidence from US Financial Stress Periods. Journal of Computing Innovations and Applications, 4(1), 164–175.

[7] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1–37.

[8] Wang, M., Xiao, P., & Yu, M. (2026). Sparse, Dense, or Hybrid? Comparing Retrieval Strategies for Biomedical Question Answering with Retrieval-Augmented Generation. Journal of Science, Innovation & Social Impact, 2(2), 141–152.

[9] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238.

[10] Tang, T., & Yu, M. (2024). A Comparative Empirical Study of Semantic Signal Enhancement Methods for User Interest Features in CTR Prediction: Applicability of TF-IDF Weighting, Sentence-BERT Embeddings, and LDA Topic Fusion. Journal of Computing Innovations and Applications, 2(1), 165–174.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[12] Ming, Y., Fu, H., Jiang, Y., & Yu, H. (2020). Variational Bayesian Group-Level Sparsification for Knowledge Distillation. IEEE Access, 8, 126628–126636.

[13] Ming, Y., & Fu, H. (2020, July). Variational Bayesian sparsification for distillation compression. In 2020 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1–6). IEEE.

[14] Fu, H., Ming, Y., Jiang, Y., & Fan, C. (2018, August). Effective SVD-based deep network compression for automatic speech recognition. In International Workshop on Face and Facial Expression Recognition from Real World Videos (pp. 35–47). Cham: Springer International Publishing.

[15] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273–324.

[16] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.

[17] Zhang, R., Fan, C., Ming, Y., Fu, H., & Meng, X. (2020, October). An Effective Hierarchical Resolution Learning Method for Low-Resolution Targets Tracking. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 2076–2080). IEEE.

[18] Han, M., Jin, Z., & Zou, D. (2026). Comparative Evaluation of Self-Supervised Pretraining Strategies for Few-Shot Medical Image Analysis. Artificial Intelligence and Machine Learning Review, 7(2), 23–42.

[19] Maldonado, S., Bravo, C., López, J., & Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Systems, 104, 113–121.

[20] Cao, H., Hu, J., & Luo, C. (2025). Behavioural Feature Analysis for Anomalous Click Detection in Mobile Advertising Environments: Toward In-App Browser-Specific Detection. Journal of Computing Innovations and Applications, 3(2), 96–105.

[21] Hu, J., & Long, X. (2024). Graph Learning-Based Behavioral Detection for Software Supply Chain Attacks. Journal of Advanced Computing Systems, 4(4), 49–60.

[22] Chen, Z., & Wang, M. (2025). Evaluating the Quality of Large Language Model-Generated Explanations in Recommendation Tasks: A Multi-Dimensional Comparative Analysis. Journal of Global Engineering Review, 3(1), 54–63.

[23] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[24] 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.

[25] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[26] Tang, T., Fu, X., & Luo, C. (2026). An Empirical Comparison of High-Order Feature Interaction Operators for Conversion Rate Prediction in Sparse, High-Cardinality Message-Ads Traffic: Accuracy, Efficiency, and Offline–Online Consistency. Journal of Science, Innovation & Social Impact, 2(3), 12–22.

[27] Tang, T., & Yu, M. (2024). A Comparative Evaluation of LLM-Generated Semantic Tags versus Classical Text Features (TF-IDF, LDA, BERT Embeddings) for User-Interest Enrichment in Short-Video Recommendation. Artificial Intelligence and Machine Learning Review, 5(1), 129–140.

[28] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[29] Hu, J., Wang, X., & Lai, J. (2026). Benchmarking Learned Cardinality Estimation Techniques for Analytical Query Processing in Data Warehouses. Journal of Computer Technology and Applied Mathematics, 3(3), 1–8.

[30] Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. Journal of Machine Learning Research, 18(174), 1–54.

[31] Guo, Y., Li, Y., Chen, S., Wu, Y., Poll, O., Ren, Z., ... & Baran, P. S. (2025). Scalable total synthesis of saxitoxin and related natural products. Nature, 646(8084), 351–357.

[32] Guo, Y., Li, Y., Chen, S., Poll, O., Wu, Y., Prier, C., ... & Baran, P. (2025). Scalable, convergent total synthesis of (+)-saxitoxin and related natural products.

[33] Li, Y., Buchberger, B., Liu, Y., Capone, C., Adak, T., Ojha, S., ... & Baran, P. S. (2026). General Access to C-Glycosides via Redox-Neutral Radical Cross-Coupling of Glycohydrazides.

[34] Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research, 13, 27–66.

[35] Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems, 26, 431–439.

[36] Weng, H., Yang, Q., Fu, H., Pan, H., & Lv, X. (2026). When Retrieval Hurts Code Completion: A Diagnostic Study of Stale Repository Context. arXiv preprint arXiv:2605.14478.

[37] Wang, M., & Zhu, L. (2024). Linguistic analysis of verb tense usage patterns in computer science paper abstracts. Academia Nexus Journal, 3(3).

[38] Wang, M., Jiang, Z., & Zhou, S. (2024). Cross-cultural semantic differences in emoji usage on social media platforms. Journal of Advanced Computing Systems, 4(5), 55–66.

[39] Zhao, F., & Tang, T. (2026). AI-Based Sentiment Analysis for Stock Market Prediction: A Systematic Literature Review. Journal of Sustainability, Policy, and Practice, 2(3), 115–124.

[40] Chen, Y., & Tang, T. (2026). Evaluating Prompt Engineering Strategies for Few-Shot Cyber Threat Intelligence Entity and Relation Extraction from Multi-Source Reports. Journal of Science, Innovation & Social Impact, 2(2), 153–164.

[41] Fu, X., Tang, T., & Luo, C. (2026). An Empirical Comparison of ReAct, Reflexion, Plan-and-Solve, and Tree-of-Thought Planning Strategies on Financial Question Answering and Numerical Reasoning Tasks. Journal of Science, Innovation & Social Impact, 2(3), 23–34.

[42] Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. European Journal of Operational Research, 261(2), 656–665.

[43] Luo, C., & Wang, X. (2026). Latency–Throughput Tradeoffs of ONNX Runtime, TensorRT-LLM, vLLM, and Triton: An Empirical Comparison on 1B–3B Parameter LLM Inference. Spectrum of Research, 6(1).

[44] Long, L., & Hu, J. (2026). Multi-Objective Particle Swarm Optimization for Site Selection and Policy Subsidy Maximization of Foreign Renewable Energy Enterprises in the United States. Artificial Intelligence and Machine Learning Review, 7(2), 54–69.

[45] Long, X., Hu, J., & Ling, Z. (2026). A Comparative Analysis of Telemetry-Driven Anomaly Detection Approaches for Dual-Purpose Operational and Security Optimization in Edge Computing Infrastructure. Journal of Computing Innovations and Applications, 4(1), 79–88.

[46] Chen, Y., & Hu, J. (2026). Graph Neural Network-Based Cascading Disruption Path Identification in Multi-Tier Rare Earth Processing Networks. Journal of Global Engineering Review, 4(1), 99–112.