Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hard and Soft Label Assignment for Cost-Sensitive Semi-Supervised Reject Inference in Credit Scoring

Abderrahim EL AMRANI

aelamrani@insea.ac.ma

Institut National de Statistique et d'Economie Appliquée

Mounir EL ANNAS

Abdelmalek Essaâdi University

Badreddine BENYACOUB

Institut National de Statistique et d'Economie Appliquée

Mohammed EL HAJ TIRARI

Institut National de Statistique et d'Economie Appliquée

Research Article

Keywords: Reject Inference, Semi-supervised learning, Cost-sensitive learning, Classification EM, Logistic regression, Credit scoring

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10168936/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hard and Soft Label Assignment for Cost-Sensitive Semi-Supervised Reject Inference in Credit Scoring

Abderrahim EL AMRANI[1*], Mounir EL ANNAS[2], Badreddine BENYACOUB[3], Mohammed EL HAJ TIRARI[4]

[1,3,4]Research Laboratory in Applied Methods in Statistics, Actuarial Science, Finance and Quantitative Economics, INSEA, Rabat, Morocco.
[2] Department of Mathematics, Faculty of Sciences of Tetouan, Abdelmalek Essaâdi University, Tetouan, Morocco.

*Corresponding author(s). E-mail(s): aelamrani@insea.ac.ma;
Contributing authors: m.elannas@uae.ac.ma; bbenyacoub@insea.ac.ma; tirari@insea.ac.ma;

Abstract

Credit scorecards trained on accepted applicants alone carry a sample selection bias and ignore the asymmetric cost of misclassifying defaulters. We propose RI-CSCEM, a Classification Expectation–Maximisation procedure that fits a cost-sensitive logistic model on both accepted and rejected applicants. Each rejected applicant receives either a hard pseudo-label, assigned by a cost-weighted balance constraint, or a soft responsibility that distributes the applicant across both classes under the same constraint. We prove monotone descent of the objective for the hard variant and consistency of the estimator under a missing-at-random rejection mechanism. On Lending Club data, across thirty stratified draws and twelve competing models, RI-CSCEM matches the best accepted-only baselines on misclassification cost, with significance confirmed by the Friedman–Nemenyi procedure. An ablation study traces this performance to two components: cost-sensitive weighting moves the model off the trivial all-non-default solution, and soft responsibilities recover the discrimination that hard pseudo-labelling destroys, raising AUC from 0.589 to 0.614.

Keywords: Reject Inference, Semi-supervised learning, Cost-sensitive learning, Classification EM, Logistic regression, Credit scoring

1 Introduction

A credit risk scorecard assigns each loan applicant a score predicting repayment, and lenders use that score to decide who receives credit. Early scorecards relied on expert judgement [1]. Statistical models replaced them: linear regression [2, 3], logistic regression [4], and linear discriminant analysis [5, 6]. Machine learning models now handle larger portfolios: support vector machines [7], decision trees [8], neural networks [9], random forests [10], XGBoost [11], and LightGBM [12].

Lenders fit these models on accepted applicants alone, because only accepted applicants have a recorded repayment outcome. Rejected applicants never enter the training data, so the scorecard inherits a sample selection bias [13]: it scores the accepted population well but the wider applicant population poorly. Reject inference infers outcomes for the rejected applicants and folds them into training [14]. The methods split into two groups: one predicts rejected outcomes with a scorecard trained on accepted data; the other reweights the accepted-data likelihood using rejected covariates [14]. Semi-supervised learning belongs to the first group and treats each rejected applicant as an instance with a missing label [15].

These methods can still mislead. They assume good and bad payers occur in similar proportions, but defaulters form a small minority. They also treat the two misclassification errors as equal in cost, so labelling a defaulter as a good payer carries the same penalty as the reverse, understating the lender's loss [16]. Two remedies exist: resample the data to balance the classes [16, 17], or weight the loss by misclassification costs, assigning the larger cost to the minority class [18].

This paper integrates semi-supervised logistic regression [19] with cost-sensitive logistic regression [18] through a Classification Expectation–Maximisation (CEM) procedure [20]. The result is a single penalised likelihood that addresses reject inference and class imbalance together, with formal guarantees of monotone descent and consistency under a missing-at-random (MAR) rejection mechanism. A secondary contribution is diagnostic: MAR ignorability implies that the accepted-only estimator is already unbiased, so the framework provides a controlled setting to measure when reject inference adds value, a question the literature has left open [21].

This paper proposes **RI-CSCEM**, which unifies the following contributions:

1. **Cost-sensitive soft assignment of rejected applicants**: at each E-step, every unlabelled applicant j is scored by the cost-weighted criterion

$$\mathrm{Diff}(j) = C_{FN} \log(1 + e^{-\boldsymbol{\beta}^\top \mathbf{x}_j}) - C_{FP} \log(1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_j}),$$

   which is mapped to a *soft* default responsibility $q_j \in (0, 1)$ calibrated so that the expected default proportion among the rejects equals the empirical default rate r of $\mathcal{D}_l$ (a balance constraint that prevents degenerate all-one or all-zero solutions). Each reject then contributes to both classes, weighted by q_j and $1 - q_j$. A hard-labelling variant (assign the $\lfloor r \cdot m \rfloor$ cheapest-as-default rejects to class one) is recovered as a special case; we show empirically that the soft assignment preserves discrimination that the hard variant destroys.

2. **Unlabelled data scaling** $\lambda \in [0, 1]$: the contribution of $\mathcal{D}_u$ to the objective is scaled by λ, providing a continuous interpolation between purely supervised ($\lambda = 0$) and fully semi-supervised ($\lambda = 1$) regimes. When $\lambda = 0$, the algorithm degenerates to cost-sensitive logistic regression on labelled data alone. An optional ℓ_1 penalty μ provides embedded feature selection and sparsity.
3. **Theoretical guarantees**: we prove that the alternating E-/M-step procedure produces a monotonically non-increasing objective and converges, and that the resulting estimator is consistent under a missing-at-random (MAR) rejection mechanism, which holds by construction for scorecard-based acceptance policies.
4. **Rigorous benchmarking with statistical testing**: we compare RI-CSCEM against twelve competitors (cost-sensitive supervised classifiers, resampling strategies, and classical and semi-supervised reject-inference baselines) over thirty stratified replications under a missing-at-random rejection mechanism, and assess significance with the Friedman test and the post-hoc Nemenyi analysis. Under MAR, RI-CSCEM matches the strongest accepted-only baselines on economic cost, consistent with the theory that reject inference adds little when the rejection mechanism is ignorable. An ablation isolates the contribution of each component, with the soft-vs-hard AUC recovery ($0.589 \to 0.614$) as the main empirical finding.

Section 2 reviews reject inference, semi-supervised credit scoring, and imbalanced classification. Section 3 develops the logistic CEM algorithm, integrates misclassification costs, and proves convergence and consistency. Section 4 describes the data, the experimental design, and the numerical results. Section 5 concludes and proposes future directions.

2 Literature Review

Credit risk scoring is a binary classification problem: class 0 groups clients who repay on time, class 1 those who default [22]. Assessment methods fall into three families: qualitative scoring [23], statistical learning [24], and machine learning [25].

All three families are supervised and require fully labelled data. The model should train on the full applicant population, but banks can only finance a subset, so they approve selected applications and reject the rest. This selective financing introduces sample selection bias [21]. A uniform acceptance rule would eliminate the bias but would ignore risk-profile differences and reduce allocation efficiency.

Feelders [26] framed this selection bias as a missing-data problem and distinguished three mechanisms. Under *missing completely at random* (MCAR) [27], the reject decision is independent of all applicant characteristics; acceptance probability is uniform, so no bias arises, but no bank operates this way. Under *missing at random* (MAR) [27], the reject decision depends on observed characteristics but not on the (future) repayment outcome; this creates selection bias that is ignorable given the covariates. Under *missing not at random* (MNAR) [27], the reject decision depends on both observed and unobserved factors; most financial institutions operate closer to this regime.

In practice, most rejected applications follow a MAR mechanism, because lending decisions rest on observed applicant characteristics and institution-specific thresholds. Feelders [26] introduced an EM framework for reject inference under this assumption.

Anderson and Hardin [28] tested an EM-based logistic regression on real consumer lending data. Li et al. [29] proposed a semi-supervised SVM for credit scoring that outperformed several supervised baselines. Mancisidor et al. [30] developed two Bayesian reject-inference formulations and showed improved results over classical benchmarks. Hájek and Olej [31] built a kernel-based framework combining supervised and semi-supervised learning for corporate and municipal credit scoring. Xia et al. [32] integrated pessimistic likelihood estimation with gradient boosting decision trees and confirmed the benefit of sample culling. El Annas et al. [33] used hidden Markov models for semi-supervised P2P reject inference. Kennedy et al. [34] showed that a semi-supervised one-class classifier outperformed the standard two-class model. Liu et al. [35] combined multiple classifiers with clustering and a dynamic selection mechanism. Anderson [36] used Bayesian networks for reject inference. Xiao et al. [37] proposed a GMDH-based semi-supervised feature selection method. Kang et al. [38] proposed a graph-based semi-supervised reject inference model and noted that the distribution gap between accepted and rejected samples can bias standard semi-supervised learning. These models overlook class imbalance. Falangis [39] reports that high-risk customers represent only about 10% of the population. Standard classifiers therefore misclassify this minority at higher rates, and in credit scoring the cost of missing a defaulter far exceeds the cost of wrongly flagging a repayer [40].

Imbalanced classification techniques fall into three families: data-level resampling, algorithm-level cost weighting, and ensemble strategies.

Baesens et al. compared seventeen classifiers on eight credit scoring datasets and found that simple linear models (logistic regression, LDA) matched more complex ones (SVMs, neural networks) [41]. The literature disagrees on which family wins: Yobas et al. found LDA beat neural networks [42], while Desai et al. found the opposite [43]. On class imbalance, Weiss and Provost showed that natural class distributions rarely produce optimal classifiers; training distributions with 50–90% minority examples performed better [44]. Chawla et al. proposed SMOTE, which generates synthetic minority samples and improves classification [17]. Batista evaluated ten resampling techniques across thirteen datasets and found oversampling outperforms undersampling, with SMOTE combined with Tomek links or ENN performing best [45]. Brown and Mues showed that random forests [46] and gradient boosting [47, 48] resist class imbalance well, using the statistical comparison framework of Demšar [49].

3 Methodology

3.1 Notation and Problem Setup

Let $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ where $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ comprises n accepted (labelled) applicants and $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$ comprises m rejected (unlabelled) applicants. Binary labels $y \in \{0, 1\}$ denote non-default and default respectively. We use the convention that $y = 1$ is the *minority positive class* (default).

Logistic model.

Each applicant is described by a feature vector $\mathbf{x} \in \mathbb{R}^p$ (with a leading constant absorbing the intercept). The logistic model is parameterised by $\boldsymbol{\beta} \in \mathbb{R}^p$:

$$\pi(\mathbf{x}; \boldsymbol{\beta}) = \sigma\left(\boldsymbol{\beta}^\top \mathbf{x}\right), \quad \sigma(t) = (1 + e^{-t})^{-1}.$$

For compactness, write $z_i = \boldsymbol{\beta}^\top \mathbf{x}_i$.

Logistic loss.

For a binary label $y \in \{0, 1\}$ and score z, the logistic loss is:

$$\ell(z, y) = \begin{cases} \log(1 + e^{-z}) & \text{if } y = 1, \\ \log(1 + e^{z}) & \text{if } y = 0. \end{cases}$$

3.1.1 Cost-Sensitive Weights

Let $C_{FN} > 0$ and $C_{FP} > 0$ denote the costs of a false negative (predicted non-default, true default) and a false positive (predicted default, true non-default), respectively. In credit scoring, $C_{FN} \gg C_{FP}$ reflects the asymmetric economic consequences of missed defaults. Define instance-level weights:

$$w_i = \begin{cases} C_{FN} & \text{if } y_i = 1, \\ C_{FP} & \text{if } y_i = 0. \end{cases}$$

The class-imbalance ratio is $\gamma = C_{FN}/C_{FP}$.

3.1.2 The Diff Criterion for Hard Label Assignment

For unlabelled applicant j with current parameter estimate $\boldsymbol{\beta}^{(t)}$, define the *cost difference*:

$$\text{Diff}(j; \boldsymbol{\beta}) = C_{FN}\, \ell(z_j, 1) - C_{FP}\, \ell(z_j, 0) = C_{FN} \log(1 + e^{-z_j}) - C_{FP} \log(1 + e^{z_j}), \quad (1)$$

where $z_j = \boldsymbol{\beta}^\top \mathbf{x}_j$. A negative $\text{Diff}(j)$ means that assigning $\tilde{y}_j = 1$ incurs *lower* cost-weighted logistic loss than $\tilde{y}_j = 0$; a positive value indicates the opposite.

Balance constraint.

Let $r = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[y_i = 1]$ be the empirical default proportion in $\mathcal{D}_l$. We enforce the balance constraint:

$$\frac{1}{m} \sum_{j=n+1}^{n+m} \mathbf{1}[\tilde{y}_j = 1] = r,$$

so that the pseudo-label distribution mirrors the observed class distribution, preventing degenerate solutions in which all unlabelled instances are assigned to one class.

Label assignment rule.

Given $\boldsymbol{\beta}^{(t)}$, sort the unlabelled indices by $\mathrm{Diff}(j;\boldsymbol{\beta}^{(t)})$ in ascending order. Assign $\tilde{y}_j = 1$ to the first $\lfloor r \cdot m \rfloor$ instances (those for which labelling as default is cheapest), and $\tilde{y}_j = 0$ to the remainder. Formally:

$$\tilde{\mathbf{Y}}_u^{(t+1)} = \arg\min_{\tilde{\mathbf{Y}}_u \in \{0,1\}^m} \sum_j \left[\tilde{y}_j \, C_{FN} \, \ell(z_j^{(t)}, 1) + (1 - \tilde{y}_j) \, C_{FP} \, \ell(z_j^{(t)}, 0)\right]$$
$$\text{s.t.} \quad \tfrac{1}{m} \sum_j \tilde{y}_j = r. \quad (2)$$

This is a 0-1 knapsack with equal item sizes; the greedy solution via sorting is optimal.

3.1.3 Soft Responsibility Assignment

Hard labelling commits each reject to a single class, and an early, poorly-estimated $\boldsymbol{\beta}^{(t)}$ can therefore inject systematically wrong labels into the boundary, degrading the discrimination of the final model. We instead assign each reject a *soft* default responsibility $q_j \in (0, 1)$ derived from the same cost-weighted criterion. Writing $s_j = -\mathrm{Diff}(j;\boldsymbol{\beta}^{(t)})$ (so that a larger s_j favours the default hypothesis), we set

$$q_j = \sigma\left(\frac{s_j}{\tau \, \hat{\sigma}_s} + b\right), \quad (3)$$

where $\hat{\sigma}_s$ is the standard deviation of $\{s_j\}$, $\tau > 0$ is a temperature, and the bias b is chosen, by a one-dimensional root-find, so that the soft analogue of the balance constraint holds in expectation,

$$\frac{1}{m} \sum_{j=n+1}^{n+m} q_j = r.$$

Each reject then contributes to *both* classes, with weights proportional to q_j (default) and $1 - q_j$ (non-default). The hard rule (2) is the $\tau \to 0$ limit of (3); intermediate τ trades commitment against the preservation of ranking information, and is selected by cross-validation.

Limitation of the balance constraint.

Forcing the expected default rate among rejects to equal the observed rate r among accepted applicants is a deliberate, conservative modelling choice: it prevents degenerate solutions but implicitly assumes that rejects default at the same rate as accepted applicants. Under selection bias the true reject default rate may be higher, and in that case the constraint suppresses the signal that reject inference could provide. This is a plausible reason why RI-CSCEM does not outperform the accepted-only baselines under the MAR mechanism studied here, and relaxing the constraint is a natural direction for MNAR extensions.

3.1.4 The RI-CSCEM Objective

Given responsibilities $\mathbf{q} = \{q_j\}$, the joint loss over labelled and unlabelled data is:

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{q}) = \underbrace{\textstyle\sum_{i=1}^{n} w_i\, \ell(z_i, y_i)}_{\text{labelled}} + \lambda \underbrace{\textstyle\sum_j \left[q_j C_{FN} \ell(z_j, 1) + (1-q_j) C_{FP} \ell(z_j, 0) \right]}_{\text{unlabelled, soft-assigned}} + \mu \|\boldsymbol{\beta}\|_1, \tag{4}$$

subject to the (expected) balance constraint $\frac{1}{m}\sum_j q_j = r$. The hard objective is the special case $q_j \in \{0, 1\}$. Two hyperparameters govern the regularisation:

- $\lambda \in [0, 1]$: **unlabelled data weight**, the relative importance of the pseudo-labelled unlabelled set. Setting $\lambda = 0$ yields cost-sensitive supervised logistic regression (CSLR); setting $\lambda = 1$ treats unlabelled data on equal footing with labelled. Intermediate values are tuned by cross-validation.
- $\mu > 0$: ℓ_1 **regularisation penalty**, controlling the sparsity of $\boldsymbol{\beta}$ independently of how much unlabelled data is incorporated.

The full optimisation problem is:

$$\min_{\boldsymbol{\beta},\, \tilde{\mathbf{Y}}_u} \mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u) \quad \text{s.t.} \quad \frac{1}{m}\sum_j \tilde{y}_j = r, \;\; \tilde{y}_j \in \{0, 1\}. \tag{5}$$

Special cases.

- $\lambda = 0$: reduces to cost-sensitive ℓ_1-penalised logistic regression on $\mathcal{D}_l$ alone.
- $\lambda > 0$, $\mu = 0$, equal costs: recovers a standard semi-supervised CEM logistic regression in the spirit of [50], but here cast in the reject-inference setting with the balance constraint; RI-CSCEM departs from it through the cost-weighted Diff criterion and the MAR consistency analysis.
- $\lambda > 0$, $C_{FN} = C_{FP}$: reduces to a balanced semi-supervised ℓ_1-logistic regression.

Relation to prior cost-sensitive semi-supervised work.

Zhu and Wan [50] combined cost-sensitive learning with semi-supervised classification in a hit-and-run analysis setting, but without the reject-inference framing, the CEM E-step, or the MAR consistency analysis. RI-CSCEM's novelty lies at the intersection of these three elements: cost-sensitivity, CEM-based reject inference, and a formal consistency guarantee tied to the MAR ignorability condition. The credit-scoring literature lacks a clean statement of when reject inference pays off. RI-CSCEM, through its consistency result and controlled MAR experiments, provides a diagnostic baseline for that question.

3.2 MAR Assumption and Consistency

The rejection indicator $R \in \{0, 1\}$ satisfies

$$P(R = 1 \mid \mathbf{x}, y) = P(R = 1 \mid \mathbf{x}),$$

i.e. conditional on observed features, rejection is independent of the unobserved default outcome.

Assumption 3.2 is the standard ignorability condition of [51]. It holds by construction in any scorecard that assigns acceptance/rejection decisions based solely on observed application variables, without direct reference to the (future) default event.

Proposition 1 (Consistency under MAR) *Assume, in addition to 3.2:*

(R1) Compact parameter space: $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^p$, *with* $\mathcal{B}$ *compact.*
(R2) Identifiability: *the population risk* $\mathbb{E}[\ell(z, y)]$ *has a unique minimiser* $\boldsymbol{\beta}^*$ *in* $\mathcal{B}$.
(R3) Correct specification: $P(y = 1 \mid \mathbf{x}) = \sigma({\boldsymbol{\beta}^*}^\top \mathbf{x})$.
(R4) Finite Fisher information: $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \pi(\mathbf{x})(1 - \pi(\mathbf{x}))]$ *is positive definite.*

Let $\hat{\boldsymbol{\beta}}$ *be the RI-CSCEM estimator obtained from objective (4) with* $\lambda \in]0, 1]$ *and* $\mu = \mu_n = O(n^{-1/2})$. *Then* $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$ *as* $n, m \to \infty$.

Proof Under Assumption 3.2, the complete-data likelihood factorises as

$$p(\mathbf{x}, y, R; \boldsymbol{\beta}) = p(y \mid \mathbf{x}; \boldsymbol{\beta})\, p(\mathbf{x})\, p(R \mid \mathbf{x}).$$

The selection mechanism $p(R \mid \mathbf{x})$ is *ignorable* in the sense of [51]: it does not depend on $\boldsymbol{\beta}$ and therefore factors out of the likelihood maximisation. Consequently, maximising the observed-data likelihood $\prod_{i:R_i=0} p(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$ (over accepted applicants with $R_i = 0$) already yields a consistent estimator of $\boldsymbol{\beta}^*$. That is, consistency holds for the accepted-only estimator as well, and reject inference does not improve the asymptotic result under MAR.

The cost weights $w_i \in \{C_{FN}, C_{FP}\}$ are fixed, user-specified constants. They scale the score equations but do not introduce bias: the weighted score equations $\sum_i w_i(y_i - \pi_i)\mathbf{x}_i = \mathbf{0}$ have the same root as the unweighted equations under correct specification (R3). More precisely, for any fixed $w_i > 0$ the weighted empirical risk converges uniformly to its population counterpart by the uniform law of large numbers over $\mathcal{B}$ (R1), and the population minimiser is $\boldsymbol{\beta}^*$ (R2).

It remains to show that the pseudo-labels converge to the true labels. As the sample size grows, $\hat{\boldsymbol{\beta}}^{(t)}$ converges to a neighbourhood of $\boldsymbol{\beta}^*$, and the Bayes-optimal decision boundary under correct specification (R3) and finite Fisher information (R4) assigns each applicant to the correct class with probability tending to one. Consequently, the fraction of pseudo-label errors $m^{-1} \sum_j \mathbf{1}[\tilde{y}_j \neq y_j] \to 0$ in probability, and the unlabelled term in (4) converges to the labelled loss evaluated on $\mathcal{D}_u$. The combined objective then equals the full cost-sensitive logistic loss on $\mathcal{D}_l \cup \mathcal{D}_u$.

The ℓ_1 penalty introduces shrinkage bias of order $O(\mu) = O(n^{-1/2})$, which vanishes asymptotically. Consistency of the penalised estimator then follows from the standard argument for ℓ_1-penalised M-estimators (see, e.g., [52]): the penalised population minimiser equals $\boldsymbol{\beta}^*$ as $\mu \to 0$, and the empirical risk converges uniformly over $\mathcal{B}$. □

3.2.1 Monotone Non-Increase of the Loss

We now establish that the joint minimisation of $\mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u)$ via alternating optimisation produces a non-increasing sequence.

Fix $\boldsymbol{\beta}^{(t)}$. Let $\tilde{\mathbf{Y}}_u^{(t+1)}$ be the hard label assignment in (2). Then:

$$\mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t)}).$$

Proof The labelled term and the ℓ_1 term are independent of $\tilde{\mathbf{Y}}_u$, so it suffices to show that the unlabelled term does not increase. The E-step solves

$$\tilde{\mathbf{Y}}_u^{(t+1)} = \arg\min_{\tilde{\mathbf{Y}}_u} \mathcal{U}(\tilde{\mathbf{Y}}_u; \boldsymbol{\beta}^{(t)}), \quad \mathcal{U}(\tilde{\mathbf{Y}}_u; \boldsymbol{\beta}) := \sum_j \left[\tilde{y}_j C_{FN} \ell(z_j, 1) + (1 - \tilde{y}_j) C_{FP} \ell(z_j, 0)\right],$$

subject to $\frac{1}{m}\sum_j \tilde{y}_j = r$. By definition of the minimiser, $\mathcal{U}(\tilde{\mathbf{Y}}_u^{(t+1)}; \boldsymbol{\beta}^{(t)}) \leq \mathcal{U}(\tilde{\mathbf{Y}}_u^{(t)}; \boldsymbol{\beta}^{(t)})$. The inequality for $\mathcal{L}$ follows immediately because the remaining terms are unchanged. □ □

Fix $\tilde{\mathbf{Y}}_u^{(t+1)}$. Let $\boldsymbol{\beta}^{(t+1)}$ be the proximal gradient update minimising $\mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u^{(t+1)})$ over $\boldsymbol{\beta}$. Then, provided $\boldsymbol{\beta}^{(t+1)} \neq \boldsymbol{\beta}^{(t)}$:

$$\mathcal{L}(\boldsymbol{\beta}^{(t+1)}, \tilde{\mathbf{Y}}_u^{(t+1)}) < \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)}).$$

Proof With $\tilde{\mathbf{Y}}_u^{(t+1)}$ fixed, the objective $\mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u^{(t+1)})$ is a sum of weighted logistic losses (each being convex and smooth in $\boldsymbol{\beta}$) plus the ℓ_1 penalty $\mu\|\boldsymbol{\beta}\|_1$. This is a proper, closed, convex function in $\boldsymbol{\beta}$.

The M-step applies a proximal gradient step:

$$\boldsymbol{\beta}^{(t+1)} = \mathrm{prox}_{\alpha\mu}\Big(\boldsymbol{\beta}^{(t)} - \alpha\, \nabla_{\boldsymbol{\beta}} \mathcal{L}_{\mathrm{smooth}}(\boldsymbol{\beta}^{(t)})\Big),$$

where $\mathcal{L}_{\mathrm{smooth}}$ denotes the smooth (logistic loss) part of (4) and $\alpha > 0$ is a step size chosen to satisfy the descent Lemma (i.e. $\alpha \leq 1/L$ with L the Lipschitz constant of $\nabla\mathcal{L}_{\mathrm{smooth}}$).

For a convex function with L-Lipschitz gradient, the proximal gradient step guarantees the sufficient decrease condition:

$$\mathcal{L}(\boldsymbol{\beta}^{(t+1)}, \tilde{\mathbf{Y}}_u^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)}) - \frac{1}{2\alpha}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2.$$

If $\boldsymbol{\beta}^{(t+1)} \neq \boldsymbol{\beta}^{(t)}$, the right-hand side is strictly less than $\mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)})$. □ □

Theorem 2 (Monotone Non-Increase) *Let $\{\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t)}\}_{t\geq 0}$ be the sequence generated by Algorithm 1. Then for all $t \geq 0$:*

$$\mathcal{L}(\boldsymbol{\beta}^{(t+1)}, \tilde{\mathbf{Y}}_u^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t)}).$$

Consequently, the loss sequence $\{\mathcal{L}^{(t)}\}$ is monotonically non-increasing and, since $\mathcal{L} \geq 0$, it converges.

Proof The second inequality is Lemma 3.2.1 (E-step does not increase the loss). The first inequality follows from Lemma 3.2.1 (M-step does not increase the loss when $\boldsymbol{\beta}$ is updated with a valid proximal gradient step). Together:

$$\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t+1/2)} \leq \mathcal{L}^{(t)},$$

where $\mathcal{L}^{(t+1/2)} := \mathcal{L}(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{Y}}_u^{(t+1)})$ denotes the loss after the E-step but before the M-step. The sequence $\{\mathcal{L}^{(t)}\}$ is non-increasing and bounded below by 0; by the monotone convergence theorem it converges to a finite limit. This mirrors the classical monotonicity argument for EM [53], here adapted to the constrained hard-assignment classification step.

□

Every accumulation point of $\{\boldsymbol{\beta}^{(t)}\}$ is a stationary point of $\mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u(\boldsymbol{\beta}))$, where $\tilde{\mathbf{Y}}_u(\boldsymbol{\beta})$ denotes the optimal hard assignment for a given $\boldsymbol{\beta}$.

Proof At a fixed point $(\boldsymbol{\beta}^*, \tilde{\mathbf{Y}}_u^*)$, neither the E-step nor the M-step can decrease the loss further, so

$$\tilde{\mathbf{Y}}_u^* = \tilde{\mathbf{Y}}_u(\boldsymbol{\beta}^*) \quad \text{and} \quad \mathbf{0} \in \partial_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}^*, \tilde{\mathbf{Y}}_u^*),$$

where $\partial_{\boldsymbol{\beta}}$ denotes the subdifferential (accounting for the ℓ_1 penalty). This is precisely the stationarity condition. □

Remark 1 (Scope of the descent guarantee) The monotone-descent guarantee of Theorem 2 covers only the *hard* E-step, whose constrained assignment (2) is an exact minimiser of the unlabelled term. The recommended *soft* variant (Section 3.1.3) replaces this exact minimisation with the expectation-style responsibilities (3); the M-step still decreases $\mathcal{L}(\cdot, \mathbf{q}^{(t+1)})$ for fixed $\mathbf{q}^{(t+1)}$, but the soft E-step is not an exact minimiser, so the overall monotone-descent guarantee does *not* transfer to the soft variant. We adopt the soft variant because it yields substantially higher AUC in practice (Section 4) and retain the hard variant, for which the full guarantee holds, as an ablation baseline.

3.2.2 RI-CSCEM Algorithm

Algorithm 1 summarises RI-CSCEM. The initialisation fits cost-sensitive ℓ_1-logistic regression on the labelled data only, providing a warm start that reflects the known class costs before any unlabelled information is incorporated.

Algorithm 1 RI-CSCEM for Reject Inference under MAR

Require: Labelled data $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, unlabelled data $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$, misclassification costs $C_{FN} \geq C_{FP} > 0$, unlabelled weight $\lambda \in [0, 1]$, ℓ_1 penalty $\mu > 0$, convergence tolerance $\varepsilon > 0$, maximum iterations T

Ensure: Parameter vector $\hat{\boldsymbol{\beta}}$

Standardise: centre and scale $\mathbf{X}_l$ and $\mathbf{X}_u$ jointly.

Initialise: Fit cost-sensitive ℓ_1-logistic regression on $\mathcal{D}_l$ with weights $\{w_i\}$ and penalty μ; set $\boldsymbol{\beta}^{(0)} \leftarrow \hat{\boldsymbol{\beta}}_{\text{init}}$.

Compute balance ratio: $r \leftarrow \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[y_i = 1]$

for $t = 0, 1, \ldots, T-1$ **do**

  E-step (Diff-based hard label assignment):

  for each $j \in \{n+1, \ldots, n+m\}$ **do**

    $z_j \leftarrow {\boldsymbol{\beta}^{(t)}}^{\top} \mathbf{x}_j$

    $\text{Diff}(j) \leftarrow C_{FN} \log(1 + e^{-z_j}) - C_{FP} \log(1 + e^{z_j})$

  end for

  Sort $\{j\}$ by $\text{Diff}(j)$ ascending.

  Assign $\tilde{y}_j \leftarrow 1$ to the first $\lfloor r \cdot m \rfloor$ indices; $\tilde{y}_j \leftarrow 0$ to the remainder.

  M-step (cost-sensitive ℓ_1-logistic regression):

  Construct augmented dataset:

$$\begin{aligned}
&\text{Labelled positives: weight } C_{FN} \\
&\text{Labelled negatives: weight } C_{FP} \\
&\text{Pseudo-positives } (\tilde{y}_j = 1)\text{: weight } \lambda \cdot C_{FN} \\
&\text{Pseudo-negatives } (\tilde{y}_j = 0)\text{: weight } \lambda \cdot C_{FP}
\end{aligned}$$

  Solve: $\boldsymbol{\beta}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \tilde{\mathbf{Y}}_u^{(t+1)})$ via proximal gradient (soft-thresholding for ℓ_1).

  if $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 < \varepsilon$ **then**

    break

  end if

end for

return $\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(t+1)}$

4 Numerical Results

4.1 Data and Variables

The data come from Lending Club loan applications filed between 2007 and 2018. After removing records with missing values or data-entry errors, 1,266,782 accepted (labelled) and 2,064,314 rejected (unlabelled) applications remain. Following [29] and [38], we draw thirty independent stratified samples preserving class proportions. Each draw yields 3,500 accepted training samples, 25,000 rejected training samples, 1,000 accepted validation samples, and 1,500 accepted test samples. The accepted and

rejected training samples form the combined training set for all reject-inference methods; the validation set tunes hyperparameters; the test set serves final evaluation only.

The data share eight raw predictors – loan amount, FICO score, debt-to-income ratio, employment length, a three-level grouping of the applicant's state, and loan purpose – and the accepted samples carry an additional binary outcome indicating whether the loan was repaid (0) or charged off (1). Because purpose is nominal, it is one-hot encoded (employment length, which is ordinal, is kept as a single ordered code), yielding $p = 21$ model features. The default rate among accepted applicants is 19.53%. The marginal distributions of accepted and rejected samples differ across most predictors [35, 54], motivating the distribution-adaptation step in the proposed model.

4.2 Experimental Design

We evaluate RI-CSCEM with two experiments. Experiment I benchmarks it against the full set of competing models, and Experiment II ablates its components. Both run under the same missing-at-random protocol and the same thirty stratified draws.

Experiment I: Baseline Benchmark Comparisons.

This experiment compares RI-CSCEM against existing supervised, resampling, and semi-supervised models. Following the comparative framework of Kang[38], we benchmark four groups of models trained on the same training set, and make every supervised model cost-sensitive so that the comparison is fair.

1. *Cost-sensitive supervised models (accepted samples only).* Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), XGBoost, LightGBM, and CatBoost are trained on the labelled accepted samples with class weights set by the misclassification costs. They discard the rejected population and so carry a sample selection bias [55].
2. *Resampling strategies (accepted samples only).* SMOTE [17] rebalances the accepted training set before fitting logistic regression (SMOTE+LR) and a random forest (SMOTE+RF).
3. *Traditional reject-inference models.* The *Augmentation* and *Parceling* methods assign pseudo-labels or weights to rejected samples before retraining a base classifier [21, 27]. The graph-based reject-inference model RI-LS of [38] represents the recent label-propagation family.
4. *Semi-supervised baselines.* The semi-supervised support vector machine (S3VM) of [29] and the base CEM logistic model (linear, equal-cost, no regularization) use both accepted and rejected samples but account for neither distributional shift nor cost asymmetry.

The full RI-CSCEM (cost-sensitive soft responsibilities, λ-weighted unlabelled data, and ℓ_1 regularization) is trained in the same setting and compared with every baseline.

The test set consists of the 1,500 accepted samples held out during sampling. Because the accepted test set is the same across all models, cost comparisons are

meaningful: the average cost of misclassifying a defaulter as a repayer is 5.3527 times that of misclassifying a repayer as a defaulter, reflecting the real loss of principal and interest versus the foregone loan-interest income [56, 57].

Experiment II: Ablation Study.

This experiment isolates the contribution of each architectural component of RI-CSCEM. Two single-component ablation variants are constructed from the base CEM logistic model:

1. **CEM + L1**: base semi-supervised CEM logistic augmented with ℓ_1 regularization only (equal costs).
2. **CEM + CS**: base semi-supervised CEM logistic augmented with cost-sensitive instance weighting and the cost-weighted Diff criterion only (no ℓ_1 regularization). The cost weights follow the asymmetric misclassification-cost framework in which defaulter misclassification incurs a higher penalty than non-defaulter misclassification [58, 59].

Each variant is compared with (a) the base CEM logistic (linear, equal-cost, no regularization), (b) the two single-component ablations, and (c) the full RI-CSCEM (both components active simultaneously). Improvement in F_1, G-mean, AUC, and total cost over the base model quantifies each component's independent contribution; the difference between the sum of individual gains and the combined gain quantifies interaction effects between components (reported in Section 4.4).

All hyperparameters of the CEM family (λ, μ, and the soft temperature τ) are selected on the held-out validation subset; the cost ratio is fixed at its empirical value $C_{FN}/C_{FP} = 5.3527$. The whole protocol is repeated over thirty independent stratified draws, so that every reported figure is a mean over thirty observations and amenable to rank-based significance testing.

4.3 Evaluation Metrics

Model performance is assessed with a comprehensive set of indicators that cover discrimination, calibration, and economic cost, following [30] and [60]. Let TP, FP, TN, and FN denote the entries of the standard confusion matrix. The evaluation indicators are:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{6}$$

$$\text{Error}_1 = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{7}$$

$$\text{Error}_2 = \frac{\text{FN}}{\text{FN} + \text{TP}}, \tag{8}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{9}$$

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}} \cdot \frac{\text{TN}}{\text{TN}+\text{FP}}}, \tag{10}$$

where Error_1 is the rate at which on-time repayers are incorrectly classified as defaulters (Type I) and Error_2 is the rate at which defaulters are incorrectly classified as on-time repayers (Type II). Because the data set is class-imbalanced, the harmonic F_1-score and the geometric G-mean are emphasized over raw accuracy, as both metrics penalize trivial solutions in which the minority class is systematically ignored [38]. The AUC is obtained via Platt scaling applied to the logistic-model outputs [30].

Table 1: Asymmetric cost matrix for credit evaluation.

	Observed class	
Predicted class	Default	On-time repayment
Default	$C_{\text{TP}} = 0$	C_{FP}
On-time repayment	C_{FN}	$C_{\text{TN}} = 0$

Table 2 lists the hyperparameter search spaces used for each model in Experiments I and II.

Table 2: Hyperparameter search ranges for all classifiers.

Classifier	**Search range**
LR	$C \in \{0.1, 0.25, 0.5, 1\}$; penalty $\in \{\ell_1, \ell_2\}$
RF	n_estimators $\in \{10, 50, 100\}$; max_depth $\in \{3, 5, 10\}$
SVM	type $\in$ {C-SVC, C-BSVC}; kernel $\in$ {linear, RBF}
XGBoost	max_depth $\in \{10, 20, 30\}$; $\eta \in \{0.3, 0.5, 1\}$
CatBoost	depth $\in \{10, 20, 30\}$; $\eta \in \{0.3, 0.5, 1\}$
LightGBM	max_depth $\in \{10, 20, 30\}$; $\eta \in \{0.3, 0.5, 1\}$
RI-LS	Classifier $\in$ {RF, LR}
Aug./Parc.	$C \in \{0.1, 0.25, 0.5, 1\}$
S3VM	$C \in \{1, 5\}$; $\gamma \in \{0.05, 0.1\}$; kernel $\in$ {RBF, linear}
CEM (base)	$\lambda = 0$; equal costs
CEM+L1	$\mu \in \{0.001, 0.01, 0.1, 1\}$; equal costs
CEM+CS	cost ratio $\in \{2, 5, 10, 20\}$; $\mu = 0$
RI-CSCEM	$\mu \in \{0.001, 0.01, 0.1, 1\}$; cost ratio empirical; $\lambda \in \{0.001, 0.01, 0.1, 1\}$

4.4 Results and Discussion

4.4.1 Benchmark comparison (Experiment I)

Table 4 reports the mean and standard deviation of each metric over thirty draws. At the 0.5 threshold, three models (SVM, Parceling, S3VM) collapse to a near-trivial all-non-default rule: F_1 and G-mean near zero, cost near the upper bound, yet AUC

remains competitive. Discrimination (AUC) and operating-point performance (cost, F_1, G-mean) measure different things; the quantity that matters economically is total misclassification cost.

On that economic criterion RI-CSCEM's mean total cost of 1120.3 is third lowest (Figure 2), behind cost-sensitive logistic regression (1070.4) and SMOTE-balanced logistic regression (1104.7), both of which discard the rejected population entirely. To test whether these differences are statistically meaningful we apply the Friedman omnibus test followed by the post-hoc Nemenyi test, treating each of the thirty draws as a paired block [49]; the resulting critical-difference diagrams are shown in Figure 1, and Table 3 reports the numerical details. The Friedman test rejects the null hypothesis of equal central tendency for every metric ($p < 10^{-3}$ in all four cases). On total cost the Nemenyi test (critical difference CD = 4.51) places RI-CSCEM (mean rank 2.87) in the same statistically indistinguishable group as LR (1.27) and SMOTE+LR (2.07); on F_1 RI-CSCEM (mean rank 3.57) again falls in the leading group with these two baselines (2.00 and 1.50, respectively).

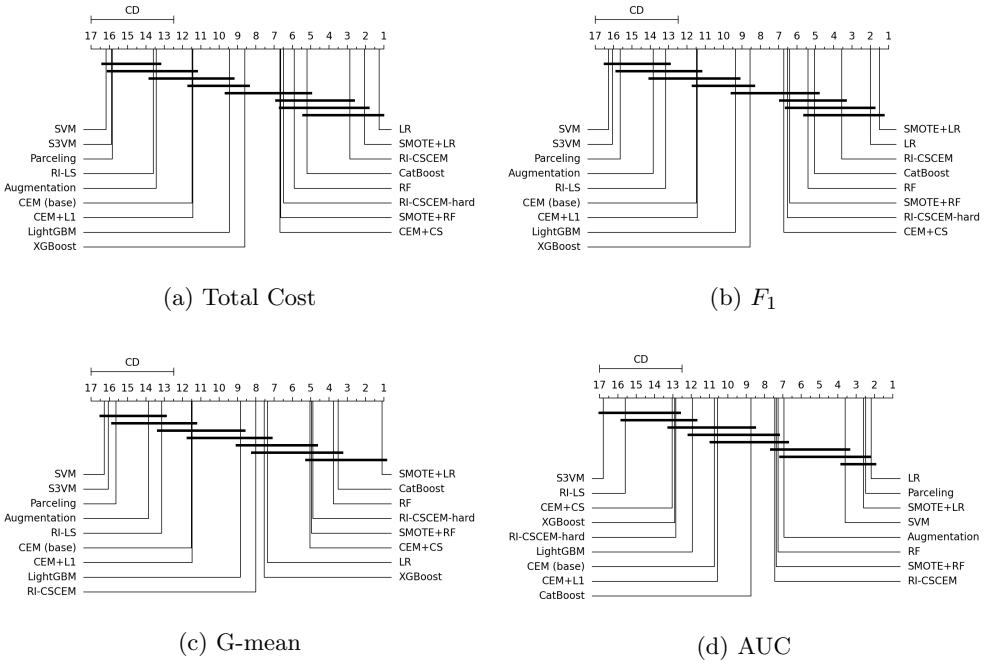
This outcome is consistent with the theoretical prediction of Proposition 1: under a MAR rejection mechanism the accepted-only estimator is already unbiased, so the rejected sample adds little additional accuracy. RI-CSCEM *matches* the strongest accepted-only baselines while additionally exploiting the rejected population; the benefit of reject inference should grow in a missing-not-at-random regime (Section 5).

Note that cost differences smaller than roughly one standard deviation ($\approx$ 50–90 units; see Table 4) should not be over-interpreted. In particular, CEM+CS (1215.6) and RI-CSCEM-hard (1214.3) differ by 1.3 units on a standard deviation of ≈ 90, i.e. they are within noise.

RI-CSCEM does not dominate on all axes. Its AUC (0.614) trails the best linear models (0.637), and its G-mean (0.524) falls below SMOTE+LR (0.597). The hyperparameters are tuned to minimise validation cost, which pushes the model toward high recall ($\text{Error}_2 = 0.224$) at the expense of specificity ($\text{Error}_1 = 0.637$), depressing G-mean. A tuning objective that balances cost against G-mean would shift this operating point.

Table 3: Friedman omnibus test and Nemenyi post-hoc critical difference for each metric ($k = 17$ models, $n = 30$ draws, $\alpha = 0.05$). Mean ranks for the three top models are listed; see the CD diagrams in Figure 1 for all models.

Metric	Friedman p	CD	LR (MR)	SMOTE+LR (MR)	RI-CSCEM (MR)
Total cost	$< 10^{-3}$	4.51	1.27	2.07	2.87
F_1	$< 10^{-3}$	4.51	2.00	1.50	3.57
G-mean	$< 10^{-3}$	4.51	7.37	1.10	8.00
AUC	$< 10^{-3}$	4.51	2.18	2.60	7.43



(a) Total Cost

(b) F_1

(c) G-mean

(d) AUC

Fig. 1: Critical-difference diagrams (Friedman + post-hoc Nemenyi, $n = 30$) for total cost, F_1, G-mean and AUC. Models joined by a bar are not significantly different at the 5% level; lower mean rank (left) is better.

4.4.2 Ablation study (Experiment II)

Table 5 isolates each component by building up from the base semi-supervised CEM logistic model. The ℓ_1 penalty alone (CEM+L1) leaves performance virtually unchanged: with only $p = 21$ features it drops almost nothing, so ℓ_1 serves sparsity and interpretability rather than accuracy. Cost-sensitive Diff weighting (CEM+CS) lowers total cost from 1401.5 to 1215.6 and raises F_1 from 0.228 to 0.334, moving the model off the trivial all-non-default solution. Note that RI-CSCEM-hard (1214.3) is within noise of CEM+CS (1215.6; difference 1.3 on a standard deviation of ≈ 90), confirming that the hard pseudo-labelling variant adds nothing beyond cost-sensitivity alone.

Replacing hard labels with soft responsibilities (the full RI-CSCEM) lowers cost to 1120.3 and recovers discrimination: AUC rises from 0.589 to 0.614 (Figure 3). This soft-vs-hard AUC recovery is the main empirical finding.

Interaction effects.

From Table 5 we can quantify component interactions on total cost. Starting from the base CEM cost of 1401.5: the ℓ_1 penalty alone reduces cost by 0.2 ($1401.5 - 1401.3$); cost-sensitive weighting alone reduces cost by 185.9 ($1401.5 - 1215.6$); the sum of

individual gains is 186.1. The full model (RI-CSCEM) reduces cost by 281.2 (1401.5 − 1120.3), so the interaction term (combined gain minus sum of individual gains) is +95.1, a synergy driven by the soft responsibility assignment.

Table 4: Benchmark comparison on the LendingClub accepted test set (mean ± std over 30 stratified draws). Metrics other than AUC use the 0.5 decision threshold; best value per column in bold.

Model	F_1	G-mean	AUC	Error_1	Error_2	Total cost
LR	0.365	0.546	**0.637**	0.627	**0.200**	**1070.4**
SVM	0.001	0.011	0.633	**0.001**	0.999	1567.9
RF	0.341	0.571	0.614	0.358	0.491	1202.7
XGBoost	0.315	0.541	0.589	0.305	0.578	1275.3
LightGBM	0.305	0.526	0.593	0.272	0.618	1298.3
CatBoost	0.345	0.575	0.607	0.407	0.442	1185.2
SMOTE+LR	**0.369**	**0.597**	0.636	0.432	0.372	1104.7
SMOTE+RF	0.336	0.563	0.613	0.314	0.536	1220.6
Augmentation	0.095	0.228	0.614	0.023	0.945	1509.7
Parceling	0.008	0.042	0.637	0.002	0.996	1565.0
S3VM	0.004	0.021	0.528	0.002	0.998	1567.1
RI-LS	0.121	0.272	0.560	0.056	0.921	1511.7
RI-CSCEM	0.353	0.524	0.614	0.637	0.224	1120.3

Table 5: Ablation of RI-CSCEM components (mean over 30 draws); best value per column in bold.

Model	F_1	G-mean	AUC	Error_1	Error_2	Total cost
CEM (base)	0.228	0.402	0.597	**0.095**	0.820	1401.5
CEM+L1	0.228	0.402	0.598	**0.095**	0.820	1401.3
CEM+CS	0.334	0.562	0.588	0.430	0.444	1215.6
RI-CSCEM-hard	0.334	**0.562**	0.589	0.431	0.442	1214.3
RI-CSCEM	**0.353**	0.524	**0.614**	0.637	**0.224**	**1120.3**

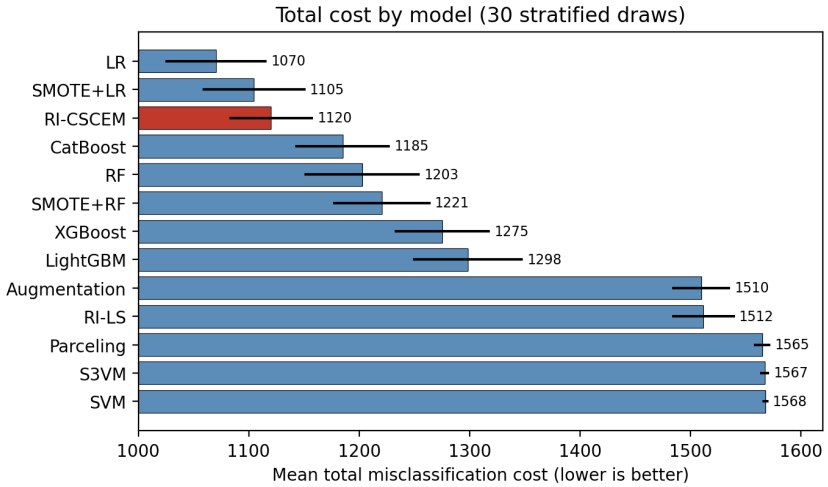


Fig. 2: Mean total misclassification cost per model over the thirty stratified draws (error bars: one standard deviation). RI-CSCEM (highlighted) attains the third-lowest cost and is statistically indistinguishable from the two leading baselines (cf. Figure 1).

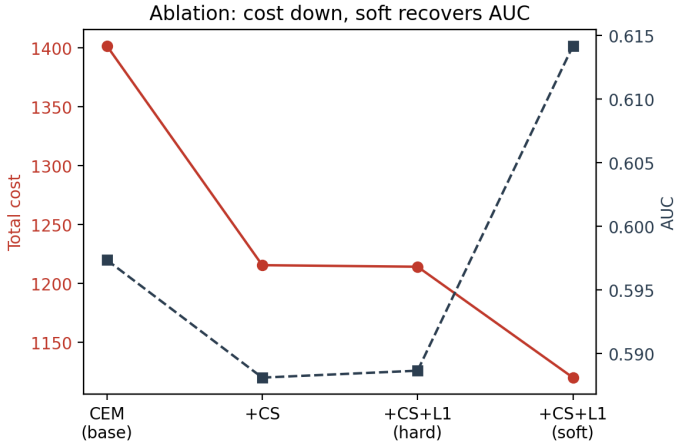


Fig. 3: Ablation progression. Cost-sensitive weighting drives the total cost down (solid), while replacing hard labels with soft responsibilities recovers the discrimination (AUC, dashed) that the hard variant loses.

5 Conclusion

RI-CSCEM fits a cost-sensitive logistic model on accepted and rejected applicants through a Classification EM procedure. The E-step assigns each rejected applicant a cost-weighted soft responsibility under a balance constraint; the M-step refits an ℓ_1-penalised cost-sensitive logistic regression. The hard-assignment variant decreases the objective monotonically, and the estimator is consistent under MAR.

On 1,266,782 accepted and 2,064,314 rejected Lending Club applications, RI-CSCEM matched the strongest accepted-only baselines on misclassification cost (mean 1120, statistically indistinguishable from cost-sensitive LR at 1070 and SMOTE+LR at 1105) across thirty stratified draws. This outcome is consistent with the theory: under MAR, the accepted-only estimator is already unbiased. The ablation traced performance to two components: cost-sensitive weighting moved the model off the trivial all-non-default solution, and soft responsibilities recovered discrimination that hard pseudo-labelling loses, raising AUC from 0.589 to 0.614. The ℓ_1 penalty changed accuracy little at $p = 21$ features.

Three directions follow. First, the rejected sample added little under MAR; the gain should grow under MNAR, when rejection depends on factors the scorecard does not observe. A controlled MNAR study, or an evaluation on a portfolio with observed reject outcomes, would measure the benefit. Second, the balance constraint forces the reject default rate to equal the accepted rate r, which may suppress signal under selection bias; relaxing it is a natural MNAR extension. Third, tuning to minimise cost pushed the model toward high recall and depressed G-mean (0.524) relative to SMOTE+LR (0.597); a multi-objective tuning criterion would address this. Replacing the logistic base learner with a calibrated gradient-boosted model inside the same CEM scheme is a further direction.

Acknowledgements

I would like to thank Prof. Badreddine Benyacoub for his supervision and guidance on the statistical methodology, the convergence analysis, and the semi-supervised learning framework underpinning this work; Prof. Mohammed El Haj Tirari for his co-supervision and insights on cost-sensitive learning and the experimental design; and Prof. Mounir El Annas for sharing his expertise in credit scoring and reject inference, which shaped the practical motivation of this study.

Declarations

Funding.

This research received no external funding.

Conflict of interest.

The authors declare that there is no conflict of interest.

References

[1] Thomas, L.C.: A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting **16**(2), 149–172 (2000) https://doi.org/10.1016/S0169-2070(00)00034-0

[2] Horrigan, J.O.: The determination of long-term credit standing with financial ratios. Journal of Accounting Research **4**, 44–62 (1966)

[3] West, R.R.: An alternative approach to predicting corporate bond ratings. Journal of Accounting Research **8**, 118–125 (1970)

[4] Peussa, A.: Credit risk scorecard estimation by logistic regression. Master's thesis, University of Helsinki, Faculty of Science, Department of Mathematics and Statistics (2016)

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

[6] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance **23**, 589–609 (1968)

[7] Bellotti, T., Crook, J.: Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications **36**, 3302–3308 (2009)

[8] Hongxia, W., Xueqin, L., Yanhui, L.: Enterprise credit rating model based on fuzzy clustering and decision tree. In: Information Science and Engineering (ISISE) International Symposium On, pp. 105–108. IEEE, ??? (2010)

[9] Khashman, A.: Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications **37**, 6233–6239 (2010)

[10] Kurniawan, R.: Application of random forest algorithm on credit risk analysis. Procedia Computer Science **245**, 740–749 (2024)

[11] Li, J., Liu, H., Yang, Z., Han, L.: A credit risk model with small sample data based on G-XGBoost. Applied Artificial Intelligence **35**(15), 1550–1566 (2021)

[12] Ponsam, J.G., Gracia, S.J.B., Geetha, G., Karpaselvi, S., Nimala, K.: Credit risk analysis using LightGBM and a comparative study of popular algorithms. In: Proceedings of the 2021 4th International Conference on Computing and Communications Technologies (ICCCT), pp. 634–641. IEEE, ??? (2021)

[13] Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)

[14] Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., Beben, S.: Reject inference methods in credit scoring. Journal of Applied Statistics **48**(13–15), 2734–2754 (2021)

[15] Hady, M.F.A., Schwenker, F.: Semi-supervised learning. In: Bianchini, M., Maggini, M., Jain, L. (eds.) Handbook on Neural Information Processing. Intelligent Systems Reference Library, vol. 49, pp. 215–239. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36657-4_7

[16] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE, ??? (2008)

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

[18] Bahnsen, A.C., Aouada, D., Ottersten, B.: Example-dependent cost-sensitive logistic regression for credit scoring. In: Proceedings of the 2014 13th International Conference on Machine Learning and Applications, pp. 263–269. IEEE, ??? (2014)

[19] Vittaut, J.N., Amini, M.R., Gallinari, P.: Learning classification with both labeled and unlabeled data. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) Machine Learning: ECML 2002. Lecture Notes in Computer Science, vol. 2430. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-36755-1_39

[20] Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis **14**(3), 315–332 (1992)

[21] Crook, J., Banasik, J.: Does reject inference really improve the performance of application scoring models? Journal of Banking & Finance **28**(4), 857–874 (2004)

[22] 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)

[23] Anderson, R.: The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press, ??? (2007)

[24] Hand, D.J.: Modelling consumer credit risk. IMA Journal of Management Mathematics **12**(2), 139–155 (2001)

[25] Galindo, J., Tamayo, P.: Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational

Economics **15**(1), 107–143 (2000)

[26] Feelders, A.J.: Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance & Management **9**(1), 1–15 (2000) https://doi.org/10.1002/(SICI)1099-1174(200003)9:1⟨1::AID-ISAF177⟩3.0.CO;2-2

[27] Banasik, J., Crook, J.N.: Reject inference, augmentation, and sample selection. European Journal of Operational Research **183**(3), 1582–1594 (2007)

[28] Anderson, B., Hardin, J.M.: Modified logistic regression using the EM algorithm for reject inference. International Journal of Data Analysis Techniques and Strategies **5**(4), 359–373 (2013) https://doi.org/10.1504/IJDATS.2013.058582

[29] Li, Z., Tian, Y., Li, K., Zhou, F., Yang, W.: Reject inference in credit scoring using semi-supervised support vector machines. Expert Systems with Applications **70**, 201–210 (2017) https://doi.org/10.1016/j.eswa.2017.01.011

[30] Mancisidor, R.A., Kampffmeyer, M., Aas, K., Jenssen, R.: Deep generative models for reject inference in credit scoring. Knowledge-Based Systems **196**, 105758 (2020)

[31] Hájek, P., Olej, V.: Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. Neural Computing and Applications **20**(6), 761–773 (2011)

[32] Xia, Y., Yang, X., Zhang, Y.: A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. Electronic Commerce Research and Applications **30**, 111–124 (2018)

[33] Annas, M.E., Benyacoub, B., Ouzineb, M.: Semi-supervised adapted HMMs for P2P credit scoring systems with reject inference. Computational Statistics **38**(1), 149–169 (2023)

[34] Kennedy, K., Namee, B.M., Delany, S.J.: Using semi-supervised classifiers for credit scoring. Journal of the Operational Research Society **64**(4), 513–529 (2013)

[35] Liu, Y., Li, X., Zhang, Z.: A new approach in reject inference using ensemble learning based on a global semi-supervised framework. Future Generation Computer Systems **109**, 382–391 (2020)

[36] Anderson, B.: Using Bayesian networks to perform reject inference. Expert Systems with Applications **137**, 349–356 (2019)

[37] Xiao, J., Cao, H., Jiang, X., Gu, X., Xie, L.: GMDH-based semi-supervised feature selection for customer classification. Knowledge-Based Systems **132**, 236–248 (2017)

[38] Kang, Y., Chen, L., Jia, N., *et al.*: A CWGAN-GP-based multi-task learning model for consumer credit scoring. Expert Systems with Applications **206**, 117650 (2022)

[39] Falangis, K.: The use of MSD model in credit scoring. Operational Research **7**(3), 481–503 (2007)

[40] Verbeke, W., Martens, D., Mues, C., Baesens, B.: Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications **38**(3), 2354–2364 (2011)

[41] Baesens, B., Gestel, T.V., 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)

[42] Yobas, M.B., Crook, J.N., Ross, P.: Credit scoring using neural and evolutionary techniques. IMA Journal of Management Mathematics **11**(2), 111–125 (2000)

[43] 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**(1), 24–37 (1996)

[44] Weiss, G.M., Provost, F.J.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research **19**, 315–354 (2003)

[45] Batista, G.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter **6**(1), 20–29 (2004)

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

[47] Friedman, J.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**(5), 1189–1232 (2001)

[48] Friedman, J.: Stochastic gradient boosting. Computational Statistics and Data Analysis **38**(4), 367–378 (2002)

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

[50] Zhu, S., Wan, J.: Cost-sensitive learning for semi-supervised hit-and-run analysis. Accident Analysis and Prevention **158**, 106199 (2021)

[51] Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)

[52] Van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, ??? (1998)

[53] Wu, C.F.J.: On the convergence properties of the EM algorithm. The Annals of Statistics **11**(1), 95–103 (1983)

[54] Shen, F., Zhao, X., Kou, G.: Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. Decision Support Systems **137**, 113366 (2020)

[55] Banasik, J., Crook, J., Thomas, L.: Sample selection bias in credit scoring models. Journal of the Operational Research Society **54**(8), 822–832 (2003)

[56] Xia, Y., Liu, C., Liu, N.: Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electronic Commerce Research and Applications **24**, 30–49 (2017) https://doi.org/10.1016/j.elerap.2017.06.004

[57] Petrides, G., Moldovan, D., Coenen, L., Guns, T., Verbeke, W.: Cost-sensitive learning for profit-driven credit scoring. Journal of the Operational Research Society **73**(2), 338–350 (2022) https://doi.org/10.1080/01605682.2020.1843975

[58] Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 973–978 (2001)

[59] Khalili, N., Rastegar, M.A.: Optimal cost-sensitive credit scoring using a new hybrid performance metric. Expert Systems with Applications **213**, 119232 (2023) https://doi.org/10.1016/j.eswa.2022.119232

[60] Verbraken, T., Bravo, C., Weber, R., Baesens, B.: Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research **238**(2), 505–513 (2014) https://doi.org/10.1016/j.ejor.2014.04.001