World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr4124

RISK-AWARE ENSEMBLE LEARNING FOR CREDIT DEFAULT PREDICTION UNDER ECONOMIC UNCERTAINTY

JiYuan Wang[1*], JingYi Liu[2], Yao Ge[3], BingYing Jiang[4]
[1]*Duke University, Durham 27708, North Carolina, USA.*
[2]*Cornell University, Ithaca 14853, New York, USA.*
[3]*University of Miami, Coral Gables 33124, Florida, USA.*
[4]*University of Wisconsin-Madison, Madison 53706, Wisconsin, USA.*
[*]*Corresponding Author: JiYuan Wang*

Abstract: Credit scoring models are estimated on past loan vintages but deployed into economic conditions that are unknown at origination. When the economy enters a downturn, ensembles fitted to average historical performance produce probability-of-default (PD) estimates that deteriorate precisely when accurate risk measurement matters most. This paper proposes TRACE (Tail-Risk-Aware Calibrated Ensemble), a stacking ensemble whose aggregation layer is explicitly risk-aware with respect to economic regimes. TRACE (i) builds a rolling-origin prediction cube in which every base-learner prediction is a genuine out-of-time forecast; (ii) derives two epistemic margins from the cube—cross-learner and cross-regime dispersion—that enter the PD as non-negative conservatism terms; (iii) fits the aggregator by minimising the conditional value-at-risk of vintage-level log-losses around a recency-anchored nominal regime distribution, which is equivalent to a distributionally robust problem over regime mixtures with bounded likelihood ratio; and (iv) learns the length of the regime memory from data through one-step-ahead validation. The resulting convex program is solved exactly. Experiments on two public loan data sets spanning the Great Recession—90,000 U.S. Small Business Administration (SBA) loans approved in 1993–2010 and 39,786 Lending Club loans issued in 2007–2011—compare TRACE with six base learners and eight ensemble, drift-adaptive and dynamic-selection benchmarks over 51 out-of-time test periods. On the SBA data, TRACE attains the lowest average log-loss, Brier score, calibration error and misclassification cost, reduces the worst-year log-loss by 11.0% relative to classical stacking, and cuts the PD under-estimation during the 2005–2008 stress vintages from 12.7 to 1.5 percentage points. On the Lending Club data, where drift is mild, it performs on par with the strongest ensembles. Ablation, sensitivity and label-delay analyses identify the adaptive regime memory as the main source of improvement and the CVaR objective as an effective control of tail risk.
Keywords: Credit scoring; Ensemble learning; Economic uncertainty; Distribution shift; Conditional value-at-risk; Probability of default

1 INTRODUCTION

Probability-of-default (PD) models are the backbone of retail and small-business lending. They are used to accept or reject applications, to price loans, to set provisions and to compute regulatory capital. A PD model is always trained on loans that were originated in the past and whose outcomes are already known, but it is applied to new loans whose outcomes will be determined by economic conditions that do not exist yet. Under stable conditions this temporal gap is harmless. Under economic uncertainty it is not: a recession, a credit crunch or a policy shock changes default rates across the whole portfolio, changes the relative riskiness of borrower segments, and changes the relationship between application characteristics and repayment behaviour [1-2]. Measures of macroeconomic and policy uncertainty spike exactly around such episodes [3-5], and the vintages originated just before them are the ones that suffer the largest losses.

Machine-learning ensembles currently provide the most accurate credit scores on tabular data [6-7]. Heterogeneous stacking, ensemble selection and gradient-boosted trees consistently outperform single classifiers in benchmark studies [8-11]. These comparisons, however, are almost always carried out with random cross-validation or with a single hold-out sample, i.e. under the implicit assumption that training and deployment data are exchangeable. The aggregation layer of an ensemble inherits this assumption in two ways. First, the meta-learner of a stacking ensemble is trained on out-of-fold predictions produced by random splits, so it never observes how base learners behave when they extrapolate into a new economic regime. Second, the weights are chosen to minimise an average loss over the pooled history, so that the ensemble is *risk-neutral* with respect to regimes: a configuration that is slightly better in benign years and much worse in a crisis is preferred to one that is uniformly reliable. The consequence is visible in our experiments. During the 2005–2008 SBA vintages, which were hit by the Great Recession, classical stacking under-estimates the realised default rate by 12.7 percentage points on average.

This paper argues that economic uncertainty should be handled where ensembles make their key decision, namely in the aggregation of base-learner opinions, and that it should be handled with the tools of risk management rather than with point forecasts of the economy. We formalise economic uncertainty as *ambiguity about which regime a new vintage will face*. The historical vintages define a menu of regimes; a nominal belief puts more weight on recent regimes

© By the Author(s) 2023, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

because economic conditions are persistent; and the lender hedges against every mixture of regimes whose likelihood ratio relative to the nominal belief is bounded by $1/\alpha$. Minimising the worst-case expected loss over this ambiguity set is equivalent to minimising the conditional value-at-risk (CVaR) of vintage-level losses [12-13]. Building on this formulation we propose TRACE, a Tail-Risk-Aware Calibrated Ensemble. The main contributions are:

1. A regime-risk formulation of ensemble aggregation. We cast stacking as a distributionally robust problem over mixtures of economic regimes, with a recency-anchored nominal distribution and a CVaR ambiguity set. We show that the resulting objective is convex in the aggregation parameters, that it interpolates between nominal-weighted empirical risk minimisation and worst-vintage optimisation, and that its optimal value bounds the expected loss of any future regime mixture in the ambiguity set (Proposition 1).
2. The TRACE algorithm. TRACE combines (a) a rolling-origin prediction cube that turns every base-learner prediction into a genuine out-of-time forecast at the cost of one model fit per learner and vintage, (b) two epistemic margins—cross-learner and cross-regime dispersion—that act as data-driven margins of conservatism, (c) an exact convex solver for the CVaR aggregation problem, and (d) a one-step-ahead validation procedure that learns the regime memory (the half-life of the nominal distribution) without access to future information.
3. A rigorous out-of-time evaluation across the Great Recession. Using 90,000 SBA 7(a) loans approved between 1993 and 2010 and 39,786 Lending Club loans issued between 2007 and 2011, we compare TRACE with 14 benchmarks over 51 test periods and three independent runs. The study includes ablations of every component, sensitivity analyses of all hyper-parameters, a label-delay robustness check, statistical tests, and decision-level evaluations based on misclassification costs and realised loan profits. We also document a target leakage in the widely used SBA data set (the loan term of charged-off loans is recorded *after* default) and remove it.

The empirical results show a clear pattern. When the economy moves through a severe regime change, as in the SBA data, TRACE delivers the best probability estimates on every proper scoring rule and calibration measure, the lowest tail losses, and the lowest misclassification cost; its advantage is concentrated in the stress vintages, where it reduces log-loss by 6.9% relative to classical stacking and almost eliminates PD under-estimation. When drift is mild, as in the Lending Club data, TRACE performs on par with the strongest ensembles. The adaptive regime memory is the main driver of these gains: as the crisis unfolds, the validation procedure shortens the memory from an infinite horizon to a half-life of one quarter of a year, which is an interpretable signature of the regime change.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 formalises the problem, Section 4 presents TRACE, and Section 5 describes the experimental design. Section 6 reports the results, Section 7 discusses implications for credit risk management, and Section 8 concludes.

2 RELATED WORK

2.1 Ensemble Learning for Credit Scoring

Large benchmark studies have established that ensembles are the most accurate classifiers for credit scoring. Lessmann et al. compared 41 classifiers on eight credit data sets and found heterogeneous ensembles, in particular ensemble selection, to be the strongest family [6]. Subsequent work refined the construction of heterogeneous ensembles through bagging-based stacking [8], consensus mechanisms [14], dynamic selection of tree-based learners [15] and the choice of base classifiers [9]. Gradient-boosted decision trees such as XGBoost, LightGBM and CatBoost are now standard components [16-18], and comparisons on tabular data show that tree ensembles remain hard to beat by deep networks [7], a conclusion confirmed for credit scoring by Gunnarsson et al. [10]. Interpretable alternatives that inject tree-based effects into logistic regression have also been proposed [11]. Ensembles have been applied to peer-to-peer (P2P) lending [19-21], to imbalanced data [22], and to mortgage default in a distressed market [23]. A common feature of these studies is that the aggregation rule is learned under an exchangeability assumption and optimised for average accuracy.

2.2 Credit Risk under Changing Economic Conditions

Credit risk research has long recognised that default behaviour depends on the economic environment. Bellotti and Crook incorporate macroeconomic variables into discrete-time survival models to forecast and stress-test credit card and retail portfolios, and use value-at-risk and expected shortfall to summarise the distribution of simulated default rates [1, 24]. Leow and Crook document that the parameters of PD models estimated before and after the 2008 credit crisis are statistically different [2]. Sousa et al. show that credit models built from sequences of data snapshots outperform static models and that a shorter memory is preferable during shocks [25], an observation that our adaptive regime memory turns into a learnable quantity. Economic uncertainty itself has been measured through newspaper-based policy uncertainty indices [5] and through the unforecastable component of a large set of macroeconomic series [4]; Bloom reviews its effects on the real economy [3]. Our approach is complementary to macro-econometric PD models: instead of forecasting the macroeconomic path a loan will face, it hedges the aggregation of machine-learning models against the set of regimes observed in the past.

2.3 Learning under Distribution Shift and Risk-Averse Learning

Concept drift adaptation is surveyed by Gama et al. [26]. Classical drift-adaptive ensembles weight members trained on recent data chunks by their current accuracy [27], and dynamic ensemble selection chooses competent members locally in the feature space [28-29]. These approaches react to drift but do not explicitly control the risk of adverse future regimes. Distributionally robust optimisation (DRO) minimises the worst-case expected loss over an uncertainty set of distributions. Duchi and Namkoong analyse DRO with f-divergence balls, including the CVaR case, and show that it improves performance on the tails of the data-generating distribution [13]; Sagawa et al. minimise the worst-group loss over pre-defined groups [30]. CVaR itself was introduced as a tractable coherent risk measure by Rockafellar and Uryasev [12]. TRACE transfers these ideas to the aggregation layer of a credit-scoring ensemble, where the "groups" are origination vintages and the uncertainty set is anchored on a nominal distribution whose memory is learned from data.

2.4 Uncertainty Quantification and Calibration

Ensembles provide a natural estimate of epistemic uncertainty through the disagreement of their members [31]. Malinin et al. extend this view to gradient boosting [32], and Ovadia et al. show that the quality of predictive uncertainty deteriorates under dataset shift and that post-hoc calibration on in-distribution data is insufficient [33]. Temperature and Platt-type scaling are effective under stationarity [34]. In credit scoring, calibration determines expected losses, provisions and profit-based decisions [35-36]. TRACE uses disagreement across learners and across training regimes as non-negative margins inside the aggregation, so that the amount of conservatism is learned jointly with the ensemble weights under the risk-aware objective.

3 PROBLEM FORMULATION

3.1 Vintages and the Out-of-Time Setting

Let $\mathcal{D} = \{(x_i, y_i, v_i)\}_{i=1}^{N}$ be a loan book, where $x_i \in \mathbb{R}^d$ collects origination-time attributes, $y_i \in \{0,1\}$ indicates default (charge-off) and $v_i \in \{1, ..., V\}$ is the origination vintage (a fiscal year or a calendar quarter). Vintages are ordered in time and define economic regimes: loans of the same vintage are exposed to the same macroeconomic path during their life. At decision time T a lender has to assign a PD $\hat{p}_i$ to every new loan of vintage T using only information available before T. Ignoring the outcome delay for the moment (Section 6.5 relaxes this), the labelled history consists of vintages $1, ..., T-1$.

We consider K heterogeneous base learners $f^1, ..., f^K$ and an aggregator g_θ that maps the base-learner outputs to a PD. The quality of a PD for vintage v is measured by the vintage log-loss

$$L_v(\theta) = \frac{1}{n_v} \sum_{i:\, v_i = v} \left[-y_i \log \hat{p}_i(\theta) - (1 - y_i) \log\left(1 - \hat{p}_i(\theta)\right) \right], \tag{1}$$

a strictly proper scoring rule that rewards both discrimination and calibration.

3.2 Economic Uncertainty as Regime Ambiguity

The joint distribution of (x, y) for the next vintage is unknown. We assume that it resembles a mixture of the regimes that have been observed, but we do not know the mixture weights. Economic conditions are persistent, so a natural nominal belief puts more weight on recent vintages. For a half-life $h \in (0, \infty]$ we define the recency-anchored nominal distribution over the labelled vintages $\mathcal{V}_T$,

$$q_v(h) = \frac{2^{-a_v/h}}{\sum_{u \in \mathcal{V}_T} 2^{-a_u/h}}, \qquad a_v = \max(\mathcal{V}_T) - v, \tag{2}$$

which is uniform for $h = \infty$ and concentrates on the latest vintage as $h \to 0$. Ambiguity about the regime of the next vintage is represented by the set of mixtures whose likelihood ratio with respect to the nominal belief is bounded by $1/\alpha$:

$$\mathcal{U}_\alpha(q) = \left\{ q' \in \Delta_{|\mathcal{V}_T|} :\ 0 \le q'_v \le q_v/\alpha \ \ \forall v \right\}, \qquad \alpha \in (0,1]. \tag{3}$$

The parameter α expresses the risk appetite of the lender: $\alpha = 1$ trusts the nominal belief completely, whereas small values allow the adversary to up-weight any past regime, including the worst crisis vintage, by a large factor. A risk-aware lender chooses the aggregation parameters θ that minimise the worst-case expected vintage loss,

$$\min_{\theta \in \Theta} \max_{q' \in \mathcal{U}_\alpha(q(h))} \sum_{v \in \mathcal{V}_T} q'_v L_v(\theta). \tag{4}$$

Two quantities thus describe economic uncertainty: the persistence of regimes (h), which is a statistical property that can be learned from the data, and the risk appetite (α), which is a managerial choice.

4 THE TRACE ENSEMBLE

Figure 1 summarises the method. TRACE consists of four components that are described in turn.

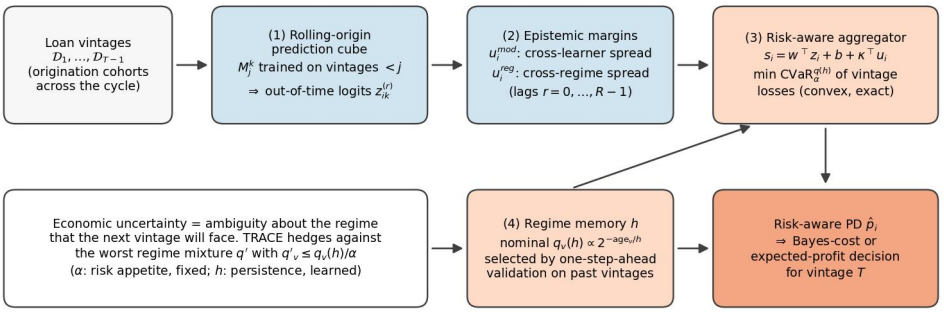


Figure 1 Overview of TRACE. Base Learners Are Re-Estimated on a Rolling Origin so That Every Prediction Used by the Aggregator Is an Out-of-Time Forecast. Disagreement across Learners and across Training Regimes Yields Two Epistemic Margins. The Aggregator Minimises the CVaR of Vintage Losses around a Recency-Anchored Nominal Regime Distribution Whose Half-Life Is Selected by One-Step-Ahead Validation

4.1 Rolling-Origin Prediction Cube

For every vintage $j \geq j_0$ and every learner k, a model M_j^k is trained on the vintages $\{\max(1, j-W), \ldots, j-1\}$, where W is the length of the training window (expanding if $W = \infty$). The model scores the A vintages $j, \ldots, j+A-1$. This yields the prediction cube

$$z_{ik}^{(r)} = \text{logit} M_{v_i - r}^k(x_i), \qquad r = 0, \ldots, R-1, \tag{5}$$

where r is the forecast lag. The entry $z_{ik}^{(0)}$ is a genuine one-step-ahead out-of-time forecast: the model that produced it has not seen vintage v_i or any later vintage. The entries with $r \geq 1$ are forecasts of the same loan by models trained on older information sets. The cube is the information set of the aggregator: for a test vintage T, the aggregator is trained on the cube entries of the labelled vintages $v < T$ and applied to the entries of vintage T. This differs from classical stacking, in which the meta-learner is trained on random out-of-fold predictions that do not reflect the extrapolation errors of the base learners. The construction requires exactly one fit per learner and vintage, because the model M_j^k serves both as the out-of-time forecaster for vintage j and as the deployed model when j is the test vintage.

4.2 Epistemic Margins

Two dispersion statistics are extracted from the cube for every loan:

$$u_i^{mod} = \text{sd}_k\left(z_{ik}^{(0)}\right) \tag{6}$$

$$u_i^{reg} = \frac{1}{K}\sum_{k=1}^{K} \text{sd}_r\left(z_{ik}^{(r)}\right). \tag{7}$$

The cross-learner margin u_i^{mod} measures model uncertainty in the sense of ensemble disagreement [31-32]. The cross-regime margin u_i^{reg} measures how strongly the PD of a loan depends on which historical regimes were available for training; a large value indicates that recent vintages have changed the model's view of this type of borrower, i.e. that the loan is exposed to regime uncertainty. Both statistics are non-negative and come at no additional training cost.

4.3 Risk-Aware Aggregation

The aggregator is a constrained logistic stacker with margins,

$$\hat{p}_i(\theta) = \sigma\big(s_i(\theta)\big), \qquad s_i(\theta) = w^\top z_i^{(0)} + b + \kappa_1 u_i^{mod} + \kappa_2 u_i^{reg}, \tag{8}$$

with $\theta = (w, b, \kappa)$, $w \in \mathbb{R}^K_{\geq 0}$, $\kappa \in \mathbb{R}^2_{\geq 0}$ and σ the logistic function. Non-negative weights on logits are a standard safeguard against over-fitting in stacking [37], and non-negative κ guarantees that the margins can only increase the PD, so that they act as data-driven margins of conservatism. The parameters are obtained from

$$\min_\theta J_{\alpha,h}(\theta) = \text{CVaR}_\alpha^{q(h)}\big(L_1(\theta), \ldots, L_{|\mathcal{V}_T|}(\theta)\big) + \lambda \parallel w - 1/K\,\mathbf{1}\parallel_2^2, \tag{9}$$

where, following Rockafellar and Uryasev [12],

$$\mathrm{CVaR}_{\alpha}^{q}(L) = \min_{\eta\in\mathbb{R}} \eta + \frac{1}{\alpha}\sum_{v} q_v \,[L_v - \eta]_{+}. \tag{10}$$

The ridge term shrinks the weights towards the equally weighted logit average and stabilises the solution when few vintages are available.

Proposition 1. Let q have strictly positive entries and $\alpha \in (0,1]$. Then:

1. $\mathrm{CVaR}_{\alpha}^{q}(L) = \max_{q'\in\mathcal{U}_{\alpha}(q)} \sum_{v} q'_v L_v$; hence problem (4) and the minimisation of (9) with $\lambda = 0$ coincide.
2. $J_{\alpha,h}$ is convex in θ, and the minimisation of (9) is equivalent to the convex program

$$\min_{\theta,\eta,t} \eta + \frac{1}{\alpha}\sum_{v} q_v\, t_v + \lambda \,\|\, w - 1/K\, 1\|^2 \quad \text{s.t.} t_v \geq L_v(\theta) - \eta,\ t_v \geq 0,\ w \geq 0,\ \kappa \geq 0. \tag{11}$$

3. For $\alpha = 1$ the objective reduces to the nominal-weighted empirical risk $\sum_{v} q_v L_v(\theta)$; for $\alpha \leq \min_v q_v$ it reduces to the worst-vintage risk $\max_v L_v(\theta)$.
4. $\mathrm{CVaR}_{\alpha}^{q}(L)$ is non-increasing in α, and for the minimiser $\theta^{\star}$ and every future regime mixture $q' \in \mathcal{U}_{\alpha}(q)$, $\sum_{v} q'_v L_v(\theta^{\star}) \leq J_{\alpha,h}(\theta^{\star})$.

Proof. (a) For fixed L, the right-hand side is a linear program over the capped simplex; its Lagrangian dual with respect to the equality constraint $\sum_{v} q'_v = 1$ is exactly the minimisation over η in (10) [12-13]. (b) The log-loss $\log(1 + e^{s}) - ys$ is convex in s, and $s_i(\theta)$ is affine in θ, so each L_v is convex. By (a), $\mathrm{CVaR}_{\alpha}^{q}\big(L(\theta)\big)$ is a pointwise maximum of non-negative combinations of convex functions and is therefore convex; the ridge term is convex and the constraints define a convex set. The epigraph reformulation follows by introducing $t_v = [L_v - \eta]_{+}$. (c) If $\alpha = 1$, the constraints $q'_v \leq q_v$ together with $\sum_{v} q'_v = 1$ force $q' = q$. If $\alpha \leq \min_v q_v$, every cap q_v/α is at least one, so the maximum over the simplex puts all mass on the largest loss. (d) The set $\mathcal{U}_{\alpha}(q)$ grows as α decreases, so the maximum in (a) cannot decrease. The bound follows from (a) and the non-negativity of the ridge term. ▫

Part (d) gives the risk-management interpretation of TRACE: the fitted ensemble comes with a certificate on its expected log-loss under every regime mixture that is at most $1/\alpha$ times more likely than the nominal belief, including mixtures dominated by the worst historical crisis vintage. Program (11) has $K + 4 + |\mathcal{V}_T|$ variables and $|\mathcal{V}_T|$ smooth convex constraints; we solve it with sequential quadratic programming using analytic gradients of the vintage losses, which takes well below one second for the problem sizes in this paper.

4.4 Learning the Regime Memory

The half-life h determines how quickly the nominal belief forgets old regimes. A long memory is preferable in stable periods because it averages over more data; a short memory is preferable during regime changes because the most recent vintages are the most informative [25]. TRACE therefore treats h as a statistical parameter and selects it by one-step-ahead validation on the labelled history: for every candidate $h \in \mathcal{H}$ and each of the last P labelled vintages $\tau \in \{T - 1, \ldots, T - P\}$, the aggregator is fitted on the vintages before τ and evaluated on τ; the candidate with the smallest average validation log-loss is used to fit the final aggregator on all labelled vintages. The risk appetite α is held fixed, because it encodes a preference rather than a property of the data. Validation fits are cached across decision times, since a fit for validation vintage τ depends only on vintages before τ. Algorithm 1 summarises the procedure.

Algorithm 1 TRACE at decision time T (outcome delay $g = 0$).

Input: cube $\{z_{ik}^{(r)}\}$ for all loans with $v_i \leq T$; labels of vintages $< T$; α, λ, R, P, grid $\mathcal{H}$.

1. Compute u_i^{mod} and u_i^{reg} by Eqs. (6)–(7) for all loans with available lags.
2. For every $h \in \mathcal{H}$ and every validation vintage $\tau = T - 1, \ldots, T - P$: fit $\theta_{h,\tau}$ by solving program (11) on vintages $< \tau$ with nominal distribution $q(h)$ (cached), and record the one-step-ahead validation loss $\ell(h,\tau) = L_{\tau}(\theta_{h,\tau})$.
3. Select $h^{\star} = \mathrm{argmin}_h \frac{1}{P}\sum_{\tau} \ell\,(h,\tau)$.
4. Fit $\theta^{\star}$ by solving program (11) on all vintages $< T$ with nominal distribution $q(h^{\star})$.
5. Return $\hat{p}_i = \sigma\big(s_i(\theta^{\star})\big)$ for all loans of vintage T.

4.5 Risk-Aware Credit Decisions

Calibrated PDs translate directly into decisions. With misclassification costs c_{FN} (accepting a defaulter) and c_{FP} (rejecting a good borrower) and cost ratio $\rho = c_{FN}/c_{FP}$, the Bayes rule accepts an application if $\hat{p}_i < 1/(1 + \rho)$. For

loans with known contractual cash flows, a lender maximising expected profit accepts loan i if $(1-\hat{p}_i)G_i - \hat{p}_i\,\widehat{\mathrm{LGD}}\,A_i > 0$, where A_i is the funded amount, G_i the scheduled interest income and $\widehat{\mathrm{LGD}}$ the loss rate of defaulted loans in the labelled history [35-36]. Both rules reward calibration in addition to ranking, so under-estimated PDs in a downturn translate directly into excess losses.

4.6 Computational Cost

The cube requires K model fits per vintage, which is the same number of fits as training the base learners once per period. Classical m-fold stacking requires mK additional fits per decision time. The aggregation problem is small: each evaluation of the vintage losses and gradients costs $O\big(n_{meta}(K+2)\big)$ operations, where n_{meta} is the number of labelled loans in the cube. With $|\mathcal{H}| = 5$ candidate half-lives and caching, the memory selection adds about five small convex fits per decision time.

5 EXPERIMENTAL DESIGN

5.1 Data

Two public data sets with origination dates spanning the Great Recession are used (Table 1, Figure 2). Only attributes that are known at origination are used as predictors.

SBA 7(a) loans. The U.S. Small Business Administration national data set published by Li et al. contains 899,164 guaranteed loans with their final status (paid in full or charged off) [38]. We use the approval fiscal years 1993–2010 as vintages and draw 5,000 loans per year (90,000 loans); two further independent random subsamples are drawn for the robustness runs. The default rate rises from 1.7% for the 1993 vintage to 43.1% for the 2007 vintage and falls to 13.5% for 2010, reflecting the exposure of the 2005–2008 vintages to the Great Recession. The predictors are the number of employees, new versus existing business, created and retained jobs, franchise status, urban/rural indicator, revolving-credit and low-documentation flags, the logarithm of the approved amount, the guaranteed share, the amount per employee, whether the bank is located in the borrower's state, the industry sector (two-digit NAICS) and the borrower's state (59 features after encoding). The loan term is deliberately excluded. Among the 2004–2006 vintages, standard contractual maturities (60, 84, 120, 180, 240 or 300 months) account for 76.1% of repaid loans but only 2.0% of charged-off loans, whose recorded terms take irregular values such as 39, 43 or 51 months. The field is therefore populated after default and constitutes target leakage: with it, a LightGBM model reaches an out-of-time AUC of about 0.97, without it about 0.64–0.78. Disbursement-related fields are excluded for the same reason.

Lending Club loans. The Lending Club 2007–2011 file contains 39,786 personal loans that meet the platform's credit policy and have a final status (fully paid or charged off), so that outcomes are fully observed. Vintages are issue quarters from 2007Q3 to 2011Q4 (the single loan of 2007Q2 is merged into 2007Q3). The quarterly default rate ranges between 10.7% and 20.7%. The 67 predictors include amount, term, interest rate, instalment, sub-grade, employment length, income, debt-to-income ratio, delinquency and inquiry history, revolving balance and utilisation, credit-history length, loan-to-income and payment-to-income ratios, home ownership, verification status, purpose and state. Realised cash flows (total payments net of collection fees) are used only for the profit evaluation.

Table 1 Data Sets and Rolling-Origin Protocol

	SBA 7(a)	Lending Club
Loans (per run)	90,000 (5,000 per vintage)	39,786
Vintage unit and span	fiscal year, 1993–2010 (18)	quarter, 2007Q3–2011Q4 (18)
Default rate per vintage	1.7%–43.1%	10.7%–20.7%
Predictors	59	67
Training window of base learners	last 10 vintages	expanding
First out-of-time vintage	1996	2009Q1
Test periods	2001–2010 (10)	2010Q2–2011Q4 (7)
Independent runs	3 (seeds and subsamples)	3 (model seeds)

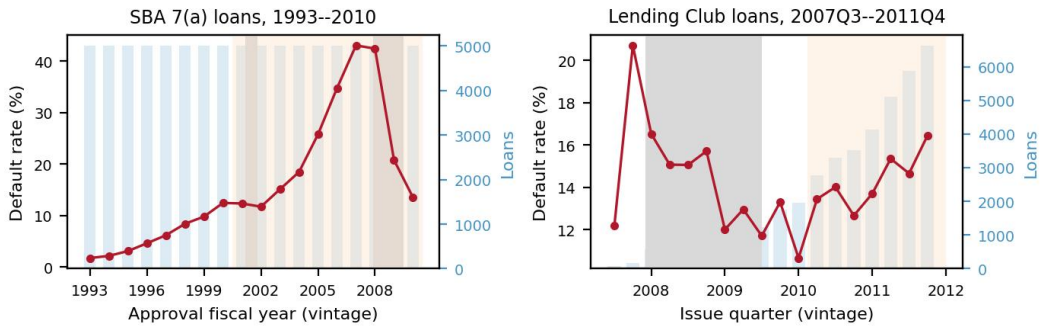


Figure 2 Default Rate (Red, Left Axis) and Number of Loans (Bars, Right Axis) per Vintage. Grey Bands Mark NBER Recessions (2001; December 2007–June 2009); the Shaded Orange Area Marks the Test Periods

5.2 Protocol

All methods are evaluated with the same rolling-origin protocol. At each test period T every model is trained only on vintages before T and evaluated on vintage T. For SBA the first models are trained on 1993–1995 and the test years are 2001–2010; for Lending Club the first models are trained on 2007Q3–2008Q4 and the test quarters are 2010Q2–2011Q4. Together with three independent runs this yields $3 \times (10 + 7) = 51$ test blocks. The aggregator of TRACE uses the cube entries of all labelled vintages with $R = 3$ available lags. In the main protocol, outcomes of all past vintages are treated as known at decision time, as is common in out-of-time credit scoring studies; Section 6.5 reports results with outcome delays of one and two vintages.

5.3 Base Learners and Benchmarks

The base learners are ℓ_2-regularised logistic regression (LR), random forest (RF, 200 trees, minimum leaf size 10), XGBoost (XGB, 300 trees, depth 4, learning rate 0.05) [16], LightGBM (LGBM, 300 trees, 31 leaves, learning rate 0.05) [17], CatBoost (CAT, 300 iterations, depth 6) [18] and a multilayer perceptron (MLP, two hidden layers of 64 and 32 units, early stopping). Hyper-parameters are fixed to commonly used values for all periods, so that no information from future periods enters model selection. The benchmarks comprise three groups. *Drift-adaptive methods*: LightGBM with exponential recency weights (LGBM-RW, half-life of two vintages); the accuracy-weighted ensemble of single-vintage LightGBM models (AWE) [27]; and LightGBM recalibrated by Platt scaling on its out-of-time predictions for the most recent vintage (LGBM-Platt). *Ensembles*: the average of the base-learner PDs (AVG); greedy ensemble selection with replacement on out-of-time log-loss (ES) [6, 39]; KNORA-Eliminate dynamic ensemble selection with the most recent vintage as dynamic selection set (KNORA-E) [28-29]; classical stacking with a logistic meta-learner trained on random three-fold out-of-fold predictions (STK-CV) [8, 37]; and logistic stacking trained on the out-of-time cube (STK-OOT), which isolates the effect of the information set. TRACE uses $\alpha = 0.5$, $\lambda = 0.01$, $R = 3$, $P = 3$ and $\mathcal{H} = \{0.25, 0.5, 1, 2, \infty\}$ (in vintage units) throughout.

5.4 Metrics and Statistical Tests

Discrimination is measured by the AUC. Probability quality is measured by the Brier score, the log-loss (LL) and the expected calibration error (ECE, ten equal-mass bins). Tail risk across the economic cycle is measured by the worst-period log-loss (wLL) and by the CVaR at level 0.3 of the period log-losses (cLL), i.e. the mean over the worst 30% of test periods. Decision quality is measured by the average misclassification cost of the Bayes rule with cost ratio $\rho = 5$ (Cost), additionally with $\rho \in \{3, 10\}$, by the charged-off principal accepted (SBA) and by the realised profit of the expected-profit rule (Lending Club). The PD bias of a period is the mean predicted PD minus the realised default rate. Following Demšar, methods are compared with the Friedman test over the 51 blocks and with one-sided Wilcoxon signed-rank tests of TRACE against each benchmark with Holm correction [40]. Stress periods are defined ex post as the SBA vintages 2005–2008 and the Lending Club quarters 2011Q2–2011Q4, i.e. the test periods with the highest default rates.

6 RESULTS

6.1 Overall Comparison

Table 2 reports the averages over test periods and runs. On the SBA data, TRACE achieves the lowest mean log-loss (0.4837), Brier score (0.1580), ECE (0.0572) and misclassification cost (0.6049) of all 15 methods, and the lowest worst-year and CVaR log-losses. Its AUC (0.6817) is within 0.001 of the best value (RF, 0.6825). Relative to classical stacking (STK-CV), the mean log-loss decreases by 2.5%, the worst-year log-loss by 11.0%, the CVaR of period losses by 8.2%, the calibration error by 27.2% and the misclassification cost by 6.8%. The two strongest benchmarks on the SBA data are LGBM-Platt and STK-OOT, both of which exploit out-of-time information; TRACE improves on them in every probability, tail and cost measure. A within-period bootstrap of the first run gives a mean log-loss difference between LGBM-Platt and TRACE of 0.0051 (95% interval 0.0031–0.0068).

On the Lending Club data, all ensembles are close. ES has the lowest mean log-loss (0.3762) and TRACE is within 0.001 of it (0.3771); TRACE shares the lowest worst-quarter log-loss (0.4087) with ES and STK-OOT, and all stacking-type ensembles clearly outperform the single learners and the drift-adaptive methods. The realised profit of the expected-profit rule differs by less than 0.3% among AVG, ES, STK-CV, STK-OOT and TRACE.

Table 2 Out-of-Time Performance Averaged over Test Periods and Three Runs. LL: Log-Loss; wLL: Worst-Period Log-Loss; cLL: CVaR $_{0.3}$ of Period Log-Losses; Cost: Bayes-Rule Misclassification Cost ($\rho = 5$); Profit: Realised Profit of the Expected-Profit Rule (USD Million, Summed over Test Quarters). Best Values in Bold

Method	AUC↑	Brier↓	LL↓	wLL↓	cLL↓	ECE↓	Cost↓	Profit↑

Method	AUC↑	Brier↓	LL↓	wLL↓	cLL↓	ECE↓	Cost↓	Profit↑
SBA 7(a) loans (10 test years × 3 runs)								
LR	0.6344	0.1712	0.5231	0.7582	0.7208	0.0928	0.7102	–
RF	**0.6825**	0.1630	0.4975	0.7219	0.6838	0.0861	0.6593	–
XGB	0.6688	0.1638	0.5066	0.7130	0.6745	0.0865	0.6694	–
LGBM	0.6705	0.1631	0.5072	0.7124	0.6720	0.0859	0.6703	–
CAT	0.6793	0.1619	0.4979	0.7156	0.6763	0.0818	0.6577	–
MLP	0.6531	0.1658	0.5130	0.7374	0.6984	0.0887	0.6970	–
LGBM-RW	0.6711	0.1613	0.5015	0.6819	0.6468	0.0827	0.6599	–
AWE	0.6684	0.1606	0.5044	0.6661	0.6397	0.0744	0.6388	–
LGBM-Platt	0.6705	0.1586	0.4884	0.6312	0.6154	0.0586	0.6350	–
AVG	0.6819	0.1624	0.4962	0.7178	0.6790	0.0815	0.6567	–
ES	0.6802	0.1625	0.4965	0.7145	0.6754	0.0827	0.6564	–
KNORA-E	0.6730	0.1636	0.5011	0.7194	0.6822	0.0848	0.6672	–
STK-CV	0.6801	0.1614	0.4960	0.7069	0.6682	0.0786	0.6488	–
STK-OOT	0.6818	0.1606	0.4898	0.6503	0.6253	0.0703	0.6296	–
TRACE (ours)	0.6817	**0.1580**	**0.4837**	**0.6294**	**0.6136**	**0.0572**	**0.6049**	–
Lending Club loans (7 test quarters × 3 runs)								
LR	0.7031	0.1152	0.3801	0.4103	0.3956	0.0168	0.5477	**43.54**
RF	0.7025	0.1149	0.3788	0.4112	0.3957	0.0158	0.5553	43.42
XGB	0.7010	0.1155	0.3804	0.4113	0.3967	0.0186	0.5560	43.32
LGBM	0.6908	0.1173	0.3883	0.4135	0.4036	0.0297	0.5690	43.10
CAT	0.6971	0.1162	0.3825	0.4133	0.3985	0.0206	0.5586	43.28
MLP	0.6741	0.1176	0.3905	0.4181	0.4074	0.0235	0.5786	43.30
LGBM-RW	0.6865	0.1185	0.3945	0.4169	0.4110	0.0382	0.5763	43.12
AWE	0.6611	0.1250	0.4479	0.5159	0.4759	0.0709	0.6158	42.67
LGBM-Platt	0.6908	0.1160	0.3831	0.4113	0.3981	0.0185	0.5762	43.43
AVG	0.7070	0.1145	0.3769	0.4092	0.3940	0.0140	0.5479	43.45
ES	0.7083	**0.1142**	**0.3762**	**0.4087**	**0.3933**	**0.0123**	0.5456	**43.54**
KNORA-E	0.6996	0.1152	0.3801	0.4105	0.3958	0.0164	0.5566	43.41
STK-CV	**0.7086**	0.1142	0.3764	0.4088	0.3934	0.0132	**0.5453**	43.50
STK-OOT	0.7080	0.1142	0.3768	0.4087	0.3933	0.0131	0.5485	43.49
TRACE (ours)	0.7078	0.1144	0.3771	**0.4087**	0.3936	0.0143	0.5466	43.45

Over all 51 blocks the Friedman test rejects equal performance for log-loss ($p = 1.1 \times 10^{-53}$), Brier score ($p = 1.3 \times 10^{-47}$) and AUC ($p = 3.5 \times 10^{-58}$). The average log-loss ranks place STK-CV (3.90), STK-OOT (4.18), TRACE (4.33) and ES (4.86) in a leading group, followed by AVG (5.98) and LGBM-Platt (7.37). Holm-corrected Wilcoxon tests show that TRACE is significantly better than every single learner, every drift-adaptive method and KNORA-E (all adjusted $p < 0.003$; e.g. 42 wins in 51 blocks against LGBM-Platt), whereas the per-block differences to STK-CV, ES and STK-OOT are not significant in rank terms (adjusted $p = 0.36$) although the mean log-loss difference favours TRACE by 0.0070 against STK-CV and 0.0072 against ES. The reason is the asymmetric profile analysed next: TRACE gains a lot in stress periods and pays a small premium in benign periods.

6.2 Behaviour across the Economic Cycle

Table 3 splits the results into stress and benign periods, and Figure 3 shows the period-by-period log-loss and PD bias of the first run. In the SBA stress vintages 2005–2008, RF, CAT, AVG, ES and STK-CV under-estimate the default rate by 12.7–14.2 percentage points. STK-OOT reduces the bias to -6.7 points, LGBM-Platt to -2.3 points and TRACE to -1.5 points, with the lowest stress-period log-loss (0.5921 versus 0.6359 for STK-CV). The reliability diagrams in Figure 4 make the difference visible for the 2007 vintage: the stacking ensembles predict PDs between 5% and 55% for loans whose realised default rates range from 14% to 63%, whereas TRACE is close to the diagonal. In the benign SBA years TRACE remains conservative (bias $+3.8$ points, mainly in the 2009–2010 recovery), which costs 0.015 in log-loss relative to the best benign model. On the Lending Club data, stress and benign periods differ much less, and TRACE has the smallest absolute stress bias (-0.37 points).

Table 3 Log-Loss and PD Bias (Percentage Points) in Stress and Benign Periods (Three Runs). Stress Periods: SBA 2005–2008, LC 2011Q2–Q4

	SBA				Lending Club			
	Stress		Benign		Stress		Benign	
Method	LL	Bias	LL	Bias	LL	Bias	LL	Bias
RF	0.6494	-14.22	**0.3963**	+0.71	0.3957	-0.81	0.3661	+0.27
CAT	0.6436	-12.98	0.4008	**+0.00**	0.3985	-1.21	0.3706	-0.37
LGBM-Platt	0.5933	-2.28	0.4184	+2.32	0.3981	-0.40	0.3719	-0.30
AVG	0.6454	-13.66	0.3967	+0.44	0.3940	-0.91	0.3641	**-0.05**
ES	0.6423	-13.58	0.3993	+0.12	**0.3933**	-0.56	**0.3635**	+0.40
STK-CV	0.6359	-12.68	0.4027	+0.11	0.3934	-0.42	0.3638	+0.60
STK-OOT	0.6015	-6.70	0.4154	+4.49	0.3933	-0.86	0.3643	-0.40

	SBA				Lending Club			
TRACE (ours)	**0.5921**	**-1.51**	0.4114	+3.84	0.3936	**-0.37**	0.3646	+0.19

Table 4 confirms this pattern statistically. In the 21 stress blocks, TRACE has a lower log-loss than STK-CV and ES in 14 blocks each, with mean improvements of 0.025 and 0.029 (Wilcoxon $p = 0.002$), and it outperforms AVG, LGBM-Platt, RF and CAT in 18–21 of the 21 blocks ($p < 0.001$). In the 30 benign blocks, the differences to the stacking ensembles are small and not significant in favour of either side (two-sided $p \geq 0.058$), while TRACE still outperforms LGBM-Platt in 24 of 30 blocks.

Table 4 TRACE Versus Benchmarks in Stress and Benign Blocks: Number of Blocks in Which TRACE Has the Lower Log-Loss, Mean Log-Loss Difference (Benchmark Minus TRACE) and One-Sided Wilcoxon p-Value

	Stress blocks			Benign blocks		
Benchmark	Wins	ΔLL	p	Wins	ΔLL	p
RF	21/21	+0.0337	<0.001	18/30	-0.0085	0.292
CAT	21/21	+0.0315	<0.001	23/30	-0.0040	0.099
LGBM-Platt	18/21	+0.0026	<0.001	24/30	+0.0071	<0.001
AVG	19/21	+0.0306	<0.001	13/30	-0.0090	0.768
ES	14/21	+0.0285	0.002	10/30	-0.0078	0.894
STK-CV	14/21	+0.0249	0.002	10/30	-0.0056	0.973
STK-OOT	11/21	+0.0052	0.073	12/30	+0.0023	0.444

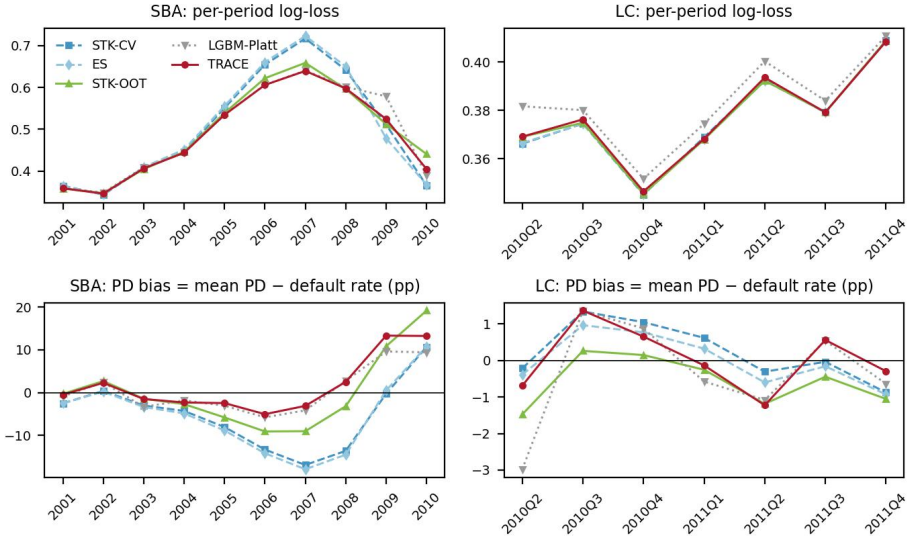


Figure 3 Per-Period Log-Loss (Top) and PD Bias (Bottom) of TRACE and the Strongest Ensembles (First Run)

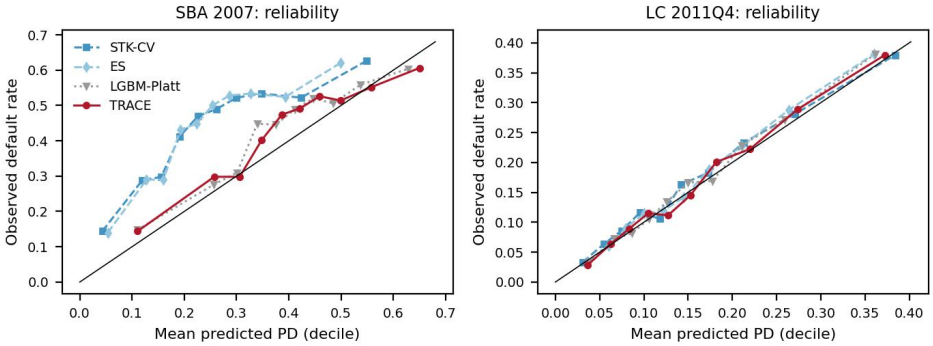


Figure 4 Reliability Diagrams (Deciles of Predicted PD) for the SBA Stress Vintage 2007 and the Lending Club Quarter 2011Q4 (First Run)

6.3 Ablation Study

Table 5 removes or replaces one component at a time. Removing the adaptive regime memory ($h = \infty$) has the largest effect on the SBA data: the mean log-loss increases from 0.4837 to 0.4909, the worst-year log-loss from 0.6294 to 0.6443, the misclassification cost from 0.6049 to 0.6315, and the stress-period bias deteriorates from -1.5 to -5.1 points. Removing the epistemic margins increases the mean log-loss to 0.4873, the ECE to 0.0602 and the cost to 0.6184; each margin alone recovers part of the effect, and the cross-learner margin mainly improves the average while

the cross-regime margin mainly improves the tail. Removing the CVaR objective ($\alpha = 1$) slightly improves the mean log-loss (0.4812) but increases the worst-year and CVaR log-losses and the stress under-estimation (-2.0 points); this is the expected trade-off of a risk-aware objective. Replacing the out-of-time cube by random out-of-fold predictions also improves the average slightly but deteriorates the tail (worst-year log-loss 0.6387) and triples the stress bias (-5.0 points), because the aggregator no longer sees how base learners behave when they extrapolate. Removing memory, CVaR and margins together yields vintage-balanced stacking on the cube, which is clearly worse in all tail measures (worst-year log-loss 0.6558, stress bias -7.5 points); on the SBA data it coincides with pooled stacking because all vintages have the same size. A regret-based CVaR (vintage losses measured relative to the best vintage-specific aggregator) and an additional signed drift feature give no consistent improvement, and the worst-vintage objective ($\alpha \to 0$) is overly conservative (mean log-loss 0.5044). On the Lending Club data all variants except the worst-vintage objective lie within 0.0008 of each other, which is consistent with the mild drift of this portfolio.

Table 5 Ablation Study (Means over Three Runs). sBias: Mean PD Bias in SBA Stress Periods (Percentage Points)

	SBA						Lending Club	
Variant	LL	wLL	cLL	ECE	Cost	sBias	LL	wLL
TRACE (full)	0.4837	0.6294	0.6136	0.0572	0.6049	-1.51	0.3771	0.4087
w/o CVaR ($\alpha = 1$)	0.4812	0.6301	0.6146	0.0561	0.6040	-2.01	0.3768	0.4087
w/o regime memory ($h = \infty$)	0.4909	0.6443	0.6206	0.0674	0.6315	-5.06	0.3770	0.4088
w/o margins ($\kappa = 0$)	0.4873	0.6302	0.6141	0.0602	0.6184	-1.46	0.3768	0.4086
margins: u^{mod} only	0.4838	0.6302	0.6142	0.0561	0.6099	-1.48	0.3770	0.4086
margins: u^{reg} only	0.4880	0.6295	0.6139	0.0590	0.6062	-1.58	0.3769	0.4086
random 3-fold OOF inputs	0.4811	0.6387	0.6196	0.0561	0.6121	-5.03	0.3775	0.4087
w/o memory, CVaR, margins	0.4907	0.6558	0.6286	0.0715	0.6290	-7.51	0.3767	0.4089
pooled ERM on OOT inputs	0.4907	0.6558	0.6286	0.0715	0.6290	-7.51	0.3767	0.4088
regret-based CVaR	0.4833	0.6326	0.6149	0.0582	0.6067	-2.36	0.3771	0.4086
worst vintage ($\alpha \to 0$)	0.5044	0.6322	0.6144	0.0837	0.6454	+0.28	0.3785	0.4104
+ signed drift feature	0.4883	0.6290	0.6124	0.0576	0.6040	-1.52	0.3772	0.4087

6.4 Sensitivity Analysis

Risk appetite. Table 6 and Figure 5 trace the effect of α (first run). On the SBA data, decreasing α from 1 to 0.1 moves the ensemble along a risk–return frontier: the CVaR of period log-losses stays at 0.614–0.615 and the worst-year log-loss at 0.6395–0.6400, while the mean log-loss increases from 0.4829 to 0.4996. The default $\alpha = 0.5$ attains the minimum tail losses at a mean log-loss of 0.4862. Every point of the TRACE path with $\alpha \geq 0.5$ dominates all benchmarks, including LGBM-Platt, in both mean and tail loss. The limiting worst-vintage objective is clearly inferior. On the Lending Club data the frontier is flat: all $\alpha \geq 0.1$ give mean log-losses between 0.3767 and 0.3774.

Table 6 Sensitivity to the Risk Appetite α (First Run)

	SBA				Lending Club			
α	LL	cLL	wLL	ECE	LL	cLL	wLL	ECE
$\to 0$ (max)	0.5118	0.6148	0.6431	0.0903	0.3788	0.3960	0.4113	0.0221
0.1	0.4996	0.6144	0.6395	0.0661	0.3773	0.3937	0.4084	0.0153
0.2	0.4915	0.6144	0.6395	0.0603	0.3774	0.3937	0.4084	0.0159
0.3	0.4914	0.6144	0.6395	0.0593	0.3773	0.3937	0.4084	0.0153
0.5 (default)	0.4862	0.6144	0.6395	0.0559	0.3773	0.3937	0.4084	0.0147
0.7	0.4843	0.6154	0.6395	0.0573	0.3770	0.3937	0.4087	0.0137
1	0.4829	0.6153	0.6400	0.0547	0.3767	0.3937	0.4085	0.0143

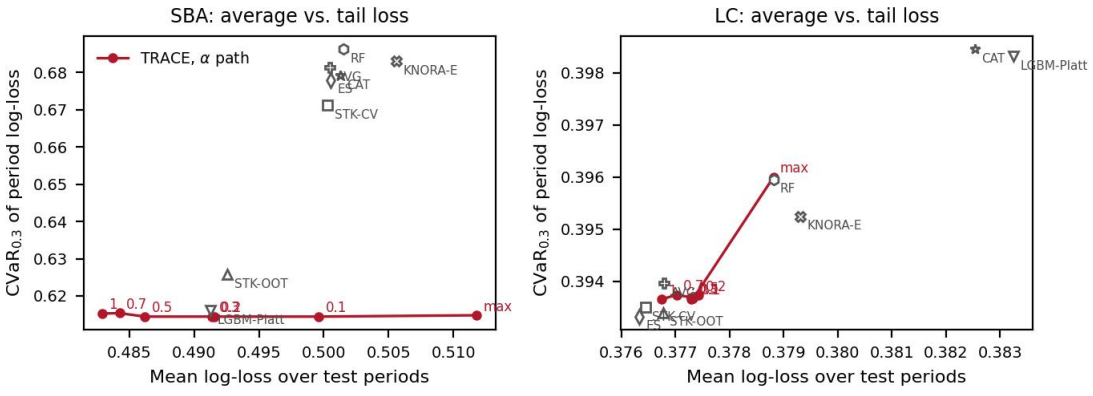


Figure 5 Average Versus Tail Loss. The Red Path Shows TRACE for $\alpha \in \{1, 0.7, 0.5, 0.3, 0.2, 0.1\}$ and the Worst-Vintage Limit ("max"); Grey Markers Show Benchmarks (First Run)

Other hyper-parameters. Figure 6 reports the mean log-loss relative to the best setting for fixed half-lives, the ridge parameter λ, the number of lags R and the validation window P. On the SBA data, the best fixed half-life in hindsight ($h = 0.25$, mean log-loss 0.4852) is matched by the data-driven selection (0.4862) without access to test information, whereas fixed half-lives between 0.5 and ∞ yield 0.4943–0.4990. The results are insensitive to $\lambda \in [0,1]$ (0.4856–0.4891) and to $P \in \{1,2,3,5\}$ (0.4862–0.4878). At least three lags are needed for an informative cross-regime margin ($R = 2$: 0.4981; $R = 3$: 0.4865; $R = 4$: 0.4848). On the Lending Club data, all settings lie within 0.2% of each other.

Base-learner pool. Dropping any single learner changes the SBA mean log-loss by at most 0.005 (range 0.4815–0.4875) and the Lending Club mean log-loss by at most 0.0005. Restricting the pool to the three gradient-boosting models deteriorates both data sets (SBA 0.4943, Lending Club 0.3793), which shows that heterogeneity of the pool is valuable for the risk-aware aggregator.

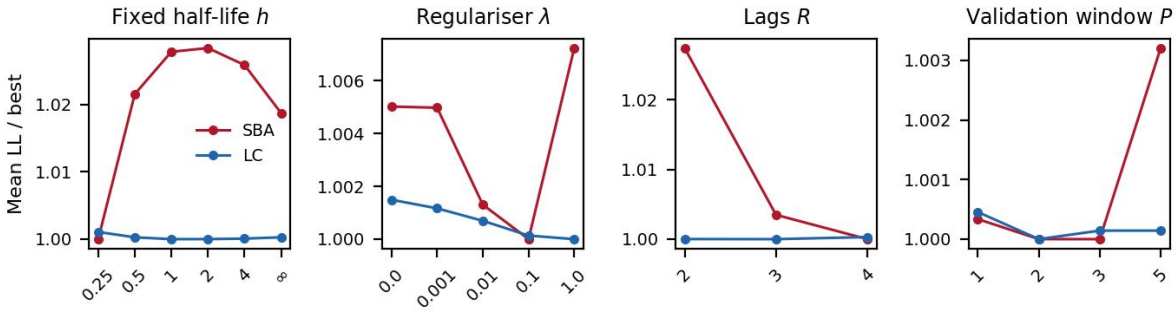


Figure 6 Sensitivity of the Mean Log-Loss (Relative to the Best Setting) to Fixed Half-Lives h, the Ridge Parameter λ, the Number of Lags R and the Validation Window P (First Run)

6.5 Robustness to Outcome Delay

In practice the outcome of a loan is observed only some time after origination. We therefore repeat the evaluation with an outcome delay of $g \in \{1,2\}$ vintages: at decision time T only vintages up to $T - 1 - g$ are labelled, base-learner predictions for vintage T are $(g + 1)$-step-ahead forecasts, and TRACE selects its memory on the labelled vintages only. Benchmarks that require re-training (STK-CV, LGBM-RW, AWE) are omitted from this analysis. Table 7 shows that TRACE retains the lowest worst-year log-loss on the SBA data at both delays (0.6733 and 0.7392, versus 0.6894 and 0.7510 for LGBM-Platt and 0.7778 and 0.8360 for ES), while its mean log-loss is within 0.7% (delay 1) and 1.6% (delay 2) of the best benchmark (STK-OOT). On the Lending Club data, ES and AVG have slightly lower mean log-losses at longer delays (differences of 0.001–0.002), and TRACE remains within 0.6% of the best method.

Table 7 Mean (LL) and Worst-Period (wLL) Log-Loss under Outcome Delays of $g = 0,1,2$ Vintages (First Run; Periods with Sufficient Labelled History)

	$g = 0$		$g = 1$		$g = 2$	
Method	LL	wLL	LL	wLL	LL	wLL
			SBA			
RF	0.5015	0.7300	0.5251	0.7825	0.5917	0.8385
LGBM-Platt	0.4913	0.6405	0.5249	0.6894	0.5910	0.7510
AVG	0.5005	0.7247	0.5269	0.7844	0.5978	0.8378
ES	0.5005	0.7232	0.5276	0.7778	0.5951	0.8360
STK-OOT	0.4926	0.6588	**0.5207**	0.6953	**0.5827**	0.7572
TRACE (ours)	**0.4862**	**0.6395**	0.5244	**0.6733**	0.5918	**0.7392**
			Lending Club			
RF	0.3788	0.4112	0.3806	0.4132	0.3838	0.4142
LGBM-Platt	0.3833	0.4107	0.3855	0.4124	0.3886	0.4158
AVG	0.3768	0.4088	0.3789	0.4102	**0.3820**	0.4118
ES	**0.3763**	0.4087	**0.3788**	**0.4095**	**0.3820**	**0.4106**
STK-OOT	0.3768	0.4087	0.3803	0.4098	0.3849	0.4114
TRACE (ours)	0.3773	**0.4084**	0.3798	0.4096	0.3840	0.4128

6.6 Decision-Level Evaluation

Table 8 evaluates the PDs in lending decisions (first run). On the SBA data, TRACE has the lowest average misclassification cost for all cost ratios ($\rho = 3$: 0.4859; $\rho = 5$: 0.6066; $\rho = 10$: 0.7013), 8.4% below ES and 7.4% below STK-CV at $\rho = 5$. Because its PDs are not biased downwards in the stress vintages, the Bayes rule accepts 40.1% of the applications instead of about 55% for the stacking ensembles, and the charged-off principal accepted over the ten test years falls from USD 374.5 million (STK-CV) and 385.7 million (ES) to USD 192.3 million. STK-OOT has

the lowest worst-year cost, followed by TRACE. On the Lending Club data the realised profit of the expected-profit rule lies between USD 43.37 million (KNORA-E) and 43.55 million (ES) for the ensembles (TRACE: 43.44 million), and the return on funded capital lies between 12.33% and 12.46% for all methods; under mild drift the decision value of the ensembles is practically identical.

Table 8 Decision-Level Results (First Run). SBA: Average and Worst-Year (wCost $_5$) Bayes-Rule Misclassification Cost for Cost Ratios ρ, Charged-Off Principal Accepted (USD Million) and Approval Rate (%) at $\rho = 5$. Lending Club: Realised Profit (USD Million), Return on Funded Capital (ROI, %) and Lowest Quarterly ROI (%) of the Expected-Profit Rule

	SBA						Lending Club		
Method	Cost $_3$	Cost $_5$	Cost $_{10}$	wCost $_5$	Charge-off	Appr.	Profit	ROI	min ROI
LGBM	0.5356	0.6705	0.7754	0.7758	378.1	55.5	43.20	**12.46**	11.20
CAT	0.5265	0.6581	0.7589	0.7816	385.9	55.5	43.24	12.41	11.23
LGBM-Platt	0.4863	0.6402	0.7291	0.7786	201.4	38.6	43.39	12.33	11.00
AVG	0.5331	0.6646	0.7220	0.7618	390.6	54.3	43.44	12.35	11.23
ES	0.5334	0.6619	0.7314	0.7604	385.7	55.1	**43.55**	12.36	**11.42**
KNORA-E	0.5412	0.6746	0.7464	0.7960	406.5	55.7	43.37	12.36	11.21
STK-CV	0.5266	0.6551	0.7418	0.7584	374.5	54.5	43.45	12.37	11.26
STK-OOT	0.5104	0.6309	0.7128	**0.7220**	233.9	39.7	43.52	12.35	11.33
TRACE (ours)	**0.4859**	**0.6066**	**0.7013**	0.7464	**192.3**	40.1	43.44	12.34	11.12

6.7 What Does TRACE Learn?

Figure 7 shows the selected half-life and the aggregation parameters over time (first run). On the SBA data the validation procedure keeps an infinite memory while the default rate drifts slowly (2001–2004), shortens it to one year when the default rate starts to accelerate (2005–2007), to half a year at the peak of the crisis (2008–2009) and to a quarter of a year in the recovery (2010). The regime memory is thus an interpretable, data-driven indicator of a regime change. The ensemble weights remain diversified across learners, with the largest weights on the tree ensembles. The cross-regime margin receives a large weight after the crisis (2008–2010), when models trained before and after the crisis disagree most; this produces the conservative PDs in the recovery years discussed above. On the Lending Club data the memory switches between short and long horizons without a persistent pattern, and logistic regression and random forest receive the largest weights.

The cross-regime margin identifies loans whose risk is under-estimated by a plain ensemble. Sorting the SBA test loans of each period into quintiles of u^{reg}, the gap between the realised default rate and the AVG ensemble PD increases monotonically from 3.1 percentage points in the lowest quintile to 10.2 points in the highest quintile, while the corresponding gap for TRACE stays between -2.4 and -0.5 points. On the Lending Club data the gap of AVG is small in all quintiles (-0.1 to 0.8 points), consistent with the mild drift of this portfolio.

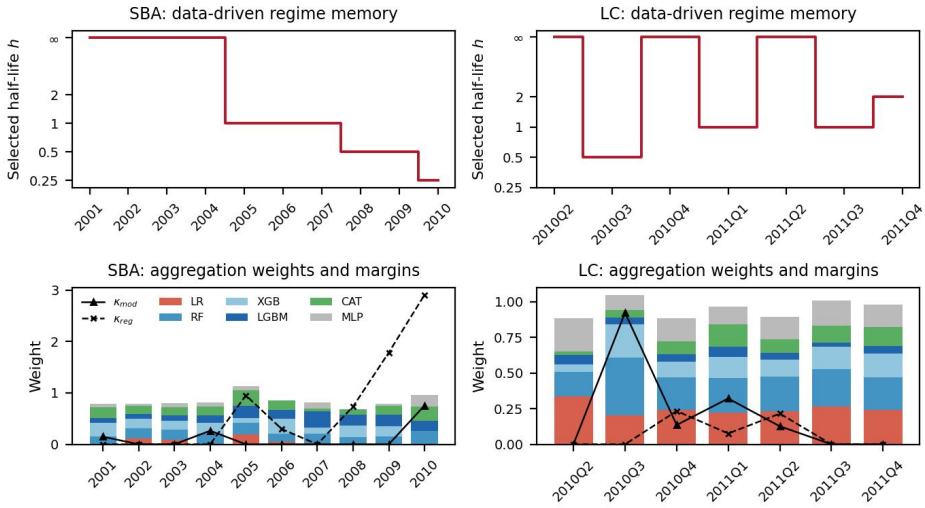


Figure 7 Data-Driven Regime Memory (Top) and Aggregation Parameters (Bottom): Stacked Bars Show the Learner Weights w, Lines the Margin Coefficients κ_{mod} and κ_{reg} (First Run)

7 DISCUSSION

Where the gains come from. The experiments separate three effects. The first is the information set of the aggregator. Out-of-time predictions reveal the systematic error that base learners make when they are applied to a new vintage; in a

period of rising defaults this error is an under-estimation, and an aggregator trained on the cube learns to correct it. The second is the regime memory. When the economy changes, the most recent vintages contain the most relevant information, and selecting the half-life by one-step-ahead validation lets the data decide how quickly old regimes are forgotten. The third is the risk-aware objective. CVaR aggregation trades a small amount of average accuracy for lower tail losses, and the epistemic margins add conservatism where models disagree. The combination yields the profile that a risk manager wants: large gains when conditions deteriorate and limited costs when conditions are benign.

Implications for credit risk management. Prudential practice requires PD estimates to remain adequate over the economic cycle and to contain a margin of conservatism that reflects estimation uncertainty. TRACE offers a transparent way to operationalise both requirements within a machine-learning ensemble: the risk appetite α is an explicit, auditable parameter; Proposition 1 provides a certificate on the expected loss under adverse regime mixtures; the margins are interpretable in terms of model and regime disagreement; and the selected regime memory documents when the model judged recent data to be more relevant than the long history. Because all components operate on the predictions of standard base learners, TRACE can be added to an existing scoring pipeline without changing the underlying models.

Mild versus severe drift. The two data sets illustrate the range of conditions under which a lender operates. The Lending Club portfolio, originated during and after the crisis by a rapidly growing platform, shows moderate changes in default rates, and all strong ensembles achieve almost identical accuracy and profit. The SBA portfolio contains vintages that were originated in a benign environment and then exposed to a severe recession, and there the aggregation rule makes a large difference. A risk-aware ensemble is designed for the second situation and performs on par with the best benchmarks in the first.

Extensions. The framework accommodates several extensions. Macroeconomic indicators or uncertainty indices can be used to tilt the nominal regime distribution towards historical vintages with similar conditions [4-5]; lifetime PD models can be obtained by applying the aggregation to survival-type base learners [1]; and the vintage losses can be replaced by profit-based losses to align the risk measure with the lender's economic objective [35].

8 CONCLUSION

This paper has proposed TRACE, a risk-aware ensemble for credit default prediction under economic uncertainty. Economic uncertainty is modelled as ambiguity about which historical regime the next loan vintage will resemble, and the ensemble is aggregated by minimising the conditional value-at-risk of vintage losses around a recency-anchored nominal distribution. Out-of-time prediction cubes, epistemic margins and a data-driven regime memory complete the method, and the aggregation problem is convex and solved exactly. On 90,000 SBA loans spanning the Great Recession, TRACE improves every probability-quality, tail-risk and decision-cost measure relative to 14 benchmarks, reduces the worst-year log-loss of classical stacking by 11.0% and nearly removes the PD under-estimation of the crisis vintages. On 39,786 Lending Club loans with mild drift it matches the strongest ensembles. The adaptive regime memory is the most important component, the CVaR objective controls tail risk, and the learned memory provides an interpretable signal of regime change. These properties make risk-aware aggregation a practical complement to the ensembles that already dominate credit scoring.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

DATA AVAILABILITY

The SBA national loan data set described by Li et al. and the Lending Club 2007–2011 loan file (LoanStats3a) are publicly available [38]. The code that reproduces all experiments, tables and figures, including the data preparation, is provided as supplementary material.

REFERENCES

[1] Bellotti T, Crook J. Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 2013, 29: 563-574.

[2] Leow M, Crook J. The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. European Journal of Operational Research, 2016, 249: 457-464. DOI: 10.1016/j.ejor.2014.09.005.

[3] Bloom N. Fluctuations in uncertainty. Journal of Economic Perspectives, 2014, 28: 153-176. DOI: 10.1257/jep.28.2.153.

[4] Jurado K, Ludvigson S C, Ng S. Measuring uncertainty. American Economic Review, 2015, 105: 1177-1216. DOI: 10.1257/aer.20131193.

[5] Baker S R, Bloom N, Davis S J. Measuring economic policy uncertainty. The Quarterly Journal of Economics, 2016, 131: 1593-1636. DOI: 10.1093/qje/qjw024.

[6] Lessmann S, Baesens B, Seow H-V, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 2015, 247: 124-136. DOI: 10.1016/j.ejor.2015.05.030.
[7] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in neural information processing systems 35 (NeurIPS 2022), datasets and benchmarks track, 2022.
[8] Xia Y, Liu C, Da B, et al. A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Systems with Applications, 2018, 93: 182-199. DOI: 10.1016/j.eswa.2017.10.022.
[9] Abellán J, Castellano J G. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 2017, 73: 1-10.
[10] Gunnarsson B R, Broucke S vanden, Baesens B, et al. Deep learning for credit scoring: Do or don't? European Journal of Operational Research, 2021, 295: 292-305.
[11] Dumitrescu E, Hué S, Hurlin C, et al. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 2022, 297: 1178-1192. DOI: 10.1016/j.ejor.2021.06.053.
[12] Rockafellar R T, Uryasev S. Optimization of conditional value-at-risk. Journal of Risk, 2000, 2: 21-41.
[13] Duchi J C, Namkoong H. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics, 2021, 49: 1378-1406. DOI: 10.1214/20-AOS2004.
[14] Ala'raj M, Abbod M F. Classifiers consensus system approach for credit scoring. Knowledge-Based Systems, 2016, 104: 89-105.
[15] Xia Y, Zhao J, He L, et al. A novel tree-based dynamic heterogeneous ensemble method for credit scoring. Expert Systems with Applications, 2020, 159: 113615. DOI: 10.1016/j.eswa.2020.113615.
[16] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016: 785-794.
[17] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems 30 (NIPS 2017), 2017: 3146-3154.
[18] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: Unbiased boosting with categorical features. In: Advances in neural information processing systems 31 (NeurIPS 2018), 2018.
[19] Malekipirbazari M, Aksakalli V. Risk assessment in social lending via random forests. Expert Systems with Applications, 2015, 42: 4621-4631. DOI: 10.1016/j.eswa.2015.02.001.
[20] Emekter R, Tu Y, Jirasakuldech B, et al. Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. Applied Economics, 2015, 47: 54-70. DOI: 10.1080/00036846.2014.962222.
[21] Jagtiani J, Lemieux C. The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. Financial Management, 2019, 48: 1009-1029.
[22] Brown I, Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2012, 39: 3446-3453.
[23] Fitzpatrick T, Mues C. An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. European Journal of Operational Research, 2016, 249: 427-439. DOI: 10.1016/j.ejor.2015.09.014.
[24] Bellotti T, Crook J. Retail credit stress testing using a discrete hazard model with macroeconomic factors. Journal of the Operational Research Society, 2014, 65: 340-350. DOI: 10.1057/jors.2013.91.
[25] Sousa M R, Gama J, Brandão E. A new dynamic modeling framework for credit risk assessment. Expert Systems with Applications, 2016, 45: 341-351. DOI: 10.1016/j.eswa.2015.09.055.
[26] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46: 44:1-44:37. DOI: 10.1145/2523813.
[27] Wang H, Fan W, Yu P S, et al. Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, 2003: 226-235.
[28] Ko A H R, Sabourin R, Britto Jr. A S. From dynamic classifier selection to dynamic ensemble selection. Pattern Recognition, 2008, 41: 1718-1731.
[29] Cruz R M O, Sabourin R, Cavalcanti G D C. Dynamic classifier selection: Recent advances and perspectives. Information Fusion, 2018, 41: 195-216.
[30] Sagawa S, Koh P W, Hashimoto T B, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International conference on learning representations (ICLR), 2020.
[31] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems 30 (NIPS 2017), 2017: 6402-6413.
[32] Malinin A, Prokhorenkova L, Ustimenko A. Uncertainty in gradient boosting via ensembles. In: International conference on learning representations (ICLR), 2021.
[33] Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in neural information processing systems 32 (NeurIPS 2019), 2019.
[34] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. In: Proceedings of the 34th international conference on machine learning (ICML), 2017: 1321-1330.

[35] Verbraken T, Bravo C, Weber R, et al. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, 2014, 238: 505-513. DOI: 10.1016/j.ejor.2014.04.001.
[36] Serrano-Cinca C, Gutiérrez-Nieto B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, 2016, 89: 113-122. DOI: 10.1016/j.dss.2016.06.014.
[37] Wolpert D H. Stacked generalization. Neural Networks, 1992, 5: 241-259.
[38] Li M, Mickel A, Taylor S. "Should This Loan be Approved or Denied?": A large dataset with class assignment guidelines. Journal of Statistics Education, 2018, 26: 55-66. DOI: 10.1080/10691898.2018.1434342.
[39] Caruana R, Niculescu-Mizil A, Crew G, et al. Ensemble selection from libraries of models. In: Proceedings of the 21st international conference on machine learning (ICML), 2004.
[40] Demšar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 2006, 7: 1-30.