How High Does Your Gini Need to Be? Deriving Minimum Scorecard Discrimination from Business KPIs

Minh Tran[a,*] and Mai Nguyen[b]

[a] Modeling Center, Mcredit, Hanoi, Vietnam. minhtc.ho@mcredit.com.vn

[b] Risk Management Division, Mcredit, Hanoi, Vietnam. maintn3.ho@mcredit.com.vn

[*] Corresponding author.

Working paper submitted to the

Credit Scoring and Credit Control Conference

Credit Research Centre, University of Edinburgh Business School

August 2026

Declarations: The authors declare no conflicts of interest. The views expressed are those of the authors and do not necessarily reflect those of Mcredit.

How High Does Your Gini Need to Be?
Deriving Minimum Scorecard Discrimination from Business KPIs

Abstract

Credit scorecards are almost universally evaluated with the Gini coefficient, yet a Gini number in isolation cannot tell a lender whether a model is good enough to meet the business objectives it will be deployed against. We invert the usual evaluation logic and ask a design-time question: given a stated business KPI, what is the minimum Gini a scorecard must achieve for that KPI to be attainable? Under a one-parameter score-separation model we link the Gini to the full accept/reject decision economics and solve the inversion for two KPI families: operating-point targets (approval rate and booked bad-rate ceiling) and profit targets; under the bi-normal model the operating-point requirement admits an exact closed form. Two layers absent from prior work are added: a robustness analysis that recomputes every requirement under bi-normal (equal and unequal variance), bi-gamma and skew-normal score distributions, and a probabilistic layer that converts the sampling variance of the AUC into the probability that a scorecard clears its required floor. On the UCI German Credit data, a weight-of-evidence logistic scorecard achieves a Gini of 0.602. Against a representative operating KPI at a 10% through-the-door bad rate the required floor is 0.666 and the scorecard falls short (clearance probability 0.15); against a profit KPI the floor is 0.536 and it passes (0.86). A score-degradation experiment and a replication on the Taiwan credit-card dataset corroborate the machinery. The framework gives risk managers a defensible go/no-go bar before a single model is built.

Keywords: credit scoring; Gini coefficient; discriminatory power; profit scoring; model validation; scorecard design

1 Introduction

For over two decades, consumer credit scoring has been organised around one dominant performance dimension: how well a model ranks good borrowers above bad ones. Since the foundational reviews of Hand and Henley (1997) and Thomas (2000), the Gini coefficient and its monotone transform the AUC have been the reporting standard, and the definitive benchmarking studies of Baesens et al. (2003) and Lessmann et al. (2015) compare dozens of algorithms almost entirely through this lens. A scorecard is quoted by its Gini in model documentation, in validation reports to regulators, in vendor comparisons, and in the internal sign-off that lets a model go live. The number has become a lingua franca precisely because it is a single, dimensionless, monotone summary of rank quality that can be compared across portfolios and across time.

Yet this very convenience conceals a gap that every risk manager confronts in practice. The usual workflow is diagnostic and *post hoc*: a model is built, its Gini is measured, and only then is the question asked whether the number is acceptable. The answer is almost always given by analogy—"0.5 is typical for this product," "last year's card was 0.55," "the vendor benchmark is 0.6"—rather than by derivation. Acceptability, however, is not a property of the Gini alone. Hand (2009) showed that the AUC is an incoherent performance summary because it implicitly averages misclassification costs using a weighting that depends on the classifier itself, so two models are effectively judged on different cost assumptions and a Gini value in isolation does not encode the lender's economics. Whether a Gini of, say, 0.55 is "good enough" cannot be answered without stating what the model is for: the same 0.55 may comfortably clear a conservative, low-approval lending plan and yet fall well short of an aggressive growth target on a riskier population.

A body of work has connected discrimination to lending economics—most directly Stein (2005) and Blöchlinger and Leippold (2006), who map the ROC curve of a *given* model into cutoffs, prices and profits, and the profit-scoring literature (Verbraken et al., 2014; Petrides et al., 2022; Maldonado et al., 2017), which chooses the operating cutoff that maximises value. All of this runs in the *forward* direction: it takes a model as input and asks how it should be used and what it is worth. It does not answer the design-time question a risk manager must settle *before* committing budget, data, and modelling effort: given the business plan the scorecard must serve—a target approval rate and a bad-rate ceiling, or a profit-per-account target—what discrimination is even necessary? Setting that bar too low guarantees a model that cannot meet the plan no matter how it is deployed; setting it too high wastes effort chasing lift the business does not need. Today that bar is set by intuition. We argue it can be derived.

We supply that derivation as an *inversion*: instead of mapping a model to a Gini and then judging it, we map a business KPI directly to the minimum Gini that makes the KPI attainable at all. Our contributions are:

1. A mapping from a business KPI to the minimum required Gini, under a one-parameter

score-separation model, for both operating-point KPIs (approval rate and bad-rate ceiling) and profit KPIs, with monotonicity arguments (Appendix A) that guarantee the floor is well-defined and unique. Under the bi-normal model the operating-point floor is available in exact closed form (Equation (9)).

2. A robustness analysis showing the requirement is stable across bi-normal, bi-gamma and skew-normal score distributions calibrated to the same Gini—including sweeps of the families' shape parameters and an unequal-variance bi-normal variant—so the floor does not hinge on an arbitrary shape assumption.

3. A probabilistic layer that turns a point Gini into a probability of clearing the floor, using the Hanley and McNeil (1982) closed-form AUC variance (with the nonparametric DeLong et al. (1988) estimator as the recommended production alternative), making the finite-sample uncertainty of the verdict explicit.

4. An empirical study on the real UCI German Credit dataset, yielding an actionable split verdict—a pass on the profit KPI and a shortfall on the operating KPI for the same scorecard—that a bare discrimination number could never deliver, together with a data-driven check of the inversion curve itself and a replication design for a larger book.

The remainder of the paper is organised as follows. Section 1.1 positions the work against the discrimination, profit-scoring, modelling, and reject-inference literatures. Section 2 develops the framework: the decision model, the separation-model linkage, the two inversion operators, and the robustness and uncertainty layers. Section 3 applies the framework end-to-end on real data. Section 4 discusses implications, limitations, and extensions, and Section 5 concludes.

1.1 Related work

Our framework sits at the intersection of four established literatures. We review each in turn and, for each, state precisely what we take from it and where we depart.

Discrimination measures and their validation. The Gini coefficient and the AUC are the currency of scorecard reporting (Hand and Henley, 1997; Baesens et al., 2003; Lessmann et al., 2015). Their appeal is that they summarise the entire ROC curve in one number that is invariant to the score's monotone re-scaling and to the class prior, which makes them portable across products and vintages. This portability, however, is exactly what severs the number from the decision it is meant to inform. Hand (2009) demonstrated the AUC's incoherence as a cost-sensitive summary—because the implicit cost weighting depends on the classifier's own score distribution—and proposed the H-measure with an explicit, fixed cost prior as a coherent alternative. The Basel validation literature likewise treats power statistics as estimates whose sampling error matters for sign-off: Engelmann et al. (2003) develop tests for the accuracy ratio, and the Basel Committee's study of rating-system validation (BCBS, 2005) catalogues

discriminatory-power measures and their confidence intervals. We share Hand's diagnosis that a rank statistic alone cannot certify a model, but we take a different remedy: rather than replacing the AUC with a better scalar, we *keep* the Gini—because it is what practitioners already report and regulators already expect—and supply the missing economics by deriving, for a stated KPI, the Gini value at which the decision becomes feasible; and we inherit the validation literature's insistence on sampling uncertainty through the clearance probability of Section 2.7.

Linking discrimination to lending economics. Two strands make the power–profit connection explicit, and both are close ancestors of this paper. The first maps a given model's ROC curve into decisions and value: Stein (2005) integrates ROC analysis with loan pricing, derives profit-maximising cutoff rules, and shows that (uniformly) more powerful models are more profitable; Blöchlinger and Leippold (2006) derive the profit-maximising cutoff regime and pricing curve from the ROC and quantify the economic benefit of more powerful scoring in a competitive loan market. The second strand reframes model evaluation and construction around business value: Verbraken et al. (2014) introduce the Expected Maximum Profit measure, which integrates over a loss-given-default distribution to select the profit-maximising fraction to accept; Petrides et al. (2022) cast scoring as cost-sensitive learning, and Maldonado et al. (2017) embed profit directly into feature selection and training. This literature establishes two premises we adopt wholesale: the cutoff and the economics—not the ranking alone—determine realised outcomes, and value is monotone in discriminatory power. But it operates in the *forward* direction: it takes a fixed model (or fixed model class) and finds its best operating point and its value. Ours is the dual problem. We fix the operating point or value implied by the KPI and ask for the minimum model quality that makes it achievable, so the two directions are complementary—theirs optimises deployment of a model in hand, ours specifies the design target that deployment will later optimise against. Prior work maps power forward into profit; to our knowledge no prior work runs the map in reverse as a design-time requirement, combines it with an explicit distributional-robustness range, and attaches a sampling-based clearance probability, which is the gap this paper fills.

Modelling algorithms. The methods used to build scorecards have broadened well beyond logistic regression—support vector machines (Bellotti and Crook, 2009), machine-learning ensembles applied across origination and collections (Khandani et al., 2010; Tran and Tran, 2025), deep learning (Gunnarsson et al., 2021), tree-augmented logistic models that preserve interpretability (Dumitrescu et al., 2022), and models exploiting alternative and big data sources (Djeundje et al., 2021; Óskarsdóttir et al., 2019). Across all of them, achieved lift is reported in Gini/AUC, and Finlay (2011) surveys how classifier architectures trade off against one another. This breadth matters to us in a specific way: whatever technology a lender adopts, the *output* our framework consumes is the same—an achieved Gini—so the inversion is agnostic to how the scorecard was built. It tells any modelling team, regardless of algorithm, what target to aim for. We anchor our empirical study on a weight-of-evidence logistic scorecard because it remains the interpretable, regulated industry baseline, but nothing in the framework depends

on that choice.

Decision population and uncertainty. Two further issues, often treated as technicalities, are first-order for an inversion and we therefore fold them into the framework rather than the discussion. First, scorecards decide who is booked, so the achieved Gini is measured on *accepted* applicants whereas the KPI is defined over the *through-the-door* population; these differ because of prior accept/reject filtering—the reject-inference problem studied by Crook and Banasik (2004) and Banasik and Crook (2007). Because our floor depends explicitly on the through-the-door bad rate, the choice of that rate (and hence reject inference) is not a footnote but the single most influential input, as our sensitivity analysis shows; the related question of whether a Gini measured on one population can be carried to another is made an explicit assumption in Section 2.8. Second, a Gini is an estimate with sampling variance, not a constant; Hanley and McNeil (1982) give the standard closed-form variance of the AUC and DeLong et al. (1988) its fully nonparametric counterpart. We use them to convert the point comparison "achieved versus required" into a probability of clearance.

2 Methods

2.1 Overview and notation

The framework has four layers, developed in turn below: (i) a decision model that maps a scorecard and a cutoff to the operating quantities a lender cares about; (ii) a one-parameter *separation model* that renders those quantities a deterministic function of the Gini coefficient, which is what makes the KPI invertible; (iii) two inversion operators, one for operating-point KPIs and one for profit KPIs, together with the monotonicity arguments that guarantee a well-defined floor; and (iv) two correctness layers that a bare inversion would omit—a robustness analysis across alternative separation families, and an uncertainty analysis that turns the point floor into a probability of clearance. Throughout, we take the perspective of a risk manager who must commit to a model-quality target *before* any model is built, so every quantity is expressed as a function of the design inputs (the KPI, the portfolio bad rate, the loan economics) rather than of a fitted model.

We use the following notation. Let $S \in \mathbb{R}$ denote the credit score of an applicant, oriented so that *higher is safer*; an applicant is booked (accepted) if and only if $S \geq c$ for a cutoff c. Each applicant is either *good* (repays) or *bad* (defaults). Write $p = \Pr(\text{bad})$ for the through-the-door (TTD) bad rate—the bad rate of the entire applicant population the scorecard will face, *not* the bad rate of any modelling sample. Let the class-conditional score densities be f_B (bad) and f_G (good) with cumulative distribution functions F_B, F_G and survival (tail) functions $\bar{F}_B = 1 - F_B$ and $\bar{F}_G = 1 - F_G$. The economics are summarised by two constants: r, the expected revenue from a booked good, and L, the expected loss from a booked bad. We treat r and L as fixed (stationary) per-account figures; the extension to instance-dependent costs is discussed in Sec-

tion 4. Throughout we consider threshold (cutoff) decision policies, which is how scorecards are deployed in practice.

2.2 Decision model: from score to operating quantities

Given a cutoff c, three quantities fully describe the operating consequences of the scorecard on the TTD population. The *approval rate* is the fraction of applicants booked,

$$A(c) \;=\; \Pr(S \geq c) \;=\; p\,\bar{F}_B(c) + (1-p)\,\bar{F}_G(c), \tag{1}$$

a mixture of the bad and good tails weighted by the population composition. Among those booked, the realised *booked bad rate* is the conditional probability of a bad given acceptance,

$$b(c) \;=\; \Pr(\text{bad} \mid S \geq c) \;=\; \frac{p\,\bar{F}_B(c)}{A(c)}, \tag{2}$$

by Bayes' rule. Finally, the expected *profit per applicant* nets revenue on booked goods against losses on booked bads,

$$\pi(c) \;=\; r\,(1-p)\,\bar{F}_G(c) \;-\; L\,p\,\bar{F}_B(c). \tag{3}$$

The discrimination of the scorecard is the probability that a random good outscores a random bad,

$$\text{AUC} \;=\; \Pr\big(S_{\text{good}} > S_{\text{bad}}\big) = \int_{-\infty}^{\infty} \bar{F}_B(s)\, f_G(s)\, ds, \qquad \text{Gini} \;=\; 2\,\text{AUC} - 1. \tag{4}$$

Equations (1)–(4) are exact and model-free: they hold for any pair of class-conditional score distributions. The difficulty is that they involve the full survival functions $\bar{F}_B, \bar{F}_G$, whereas the risk manager has only a *scalar* summary of model quality—the Gini. A single Gini is consistent with infinitely many distribution pairs, and hence with a range of operating outcomes. Closing this gap requires a model that ties the whole distribution to the scalar.

2.3 The separation model and the Gini linkage

We adopt a *one-parameter separation family*: a family of distribution pairs $\{(f_B^\theta, f_G^\theta)\}_{\theta \geq 0}$ indexed by a single separation parameter θ that controls how far apart the good and bad score distributions lie, with $\theta = 0$ corresponding to no separation (Gini $= 0$) and Gini increasing in θ. Fixing a target Gini pins down θ, and hence the entire pair of distributions, which in turn makes $A(c)$, $b(c)$ and $\pi(c)$ deterministic functions of the cutoff. This is the device that makes a scalar KPI invertible into a scalar Gini requirement.

The canonical member is the *bi-normal* family, standard in the ROC literature (Hanley and McNeil, 1982; Hand, 2009): the bad scores are standard normal and the good scores are shifted

by the separation $\theta = d'$,

$$S \mid \text{bad} \sim N(0,1), \qquad S \mid \text{good} \sim N(d',1). \tag{5}$$

Under this model the survival functions are $\bar{F}_B(c) = \Phi(-c)$ and $\bar{F}_G(c) = \Phi(d'-c)$, where Φ is the standard normal CDF. Substituting into (4), the difference $S_{\text{good}} - S_{\text{bad}} \sim N(d',2)$, so

$$\text{AUC} = \Pr\big(N(d',2) > 0\big) = \Phi\left(\frac{d'}{\sqrt{2}}\right), \qquad \Longrightarrow \qquad d' = \sqrt{2}\,\Phi^{-1}\left(\frac{G+1}{2}\right), \tag{6}$$

a closed-form, invertible map between the Gini G and the separation d'. Given G, Equation (6) yields d'; the survival functions then follow, and (1)–(3) become explicit functions of c alone. We take the equal-variance bi-normal as the baseline; an unequal-variance variant, in which the good and bad classes have different score dispersions, is treated as part of the robustness analysis of Section 2.6. Panel A of Figure 1 plots this Gini–AUC–d' linkage; it is the backbone of the entire framework.

2.4 Inverting the operating-point KPI

An operating-point KPI specifies a target approval rate A^* (how much of the market the lender intends to book) and a booked bad-rate ceiling BR^* (the maximum default rate the portfolio can tolerate). The inversion proceeds in two nested steps for each candidate Gini G.

Step 1—pin the cutoff to the approval target. Since $A(c)$ in (1) is continuous and strictly decreasing in c (both tails shrink as the cutoff rises), there is a unique cutoff $c^\star(G)$ solving

$$A\big(c^\star(G)\big) = A^*, \tag{7}$$

found by a one-dimensional root solve (bisection/Brent). Fixing the *approval* rather than the cutoff is what makes different Gini values comparable: we hold the volume of lending constant and ask how the *quality* of the booked pool changes with discrimination.

Step 2—read the booked bad rate and solve for the floor. Evaluating (2) at $c^\star(G)$ gives the achieved booked bad rate at that Gini, $b(G) := b\big(c^\star(G)\big)$. A sharper scorecard, holding approval fixed, concentrates the booked pool among genuinely good applicants, so $b(G)$ is decreasing in G (Proposition 1 in Appendix A). The minimum required Gini is the smallest G at which the ceiling is met:

$$G^*_{\text{op}} = \min\{\, G \in [0,1) : b(G) \le \text{BR}^* \,\}. \tag{8}$$

Because $b(\cdot)$ is continuous and monotonically decreasing, the set in (8) is an interval and the minimum is attained at the unique root of $b(G) = \text{BR}^*$ (found by Brent's method), provided a root exists in $[0,1)$. If $b(0) \le \text{BR}^*$ the KPI is met even by a random scorecard and we report

$G^*_{\text{op}} = 0$ (a *non-binding* requirement); if $b(G) > \text{BR}^*$ for all attainable G the KPI is infeasible under the stated approval target and we report this rather than an inflated number. Panel B of Figure 1 illustrates the construction: booked bad rate falls with Gini and the floor is the abscissa at which the curve crosses the KPI ceiling.

An exact closed form under the bi-normal model. For the equal-variance bi-normal family the outer root solve can be dispensed with entirely. When the requirement binds, the ceiling holds with equality, so (7) and $b(c^\star) = \text{BR}^*$ jointly pin down both tails: $p\,\bar{F}_B(c^\star) = A^*\text{BR}^*$ and $(1-p)\,\bar{F}_G(c^\star) = A^*(1-\text{BR}^*)$. Substituting the bi-normal survival functions and eliminating $c^\star$ gives the required separation $d'^* = \Phi^{-1}\big(A^*(1-\text{BR}^*)/(1-p)\big) - \Phi^{-1}\big(A^*\text{BR}^*/p\big)$, and hence, via (6),

$$G^*_{\text{op}} = 2\,\Phi\left(\frac{\Phi^{-1}\big(\frac{A^*(1-\text{BR}^*)}{1-p}\big) - \Phi^{-1}\big(\frac{A^*\text{BR}^*}{p}\big)}{\sqrt{2}}\right) - 1. \tag{9}$$

The two boundary regimes of (8) appear as the natural domain conditions of (9): if $A^*\text{BR}^* \geq p$ then even booking every bad applicant leaves the booked bad rate below the ceiling, the requirement is non-binding and $G^*_{\text{op}} = 0$; if $A^*(1-\text{BR}^*) \geq 1-p$ there are not enough good applicants in the population to fill the book at the required purity and the KPI is infeasible. Equation (9) reproduces the numerical root solve to the fourth decimal at every design point we evaluate (Section 3.3), and makes the floor's dependence on (A^*, BR^*, p) transparent enough for back-of-envelope use.

2.5 Inverting the profit KPI

A profit KPI specifies a target expected profit per applicant π^*. Here the lender does not fix a cutoff externally; instead, for any given model the cutoff is chosen to maximise profit. For each candidate Gini G we therefore solve an inner optimisation

$$c^{\text{opt}}(G) = \arg\max_c \ \pi(c), \qquad \pi_{\max}(G) = \pi\big(c^{\text{opt}}(G)\big). \tag{10}$$

Differentiating (3), the first-order condition $r(1-p)f_G(c) = Lp\,f_B(c)$ is the classical Bayes decision rule: accept where the posterior odds of being good exceed the loss-to-revenue ratio L/r. The maximised profit $\pi_{\max}(G)$ is non-decreasing in G, because a more discriminating score can only enlarge the set of profitably bookable applicants (Proposition 1); this is the same monotonicity of value in power established, for a model in hand, by Stein (2005). The minimum required Gini is

$$G^*_\pi = \min\{\, G \in [0,1) : \pi_{\max}(G) \geq \pi^* \,\}. \tag{11}$$

Two boundary behaviours are worth stating explicitly. When the economics are generous (L/r small) the target may already be met at $G = 0$: the KPI imposes *no binding* discrimination requirement, which is itself a reportable finding—the business goal is achievable with an uninformative model, so discrimination is not the binding constraint. Conversely, if $\pi_{\max}(G) < \pi^*$ for all attainable G, the profit target is unreachable at the stated economics and portfolio mix. We flag both cases rather than forcing a number. A subtle but important computational point: for skewed separation families the profit function $\pi(c)$ can be sharply peaked, so a naive local optimiser initialised far from the peak can return the accept-all boundary; we therefore locate c^{opt} with a coarse grid search over the plausible cutoff range followed by a local refinement, which is robust across all families we consider. Panel C of Figure 1 shows maximised profit rising with Gini until it meets the target, defining G^*_π.

Remark 1 (Multiple simultaneous KPIs)**.** *When the lending plan states several KPIs at once—for example an approval target with a bad-rate ceiling* and *a profit-per-applicant goal—each KPI induces its own floor, and by the monotonicity of Proposition 1 each feasible set is an interval* $[G^*_k, 1)$*. The joint requirement is therefore simply* $G^* = \max_k G^*_k$*: the most demanding KPI is the binding one, and the framework identifies it by name.*

2.6 Robustness across separation families

The bi-normal model is a modelling choice, not a law, so a floor derived from it alone would inherit an unquantified distributional risk. We therefore recompute every floor under two additional one-parameter families, each calibrated to the *same* Gini, and report the spread as the framework's sensitivity to the shape assumption.

The *bi-gamma* family models scores as gamma-distributed with a common rate and a shape gap: $S \mid \text{bad} \sim \Gamma(k_0, 1)$ and $S \mid \text{good} \sim \Gamma(k_0 + \theta, 1)$ with baseline shape $k_0 = 4$. Gamma scores are right-skewed and bounded below, capturing portfolios where a mass of clearly-good applicants sits in a long upper tail. The *skew-normal* family uses $S \mid \text{bad} \sim \text{SN}(0, 1, \alpha)$ and $S \mid \text{good} \sim \text{SN}(\theta, 1, \alpha)$ with shape $\alpha = 4$, a location-shifted family that departs from symmetry while remaining unbounded. The baseline shape parameters are chosen to represent moderate, empirically plausible asymmetry—$k_0 = 4$ gives the bad class a skewness of 1, and $\alpha = 4$ a skewness of about 0.78, magnitudes typical of fitted scorecard log-odds distributions—but they are choices, so we additionally sweep $k_0 \in \{2, 8\}$ and $\alpha \in \{2, 8\}$, and add an *unequal-variance bi-normal* variant $S \mid \text{good} \sim N(\theta, \sigma^2)$ with $\sigma \in \{0.8, 1.25\}$, the standard two-parameter bi-normal ROC model with the variance ratio held fixed while θ carries the separation. The full set of nine variants is reported in Section 3.2. For the non-normal families the AUC has no elementary closed form, so we obtain it by numerical integration of (4) with an adaptive upper limit chosen from a high quantile of the good distribution (to keep the quadrature accurate as θ

grows). Calibration then solves, for a target Gini G,

$$\theta^*(G): \quad \mathrm{AUC}_{\mathrm{family}}\big(\theta^*(G)\big) = \frac{G+1}{2}, \tag{12}$$

by a one-dimensional root solve over θ. With $\theta^*(G)$ in hand the family's survival functions are known and the identical inversion machinery of (7)–(11) applies unchanged; Appendix A verifies that every family used satisfies the nested-ROC property that underwrites the monotonicity, and hence uniqueness, of the floor. Reporting max over the families gives a conservative floor, and the family-to-family range is a direct, honest measure of how much the requirement depends on the score-shape assumption.

2.7 Uncertainty: from a point floor to a clearance probability

A required floor is a single number, but an achieved Gini is an *estimate* with sampling variance, computed on a finite test book of n_G goods and n_B bads. Ignoring this variance would let a scorecard whose point estimate merely grazes the floor be declared a pass when it might well be below it. We quantify the gap probabilistically using the closed-form variance of the AUC of Hanley and McNeil (1982),

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{AUC}}) = \frac{\widehat{A}(1-\widehat{A}) + (n_B-1)(Q_1-\widehat{A}^2) + (n_G-1)(Q_2-\widehat{A}^2)}{n_B\, n_G}, \tag{13}$$

with $\widehat{A}$ the estimated AUC, $Q_1 = \widehat{A}/(2-\widehat{A})$ and $Q_2 = 2\widehat{A}^2/(1+\widehat{A})$. We use the Hanley–McNeil form because it requires only the triple $(\widehat{A}, n_B, n_G)$ and therefore keeps the whole framework free of any fitted model; when the individual test-set scores are available, the fully nonparametric estimator of DeLong et al. (1988), computed from placement values, is the recommended production alternative, and we report both in Section 3.1. Writing $\mathrm{SE} = \sqrt{\widehat{\mathrm{Var}}}$ and $\mathrm{AUC}^* = (G^*+1)/2$ for the AUC implied by the required floor, and invoking the asymptotic normality of the AUC estimator, the probability that the scorecard actually clears the floor is

$$\Pr(\mathrm{clear}) = \Phi\left(\frac{\widehat{\mathrm{AUC}} - \mathrm{AUC}^*}{\mathrm{SE}}\right). \tag{14}$$

This converts the binary "pass/fail" of a point comparison into a calibrated clearance probability that shrinks toward 0.5 as the achieved Gini approaches the floor and as the test book shrinks. Two interpretive caveats are in order. First, (14) is an asymptotic confidence statement about the estimator; read literally as the probability that the *true* AUC exceeds the floor, it is a Bayesian posterior probability under a flat prior on the AUC. Second, it is *conditional* on the design inputs: the floor G^* is treated as known given (p, r, L) and the separation family, so (14) propagates the sampling uncertainty of the achieved Gini but not any uncertainty in p, r or L. We address the latter through the sensitivity analysis of Section 3.3—which brackets

the floor across a plausible range of p—and flag full joint propagation as future work. With those caveats, the clearance probability is the quantity a risk committee should actually weigh, and—as Section 3 shows—a scorecard can sit above a floor on the point estimate yet clear it with uncomfortably low probability on a small book.

2.8 The decision population, reject inference, and Gini portability

One assumption deserves emphasis because it is the framework's most consequential input. The KPI and the floor are defined over the *through-the-door* population via the TTD bad rate p, whereas an achieved Gini is typically measured on *accepted* applicants only—the population that survived a previous accept/reject policy. These two populations differ systematically, the classic reject-inference problem (Crook and Banasik, 2004; Banasik and Crook, 2007). The dependence enters the framework in two distinct places, and it is worth separating them.

First, p appears directly in (1)–(3), and Section 3.3 shows the floor is highly sensitive to it. We therefore recommend that p be set from a reject-inference-corrected estimate of the TTD bad rate, not from the accepted-sample bad rate, and we treat p as the primary axis of the sensitivity analysis. Where a corrected p is unavailable, the sensitivity table bounds the floor across a plausible range of p.

Second—and less often made explicit—comparing an *achieved* Gini against a floor defined at the TTD rate p invokes a *portability assumption*: that the class-conditional score distributions f_B and f_G, and hence the Gini, are the same in the population where the Gini was measured and in the TTD population where the KPI lives. The Gini is invariant to the class *prior* (a re-weighting of goods versus bads leaves the ROC unchanged), so a shift in bad rate alone is harmless; what is *not* harmless is selection that truncates or reshapes the class-conditional distributions themselves, as acceptance on a correlated legacy score does. Under such selection the accepted-sample Gini can be biased in either direction relative to the TTD Gini. The honest statement of the framework's verdict is therefore: *conditional on the measured Gini being a valid estimate of the TTD Gini*, the scorecard clears (or misses) the floor with the stated probability. In practice this argues for measuring the achieved Gini on as unselected a population as possible—including overridden and above-cutoff-declined accounts where performance is observed—or for applying a reject-inference adjustment to the ROC itself, in the same spirit as the adjustment to p. Section 3.1 states how this assumption plays out in our empirical study.

2.9 Computational procedure

The complete procedure is summarised as follows. *Inputs:* a KPI (either (A^*, BR^*) or π^* with economics r, L), the TTD bad rate p, and a choice of separation family. *For each candidate Gini G* on a grid spanning $[0, 1)$: (1) calibrate the family's separation parameter to G—by the closed form (6) for bi-normal, or a root solve for bi-gamma and skew-normal; (2) for an operating KPI, solve (7) for the approval-matching cutoff and evaluate the booked bad rate

(2); for a profit KPI, solve the inner optimisation (10). *Then* solve the outer root problem (8) or (11) for the floor G^*, guarding the non-binding ($G^* = 0$) and infeasible cases—or, for the bi-normal operating KPI, evaluate the closed form (9) directly. *Finally* repeat across the families to obtain the robustness range, and apply (14) with the achieved AUC and test-book sizes to obtain the clearance probability. Every step is either a closed-form evaluation or a well-conditioned one-dimensional root solve, so the entire mapping—from a business plan to a probabilistic model-quality specification—runs in well under a second and requires no fitted model as input.

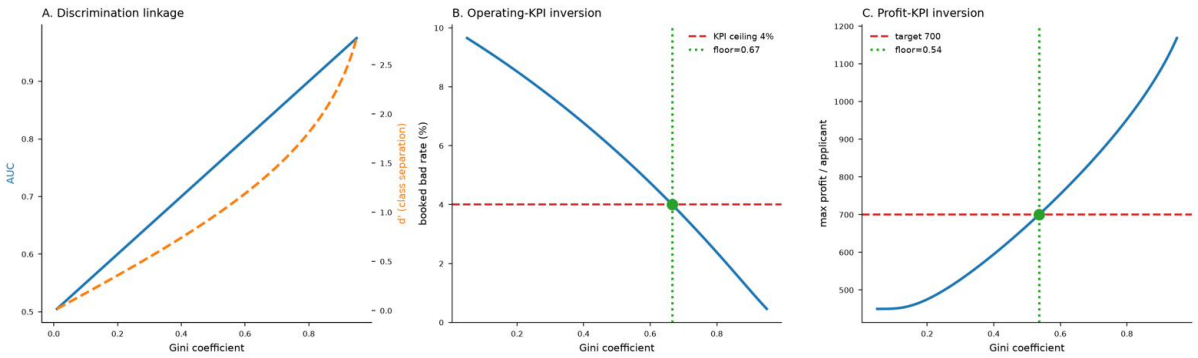

Figure 1: The inversion mechanism. (A) the Gini–AUC–d' linkage. (B) operating-KPI inversion; booked bad rate falls with Gini until it meets the KPI ceiling, defining the floor. (C) profit-KPI inversion; maximum profit rises with Gini until it meets the target.

3 Results

3.1 Data and scorecard

We use the real UCI Statlog German Credit dataset (1000 applications, 20 attributes, 30% sample bad rate). Because that base rate is a research-sample artefact, we treat the TTD bad rate p as a swept parameter; the role of the data is therefore twofold—to establish a realistic, achievable Gini for a scorecard of this type, and to validate the score-distribution assumption on which the inversion rests—and, in Section 3.4, to check the inversion curve itself against data. A weight-of-evidence logistic scorecard, trained on 70% and tested on the held-out 30% (300 loans, 90 bad / 210 good), achieves **AUC 0.801, Gini 0.602** (95% CI [0.483, 0.721]); full construction details (binning scheme, variable screening, and final model composition) are given in Appendix B. Across 50 repeated random 70/30 splits the test Gini averages 0.555 with standard deviation 0.050, so the headline split sits toward the upper end of the sampling range (about one standard deviation above the mean); it is exactly this split-to-split variability that the clearance layer of Section 2.7 carries into the verdict, so no conclusion below rests on the point estimate alone. The Hanley–McNeil form (13) gives $\mathrm{SE}(\widehat{\mathrm{AUC}}) = 0.030$; the nonparametric DeLong estimator computed on the same test scores gives 0.027, so the closed form is mildly conservative on this book. A gradient-boosting benchmark reaches Gini 0.596 on the

same split—essentially level with the scorecard—consistent with the finding that boosted ensembles need substantially larger samples to open a material gap over a well-binned logistic model (Lessmann et al., 2015).

The bi-normal assumption holds well empirically: the fitted class separation implies a Gini of 0.582, closely matching the ROC Gini of 0.602, and the per-class Q–Q plots track normality, with per-class Q–Q correlations of 0.99 (Figure 2).

Two caveats scope what this study demonstrates. First, because p is swept rather than estimated from the sample, the exercise is an end-to-end *application* of the framework at representative design inputs, not a validation of any particular value of p. Second, the portability assumption of Section 2.8 here takes a specific form: the Gini of 0.602 is measured on the research sample (30% bad rate), while the floors are stated at a hypothetical TTD rate of $p = 10\%$. Carrying the Gini across relies on the class-conditional score distributions being stable under the change of mix—exactly the prior-invariance that holds when composition, not selection, differs between the two populations. The German Credit sample was not filtered by a score correlated with ours, so within this study the assumption is innocuous; in production use, where the Gini is measured on a book filtered by a legacy scorecard, it is not, and the remedies of Section 2.8 apply.

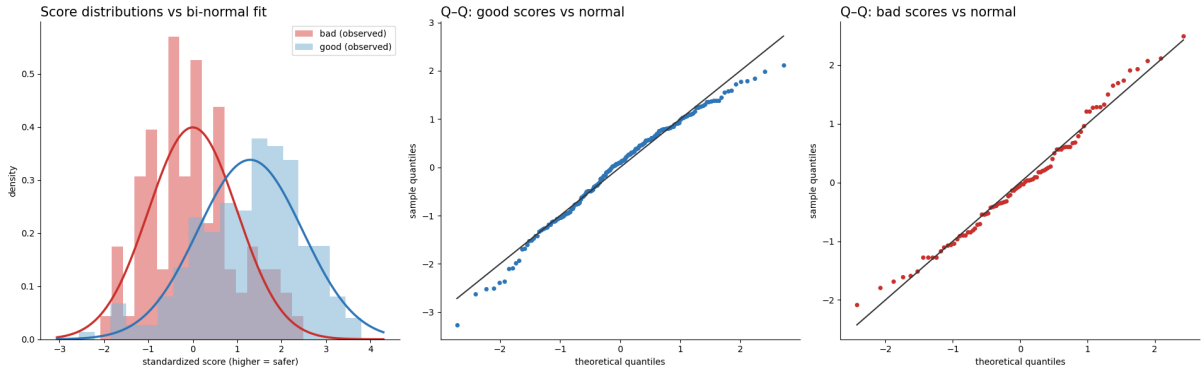

Figure 2: Score-distribution validation. Left: observed good/bad score densities against the fitted bi-normal. Centre and right: Q–Q plots of good and bad scores against the normal, supporting the one-parameter separation model.

3.2 The inversion verdict

We apply two representative KPIs at $p = 10\%$: an operating KPI (approve 75%, booked bad rate $\leq 4\%$) and a profit KPI (target profit of 700 per applicant, with $r = 1500$ and $L = 9000$). Monetary quantities are in stylised currency units; the economically meaningful inputs are the ratios $L/r = 6$ and $\pi^*/r \approx 0.47$, and every floor is invariant to a common re-scaling of (r, L, π^*). The floors and verdicts across the three baseline distributional families are shown in Table 1.

Table 1: Minimum required Gini and clearance verdict by KPI and score distribution ($p = 10\%$).

KPI	Distribution	Min Gini	Achieved	Gap	Pr(clear)	Verdict
Operating (approve 75%, bad$\leq$4%)	bi-normal	0.666	0.602	-0.064	0.146	**SHORTFALL**
Profit (target 700/applicant)	bi-normal	0.536	0.602	$+0.066$	0.862	**PASS**
Operating (approve 75%, bad$\leq$4%)	bi-gamma	0.661	0.602	-0.059	0.165	**SHORTFALL**
Profit (target 700/applicant)	bi-gamma	0.522	0.602	$+0.080$	0.906	**PASS**
Operating (approve 75%, bad$\leq$4%)	skew-normal	0.650	0.602	-0.048	0.216	**SHORTFALL**
Profit (target 700/applicant)	skew-normal	0.455	0.602	$+0.147$	0.993	**PASS**

The result is a clear split: the scorecard **passes the profit KPI** (floor 0.536, clearance probability 0.86) but **falls short of the operating KPI** (floor 0.666, clearance probability only 0.15). This is exactly the kind of actionable, economics-aware verdict a bare Gini cannot deliver.

The two floors are not equally robust to the shape assumption, and the asymmetry is informative. Across the three baseline families the operating floor varies by only 0.016 Gini points, while the profit floor varies by 0.081. Table 2 extends the check to sweeps of the shape parameters ($k_0 \in \{2,4,8\}$, $\alpha \in \{2,4,8\}$) and to the unequal-variance bi-normal ($\sigma \in \{0.8,1,1.25\}$): across all nine variants the operating floor stays within $[0.650, 0.675]$ (spread 0.025) whereas the profit floor spans $[0.415, 0.560]$. The mechanism is the location of the cutoff each KPI pins down. The operating KPI fixes the cutoff at the 75% approval quantile—in the body of the score mixture, where families calibrated to the same AUC nearly coincide—so the floor is close to a property of the AUC alone. The profit-optimal cutoff, by contrast, sits where the posterior odds equal $L/r = 6$, deep in the upper tail, precisely where equal-AUC families differ most; right-skewed families place relatively more clearly-good mass in that tail, lowering the profit floor. The practical reading is that operating-point floors can be quoted with little distributional qualification, while profit floors should be reported with their distributional range—or conservatively, as the maximum over families—and, where score history exists, with the family calibrated to the portfolio's own scores.

Table 2: Extended shape robustness: minimum required Gini across nine separation-family variants ($p = 10\%$; same KPIs as Table 1). Baseline variants in bold.

Variant	Floor (operating)	Floor (profit)
bi-normal, equal variance ($\sigma = 1$)	**0.666**	**0.536**
bi-normal, $\sigma = 0.8$	0.650	0.485
bi-normal, $\sigma = 1.25$	0.675	0.560
bi-gamma, $k_0 = 2$	0.659	0.517
bi-gamma, $k_0 = 4$	**0.661**	**0.522**
bi-gamma, $k_0 = 8$	0.662	0.526
skew-normal, $\alpha = 2$	0.651	0.499
skew-normal, $\alpha = 4$	**0.650**	**0.455**
skew-normal, $\alpha = 8$	0.657	0.415

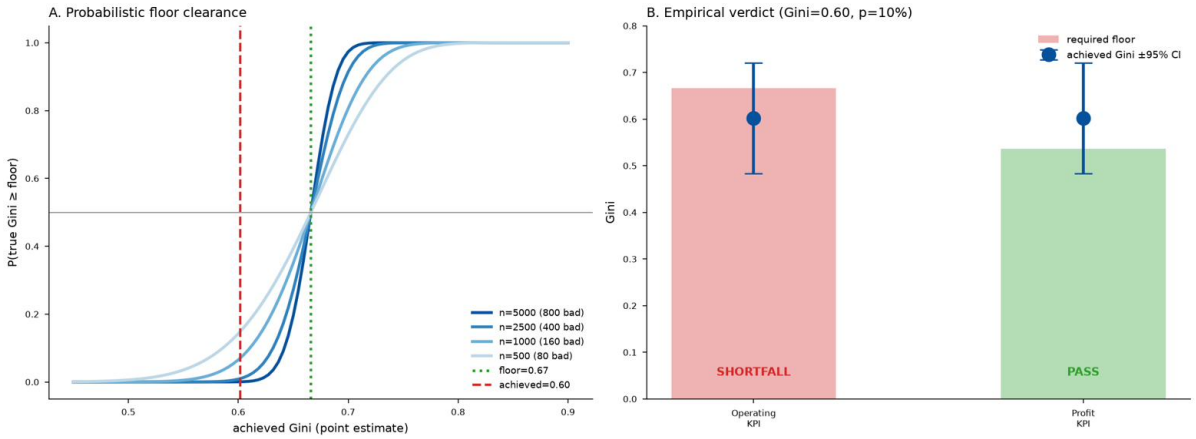

Figure 3: Uncertainty and verdict. Left: clearance probability for the operating floor as a function of achieved Gini, for several test-book sizes, showing that small books make the verdict uncertain even above the floor. Right: achieved Gini with 95% CI against the two required floors.

3.3 Sensitivity

The required floor is highly sensitive to the TTD bad rate—the single most important input. Table 3 sweeps p under the operating and profit KPIs; the operating column is reproduced exactly by the closed form (9).

Table 3: Minimum required Gini vs through-the-door bad rate p (bi-normal).

p (TTD bad rate)	Floor (operating)	Floor (profit)
0.04	0.000	0.000
0.06	0.392	0.000
0.08	0.563	0.326
0.10	0.666	0.536
0.12	0.737	0.635
0.15	0.813	0.727
0.20	0.899	0.819

At $p = 4\%$ no discrimination is required (any model meets the KPI); by $p = 20\%$ a Gini near 0.90 is needed. This quantifies a familiar intuition—that riskier portfolios demand sharper models—on an exact scale.

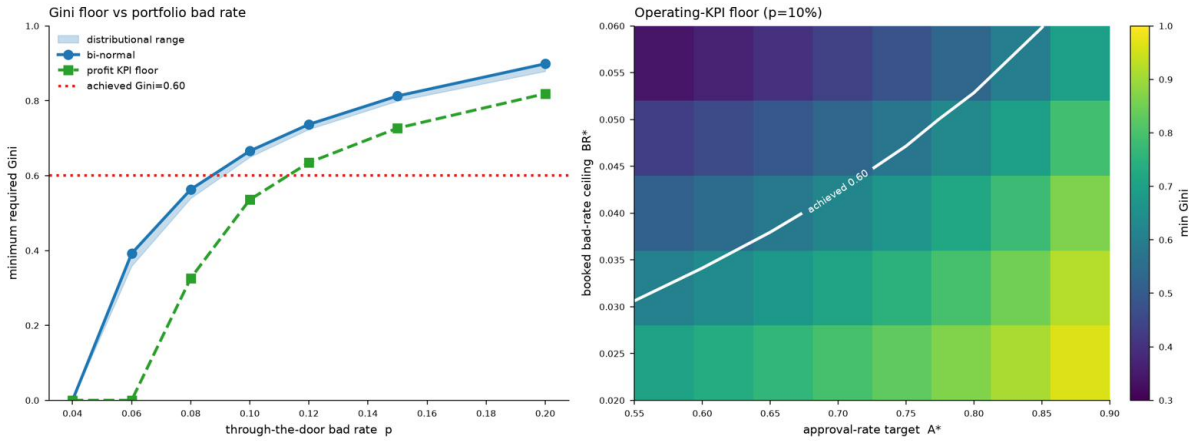

Figure 4: Sensitivity and robustness. Left: minimum Gini vs through-the-door bad rate, with the distributional range (band) across the three families and the achieved Gini for reference. Right: operating-KPI floor over the approval-target by bad-rate-ceiling grid, with the white contour marking where the achieved Gini exactly meets the requirement.

3.4 Checking the inversion curve against data

The tables above apply the framework; this subsection tests its central object—the curve $b(G)$ of Section 2.4, whose crossing with the KPI ceiling *is* the floor—against data. The design is a controlled score degradation. Starting from the fitted scorecard's test-set scores, we corrupt them with independent Gaussian noise, $\tilde{S} = S + \sigma_\varepsilon Z$ with $Z \sim N(0,1)$, choosing σ_ε so that the corrupted Gini sweeps a grid from 0.60 down to 0.10 in steps of 0.05. At each grid point we (i) re-weight the test sample so that goods and bads carry importance weights matching a TTD composition of $p = 10\%$, (ii) set the cutoff at the weighted 75% approval quantile, and (iii) record the weighted booked bad rate, averaging over 200 noise draws. This produces an

empirical booked-bad-rate curve $\widehat{b}(G)$ from real repayment outcomes, with no distributional assumption, to set against the theoretical bi-normal curve $b(G)$ that defines the floor. The agreement is close: over the sweep the mean absolute deviation between $\widehat{b}(G)$ and $b(G)$ is 0.05 percentage points, with a maximum of 0.13 points (Figure 5). Because noise can only degrade the scorecard, the sweep necessarily covers the region below the achieved Gini of 0.602; on this entire range the empirical curve stays above the 4% ceiling, exactly as the theory requires below the floor of 0.666, and at the achieved Gini the un-noised scorecard books a weighted bad rate of 4.5% against a theoretical 4.7%—both above the ceiling, consistent with the shortfall verdict of Table 1. Because the degraded scores are noised versions of an approximately bi-normal score, this is a test of the inversion machinery and of the adequacy of the one-parameter reduction on real outcome data, not of the bi-normal assumption in isolation (which Figure 2 addresses directly).

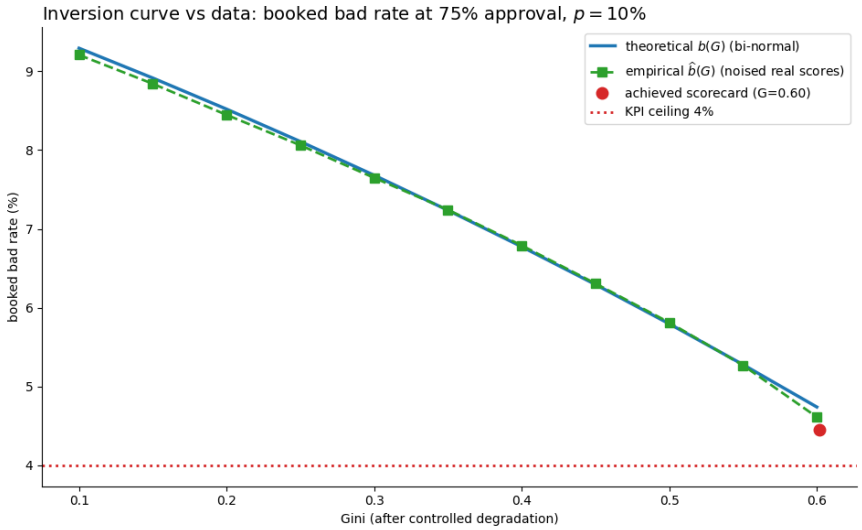

Figure 5: Checking the inversion curve against data. The theoretical booked-bad-rate curve $b(G)$ (bi-normal, 75% approval, $p = 10\%$) is overlaid with the empirical curve $\widehat{b}(G)$ obtained by noise-degrading the real test-set scores and re-weighting outcomes to the TTD composition. Mean absolute deviation 0.05 percentage points; maximum 0.13.

3.5 Replication on a larger book

The German Credit sample is small, and it is the width of the resulting confidence interval—not any property of the framework—that drives the low clearance probability on the operating KPI. To show how a larger book sharpens the verdict, and that the score-shape validation is not particular to one dataset, we replicate the full pipeline (scorecard, Q–Q validation, floors, clearance) on the UCI Taiwan default-of-credit-card-clients dataset (30,000 accounts, 22% sample bad rate), holding the KPIs and $p = 10\%$ fixed so that the floors of Table 1 carry over unchanged and only the achieved Gini and its uncertainty are re-estimated. Table 4 reports the result.

Table 4: Replication on the Taiwan credit-card dataset (same KPIs and $p = 10\%$; floors as in Table 1, bi-normal).

Dataset	n_{test} (bad/good)	Gini [95% CI]	Pr(clear) op.	Pr(clear) profit	Bi-normal fit
German Credit	300 (90/210)	0.602 [0.483, 0.721]	0.146	0.862	Q–Q $r = 0.99$
Taiwan cards	9000 (1991/7009)	0.471 [0.444, 0.498]	<0.001	<0.001	Q–Q $r = 0.94$–0.96

The Taiwan scorecard, built with the identical pipeline on a different product, achieves a lower Gini of 0.471—behavioural separation on revolving credit-card accounts from application-style characteristics is known to be harder—and on a test book of 9,000 accounts the verdict is decisive in the direction of failure: the clearance probability is below 0.001 for *both* KPIs, leaving no ambiguity for a committee to debate. This is the framework's promise seen from the opposite side of the German Credit case: a large book does not guarantee a pass, it guarantees a *sharp* answer, and the same floors that were merely improbable for the German scorecard are unambiguously out of reach here unless the KPI itself is relaxed (for example, by lowering the approval target along the contour of Figure 4). The per-class Q–Q correlations of 0.94–0.96 indicate that the bi-normal shape is a rougher, though still serviceable, approximation on this portfolio, which is precisely the situation the multi-family range of Table 2 is designed to absorb.

4 Discussion

For practice. The framework changes *when* the model-quality question is asked. Instead of building a scorecard and then debating whether its Gini is adequate, a risk team can state its lending plan—approval target, bad-rate ceiling, or profit-per-account goal—and read off the minimum Gini before a line of modelling code is written. This has three concrete uses. In *project scoping*, the floor tells the team whether the plan is reachable with available data at all, or whether the KPI itself must be relaxed; a floor near 0.9 (as arises for a 20% through-the-door bad rate in Table 3) is a warning that no realistic scorecard will rescue an over-ambitious approval target, and the closed form (9) makes this a one-line calculation. In *model governance*, the clearance probability of Equation (14) gives validators a defensible, quantitative sign-off criterion—"the scorecard clears the profit floor with clearance probability 0.86 on the current book"—in place of an eyeballed comparison to a benchmark. In *vendor and build-versus-buy decisions*, competing scorecards can be judged against the same KPI-derived bar rather than against each other's headline Gini. Across all three, the strong dependence of the floor on the through-the-door bad rate p is the practical headline: the target must be set on a realistic TTD rate, not the modelling-sample rate, which directly implicates reject inference (Crook and Banasik, 2004) and argues for investing in a corrected estimate of p before trusting any floor.

Limitations. Four simplifications bound the framework's claims. First, the one-parameter

separation model compresses an entire score distribution to a single knob; we mitigate this with the multi-family robustness check—bi-normal (equal and unequal variance), bi-gamma and skew-normal, with shape-parameter sweeps—and with the empirical Q–Q validation of Figure 2. The operating floor proves insensitive to the shape assumption, but the profit floor is materially more shape-dependent (Table 2), and a bespoke portfolio with a strongly multi-modal or heavy-tailed score could depart from all the families considered; the honest response is to calibrate the family to that portfolio's own score history (below). Second, the economics are stationary: r and L are treated as constants, whereas real revenue and loss-given-default vary by applicant, exposure, and macroeconomic state, which instance-dependent cost-sensitive methods (Petrides et al., 2022; Maldonado et al., 2017) are designed to capture. Our profit inversion should therefore be read as a first-order, portfolio-average statement. Third, the empirical anchor is the German Credit sample of 1,000 loans, which is small; the resulting confidence interval on the achieved Gini is wide (roughly ± 0.12), and it is this width—not any weakness of the framework—that drives the low clearance probability on the operating KPI. As Figure 3 makes explicit, and the replication of Section 3.5 demonstrates on a book thirty times larger, a larger book sharpens the verdict without changing the floor—in the Taiwan case, sharpening it to a decisive shortfall. Fourth, the verdict rests on the portability assumption of Section 2.8: the achieved Gini must be a valid estimate of the TTD Gini. Where the measurement population has been filtered by a correlated legacy score, the comparison inherits the reject-inference problem on the ROC itself, not only on p.

Future work. Several extensions follow directly. The profit model can be generalised to instance-dependent costs by replacing the scalar r, L with per-applicant distributions and integrating, in the spirit of the Expected Maximum Profit measure (Verbraken et al., 2014). The reject-inference correction can be embedded directly into the p that enters the floor, so that the design target and the population estimate are consistent by construction rather than supplied separately, and extended to the achieved ROC itself as Section 2.8 argues. Uncertainty in the design inputs (p, r, L) can be propagated jointly with the sampling variance of the achieved Gini, turning the clearance probability into a fully marginal statement. The separation family can be fitted to a bank's own historical score distributions, turning the robustness range from a fixed multi-family band into a portfolio-specific credible interval. Finally, the same inversion logic extends naturally to multi-segment portfolios—deriving a per-segment floor and aggregating via Remark 1—and to through-the-cycle targets that let p itself vary with the economic cycle.

5 Conclusion

We turned scorecard evaluation on its head. The prevailing practice measures a Gini and then asks, by analogy to past models, whether it is enough. We instead start from the business KPI the scorecard must serve and derive the minimum Gini that makes the KPI attainable, closing the long-standing gap—articulated most sharply by Hand (2009)—between a discrimination

number and the economics of the decision it informs. Where prior work maps a given model's power forward into cutoffs and profit (Stein, 2005; Blöchlinger and Leippold, 2006; Verbraken et al., 2014), we run the map in reverse, at design time. The derivation is exact and inexpensive: a one-parameter separation model turns the KPI into a well-conditioned root-find whose solution is the required floor—available in closed form for the operating KPI under the bi-normal model—and the same machinery covers both operating-point and profit KPIs.

Two design choices make the result trustworthy rather than merely elegant. Recomputing every floor under bi-normal, bi-gamma and skew-normal score distributions—including shape-parameter sweeps and an unequal-variance variant—shows that the operating requirement is a property of the KPI and the portfolio rather than of a shape assumption, while honestly flagging that profit requirements carry a wider distributional range. And carrying the sampling variance of the achieved AUC (Hanley and McNeil, 1982; DeLong et al., 1988) through to a clearance probability converts a brittle point comparison into a statement a risk committee can weigh. On the real UCI German Credit data the framework delivered a concrete, economics-aware verdict for a single scorecard—a comfortable pass on the profit KPI and a clear shortfall on the operating KPI—of a kind no headline Gini could produce, and it localised the shortfall's uncertainty to the small size of the test book rather than to any ambiguity in the method.

Beyond the specific numbers, the contribution is a change of workflow: a principled, economics-aware bar that credit-risk teams can set *before* they build, use to scope projects and govern models, and defend to validators and regulators in the same Gini language they already speak. We hope it moves the field's central question from "how high is our Gini?" to "how high does our Gini need to be?"—and gives a derivable answer in place of an intuition.

Data and code availability

The German Credit and Taiwan credit-card datasets are publicly available from the UCI Machine Learning Repository. Python code reproducing every floor, table and figure in this paper—the inversion framework, the scorecard pipeline, the degradation experiment and the replication—accompanies the submission and is available from the corresponding author.

A Monotonicity of the inversion

Both inversion operators rest on the monotonicity, in the separation parameter, of the booked bad rate at fixed approval and of the maximised profit. The following proposition establishes both from a single ordering property of the family's ROC curves, and we then verify that property for every family used in the paper. Recall that for a separation family $\{(f_B^\theta, f_G^\theta)\}$ the ROC curve at separation θ is $R_\theta(u) = \bar{F}_G^\theta\big((\bar{F}_B^\theta)^{-1}(u)\big)$, $u \in [0,1]$, the good-tail (true-positive) rate as a function of the bad-tail (false-positive) rate under threshold policies.

Proposition 1. *Suppose the family's ROC curves are* nested*: $R_{\theta'}(u) \geq R_\theta(u)$ for all $u \in [0,1]$ whenever $\theta' > \theta$, with strict inequality on a set of positive measure. Then, restricting to threshold policies: (i) at any fixed approval rate $A^* \in (0,1)$, the booked bad rate b is nonincreasing in θ; (ii) the maximised profit $\pi_{\max}$ is nondecreasing in θ. Since $\mathrm{AUC}(\theta) = \int_0^1 R_\theta(u)\,du$ is strictly increasing under nesting, both quantities are correspondingly monotone in the Gini, so the feasible sets in* (8) *and* (11) *are intervals and the floors are unique.*

Proof. Parametrise threshold policies by the bad acceptance rate $u = \bar{F}_B(c)$. (i) The approval constraint reads $\psi_\theta(u) := p\,u + (1-p)\,R_\theta(u) = A^*$. For each θ, ψ_θ is continuous and strictly increasing in u, so the constraint has a unique solution u_θ; and ψ_θ is nondecreasing in θ pointwise by nesting, so $u_{\theta'} \leq u_\theta$ for $\theta' > \theta$. The booked bad rate is $b = p\,u_\theta / A^*$, which is therefore nonincreasing in θ. (ii) Profit under the policy u is $\pi_\theta(u) = r(1-p)R_\theta(u) - L\,p\,u$. By nesting, $\pi_{\theta'}(u) \geq \pi_\theta(u)$ for every u, hence $\max_u \pi_{\theta'}(u) \geq \max_u \pi_\theta(u)$. □

Verification for the families used. (a) *Bi-normal (any fixed variance ratio σ):* $u = \Phi(-c)$ gives $c = -\Phi^{-1}(u)$ and $R_\theta(u) = \Phi\big((\theta + \Phi^{-1}(u))/\sigma\big)$, strictly increasing in θ. (b) *Bi-gamma:* $R_\theta(u) = \bar{F}_{\Gamma(k_0+\theta)}\big(\bar{F}^{-1}_{\Gamma(k_0)}(u)\big)$; a gamma variable with larger shape (and common rate) is the sum of the smaller-shape variable and an independent positive increment, hence stochastically larger, so its survival function at any fixed point is increasing in θ. (c) *Skew-normal (location shift):* $R_\theta(u) = \bar{F}_0\big(\bar{F}_0^{-1}(u) - \theta\big)$ for the fixed baseline CDF F_0, strictly increasing in θ. In all cases nesting holds, and within each pair at fixed θ the equal-variance bi-normal, bi-gamma and skew-normal families additionally have a monotone likelihood ratio in s, so the Bayes-optimal policy in (10) is indeed a threshold rule; for the unequal-variance bi-normal the likelihood ratio is not globally monotone and the restriction to threshold policies in Proposition 1 is substantive—but it matches deployment practice, where scorecards are operated with a single cutoff.

B Scorecard construction details

The weight-of-evidence logistic scorecard of Section 3.1 was built as follows. The seven numerical attributes were binned into five equal-frequency (quintile) bins, with bin edges computed on the training split only; each of the thirteen categorical attributes used its own categories as bins. Weights of evidence were computed on the training split with Laplace smoothing ($\varepsilon = 0.5$ added to each cell), oriented so that higher weight of evidence indicates lower default risk; bins unseen in training map to zero. Characteristics were screened by information value with threshold 0.02, retaining 15 of the 20 attributes (dropped: instalment rate, years at present residence, job, number of dependants, telephone); the strongest retained characteristics by information value are checking-account status (0.71), credit history (0.31) and duration (0.24). The logistic regression was fitted by unpenalised maximum likelihood (scikit-learn, `lbfgs`) on the weight-of-evidence-transformed characteristics, using a stratified 70/30 train/test split

with random seed 86; the score is the negative of the fitted log-odds of default, so that higher is safer, and all reported test quantities use the held-out 30%. The gradient-boosting benchmark is scikit-learn's `GradientBoostingClassifier` with default hyperparameters (100 trees, depth 3, learning rate 0.1, seed 0) on one-hot-encoded attributes and the same split. The Taiwan replication of Section 3.5 uses the identical pipeline (sex, education and marital status as categorical; all remaining attributes quintile-binned) with seed 42, retaining 14 of 23 characteristics; the screened-out characteristics are sex, marital status, age, and the six raw bill-amount fields, whose information values fall below the threshold once the payment-status and payment-amount histories are present.

References

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

Banasik, J. and Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3), 1582–1594. doi:10.1016/j.ejor.2006.06.072

Basel Committee on Banking Supervision (2005). Studies on the validation of internal rating systems. Working Paper No. 14, Bank for International Settlements. `https://www.bis.org/publ/bcbs_wp14.htm`

Bellotti, T. and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2), 3302–3308. doi:10.1016/j.eswa.2008.01.005

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3), 851–873. doi:10.1016/j.jbankfin.2005.07.014

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4), 857–874. doi:10.1016/j.jbankfin.2003.10.010

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845. doi:10.2307/2531595

Djeundje, V.B., Crook, J., Calabrese, R. and Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163, 113766. doi:10.1016/j.eswa.2020.113766

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192. doi:10.1016/j.ejor.2021.06.053

Engelmann, B., Hayden, E. and Tasche, D. (2003). Testing rating accuracy. *Risk*, 16(1), 82–86.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368–378. doi:10.1016/j.ejor.2010.09.029

Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M. and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1), 292–305. doi:10.1016/j.ejor.2021.03.006

Hand, D.J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123. doi:10.1007/s10994-009-5119-5

Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541. doi:10.1111/j.1467-985x.1997.00078.x

Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. doi:10.1148/radiology.143.1.7063747

Khandani, A.E., Kim, A.J. and Lo, A.W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787. doi:10.1016/j.jbankfin.2010.06.001

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

Maldonado, S., Bravo, C., López, J. and Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems*, 104, 113–121. doi:10.1016/j.dss.2017.10.007

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J. and Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74, 26–39. doi:10.1016/j.asoc.2018.10.004

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

Stein, R.M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking & Finance*, 29(5), 1213–1236. doi:10.1016/j.jbankfin.2004.04.008

Thomas, L.C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172. doi:10.1016/s0169-2070(00)00034-0

Tran, C.M. and Tran, T.H. (2025). Optimizing recovery debt collection process by using machine learning and operation research. In: Nguyen, N.T. et al. (eds), *Recent Challenges in*

Intelligent Information and Database Systems (ACIIDS 2025), Communications in Computer and Information Science, vol. 2495. Springer, Singapore. doi:10.1007/978-981-96-5887-9_15

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