**The Eastasouth Journal of Information System and Computer Science**
Vol. 3, No. 01, August 2025, pp. 224-236
ISSN: 3025-566X, DOI: 10.58812/esiscs.v3i01

# Cost Sensitive Machine Learning for Credit Risk Assessment and Lending Decisions

**Muhammad Imran Syed[1], Mitu Akter[2], Semanto Saha Utsab[3], Md Tanvirul Islam[4], Yeasmin Jahan[5]**

[1] University of St. Thomas, Houston, TX, USA
[2] Zicklin School of Business, Baruch College, New York, NY 10010, USA
[3] Montclair State University, Montclair, New Jersey, United States
[4] Atlantis University, Miami, Florida, USA
[5] Metropolitan University, Sylhet, Bangladesh

**Article Info**

***Article history:***

Received August, 2025
Revised August, 2025
Accepted August, 2025

***Keywords:***

Cost-Sensitive Classification;
Credit Risk Assessment;
Ensemble Learning;
Financial Analytics;
Machine Learning;
Probability Calibration

**ABSTRACT**

Machine learning can improve credit risk prediction, yet predictive gains do not automatically translate into better lending decisions. The financial value of a score depends on which errors it makes, whether its probability is calibrated, and how its operating threshold reflects the costs of approval and rejection. This study develops an analytical framework linking these quantities for statistical and ensemble credit scoring models. Published findings on classifier benchmarking, imbalanced credit data, behavioral information, cost-sensitive learning, and distributional effects establish the research context. The original analysis derives feasible loss bounds from accuracy and class prevalence, identifies conditions for economic dominance between classifiers, and examines sensitivity to error costs and probability uncertainty. Deterministic scenarios use adverse class prevalences from 5% to 50%, relative error penalties from one to ten, and classifier accuracies from 70% to 95%. No borrower-level models are newly trained. At 85% accuracy, 30% adverse prevalence, and a five-to-one error penalty, feasible loss ranges from 150 to 750 units per 1,000 applications. Two feasible classifiers with accuracies of 85% and 80% can incur losses of 750 and 200 units, respectively, demonstrating a possible reversal between statistical and economic rankings. Expected loss minimization gives a probability threshold of one sixth at a five-to-one penalty ratio. A general regret expression also shows why probability uncertainty near that threshold deserves particular attention. The framework explains how findings from credit scoring research can inform model selection without treating results from different populations as directly comparable. It supports evaluating the combined prediction and decision process through discrimination, calibration, financial consequences, and distributional outcomes.

*This is an open access article under the CC BY-SA license.*

***Corresponding Author:***

Name: Muhammad Imran Syed
Institution: University of St. Thomas, Houston, TX, USA
Email: emraan.kazmi5@gmail.com

## 1. INTRODUCTION

Credit risk assessment requires an institution to make a financial commitment before the borrower's future repayment behavior is known. Statistical models help estimate uncertainty from financial histories and applicant characteristics. In recent years, machine learning has expanded the range of relationships that can be represented, particularly when risk depends on interactions among several variables. The operational question, however, remains whether the resulting decisions improve lending outcomes under the institution's actual costs and constraints.

The distinction between prediction and action has roots in the statistical credit scoring literature. [1] describe the classification setting and the practical issues involved in applying statistical methods to consumer credit. [2] subsequently compare advanced classification techniques, and [3] extend this benchmarking tradition. These contributions establish that algorithm comparisons should be broad enough to identify meaningful alternatives. They also encourage attention to the connection between statistical performance and the business value of a scorecard.

Empirical studies provide several reasons to investigate flexible models. [4] find that Random Forest and gradient boosting perform well in their imbalanced credit scoring experiments. [5] use transaction and credit bureau information to forecast consumer credit outcomes and estimate economically meaningful benefits from decisions based on those forecasts. [6] demonstrate that digital footprints contain predictive information that can complement conventional credit information. These findings support further research, although the population, outcome, and decision setting differ across studies.

A separate line of research demonstrates why better prediction is insufficient as a complete objective. [7] incorporate observation-specific costs into decision trees and evaluate financial savings. [8] examine how predictive technology changes the distribution of benefits in credit markets. These studies direct attention to consequences that an overall accuracy measure can conceal. A model can classify more applicants correctly while increasing expensive adverse approvals, and a reduction in aggregate financial loss can coexist with unequal outcomes across groups.

The research problem addressed here is therefore the incomplete translation of model performance into lending policy. Discrimination measures describe how well a score separates outcome classes. Calibration concerns the meaning of its probability estimates. A threshold converts those probabilities into actions, while an economic model attaches consequences to the resulting decisions. Evaluating these parts independently can obscure failures in the combined process. A strong classifier can be deployed with an unsuitable threshold, and a theoretically appropriate threshold can fail when the underlying probabilities are inaccurate.

This paper develops an analytical framework connecting these components. Its original work consists of feasible error allocations, loss bounds, conditions for comparing policies, and deterministic sensitivity calculations. Published research supplies empirical context and motivates the variables examined; it is not pooled into a synthetic dataset. The analysis does not claim a new machine learning algorithm or newly measured borrower-level performance. Instead, it establishes what information is necessary to assess the economic meaning of a classification result.

Three questions guide the study. How much financial uncertainty remains when accuracy and class prevalence are known but error composition is unavailable? Under what conditions does a more accurate model necessarily have lower benchmark loss? How do unequal costs and probability uncertainty alter the decision threshold and the consequences of acting on it? Addressing these questions offers a common framework for interpreting findings across conventional scoring, ensemble learning, and financial analytics research.

## 2. RELATED RESEARCH AND RESEARCH GAP

### 2.1 *Statistical and Ensemble Credit Scoring*

Comparative evidence supports testing multiple model families rather than selecting an algorithm solely because it is newer. [3] evaluate 41 classifiers across eight credit scoring datasets and report particularly favorable results for heterogeneous ensembles. Their study also considers financial implications and several evaluation measures. The breadth of that comparison provides a useful counterweight to conclusions based on one learner and one train-test division. It does not imply that the same ranking holds for every lending portfolio.

[4] examine classification under varying degrees of imbalance. Their findings favor Random Forest and gradient boosting relative to several alternatives in the studied settings, including pronounced imbalance. The underlying issue is important because a majority-class prediction can produce apparently respectable accuracy while failing to identify many adverse cases. [9] explains the Random Forest construction, in which randomized trees are aggregated to reduce reliance on a particular tree. The resulting flexibility is useful, but the fitted ensemble still needs validation under the intended operating conditions.

[10] provides a further credit scoring example using exploratory analysis, preprocessing, PySpark, decision trees, and Random Forest. For 1,000 German Credit records and 20 predictors, they report accuracies of 84.5% for the tuned forest and 71.6% for the restricted decision tree. Their use of an asymmetric penalty matrix makes the study relevant to economic evaluation. These are findings from their reported experiment; they are not new results of the present paper or directly comparable with performance percentages from unrelated datasets.

The collective implication is that ensembles are credible candidates for credit risk modeling, while their practical advantage must be established locally. Gradient boosting also merits consideration, and [11] provide a scalable implementation through XGBoost. A transparent statistical baseline remains valuable for identifying the incremental contribution of complexity. The purpose of a benchmark should be to identify a useful and reproducible decision process, rather than to confirm a preferred algorithm.

### 2.2 *Financial Information and Economic Consequences*

[5] combine customer transactions and credit bureau information from a commercial bank to develop out-of-sample consumer risk forecasts. Under their specified assumptions about credit-line decisions, they estimate potential savings of 6% to 25% of total losses. This is evidence that forecasts can have economic value when connected to an explicit action. The range is conditional on their setting and assumptions; it cannot be transferred as an expected savings rate for a different portfolio.

[6] show that digital footprints offer useful predictive information and can complement established credit information. This finding broadens the possible information base for scoring, particularly when conventional histories are limited. It also introduces questions about data availability, stability, and whether observed relationships will persist. Greater predictive information should be evaluated through prospective decision performance rather than treated as a sufficient reason to collect every accessible applicant attribute.

Cost-sensitive research addresses the financial objective more directly. [12] develops the foundations of classification with unequal costs. [7] incorporate example-dependent costs into tree construction and pruning and report improvements in savings in their

applications. Together, these studies justify treating financial consequences as part of the learning and decision problem. They also show why a constant error penalty is only a starting point when loans differ in exposure and recovery prospects.

### 2.3 *Discrimination and Probability of Reliability*

ROC analysis describes performance across thresholds, allowing comparison of ranking behavior [13]. Precision-recall analysis adds information about performance in a class of interest when classes are unevenly represented [14]. These perspectives are useful for screening models, but neither directly verifies that predicted probabilities agree with observed frequencies. A score may rank borrowers well while systematically overstating or understating their risks.

Probability of reliability deserves separate treatment because financial decisions often depend on the absolute level of estimated risk. [15] provides a squared-error score for probabilistic forecasts, and [16] explain the role of strictly proper scoring rules. [17] find that different supervised learners produce different probability distortions and examine calibration methods. [18] provide additional evidence of calibration problems in neural networks, a different model family. Their work reinforces the need for assessment, without establishing a universal calibration method for lending.

### 2.4 *Explanations and Distributional Outcomes*

Model explanations can help an analyst inspect the information driving a prediction. [19] develop local explanatory approximations, while [20] formulate an additive attribution approach. [21] discuss explainable methods in financial risk management. Such tools can assist investigation and communication, but an explanation of a model's output does not establish that the underlying decision rule is economically suitable or that an input has a causal effect on repayment.

Distributional effects create an additional evaluation requirement. [8] find that Black and Hispanic borrowers are disproportionately less likely to benefit from the introduction of machine learning in their studied credit-market setting. [22] explain mechanisms through which data-based decisions can reproduce disadvantages. [23] formalize an equality-of-opportunity criterion, while [24] show that certain risk-score fairness conditions cannot generally be satisfied together. These results support explicit reporting of the chosen objectives and their consequences.

### 2.5 *The Unresolved Connection Between Prediction and Policy*

Literature provides strong foundations for model comparison, economic evaluation, and responsible use. The gap pursued in this paper is more specific than the absence of another classifier. It concerns how to interpret a performance summary when the operating threshold, error composition, and probability reliability are only partly known. An accuracy increase can be clear while its economic implications remain unidentified. Likewise, evidence that an algorithm performs well in a benchmark does not specify how it should be operated in a different portfolio.

The framework developed below connects those missing quantities through a single set of definitions. It provides general loss bounds, explicit comparison conditions, and a threshold uncertainty analysis. These are established decision-theoretically building blocks assembled into a reproducible credit assessment study. The contribution lies in their integrated application and interpretation, rather than a claim that unequal-cost classification or calibration is itself new.

## 3. RESEARCH DESIGN AND ANALYTICAL FRAMEWORK

### 3.1 *Study Design and Analytical Inputs*

The study combines a focused interpretation of published findings with original mathematical analysis and deterministic scenario evaluation. The literature establishes the relevance of statistical models, ensembles, alternative information, asymmetric costs, and distributional effects. The analytical results are calculated independently of any one published model. No participant records are pooled across studies, and no meta-analytic effect size is estimated. Differences in populations and evaluation procedures make a common accuracy ranking across those publications inappropriate.

The scenario design varies adverse-class prevalence, relative error cost, and accuracy. The principal grid uses prevalence values of 0.05, 0.10, 0.20, 0.30, 0.40, and 0.50; relative adverse-approval penalties of 1, 2, 5, and 10; and accuracies of 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95. This produces 144 parameter combinations. Feasible error allocations are determined analytically for every combination. Smooth curves in the figures use denser grids for display, with the same equations.

These inputs are deliberately transparent scenarios, not estimated population parameters. The 30% prevalence and five-to-one cost ratio also provide a familiar benchmark connection to German Credit, whose repository includes an asymmetric cost matrix [25]. The broader grid prevents the analysis from depending on that one case. All financial results are expressed in normalized penalty units unless an explicit payoff extension is being discussed.

Figure 1 presents the framework. Published findings inform the choice of modeling candidates and evaluation questions. The analytical branch identifies what can be deduced from accuracy, prevalence, costs, and probabilities. The decision layer then combines those deductions with the evidence required from future borrower-level validation. The diagram does not represent a newly trained or deployed scoring system.

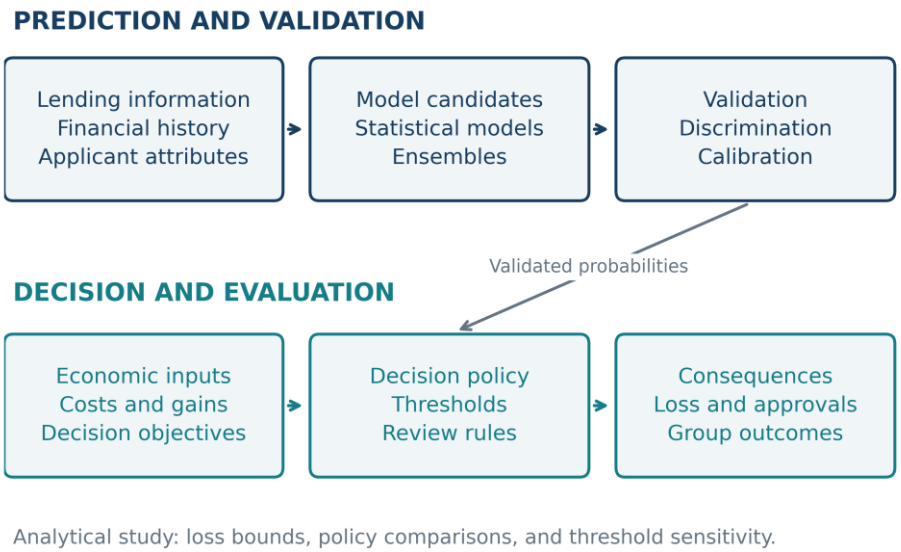



Figure 1. Framework Linking Credit Prediction with Lending Decisions

### *3.2 Outcome and Action Definitions*

Let Y = 1 denote an adverse credit outcome within a specified assessment horizon, and Y = 0 denote a favorable outcome. In an empirical application, the adverse event must be defined precisely, such as a specified delinquency level or default. For the present binary analysis, a prediction of 1 lead to rejection and a prediction of 0 leads to approval. This convention keeps the positive class aligned with the adverse event.

Let FN be adverse applicants approved and FP be favorable applicants rejected. TP and TN are correctly rejected adverse applicants and correctly approved favorable applicants. Let N denote the number of applications and pi the adverse-class prevalence. An adverse approval receives penalty c, a favorable rejection receives penalty one, and both correct decisions receive zero penalty. The resulting loss is a benchmark weighted error, not a complete accounting measure of bank profit.

Accuracy A and average penalty ell are:

$$A = (TP + TN)/N; \quad ell = (c\,FN + FP)/N. \quad (1)$$

Define e = 1 - A, f = FN/N, and g = FP/N. Since f + g = e,

$$ell = e + (c - 1)f. \quad (2)$$

Equation 2 makes the decision problem explicit. Accuracy determines total error mass e, while the allocation f determines how much of it receives the larger penalty. Unless c = 1, knowledge of accuracy alone does not identify average loss.

### *3.3 Feasible Loss Bounds*

Class sizes restrict how the errors can be distributed. Adverse approvals cannot exceed adverse applicants, and favorable rejections cannot exceed favorable applicants. Thus,

$$f_{min} = \max(0, e - (1 - pi)); \quad f_{max} = \min(e, pi). \quad (3)$$

For c >= 1, the corresponding loss bounds are

$$ell_{min} = e + (c - 1)f_{min}; \quad ell_{max} = e + (c - 1)f_{max}. \quad (4)$$

These bounds are sharp for normalized error allocations: each endpoint corresponds to a feasible confusion allocation under the specified inputs. Their interpretation is identification uncertainty, arising from unknown error composition. They are not confidence intervals, prediction intervals, or distributions of expected outcomes. Finite samples also impose integer-count restrictions; normalized scenarios avoid pretending that every decimal combination describes an actual dataset.

When e is no larger than both class proportions, Equation 4 simplifies to e <= ell <= ce. The spread is (c - 1)e. Economic ambiguity therefore grows with both error frequency and cost asymmetry. When one class is small, class-size constraints can narrow the envelope. This dependence is examined explicitly rather than assuming that every possible error composition remains feasible at every prevalence.

### *3.4 Reference Policies and Economic Dominance*

Approving every application yields average loss c pi and accuracy 1 - pi. Rejecting every application yields average loss 1 - pi and accuracy pi. They have equal benchmark loss when:

$$pi = 1/(c + 1). \quad (5)$$

These policies are reference points, not recommendations for responsible lending. They identify whether a learned policy improves on a trivial decision under the chosen cost model. Figure 2 maps the reference-policy boundary as prevalence and costs vary. Above the boundary, reject-all has lower benchmark loss; below it, approve-all has lower benchmark loss.

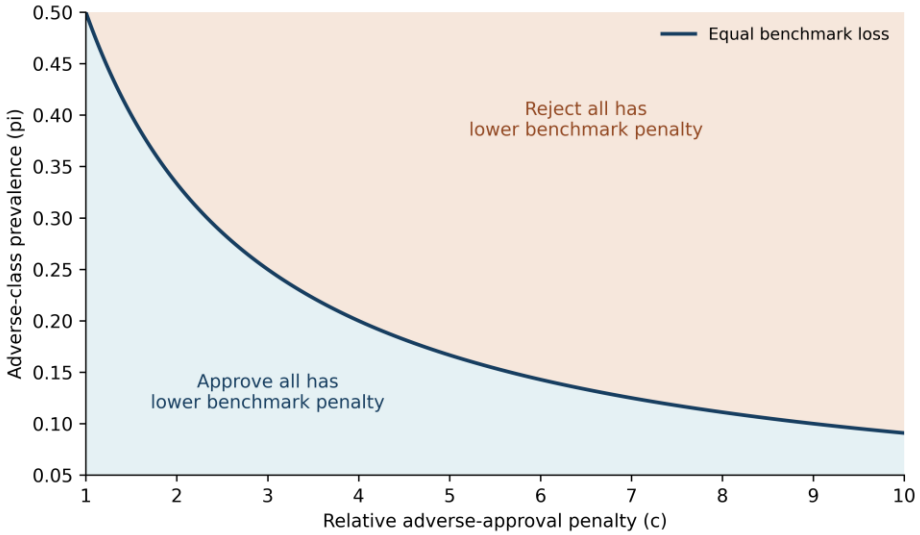



Figure 2. Reference Policy Boundary Under Changing Prevalence and Costs

For two models, let model 1 have higher accuracy than model 2, so e_1 < e_2. Their average loss difference is:

ell_1 - ell_2 = (e_1 - e_2) + (c - 1)(f_1 - f_2). (6)

For c > 1, model 1 has lower loss exactly when f_1 - f_2 < (e_2 - e_1)/(c - 1). A sufficient and necessary condition for strict dominance over every feasible pairing of error allocations is ell_1,max < ell_2,min. This stronger requirement distinguishes a model that is actually cheaper at observed operating points from one that is guaranteed cheaper using only aggregate information.

### 3.5 *Threshold Selection and Probability Uncertainty*

Let p be an applicant's calibrated adverse-event probability. Approval has expected penalty cp, while rejection has expected penalty 1 - p. The loss-minimizing action is rejection when:

p > t; t = 1/(c + 1). (7)

If the costs vary across applicants, with adverse-approval cost c_i and favorable-rejection cost r_i, the threshold becomes r_i/(c_i + r_i). These expressions assume zero penalties for correct decisions and positive error costs. A richer payoff matrix requires a new comparison of expected action values. Neither equation establishes the correct threshold for a bank without a defensible specification of those costs.

Suppose a decision uses an estimated probability s instead of the true conditional probability p. If s and p lie on opposite sides of t, the expected excess penalty, or regret, is:

R(p, s) = |(c + 1)p - 1| = (c + 1)|p - t|. (8)

Otherwise regret is zero, with threshold ties having equal expected penalty. If the pointwise probability error satisfies |s - p| <= epsilon, an incorrect action can occur only within epsilon of t, and regret is bounded above by (c + 1)epsilon. This is a conditional theoretical bound. Aggregate calibration metrics alone do not establish a pointwise error guarantee for every applicant.

### 3.6 *Reproducibility and Empirical Evaluation Requirements*

The accompanying code evaluates the 144 scenario combinations, records the inputs, verifies feasible bounds, and produces all four figures. No random seed is necessary for these deterministic calculations. Illustrative confusion allocations are explicitly identified as constructed examples, and none is attributed to a published study. The code does not train classifiers or recreate confidential lending outcomes.

A borrower-level extension would require a validation design that separates model selection from final

evaluation. [26] show how using cross-validation for selection and error estimation can create optimistic results. [27] analyze the broader problem of selection overfitting. These findings justify performing feature selection, tuning, calibration, and threshold choice within training data, while keeping final evaluation observations untouched.

Statistical comparisons should also respect their unit of analysis. [28] addresses comparisons of classifiers over multiple datasets. Repeated folds from the same sample are dependent and should not simply be treated as independent datasets. In an applied extension, uncertainty estimates should accompany policy differences, and investigators should retain predictions and scores so that economic results can be independently checked.

## 4. ANALYTICAL RESULTS

### *4.1 How Prevalence Changes The Reference Policy*

The 24 prevalence-cost combinations identify several changes in the preferred constant policy. At c = 1, the crossover prevalence is 50%. At c = 2, it is one third; at c = 5, one sixth; and at c = 10, approximately 9.09%. A population can therefore have a clear majority of favorable outcomes while a reject-all policy has lower benchmark penalty than approve-all. This occurs because the expensive errors can outweigh the more numerous inexpensive opportunities lost.

For example, at pi = 0.10 and c = 5, approve-all incurs 0.50 units per application, below the reject-all value of 0.90. At pi = 0.30 with the same cost ratio, approve-all incurs 1.50, above the reject-all value of 0.70. These examples demonstrate why comparing a learner only with a majority-class accuracy baseline leaves the economic comparison incomplete. The relevant baseline depends on the stated objective and population.

Figure 2 shows a continuous boundary rather than an estimated relationship. The region labels follow exactly from Equation 5. They do not describe actual approval policies, lending volumes, or realized bank behavior. Their value is to establish a transparent reference against which a model's observed decisions could later be evaluated.

### *4.2 Loss Uncertainty at A Fixed Accuracy*

At 85% accuracy, e = 0.15. With pi = 0.30 and c = 5, Equation 4 gives ell between 0.15 and 0.75, or 150 to 750 units per 1,000 applications. Both endpoints are feasible because the favorable and adverse classes each contain enough observations to accommodate all the errors. The ratio between the endpoints is five, matching the ratio between the two error penalties. The underlying distinction is which errors the model makes.

The same accuracy produces a narrower envelope when the adverse class is rare. At pi = 0.05 and c = 5, f cannot exceed 0.05, so average loss ranges from 0.15 to 0.35. At pi = 0.10, the upper limit is 0.55. Once adverse prevalence reaches 0.15 and favorable prevalence remains at least 0.15, the upper limit reaches 0.75. These are class-size constraints, not evidence that a rare-event classifier performs better.

Figure 3 presents loss envelopes for three accuracy scenarios across the prevalence range. The bands narrow when the smaller class cannot accommodate all errors. At c = 1, every envelope collapses to the error rate, confirming that accuracy identifies this loss only when errors carry equal weight. At c = 10, the envelope for 85% accuracy and 30% prevalence expands to 0.15 through 1.50, showing how greater asymmetry increases the need for error-specific reporting.

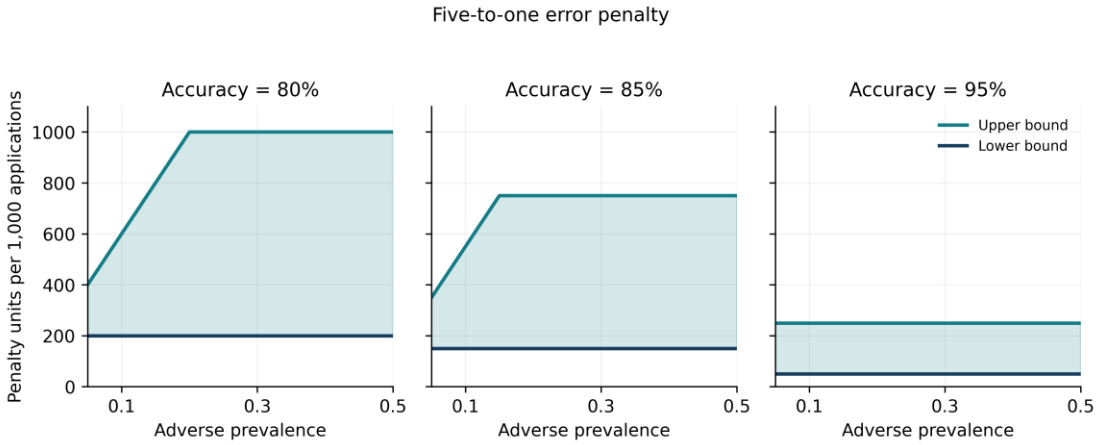



Figure 3. Feasible Loss Envelopes Across Accuracy and Prevalence Scenarios

### *4.3 A Reversal Between Accuracy and Economic Ranking*

Consider 1,000 normalized applications with 300 adverse and 700 favorable outcomes. Model A makes 150 adverse approvals and no favorable rejections. It has TP = 150 and TN = 700, giving 85% accuracy and loss of 750 at c = 5. Model B makes no adverse approvals and 200 favorable rejections. It has TP = 300 and TN = 500, giving 80% accuracy and loss of 200. Both confusion allocations are feasible in the stated population.

Thus, a five-percentage-point accuracy advantage can coexist with 550 additional penalty units. This constructed example is a counterexample to a general ranking rule, not a claim about how actual Random Forest or regression models behave. Its purpose is to show that improved accuracy does not logically entail improved economic outcomes. Naming an algorithm would add no evidence because the result follows from error allocation rather than model architecture.

Equation 6 gives the exact comparison condition. With e_A = 0.15, e_B = 0.20, and c = 5, the more accurate model is cheaper only if f_A - f_B < 0.0125. Equivalently, per 1,000 applications, it may make fewer than 12.5 additional adverse approvals. Its accuracy advantage can be overcome by a relatively small movement toward the more expensive error. Actual policy selection therefore requires the relevant joint predictions or confusion counts.

### *4.4 Conditions For A Conclusion from Aggregate Results*

Loss envelopes can sometimes establish a ranking even without exact confusion counts. At 30% adverse prevalence, c = 2, and accuracies of 95% and 80%, the first model's loss is between 0.05 and 0.10, while the second model's is between 0.20 and 0.40. The first is strictly cheaper for every feasible allocation. In this setting the accuracy advantage is large enough, relative to cost asymmetry, to resolve the economic comparison.

At c = 5, the corresponding intervals are 0.05 to 0.25 and 0.20 to 1.00. They overlap, so accuracy and prevalence no longer guarantee the ranking. This does not mean the better classifier is unlikely to have lower loss. It means the available aggregate information permits both rankings. The distinction between uncertainty about a ranking and evidence against a model is essential when interpreting incomplete performance reports.

The same reasoning applies to constant policies. For a model to be guaranteed cheaper than the better constant policy, its upper loss bound must be below min(c pi, 1 - pi). When that condition fails, observed error composition may still establish a benefit. Bounds are therefore a screening device

for deciding what can already be concluded and what additional evidence is needed.

### 4.5 *Threshold Sensitivity and Uncertainty Near The Decision Boundary*

The optimal benchmark threshold decreases as adverse approval becomes more expensive. It equals 0.50 at c = 1, approximately 0.333 at c = 2, 0.167 at c = 5, and 0.091 at c = 10. These values follow from expected-loss minimization and calibrated probabilities. They do not imply corresponding approval rates because the distribution of applicant probabilities has not been specified.

Figure 4 adds a probability uncertainty band of epsilon = 0.05 around the theoretical threshold. At c = 5, this band extends from approximately 0.117 to 0.217. Under the assumed pointwise error limit, decisions outside the band cannot cross the true optimal boundary merely because of probability error. Inside it, an action can change, so probability reliability has a direct decision consequence. The band is an analytical construction, not an estimated confidence interval.

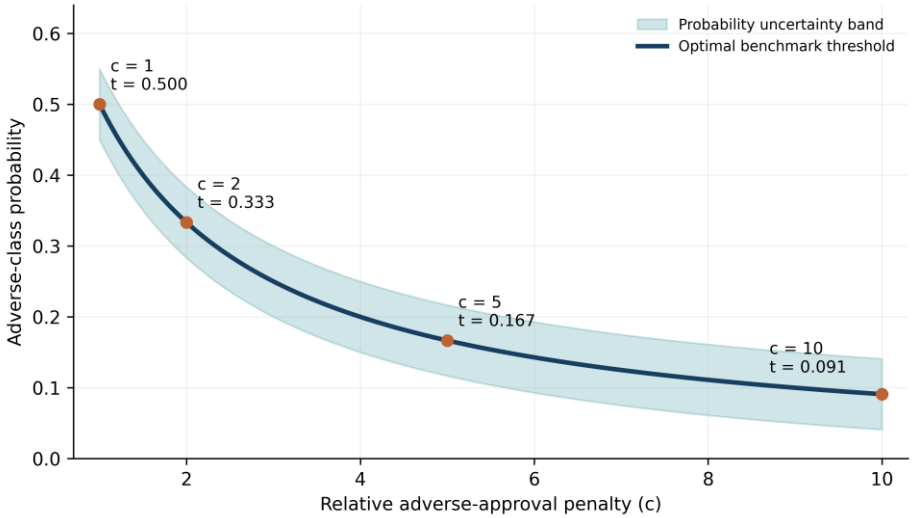



Figure 4. Threshold Sensitivity with A Probability Uncertainty Band

For a concrete example, let p = 0.20 while the estimated score is s = 0.15 at c = 5. The optimal action under the benchmark is rejection, with expected penalty 0.80. Acting on s would approve, producing expected penalty 1.00. Regret is 0.20, consistent with Equation 8. The example also meets the assumed probability error limit of 0.05. It shows why the location of probability error matters, rather than merely its average magnitude across a portfolio.

At c = 5 and epsilon = 0.05, the general regret bound is 0.30 units per application. At c = 10, it becomes 0.55. These upper limits are conditional on a strong pointwise assumption that should not be confused with good average calibration. They identify the information required for a decision guarantee, not a guarantee established for any trained model in this study.

## 5. DISCUSSION

### 5.1 *Connecting Empirical Findings with Decision Consequences*

The analytical results help interpret several strands of credit risk research without collapsing them into one performance leaderboard. Evidence that ensembles perform well supports them in comparative modeling. Evidence that behavioral information improves forecasting supports evaluating its incremental contribution. Evidence that cost-aware learning improves savings

supports aligning the training objective with the business decision. Each finding addresses part of the process, while the present equations explain how their practical implications depend on other quantities.

This perspective also clarifies why published percentages should retain their original context. The accuracy difference reported by [10] is evidence from a particular German Credit experiment. The estimated savings reported by [5] concern their credit-line intervention assumptions. The benchmark results of [3] span several datasets and measures. None is a numerical substitute for another. The appropriate synthesis is to compare what each study establishes and use that evidence to design a locally valid evaluation.

The framework gives a precise reason to report economic and statistical outcomes together. Accuracy alone can leave a wide loss envelope, and a higher score may fail to establish an economic ranking. Conversely, sufficiently separated envelopes can establish dominance under stated assumptions. This distinction allows researchers to make stronger claims when the evidence permits them and more conditional claims when important information remains unavailable.

### *5.2 Model Selection Should Evaluate A Complete Policy*

A credit decision policy includes the information collected, the fitted model, its probability adjustment, and the rule converting scores into actions. Two policies using the same model can produce different outcomes through different thresholds. Likewise, two models with similar ranking performance can differ in calibration and therefore in decisions made at a fixed cost-derived threshold. Selecting a learner independently of the operating rule can therefore miss the actual source of economic performance.

An empirical evaluation should fix the adverse event and prediction horizon before fitting. Applicant information must be available at the decision time; post-approval variables or later outcomes can introduce leakage. Models should then be compared under common partitions, with all preprocessing and selection fitted within training data. A conventional statistical model, a single tree, Random Forest, and a boosting model would provide a useful range of complexities. The final choice should follow the prespecified decision objective and the uncertainty around estimated differences.

The reporting set should include confusion counts, discrimination, probability quality, approval rate, and the stated financial objective. ROC AUC and an explicitly defined precision-recall summary describe different aspects of ranking. Brier score and calibration diagnostics assess probability behavior. Average benchmark penalty connects predictions with unequal costs. When financial exposure information is available, an applicant-level payoff analysis is preferable to assuming that every adverse approval has the same consequence.

### *5.3 Moving from Benchmark Penalties to Lending Value*

The zero-diagonal penalty matrix intentionally simplifies the economics. Approving a favorable borrower can generate income, while rejecting an adverse borrower can avoid a loss but also forgo other potential value. Exposure, recovery, funding, servicing, and pricing can vary across applicants. A realistic decision model should assign values to every action-outcome combination and document which items are observed, estimated, or assumed.

For applicant i, one possible extension compares expected approval value, (1 - p_i)G_i - p_i D_i - K_i, with the value assigned to rejection. Here G_i represents the gain from a favorable

approved loan, D_i the loss from an adverse approved loan, and K_i a specified processing or funding cost under the chosen convention. This expression is illustrative and requires careful accounting to avoid double-counting. It shows why a universal threshold may be unsuitable when loan sizes and economic conditions differ.

Cost-aware evaluation can also reveal where additional information has value. A variable that changes few decisions may add little economic benefit even if it improves a global ranking measure. A modest probability correction around the operating threshold may have greater practical value. This follows from the regret expression: decision consequences depend on the margin between the true probability and the action boundary, as well as on cost asymmetry.

### *5.4 Explainability and Fair Access to Credit*

Economic improvement should not be treated as a complete account of institutional or applicant welfare. The distributional findings in credit-market research show that technology can change who benefits from lending decisions. Researchers should therefore report relevant group outcomes alongside aggregate loss where the data and sample sizes permit meaningful analysis. Differences in approval rates, favorable rejections, and adverse approvals answer different questions and should be identified explicitly.

Explanations are useful for investigating these outcomes, but they need their own validation. Impurity-based feature importance can favor certain variable structures, as shown by [29]. Local attribution methods may change with the model, background sample, or correlated predictors. A clear reason for a prediction should therefore be accompanied by checks on stability and the meaning of the underlying variable. A predictive association is insufficient evidence that changing that variable would improve a borrower's outcome.

The decision framework can accommodate constraints on group performance, review capacity, or access objectives, but those constraints must be specified rather than assumed. A policy minimizing one weighted error function may conflict with another institutional objective. The analytical bounds describe the stated objective only; they do not remove the need to choose and justify the values embedded in the decision process.

### *5.5 Computational Scale and Changing Populations*

Big data infrastructure can support large training samples and repeatable processing. [30] describes Apache Spark as a unified distributed processing engine. Its usefulness for a particular credit workload should still be demonstrated through measured runtime, memory use, and throughput. Installing a distributed framework or processing a small benchmark with it does not by itself establish a computational contribution.

Population change also affects the decision boundary's interpretation. A cost ratio can change when recovery prospects or funding conditions change. A model's probability estimates can lose reliability when applicant composition or repayment behavior changes. The sensitivity results show why both economic inputs and predictive reliability need monitoring. More observations may reduce estimation uncertainty, but they do not guarantee perfect classification or preserve performance under a different economic environment.

## 6. LIMITATIONS AND FURTHER RESEARCH

This study provides analytical results and deterministic scenarios rather than a new borrower-level model comparison. It does not estimate realized savings, approve loans, or validate a deployed scoring system.

Published studies are used selectively to motivate the research problem, not as an exhaustive systematic review. Their results retain their original populations and assumptions, and no pooled performance estimate is produced.

The scenario grid is intentionally manageable and does not represent the full range of lending markets. The loss bounds assume binary outcomes, fixed class prevalence for each scenario, and known relative penalties. They exclude uncertainty in the costs themselves. Finite datasets impose integer restrictions that normalized allocations do not capture. The resulting envelopes identify feasible economic consequences under stated inputs; they do not tell how likely each allocation is for a particular learning algorithm.

The probability of uncertainty analysis also relies on a strong assumption. A pointwise error bound is more demanding than aggregate calibration, and this paper does not establish one for any model. A practical extension could evaluate empirical decision instability near the threshold while reporting uncertainty from limited sample sizes. Temporal validation would be needed to determine whether a selected calibration and threshold procedure remains reliable as the lending environment changes.

Further research should evaluate the complete framework on several lending datasets with documented outcome horizons and applicant-level cost information. Candidate policies should be compared using shared partitions, retained score files, and prespecified economic objectives. External validation would test portability across institutions. Separate analyses should examine the stability of explanations and distributional outcomes. These experiments would assess how much analytical uncertainty can be reduced with better empirical reporting.

## 7. CONCLUSION

Machine learning offers useful tools for credit risk prediction, and published research supports evaluating ensembles, behavioral information, and cost-sensitive methods. Their operational value depends on how predictions are converted into lending decisions. The framework developed here makes that connection explicit through feasible error allocations, loss bounds, economic comparison conditions, and probability thresholds.

The deterministic results show that accuracy can leave substantial uncertainty about financial consequences. Under one transparent scenario, a classifier with 85% accuracy can incur more benchmark loss than a feasible classifier with 80% accuracy. The analysis also identifies cases in which aggregate results are sufficient to establish economic dominance. Probability uncertainty near the operating threshold introduces another decision consequence that global discrimination metrics do not resolve.

Credit scoring research should therefore evaluate the combined model and action policy, preserve the context of published findings, and report financial and distributional outcomes alongside predictive performance. Empirical validation remains necessary, but explicit decision analysis helps determine what that validation must establish and prevents a statistical improvement from being mistaken for an economic conclusion.

## REFERENCES

[1] D. J. Hand and W. E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: a Review," *J. R. Stat. Soc. Ser. A (Statistics Soc.*, vol. 160, no. 3, pp. 523–541, 1997, doi: https://doi.org/10.1111/j.1467-985X.1997.00078.x.

[2] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *J. Oper. Res. Soc.*, vol. 54, no. 6, pp. 627–635, 2003, doi: https://doi.org/10.1057/palgrave.jors.2601545.

[3] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015, doi: https://doi.org/10.1016/j.ejor.2015.05.030.

[4] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012, doi: https://doi.org/10.1016/j.eswa.2011.09.033.

[5] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *J. Bank. Financ.*, vol. 34, no. 11, pp. 2767–2787, 2010, doi: https://doi.org/10.1016/j.jbankfin.2010.06.001.

[6] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," *Rev. Financ. Stud.*, vol. 33, no. 7, pp. 2845–2897, 2020, doi: https://doi.org/10.1093/rfs/hhz099.

[7] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6609–6619, 2015, doi: https://doi.org/10.1016/j.eswa.2015.04.042.

[8] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *J. Finance*, vol. 77, no. 1, pp. 5–47, 2022, doi: https://doi.org/10.1111/jofi.13090.

[9] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[10] M. M. T. G. Manik, A. S. M. Saimon, M. S. Islam, M. Moniruzzaman, E. Rozario, and E. Hossin, *Big Data Analytics for Credit Risk Assessment*. 2025. doi: https://doi.org/10.1109/ICMLAS64557.2025.10967667.

[11] T. Chen, "XGBoost: A Scalable Tree Boosting System," *Cornell Univ.*, 2016, doi: https://doi.org/10.1145/2939672.2939785.

[12] C. Elkan, "The Foundations of Cost-Sensitive Learning," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. pp. 973–978, 2001. [Online]. Available: https://mlanthology.org/ijcai/2001/elkan2001ijcai-foundations/

[13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[14] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015, doi: https://doi.org/10.1371/journal.pone.0118432.

[15] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950, doi: https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[16] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007, doi: https://doi.org/10.1198/016214506000001437.

[17] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," *Proceedings of the 22nd International Conference on Machine Learning*. ACM, pp. 625–632, 2005. doi: https://doi.org/10.1145/1102351.1102430.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. PMLR, pp. 1321–1330, 2017. doi: https://proceedings.mlr.press/v70/guo17a.html.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi: https://doi.org/10.1145/2939672.2939778.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[21] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in Fintech Risk Management," *Front. Artif. Intell.*, vol. 3, p. 26, 2020, doi: https://doi.org/10.3389/frai.2020.00026.

[22] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016, doi: https://doi.org/10.15779/Z38BG31.

[23] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems 29*. 2016. doi: https://proceedings.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html.

[24] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. pp. 43:1-43:23, 2017. doi: https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

[25] H. Hofmann, "Statlog (German Credit Data)." UCI Machine Learning Repository, 1994. doi: https://doi.org/10.24432/C5NC77.

[26] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, p. 91, 2006, doi: https://doi.org/10.1186/1471-2105-7-91.

[27] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010, doi: https://jmlr.org/papers/v11/cawley10a.html.

[28] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006, doi: https://jmlr.org/papers/v7/demsar06a.html.

[29] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007, doi: https://doi.org/10.1186/1471-2105-8-25.

[30] M. Zaharia *et al.*, "Apache Spark: A Unified Engine for Big Data Processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016, doi: https://doi.org/10.1145/2934664.