

Beyond ROC-AUC: Decision-Adjusted Analytical Value for Capacity-Constrained Management Analytics under Temporal Distribution Shift

Author(s): Kwan Hong TAN

Singapore University of Social Sciences, Singapore.

Corresponding Author: khtan055@suss.edu.sg

Received: August, 2026, Published: October, 2026

ARTICLE INFO

Keywords:

Business analytics; Decision intelligence; Distribution shift; Model evaluation; Resource allocation

© 2026 by the Author(s). This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license, making research freely available to the public and supporting a greater global exchange of knowledge and human experiments.



ABSTRACT

Management analytics is commonly evaluated with model-centric statistics such as the area under the receiver operating characteristic curve (ROC-AUC), yet managerial deployment converts scores into costly actions under finite capacity, changing base rates, and asymmetric economic consequences. This study introduces Decision-Adjusted Analytical Value (DAAV), a normalized framework that evaluates a predictive policy against random-capacity and attainable-oracle benchmarks after incorporating an economic action threshold and an operational capacity constraint. The framework is tested on the UCI Bank Marketing dataset using 41,188 chronologically ordered campaign records. To prevent post-contact leakage, call duration is excluded. Logistic regression, random forest, and Extreme Gradient Boosting (XGBoost) models are trained on the first 75% of observations, calibrated on the next 10%, and evaluated on the final 15% temporal holdout. Performance is assessed across 20 benefit-cost and capacity scenarios, with paired nonparametric bootstrapping for the focal policy comparison. The temporal holdout exhibits a pronounced prevalence shift, from 6.8% in validation to 38.5% in testing. Random forest attains the highest ROC-AUC (0.653), but is not the value-maximizing model in 16 of 20 operating scenarios. At a benefit-cost ratio of 3 and 50% capacity, logistic regression produces 226.9 normalized net-value units per 1,000 cases versus 95.6 for random forest; the paired difference is 131.0 units (95% bootstrap interval: 110.4 to 151.3). Static Platt calibration fitted before the shift further reduces downstream value despite leaving discrimination unchanged. The findings show that ranking models by predictive accuracy can misallocate managerial attention when economic thresholds and resource constraints bind. DAAV provides managers, analytics teams, and governance functions with a transparent, decision-centered complement to conventional predictive metrics for model selection and deployment under temporal shift.

1.0 Introduction

Predictive analytics has become embedded in managerial processes ranging from customer targeting and credit triage to retention, procurement, staffing, and risk prioritization. Yet the analytical model is rarely the final decision. Managers act on scores by selecting a limited number of cases, setting action thresholds, allocating scarce attention, and accepting asymmetric consequences for false positives and false negatives. The distinction matters because a model can rank cases well while producing probability estimates that induce poor operational decisions. Conversely, a model with only modestly lower discrimination can create greater realized value if its score scale, threshold behavior, and coverage better match the economics and capacity of the deployment environment.

The dominant evaluation practice nevertheless remains model-centric. Receiver operating characteristic area under the curve (ROC-AUC), accuracy, F1 scores, and related measures summarize aspects of predictive performance, but they do not directly encode the

managerial value of the resulting action policy. Hand (2009) showed that ROC-AUC aggregates performance across implicit cost conditions that need not match a real decision problem, while Hernandez-Orallo et al. (2012) formalized how threshold choice changes expected classification loss. Earlier cost-sensitive evaluation research likewise emphasized that classifier choice depends on class distributions, costs, and operating conditions (Provost & Fawcett, 2001; Drummond & Holte, 2006). In management settings, this problem is intensified by capacity constraints. A marketing team may contact only 10% or 30% of a lead pool, an assurance function may review only a fixed number of alerts, and an executive team may fund only a limited set of interventions. Model evaluation therefore becomes inseparable from resource allocation.

A second difficulty is temporal instability. Models are commonly trained and calibrated on historical data, but the base rate of the target event, the mix of customers, and the mapping from features to outcomes can change. Ovadia et al. (2019) demonstrate that predictive uncertainty can deteriorate under dataset shift, and Tan (2026) similarly emphasizes the need for shift-responsive reliability controls in selective classification. Static post-hoc calibration may improve probability quality on a validation distribution but need not remain appropriate after a material change in prevalence. For managers, the consequence is not merely statistical miscalibration. A shifted score distribution can cause a nominally profitable threshold to contact too few or too many cases, thereby changing the use of scarce capacity and the realized economics of an analytics program.

This paper addresses the prediction-to-action gap by introducing Decision-Adjusted Analytical Value (DAAV), a decision-centered evaluation framework for binary management analytics. DAAV combines predicted probabilities, an explicit benefit-cost ratio, a maximum operational capacity, and realized outcomes. It evaluates the net value of the induced action policy relative to an expected random-capacity benchmark and an attainable-oracle benchmark. The statistic is intended as a complement to, rather than a replacement for, discrimination and calibration metrics. Its purpose is to make the operational assumptions of model selection explicit and auditable.

The framework is evaluated using the UCI Bank Marketing dataset, a business classification dataset containing direct-marketing records from a Portuguese bank (Moro et al., 2014a). The full bank-additional file contains 41,188 observations ordered by date from May 2008 to November 2010. Three standard model families are compared under a chronological train-validation-test design: logistic regression, random forest, and XGBoost. The analysis then subjects these models to 20 decision scenarios that vary both the benefit-cost ratio and the maximum share of cases that can be contacted. This design is deliberately model-agnostic. It isolates an evaluation problem that can occur even when organizations use familiar, well-understood predictive models.

The study makes four contributions. First, it formalizes a post hoc value metric that combines economic screening and finite capacity in a single auditable evaluation rule. Second, it distinguishes DAAV from expected-cost, cost-curve, decision-curve, and decision-focused learning approaches by using realized policy value and scenario-specific random-capacity and attainable-oracle normalization. Third, it documents utilization-sensitive rank reversal, in which a model with a slightly higher discrimination point estimate can create less realized value because too few cases clear the economic threshold. Fourth, it shows that pre-shift probability calibration can preserve discrimination while reducing downstream policy value after a substantial change in temporal prevalence.

The study seeks to answer the following research questions:

RQ1: Does the model with the best conventional predictive discrimination also maximize realized managerial value?

RQ2: How do model rankings vary under different economic payoff structures and operational capacity constraints?

RQ3: Does static probability calibration fitted before a temporal prevalence shift necessarily improve downstream decision value?

The remainder of the paper is structured as follows. Section 2 reviews predictive, cost-sensitive, decision-theoretic, and capacity-aware evaluation and positions DAAV relative to those traditions. Section 3 describes the data, models, action rule, DAAV formulation, and uncertainty analysis. Section 4 reports the temporal shift, predictive and decision results, sensitivity analyses, and managerial implications. Section 5 concludes with the main contribution and scope conditions.

2.0 Literature Review

2.1 From Predictive Performance to Decision Performance

The conceptual separation between prediction and decision has a long history in statistics, economics, and operations research. In modern machine learning, a prediction produces an estimate of an unknown quantity, whereas a decision maps that estimate into an action under an objective and a set of constraints. Kleinberg et al. (2018) make this distinction explicit in a consequential human-decision setting, arguing that the value of machine prediction can be assessed only after the link among prediction, preferences, and counterfactual outcomes is specified. Elmachtoub and Grigas (2022) formalize the same principle in predict-then-optimize problems by showing that minimizing prediction error can be misaligned with minimizing downstream decision loss.

Management analytics often retains the predictive metric as the terminal criterion. A model development team may select the highest ROC-AUC model and pass its scores to an operations unit, where thresholds and workload limits are added later. This modular practice is convenient, but it can hide an objective mismatch. ROC-AUC is threshold-independent in the sense that it summarizes ranking performance across thresholds; managerial action is threshold-dependent because a particular cost structure and resource limit determine which cases are acted upon. Hand (2009) criticizes ROC-AUC as a universal comparison measure because its implicit weighting across misclassification costs varies with the score distributions being compared. Hernandez-Orallo et al. (2012) further show that classifier performance depends on how thresholds are chosen and how operating conditions are distributed.

Decision curve analysis offers a related precedent for translating predictive output into net benefit. Vickers and Elkin (2006) evaluate prediction models based on the consequences of threshold probabilities rather than on discrimination alone. The present study adopts the same broad decision-theoretic intuition but modifies it for management analytics by making finite action capacity explicit and by normalizing realized value against capacity-matched random and attainable-oracle benchmarks.

2.2 Cost-Sensitive Evaluation and the Position of DAAV

The closest intellectual antecedents to DAAV are cost-sensitive classifier evaluation methods. Provost and Fawcett (2001) show that classifier performance should be assessed against uncertain class distributions and misclassification costs, while Drummond and Holte (2006) use cost curves to visualize expected cost over changing operating conditions. These approaches establish that no classifier is uniformly best when costs and priors vary. DAAV extends that logic to a managerial policy in which an economic threshold is applied before a hard capacity ceiling. Its target is not expected classification cost over a distribution of possible operating conditions, but realized net decision value for a specified operating scenario.

DAAV also differs from decision-focused and predict-then-optimize methods. Task-based and decision-focused learning explicitly trains predictive models against downstream optimization objectives (Donti et al., 2017; Wilder et al., 2019), and the broader end-to-end constrained optimization literature integrates learning and optimization during model

construction (Kotary et al., 2021). A recent survey and benchmark by Mandi et al. (2024) consolidates this decision-focused learning literature and emphasizes downstream decision quality as the training objective. Smart Predict-then-Optimize similarly modifies training loss to reflect downstream optimization quality (Elmachtoub & Grigas, 2022). DAAV does not alter the learning objective or require retraining. It is an ex post evaluation layer for already trained score-producing models, designed to compare model-policy combinations under explicit payoff and capacity assumptions. This narrower scope is useful in governance settings where model development and deployment approval are institutionally separated.

2.3 Capacity Constraints as a Managerial Evaluation Variable

Operational capacity is not a peripheral implementation detail. It changes the feasible policy set. In direct marketing, call-center hours and staffing cap the number of customers that can be contacted. In fraud operations, investigators can inspect only a subset of alerts. In procurement, managers can renegotiate only a limited number of supplier contracts at one time. Under such conditions, the same predictive model can induce materially different value at 10% versus 50% capacity. A purely rank-based metric does not reveal whether economically eligible cases are concentrated within the available capacity or whether the probability scale leaves profitable capacity unused.

This distinction creates what this paper terms utilization-sensitive rank reversal. Suppose Model A has better overall ranking but produces conservative probabilities after a distribution shift. If the economic threshold is applied before the capacity cap, Model A may classify relatively few cases as eligible, even when the organization can act on more cases. Model B may have slightly lower ranking quality but assign more true opportunities above the economic threshold. Model B can then generate greater total value despite lower precision among selected cases. The question is not which model is intrinsically superior. The question is which model-policy combination produces more value under the current economics and operational envelope.

2.4 Calibration, Prior Shift, and Action Thresholds

Probability calibration is central when thresholds represent economic break-even points. A calibrated probability of 0.20 should correspond, within an appropriate reference population, to approximately a 20% event frequency. Brier scores and logarithmic loss assess probabilistic accuracy, and post-hoc methods such as Platt scaling or temperature scaling are widely used to improve calibration (Guo et al., 2017). Yet calibration is distribution-dependent. If the class prior or conditional distributions change after calibration, the same mapping between score and event probability may no longer hold.

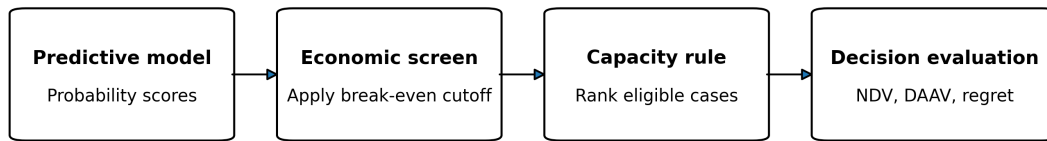
Prior-probability shift is especially relevant to management systems because demand, campaigns, economic conditions, and customer selection practices change over time. Saerens et al. (2002) show that classifier outputs may require adjustment when class priors change. Ovadia et al. (2019) demonstrate more broadly that uncertainty quality degrades under dataset shift. The managerial consequence is direct: when an economic threshold is expressed in probability units, miscalibration changes the eligibility of actions. A model can retain an identical ROC-AUC after monotonic calibration while inducing a very different number of actions. This makes calibration both a policy variable and a statistical property.

2.5 Research Gap and Contribution

Existing literature provides important components of the problem: management research explains changing human-AI decision structures (Shrestha et al., 2019; Raisch & Krakowski, 2021), cost-sensitive evaluation examines operating conditions (Provost & Fawcett, 2001; Drummond & Holte, 2006), machine-learning research studies discrimination and calibration, and decision-focused learning integrates predictions with downstream objectives (Donti et al., 2017; Wilder et al., 2019; Elmachoub & Grigas, 2022). A practical gap remains in ordinary management analytics workflows, where organizations already

possess probability scores and must decide which trained model to deploy under simple yet binding economic and capacity constraints.

This study contributes a lightweight policy-evaluation layer that can be applied after model training without requiring a new learning algorithm. DAAV makes the economic threshold and resource constraint explicit, reports realized net decision value, and normalizes that value against two scenario-matched benchmarks. Its distinctive contribution is therefore the combination of threshold economics, finite capacity, realized value, and opportunity-capture normalization within one post-hoc management analytics audit. The empirical analysis also contributes a caution regarding calibration under temporal shift: calibration validated on an earlier period may be action-destructive when a later period has a substantially different event rate. Fig. 1 summarizes the resulting evaluation pipeline from predictive scores to policy value.



Model quality is evaluated after scores are converted into operational actions.

Fig. 1 Decision-adjusted evaluation pipeline from predictive scores to policy value.

3.0 Methodology

3.1 Data Source and Temporal Design

The empirical study uses the UCI Bank Marketing dataset associated with direct-marketing telephone campaigns of a Portuguese banking institution (Moro et al., 2014a, 2014b). The bank-additional-full file contains 41,188 observations and 20 input variables, and UCI documents that the observations are ordered by date from May 2008 to November 2010. The binary outcome indicates whether the client subscribed to a term deposit. The chronological ordering permits a realistic temporal holdout design rather than a random split that would mix later market conditions into the training set.

The first 75% of observations ($n = 30,891$) form the training set, the next 10% ($n = 4,118$) form the validation/calibration set, and the final 15% ($n = 6,179$) form the test set. The event rates are 6.42%, 6.85%, and 38.45%, respectively. The sharp increase in the final holdout is not artificially constructed; it emerges from the documented chronological ordering of the source file. Because the dataset does not include an explicit year field for each row, the study refers to ordered temporal windows rather than assigning exact calendar dates to the split boundaries. Table 1 summarizes the resulting sample sizes, roles, and event rates.

Call duration is excluded from all predictive models. UCI defines duration as duration of the last contact, in seconds. That quantity is only known after a contact has occurred and therefore cannot be used to choose whom to contact in a pre-call targeting policy. Retaining it would create post-contact leakage and would inflate the apparent deployability of the models. All other available bank-additional predictors are retained. Categorical variables are one-hot encoded, with unseen categories ignored during transformation; numeric variables are standardized for logistic regression via a shared preprocessing pipeline.

Table 1. Data and Chronological Split

Window	Observations	Share	Event rate	Role
Training	30,891	75%	6.42%	Model fitting

Window	Observations	Share	Event rate	Role
Validation	4,118	10%	6.85%	Static calibration
Temporal test	6,179	15%	38.45%	Final evaluation
Total	41,188	100%	11.27%	Ordered source file

3.2 Models and Reproducibility

Three model families are used to represent common levels of functional complexity. Logistic regression provides a linear probabilistic baseline. Random forest represents a bagged nonlinear tree ensemble (Breiman, 2001). XGBoost is a regularized gradient-boosted tree algorithm (Chen & Guestrin, 2016). The purpose is not to identify the state-of-the-art classifier for this dataset. Instead, the heterogeneous models create different ranking and probability behaviors through which the proposed evaluation framework can be tested.

All models are fitted only on the training window. Hyperparameters were selected before inspecting the final test-set results using standard, literature-informed specifications intended to provide stable representative models rather than optimized leaderboard performance. The validation window was reserved for the static Platt calibration exercise and was not used for hyperparameter search. Logistic regression uses the liblinear solver with a maximum of 1,000 iterations. Random forest uses 500 trees, a minimum leaf size of 10, and square-root feature sampling. XGBoost uses 500 estimators, maximum depth 4, learning rate 0.03, subsample 0.80, column subsample 0.80, and L2 regularization parameter 2. Random seeds are fixed at 42. The analysis was executed with scikit-learn 1.8.0 and XGBoost 3.1.3. Table 2 reports the complete specification and selection basis. These implementation details support replication; no claim is made that the findings depend on these exact hyperparameters.

Table 2. Model Hyperparameters and Selection Basis

Model	Hyperparameter	Value	Selection basis
Logistic regression	Solver	liblinear	Standard stable binary-classification solver
Logistic regression	Maximum iterations	1,000	Conservative convergence allowance
Random forest	Number of trees	500	Stability-oriented ensemble size
Random forest	Minimum leaf size	10	Moderate regularization
Random forest	Feature sampling	sqrt	Standard random-forest feature subsampling
XGBoost	Estimators	500	Moderate boosted-ensemble size
XGBoost	Maximum depth	4	Moderate tree complexity
XGBoost	Learning rate	0.03	Conservative shrinkage
XGBoost	Subsample	0.80	Standard stochastic subsampling
XGBoost	Column subsample	0.80	Standard feature subsampling
XGBoost	L2 regularization	2	Moderate weight regularization
All stochastic models	Random seed	42	Fixed for exact reproducibility

Table 3. Final-Holdout Predictive Performance

Model	ROC-AUC	AP	Brier	Log loss	Mean probability
Logistic regression	0.648	0.516	0.235	0.680	0.336
Random forest	0.653	0.526	0.246	0.694	0.232

XGBoost	0.644	0.536	0.243	0.703	0.239
---------	-------	-------	-------	-------	-------

3.3 Conventional Predictive Evaluation

Test-set discrimination is assessed with ROC-AUC and average precision (AP). Probability quality is assessed with the Brier score and logarithmic loss. The mean predicted probability is also compared with the observed event rate. Table 3 reports the final holdout predictive results. For diagnostic calibration, a logistic calibration regression is fitted on the test set using the logit of the predicted probability as the independent variable; an ideal calibration intercept is 0 and slope is 1. This test-set calibration regression is descriptive only and is not used to modify predictions.

A separate static Platt calibration exercise is conducted to evaluate RQ3. For each model, a logistic mapping from the validation-set logit score to the validation outcome is estimated and then applied unchanged to the final temporal holdout. Because Platt scaling is monotonic, it does not change the ROC-AUC ranking, but it can materially change probability thresholds and, therefore, the number of cases that qualify for action. This makes it a useful stress test for distinguishing between discrimination-preserving and decision-preserving transformations.

3.4 Economic Action Rule and Capacity Constraint

For each case, the realized subscription outcome is binary, and the model supplies a predicted probability. Let one contact have a positive cost and a successful subscription have a larger positive value. The benefit-cost ratio is defined as the successful-subscription value divided by the contact cost. Since multiplying both quantities by the same positive constant leaves the action decision unchanged, the contact cost is normalized to one. The expected net value of a contact is positive only when the predicted probability exceeds the corresponding break-even threshold. Equation (1) states this rule formally.

$$\hat{p}_i V - C > 0 \iff \hat{p}_i > \frac{C}{V} = \frac{1}{R} \quad (1)$$

Maximum capacity is defined as a fraction of the available cases. The implemented policy first filters cases by the economic threshold and then ranks eligible cases by predicted probability. It acts on the highest-ranked eligible cases only up to the capacity ceiling. Thus, capacity is a ceiling rather than a requirement to fill all available slots. Equation (2) gives the formal action indicator.

$$a_i(\hat{p}_i; R, b) = \mathbf{1}\{\hat{p}_i \geq \frac{1}{R}\} \mathbf{1}\{\text{rank}_i \leq \lfloor bN \rfloor\} \quad (2)$$

3.5 Net Decision Value and Decision-Adjusted Analytical Value

For a candidate action policy, realized net decision value (NDV) sums the normalized benefit from successful selected cases and subtracts the normalized cost of every selected case. To facilitate interpretation across samples, NDV is also reported per 1,000 test cases. The policy may select fewer than the nominal capacity if too few cases satisfy the economic threshold. This feature is intentional because an economically screened manager should not fill capacity with cases whose predicted expected value is negative merely to meet a quota.

DAAV normalizes the realized policy value against two benchmarks. The random-capacity benchmark is the expected value of uniformly selecting exactly as many cases as the capacity ceiling permits, using the observed holdout event rate as an ex post diagnostic baseline. The attainable-oracle benchmark assumes perfect knowledge of realized outcomes and selects positive cases up to capacity while never spending contact cost on

a known negative case. This oracle is deliberately constrained by the same capacity limit; it is not an unconstrained upper bound.

DAAV equals 0 when the policy performs at the expected random-capacity benchmark and 1 when it reaches the attainable oracle. Values below 0 indicate performance worse than that of the random-capacity baseline. DAAV is undefined when the oracle and random benchmarks coincide, because the denominator in Equation (5) is zero. Under the maintained assumptions of positive capacity, a benefit-cost ratio above one, and a nondegenerate outcome distribution, this does not occur in the present experiment. In general, it can arise in degenerate cases such as zero effective capacity or an all-positive evaluation sample. Because the denominator depends on the economic and capacity scenario, DAAV is most appropriately interpreted within a scenario rather than as a universal property of a model.

$$V_{\pi}(R, b) = \sum_{i=1}^N a_i(Ry_i - 1), \quad V_{\pi, 1000} = \frac{1000}{N} V_{\pi} \quad (3)$$

$$V_{\text{rnd}} = (R\bar{y} - 1)\lfloor bN \rfloor, \quad V_{\text{orc}} = (R - 1)\min\left(\sum_i y_i, \lfloor bN \rfloor\right) \quad (4)$$

$$\text{DAAV}_{\pi} = \frac{V_{\pi} - V_{\text{rnd}}}{V_{\text{orc}} - V_{\text{rnd}}} \quad (5)$$

For scenario-level model selection, the decision regret of a conventionally selected model is defined as the proportional loss relative to the best realized policy value in the same scenario. Regret is zero when the conventionally selected model achieves the best value and is positive otherwise, measuring the fraction of the best attainable candidate value that remains unrealized. Equation (6) formally states the calculation.

In Equations (1) to (6), i indexes cases; y_i in $\{0,1\}$ is the realized outcome; $\hat{p}_i$ is the predicted event probability; V is the gross value of a successful subscription; C is contact cost; $R = V/C$ is the benefit-cost ratio; b is the maximum capacity share; N is the number of evaluation cases; $\text{floor}(bN)$ is the capacity ceiling; rank_i is the descending probability rank among economically eligible cases; a_i is the action indicator; $\bar{y}$ is the observed holdout event rate; V_{π} is realized policy NDV; V_{rnd} and V_{orc} are the random-capacity and attainable-oracle benchmarks; V^* is the best realized candidate value in a scenario; and V_m is the value of the conventionally selected model m . Contact cost is normalized to $C = 1$ without loss of generality for ranking policies.

$$\text{Regret}_m = \frac{V^* - V_m}{V^*} \quad (6)$$

3.6 Scenario Grid and Statistical Uncertainty

The main analysis varies the benefit-cost ratio over 3, 5, 8, 12, and 20 and the maximum capacity share over 10%, 20%, 30%, and 50%, producing 20 operating scenarios. This range deliberately includes a low-margin regime in which probability calibration and eligibility are consequential, as well as higher-return regimes in which more cases are economically attractive. For each scenario, the policy is evaluated separately for all three models.

Statistical uncertainty is quantified for the focal low-margin, high-capacity scenario with a benefit-cost ratio of 3 and a maximum capacity of 50% using 1,000 paired nonparametric bootstrap resamples of the final temporal holdout. In each resample, the outcomes and all three model scores are resampled jointly, preserving paired comparisons. The 2.5th and 97.5th percentiles of the bootstrap distribution define the lower and upper bounds of the 95% percentile bootstrap confidence interval. The focal comparison is selected because it

produces the clearest utilization-sensitive rank reversal and therefore directly tests the central decision-value mechanism.

4.0 Analysis and Discussion

4.1 Temporal Shift in the Holdout

The chronological structure produces a substantial shift in target prevalence. The training and validation windows contain subscription rates of 6.42% and 6.85%, whereas the final 15% holdout contains 38.45%. Fig. 2 shows the event rate by chronological decile across the entire file. The first six deciles remain below 7%, followed by increases to 10.17%, 11.97%, 15.74%, and finally 45.94% in the last decile. The final holdout therefore represents a materially different operating environment rather than a routine iid test set.

This shift is analytically useful because it exposes the difference between ranking stability and probability transportability. A model may continue to rank positive cases above negative cases reasonably well while its absolute probabilities lag the new base rate. Such a situation is especially consequential for management policies that translate a break-even economic threshold into a probability cutoff.

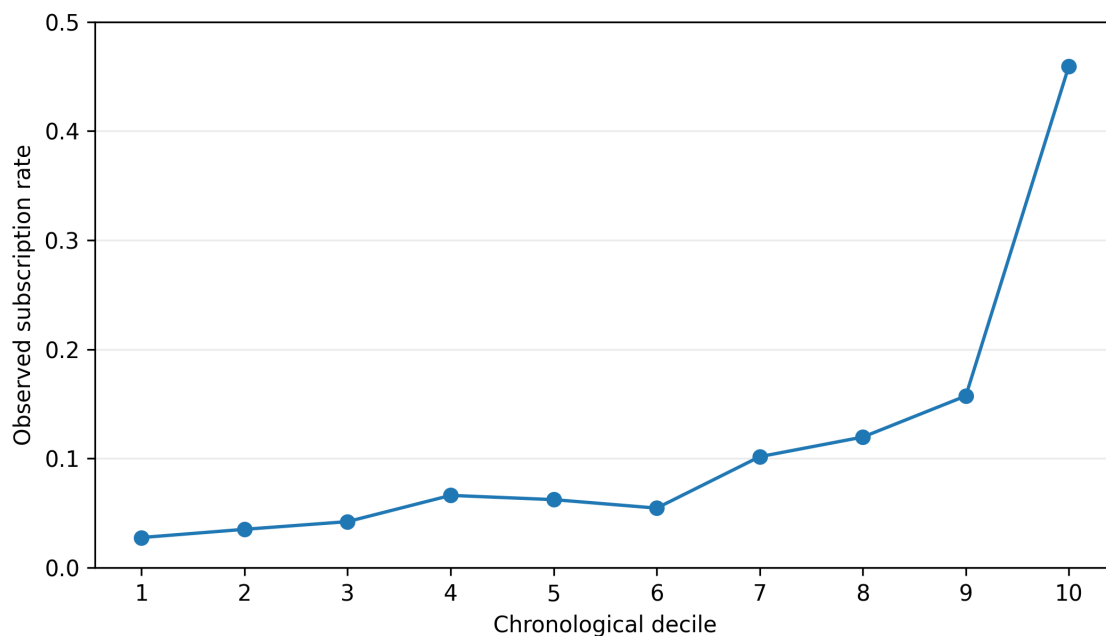


Fig. 2 Observed subscription rate by chronological decile. The final decile exhibits a pronounced prevalence increase.

4.2 Predictive and Calibration Performance

Table 3 reports the conventional predictive results. Random forest has the highest ROC-AUC point estimate at 0.653, followed by logistic regression at 0.648 and XGBoost at 0.644. The absolute differences are small: 0.005 between random forest and logistic regression and 0.009 between random forest and XGBoost. No pairwise significance test of ROC-AUC differences is used, so the manuscript does not interpret the random forest as statistically superior in discrimination. Instead, it is treated only as the model that would be selected if one mechanically chose the largest observed ROC-AUC. XGBoost has the highest AP at 0.536. These rankings already illustrate that model preference depends on the predictive metric. More importantly, all three models underpredict the shifted test prevalence on average. Mean predicted probabilities are 0.336 for logistic regression, 0.232 for random forest, and 0.239 for XGBoost compared with the realized rate of 0.385.

The descriptive calibration regression shows additional differences in probability transport. Logistic regression has intercept -0.128 and slope 0.447; random forest has intercept 1.049

and slope 1.245; XGBoost has intercept 0.342 and slope 0.636. None approximates the ideal intercept 0 and slope 1 on the shifted period. Thus, the model that ranks best by ROC-AUC is not necessarily the model whose absolute probabilities are most useful for an economic action threshold. Table 4 reports the calibration intercepts and slopes for all three models.

Table 4. Descriptive Calibration Diagnostics on the Final Temporal Holdout

Model	Observed rate	Mean probability	Calibration intercept	Calibration slope
Logistic regression	0.385	0.336	-0.128	0.447
Random forest	0.385	0.232	1.049	1.245
XGBoost	0.385	0.239	0.342	0.636

4.3 Decision Value in the Focal Scenario

The focal scenario uses a benefit-cost ratio of 3 and a maximum capacity of 50%. With normalized contact cost 1, the break-even probability threshold is $1/3$. Although the organization can contact up to half of the test population, the models do not all identify enough cases above that threshold to use the available capacity. Logistic regression selects 41.11% of cases, random forest selects only 12.19%, and XGBoost selects 22.45%. Random forest obtains the highest selected-case precision among the three at 59.50%, but its conservative eligibility set leaves most of the nominal capacity unused.

The resulting value ranking reverses the observed ROC-AUC point-estimate ranking. As reported in Table 5, logistic regression generates 226.9 NDV units per 1,000 cases, XGBoost generates 170.3, and random forest generates 95.6. The random-capacity benchmark for this scenario is 76.8 per 1,000 and the attainable oracle is 769.1. Corresponding DAAV values are 0.217 for logistic regression, 0.135 for XGBoost, and 0.027 for random forest. Fig. 3 visualizes the divergence between discrimination and realized decision value. The result is substantively important because the random forest is not failing through low precision. It is failing through opportunity under-capture at the economically relevant threshold.

The paired bootstrap supports the robustness of this focal difference. Logistic regression has mean bootstrapped NDV of 226.6 per 1,000 with a 95% interval of 202.1 to 251.8. Random forest has 95.5 with an interval of 80.8 to 110.5, and XGBoost has 170.1 with an interval of 152.3 to 189.2. The paired logistic-minus-random-forest difference averages 131.0 units per 1,000, with a 95% interval of 110.4 to 151.3, and is positive in all 1,000 bootstrap resamples. Table 6 summarizes these bootstrap estimates and percentile intervals.

Table 5. Decision Performance at Benefit-Cost Ratio 3 and Maximum Capacity 50%

Model	Contact rate	Selected precision	NDV / 1,000	DAAV
Logistic regression	41.11%	51.73%	226.9	0.217
Random forest	12.19%	59.50%	95.6	0.027
XGBoost	22.45%	58.62%	170.3	0.135

For reference, the expected random-capacity benchmark is 76.8 NDV units per 1,000 and the attainable-oracle benchmark is 769.1. The logistic policy therefore captures materially more of the available decision opportunity even though its selected-case precision is lower than that of the two tree ensembles.

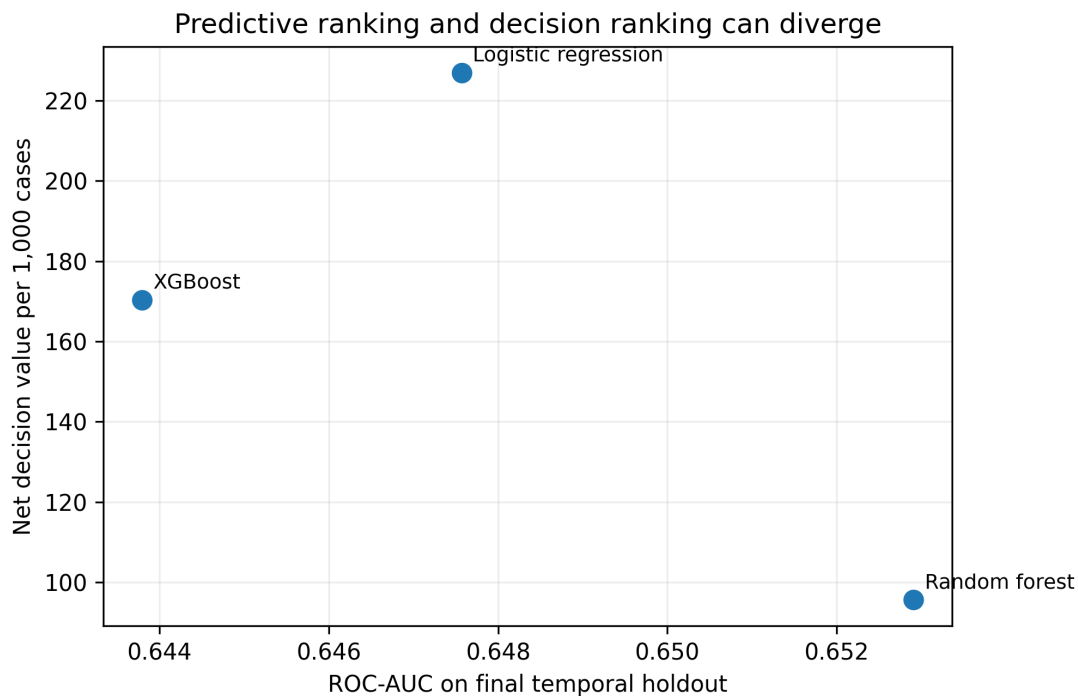


Fig. 3 ROC-AUC versus realized net decision value in the focal scenario. The predictive and decision rankings differ.

Table 6. Paired Bootstrap Uncertainty for the Focal Scenario

Quantity	Bootstrap mean	95% percentile interval
Logistic regression NDV / 1,000	226.6	202.1 to 251.8
Random forest NDV / 1,000	95.5	80.8 to 110.5
XGBoost NDV / 1,000	170.1	152.3 to 189.2
Logistic - random forest	131.0	110.4 to 151.3

4.4 Sensitivity Across Economic and Capacity Conditions

The focal result is not an isolated threshold artifact. Table 7 reports raw NDV across the full 20-scenario grid, while Table 8 reports the corresponding DAAV values obtained from the same scenario-specific random and attainable-oracle benchmarks. Across the grid, random forest, the model with the highest ROC-AUC point estimate, is not the NDV-maximizing model in 16 of 20 scenarios. XGBoost is strongest at low 10% capacity in every tested benefit-cost regime and at a benefit-cost ratio of 3 with 20% capacity. Random forest is strongest at 20% capacity for benefit-cost ratios of 5, 8, 12, and 20. Logistic regression performs best at 30% and 50% capacity across all tested benefit-cost ratios, with a particularly large advantage in the low-margin ratio-3 regime.

The pattern reveals two separate operational roles for predictive quality. At a very tight capacity, ranking the most promising cases dominates, and the boosted model performs well. At moderate capacity and higher returns, the random forest can capture strong value. At wider capacity, the absolute probability scale and economic eligibility become more influential because the policy must identify a larger pool of cases whose expected benefit exceeds cost. Logistic regression then produces a greater realized value despite a lower ROC-AUC. Fig. 4 displays this rank reversal at a benefit-cost ratio of 3.

Selecting the largest observed ROC-AUC point estimate would create positive decision regret in 16 scenarios. Relative to the best model in each scenario, the mean nonnegative

regret of the ROC-AUC-selected random forest is 10.1%, with a maximum of 57.8% at a benefit-cost ratio of 3 and 50% maximum capacity. Table 8 shows that the normalized opportunity-capture pattern is equally systematic: at 10% capacity, XGBoost reaches DAAV values near 0.408, whereas at 30% capacity, logistic regression is consistently near 0.284 and exceeds the two tree ensembles. These figures should not be interpreted as a universal penalty for random forests. They quantify the cost of choosing a model solely based on a predictive point estimate when the operating policy includes threshold and capacity constraints.

Table 7. Net Decision Value per 1,000 Cases Across 20 Operating Scenarios

Benefit-cost ratio	Capacity	LR	RF	XGB	Best
3	10%	64.7	80.3	90.5	XGB
3	20%	137.1	95.6	154.6	XGB
3	30%	203.1	95.6	170.3	LR
3	50%	226.9	95.6	170.3	LR
5	10%	174.5	200.4	217.3	XGB
5	20%	361.7	394.1	390.8	RF
5	30%	538.4	519.0	514.2	LR
5	50%	731.7	723.6	707.1	LR
8	10%	339.1	380.5	407.7	XGB
8	20%	698.7	750.4	745.3	RF
8	30%	1041.4	1010.4	1002.6	LR
8	50%	1470.6	1457.7	1435.7	LR
12	10%	558.5	620.7	661.4	XGB
12	20%	1147.9	1225.6	1217.8	RF
12	30%	1712.1	1665.5	1653.8	LR
12	50%	2455.9	2436.5	2403.5	LR
20	10%	997.4	1101.0	1169.0	XGB
20	20%	2046.4	2175.9	2163.0	RF
20	30%	3053.4	2975.7	2956.3	LR
20	50%	4426.4	4394.1	4339.1	LR

Table 8. Decision-Adjusted Analytical Value Across 20 Operating Scenarios

Benefit-cost ratio	Capacity	Random benchmark	Oracle benchmark	LR DAAV	RF DAAV	XGB DAAV
3	10%	15.3	199.7	0.268	0.352	0.408
3	20%	30.7	399.7	0.288	0.176	0.336
3	30%	46.1	599.8	0.284	0.089	0.224
3	50%	76.8	769.1	0.217	0.027	0.135
5	10%	92.1	399.4	0.268	0.352	0.407
5	20%	184.4	799.5	0.288	0.341	0.336
5	30%	276.7	1199.5	0.284	0.263	0.257
5	50%	461.2	1538.1	0.251	0.244	0.228
8	10%	207.3	699.0	0.268	0.352	0.408
8	20%	415.0	1399.1	0.288	0.341	0.336
8	30%	622.6	2099.2	0.284	0.263	0.257
8	50%	1037.9	2691.7	0.262	0.254	0.241
12	10%	360.9	1098.4	0.268	0.352	0.407
12	20%	722.4	2198.6	0.288	0.341	0.336
12	30%	1083.9	3298.8	0.284	0.263	0.257
12	50%	1806.9	4229.8	0.268	0.260	0.246
20	10%	668.1	1897.2	0.268	0.352	0.408
20	20%	1337.2	3797.5	0.288	0.341	0.336
20	30%	2006.4	5697.8	0.284	0.263	0.257
20	50%	3344.7	7306.0	0.273	0.265	0.251

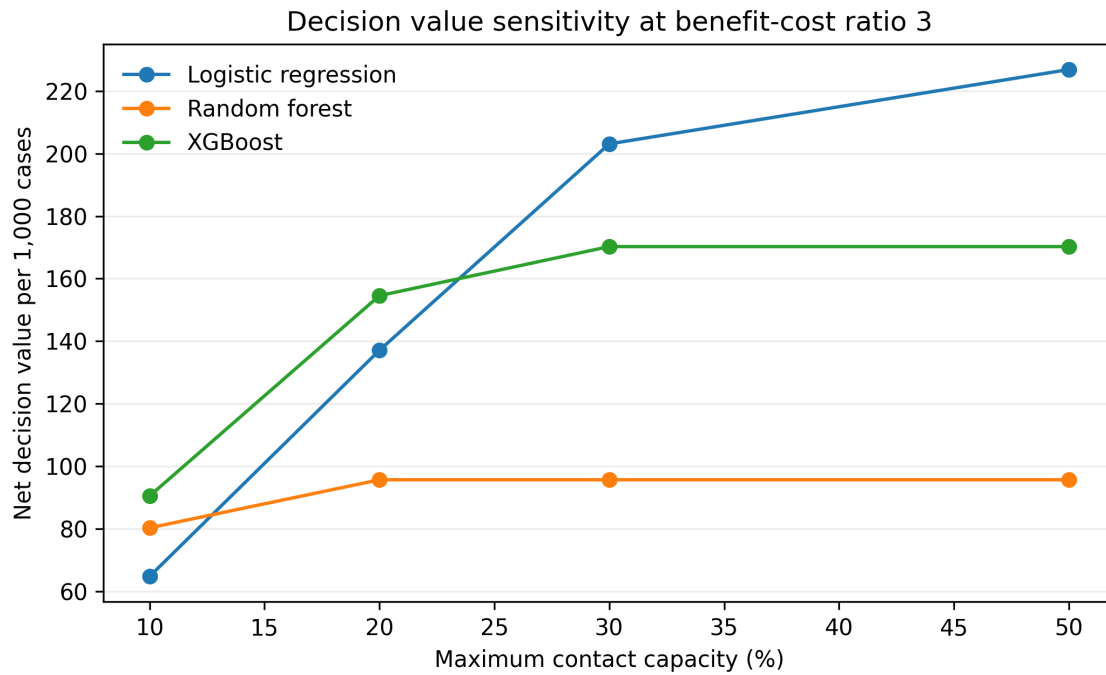


Fig. 4 Decision-value sensitivity to maximum contact capacity at benefit-cost ratio 3.

4.5 Static Calibration under Abrupt Shift

The Platt-scaling stress test produces a second policy-level caution. Validation prevalence is only 6.85%, close to the training rate but far below the 38.45% test rate. Calibration fitted to that earlier validation environment therefore pulls predicted probabilities downward relative to the later test period. ROC-AUC is unchanged under the monotonic transformation, yet the Brier score and log loss worsen for all three models on the final holdout set. Table 9 reports the raw and Platt-scaled probabilities and focal-scenario decision outcomes.

The downstream effect is severe in the low-margin focal scenario. Platt-calibrated logistic regression selects 18.45% of cases and yields 122.8 NDV units per 1,000, well below its raw-score value of 226.9. The calibrated random forest and XGBoost assign no cases above the one-third break-even threshold, producing zero realized NDV and negative DAAV relative to the random-capacity benchmark. The result does not imply that calibration is generally harmful. It shows that calibration must itself be monitored for transportability. A statistically sensible mapping on a pre-shift validation set can become operationally destructive after an abrupt change in event prevalence.

Table 9. Static Platt Calibration Fitted on Validation and Applied to the Final Holdout

Model	Raw Brier	Platt Brier	Raw mean probability	Platt mean probability	Platt NDV / 1,000 at ratio 3, capacity 50%
Logistic regression	0.235	0.266	0.336	0.191	122.8
Random forest	0.246	0.319	0.232	0.089	0.0
XGBoost	0.243	0.312	0.239	0.095	0.0

4.6 Theoretical Implications for Management Analytics

The findings extend the prediction-versus-decision distinction into a practical management-analytics evaluation problem. First, model value is policy contingent. The same model cannot be assigned a single managerial value independent of the benefit-cost ratio and the capacity envelope. This supports the broader argument in AI management research that

organizational structure, task contingencies, and human objectives determine how analytical systems should be deployed (Shrestha et al., 2019; Raisch & Krakowski, 2021).

Second, the results identify utilization as a missing bridge between calibration and resource allocation. Traditional calibration asks whether predicted probabilities correspond to observed frequencies. The present analysis shows why that correspondence matters operationally: a shifted probability scale alters the set of cases that clears a break-even threshold, which in turn determines whether scarce capacity is used and where it is used. In this sense, capacity utilization is an intermediate mechanism by which probabilistic error translates into managerial value loss.

Third, DAAV complements rather than competes with cost-sensitive and decision-focused learning. Cost curves and expected-cost methods characterize classifier performance as costs and class distributions vary (Provost & Fawcett, 2001; Drummond & Holte, 2006). Decision-focused approaches alter training to optimize predictive parameters for downstream decisions (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022). DAAV instead audits already trained model-policy pairs after economic screening and capacity restriction. Its normalized random-to-oracle scale is therefore best interpreted as retrospective opportunity capture under a specified management scenario, not as a new training loss, a causal estimand, or a universal replacement for net benefit or expected cost.

4.7 Managerial and Governance Implications

A direct implication is that model approval committees should request an operating-value matrix in addition to ROC-AUC, AP, and calibration statistics. Before deployment, managers can define plausible benefit-cost ratios and capacity levels, then compare candidate models on realized or backtested NDV and DAAV. A model should not be labeled best without specifying the scenario in which the label applies. This practice also makes economic assumptions visible to nontechnical stakeholders and can expose disagreement that would otherwise be hidden inside a statistical threshold.

A second implication concerns monitoring. Organizations often monitor drift, discrimination, and calibration separately from operations. DAAV suggests a joint control chart or dashboard containing at least four layers: event prevalence, probability calibration, selected-case coverage, and realized decision value. An abrupt decline in selected coverage, for example, may signal not a shortage of operational capacity but a probability-scale shift that is screening out cases at the economic threshold.

A third implication is governance of recalibration. A validation-based calibration step should not be considered permanently safe simply because it improved Brier score at the time of development. Calibration maps should have an expiration logic or shift trigger, especially when thresholds have direct economic meaning. Under substantial prevalence change, organizations should reassess the probability mapping, the action threshold, or both before allowing an automated policy to suppress or expand the action set.

Finally, managers should separate a model's statistical reputation from its current policy fit. Complex ensemble models can be excellent rankers and still be poorly aligned with a particular threshold-capacity configuration. Simpler models can sometimes create greater operational value because their probability distribution interacts more favorably with the economic rule. This is not an argument for simplicity per se. It is an argument for evaluating the full model-policy system.

4.8 Limitations and Future Research

Several limitations define the scope of the findings. First, the empirical evidence comes from one historical banking dataset. The large change in temporal prevalence is valuable as a stress case, but it should not be assumed to represent all marketing deployments. Replication is needed in credit, churn, fraud, maintenance, workforce, and procurement settings, with varying forms of drift and action constraints.

Second, the study uses scenario-based normalized economics rather than confidential bank profit data. The tested benefit-cost ratios are therefore sensitivity parameters, not estimates of the Portuguese bank's actual margins. This design improves transparency and transferability but limits any claim about the monetary return of the original campaign. Future work can embed organization-specific contribution margins, downstream retention value, contact-channel costs, and heterogeneous customer value.

Third, the outcome is observed for customers who were historically contacted. The analysis retrospectively asks which of those contacted cases a policy would have prioritized. It does not estimate the causal incremental effect of contact relative to no contact. A causal targeting problem would require treatment-control variation, uplift modeling, or another credible identification strategy. DAAV can be extended to such settings by replacing the outcome probability with an estimated incremental treatment effect and defining value accordingly.

Fourth, the random-capacity benchmark uses the observed test prevalence and is therefore an ex post diagnostic normalization rather than a deployable ex ante forecast. In live operations, the expected random benchmark would need to use a contemporaneous prevalence estimate with uncertainty. Likewise, the oracle benchmark is attainable only conceptually and should be interpreted as a ceiling for retrospective opportunity capture.

Fifth, the study intentionally holds model development simple. Hyperparameter optimization, alternative recalibration methods, prior-shift correction, conformal selection, and online adaptation could change absolute performance. The reported stochastic models use a fixed seed to make the analysis exactly reproducible; a multi-seed stability analysis would provide an additional robustness check for whether small rank differences persist across stochastic fits. Likewise, bootstrap uncertainty is concentrated on the focal policy comparison rather than every cell of the 20-scenario grid. Extending paired resampling to the tight-capacity scenarios would better quantify uncertainty in grid-wide rank reversals. These are useful robustness extensions, but they do not alter the present paper's more limited claim that decision ranking can diverge from predictive ranking when thresholds, capacity, and shift interact.

5.0 Conclusion

Management analytics creates value only when predictions are converted into actions. This study shows that a model-centric evaluation can obscure that conversion. In a chronologically shifted bank-marketing application, the random forest produced the highest ROC-AUC point estimate, although the discrimination differences were small and were not tested for pairwise statistical significance. It nevertheless failed to maximize realized net decision value in 16 of 20 benefit-cost and capacity scenarios. The largest divergence arose in a low-margin, high-capacity regime where the random forest maintained high selected-case precision but left economically usable capacity underutilized. Logistic regression generated substantially greater value because it identified a broader set of profitable cases above the break-even threshold.

The proposed Decision-Adjusted Analytical Value framework makes this mechanism measurable. By combining an explicit economic threshold, a capacity rule, realized NDV, and normalization against random-capacity and attainable-oracle benchmarks, DAAV reports how much of the available decision opportunity a model-policy pair captures. The calibration stress test further shows that monotonic probability recalibration can preserve discrimination while harming action value following a sharp shift in temporal prevalence. The practical conclusion is straightforward: organizations should select and govern predictive models as components of decision policies, not as isolated statistical objects. Predictive metrics remain necessary, but deployment approval should also ask what actions the scores cause, how much capacity those actions consume, and what value is realized under plausible operating conditions.

Acknowledgment

The author acknowledges the UCI Machine Learning Repository and the original dataset creators for making the Bank Marketing data publicly available. No external funding was received for this study.

Data Availability

The Bank Marketing dataset is publicly available from the UCI Machine Learning Repository under a CC BY 4.0 license at DOI 10.24432/C5K306. The analysis uses the bank-additional-full file and the row ordering documented by UCI.

Ethics Statement

This study uses a publicly available secondary dataset and does not involve new interaction with human participants. No personally identifying information is analyzed in this manuscript.

Conflict of Interest

The author declares no conflict of interest.

References

- [1] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>
- [2] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>
- [4] Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637-643. <https://doi.org/10.1007/s12599-019-00595-2>
- [5] Donti, P. L., Amos, B., & Kolter, J. Z. (2017). Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, 30, 5484-5494.
- [6] Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65, 95-130. <https://doi.org/10.1007/s10994-006-8199-5>
- [7] Elmachtoub, A. N., & Grigas, P. (2022). Smart "Predict, then Optimize". *Management Science*, 68(1), 9-26. <https://doi.org/10.1287/mnsc.2020.3922>
- [8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70, 1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [9] Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103-123. <https://doi.org/10.1007/s10994-009-5119-5>
- [10] Hernandez-Orallo, J., Flach, P. A., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13, 2813-2869. <https://www.jmlr.org/papers/v13/hernandez-orallo12a.html>
- [11] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237-293. <https://doi.org/10.1093/qje/qjx032>
- [12] Kotary, J., Fioretto, F., Van Hentenryck, P., & Wilder, B. (2021). End-to-end constrained optimization learning: A survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4475-4482. <https://doi.org/10.24963/ijcai.2021/610>

- [13] Mandi, J., Kotary, J., Berden, S., Mulamba, M., Bucarey, V., Guns, T., & Fioretto, F. (2024). Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80, 1623-1701. <https://doi.org/10.1613/JAIR.1.15320>
- [14] Moro, S., Rita, P., & Cortez, P. (2014a). Bank Marketing [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5K306>
- [15] Moro, S., Cortez, P., & Rita, P. (2014b). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31. <https://doi.org/10.1016/j.dss.2014.03.001>
- [16] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 13991-14002.
- [17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- [18] Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3), 203-231. <https://doi.org/10.1023/A:1007601015854>
- [19] Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46(1), 192-210. <https://doi.org/10.5465/amr.2018.0072>
- [20] Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1), 21-41. <https://doi.org/10.1162/089976602753284446>
- [21] Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66-83. <https://doi.org/10.1177/0008125619862257>
- [22] Tan, K. H. (2026). Shift-responsive conformal ensemble for reliable selective classification under distribution shift. *International Journal of Progressive Research in Engineering Management and Science*, 6(7), 1664-1675. <https://doi.org/10.58257/IJPREMS54397>
- [23] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565-574. <https://doi.org/10.1177/0272989X06295361>
- [24] Wilder, B., Dilkina, B., & Tambe, M. (2019). Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 1658-1665. <https://doi.org/10.1609/aaai.v33i01.33011658>

Appendix A. Reproducibility Specification

Implementation sequence:

- (1) read bank-additional-full.csv with a semicolon delimiter;
- (2) encode the subscription outcome as 1 for "yes" and 0 otherwise;
- (3) remove the outcome and call duration from the predictor matrix;
- (4) use the first 30,891 rows for training, the next 4,118 rows for validation, and the final 6,179 rows for testing;
- (5) fit preprocessing and each model on training data only using the fixed specifications in Table 2;
- (6) compute raw validation and test probabilities;
- (7) fit Platt mappings on validation logits only;
- (8) evaluate conventional test metrics in Table 3;
- (9) for each benefit-cost and capacity scenario, screen cases at the break-even probability, rank eligible cases by predicted probability, and select no more than the capacity ceiling;
- (10) compute NDV, DAAV, and scenario rankings, including the full DAAV grid in Table 8; and
- (11) conduct 1,000 paired bootstrap resamples with seed 20260816 for the focal scenario.