

PREDICTIVE ANALYTICS USING MACHINE LEARNING TECHNIQUES FOR INTELLIGENT DECISION SUPPORT

¹Gawde Vilas Anant, ²Dr. R M Krishna Sureddi

¹Research Scholar, ²Research Supervisor

¹⁻²Department of Computer Application, Arunodaya University, Itanagar, Arunachal Pradesh

Abstract

Machine learning is widely used to predict events such as customer response, loan default, employee attrition and disease, but a prediction becomes useful only when it is turned into an action. Most evaluations stop at accuracy or the area under the ROC curve (AUC) and leave open how predictions should be converted into decisions, how far the probabilities can be trusted, which cases should be left to people, and whether the explanations are stable. This paper proposes a decision-centred framework for intelligent decision support that links five layers: prediction, probability calibration, cost-based decision rules, uncertainty-based deferral to human reviewers, and explanation. The framework was tested on five public decision problems: bank telemarketing, credit approval, employee retention, cardiac referral and daily bicycle-fleet provisioning. Four learners (logistic regression, random forest, LightGBM and a multilayer perceptron) were evaluated in raw, Platt-scaled and isotonic-calibrated form by five-fold cross-validation repeated twice, and the provisioning task was evaluated by monthly rolling-origin forecasting. Differences in AUC between the learners were small (at most 0.03 on three of the four classification tasks), whereas the decision rule changed outcomes substantially: replacing the default threshold of 0.5 with the threshold implied by the costs of errors raised the net gain per bank customer from 0.62 to 2.02 cost units over the best model-free rule, cut the expected cost per credit applicant by 32 per cent, and turned the cardiac referral model from worse than referring every patient into better than it. Calibration narrowed the gap between the theoretical cost-based threshold and the best threshold found in hindsight, most strongly on the smallest data set. Deferring the 20 per cent of cases with the highest expected regret removed between 31 and 48 per cent of decision regret. In provisioning, a point forecast corrected by a held-out residual quantile reduced the daily mis-provisioning cost by about a quarter compared with the best point forecast. The results show that the value of predictive analytics for decision support depends less on the choice of learner than on calibration, the decision rule and the design of the human-machine division of work.

Keywords: Predictive analytics; machine learning; decision support systems; probability calibration; cost-sensitive decision; selective prediction; quantile forecasting; explainable AI

1. Introduction

Predictive analytics uses historical data to estimate the likelihood of future events or the value of future quantities, and machine learning has become its main technical engine (Shmueli & Koppius, 2011). Organizations now apply these methods to choose which customers to contact, which loans to approve, which employees to support and which patients to refer for further testing. In each of these cases the prediction is not the end product. It is an input to a decision, and the decision has costs and benefits that are usually unequal for different kinds of error. A model that is very accurate in a statistical sense can still lead to poor decisions if its probabilities are miscalibrated, if the decision threshold ignores the costs of errors, or if the system acts automatically on cases where it is effectively guessing.

Decision support systems have long been understood as systems that help people make semi-structured decisions by combining data, models and human judgement (Power, 2002). Machine learning has made the model component far more powerful, but the way models are evaluated has not always kept pace with the way they are used. Benchmark studies usually rank learners by accuracy or AUC (Fernández-Delgado et al., 2014; Lessmann et al., 2015). Hand (2009) showed that AUC implicitly averages over misclassification costs in a way that depends on the classifier itself, so a higher AUC does not guarantee a better decision for a given cost structure. Work on prescriptive analytics has argued that prediction and optimization should be designed together rather than in sequence (Bertsimas & Kallus, 2020; Elmachtoub & Grigas, 2022).

This paper treats intelligent decision support as a chain of design choices and asks how much each link contributes. Four questions guide the analysis:

1. How large are the differences in predictive quality between commonly used learners on typical tabular decision problems?

2. How much decision value is gained or lost through the decision rule, and does probability calibration make cost-based rules more reliable?
3. Can uncertainty estimates identify the cases that should be deferred to human judgement, and how much regret does deferral remove?
4. For decisions about quantities rather than classes, does forecasting the right quantile matter more than forecasting the mean accurately?

A fifth, supporting question concerns explanation: whether the variables that a model presents as important stay the same when the data change slightly, since unstable explanations weaken the trust that decision support depends on. The contribution of the paper is a five-layer framework and an empirical test of it on five real decision problems from marketing, finance, human resources, health care and operations, with all results computed from repeated out-of-sample evaluation.

2. Background and Related Work

2.1 From prediction to decision support

Davenport and Harris (2007) described analytics as a source of competitive advantage when it is built into decision processes rather than kept in reports. Sculley et al. (2015) pointed out that most of the effort in deployed machine learning systems lies outside the model: in data pipelines, configuration, monitoring and the interfaces that connect predictions to actions. Profit-driven evaluation has been proposed in specific domains, such as churn management (Verbeke et al., 2012) and credit scoring (Lessmann et al., 2015), but the different design layers are rarely examined together.

2.2 Learners for tabular data

Logistic regression remains the standard reference model for binary outcomes because it is fast, stable and directly interpretable. Random forests average many decorrelated decision trees and are robust to irrelevant variables (Breiman, 2001). Gradient-boosted trees, of which LightGBM is an efficient implementation (Ke et al., 2017), fit trees sequentially to the errors of earlier trees and often achieve the highest accuracy on tabular data. Large benchmarks find that the best learners are often statistically indistinguishable and that no learner dominates on every problem (Fernández-Delgado et al., 2014).

2.3 Probability calibration

A classifier is calibrated when, among cases given a probability of 0.3, about 30 per cent are positive. Calibration is essential whenever probabilities are compared with a threshold derived from costs or combined with other quantities. Niculescu-Mizil and Caruana (2005) showed that boosted trees and some other learners produce distorted probabilities that can be corrected by post-processing. Platt (1999) proposed fitting a logistic function to classifier scores, and Zadrozny and Elkan (2002) proposed isotonic regression, a non-parametric monotone mapping. Guo et al. (2017) found that modern neural networks tend to be over-confident. Calibration is usually summarized by the Brier score (Brier, 1950) and by the expected calibration error (Naeini et al., 2015).

2.4 Cost-sensitive decisions and quantile forecasting

When the costs of false positives and false negatives differ, the optimal decision is to act whenever the predicted probability exceeds a threshold determined by the costs (Elkan, 2001). For quantities, the classic newsvendor result states that when shortage and surplus have different unit costs, the optimal stock level is a quantile of the demand distribution rather than its mean (Khouja, 1999). Quantile regression estimates such quantiles directly (Koenker & Bassett, 1978). Conformal prediction offers a distribution-free alternative that adjusts a point forecast using the empirical distribution of held-out residuals (Vovk et al., 2005).

2.5 Selective prediction, deferral and explanation

Chow (1970) showed that allowing a classifier to reject uncertain cases reduces the error rate on the cases it accepts. The idea has been revived as selective classification (Geifman & El-Yaniv, 2017) and as learning to defer to a human expert (Madras et al., 2018). On the explanation side, permutation importance measures how much a model's performance falls when a variable is randomly shuffled (Breiman, 2001; Fisher et al., 2019), and SHAP values attribute individual predictions to variables (Lundberg & Lee, 2017). Strobl et al. (2008) showed that importance measures are biased when predictors are correlated. Rudin (2019) argued that in high-stakes decisions interpretable models should be preferred over explained black boxes.

3. A Decision-Centred Framework

The framework proposed here organizes an intelligent decision support system into five layers. Each layer has its own design choices and its own evaluation measure, and a weakness in one layer cannot be repaired by strength in another.

1. **Prediction layer.** A learner estimates the probability of an event or the distribution of a quantity from tabular data. It is judged by discrimination (AUC) for classes and by forecast error for quantities.
2. **Calibration layer.** Scores are mapped to probabilities that match observed frequencies, using Platt scaling or isotonic regression fitted on internal cross-validation folds. It is judged by the Brier score and the expected

calibration error.

3. **Decision layer.** Probabilities are converted into actions by minimising expected cost, or quantities into order levels by choosing the cost-optimal quantile. It is judged by realised cost per case.
4. **Deferral layer.** Cases whose expected regret is highest are routed to a human reviewer, and the remaining cases are decided automatically. It is judged by the share of regret that stays in the automated stream at a given level of coverage.
5. **Explanation layer.** The variables that drive the predictions are reported to decision makers. This layer is judged by the stability of the reported variables across resamples of the data.

Let C_{ij} denote the cost of taking action i ($1 = \text{act}$, $0 = \text{do not act}$) when the true class is j . For a calibrated probability p that the event occurs, the expected costs of the two actions are:

$$E[\text{cost} | \text{act}] = p \cdot C_{11} + (1 - p) \cdot C_{10} \quad E[\text{cost} | \text{not act}] = p \cdot C_{01} + (1 - p) \cdot C_{00} \quad (1)$$

Acting is optimal when the first expression is smaller than the second, which gives the cost-based threshold

$$p^* = (C_{10} - C_{00}) / [(C_{10} - C_{00}) + (C_{01} - C_{11})] \quad (2)$$

Equation (2) holds only if p is a true probability, which is why the calibration layer comes before the decision layer. Calibration is measured by the Brier score and by the expected calibration error over $B = 10$ equal-frequency bins:

$$\text{Brier} = (1/n) \sum (p_i - y_i)^2 \quad \text{ECE} = \sum_b (n_b/n) \cdot |\bar{y}_b - \bar{p}_b| \quad (3)$$

Expected regret measures how much a decision is expected to lose compared with a decision made with perfect knowledge of the outcome:

$$r(p) = \min\{E[\text{cost} | \text{act}], E[\text{cost} | \text{not act}]\} - [p \cdot \min(C_{11}, C_{01}) + (1 - p) \cdot \min(C_{10}, C_{00})] \quad (4)$$

Regret is zero when p is 0 or 1 and highest near p^* , so it ranks cases by how much the decision depends on information the model does not have. The deferral layer sends the cases with the highest $r(p)$ to a reviewer. For quantity decisions, with unit shortage cost c_u and unit surplus cost c_o , the provisioning level s that minimizes expected cost is the q^* -quantile of demand, and the realized daily cost is:

$$q^* = c_u / (c_u + c_o) \quad \text{cost}(s, y) = c_u \cdot \max(y - s, 0) + c_o \cdot \max(s - y, 0) \quad (5)$$

Explanation stability is measured by the mean Jaccard similarity between the sets of the five most important variables obtained on different folds, $J = |A \cap B| / |A \cup B|$. A value of 1 means the same five variables appear on every fold.

4. Experimental Design

4.1 Decision tasks

Five public data sets were chosen to cover different domains, sizes and class balances (Table 1). Each was paired with a decision and a stylized cost structure. The costs units chosen to reflect the usual direction and rough size of the asymmetry in each domain; they are not taken from any particular organization. In the bank data, call duration was removed because it is known only after the call and would leak the outcome (Moro et al., 2014). In the credit data, the code 99999999 used for missing income, assets and debt was treated as missing.

Table 1. Decision tasks, data and cost structures

Task and data	Cases	Vari-ables	Positiv-e rate	Decision (act)	Costs (units)	p^*
Bank telemarketing (Moro et al., 2014)	41,188	19	11.3%	Call the client	Call 5; subscription worth 50	0.10
Credit approval (credit scoring data)	4,454	13	28.2%	Reject the application	Bad loan accepted 4; good loan rejected 1	0.20
Employee retention (IBM HR attrition)	1,470	30	16.1%	Make a retention offer	Offer 1; leaver 6; offer halves leaver cost	0.33
Cardiac referral (Detrano et al., 1989)	303	13	54.5%	Refer for further testing	Missed case 10; unnecessary referral 1	0.09
Bicycle fleet provisioning (Fanaee-T & Gama, 2014)	728 days	12	–	Number of bicycles made available	Shortage c_u ; surplus c_o = 1	q^* 0.5– 0.9

4.2 Learners and calibration

Four learners were used: logistic regression; a random forest with 200 trees and a minimum of three cases per leaf; LightGBM with 300 trees, 15 leaves and a learning rate of 0.05; and a multilayer perceptron with one hidden layer of 64 units and early stopping. All learners were placed in the same pipeline, which imputed missing numerical values by the median, standardized numerical variables and one-hot encoded categorical variables, with every step fitted only on training data. Each learner was used in three forms: raw, calibrated by Platt scaling (sigmoid) and calibrated by isotonic regression, giving twelve variants per task. Calibration used three shuffled internal folds of the training data,

so the calibration map was always fitted on data the base model had not seen.

4.3 Evaluation protocol

The classification tasks were evaluated by stratified five-fold cross-validation repeated twice with different random partitions, so every case received two out-of-fold probabilities from every variant. All measures were computed on these out-of-fold probabilities and averaged over the two repetitions. Three decision rules were compared: the default threshold of 0.5, the cost-based threshold p^* from equation (2), and an oracle threshold chosen in hindsight from a grid of 99 values to minimise realized cost. The oracle is not a usable rule, because it uses the test outcomes; it marks the best that any single threshold could have achieved with the given probabilities. Two model-free rules were also computed: acting on every case and acting on no case.

The provisioning task used daily rental counts for 2011 and 2012 with calendar, weather and lagged-demand variables. The first 544 days formed the initial training window, and the last 184 days (July to December 2012) were forecast month by month, with all models re-fitted at the start of each month on all earlier data. Six provisioning policies were compared: a naive policy using the demand of two days earlier; the LightGBM point forecast; linear and LightGBM quantile regression at the target quantile; and linear and LightGBM point forecasts shifted by the target quantile of residuals measured on the most recent 90 days of the training window, using a model fitted on the earlier days. Three cost ratios were used, giving target quantiles of 0.5, 0.75 and 0.9.

Explanation stability was assessed for logistic regression and LightGBM on each classification task by computing permutation importance (drop in AUC, five permutations) on each test fold of a five-fold split and comparing the five most important variables across folds. The experiments were run in Python with scikit-learn 1.9 (Pedregosa et al., 2011) and LightGBM 4.7, with fixed random seeds.

5. Results

5.1 Discrimination and calibration

Table 2 reports AUC and calibration for the sixteen task–learner combinations. On three of the four tasks the four learners lay within 0.03 AUC of each other: LightGBM was best on the bank task (0.804), all learners were between 0.835 and 0.840 on the credit task, and the random forest was best on the cardiac task (0.912). The retention task was the exception: logistic regression (0.835) clearly outperformed the tree ensembles (0.804–0.810) and the neural network (0.759), whose AUC also varied widely across folds (standard deviation 0.091).

Table 2. Discrimination and calibration of the raw and calibrated learners (mean over two repetitions of five-fold cross-validation)

Task	Learner	AUC (SD)	Brier (raw)	ECE raw	ECE Platt	ECE isotonic
Bank marketing	Logistic regression	0.791 (0.006)	0.078	0.006	0.006	0.003
	Random forest	0.792 (0.007)	0.077	0.014	0.013	0.004
	LightGBM	0.804 (0.007)	0.076	0.005	0.007	0.003
	Neural network (MLP)	0.792 (0.010)	0.078	0.006	0.015	0.003
Credit scoring	Logistic regression	0.837 (0.010)	0.139	0.015	0.017	0.014
	Random forest	0.838 (0.012)	0.142	0.027	0.026	0.008
	LightGBM	0.835 (0.010)	0.143	0.031	0.025	0.017
	Neural network (MLP)	0.840 (0.008)	0.139	0.013	0.028	0.012
Employee attrition	Logistic regression	0.835 (0.033)	0.094	0.029	0.033	0.017
	Random forest	0.810 (0.029)	0.106	0.045	0.022	0.028
	LightGBM	0.804 (0.027)	0.107	0.062	0.040	0.027
	Neural network (MLP)	0.759 (0.091)	0.119	0.051	0.031	0.028
Heart disease	Logistic regression	0.905 (0.021)	0.124	0.031	0.039	0.043
	Random forest	0.912 (0.030)	0.123	0.069	0.045	0.030
	LightGBM	0.884 (0.021)	0.148	0.094	0.058	0.049
	Neural network (MLP)	0.884 (0.028)	0.163	0.127	0.074	0.079

Calibration varied far more than discrimination. On the bank task all raw learners were already well calibrated (ECE between 0.005 and 0.014), because the data set is large and the learners optimize log-loss or average many trees. On the cardiac task the raw LightGBM model and the neural network were poorly calibrated (ECE 0.094 and 0.127). Isotonic calibration reduced these errors to 0.049 and 0.079, and Platt scaling to 0.058 and 0.074. Platt scaling and isotonic regression usually reduced ECE where it was high and changed little where it was already low. Calibration also left AUC almost unchanged, except for the neural network on the two small data sets, where the internal folds

left too little data for the base model.

5.2 Decision rules and the value of calibration

Table 3 converts the probabilities into decisions. Expected cost per case is reported in the units of Table 1, and negative values on the bank task are net gains. For every learner on the bank, credit and cardiac tasks, the cost-based threshold gave a lower cost than the default threshold of 0.5, and the difference was large compared with the differences between learners. On the bank task the cost-based threshold more than doubled the net gain of every learner. On the cardiac task it cut the cost of every raw learner by between 39 and 50 per cent. On the retention task, where $p^* = 0.33$ is close to 0.5, the two rules gave similar costs.

Table 3. Expected cost per case under three decision rules (negative values are net gains)

Task	Learner	Threshold 0.5 (raw)	p^* raw	p^* Platt	p^* isotonic	Oracle threshold
Bank marketing	Logistic regression	-1.089	-2.534	-2.531	-2.541	-2.544
	Random forest	-1.243	-2.584	-2.612	-2.618	-2.624
	LightGBM	-1.253	-2.650	-2.643	-2.634	-2.656
	Neural network (MLP)	-1.128	-2.529	-2.537	-2.557	-2.557
Credit scoring	Logistic regression	0.615	0.422	0.424	0.423	0.421
	Random forest	0.652	0.414	0.406	0.407	0.404
	LightGBM	0.599	0.428	0.409	0.412	0.409
	Neural network (MLP)	0.624	0.432	0.427	0.430	0.422
Employee attrition	Logistic regression	0.854	0.857	0.855	0.855	0.852
	Random forest	0.922	0.883	0.877	0.871	0.871
	LightGBM	0.894	0.883	0.873	0.877	0.866
	Neural network (MLP)	0.922	0.947	0.913	0.915	0.910
Heart disease	Logistic regression	0.746	0.411	0.358	0.386	0.335
	Random forest	0.751	0.375	0.347	0.356	0.284
	LightGBM	0.975	0.507	0.399	0.363	0.338
	Neural network (MLP)	0.721	0.442	0.426	0.411	0.375

Calibration matters because equation (2) assumes true probabilities. The difference between the cost at p^* and the cost at the oracle threshold measures how much is lost because the probabilities are not exactly right. Averaged over the four learners, isotonic calibration reduced this gap from 0.078 to 0.022 on the cardiac task, from 0.014 to 0.003 on the retention task and from 0.018 to 0.004 on the bank task; on the credit task the gap was already small (0.005). The clearest single case was raw LightGBM on the cardiac task, whose cost at p^* fell from 0.507 to 0.363 after isotonic calibration, a reduction of 28 per cent with no change in the underlying model. Calibration was thus most valuable where the data were smallest and the threshold furthest from 0.5, which is where miscalibrated probabilities are most likely to fall on the wrong side of the threshold.

Table 4 summarizes the decision value of each task by comparing the best variant under each rule with the better of the two model-free rules. The gain over the model-free rule is the value that predictive analytics adds to the decision.

Table 4. Best achievable cost per case by decision rule, compared with the best model-free rule

Task	Best model-free rule	Best at 0.5	Gain at 0.5	Best at p^*	Gain at p^*	Oracle
Bank marketing	-0.633 (act on every case)	-1.253	+0.621	-2.650	+2.017	-2.656
Credit scoring	0.718 (act on every case)	0.599	+0.119	0.406	+0.312	0.404
Employee attrition	0.967 (act on no case)	0.854	+0.113	0.855	+0.112	0.852
Heart disease	0.455 (act on every case)	0.721	-0.266	0.347	+0.109	0.284

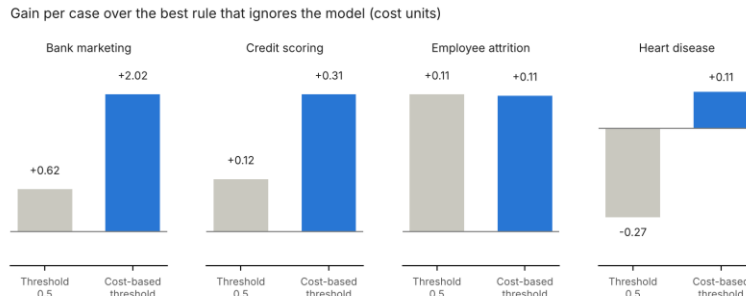


Figure 1. Gain per case over the best model-free rule under the default and the cost-based threshold. The cardiac task shows most clearly why the decision rule matters. Referring every patient costs 0.455 per patient, because nearly all cases that matter are caught and unnecessary referrals are cheap. With the threshold of 0.5, even the best variant cost 0.721, so the model made decisions worse than the simple policy it was meant to improve on. With the cost-based threshold, the best variant (a calibrated random forest) cost 0.347, a gain of 24 per cent over referring everyone. On the bank task, the best model-free rule was to call every client (net gain 0.633 per client); a model with threshold 0.5 added 0.62 units per client, and the same model with the cost-based threshold added 2.02, more than three times as much. On the retention task the gain from the model was modest under either rule (0.112 units per employee, or 12 per cent of the cost of doing nothing). On the bank, credit and retention tasks the best cost at p^* was within 0.006 of the oracle threshold, so a correctly specified cost rule left almost nothing on the table; on the cardiac task the gap was 0.063, reflecting how noisy any threshold is on 303 cases.

5.3 Deferral to human reviewers

For each task, the variant with the lowest cost at p^* was used to test deferral. Cases were ranked by expected regret, equation (4), and the highest-ranked 10 to 40 per cent were assumed to be passed to a reviewer. Table 5 reports the share of the total realized regret that remained in the automated decisions, and the error rate of those decisions. If deferral were random, deferring 20 per cent of cases would leave 80 per cent of the regret.

Table 5. Regret kept by the automated stream and error rate of automated decisions, by share of cases deferred

Task	Variant used	0% deferred	10%	20%	30%	40%
Bank marketing	LightGBM (raw)	100% / 18.4%	83% / 12.9%	69% / 10.2%	56% / 7.8%	44% / 6.7%
Credit scoring	Random forest (Platt)	100% / 27.1%	82% / 23.5%	66% / 20.2%	50% / 17.1%	38% / 14.4%
Employee attrition	Logistic regression (isotonic)	100% / 13.2%	71% / 9.4%	52% / 7.4%	38% / 6.1%	27% / 4.8%
Heart disease	Random forest (Platt)	100% / 31.7%	76% / 25.8%	60% / 22.3%	39% / 17.2%	26% / 14.8%

Deferring the most uncertain cases removes regret far faster than deferring at random

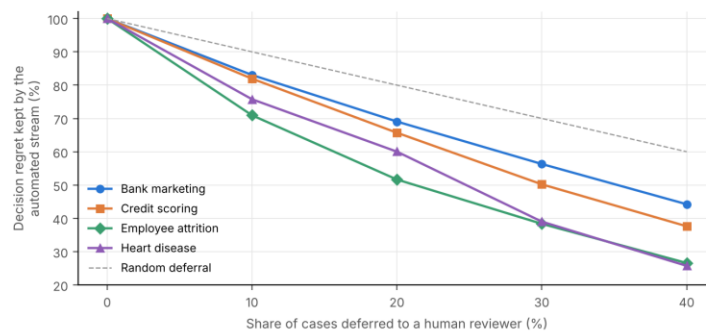


Figure 2. Share of decision regret kept by the automated stream as more uncertain cases are deferred. Regret-based deferral was far more effective than random deferral on all four tasks. Deferring 20 per cent of cases removed 31 per cent of the regret on the bank task, 34 per cent on the credit task, 40 per cent on the cardiac task and 48 per cent on the retention task. Deferring 40 per cent removed between 56 and 74 per cent. The error rate of

automated decisions fell correspondingly: on the bank task from 18.4 per cent with no deferral to 10.2 per cent with 20 per cent deferred, and on the retention task from 13.2 to 7.4 per cent. The retention task benefited most because most employees received low predicted probabilities, so the cases near the threshold formed a small, well-defined group.

5.4 Provisioning under asymmetric costs

Among the point forecasts, LightGBM had the lowest mean absolute error (679 rentals per day, compared with 764 for linear regression, 781 for the random forest and 1,099 for the naive forecast; mean daily demand in the test period was 6,043). The test period included the closure caused by Hurricane Sandy (29 October 2012, 22 rentals), which none of the models could anticipate. Table 6 reports the daily mis-provisioning cost and the share of days on which demand was fully met, for the three cost ratios.

Table 6. Mean daily provisioning cost (surplus cost = 1 per bicycle) and service level, July–December 2012

Policy	$q^* = 0.50$ ($c_u = 1$)	$q^* = 0.75$ ($c_u = 3$)	$q^* = 0.90$ ($c_u = 9$)
Naive (two days earlier)	1,099 / 48%	2,162 / 48%	5,351 / 48%
LightGBM point forecast	679 / 41%	1,360 / 41%	3,404 / 41%
Linear quantile regression	787 / 48%	1,103 / 66%	1,468 / 79%
LightGBM quantile regression	809 / 48%	1,337 / 61%	1,898 / 72%
Linear + residual quantile	779 / 60%	1,079 / 79%	1,365 / 91%
LightGBM + residual quantile	691 / 57%	1,000 / 78%	1,292 / 91%

Mean daily cost of mis-provisioning: a calibrated quantile beats a better point forecast

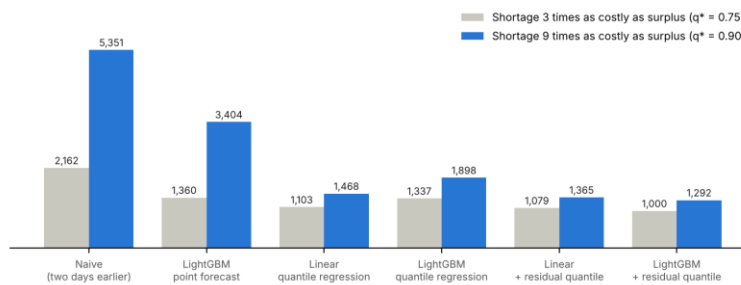


Figure 3. Mean daily mis-provisioning cost of six policies under two shortage-to-surplus cost ratios. When shortage and surplus cost the same ($q^* = 0.5$), the point forecast was the best policy, as expected. As soon as shortages became more expensive, the point forecast lost its advantage, because it under-provided on about 59 per cent of days. With a cost ratio of 3, shifting the LightGBM forecast by the 0.75 quantile of recent residuals reduced the daily cost from 1,360 to 1,000 (27 per cent lower); with a ratio of 9 the reduction was from 3,404 to 1,292 (62 per cent). The direct LightGBM quantile model performed worse than expected: at the 0.75 and 0.90 targets it met demand on only 61 and 72 per cent of days, well short of its nominal level. Demand grew strongly during 2012 and tree models cannot extrapolate a trend, so the quantile model learned its quantiles from a lower level of demand. The residual-quantile policies achieved service levels close to the target (78 and 91 per cent), because the residuals were taken from the most recent period.

5.5 Stability of explanations

Table 7 lists, for each task and model, the four variables with the highest mean permutation importance, with the number of folds (out of five) on which each appeared among the top five, and the mean Jaccard similarity of the top-five sets across folds.

Table 7. Most important variables and stability of the top-five set across five folds

Task	Model	Four most important variables (folds in top five)	Jaccard
Bank marketing	Logistic regression	employment variation rate (5), consumer price index (5), contact channel (5), month of contact (5)	0.73
	LightGBM	number employed (5), three-month Euribor (5), contact channel (5), consumer price index (3)	0.77
Credit scoring	Logistic regression	loan amount (5), arrears record (5), income (5), job seniority (5)	0.87
	LightGBM	loan amount (5), income (5), arrears record (5), job seniority (5)	0.87
Employee attrition	Logistic regression	overtime (5), job role (3), environment satisfaction (2), job satisfaction (2)	0.29
	LightGBM	overtime (5), monthly income (3), stock option level (4), job satisfaction (3)	0.39
Heart disease	Logistic regression	chest pain type (cp) (4), vessels coloured (ca) (5), thalassaemia test (thal) (5), sex (5)	0.77
	LightGBM	vessels coloured (ca) (5), thalassaemia test (thal) (5), chest pain type (cp) (4), ST depression (oldpeak) (3)	0.59

Explanations were stable on the credit task (Jaccard 0.87 for both models), where loan amount, arrears record, income and job seniority were identified on every fold by both models. They were unstable on the retention task (Jaccard 0.29 and 0.39): overtime was always the leading variable, but the rest of the top five changed from fold to fold, because many weak and inter-related variables (satisfaction scores, tenure measures, income and job level) each carry a small share of the signal. On the bank task the two models agreed on the importance of macroeconomic conditions but disagreed on which indicator mattered: logistic regression relied on the employment variation rate, LightGBM on the number employed and the Euribor rate. These three indicators are strongly correlated in the data (pairwise correlations between 0.91 and 0.97), so the models were sharing one underlying signal in different ways, a known weakness of importance measures under correlated predictors (Strobl et al., 2008).

6. Discussion

6.1 Where the value of predictive analytics comes from

The results show a consistent ordering of design choices by their effect on decision quality. Choice of learner had a small effect on three of the four classification tasks, with the leading learners within 0.03 AUC of each other, which agrees with large benchmark studies (Fernández-Delgado et al., 2014; Lessmann et al., 2015). The decision rule had a large effect on every task where the cost-based threshold lay far from 0.5. On the bank task it more than tripled the value added by the model, and on the cardiac task it decided whether the model was worth using at all. Calibration had a moderate effect overall but a large one in the setting where it was most needed: a small data set, a learner prone to over-confidence and a threshold far from 0.5. Deferral produced a further large reduction in regret for a modest share of human effort. In practice, the effort put into choosing among strong learners is often much larger than the effort put into the decision layer. These findings suggest the balance should be reversed. AUC measures ranking quality across all thresholds, but a decision uses one threshold: on the cardiac task, models with AUC above 0.90 were worse than referring every patient when the default threshold was used. This supports the argument of Hand (2009) that evaluation should use the costs of the actual decision.

6.2 Calibration as a design layer

Calibration is often treated as an optional refinement, but the results suggest it should be a standard layer whenever a decision depends on a probability threshold, and its cost is only an internal cross-validation step. The benefit is greatest for learners that are known to be miscalibrated, such as boosted trees and neural networks (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017), and on small data sets. The experiments also showed a limitation. On the two smallest data sets the neural network lost discrimination when it was calibrated, because each base model was trained on two-thirds of the training data. Isotonic regression, being non-parametric, also needs enough data to estimate its mapping. For small data, Platt scaling or a well-calibrated learner such as logistic regression may be safer.

6.3 Human-machine division of work

The deferral results give a quantitative basis for designing the division of work between a model and its users. Instead of a fixed rule that every case is reviewed or none is, an organization can choose a deferral rate from a curve like Figure 2 according to the time reviewers have. Because regret is concentrated near the threshold, a small review effort removes a large share of it. This agrees with the theory of the reject option (Chow, 1970) and with the selective prediction literature (Geifman & El-Yaniv, 2017). Two conditions apply. First, deferral helps only if reviewers decide

deferred cases better than the model, which must be checked in practice; Bansal et al. (2021) found that human–AI teams do not automatically outperform their best member. Second, regret ranking relies on calibrated probabilities, so the calibration layer supports the deferral layer as well as the decision layer.

6.4 Forecasting the decision quantity

The provisioning task showed in a different form the same principle as the classification tasks: the decision needs the right quantity, not only an accurate one. The most accurate point forecast was a poor provisioning policy whenever shortage cost more than surplus. The best policy combined a good point forecast with an estimate of its recent errors, which is the logic of split conformal prediction (Vovk et al., 2005) and of the newsvendor result (Khouja, 1999). The weak direct tree-based quantile model is a warning for trending demand. Monitoring the actual service level against the target is therefore an essential part of the decision layer.

6.5 Explanations that decision makers can rely on

Explanations help decision makers to accept, question and correct model outputs only if they are stable. On the retention task a manager shown one fold's list could draw different conclusions from a manager shown another's, and on the bank task correlated indicators made the choice among them arbitrary. Practical guidance follows: report importance aggregated over resamples together with a stability measure, group strongly correlated variables before computing importance, and where stakes are high, prefer models whose structure is itself interpretable, which in these experiments cost little or no accuracy (Rudin, 2019).

6.6 Threats to validity

Several limitations qualify the findings. The costs were stylized and chosen by the author rather than measured, although they reflect the usual direction and size of the asymmetries in each domain, and the qualitative conclusions do not depend on their exact values. The learners were used with fixed, reasonable settings rather than tuned, which could change the small differences in AUC but is unlikely to change the larger effects of the decision layer. The cardiac data set is small, so its estimates are less precise. The deferral analysis assumed that reviewers decide deferred cases correctly, which gives an upper bound on the benefit of deferral. The provisioning evaluation covered one six-month period and one location.

7. Implications for Practice

The findings can be translated into a set of design principles for organizations that build decision support on predictive models:

- **Specify the decision before the model.** Write down the possible actions, the cost or value of each outcome and the resulting threshold or target quantile, and evaluate every model by the cost of the decisions it supports.
- **Always compare with model-free rules.** Report the gain over the best rule that ignores the model, such as acting on every case or on none. A model that does not beat such a rule should not be deployed.
- **Calibrate before applying a threshold.** Add Platt scaling or isotonic calibration on internal folds, check the expected calibration error, and prefer Platt scaling or naturally calibrated learners on small data sets.
- **Use expected regret to route cases.** Send the cases with the highest expected regret to human reviewers and set the deferral rate according to reviewer capacity, using a curve of regret against coverage.
- **Forecast the quantile the decision needs.** For provisioning, staffing and inventory, derive the target quantile from the cost ratio and adjust point forecasts with recent residuals; monitor the achieved service level.
- **Report explanations with their stability.** Aggregate importance over resamples, state how stable the leading variables are, and treat correlated variables as groups.

8. Conclusion

This paper proposed a five-layer framework for intelligent decision support based on predictive analytics and tested it on five public decision problems. The choice of learner mattered least: the leading learners were within 0.03 AUC on three of four classification tasks. The decision rule mattered most. A threshold derived from the costs of errors more than tripled the value added by the model in bank telemarketing, cut the expected cost of credit decisions by about a third, and turned a cardiac referral model from harmful into useful. Calibration made the cost-based rule reliable, especially on small data. Regret-based deferral removed a large share of the remaining regret with a modest review effort. In provisioning, the right quantile mattered more than the most accurate mean. Explanations were stable in some domains and unstable in others, which decision makers need to know. Predictive analytics supports good decisions when the model, the probabilities, the decision rule, the human role and the explanation are designed together.

Future work. Future studies should use costs measured in real organizations, include hyperparameter tuning within nested cross-validation, evaluate deferral with real reviewers rather than assuming correct review, extend the provisioning analysis to longer periods with trend-aware models, and test whether stability-aware explanations improve the decisions of human users.

References

- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.
- Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1), 41–46.
- Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business School Press.
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5), 304–310.
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 973–978).
- Elmachtoub, A. N., & Grigas, P. (2022). Smart “predict, then optimize”. *Management Science*, 68(1), 9–26.
- Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3), 113–127.
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30* (pp. 4878–4887).
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146–3154).
- Khouja, M. (1999). The single-period (news-vendor) problem: Literature review and suggestions for future research. *Omega*, 27(5), 537–553.
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems 31* (pp. 6147–6157).
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31.
- Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (pp. 2901–2907).
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.
- Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. Quorum Books.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28* (pp. 2503–2511).
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553–572.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307.
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229.
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699).