INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Cost-Aware Calibration of Classifiers

Mochen Yang, Xuan Bi

To cite this article:
Mochen Yang, Xuan Bi (2025) Cost-Aware Calibration of Classifiers. INFORMS Journal on Data Science 4(2):101-113.
https://doi.org/10.1287/ijds.2024.0038

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 4, No. 2, April–June 2025, pp. 101–113
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Cost-Aware Calibration of Classifiers

Mochen Yang,[a,*] **Xuan Bi**[a]

[a] Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455
*Corresponding author
Contact: yang3653@umn.edu, https://orcid.org/0000-0001-5101-9041 (MY); xbi@umn.edu, https://orcid.org/0000-0002-4683-1411 (XB)

Received: May 22, 2024
Revised: September 6, 2024
Accepted: September 18, 2024
Published Online in Articles in Advance: December 9, 2024

https://doi.org/10.1287/ijds.2024.0038

Abstract. Most classification techniques in machine learning are able to produce probability predictions in addition to class predictions. However, these predicted probabilities are often not well calibrated in that they deviate from the actual outcome rates (i.e., the proportion of data instances that actually belong to a certain class). A lack of calibration can jeopardize downstream decision tasks that rely on accurate probability predictions. Although several post hoc calibration methods have been proposed, they generally do not consider the potentially asymmetric costs associated with overprediction versus underprediction. In this research, we formally define the problem of cost-aware calibration and propose a metric to quantify the cost of miscalibration for a given classifier. Next, we propose three approaches to achieve cost-aware calibration, two of which are cost-aware adaptations of existing calibration algorithms; the third one (named `MetaCal`) is a Bayes optimal learning algorithm inspired by prior work on cost-aware classification. We carry out systematic empirical evaluations on multiple public data sets to demonstrate the effectiveness of the proposed approaches in reducing the cost of miscalibration. Finally, we generalize the definition and metric as well as solution algorithms of cost-aware calibration to account for nonlinear cost structures that may arise in real-world decision tasks.

History: David Martens served as the senior editor for this article.
Data Ethics & Reproducibility Note: There are no data ethics considerations. The code capsule is available on Code Ocean at https://doi.org/10.24433/CO.8552538.v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2024.0038).

Keywords: classification • probability prediction • calibration • cost-aware learning

1. Introduction

Imagine visiting a clinic for a health check. After taking multiple physiological measures and a comprehensive blood panel, a risk assessment of future heart attack is reported by a predictive model designed to detect the risk of cardiovascular disease. Suppose a patient receives a risk score of 70%, but across all patients who receive this score, only 60% of them actually developed heart attacks later on. In this case, the "70%" risk score reported by the model would be considered inaccurate because it overestimates the actual risk of cardiovascular disease.

The above scenario represents what happened with the QRISK-3 (Cardiovascular Risk Score) risk tool, which has been found to overestimate the actual risk of cardiovascular disease among patients in England (Pate et al. 2020). More generally, discrepancies between a model's predicted probabilities and the actual outcome rates are known as the lack of *calibration*. In decision domains where probability predictions (as a form of risk assessment or quantification of prediction uncertainty) are of primary importance, having well-calibrated predictive models is crucial because the predicted probabilities can affect subsequent decision making (e.g., patients may decide whether to request more tests or interventions based on predicted risk levels). Therefore, calibration has been recognized as an important property of predictive models across many domains, such as medical diagnosis (Bayati et al. 2014, Shah et al. 2018, Van Calster et al. 2019), meteorological forecasting (Voudouri et al. 2017), recidivism prediction (Kleinberg et al. 2017), and natural language processing (Nguyen and O'Connor 2015, Card and Smith 2018).

Unfortunately, many supervised learning techniques, including modern deep neural networks, produce poorly calibrated probability predictions (Niculescu-Mizil and Caruana 2005, Guo et al. 2017) and need to be calibrated post hoc. Moreover, although a number of calibration methods have been studied in the machine learning literature, they generally place equal weights on overprediction and underprediction of probabilities. In practice, however, overpredictions versus underpredictions often have asymmetric costs to stakeholders. For example, underprediction of risk levels in a medical diagnosis is usually more costly to patients than overprediction because the former may cause (sometimes irremediable) delay in necessary interventions.

In this research, we study the problem of *cost-aware calibration* of classification models. The objective of cost-aware calibration is to *calibrate a classifier's probability predictions, taking into account the potentially asymmetric costs of overprediction and underprediction of probabilities in order to minimize the overall costs of miscalibration*.[1] This is

analogous to but fundamentally different from the *cost-aware classification* problem, where the objective is to adjust class predictions based on potentially asymmetric costs of false positives and false negatives (which can be a different set of costs than those relevant for calibration) to minimize overall misclassification costs.

We start by providing a formal definition of the cost-aware calibration problem as well as a metric to quantify calibration cost for a given classifier. Then, we propose two cost-aware calibration methods, namely cost-aware Platt scaling (CAPS) and cost-aware temperature scaling (CATS). These two new methods are, respectively, based on Platt scaling (PS) (Platt 1999) and temperature scaling (TS) (Guo et al. 2017), two commonly used parametric calibration methods, but they are adapted to be cost aware. Additionally, we also develop a novel algorithm, `MetaCal`, that achieves cost-aware calibration in a principled manner based on the idea of Bayes optimal learning, which is independent of existing calibration methods. `MetaCal` is inspired by the `MetaCost` algorithm for cost-aware classification (Domingos 1999). It leverages a bootstrapping procedure to estimate the cost-minimizing predicted probability for each data instance; then, it learns a model that predicts those cost-minimizing probabilities.

We conduct empirical experiments on a number of public data sets to evaluate the performance of different cost-aware calibration methods. We find that all three methods are generally effective at reducing calibration costs. `MetaCal` outperforms the scaling-based methods on several data sets and consistently has the smallest performance variance (i.e., greater performance stability).

Finally, we generalize our proposed cost-aware calibration methods to account for more flexible cost structures, where the amount of calibration costs can be a nonlinear function of the absolute difference between predicted probability and actual outcome rate. We repeat the empirical evaluations under several different (nonlinear) cost structures, and again, we confirm the superior performance of `MetaCal` in reducing calibration costs.

We make several notable contributions to the machine learning literature on calibration. First, we extend the standard definition of calibration to account for asymmetric costs of overprediction versus underprediction, which is more realistic than symmetric costs for many real-world decision tasks. Such a cost-aware extension of calibration aligns with the "utility-based machine learning" perspective (Provost 2005, p. 1, Berardi et al. 2015), where the objective of building a machine learning model is not necessarily to optimize predictive performance but to optimize the "utilities" (in this case, minimize costs of miscalibration) associated with downstream decision tasks. Similar considerations of cost-aware learning have also been studied in other contexts, such as the "big data newsvendor" problem (Ban and Rudin 2019, Huber et al. 2019) and prescriptive analytics (Bertsimas and Kallus 2020). Second, we solve the cost-aware calibration problem by making adaptations to existing calibration methods (Platt scaling and temperature scaling) as well as proposing a new general-purpose algorithm named `MetaCal`. The `MetaCal` algorithm, in particular, has a clear theoretical foundation (i.e., Bayes optimal learning) and achieves superior performance based on our systematic evaluations. Finally, we take a step further and generalize the notion of calibration to consider nonlinear cost structures, and we demonstrate how our proposed approaches are readily adaptable to this case as well. Such a generalization further enhances the relevance of our work in decision tasks with nuanced cost structures that cannot be reduced to a simple linear form.

2. Theoretical Background and Problem Definition

In this section, we formally define the calibration problem and review several widely used calibration methods proposed in the prior literature that are relevant for us (Section 2.1). We then extend the standard definition of calibration to account for the potentially asymmetric costs of overprediction versus underprediction (Section 2.2). We focus on cost-aware calibration of *binary classifiers*. This is because simultaneous calibration of multiple classes has been shown to be generally infeasible, and the common strategy is to select one class of interest and convert a multiclass classification problem to a binary classification problem (Guo et al. 2017, Zhao et al. 2021).

2.1. Calibration of a Binary Classifier

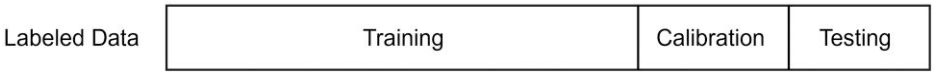
Consider a binary classification problem $\mathcal{X} \to \mathcal{Y} = \{0, 1\}$. Without loss of generality, we assume that class 1 is the main outcome of interest to decision makers (e.g., class "positive" in a medical diagnosis). The training, calibration, and evaluation of a classifier are done using a labeled data set. Figure 1 shows the different partitions of the labeled data set that correspond to each of these steps. Specifically, training data are used to train the classifier (it may be further partitioned for parameter tuning purposes), calibration data are used for calibrating probability predictions, and testing data are used for the final out-of-sample model evaluation.[2]

For a given data instance with features $x \sim \mathcal{X}$ and label $y \in \{0, 1\}$, suppose a classifier has been trained to produce a probability prediction of class 1, which we denote as $\Pr(y = 1|x)$ and write as $\Pr(1|x)$ as an abbreviation. The degree to which this classifier is calibrated is defined as the discrepancy between the predicted probabilities and the actual occurrence rates of class 1 (Platt 1999, Guo et al. 2017, Kuleshov et al. 2018). Formally, the classifier is perfectly calibrated if the following condition is satisfied:

$$\mathbb{E}(y|\Pr(1|x) = p) = p, \quad \forall p \in [0, 1] \tag{1}$$

For instance, among all data instances that receive an 80% probability prediction from a perfectly calibrated

Figure 1. Labeled Data Partitions for Training, Calibration, and Testing of a Predictive Model



Source. Adapted from Tuomo et al. (2020, figure 1).

classifier, we should expect that 80% of those data instances actually belong to class 1. In practice, it is generally infeasible to pursue perfect calibration in a continuous sense (as defined above) because not all possible values of $p \in [0,1]$ can be covered within a finite sample (Guo et al. 2017, Kuleshov et al. 2018).

Instead, a standard approach is to discretize the probability scores into B "bins" and pursue calibration within each bin. The total number of bins is typically informed by domain expertise and reflects the appropriate granularity in the domain of interest (e.g., in medical diagnoses, a small B creates a small number of risk levels, whereas a large B creates greater risk granularity). Specifically, each bin $b \in \{1,\dots,B\}$ uniquely corresponds to an interval $((b-1)/B, b/B]$. On the calibration data set, denote D_b as the subset of instances whose probabilities predictions fall into the interval $((b-1)/B, b/B]$.[3] Then, calibration within bin b requires that the average predicted probability equals the actual occurrence rate of class 1, that is,

$$\frac{1}{|D_b|}\sum_{x\in D_b} \Pr(1|x) = \frac{1}{|D_b|}\sum_{x\in D_b} y, \quad \forall b \in \{1,\dots,B\}. \tag{2}$$

For notational simplicity, we denote the left-hand side and the right-hand side of Equation (2) as $\overline{P}_b$ and $\overline{A}_b$, respectively. Such a discretized definition of calibration also naturally gives rise to a measurement of calibration error. The expected calibration error (ECE) (Niculescu-Mizil and Caruana 2005) is the weighted sum of absolute calibration error across all bins, that is,

$$ECE = \sum_{b=1}^{B} \frac{|D_b|}{|D|} \cdot |\overline{P}_b - \overline{A}_b|. \tag{3}$$

Several methods have been proposed in the literature to calibrate binary classifiers, and we discuss three of them that are particularly relevant to the design and evaluation of our proposed methods. First, the *isotonic regression* (IR) (Zadrozny and Elkan 2002) is a nonparametric calibration method. It fits a piecewise linear and nondecreasing function to the training data using the outcome labels as the dependent variable and the original (uncalibrated) probability predictions as the independent variable. Isotonic regression minimizes the mean squared error while preserving the order determined by the independent variable values, thereby making sure that the calibration does not change the ranking among data. Second, *Platt scaling* (Platt 1999) and *temperature scaling* (Guo et al. 2017) are two scaling-based parametric calibration methods. Both seek to transform the original predicted probabilities to become more calibrated. For example, consider a binary classifier that produces p as the predicted class 1 probability, and denote $\text{logit}(p) = \log(p/(1-p))$ as the corresponding logit value. Temperature scaling with parameter λ will transform the predicted probability to $\sigma(\lambda \cdot \text{logit}(p))$, and Platt scaling with parameter (A, B) will transform the predicted probability to $\sigma(A \cdot \text{logit}(p) + B)$, where $\sigma(.)$ is the standard sigmoid function. The parameters of these two methods can be chosen to minimize the ECE measure on the calibration data set. For Platt scaling, it is also common to estimate (A, B) via a logistic regression. Finally, some research also focuses on studying calibration under more specialized considerations, such as the calibration on small data sets (Tuomo et al. 2020) or the sample efficiency of different calibration algorithms (Kumar et al. 2019).

2.2. Cost-Aware Calibration

Next, we extend the standard definition of calibration error to account for the (potentially asymmetric) calibration costs. Let C_o and C_u represent the unit costs for overprediction and underprediction, respectively.[4] Given a data instance $x \sim \mathcal{X}$, suppose the predicted class 1 probability, $\Pr(1|x)$, takes value p and the actual class 1 probability, $\mathbb{E}(y|x)$, takes value a.[5] We define the calibration cost incurred by producing a probability prediction of p when the actual probability is a via the following cost function:

$$C(p,a) = C_o \mathbb{1}(p > a) \cdot |p-a| + C_u \mathbb{1}(p \le a) \cdot |p-a|, \tag{4}$$

where $\mathbb{1}()$ is the indicator function. As an illustrative example, consider a cancer detection scenario with $C_o = 1$ and $C_u = 5$; then, predicting a risk of 70% when the actual risk is 60% (i.e., overprediction) would incur a cost of $1 \cdot (0.7 - 0.6) = 0.1$, whereas predicting a risk of 70% when the actual risk is 80% (i.e., underprediction) would incur a cost of $5 \cdot (0.8 - 0.7) = 0.5$. Note that despite the asymmetric unit costs of overprediction versus underprediction, the cost associated with a specific pair of predicted and actual class 1 probabilities is still (linearly) proportional to their absolute difference (i.e., $|p-a|$); this is consistent with the standard formulation of calibration error (e.g., Equation (3)). Later in Section 5, we will consider more general forms of cost functions where the cost can depend on the absolute probability difference in a nonlinear manner.

Incorporating this cost function into the formulation of calibration error, we can then extend the ECE measure to quantify the calibration cost of a classifier on the calibration data set with B bins. Specifically, we define the expected calibration cost (ECC) measure as

$$ECC = \sum_{b=1}^{B} \frac{|D_b|}{|D|} \cdot C(\overline{P}_b, \overline{A}_b). \tag{5}$$

In the following illustration, we walk through a concrete example in a hypothetical context of cancer detection to showcase the calculation of ECE and ECC.

Example 1. Consider a cancer detection classifier that predicts a patient's probability of having malignant cancer (class 1) or not (class 0) based on patient characteristics and various testing results as input features. Suppose that the classifier's predicted probabilities are discretized into five "bins" $\{[0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1]\}$ corresponding to increasing levels of risk.

Now, suppose the classifier is evaluated on 100 patients in the calibration data set. The average predicted probability in each bin is $(0.1, 0.35, 0.45, 0.7, 0.95)$, the number of patients who fall into each bin is $(45, 10, 10, 10, 25)$, and the actual cancer rate of patients in each bin is $(0.05, 0.3, 0.6, 0.7, 0.85)$.

First, the classifier is clearly not fully calibrated. More specifically, it overpredicts the risk of cancer in bins 1, 2, and 5, and it underpredicts the risk of cancer in bin 3. The degree of miscalibration can be quantified by the ECE measure:

$$\begin{aligned} ECE &= \frac{45}{100} \cdot |0.1 - 0.05| + \frac{10}{100} \cdot |0.35 - 0.3| + \frac{10}{100} \\ &\quad \cdot |0.45 - 0.6| + \frac{25}{100} \cdot |0.95 - 0.85| \\ &= 0.0675. \end{aligned}$$

However, the ECE measure does not account for the asymmetric cost of overprediction versus underprediction. In the context of cancer detection, underprediction is arguably more costly than overprediction because underestimating a patient's actual risk of getting cancer can mislead them to ignore or postpone necessary testing or treatments until it is too late. Suppose $C_o = 1$ and $C_u = 5$ (i.e., underprediction is five times as costly as overprediction); the cost-aware ECC measure can be computed as follows:

$$\begin{aligned} ECC &= \frac{45}{100} \cdot (0.1 - 0.05) + \frac{10}{100} \cdot (0.35 - 0.3) + \frac{10}{100} \cdot 5 \\ &\quad \cdot (0.6 - 0.45) + \frac{25}{100} \cdot (0.95 - 0.85) \\ &= 0.1275. \end{aligned}$$

Finally, it is worth noting that cost-aware calibration is fundamentally different from the (seemingly similar) problem of cost-aware classification. Specifically, a binary classifier can make a class prediction by comparing the corresponding probability prediction with a prespecified cutoff threshold (e.g., 0.5). A cost-aware classification problem is typically formulated by considering the (potentially asymmetric) costs of different types of misclassifications, namely false positives and false negatives. In contrast, the cost-aware calibration problem that we study in this paper does not deal with class predictions per se (although the calibrated probabilities can certainly be converted into class predictions later if needed).[6]

3. Algorithm Design

We now propose two sets of algorithms to achieve cost-aware calibration. First, we adapt the existing scaling-based calibration methods for cost-aware calibration. Second, we propose a new general-purpose algorithm that enjoys a clearer theoretical foundation.

3.1. Cost-Aware Scaling

The ability to quantify calibration cost also enables us to adapt existing parametric calibration methods, most notably temperature scaling and Platt scaling, making them cost aware. Analogous to the practice of trying different cutoff thresholds of a binary classifier to find the one that reduces misclassification cost (Viaene and Dedene 2005), one can try different parameters of temperature/Platt scaling and search for the cost-minimizing parameter values.

Specifically, denote G as a set of candidate parameter values (e.g., generated via a grid-search process) of a given calibration method. For temperature scaling, one can search through $\lambda \in G$ for the scaling parameter that minimizes ECC. Similarly, for Platt scaling, one can search through $(A, B) \in G$ (instead of estimating (A, B) via a logistic regression). Note that this is essentially parameter tuning (with ECC as the selection criterion), and therefore, it can be achieved via a standard training-validation split or crossvalidation procedure (by further splitting the "training" data set shown in Figure 1).

Although such cost-aware adaptation of the popular scaling methods is quite straightforward, it is a heuristic approach in nature and lacks a clear theoretical foundation. With this limitation in mind, we next present a new cost-aware calibration algorithm based on the idea of Bayes optimal learning.

3.2. `MetaCal`

We now turn to the design of a general-purpose algorithm, which we call `MetaCal`, for cost-aware calibration. We start by describing the theoretical underpinning of `MetaCal`, which is inspired by the `MetaCost` algorithm for cost-aware classification (Domingos 1999).

For a given data instance $x \sim \mathcal{X}$, to achieve low calibration cost, we would ideally like the classifier-predicted class 1 probability (i.e., $\Pr(1|x)$) to be close to the (unknown) actual class 1 probability. Following the Bayes optimal learning approach (Dembczynski et al. 2010), we denote $f(.|x)$ as the density function of actual class 1 probability; then, the expected calibration cost associated with a particular probability prediction $\Pr(1|x) = p$ can be written as

$$ECC(p) = \int_0^1 C(p,a) \cdot f(a|x)da. \tag{6}$$

It follows that the cost-minimizing predicted probability would be $\tilde{p} = \arg\min_p ECC(p)$. In other words, for the purpose of minimizing calibration cost, it is more advantageous to learn to predict $\tilde{p}$ than the original binary class label. Note that the cost-minimizing $\tilde{p}$ can be considered as a form of risk-minimizing Bayes optimal prediction (Dembczynski et al. 2010); instead of trying to predict the original binary label, predicting $\tilde{p}$ can lead to lower risk (i.e., calibration cost in the current context).

Such a theoretical foundation of `MetaCal` naturally gives rise to its algorithmic design. At a high level, the `MetaCal` algorithm consists of three steps: (1) a *bootstrapping* step to empirically approximate the density of class 1 probability of each training instance (because the actual density function is not directly observable), (2) a *relabeling* step to produce cost-minimizing probabilities based on the prespecified cost function, and (3) a *learning* step to build a supervised model that predicts the cost-minimizing probabilities based on input features. We present the pseudocode for `MetaCal` in Algorithm 1.

Algorithm 1 (`MetaCal` Algorithm)

Data: Training data D_{train}; number of bootstrapping rounds R; unit costs of miscalibration C_o, C_u.

`//Bootstrapping step`

for each r *in* 1 *to* R **do**
  Sample with replacement from D_{train} to create $D_{train}^{(r)}$;
  Learn classifier $L^{(r)}$ on D_{train};
  $\forall x \in D_{train}$, use $L^{(r)}$ to calculate $\Pr^{(r)}(1|x)$;
end

`//Relabeling step`

$k \leftarrow \lceil \frac{C_u \cdot R}{C_u + C_o} \rceil$;
for each $x \in D_{train}$ **do**
  Rank vector $(\Pr^{(1)}(1|x), \ldots, \Pr^{(R)}(1|x))$ in ascending order. Denote the ranked vector as $(\Pr^{(i_1)}(1|x), \ldots, \Pr^{(i_R)}(1|x))$;
  Assign $\widetilde{\Pr}(1|x) \leftarrow \Pr^{(i_k)}(1|x)$;
end

`//Learning step`

Learn a numeric prediction model on D_{train} with $\widetilde{\Pr}(1|x)$ as numeric labels;

Output: The numeric prediction model that predicts class 1 probability given data instance $x \sim \mathcal{X}$.

We now provide a step-by-step discussion of the algorithm design. Let D_{train} denote a training data set for the binary classification problem under consideration, and let L denote a particular learning algorithm (e.g., decision tree). In the bootstrapping step, we sample from D_{train} with replacement to create R bootstrap samples $\{D_{train}^{(1)}, \ldots, D_{train}^{(R)}\}$. For each bootstrap sample, apply algorithm L to learn a classifier $L^{(r)}$ and then, use it to predict the class 1 probability of each data instance x in D_{train}, which we denote as $\Pr^{(r)}(1|x)$.

Next, in the relabeling step, we seek to create a new label for data instance x that minimizes the expected calibration cost. In contrast to `MetaCost` where the cost-aware label for each training data instance is chosen to be the class that minimizes the expected misclassification cost, a class-level relabeling approach cannot be directly applied in our case. This is because calibration is measured based on predicted probabilities rather than predicted classes. Instead, we propose to rely on Equation (6) and replace the unknown density function $f(.|x)$ by its empirical estimate, namely the sample class 1 probabilities obtained from the bootstrapping step:

$$\widetilde{\Pr}(1|x) = \arg\min_{q \in [0,1]} \frac{1}{R} \sum_{r=1}^{R} C(q, \Pr^{(r)}(1|x)). \tag{7}$$

The minimand on the right-hand side is essentially the average calibration cost of data instance x by treating $\Pr^{(r)}(1|x)$ as samples from the distribution of class 1 probability and q as the cost-minimizing predicted probability to be determined.

The minimization problem in Equation (7) can be solved via grid search or existing optimization approaches (e.g., gradient descent) at the expense of computational inefficiency or potential lack of optimality guarantee. Instead, we are able to derive the following Theorem 1 to show a closed-form solution to this minimization problem.

Theorem 1. *Let* $(\Pr^{(i_1)}(1|x), \ldots, \Pr^{(i_R)}(1|x))$ *be the ranked vector of* $(\Pr^{(1)}(1|x), \ldots, \Pr^{(R)}(1|x))$ *in ascending order. Let* $k = \lceil (C_u \cdot R)/(C_u + C_o) \rceil$. *Then,* $\Pr^{(i_k)}(1|x)$ *minimizes the expected calibration cost for data instance* $x \in D_{train}$.

The proof of this theorem is included in the Appendix. We offer two remarks on this result. First, it can be seen that as R increases, the cost-minimizing $\Pr^{(i_k)}(1|x)$ with $k = \lceil (C_u \cdot R)/(C_u + C_o) \rceil$ would tend to the $(C_u/(C_u + C_o))$th quantile of $f(.|x)$ in the limit.[7] Second, we further note that k (i.e., the index of the cost-minimizing probability) does not depend on individual values of C_o and C_u; it only depends on the ratio C_o/C_u. Therefore, when applying `MetaCal` in practice, users generally do not need to solicit exact amounts of the costs (i.e., how much do overprediction and underprediction cost, respectively) and only need to specify their ratio

(i.e., how many times more or less costly is overprediction compared with underprediction). This enhances the applicability of `MetaCal` in domains where explicit costs are hard to estimate.

Finally, in the learning step, a new predictive model is trained on D_{train}, with the derived cost-minimizing probabilities from the previous steps as labels. Because the derived labels take numeric values in $[0, 1]$, this step amounts to building a numeric prediction model. At deployment time, for an unlabeled data instance, one can directly apply the numeric prediction model to obtain a probability prediction (which can then be discretized into a class prediction if desired).

4. Empirical Evaluations

In this section, we empirically evaluate the performance of cost-aware calibration methods (including cost-aware scaling and `MetaCal`). We carry out systematic evaluations on both tabular data sets with classical machine learning techniques as well as unstructured text and image data sets with modern deep learning techniques. We also explore the performance sensitivity of `MetaCal` with respect to the number of bootstrapping rounds.

4.1. Evaluations on Tabular Data Sets

We apply the cost-aware calibration methods to five public tabular data sets from the University of California, Irvine Machine Learning Repository and Kaggle.com: Wisconsin Breast Cancer (Wolberg and Mangasarian 1990), Cardiotocography (Ayres-de Campos et al. 2000), Customer Churn (Vafeiadis et al. 2015), Default of Credit Card Clients (Yeh and Lien 2009), and Bank Marketing (Moro et al. 2014). These data sets pertain to binary classification problems in healthcare, customer churn management, finance, and marketing. We provide a detailed description of them as follows.

The Wisconsin Breast Cancer (Cancer in short) data set ($N = 683$) contains cytological characteristics of breast fine-needle aspirates, which can be used to diagnose breast cancer. The task is to predict breast cancer diagnostic results (malignant versus benign) based on nine cytological characteristics.

The Cardiotocography (CTG in short) data set ($N = 2{,}126$) provides measurements of fetal heart rate and uterine contraction features on cardiotocograms as well as diagnoses of overall fetal states determined by expert obstetricians. The task is to detect suspicious or pathological fetal states based on 21 heart rate and uterine contraction features. Note that fetal state in the original data set can take one of three values: normal, suspect, or pathological. We pool together the "suspect" and "pathological" instances into a single category to formulate a binary classification task (i.e., predicting normal versus abnormal fetal state).

The Customer Churn (Churn in short) data set ($N = 5{,}000$) contains service data on customers of a telecommunications provider (e.g., daily number of calls and 16 other features), and the task is to predict whether a customer will stay with the provider or churn.

The Default of Credit Card Clients (Default in short) data set ($N = 30{,}000$) includes 23 features of customers' demographics and transaction details to predict default payments on their credit cards.

Finally, the Bank Marketing (Bank in short) data set ($N = 45{,}211$) records 16 features of socioeconomic characteristics of a Portuguese bank's clients, and the task is to predict whether a client will make a term deposit in response to the bank's telemarketing campaigns.

4.1.1. Evaluation Setup and Calibration Methods.

Each data set is randomly partitioned into 70% training for building the binary classifier and 30% calibration for evaluating the calibration performance based on the ECC measure. Importantly, regardless of which calibration method we use, the binary classifier is always trained on the same 70% partition using a decision tree algorithm, and the 30% calibration partition also remains the same. Doing so ensures the fairness of comparison.

We evaluate seven calibration methods; the first four are cost unaware (for benchmarking purposes), and the last three are cost aware. Their configurations are described as follows.

1. NoCal. Directly calculate the ECC of a binary classifier without performing any calibration. This serves as a benchmark to show the calibration cost without any explicit calibration effort.
2. IR. Apply the isotonic regression method. Because of the nonparametric nature of isotonic regression, it cannot easily be adapted to become cost aware.
3. TS. Apply the cost-unaware temperature scaling method.
4. PS. Apply the cost-unaware Platt scaling method.

The IR, TS, and PS are all designed to reduce ECE (rather than ECC). They serve as benchmarking calibration methods that do not take into account the asymmetric calibration costs.

5. CATS. Apply the cost-aware temperature scaling method.
6. CAPS. Apply the cost-aware Platt scaling method.
7. `MetaCal`. Apply the `MetaCal` method with 100 bootstrapping rounds ($R = 100$), and use a decision tree to learn the numeric prediction model during the "learning step" (i.e., line 11 in Algorithm 1).

For the scaling-based methods, including TS, PS, CATS, and CAPS, we need to search for an advantageous set of scaling parameters. This is done by further partitioning the training data (70% of the entire data) into two parts; 50% of the entire data is used for building the classifier, and 20% of the entire data is used for parameter tuning. For TS and CATS, we search through

200 parameter values $\lambda \in \{0.01, 0.02, \ldots, 2.00\}$ to identify the one that achieves the best calibration performance on the 20% data partition (i.e., the lowest ECE for TS or the lowest ECC for CATS). For Platt scaling, we follow the same procedure to search through 200 pairs of parameters (A, B), where both A and B are randomly sampled from $\{0.01, 0.02, \ldots, 2.00\}$. Importantly, this partitioning of the training data is only done to search for scaling parameter values of the four scaling-based methods. When we evaluate the calibration performance of these methods, the classifiers are still trained on the 70% training data, and the ECC measures are calculated on the 30% calibration data.

Across all methods, we set $C_o = 1$ and $C_u = 5$, and we use 10 bins ($B = 10$) when calculating calibration costs. We repeat the experiments for 100 iterations, each time with a different random data partition.

4.1.2. Evaluation Results. The means and standard deviations of ECC for each method are reported in Table 1. The second through fifth columns report the performance of cost-unaware calibration methods, and the last three columns report the performance of cost-aware calibration methods. In addition, we perform pair-wise t tests comparing the ECC of `MetaCal` with each of the other methods, and we report the test results in Table 1 as well.

We highlight several observations based on the evaluation results. First, the ECC of isotonic regression is substantially higher than that of no calibration. In other words, when overprediction and underprediction have asymmetric costs, performing cost-unaware calibration (i.e., assuming equal costs of overprediction and underprediction) can be counterproductive and worse than not performing calibration at all. This is also true for the cost-unaware temperature scaling and Platt scaling methods in a number of cases. Second, among the three cost-aware calibration methods, `MetaCal` consistently outperforms the no calibration benchmark (and by extension, the isotonic regression benchmark), and the differences are statistically significant at the 0.1% level. Third, cost-aware temperature scaling and Platt scaling outperform the no calibration benchmark on two of the five data sets but fail to do so on the other three data sets. When the two cost-aware scaling methods outperform no calibration, their ECCs are actually statistically indifferent from that of `MetaCal`. Compared with their cost-unaware counterparts, the two cost-aware scaling methods generally result in statistically indifferent or lower ECC. Finally, across all five data sets, the standard deviations of calibration costs associated with `MetaCal` are always lower than those associated with the scaling-based approaches, indicating that `MetaCal`'s performance is more stable. Overall, this set of experiments clearly demonstrates `MetaCal`'s effectiveness in reducing calibration costs. It is able to (1) outperform the no calibration benchmark (and by extension, isotonic regression), (2) perform at least as good as (and often better than) the scaling-based approaches on these data sets, and (3) have more stable performance than the scaling-based approaches.

Next, we repeat the above experiments using random forest (a representative ensemble learning model) as the binary classification technique. All other experimental setups are kept the same, and the results are reported in Table 2. We see that all three cost-aware calibration methods outperform the four benchmarks (i.e., NoCal, IR, TS, and PS) on all data sets but one (Default). `MetaCal` and cost-aware Platt scaling achieve comparable performance on four of five data sets, whereas the two scaling-based methods perform better on the credit default data set. `MetaCal` continues to have the smallest performance variance (even on the credit default data set).

4.2. Evaluations on Textual Data

Recognizing the prevalence of deep learning techniques for language and vision tasks, we further carry out

Table 1. Cost-Aware Calibration Evaluation Results ($C_o = 1, C_u = 5$)

Data set	NoCal	IR	TS	PS	CATS	CAPS	`MetaCal`
Cancer	0.0822***	0.1561***	0.1033***	0.0967***	0.1015***	0.1035***	0.0577
	(0.0449)	(0.1484)	(0.0596)	(0.0938)	(0.0599)	(0.1098)	(0.0182)
CTG	0.0841***	0.2801***	0.0626	0.0574	0.0667	0.0646	0.0670
	(0.0401)	(0.1238)	(0.0333)	(0.0285)	(0.0353)	(0.0306)	(0.0135)
Churn	0.0472***	0.0992***	0.0421*	0.0431**	0.0366	0.0397	0.0369
	(0.0203)	(0.0707)	(0.0227)	(0.0227)	(0.0193)	(0.0161)	(0.0071)
Default	0.0102***	0.4495***	0.0134***	0.0207***	0.0115***	0.0156***	0.0071
	(0.0070)	(0.4007)	(0.0091)	(0.0178)	(0.0090)	(0.0100)	(0.0055)
Bank	0.0046***	0.1674***	0.0064***	0.0115***	0.0061***	0.0082***	0.0034
	(0.0037)	(0.2344)	(0.0055)	(0.0090)	(0.0046)	(0.0048)	(0.0020)

*$p < 0.05$ for pair-wise t tests comparing `MetaCal` with other methods; **$p < 0.01$ for pair-wise t tests comparing `MetaCal` with other methods; ***$p < 0.001$ for pair-wise t tests comparing `MetaCal` with other methods.

Table 2. Cost-Aware Calibration Evaluation Results ($C_o = 1, C_u = 5$, Random Forest Classifier)

Data set	NoCal	IR	TS	PS	CATS	CAPS	MetaCal
Cancer	0.1281***	0.0963***	0.1120***	0.0786**	0.1029***	0.0697*	0.0601
	(0.0353)	(0.0426)	(0.0441)	(0.0529)	(0.0353)	(0.0533)	(0.0292)
CTG	0.1158***	0.0715***	0.0960***	0.0681***	0.0949***	0.0496	0.0539
	(0.0255)	(0.0271)	(0.0311)	(0.0286)	(0.0310)	(0.0207)	(0.0130)
Churn	0.1069***	0.0616	0.1029***	0.0700***	0.1016***	0.0473	0.0574
	(0.0190)	(0.0288)	(0.0193)	(0.0225)	(0.0191)	(0.0142)	(0.0064)
Default	0.0601	0.2686***	0.0694	0.0804	0.0529	0.0548	0.1250
	(0.0095)	(0.1413)	(0.0178)	(0.0263)	(0.0097)	(0.0079)	(0.0028)
Bank	0.1238***	0.3168***	0.0486***	0.0590***	0.0294	0.0340	0.0328
	(0.0082)	(0.0807)	(0.0143)	(0.0215)	(0.0072)	(0.0083)	(0.0018)

*$p < 0.05$ for pair-wise t tests comparing MetaCal with other methods; **$p < 0.01$ for pair-wise t tests comparing MetaCal with other methods; ***$p < 0.001$ for pair-wise t tests comparing MetaCal with other methods.

evaluations on a textual data set. Specifically, we use the Internet Movie Database Movie Review data set (Maas et al. 2011), where each movie review has been labeled as either positive or negative. For each review, we employ the popular pretrained Bidirectional encoder representations from transformers model (Devlin et al. 2018) to obtain its embedding (i.e., average of the word-level embeddings), which is subsequently fed into a SoftMax layer to predict its sentiment. We follow the same evaluation procedures as before to apply the three cost-aware calibration algorithms and compare their performances with the four cost-unaware benchmarks. As summarized in Table 3, all three cost-aware calibration algorithms outperform the no calibration or isotonic regression benchmarks. MetaCal achieves similar performance as cost-aware temperature scaling but underperforms cost-aware Platt scaling; nevertheless, MetaCal again has the smallest performance variance. The two cost-aware scaling methods have lower ECC than their cost-unaware counterparts, although the differences are small.

4.3. Sensitivity Analyses of MetaCal

When applying the MetaCal approach, users need to choose the number of bootstrapping rounds (R). In this subsection, we carry out a set of sensitivity analyses to understand how MetaCal's performance changes with different numbers of bootstrapping rounds. Specifically, we vary $R \in \{50, 100, 200\}$ to represent an increasing level of bootstrapping intensity. The sensitivity analyses are performed on all five data sets discussed in Section 4.1 with the same experimental procedures. We report the results in Table 4.

We find minimal differences in MetaCal's performance with different bootstrapping rounds, indicating that a modest number of bootstrapping rounds is usually enough to approximate the empirical distribution of class 1 probability. Using a large number of bootstrapping rounds (e.g., $R = 200$) incurs more computational overhead without clear performance benefit. Of course, how small R can be without hurting calibration performance is an empirical question. If needed, users of MetaCal could select R by following a standard parameter-tuning procedure, similar to what has been done for the two cost-aware scaling methods.

5. Generalization to Nonlinear Cost Structures

So far, our theoretical discussions and algorithm designs of cost-aware calibration have assumed the cost function specified in Equation (4), which corresponds to a *linear* cost structure as the calibration cost grows proportionally with respect to the absolute difference between predicted and actual probabilities. In this section, we generalize our key results to accommodate other types of cost structures, where the cost can grow in a nonlinear fashion as the absolute difference between predicted and actual probabilities increases.

5.1. General Cost Function

Keeping the unit cost of overprediction/underprediction unchanged, we generalize the cost function to

$$C(p, a) = C_o \mathbb{1}(p > a) \cdot d(p, a) + C_u \mathbb{1}(p \leq a) \cdot d(p, a), \quad (8)$$

where $d(p, a)$ is a function that measures the *distance*

Table 3. Sentiment Data Set Evaluation Results ($C_o = 1, C_u = 5$)

Data set	NoCal	IR	TS	PS	CATS	CAPS	MetaCal
Sentiment	0.0500***	0.0516***	0.0456	0.0335	0.0442	0.0329	0.0468
	(0.0077)	(0.0068)	(0.0169)	(0.0122)	(0.0173)	(0.0177)	(0.0026)

***$p < 0.001$ for pair-wise t tests comparing MetaCal with other methods.

Table 4. Sensitivity Analyses of `MetaCal`: Bootstrapping Rounds

Data set	$R = 50$	$R = 100$	$R = 200$
Cancer	0.0561 (0.0199)	0.0577 (0.0182)	0.0561 (0.0179)
CTG	0.0672 (0.0145)	0.0670 (0.0135)	0.0674 (0.0148)
Churn	0.0366 (0.0076)	0.0369 (0.0071)	0.0367 (0.0075)
Default	0.0070 (0.0054)	0.0071 (0.0055)	0.0071 (0.0055)
Bank	0.0034 (0.0020)	0.0034 (0.0020)	0.0034 (0.0020)

between predicted probability p and actual probability a. By definition, the distance function $d(p, a)$ has two properties. (1) Its values are always nonnegative (i.e., $d(p,a) \geq 0$), and (2) it takes a value of zero if the predicted probability equals actual probability (i.e., $\forall p \in [0,1]$, $d(p,p) = 0$).

Choosing $d(p,a) = |p-a|$ would produce the linear cost structure that we have focused on before. However, depending on the specific decision domains and application contexts, one can choose different forms of $d(p, a)$ to reflect how calibration costs vary with different degrees of discrepancy between p and a. Below, we illustrate a few categories of nonlinear distance functions and their corresponding interpretations.

1. Polynomial: $|p-a|^k$, $k > 0$. For a fixed value of $|p-a|$, it magnifies the distance when $0 < k < 1$ and attenuates the distance when $k > 1$. It is concave in $[0,1]$ when $0 < k < 1$ and convex when $k > 1$.

2. Exponential: $\exp(|p-a|) - 1$. For a fixed value of $|p-a|$, it magnifies the distance based on the exponential function. It is convex in $[0,1]$.

3. Logarithm: $\ln(|p-a|+1)$. For a fixed value of $|p-a|$, it attenuates the distance based on the logarithm function. It is concave in $[0,1]$.

These distance functions offer greater flexibility to model nuanced, nonlinear cost structures. Given a fixed absolute difference between the predicted and actual probabilities (i.e., fixed $|p-a|$), choosing different distance functions affects not only how much calibration cost is incurred but also, how the cost varies with different values of $|p-a|$ (in a nonlinear way). In particular, the polynomial function with $0 < k < 1$ and the exponential function magnify the cost for a fixed $|p-a|$. They are useful to model application scenarios with high cost sensitivity, where even a relatively small $|p-a|$ can result in large calibration cost. Moreover, the polynomial function with $0 < k < 1$ is concave in $[0,1]$, whereas the exponential function is convex. Therefore, the former can be used to model "marginally decreasing" cost as $|p-a|$ increases, and the latter can be used to model "marginally accelerating" cost. Conversely, the polynomial function with $k > 1$ and the logarithm function attenuate the cost for a fixed $|p-a|$, which is useful in application scenarios with more "tolerance" of miscalibration. The two functions also differ in their concavity; the polynomial function with $k > 1$ is convex in $[0,1]$, whereas the logarithm function is concave.

5.2. Cost-Aware Calibration Under Nonlinear Cost Structures

Our proposed cost-aware scaling approaches and `MetaCal` can be readily adapted to achieve cost-aware calibration under nonlinear cost structures. First, we note that the measurement of ECC follows the same Equation (5) simply by plugging in the desirable cost function. For instance, suppose a quadratic distance function is chosen for a certain application; then, $ECC = \sum_{b=1}^{B} |D_b|/|D| \cdot C(\overline{P}_b, \overline{A}_b)$, where $C(\overline{P}_b, \overline{A}_b) = C_o \mathbb{1}(\overline{P}_b > \overline{A}_b) \cdot |\overline{P}_b - \overline{A}_b|^2 + C_u \mathbb{1}(\overline{P}_b \leq \overline{A}_b) \cdot |\overline{P}_b - \overline{A}_b|^2$. As a result, cost-aware temperature scaling or Platt scaling can be directly carried out using the appropriate ECC for parameter tuning.

Second, to apply `MetaCal` under nonlinear cost structures, only the relabeling step needs to be modified. Although the cost-minimizing probabilities can still be obtained by solving the same optimization problem specified in Equation (7) (with the desirable cost function plugged in), the closed-form solution in Theorem 1 no longer holds for nonlinear cost functions. In fact, unlike the result under a linear cost function, the cost-minimizing probabilities under nonlinear cost functions may not equal any of $\{\Pr^{(1)}(1|x), \ldots, \Pr^{(R)}(1|x)\}$ (i.e., the sample class 1 probabilities from bootstrapping). Therefore, the cost minimization problem in Equation (7) would need to be dealt with as a univariate constrained optimization problem. Accordingly, lines 6–9 in Algorithm 1 need to be replaced by the procedure of solving such an optimization problem. Fortunately, many off-the-shelf optimization algorithms, such as Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS-B) (Zhu et al. 1997), are readily applicable.

5.3. Empirical Evaluation

We empirically evaluate the performance of the two scaling-based approaches as well as `MetaCal` under three different nonlinear cost structures. Specifically, we pick one specific distance function from each category discussed in Section 5.1, including (1) a polynomial (square-root) distance function $d(p,a) = |p-a|^{1/2}$, (2) an exponential distance function $d(p,a) = \exp(|p-a|) - 1$, and (3) a logarithm distance function $d(p,a) = \ln(|p-a|+1)$. For `MetaCal`, we use the L-BFGS-B algorithm to solve for the cost-minimizing probabilities. The evaluations are carried out using the same five data sets and benchmarking setup as before, and the results are included in Tables 5–7.

Table 5. Cost-Aware Calibration Evaluation Results for $d(p,a) = |p-a|^{1/2}$

Data set	NoCal	IR	TS	PS	CATS	CAPS	MetaCal
Cancer	0.4701***	0.7183***	0.5180***	0.4165***	0.4922***	0.4013***	0.2469
	(0.1995)	(0.3580)	(0.2048)	(0.2181)	(0.1997)	(0.2088)	(0.0489)
CTG	0.4957***	0.8989***	0.3402***	0.3184*	0.3527***	0.3061*	0.2763
	(0.2084)	(0.3297)	(0.1828)	(0.1820)	(0.1937)	(0.1606)	(0.0263)
Churn	0.3428***	0.4953***	0.2756***	0.2692***	0.2665***	0.2487***	0.1879
	(0.1410)	(0.3087)	(0.1599)	(0.1664)	(0.1407)	(0.1332)	(0.0222)
Default	0.1378***	1.0943***	0.1674***	0.1727***	0.1424***	0.1357***	0.0891
	(0.0761)	(0.9908)	(0.1052)	(0.1328)	(0.0804)	(0.0925)	(0.0380)
Bank	0.0903***	0.5334***	0.1169***	0.1302***	0.0968***	0.0875***	0.0582
	(0.0685)	(0.7469)	(0.0893)	(0.0945)	(0.0722)	(0.0661)	(0.0159)
Sentiment	0.3179***	0.4690***	0.3098***	0.2037	0.3097***	0.2037	0.2632
	(0.0817)	(0.0393)	(0.0888)	(0.0696)	(0.0870)	(0.0696)	(0.0053)

*$p < 0.05$ for pair-wise t tests comparing MetaCal with other methods; ***$p < 0.001$ for pair-wise t tests comparing MetaCal with other methods.

We observe largely consistent patterns as in the case of linear cost (i.e., when $d(p,a) = |p-a|$). Compared with the no calibration baseline, the cost-unaware isotonic regression further exacerbates the calibration costs, the two scaling-based cost-aware approaches are able to reduce calibration costs on some of the data sets, and MetaCal is consistently the most effective approach that results in the lowest calibration costs as well as the highest performance stability (i.e., having the smallest standard deviations of calibration costs). The cost-aware scaling methods result in lower ECC than their cost-unaware counterparts in many cases. Notably, under the square-root distance function (Table 5), MetaCal achieves significantly smaller calibration costs than almost all other benchmarks, with the only exception being the (textual) Sentiment data set, on which Platt scaling outperforms MetaCal.

6. Discussion

Practical applications of predictive machine learning models often require not only accurate predictions but also, the ability to adequately quantify the uncertainty levels in those predictions. For binary classifiers specifically, this means having well-calibrated probability predictions in addition to accurate class predictions. Although several calibration methods have been developed in the machine learning literature, they generally treat overpredictions and underpredictions equally without accounting for the potentially asymmetric costs associated with the two types of miscalibration. In this research, we fill the gap by addressing the problem of cost-aware calibration.

First, we formally define the cost-aware calibration problem for a binary classifier and the expected calibration cost measurement, which serve as the necessary

Table 6. Cost-Aware Calibration Evaluation Results for $d(p,a) = \exp(|p-a|) - 1$

Data set	NoCal	IR	TS	PS	CATS	CAPS	MetaCal
Cancer	0.0850***	0.1676***	0.1070***	0.1066***	0.1059***	0.1142***	0.0599
	(0.0464)	(0.1736)	(0.0615)	(0.1245)	(0.0631)	(0.1450)	(0.0207)
CTG	0.0875***	0.3092***	0.0660	0.0599	0.0706	0.0683	0.0670
	(0.0413)	(0.1502)	(0.0348)	(0.0293)	(0.0364)	(0.0320)	(0.0143)
Churn	0.0489***	0.1048***	0.0437*	0.0442**	0.0385	0.0405	0.0377
	(0.0211)	(0.0757)	(0.0234)	(0.0227)	(0.0209)	(0.0170)	(0.0085)
Default	0.0102***	0.4901***	0.0135***	0.0222***	0.0116***	0.0157***	0.0072
	(0.0071)	(0.4352)	(0.0092)	(0.0208)	(0.0091)	(0.0101)	(0.0056)
Bank	0.0047***	0.1759***	0.0065***	0.0117***	0.0061***	0.0084***	0.0034
	(0.0038)	(0.2463)	(0.0055)	(0.0091)	(0.0046)	(0.0049)	(0.0020)
Sentiment	0.1331***	0.1911***	0.1322***	0.0902	0.1320***	0.0901	0.1189
	(0.1396)	(0.1982)	(0.1364)	(0.0909)	(0.1360)	(0.0909)	(0.1023)

*$p < 0.05$ for pair-wise t tests comparing MetaCal with other methods; **$p < 0.01$ for pair-wise t tests comparing MetaCal with other methods; ***$p < 0.001$ for pair-wise t tests comparing MetaCal with other methods.

Table 7. Cost-Aware Calibration Evaluation Results for $d(p,a) = \ln(|p-a|+1)$

Data set	NoCal	IR	TS	PS	CATS	CAPS	MetaCal
Cancer	0.0797***	0.1475***	0.101***	0.090***	0.0977***	0.0960***	0.0543
	(0.0436)	(0.1313)	(0.0565)	(0.0775)	(0.0571)	(0.0898)	(0.0171)
CTG	0.0813***	0.2583***	0.0598	0.0555	0.0634	0.0618	0.0649
	(0.0391)	(0.1077)	(0.0324)	(0.0273)	(0.0336)	(0.0295)	(0.0144)
Churn	0.0457***	0.0947***	0.0408*	0.0415*	0.0355	0.0375	0.0362
	(0.0196)	(0.0670)	(0.0220)	(0.0223)	(0.0184)	(0.0153)	(0.0072)
Default	0.0101***	0.4151***	0.0133***	0.0202***	0.0114***	0.0155***	0.0071
	(0.0069)	(0.3712)	(0.0090)	(0.0176)	(0.0089)	(0.0098)	(0.0054)
Bank	0.0046***	0.1596***	0.0064***	0.0114***	0.0061***	0.0079***	0.0034
	(0.0037)	(0.2235)	(0.0055)	(0.0089)	(0.0045)	(0.0045)	(0.0020)
Sentiment	0.1790***	0.2603***	0.1763***	0.1182	0.1763***	0.1182	0.1551
	(0.1510)	(0.2111)	(0.1481)	(0.0994)	(0.1481)	(0.0994)	(0.1084)

*$p < 0.05$ for pair-wise t tests comparing MetaCal with other methods; ***$p < 0.001$ for pair-wise t tests comparing MetaCal with other methods.

theoretical foundation for algorithm design. Second, we propose cost-aware adaptations of temperature/Platt scaling. These cost-aware scaling-based approaches are algorithmically straightforward and can already reduce calibration costs effectively. Third, beyond the two heuristic scaling-based approaches, we also propose a general-purpose algorithm, MetaCal, that has a more systematic theoretical underpinning (i.e., risk-minimizing Bayes optimal learning). We empirically demonstrate MetaCal's effectiveness on multiple data sets. Compared with the cost-aware scaling approaches, MetaCal achieves lower calibration costs on several data sets. Even in the occasional cases when MetaCal underperforms the scaling-based approaches, it still consistently has greater performance stability. Further sensitivity analyses suggest that a modest number of bootstrapping rounds is usually sufficient when using MetaCal. Meanwhile, MetaCal is computationally more demanding, and therefore, the two cost-aware scaling methods can be used on large data sets or with complex classification techniques.

We then generalize the cost-aware calibration problem to account for other types of cost structures, where the calibration costs incurred may not simply be a linear function of the absolute difference between predicted and actual class 1 probabilities. Our proposed ECC measure and the two scaling-based calibration approaches as well as MetaCal are all extensible to the cases of general cost structures. Through empirical evaluations, we again demonstrate their effectiveness in reducing calibration costs.

We conclude by discussing a few interesting directions for future research. First, researchers can study the cost-aware calibration problem under more intricate cost structures. For example, perhaps the cost of predicting a probability p when the actual outcome rate is a may not be a continuous function of $|p-a|$ but rather, a step function. This can be relevant in decision tasks where the calibration cost depends on the "level" of $|p-a|$, where low, medium, and high levels of $|p-a|$ can incur very different costs. Second, if external domain knowledge on the distribution of class probability (i.e., $f(.|x)$ in Equation (6)) can be obtained, one can seek to do away with the bootstrapping step in MetaCal, which can potentially improve the computational efficiency of the method. Third, although calibration is an intuitive concept for classification tasks, it is less well studied for numeric prediction tasks, where the target outcomes to be predicted are continuous. Formulating a practically useful calibration notion for general numeric prediction models and accounting for the costs of miscalibration also represent a promising research direction. Finally, users who engage in model calibration may also need to consider the model's classification performance, and it is important to understand the classification performance impact of different calibration methods. Suppose we fix a threshold for obtaining class predictions (e.g., at the default value of 0.5). Do different calibration methods (both cost aware and cost unaware) hurt or benefit the model's classification performance? We encourage future research to explore this question under different data sets or contexts.

Appendix. Proof of Theorem 1

For notational simplicity, we write all $\Pr^{(\cdot)}(1|x)$ simply as $p^{(\cdot)}$. Recall the cost minimization problem:

$$\min_{q\in[0,1]} \frac{1}{R}\sum_{r=1}^{R}[C_o \cdot \mathbb{1}(q > p^{(r)}) \cdot (q - p^{(r)}) + C_u \cdot \mathbb{1}(q \le p^{(r)}) \cdot (p^{(r)} - q)].$$

We first point out that the optimal value of q can only be one of $\{p^{(1)},\ldots,p^{(R)}\}$ because the above minimand is piecewise monotonic. Let $(p^{(i_1)},\ldots,p^{(i_R)})$ be the ranked vector of $(p^{(1)},\ldots,p^{(R)})$ in ascending order, and consider the intervals

$[0, p^{(i_1)}], [p^{(i_1)}, p^{(i_2)}], \ldots, [p^{(i_R)}, 1]$. Note that $\forall q \in [0, p^{(i_1)}]$, the cost-minimizing q should clearly be $p^{(i_1)}$; similarly, $\forall q \in [p^{(i_R)}, 1]$, the cost-minimizing q should clearly be $p^{(i_R)}$. Furthermore, $\forall q \in [p^{(i_k)}, p^{(i_{k+1})}]$ for $k \geq 1$, the minimization problem can be written as

$$\min_{q \in [p^{(i_k)}, p^{(i_{k+1})}]} \frac{1}{R}\left[(kC_o - (R-k)C_u)q - \sum_{r=1}^{k} C_o \cdot p^{(i_r)} + \sum_{r=k+1}^{R} C_u \cdot p^{(i_r)}\right].$$

The cost-minimizing q in this interval depends on the relative magnitude of C_o versus C_u; specifically, q should be set to $p^{(i_k)}$ if $kC_o > (R-k)C_u$ and to $p^{(i_{k+1})}$ otherwise.

Therefore, the original continuous minimization problem is equivalent to the following discrete form:

$$\begin{aligned}
&\min_{k \in \{1, \ldots, R\}} \frac{1}{R} \sum_{r=1}^{R} [C_o \cdot \mathbb{1}(p^{(i_r)} \leq p^{(i_k)}) \cdot (p^{(i_k)} - p^{(i_r)}) \\
&\qquad + C_u \cdot \mathbb{1}(p^{(i_r)} > p^{(i_k)}) \cdot (p^{(i_r)} - p^{(i_k)})] \\
&= \min_{k \in \{1, \ldots, R\}} \frac{1}{R}\left[\sum_{r=1}^{k-1} C_o \cdot (p^{(i_k)} - p^{(i_r)}) + \sum_{r=k+1}^{R} C_u \cdot (p^{(i_r)} - p^{(i_k)})\right] \\
&\overset{\text{def}}{=} \min_{k \in \{1, \ldots, R\}} \frac{1}{R} \cdot L_k.
\end{aligned}$$

Next, $\forall k \in \{1, \ldots, R-1\}$, we have

$$\begin{aligned}
L_{k+1} - L_k &= \left[\sum_{r=1}^{k} C_o \cdot (p^{(i_{k+1})} - p^{(i_r)}) + \sum_{r=k+2}^{R} C_u \cdot (p^{(i_r)} - p^{(i_{k+1})})\right] \\
&\quad - \left[\sum_{r=1}^{k-1} C_o \cdot (p^{(i_k)} - p^{(i_r)}) + \sum_{r=k+1}^{R} C_u \cdot (p^{(i_r)} - p^{(i_k)})\right] \\
&= \left[\sum_{r=1}^{k} C_o \cdot (p^{(i_{k+1})} - p^{(i_r)}) - \sum_{r=1}^{k-1} C_o \cdot (p^{(i_k)} - p^{(i_r)})\right] \\
&\quad + \left[\sum_{r=k+2}^{R} C_u \cdot (p^{(i_r)} - p^{(i_{k+1})}) - \sum_{r=k+1}^{R} C_u \cdot (p^{(i_r)} - p^{(i_k)})\right] \\
&= C_o \cdot k \cdot (p^{(i_{k+1})} - p^{(i_k)}) - C_u \cdot (R-k) \cdot (p^{(i_{k+1})} - p^{(i_k)}) \\
&= [kC_o - (R-k)C_u](p^{(i_{k+1})} - p^{(i_k)}).
\end{aligned}$$

Because $p^{(i_{k+1})} > p^{(i_k)}$ by definition, the minimal is obtained when k takes the smallest possible integer value such that $kC_o - (R-k)C_u > 0$. This gives $k^* = \lceil \frac{C_u \cdot R}{C_u + C_o} \rceil$.

Endnotes

[1] We note that the incorporation of (potentially asymmetric) costs in machine learning problems has also been referred to as "cost-*sensitive*" learning (e.g., Domingos 1999, Herńndez-Orallo 2014). For consistency, we use the term "cost-*aware*" learning throughout this paper.

[2] We note that the term "calibration data" is sometimes used equivalently as the validation data for parameter tuning (e.g., Zumel and Mount 2019). To avoid confusion, we do not use these terms interchangeably in this paper.

[3] For completeness, we define a probability prediction of exactly zero to fall into the first interval.

[4] If $C_o \approx C_u$, then it is apparent that there is little need for cost-aware calibration. Therefore, we consider the situation where C_o and C_u are meaningfully different in this paper.

[5] By "actual class 1 probability," we refer to the inherent uncertainty about the binary outcome given the observable features. Although the actual outcome of any data instance is either zero or one, there will be some degree of uncertainty in the binary outcome, conditional on a limited set of observable features. In the (hypothetical) scenario where one observes all relevant features that fully determine the outcome, then the actual probability will equal the true binary outcome.

[6] Some research has indicated that calibrating a classifier's predicted probabilities can also subsequently reduce the false-positive rate (e.g., Watson et al. 2021 in the context of privacy-preserving machine learning). However, understanding the impact of calibration on (different types of) misclassification is beyond the scope of this paper and represents an interesting future research direction.

[7] In the special case where $C_o = C_u$ (i.e., symmetric costs), this remark implies that the cost-minimizing predicted probability should be the median of distribution $f(.|x)$. This is indeed the correct choice because the median value minimizes the sum of absolute deviations (which is equivalent to minimizing calibration cost in this case).

References

Ayres-de Campos D, Bernardes J, Garrido A, Marques-de Sa J, Pereira-Leite L (2000) Sisporto 2.0: A program for automated analysis of cardiotocograms. *J. Maternal-Fetal Medicine* 9(5):311–318.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E (2014) Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One* 9(10):e109264.

Berardi G, Esuli A, Sebastiani F (2015) Utility-theoretic ranking for semiautomated text classification. *ACM Trans. Knowledge Discovery Data* 10(1):6.

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Sci.* 66(3):1025–1044.

Card D, Smith NA (2018) The importance of calibration for estimating proportions from annotations. Walker M, Ji H, Stent A, eds. *Proc. 2018 Conf. North American Chapter Assoc. Computational Linguistics Human Language Tech.*, vol. 1, Long Papers (Association for Computational Linguistics, New Orleans), 1636–1646.

Dembczynski K, Cheng W, Hüllermeier E (2010) Bayes optimal multilabel classification via probabilistic classifier chains. *Proc. 27th Internat. Conf. Machine Learn.* (Omnipress, Madison, WI), 279–286.

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint, submitted October 11, https://arxiv.org/abs/1810.04805?amp=1.

Domingos P (1999) Metacost: A general method for making classifiers cost-sensitive. *Proc. Fifth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 155–164.

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. Precup D, Teh YW, eds. *Proc. 34th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 70 (PMLR, New York), 1321–1330.

Herńndez-Orallo J (2014) Probabilistic reframing for cost-sensitive regression. *ACM Trans. Knowledge Discovery Data* 8(4):17.

Huber J, Müller S, Fleischmann M, Stuckenschmidt H (2019) A data-driven newsvendor problem: From data to decision. *Eur. J. Oper. Res.* 278(3):904–915.

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. Papadimitriou CH, ed. *8th Innovations Theoret. Comput. Sci. Conf. (ITCS 2017)*, Leibniz

International Proceedings in Informatics (LIPIcs), vol. 67 (Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl, Germany), 43:1–43:23.

Kuleshov V, Fenner N, Ermon S (2018) Accurate uncertainties for deep learning using calibrated regression. Dy J, Krause A, eds. *Proc. 35th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 80 (PMLR, New York), 2796–2804.

Kumar A, Liang PS, Ma T (2019) Verified uncertainty calibration. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates Inc., Red Hook, NY), 3792–3803.

Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. *Proc. 49th Annual Meeting Assoc. Computational Linguistics Human Language Tech.*, HLT '11, vol. 1 (Association for Computational Linguistics, Cambridge, MA), 142–150.

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62:22–31.

Nguyen K, O'Connor B (2015) Posterior calibration and exploratory analysis for natural language processing models. Màrquez L, Callison-Burch C, Su J, eds. *Proc. 2015 Conf. Empirical Methods Natural Language Processing* (Association for Computational Linguistics, Lisbon, Portugal), 1587–1598.

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. *Proc. 22nd Internat. Conf. Machine Learn.* (ACM, New York), 625–632.

Pate A, Van Staa T, Emsley R (2020) An assessment of the potential miscalibration of cardiovascular disease risk predictions caused by a secular trend in cardiovascular disease in England. *BMC Medical Res. Methodology* 20(1):1–12.

Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classifiers* 10(3):61–74.

Provost F (2005) Toward economic machine learning and utility-based data mining. *Proc. 1st Internat. Workshop Utility-Based Data Mining* (Association for Computing Machinery, New York), 1.

Shah ND, Steyerberg EW, Kent DM (2018) Big data and predictive analytics: Recalibrating expectations. *J. Amer. Medical Assoc.* 320(1):27–28.

Tuomo A, Suutala J, Röning J, Koskimäki H (2020) Better classifier calibration for small datasets. *ACM Trans. Knowledge Discovery Data* 14(3):1–19.

Vafeiadis T, Diamantaras KI, Sarigiannidis G, Chatzisavvas KC (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Model. Practice Theory* 55:1–9.

Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 17(1):1–7.

Viaene S, Dedene G (2005) Cost-sensitive learning and decision making revisited. *Eur. J. Oper. Res.* 166(1):212–220.

Voudouri A, Khain P, Carmona I, Bellprat O, Grazzini F, Avgoustoglou E, Bettems J, Kaufmann P (2017) Objective calibration of numerical weather prediction models. *Atmospheric Res.* 190: 128–140.

Watson L, Guo C, Cormode G, Sablayrolles A (2021) On the importance of difficulty calibration in membership inference attacks. Preprint, submitted November 15, https://arxiv.org/abs/2111.08440.

Wolberg WH, Mangasarian OL (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci. USA* 87(23):9193–9196.

Yeh IC, Lien Ch (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems Appl.* 36(2):2473–2480.

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. *Proc. Eighth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 694–699.

Zhao S, Kim M, Sahoo R, Ma T, Ermon S (2021) Calibrating predictions to decisions: A novel approach to multi-class calibration. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Advances in Neural Information Processing Systems*, vol. 34 (Curran Associates Inc., Red Hook, NY), 22313–22324.

Zhu C, Byrd RH, Lu P, Nocedal J (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software* 23(4):550–560.

Zumel N, Mount J (2019) *Practical Data Science with R* (Manning, Shelter Island, NY).