The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Variable-Based Calibration for Machine Learning Classifiers

Markelle Kelly, Padhraic Smyth

University of California, Irvine
kmarke@uci.edu, smyth@ics.uci.edu

Abstract

The deployment of machine learning classifiers in high-stakes domains requires well-calibrated confidence scores for model predictions. In this paper we introduce the notion of variable-based calibration to characterize calibration properties of a model with respect to a variable of interest, generalizing traditional score-based metrics such as expected calibration error (ECE). In particular, we find that models with near-perfect ECE can exhibit significant miscalibration as a function of features of the data. We demonstrate this phenomenon both theoretically and in practice on multiple well-known datasets, and show that it can persist after the application of existing calibration methods. To mitigate this issue, we propose strategies for detection, visualization, and quantification of variable-based calibration error. We then examine the limitations of current score-based calibration methods and explore potential modifications. Finally, we discuss the implications of these findings, emphasizing that an understanding of calibration beyond simple aggregate measures is crucial for endeavors such as fairness and model interpretability.

1 Introduction

Predictive models built by machine learning algorithms are increasingly informing decisions across high-stakes applications such as medicine (Rajkomar, Dean, and Kohane 2019), employment (Chalfin et al. 2016), and criminal justice (Završnik 2021). There is also broad recent interest in developing systems where humans and machine learning models collaborate to make predictions and decisions (Kleinberg et al. 2018; Bansal et al. 2021; De et al. 2021; Steyvers et al. 2022). A critical aspect of using model predictions in such contexts is calibration. In particular, in order to trust the predictions from a machine learning classifier, these predictions must be accompanied by well-calibrated confidence scores.

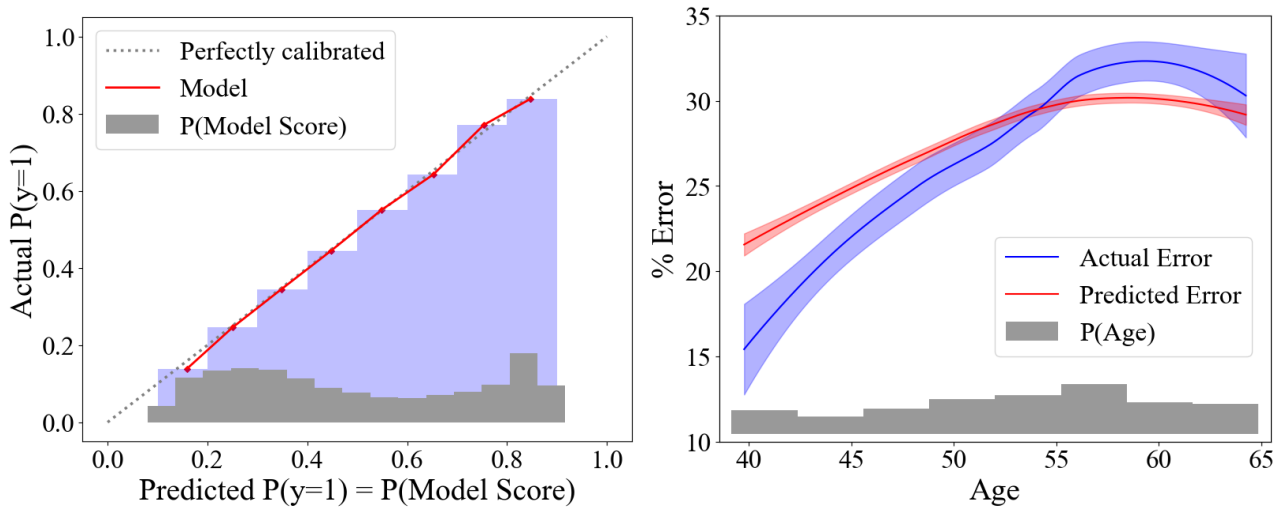
In practice, however, it has been well-documented that machine learning classifiers such as deep neural networks can produce poorly-calibrated class probabilities (Guo et al. 2017; Vaicenavicius et al. 2019; Ovadia et al. 2019). As a result, a variety of calibration methods have been developed, which aim to ensure that a model's confidence (or score) matches its true accuracy. A widely used approach is post-hoc calibration: methods which use a separate labeled dataset to learn a mapping from the original model's class probabilities to calibrated probabilities, often with a relatively simple one-dimensional mapping (e.g., Platt (1999); Kull, Filho, and Flach (2017); Kumar, Liang, and Ma (2019)). These methods have been shown to generally improve the the empirical calibration error of a model, as commonly measured by the expected calibration error (ECE).

However, as we show in this paper, aggregate measures of score-based calibration error such as ECE can hide significant systematic miscalibration in other dimensions of a model's performance. To address this issue we introduce the notion of *variable-based calibration* to better understand how the calibration error of a model can vary as a function of a variable of interest, such as an input variable to the model or some other metadata variable. We focus in particular in this paper on real-valued variables. For example, in prediction problems involving individuals (e.g., credit-scoring or medical diagnosis) one such variable could be *Age*. Detecting systematic miscalibration is important for problems such as assessing the fairness of a model, for instance detecting that a model is significantly overconfident for some age ranges and underconfident for others.

As an illustrative example, consider a simple classifier trained to predict the presence of cardiovascular disease[1]. After the application of Platt scaling, a standard post-hoc calibration method, this model attains a relatively low ECE of 0.74%. This low ECE is reflected in the reliability diagram shown in Figure 1a, which shows near-perfect alignment with the diagonal. If a user of this model were to only consider aggregate metrics such as ECE, they might reasonably conclude that the model is generally well-calibrated. However, evaluating model error and predicted error with respect to the variable *Patient Age* reveals an undesirable and systematic miscalibration pattern with respect to this variable, as illustrated in Figure 1b: the model is underconfident by upwards of five percentage points for younger patients, and is significantly overconfident for older patients.

In this paper, we systematically investigate variable-based calibration for classification models, from both theoretical and empirical perspectives. In particular, our contributions are as follows:

[1] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset



(a) Reliability diagram (for accuracy)

(b) Variable-based calibration plot (for error)

Figure 1: Calibration plots for a neural network predicting cardiovascular disease, after calibration with Platt scaling: (a) reliability diagram, (b) LOESS-smoothed estimates with confidence intervals of actual and model-predicted error as a function of patient age. This dataset consists of 70,000 records of patient data (49,000 train, 6,000 validation, 15,000 test), with a binary prediction task of determining the presence of cardiovascular disease.

1. We introduce the notion of *variable-based calibration* and define a per-variable calibration metric (VECE).
2. We characterize theoretically the relationship between variable-based miscalibration measured via VECE and traditional score-based miscalibration measured via ECE.
3. We demonstrate, across multiple well-known tabular, text, and image datasets and a variety of models, that significant variable-based miscalibration can exist in practice, even after the application of standard score-based calibration methods.
4. We investigate *variable-based calibration methods* and demonstrate empirically that these methods can simultaneously reduce both ECE and VECE. [2]

2 Related Work

Visualizing Model Performance by Variable: In prior work a number of different techniques have been developed for visual understanding and diagnosis of model performance with respect to a particular variable of interest. One such technique is partial dependence plots (Friedman 2001; Molnar 2020), which visualize the effect of an input feature of interest on model predictions. Another approach is dashboards such as FairVis (Cabrera et al. 2019) which enable the exploration of model performance (e.g., accuracy, false positive rate) across various data subgroups. However, none of this prior work investigates the visualization of per-variable calibration properties of a model, i.e., how a model's own predictions of accuracy (or error) vary as a function of a particular variable.

[2]Our code is available online at https://github.com/markellekelly/variable-wise-calibration.

Quantifying Model Calibration by Variable: Work on calibration for machine learning classifiers has largely focused on score-based calibration: reliability diagrams, the ECE, and standard calibration methods are all defined with respect to confidence scores (Murphy and Winkler 1977; Huang et al. 2020; Song et al. 2021). An exception to this is in the fairness literature, where researchers have broadly called for disaggregated model evaluation, e.g. computing metrics of interest individually for sensitive sub-populations (Mitchell et al. 2019; Raji et al. 2020). To this end, several notions of calibration that move beyond standard aggregate measures have been introduced: Hébert-Johnson et al. (2018) check calibration across all identifiable subpopulations of the data, Pan et al. (2020) evaluate calibration over data subsets corresponding to a categorical variable of interest, and Luo et al. (2022) compute "local calibration" using the average classification error on similar samples. Our paper expands on this prior work in two ways. First, we shift the focus from categorical to real-valued variables—our methods operate on a continuous basis, estimating calibration for an entire population rather than for various subgroups. Second, we center on diagnosing calibration; we present visualization and estimation techniques for understanding an existing classifier rather than prescriptive conditions for model training or selection.

3 Background on Score-Based ECE

Consider a classification problem mapping inputs x to predictions for labels $y \in \{1, \ldots, K\}$. Let f be a black-box classifier which outputs label probabilities $f(x) \in [0,1]^K$ for each $x \in X$. Then, for the standard 0-1 loss function, the predicted label is $\hat{y} = \text{argmax}(f(x)) \in \{1, \ldots, K\}$ and the corresponding confidence score is $s = s(x) = P_f(y =$

$\hat{y}|x) = \max(f(x))$. It is of interest to determine whether such a model is *well-calibrated*, that is, whether its confidence matches the true probability that a prediction is correct.

For a given confidence score s, we define $\text{Acc}(s) = P(y = \hat{y}|s) = \mathbb{E}\left[\mathbb{I}[y = \hat{y}|s]\right]$. Then the ℓ_p calibration error (CE), as a function of the confidence score s, is defined as the difference between accuracy and confidence score (Kumar, Liang, and Ma 2019):

$$\text{CE}(s) = |P(y = \hat{y}|s) - s|^p = |\text{Acc}(s) - s|^p \tag{1}$$

where $p \geq 1$. In this paper, we will focus on the expectation of the ℓ_1 calibration error with $p = 1$, known as the ECE:

$$\text{ECE} = \mathbb{E}[\text{CE}(s)] = \int_s P(s)|\text{Acc}(s) - s|ds \tag{2}$$

where an ECE of zero corresponds to perfect calibration. In practice, ECE is often estimated empirically on a labeled test dataset by creating B bins over s according to some binning scheme (e.g., Guo et al. (2017)):

$$\widehat{\text{ECE}} = \sum_{b=1}^{B} \frac{n_b}{n} |\text{Acc}_b - \text{Conf}_b| \tag{3}$$

where n_b is the number of datapoints in bin b, n is the total number of datapoints, and Acc_b and Conf_b are the estimated accuracy and estimated average value of confidence, respectively, in bin $b = 1, \ldots, B$.

4 Variable-Based Calibration Error

In many applications, we may be motivated to understand the calibration properties of a classification model f relative to one or more particular variables of interest. For instance, traditional reliability diagrams and the ECE measure may be insufficient to fully characterize the type of variable-based miscalibration shown in Figure 1.

Consider a real-valued variable V taking values v. V could be a variable related to the inputs X of the model, such as one of the input features, another feature (e.g., metadata) defined per instance but not used in the model, or some function of inputs x. To evaluate model calibration with respect to V, we introduce the notion of *variable-based calibration error* (VCE), defined pointwise as a function of v:

$$\text{VCE}(v) = \left|\text{Acc}(v) - \mathbb{E}[s|v]\right| \tag{4}$$

where $\text{Acc}(v) = P(y = \hat{y}|v)$ is the accuracy of the model conditioned on $V = v$, marginalizing over inputs to the model that do not involve V. $\mathbb{E}[s|v]$ is the expected model score conditioned on a particular value v:

$$\mathbb{E}[s|v] = \int_s s \cdot P(s|v)ds \tag{5}$$

In general, conditioning on v will induce a distribution over inputs x, which in turn induces a distribution $P(s|v)$ over scores s and predictions $\hat{y}$. As an example of $\text{VCE}(v)$, in the context of Figure 1b, at $v = 45$, the model accuracy $P(y = \hat{y}|v)$ is estimated to be $100 - 21 = 79\%$ and the expected score $\mathbb{E}[s|v]$ is estimated to be 76%, so the $\text{VCE}(v)$ is approximately 3%.

The expected value of $\text{VCE}(v)$, with respect to V, is defined as:

$$\text{VECE} = \mathbb{E}[\text{VCE}(v)] = \int_v P(v)\text{VCE}(v)dv \tag{6}$$

Comment Note that CE (and ECE) can be seen as a special case of VCE (and VECE) given the correspondence of Equations 1 and 2 with Equations 4 and 6 when V is the model score (i.e., $V = s$). In the rest of the paper, however, we view CE and ECE as being distinct from VCE and VECE in order to highlight the differences between score-based and variable-based calibration.

As with ECE, a practical way to compute an empirical estimate of VECE is by binning, where bins b are defined by some binning scheme (e.g., equal weight) over values v of the variable V (rather than over scores s):

$$\widehat{\text{VECE}} = \sum_{b=1}^{B} \frac{n_b}{n} |\text{Acc}_b - \text{Conf}_b|. \tag{7}$$

Here b is a bin corresponding to some sub-range of V, n_b is the number of points within this bin, and Acc_b and Conf_b are empirical estimates of the model's accuracy and the model's average confidence within bin b. For example, the $\widehat{\text{ECE}}$ in Figure 1 is 0.74%, while the $\widehat{\text{VECE}}$ is 2.04%.

The definitions of $\text{VCE}(v)$ and VECE above are in terms of a continuous variable V, which is our primary focus in this paper. In general, the definitions above and the theoretical results in Section 5 also apply to discrete-valued V, as well as to multivariate V.

5 Theoretical Results

In this section, we establish a number of results on the relationship between ECE and VECE. All proofs can be found in Appendix A (Kelly and Smyth 2022)[3].

First, we show that the ECE and VECE can differ by a gap of up to 50%.

Theorem 5.1 (VECE bound)**.** *There exist K-ary classifiers f and variables V such that the classifier f has both ECE = 0 and variable-based VECE $= 0.5 - \frac{1}{2K}$.*

For example, in the binary case with $K = 2$, the difference between ECE and VECE can be as large as 0.25. As the number of classes K grows, this gap approaches 0.5. Thus, we can have models f that are perfectly calibrated according to ECE (i.e. with ECE = 0) but that can have VECE ranging from 0.25 to 0.5. We will show later in Section 7 that this type of gap is not just a theoretical artifact but also exists in real-world datasets, for real-world classifiers f and for specific variables V of interest. The proof of Theorem 5.1 is by construction, using a model f that is very underconfident for certain regions of v and very overconfident in other regions of v, but perfectly calibrated with respect to s.

In the context of analyzing properties of ECE, Kumar, Liang, and Ma (2019) proved that the binned empirical estimator $\widehat{\text{ECE}}$ consistently underestimates the true ECE, and showed by construction that this gap can approach 0.5. Our results complement this work in that we are concerned with the true theoretical relationship between two different measures of calibration, namely ECE and VECE, whereas Kumar,

[3]The Appendix referenced throughout this paper can be found in the cited arXiv version of this paper.

Liang, and Ma (2019) relate the estimate $\widehat{\text{ECE}}$ (Equation 3) with the true ECE (Equation 2).

Theorem 5.2 (ECE bound)**.** *There exist K-ary classifiers f and variables V such that the classifier f has VECE $= 0$ and ECE $= 0.5 - \frac{1}{2K}$.*

We prove this by construction, where f is well-calibrated with respect to a variable V, but its low scores are very underconfident and its high scores are very overconfident.

The results above illustrate that the ECE and VECE measures can be very different for the same model f. In our experimental results we will also show that it is not uncommon (particularly for uncalibrated models) for ECE and VECE to be equal. To understand the case of equality, we first define the notion of *consistent over- or under-confidence* with respect to a variable:

Definition 5.3 (Consistent overconfidence)**.** Let f be a classifier with scores s. For a variable V taking values v, f is *consistently overconfident* if $\mathbb{E}[s|v] > P(y = \hat{y}|v), \forall v$, i.e., the expected value of the model's scores f as a function of v is always greater than the true accuracy as a function of v.

Consistent underconfidence can be defined analogously, using $\mathbb{E}[s|v] < P(y = \hat{y}|v), \forall v$. In the special case where the variable V is defined as the score itself, we have the condition $s > P(y = \hat{y}|s), \forall s$, leading to consistent overconfidence for the scores.

For the case of consistent over- or under- confidence for a model f, we have the following result:

Theorem 5.4 (Equality conditions of ECE and VECE)**.** *Let f be a classifier that is consistently under- or over- confident with respect both to s and to a variable V. Then the ECE and VECE of f are equal.*

The results above provide insight into the relationship between ECE and VECE. Specifically, if the miscalibration is "one-sided" (i.e., consistently over- or under-confident for both the score s and a variable V) then ECE and VECE will be in agreement. However, when the classifier f is both over- and under-confident (as a function of either s or v), then ECE and VECE can differ significantly and, as a result, ECE can mask significant systematic miscalibration with respect to variables of interest.

6 Mitigating Variable-Based Miscalibration

Diagnosis of Variable-Based Miscalibration

In order to better detect and characterize per-variable miscalibration, we discuss below *variable-based calibration plots*, which we have found useful in practice. Figure 1b shows an example of a variable-based calibration plot for age. In Section 7, we explore how these plots can be used to characterize miscalibration across different classifiers, datasets, and variables of interest.

For ease of interpretation in the results below we focus on the model's error rate and predicted error, rather than accuracy and confidence, although they are equivalent. Particularly for models with high accuracy, we find that it is more intuitive to discuss differences in error rate than in accuracy.

To generate these types of plots, we first compute the individual error $\mathbb{I}[y \neq \hat{y}]$ and predicted error $1 - s(x) = 1 - \max(f(x))$ for each observation. We then construct non-parametric error curves with LOESS. (Further details are available in Appendix B.) This approach allows us to obtain 95% confidence bars for the error rate and mean predicted error, based on standard error, thus putting the differences in curves into perspective.

Beyond visualization, we can use VECE scores to discover which variables for a dataset have the largest systematic variable-based miscalibration. In particular, ranking features in order of decreasing VECE highlights variables that may be worth investigating. An example of such a ranking for the Adult Income dataset[4], based on a neural network with post-hoc beta calibration (Kull, Filho, and Flach 2017), is shown in Table 1. The *Years of education* and *Age* variables rank highest in VECE, so a model developer or a user of a model might find it useful to generate a variable-based calibration plot for each of these. The *Weekly work hours* and *Census weight* variables are of lesser concern, but could also be explored. We will perform an in-depth investigation of miscalibration with respect to the variable *Age* in Section 7.

	VECE	**VCE(v^*)**
Years of education	9.95%	20.13%
Age	9.59%	23.44%
Weekly work hours	7.94%	18.21%
Census weight	5.06%	12.08%

Table 1: Variable-based calibration error of Adult Income dataset features

It is also possible to define the maximum value of VCE(v), i.e, the *worst-case calibration error*, as well as the value v^* that incurs this worst-case error:

$$\begin{aligned} v^* &= \arg\max_v \{\text{VCE}(v)\} \\ &= \arg\max_v \{\big|P(y = \hat{y}|v) - \mathbb{E}[s(v)]\big|\} \end{aligned} \tag{8}$$

Estimating either v^* or VCE(v^*) accurately may be difficult in practice, particularly for small sample sizes n, since it involves the non-parametric estimation of the difference of two curves (Bowman and Young 1996) as a function of v (as the shapes of the curves need not follow any convenient parametric form, e.g., see Figure 1b). One simple estimation strategy is to smooth both curves with LOESS and compute the maximum difference between the two estimated curves. Using this approach, worst-case calibration errors VCE(v^*) for the Adult Income model are also shown in Table 1.

Calibration Methods

We found empirically, across multiple datasets, that standard score-based calibration methods often reduce ECE while neglecting variable-based systematic miscalibration. Because calibration error can vary as a function of a feature of interest V, we propose incorporating information about V into post-hoc calibration. In particular, we introduce the concept of *variable-based calibration methods*, a family of calibration

[4]https://archive.ics.uci.edu/ml/datasets/adult

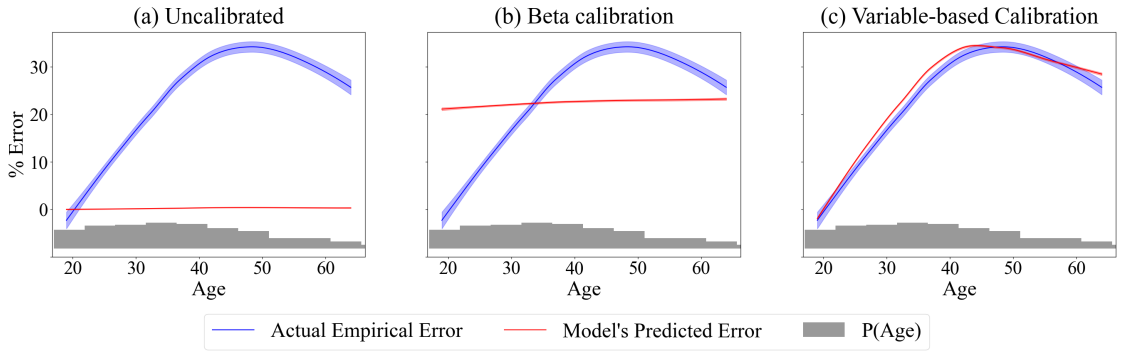


Figure 2: Variable-based calibration plots for *Age* for the Adult Income model

methods that adjust confidence scores with respect to some variable of interest V. As an illustrative example, we perform experiments in Section 7 with a modification of probability calibration trees (Leathart et al. 2017). This technique involves performing logistic calibration separately for data splits defined by decision trees trained over the input space. We alter the method to train decision trees for y with only v as input, with a minimum leaf size of one-tenth of the total calibration set size. We then perform beta calibration at each leaf (Kull, Filho, and Flach 2017), as we found in our experiments that it performs empirically better than logistic calibration. In the multi-class case, we use Dirichlet calibration, an extension of beta calibration for k-class classification (Kull et al. 2019). Our split-based calibration method using decision trees is intended to provide a straightforward illustration of the potential benefits of variable-based calibration, rather than a state-of-the-art methodology that can balance ECE and VECE (which we leave to future work). We also investigated variable-based calibration methods that operate continuously over V (rather than on separate data splits) using extensions of logistic and beta calibration, but found that these were not as reliable in our experiments as the tree-based approach (see Appendix C for details).

7 Variable-Based Miscalibration in Practice

In this section, we explore several examples where the ECE obscures systematic miscalibration relative to some variable of interest, particularly after post-hoc score-based calibration. In our experiments we use four datasets that span tabular, text, and image data. For each dataset and variable of interest V, we investigate both (1) several score-based calibration methods and (2) our variable-based calibration method (the tree-based technique described in Section 6), comparing the resulting ECE, VECE, and variable-based calibration plots. In particular, we calibrate with scaling-binning (Kumar, Liang, and Ma 2019), Platt scaling (Platt 1999), beta calibration (Kull, Filho, and Flach 2017), and, for the multi-class case, Dirichlet calibration (Kull et al. 2019). The datasets are split into training, calibration, and test sets. Each calibration method is trained on the same calibration set, and all metrics and figures are produced from the final test set. The ECE and VECE are computed with an equal-support binning scheme, with $B = 10$. Further details regarding datasets, models, and calibration methods are in Appendix B.

Adult Census Records: Predicting Income

The Adult Income dataset consists of 1994 Census records; the goal is to predict whether an individual's annual income is greater than $50,000. We model this data with a simple feed-forward neural network and evaluate the model's calibration error with respect to age (i.e. let V=age). Uncalibrated, this model has an ECE and VECE of 20.67% (see Table 2). The ECE and VECE are equal precisely because of the model's consistent overconfidence as a function of both the confidence score and V (see Definition 5.3). This overconfidence with respect to age is reflected in the variable-based calibration plot (Figure 2a). The model's error rate varies significantly as a function of age, with very high error for individuals around age 50, and much lower error for younger and older people. However, its confidence remains nearly constant at close to 100% (i.e., a predicted error close to 0%) across all ages.

	ECE	**VECE**
Uncalibrated	20.67%	20.67%
Scaling-binning	2.27%	9.25%
Platt scaling	4.57%	10.13%
Beta calibration	1.65%	9.59%
Variable-based calibration	**1.64%**	**2.11%**

Table 2: Adult Income model calibration error

After calibrating, the ECE is dramatically reduced, with beta calibration achieving an ECE of 1.65%. However, the corresponding VECE is still very high (over 9%). As shown in Figure 2b, the model's self-predicted error has increased substantially, but remains near constant as a function of age. Thus, despite a significant improvement in ECE, the model still harbors unfairness with respect to age, exhibiting overconfidence in its predictions for individuals in the 35-65 age range, and underconfidence for those outside of it. As the model is no longer consistently overconfident, the ECE and VECE diverge, as predicted theoretically.

Variable-based calibration obtains a significantly lower VECE of 2.11%, while simultaneously reducing the ECE. This improvement in VECE is reflected in Figure 2c. The model's predicted error now varies with age to match the true

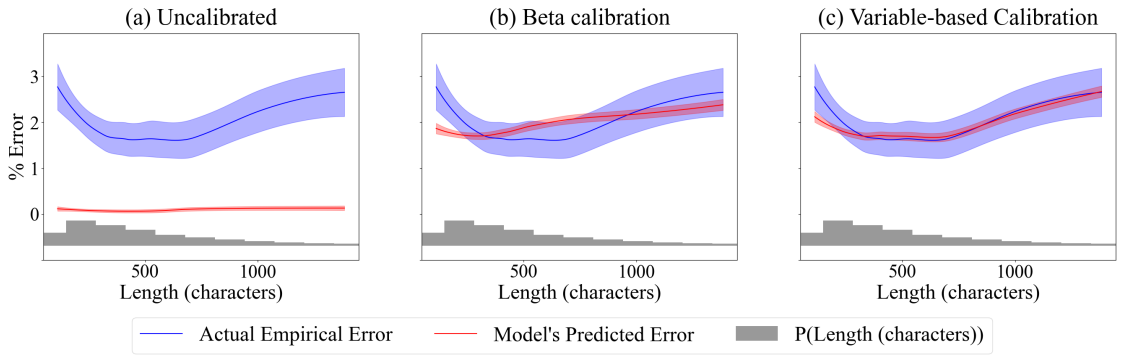


Figure 3: Variable-based calibration plots for the Yelp model for *Review Length*

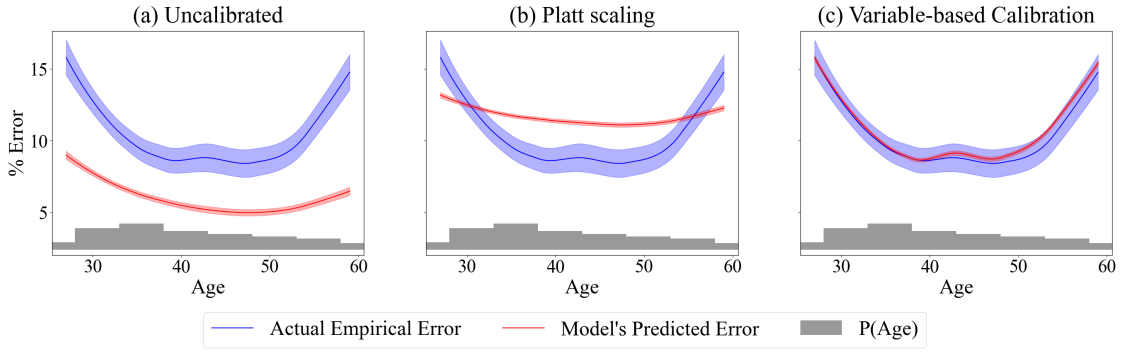


Figure 4: Variable-based calibration plots for the Bank Marketing model for *Age*

error rate. In this case, a simple variable-based calibration method improves the age-wise systematic miscalibration of the model, without detriment to the overall calibration error.

Yelp Reviews: Predicting Sentiment

To explore variable-based calibration in an NLP context, we use a fine-tuned large language model, BERT (Kenton and Toutanova 2019), on the Yelp review dataset[5]. The model predicts whether a review has a positive or negative rating based on its text. In this case there are no easily-interpretable features directly input to the model. Instead, to better diagnose model behavior, we can analyze real-valued characteristics of the text, such as the length of each review or part-of-speech statistics. Here we focus on review length in characters.

Figure 3 shows the model's error and predicted error with respect to review length. The error rate is lowest for reviews around 300-700 characters, around the median review length. Very short and very long reviews are associated with a higher error rate. Uncalibrated, this model is consistently overconfident, with an ECE and VECE of 1.93% (see Table 3).

After beta calibration, the ECE and VECE drop to 1.73% and 0.37%, respectively. Figure 3b reflects this: the model's predicted error aligns more closely with its actual error rate, although it is still overconfident for very short reviews.

Our variable-based calibration method further reduces the VECE and yields a small improvement to the ECE. The new predicted error curve matches the true relationship between review length and error rate more faithfully (Figure 3c), reducing overconfidence for short reviews.

	ECE	VECE
Uncalibrated	1.93%	1.93%
Scaling-binning	4.23%	4.23%
Platt scaling	3.04%	0.64%
Beta calibration	1.73%	0.37%
Variable-based calibration	**1.70%**	**0.23%**

Table 3: Yelp model calibration error

Bank Marketing: Predicting Subscriptions

We also investigate miscalibration on a simple neural network modeling the Bank Marketing dataset[6]. The model predicts whether a bank customer will subscribe to a bank term deposit as a result of direct marketing. Uncalibrated, the model is overconfident, with both ECE and VECE over 4.5% (see Table 4). Consider the calibration error with respect to customer age before (Figure 4a) and after (Figure 4b) Platt scaling. Platt scaling, which is the best-performing score-based calibration method, uniformly increases the predicted error across age, reducing both ECE and VECE, but resulting in underconfidence for most ages and overconfidence at the

[5]https://www.yelp.com/dataset

[6]https://archive.ics.uci.edu/ml/datasets/bank+marketing

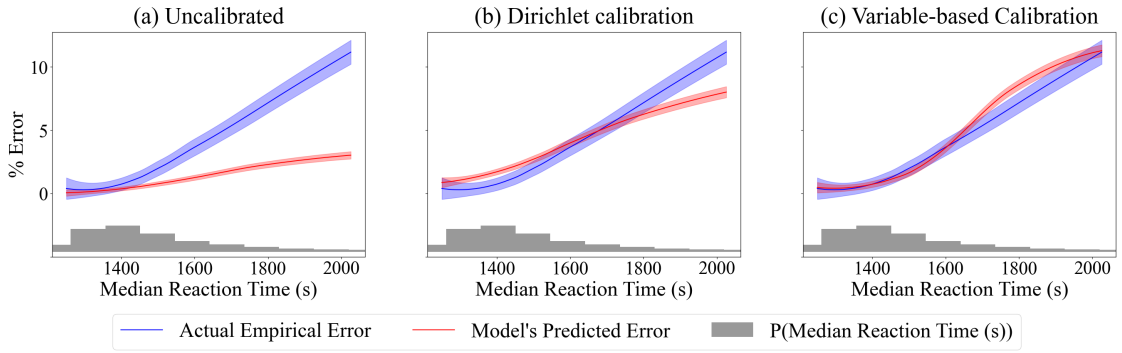


Figure 5: Variable-based calibration plots for the CIFAR-10H model for *Median Reaction Time*

edges of the distribution.

	ECE	VECE
Uncalibrated	4.69%	4.69%
Scaling-binning	4.37%	3.39%
Platt scaling	2.38%	2.83%
Beta calibration	2.48%	2.77%
Variable-based calibration	**2.10%**	**0.52%**

Table 4: Bank Marketing model calibration error

The variable-based calibration method achieves competitive ECE, while reducing VECE to about half of one percent. The calibration plot reflects this improvement: the predicted error matches the true error rate more closely, reducing the miscalibration with respect to customer age.

CIFAR-10H: Image Classification

As a multi-class example, we investigate variable-based miscalibration on CIFAR-10H, a 10-class image dataset including labels and reaction times from human annotators (Peterson et al. 2019). We use a standard deep learning image classification architecture (a DenseNet model) to predict the image category, and investigate median annotator reaction times, metadata that are not provided to the model. Instead of Platt scaling and beta calibration, here we use Dirichlet calibration (to accomodate the multiple classes).

In this case, Dirichlet calibration achieves the lowest overall ECE and variable-based calibration obtains the lowest VECE (see Table 5). The variable-based calibration plots are shown in Figure 5. We see that the variable-based calibration method reduces underconfidence for examples with low median reaction times (where the majority of data points lie).

	ECE	VECE
Uncalibrated	1.90%	1.92%
Scaling-binning	3.83%	3.60%
Dirichlet calibration	**0.80%**	1.12%
Variable-based calibration	1.18%	**0.86%**

Table 5: CIFAR-10H model calibration error

Summary of Experimental Results Our results demonstrate the potential of variable-based calibration. While score-based calibration methods generally improved the ECE, variable-based calibration methods performed better across datasets in terms of simultaneously reducing both the ECE and VECE, without any significant increase in model error rate or the VECE for other variables (details in Appendix B). The results also illustrate that variable-based calibration plots enable meaningful characterization of the relationships between variables of interest and predicted/true error, providing more detailed insight into a model's performance than a single number (i.e., ECE or VECE).

8 Discussion and Conclusions

Discussion of Limitations There are several potential limitations of this work. First, we focused on the mitigation of miscalibration for one variable V at a time. Although we did not observe higher VECE for other variables after applying our variable-based calibration method, this behavior has not been analyzed theoretically. Further, a more thorough investigation of miscalibration across intersections of variables is still needed. We also emphasize that the variable-based calibration method used in the paper is primarily for illustration; the development of new methods for simultaneously reducing score-based and variable-based miscalibration is a useful direction for future work.

Conclusions In this paper we demonstrated theoretically and empirically that ECE can obscure significant miscalibration with respect to variables of potential importance to a developer or user of a classification model. To better detect and characterize this type of miscalibration, we introduced the VECE measure and corresponding variable-based calibration plots, and we characterized the theoretical relationship between VECE and ECE. In a case study across several datasets and models, we showed that VECE, variable-based calibration plots, and variable-based calibration methods are all useful tools for understanding and mitigating miscalibration on a per-variable level. Looking forward, to mitigate biases in calibration error, we recommend moving beyond purely score-based calibration analysis. In addition to promoting fairness, these techniques offer new insight into model behavior and provide actionable avenues for improvement.

Acknowledgements

This material is based upon work supported in part by the HPI Research Center in Machine Learning and Data Science at UC Irvine, by the National Science Foundation under grants number 1900644 and 1927245, and by a Qualcomm Faculty Award.

References

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11405–11414.

Bowman, A.; and Young, S. 1996. Graphical Comparison of Nonparametric Curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(1): 83–98.

Cabrera, Á. A.; Epperson, W.; Hohman, F.; Kahng, M.; Morgenstern, J.; and Chau, D. H. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 46–56.

Chalfin, A.; Danieli, O.; Hillis, A.; Jelveh, Z.; Luca, M.; Ludwig, J.; and Mullainathan, S. 2016. Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*, 106(5): 124–27.

De, A.; Okati, N.; Zarezade, A.; and Rodriguez, M. G. 2021. Classification under Human Assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5905–5913.

Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

Hébert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *International Conference on Machine Learning*, 1939–1948.

Huang, Y.; Li, W.; Macheret, F.; Gabriel, R. A.; and Ohno-Machado, L. 2020. A Tutorial on Calibration Measurements and Calibration Models for Clinical Prediction Models. *Journal of the American Medical Informatics Association*, 27(4): 621–633.

Kelly, M.; and Smyth, P. 2022. Variable-Based Calibration for Machine Learning Classifiers. *arXiv preprint arXiv:2209.15154*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1): 237–293.

Kull, M.; Filho, T. S.; and Flach, P. 2017. Beta Calibration: A Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, 623–631.

Kull, M.; Perello-Nieto, M.; Kängsepp, M.; Filho, T. S.; Song, H.; and Flach, P. 2019. Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 12316–12326.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems*, 3787–3798.

Leathart, T.; Frank, E.; Holmes, G.; and Pfahringer, B. 2017. Probability Calibration Trees. In *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77, 145–160.

Luo, R.; Bhatnagar, A.; Wang, H.; Xiong, C.; Savarese, S.; Bai, Y.; Zhao, S.; and Ermon, S. 2022. Localized Calibration: Metrics and Recalibration. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, 1286–1295.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Molnar, C. 2020. *Interpretable Machine Learning*. Lulu.com.

Murphy, A. H.; and Winkler, R. L. 1977. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1): 41–47.

Ovadia, Y.; Fertig, E.; Lakshminarayanan, B.; Nowozin, S.; Sculley, D.; Dillon, J.; Ren, J.; Nado, Z.; and Snoek, J. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 14003–14014.

Pan, F.; Ao, X.; Tang, P.; Lu, M.; Liu, D.; Xiao, L.; and He, Q. 2020. Field-aware Calibration: A Simple and Empirically Strong Method for Reliable Probabilistic Predictions. In *Proceedings of The Web Conference: WWW 2020*, 729–739.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human Uncertainty Makes Classification More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Platt, J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 61–74.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.

Rajkomar, A.; Dean, J.; and Kohane, I. 2019. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14): 1347–1358.

Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; Flach, P.; et al. 2021. Classifier Calibration: How to assess and improve predicted class probabilities: a survey. *arXiv preprint arXiv:2112.10327*.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian Modeling of Human-AI Complementarity. *Proceedings of the National Academy of Sciences*, 119(11).

Vaicenavicius, J.; Widmann, D.; Andersson, C.; Lindsten, F.; Roll, J.; and Schön, T. 2019. Evaluating Model Calibration in Classification. In *International Conference on Artificial Intelligence and Statistics*, 3459–3467.

Završnik, A. 2021. Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings. *European Journal of Criminology*, 18(5): 623–642.