



Challenging Calibration of AI Scoring Algorithms

Agathe Fernandes Machado, Arthur Charpentier, Emmanuel Flachaire, Ewen Gallic, François Hu

Université du Québec à Montréal, <https://fer-agathe.github.io>

“The phrase ‘probability of death’, when it refers to a single person, has no meaning for us at all.” (Von Mises et al., 1939)

Motivation: Underlying risk of weather forecasts

“Weather forecasters routinely make predictions such as ‘the precipitation probability for Denver today is 30 percent’. The probabilities quoted refer to the forecasters’ subjective ‘degree of belief’ given their information at the time of the forecast (...) it is rarely appropriate to interpret a subjective probability forecast as an estimate of some underlying ‘objective’ probability ; it is usually better considered as an estimate of the forecast event itself. A forecaster is well calibrated if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent (...) we concentrate exclusively on the criterion of calibration (sometimes termed reliability)” (Dawid, 1982).

Likelihood of a loan default

“investors assess the borrowers’ credit risk,”

Here, the focus extends beyond the predicted class to the associated likelihood of loan repayment defaults. The scores generated by binary classifiers are often viewed as event probabilities. However, to ensure these scores are accurately interpreted as probabilities, the model must be properly calibrated.

“In recent years, machine learning algorithms have been frequently applied to predict borrower default probability” (Liu et al., 2021).

Probability of accident

In a motor insurance contract, the probability for an insured to have an accident within the next year can be used to estimate its premium, as risk transfer pricing is usually tied directly to event probabilities.

“This appendix identifies the information a state insurance regulator may need to review a Tree-based predictive model used by an insurer to support a personal automobile or home insurance rating plan. Tree-based predictive models include Random Forest (RF) and Gradient Boosting Machines (GBM).” (NAIC, 2022).

Mathematical formalism

Consider a binary variable Y that takes the value 1 if an event occurs and 0 otherwise. In this context, the probability of the event depends on individual characteristics, i.e., $p_i = s(\mathbf{x}_i)$, where, with sample size $n > 0$, $i = 1, \dots, n$ represents individuals, and \mathbf{x}_i the characteristics. The goal is to estimate this probability by $\hat{y}_i = \hat{s}(\mathbf{x}_i) \in [0, 1]$ using a Machine Learning (ML) model. A binary classifier is said to be well-calibrated when

$$\mathbb{P}(Y = 1 \mid \hat{s}(\mathbf{x}) = p) = p, \quad \forall p \in [0, 1],$$

which is equivalent to:

$$\mathbb{E}[Y \mid \hat{s}(\mathbf{x}) = p] = p, \quad \forall p \in [0, 1].$$

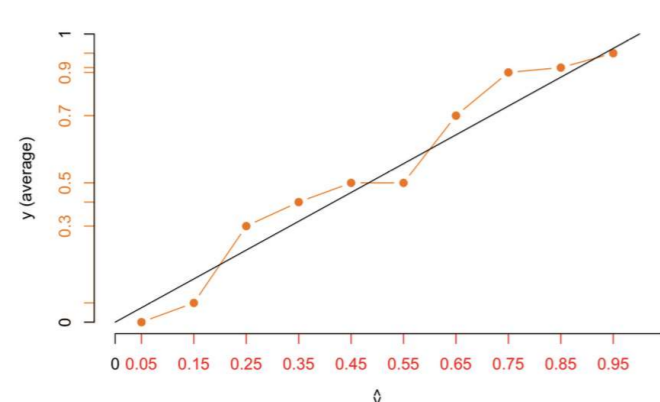
Visualization tools

Calibration curve of a binary classifier,

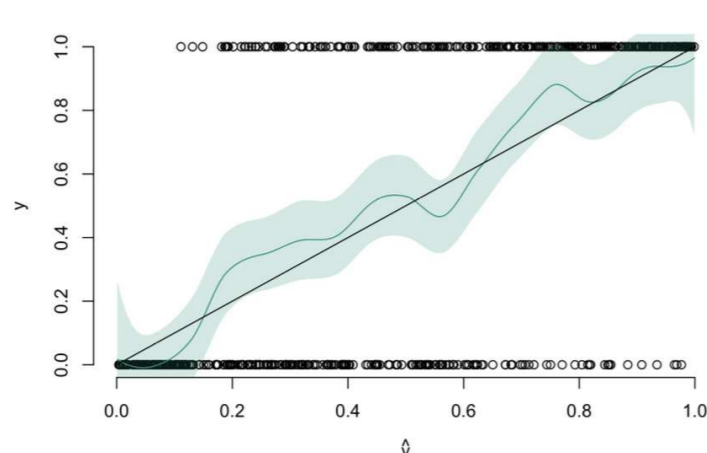
$$g : \begin{cases} [0, 1] \rightarrow [0, 1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \hat{s}(\mathbf{X}) = p] \end{cases} \quad (1)$$

The g function for a well-calibrated model is the identity function $g(p) = p$.

Reliability diagram: Based on bins defined by quantiles of $\hat{s}(\mathbf{X})$.



Local regression: Based on nearest neighbors.



Calibration and Machine Learning

Simple classifiers such as Logistic Regression models typically exhibit overall calibration (Mildenhall, 1999) due to their design in the empirical risk minimization problem, and also exhibit local calibration when the model is well-specified (Machado et al., 2024). When using more opaque models, such as Random Forest (RF) or Neural Networks (NN), the interpretability of calibration becomes more nuanced, with differing views on their potential (mis)calibration.

“Other models such as neural nets and bagged trees do not have these biases and predict well calibrated probabilities” (Niculescu-Mizil and Caruana, 2005),

“We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated” (Guo et al., 2017),

“The probabilistic predictions of RFs are known to be usually well calibrated” (Hänsch, 2020),

“Several machine learning approaches such as Naive Bayes, decision trees, and artificial neural networks have been shown to have exhibit poor calibration” Park and Ho (2020).

Calibration metrics

The Brier score is expressed as

$$BS = \frac{1}{n} \sum_{i=1}^n (\hat{s}(\mathbf{x}_i) - y_i)^2,$$

The Expected Calibration Error is determined by using quantile-binning on predicted scores $\hat{s}(\mathbf{x})$ where B denotes the chosen number of bins, $acc(b)$ the accuracy in bin b and $conf(b)$ the model’s average confidence within bin b

$$ECE = \sum_{b=1}^B \frac{n_b}{n} |acc(b) - conf(b)|$$

The Integrated Calibration Index is based on the calibration curve defined with local regression techniques and is defined as

$$ICI = \int_0^1 f(p)\phi(p) dp$$

where $f(p) = |p - g(p)|$ is the absolute difference between the calibration curve and the bisector where p denotes a predicted score. The density function of the distribution of predicted scores is denoted $\phi(p)$.

Calibration and Fairness

Here, we consider calibration within two groups of individuals based on a sensitive attribute S (Baumann, 2023): $\{S = A\}$ or $\{S = B\}$.

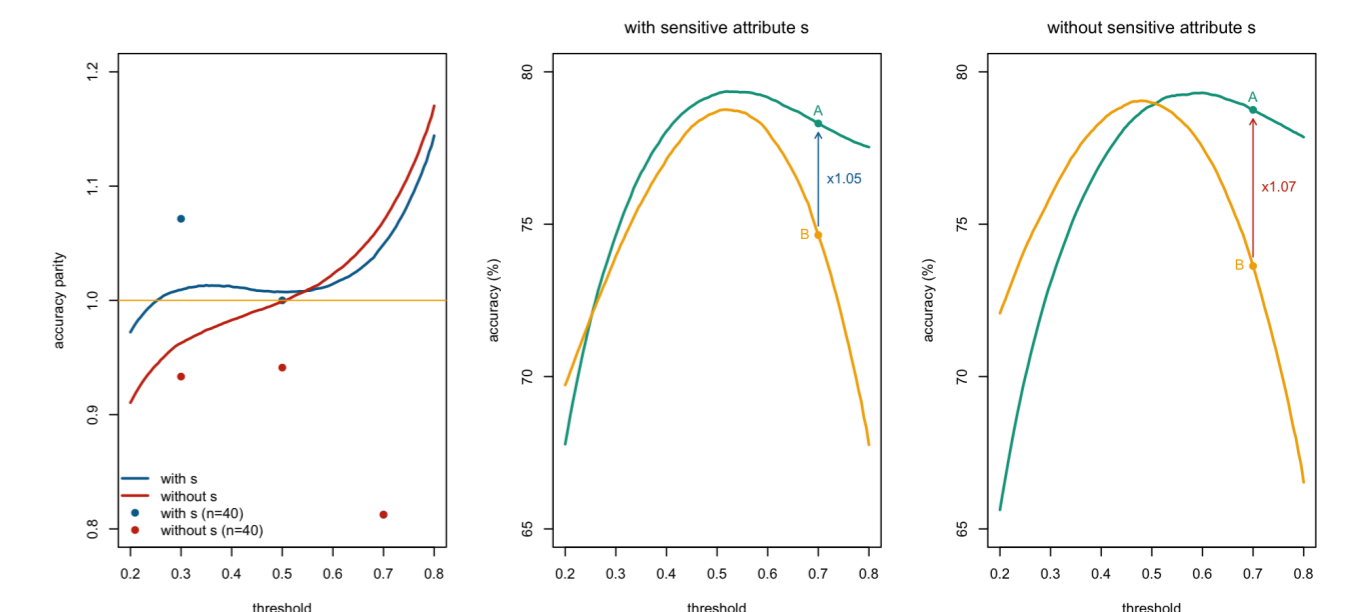
Calibration parity is met if $\forall t \in [0, 1]$,

$$\begin{aligned} \mathbb{P}(Y = 1 \mid \hat{s}(X) = p, S = A) \\ = \mathbb{P}(Y = 1 \mid \hat{s}(X) = p, S = B) \end{aligned}$$

Fairness of good calibration is met if $\forall t \in [0, 1]$,

$$\begin{aligned} \mathbb{P}(Y = 1 \mid \hat{s}(X) = p, S = A) \\ = \mathbb{P}(Y = 1 \mid \hat{s}(X) = p, S = B) = t \end{aligned}$$

Evolution of accuracy, in groups A and B:



References

- Baumann, J., . L. M. (2023). Fairness and Risk: An Ethical Argument for a Group Fairness Definition Insurers Can Use. *Philosophy technology*, 36(3).
- Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.
- Hänsch, R. (2020). Stacked random forests: More accurate and better calibrated. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 1751–1754.
- Liu, Y., Menglong, Y., Wang, Y., Li, Y., and Xiong, T. (2021). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from china. *International Review of Financial Analysis*, 79:101971.
- Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024). From uncertainty to precision: Enhancing binary classifier performance through calibration.
- Mildenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. In *Journal Proceedings of the Casualty Actuarial Society*, volume 86, pages 393–487.
- NAIC (2022). Appendix b-trees – information elements and guidance for a regulator to meet best practices’ objectives (when reviewing tree-based models).
- Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning, ICML ’05*, page 625–632, New York, NY, USA. Association for Computing Machinery.
- Park, Y. and Ho, J. C. (2020). Califorest: Calibrated random forest for health data. *Proceedings of the ACM Conference on Health, Inference, and Learning 2020*, pages 40–50.
- Von Mises, R., Neyman, J., Sholl, D., and Rabinowitsch, E. (1939). *Probability, Statistics and Truth*. Macmillan.