Challenging Calibration of Insurance Scoring Algorithms

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

Tuesday, June 18th, 2024

Setup

- Let us consider a **binary event** D whose observations are denoted $d_i = 1$ if the event occurs, and $d_i = 0$ otherwise, where i denotes the ith observations.

Setup

- Let us consider a **binary event** D whose observations are denoted $d_i = 1$ if the event occurs, and $d_i = 0$ otherwise, where i denotes the ith observations.
- Let us further assume that the (**unobserved**) probability of the event $d_i = 1$ depends on **individual characteristics**:

$$p_i = s(\mathbf{x}_i)$$

where, with sample size $n > 0$, $i = 1, \ldots, n$ represents individuals, and $\mathbf{x}_i$ the characteristics.

Setup

- Let us consider a **binary event** D whose observations are denoted $d_i = 1$ if the event occurs, and $d_i = 0$ otherwise, where i denotes the ith observations.
- Let us further assume that the (**unobserved**) probability of the event $d_i = 1$ depends on **individual characteristics**:

$$p_i = s(\mathbf{x}_i)$$

  where, with sample size $n > 0$, $i = 1, \ldots, n$ represents individuals, and $\mathbf{x}_i$ the characteristics.
- To **estimate this probability**, we can use a statistical model (*e.g.*, a GLM) or a machine learning model (*e.g.*, a random forest).

Motivation

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:

Motivation

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:
  - what is the probability for **this** insured to have an accident within the next year?

Motivation

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:
  - what is the probability for **this** insured to have an accident within the next year?
  - what is the probability of death of **this** individual within the year?

Motivation

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:
  - what is the probability for **this** insured to have an accident within the next year?
  - what is the probability of death of **this** individual within the year?

"*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

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:
  - what is the probability for **this** insured to have an accident within the next year?
  - what is the probability of death of **this** individual within the year?

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

- In such cases, it is important that the **estimated scores** can be interpreted as **probabilities**.

Motivation

- In **insurance**, we find cases where we are more interested in the **underlying risk** than on being able to **discriminate** between the occurrence/non-occurrence of an event:
  - what is the probability for **this** insured to have an accident within the next year?
  - what is the probability of death of **this** individual within the year?

  "*The phrase 'probability of death', when it refers to a single person, has no meaning for us at all.*" Von Mises et al. (1939)
- In such cases, it is important that the **estimated scores** can be interpreted as **probabilities**.
- This might become a problem when using **tree-based classifiers** (Niculescu-Mizil and Caruana, 2005; Park and Ho, 2020; Hänsch, 2020) rather than **logistic regression models** (Machado et al., 2024).

Roadmap

Definition

Calibration of a Binary Classifier (Schervish (1989))

For a binary variable D, a model is well-calibrated when

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

Definition

Calibration of a Binary Classifier (Schervish (1989))

For a binary variable D, a model is well-calibrated when

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

Note: conditioning by $\{\hat{s}(\mathbf{x}) = p\}$ leads to the concept of (local) calibration; however, as discussed by Bai et al. (2021), $\{\hat{s}(\mathbf{x}) = p\}$ is *a.s.* a null mass event. Thus, calibration should be understood in the sense that

$$\mathbb{E}[D \mid \hat{s}(\mathbf{X}) = p] \overset{a.s.}{\to} p \text{ when } n \to \infty \ ,$$

meaning that, asymptotically, the model is well-calibrated, or locally well-calibrated in p, for any p.

Calibration curve

- Estimation of $\mathrm{g}(\cdot)$ (which measures **miscalibration** on predicted scores $\hat{s}(\mathbf{x})$):

$$\mathrm{g}:\begin{cases}[0,1]\to[0,1]\\ p\mapsto \mathrm{g}(p):=\mathbb{E}[D\mid \hat{s}(\mathbf{x})=p]\end{cases}. \tag{2}$$

- **Challenge**: having enough observations with identical scores is difficult.
- **Solutions**:
  1. **Reliability diagram** (Wilks, 1990): grouping obs. into B **bins**, defined by the **quantiles** of predicted scores,
  2. Using a smoother representation with **local regression** techniques, which estimates a conditional expectation within a **specified neighborhood** of predicted scores (Denuit et al., 2021).

Metrics

Brier Score (Brier (1950))

The **Brier Score** does not depend on bins but directly on observations, and is defined as:

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

where d_i is the observed event and $\hat{s}(\mathbf{x}_i)$ the estimated score.

Metrics

Brier Score (Brier (1950))

The **Brier Score** does not depend on bins but directly on observations, and is defined as:

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

where d_i is the observed event and $\hat{s}(\mathbf{x}_i)$ the estimated score.

Integrated Calibration Index or ICI (Austin and Steyerberg (2019))

The **ICI** is based on the calibration curve $\hat{g}$ estimated with **local regression techniques** and is defined as

$$\text{ICI} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(\hat{s}(\mathbf{x}_i)\right) - \hat{s}(\mathbf{x}_i)\right|$$

where $\hat{g}\left(\hat{s}(\mathbf{x}_i)\right)$ represents the prediction obtained from the local regression fit on the estimated score $\hat{s}(\mathbf{x}_i)$.

Illustrative example

- Consider the `frenchmotor` dataset from `InsurFair` (Charpentier, 2014), where we aim to estimate the **probability of accident** for insureds within a year ($n = 12,437$ and 17 explanatory variables), by predicting the **binary response variable** D, indicating the occurrence of an accident.

Illustrative example

- Consider the `frenchmotor` dataset from `InsurFair` (Charpentier, 2014), where we aim to estimate the **probability of accident** for insureds within a year ($n = 12,437$ and 17 explanatory variables), by predicting the **binary response variable** D, indicating the occurrence of an accident.
- We compare predictions from a **GLM** and a **GAM** to those from a **random forest** (RF) regressor, increasingly used in insurance (NAIC, 2022).

Illustrative example

- Consider the `frenchmotor` dataset from `InsurFair` (Charpentier, 2014), where we aim to estimate the **probability of accident** for insureds within a year ($n = 12,437$ and 17 explanatory variables), by predicting the **binary response variable** D, indicating the occurrence of an accident.
- We compare predictions from a **GLM** and a **GAM** to those from a **random forest** (RF) regressor, increasingly used in insurance (NAIC, 2022).

Table 1: Performance and calibration metrics on test set.

Model	AUC	Brier score	ICI
GLM	0.61 ± 0.03	0.08 ± 0.03	0.04 ± 0.03
GAM	0.61 ± 0.03	0.08 ± 0.03	0.04 ± 0.03
RF	0.88 ± 0.03	0.07 ± 0.02	0.05 ± 0.03

Calibration curves (1/2)

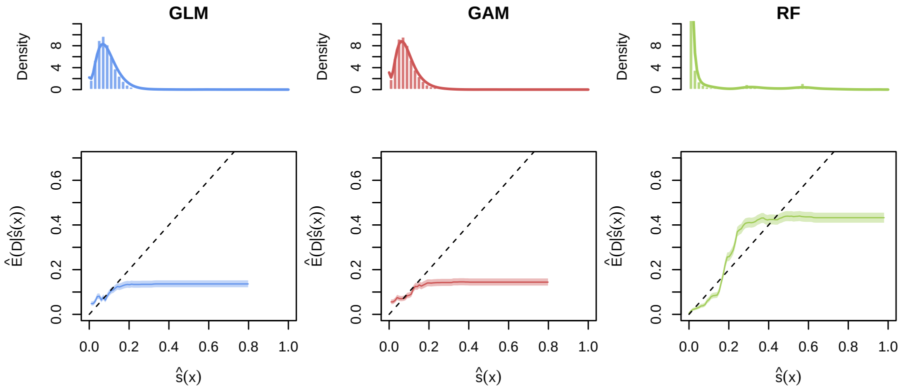

Figure 1: Distribution of estimated scores for the three models, along with their calibration curves generated using `locfit`.

Calibration curves (2/2)

The **lack of score heterogeneity** observed in RF model compared to GLM and GAM is not assessed by calibration metrics.

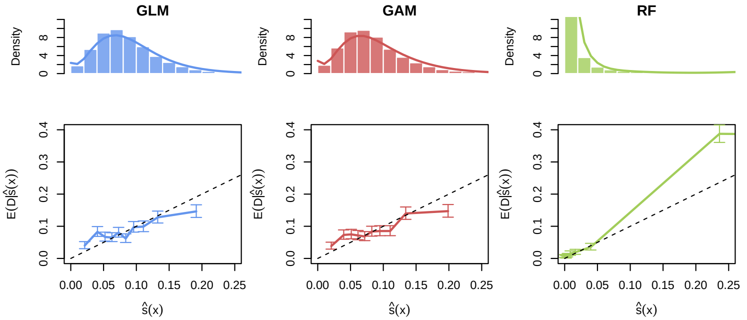

Figure 2: Distribution of estimated scores for the three models, along with their zoomed reliability diagrams.

Overview for decision trees

Here, we consider a **simulated environment** for $D_i \sim \mathcal{B}(p_i)$, with p_i the **true underlying probability distribution**.

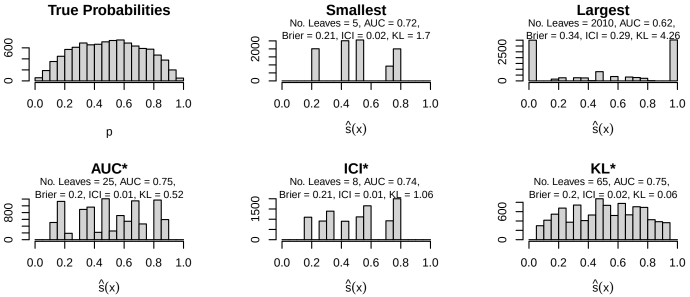

Figure 3: Distribution of true probabilities and estimated scores for trees of interest. The Kullback–Leibler divergence (KL) of ϕ from ψ is defined by $D_{KL}(\phi||\psi)=\sum_{i=1}^{m} h_\phi(i) \log \frac{h_\phi(i)}{h_\psi(i)}$.

Bayesian framework: back to the `frenchmotor` dataset

- The **true underlying data distribution** of D is **not observable**.

Bayesian framework: back to the `frenchmotor` dataset

- The **true underlying data distribution** of D is **not observable**.
- Expert opinion: **Beta prior** to model the underlying data distribution.

Bayesian framework: back to the `frenchmotor` dataset

- The **true underlying data distribution** of D is **not observable**.
- Expert opinion: **Beta prior** to model the underlying data distribution.

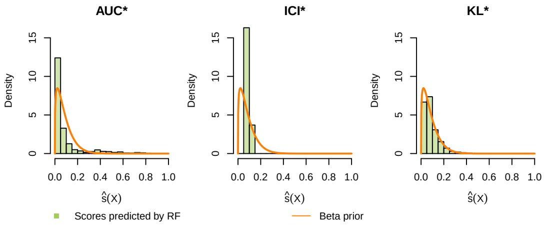

Figure 4: Distribution of RF predicted scores when optimizing hyperparameters for AUC (**AUC***), ICI (**ICI***) and KL (**KL***).

Bayesian framework: back to the `frenchmotor` dataset

- The **true underlying data distribution** of D is **not observable**.
- Expert opinion: **Beta prior** to model the underlying data distribution.

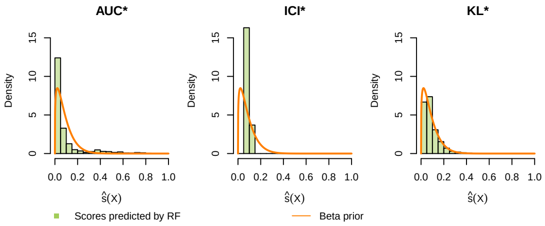

Figure 4: Distribution of RF predicted scores when optimizing hyperparameters for AUC (**AUC***), ICI (**ICI***) and KL (**KL***).

Table 2: Difference in validation set metrics between **ICI***, **KL*** and the reference model: **AUC***.

Optim.	ΔAUC	ΔICI	ΔKL
ICI*	-0.23	-0.02	$+0.44$
KL*	-0.05	$+0.01$	-0.77

Wrap-up

- **Calibration matters**: when training classifiers, looking at calibration of models should not be disregarded.

Wrap-up

- **Calibration matters**: when training classifiers, looking at calibration of models should not be disregarded.
- **Calibration may not be sufficient** for **tree-based methods**: for RF, when score heterogeneity is lacking, metrics such as KL should complement the commonly used calibration metrics.

Wrap-up

- **Calibration matters**: when training classifiers, looking at calibration of models should not be disregarded.
- **Calibration may not be sufficient** for **tree-based methods**: for RF, when score heterogeneity is lacking, metrics such as KL should complement the commonly used calibration metrics.
- Next steps: In particular, for private insurance, **calibration** (or **sufficiency**) emerges as the most suitable metric for evaluating **group fairness**, as highlighted by Baumann and Loi (2023).

Wrap-up

- **Calibration matters**: when training classifiers, looking at calibration of models should not be disregarded.
- **Calibration may not be sufficient** for **tree-based methods**: for RF, when score heterogeneity is lacking, metrics such as KL should complement the commonly used calibration metrics.
- Next steps: In particular, for private insurance, **calibration** (or **sufficiency**) emerges as the most suitable metric for evaluating **group fairness**, as highlighted by Baumann and Loi (2023).

Comments are welcome: `fernandes_machado.agathe@courrier.uqam.ca`

5 Appendix

References I

Austin, P. C. and Steyerberg, E. W. (2019). The integrated calibration index (ici) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* 38: 4051–4065, doi:10.1002/sim.8281.

Bai, Y., Mei, S., Wang, H. and Xiong, C. (2021). Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In *International Conference on Machine Learning*. PMLR, 566–576.

Baumann, J. and Loi, M. (2023). Fairness and Risk: An Ethical Argument for a Group Fairness Definition Insurers Can Use. *Philosophy & technology* 36, doi:https://doi.org/10.1007/s13347-023-00624-9.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1–3.

Charpentier, A. (2014). *Computational Actuarial Science*. CRC Press.

Denuit, M., Charpentier, A. and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics* 101: 485–497, doi:https://doi.org/10.1016/j.insmatheco.2021.09.001.

Hänsch, R. (2020). Stacked Random Forests: More Accurate and Better Calibrated. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 1751–1754.

References II

Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E. and Hu, F. (2024). From uncertainty to precision: Enhancing binary classifier performance through calibration.

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. New York, NY, USA: Association for Computing Machinery, 625–632, doi:10.1145/1102351.1102430.

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* : 40–50.

Schervish, M. J. (1989). A General Method for Comparing Probability Assessors. *The Annals of Statistics* 17: 1856–1879, doi:10.1214/aos/1176347398.

Von Mises, R., Neyman, J., Sholl, D. and Rabinowitsch, E. (1939). *Probability, Statistics and Truth*. Macmillan.

Wilks, D. S. (1990). On the combination of forecast probabilities for consecutive precipitation periods. *Weather and Forecasting* 5: 640 – 650, doi:10.1175/1520-0434(1990)005<0640:OTCOFP>2.0.CO;2.

5 Appendix

Calculation of Performance Metrics

Simulated Environment for Score Heterogeneity

Random Forest Optimization on `frenchmotor` dataset

(Mis-)Calibration and standard metrics

Table 3: Confusion Table

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

where

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

AUC (Area Under Curve): TPR and TFP for various prob. threshold τ

5 Appendix

Calculation of Performance Metrics

Simulated Environment for Score Heterogeneity

Random Forest Optimization on `frenchmotor` dataset

Data Generating Process for Score Heterogeneity

$$D_i \sim \mathcal{B}(p_i),$$

where individual probabilities are obtained using a logistic sigmoid function:

$$p_i = \frac{1}{1 + \exp(-\eta_i)},$$
$$\eta_i = \mathbf{a}\mathbf{x}_i$$

with $\mathbf{a} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} = \begin{bmatrix} 0.5 & 1 \end{bmatrix}$ and $\mathbf{x}_i = \begin{bmatrix} x_{1,i} & x_{2,i} \end{bmatrix}^\top$. The observations $\mathbf{x}_i$ are drawn from a $\mathcal{N}(0, 1)$.

5 Appendix

Calculation of Performance Metrics

Simulated Environment for Score Heterogeneity

Random Forest Optimization on `frenchmotor` dataset

Parameters of RF for different optimization objectives

Table 4: RF parameters for different optimization objectives.

Optim.	*mtry*	*num_trees*	*min_node_size*
AUC*	10	500	2
KL*	10	500	18
ICI*	4	500	512
Brier*	2	500	2

Metrics of RF optimization on validation set

Table 5: AUC, ICI and KL calculations for different RF optimization objectives.

Optim.	AUC	ICI	KL
AUC***	0.78	0.03	0.80
ICI***	0.55	0.002	1.24
KL***	0.73	0.03	0.03