From Uncertainty to Precision: Enhancing Binary Classifier Performance through Calibration.

A. Fernandes Machado[1] A. Charpentier[1] E. Flachaire[2] E. Gallic[2] F. Hu[3]

[1]Université du Québec à Montréal

[2]Aix-Marseille School of Economics, Aix-Marseille Univ.

[3]Milliman France

58th Annual Canadian Economics Association Meetings, May 31st, 2024

Calibration: intuition (1/2)

“ ¦ \0}0 _ 9 ¨· t H\9jH0
nT }9_j §nhn}}n°u”
(Dawid, 1982)

Calibration: intuition (1/2)

“ ¦ \0}0 _ 9 ¨· t H\9jH0
nT }9_j §nhn}}n°u”
(Dawid, 1982)

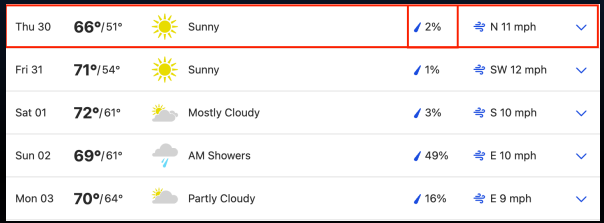
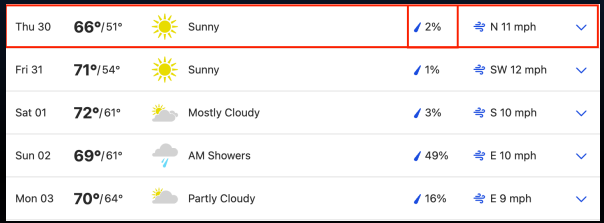


Figure 1: Weather Forecasts on Tuesday, March 2024. Source: The Weather Channel.

Calibration: intuition (1/2)

"¦ \0}0 _ 9 ¨· t H\9jH0
nT }9_j §nhn}}n°u"
(Dawid, 1982)



Thu 30	66°/51°	Sunny	2%	N 11 mph
Fri 31	71°/54°	Sunny	1%	SW 12 mph
Sat 01	72°/61°	Mostly Cloudy	3%	S 10 mph
Sun 02	69°/61°	AM Showers	49%	E 10 mph
Mon 03	70°/64°	Partly Cloudy	16%	E 9 mph

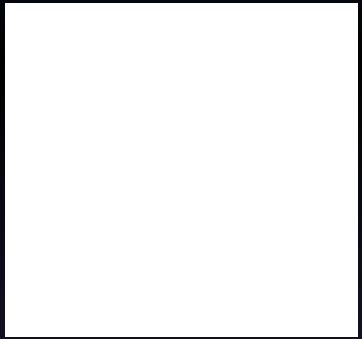
Figure 1: Weather Forecasts on Tuesday, March 2024. Source: The Weather Channel.

Consider a sequence of weather forecasts $\hat{s}(\mathbf{z}_t)$, where $t = 1, \ldots, T$ denotes the days of forecast and $\mathbf{z}$ represents characteristics used in forecasting.

Calibration: intuition (2/2)

Within this sequence, we focus on days where $\widehat{s}(\mathbf{2}_i)$ closely approximates 30%.

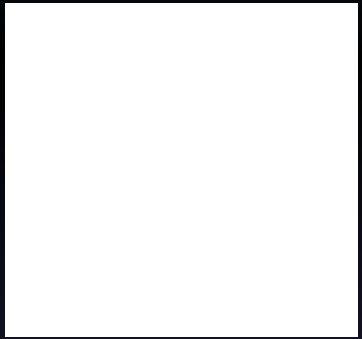
By assuming an infinite sequence, we can determine the long-term proportion p of days where the forecasted event actually occurred.

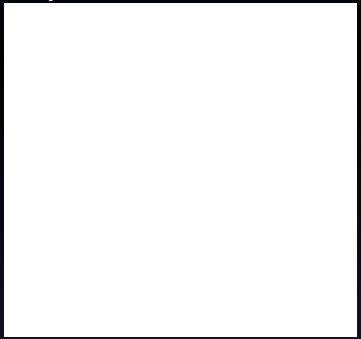


Calibration: intuition (2/2)

Within this sequence, we focus on days where $\widehat{s}(\mathbf{2}_i)$ closely approximates 30%.

By assuming an infinite sequence, we can determine the long-term proportion p of days where the forecasted event actually occurred.





Calibration: intuition (2/2)

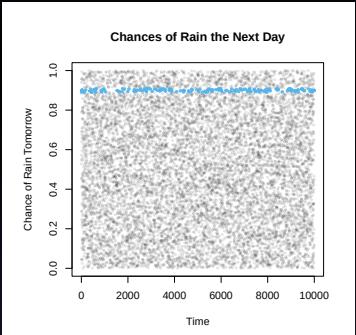
Within this sequence, we focus on days where $\widehat{s}(\mathbf{2}_i)$ closely approximates 30%.

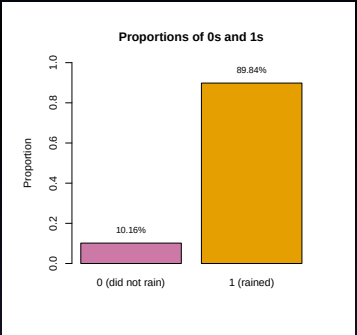
By assuming an infinite sequence, we can determine the long-term proportion p of days where the forecasted event actually occurred.

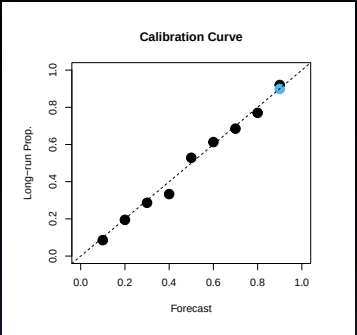
Calibration: intuition (2/2)

Within this sequence, we focus on days where $\hat{s}(\mathbf{2}_i)$ closely approximates 30%.

By assuming an infinite sequence, we can determine the long-term proportion p of days where the forecasted event actually occurred.







Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jLO}ę´_jX }_ c. Other examples include:

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jLO}ę´_jX }_ c. Other examples include:
  - does this patient have a disease or not (Van Calster et al., 2019)?

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jLO}ę´_jX }_ c. Other examples include:
  - does this patient have a disease or not (Van Calster et al., 2019)?
  - will this insured have an accident within the next year?

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jLO}ę´_jX }_ c. Other examples include:
  - does this patient have a disease or not (Van Calster et al., 2019)?
  - will this insured have an accident within the next year?
  - what is the probability for this individual to receive the treatment/control?

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jLO}ℓ´_jX }_ c. Other examples include:
  - does this patient have a disease or not (Van Calster et al., 2019)?
  - will this insured have an accident within the next year?
  - what is the probability for this individual to receive the treatment/control?
- In such cases, it is important that the 0 §_h9§0L Hn}0 can be interpreted as q}n@9@_e_§_0 .

Motivations

- We are interested in being able to L_ H}_h_j9§0 between rainy/not rainy days.
- We are also interested in the «jL0}e´_jX }_ c. Other examples include:
  - does this patient have a disease or not (Van Calster et al., 2019)?
  - will this insured have an accident within the next year?
  - what is the probability for this individual to receive the treatment/control?
- In such cases, it is important that the 0 §_h9§0L Hn}0 can be interpreted as q}n@9@_e_§_0 .
- This might become a problem when using h9H_j0 e09}j_jX He9 _»0} based on 0j 0h@e0 h0§\nL .

This Talk

What the remainder of the talk is about:

- Reviewing of ways to h09 «}0 9jL ®_ «9e_ 0 H9e_@}9§_nj for a @_j9}´ He9 _»0}.

This Talk

What the remainder of the talk is about:

- Reviewing of ways to h09 «}0 9jL ®_ «9e_ 0 H9e_@}9§_nj for a @_j9}´ He9 _»0}.
- Proposing a new metric based on enH9e }0X}0 _nj: the Local Calibration Score.

This Talk

What the remainder of the talk is about:

- Reviewing of ways to h09 «}0 9jL ®_ «9e_ 0 H9e_@}9§_nj for a @_j9}´ He9 _»0}.
- Proposing a new metric based on enH9e }0X}0 _nj: the Local Calibration Score.
- Observing the _hq9H§ nT 9 qnn} H9e_@}9§_nj on standard performance metrics.

This Talk

What the remainder of the talk is about:

- Reviewing of ways to h09 «}0 9jL ®_ «9e_ 0 H9e_@}9§_nj for a @_j9}´ He9 _»0}.
- Proposing a new metric based on enH9e }0X}0 _nj: the Local Calibration Score.
- Observing the _hq9H§ nT 9 qnn} H9e_@}9§_nj on standard performance metrics.
- Examining calibration for §}00]@9 0L h0§\nL .

Take away results

- Our new metric, the **enH9e H9e_@}9§_nj Hn}O** offers a more flexible way to visualise and measure calibration than methods based on empirical quantiles.
- **G9e_@}9§_nj h9§§O}** : when training classifiers, looking at calibration of models should not be disregarded.

Roadmad

Roadmap

Calibration

Setup

- Let us consider a @_j9}´ 0®0j§ 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 @_j9}´ O®Oj§ 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 («jn@ O}®OL) probability of the event $d_i = 1$ depends on _jL_®_L«9e H\9}9H§O}_ §_H :

$$p_i = s(^2{}_i)$$

where, with sample size $n > 0$, $i = 1, \dots, n$ represents individuals, and $^2{}_i$ the characteristics.

Predicting risks

- To 0 §_h9§0 §_ q}n@9@_e_§´, we can use a statistical model (0uXu, a GLM) or a machine learning model (0uXu, a random forest).

Predicting risks

- To 0 §_h9§0 §_ q}n@9@_e_§´, we can use a statistical model (0uXu, a GLM) or a machine learning model (0uXu, a random forest).
- These models 0 §_h9§0 9 Hn}0, $\widehat{s}(^{2}{}_{i}) \in [0, 1]$, allowing the He9 _»H9§_nj of observations based on the estimated probability of the event.

Predicting risks

- To 0 §_h9§0 §_ q}n@9@_e_§´, we can use a statistical model (0uXu, a GLM) or a machine learning model (0uXu, a random forest).
- These models 0 §_h9§0 9 Hn}0, $\hat{s}(^{2}{}_{i}) \in [0, 1]$, allowing the He9 _»H9§_nj of observations based on the estimated probability of the event.
- By setting a probability threshold τ in $[0, 1]$, one can predict the class of each observation: 1 if the event occurs, and 0 otherwise:

$$\hat{d}_i = \begin{cases} 1, & \text{if } \hat{s}(^{2}{}_{i}) \geq \tau \\ 0, & \text{if } \hat{s}(^{2}{}_{i}) < \tau \end{cases} .$$

Predicting risks

- To 0 §_h9§0 §_ q}n@9@_e_§´, we can use a statistical model (0uXu, a GLM) or a machine learning model (0uXu, a random forest).
- These models 0 §_h9§0 9 Hn}0, $\hat{s}(^2{}_i) \in [0, 1]$, allowing the He9 _»H9§_nj of observations based on the estimated probability of the event.
- By setting a probability threshold τ in $[0, 1]$, one can predict the class of each observation: 1 if the event occurs, and 0 otherwise:

$$\hat{d}_i = \begin{cases} 1, & \text{if } \hat{s}(^2{}_i) \geq \tau \\ 0, & \text{if } \hat{s}(^2{}_i) < \tau \end{cases} .$$

- However, if the model is not °Oee H9e_@}9§0L, the scores cannot be interpreted as probabilities.

Definition

Calibration of a Binary Classifier (Schervish, 1989)

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

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

Definition

Calibration of a Binary Classifier (Schervish, 1989)

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

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

Note: conditioning by $\{\hat{s}(^2) = p\}$ leads to the concept of (local) calibration; however, as discussed by Bai et al., 2021, $\{\hat{s}(^2) = p\}$ is 9u u a null mass event. Thus, calibration should be understood in the sense that

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

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

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures $\mathbf{h\_\ H9e\_@}\}\mathbf{9\S\_nj}$ on predicted scores $\widehat{s}(^{2})$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto\ ;(p) := \mathbb{E}[D \mid \widehat{s}(^{2}) = p] \end{cases} . \tag{2}$$

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures **h_ H9e_@}9§_nj** on predicted scores $\widehat{s}(^{2})$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto ;(p) := \mathbb{E}[D \mid \widehat{s}(^{2}) = p] \end{cases} . \tag{2}$$

- **G\9ee0jXO**: having enough observations with identical scores is difficult.

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures **h_ H9e_@}9§_nj** on predicted scores $\widehat{s}(^2)$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto ;(p) := \mathbb{E}[D \mid \widehat{s}(^2) = p] \end{cases} . \tag{2}$$

- **G\9ee0jX0**: having enough observations with identical scores is difficult.
- **~ne«§_nj**: grouping obs. into B **@_j** , defined by the **x«9j§_e0** of predicted scores:

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures **h_ H9e_@}9§_nj** on predicted scores $\widehat{s}(^2)$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto ;(p) := \mathbb{E}[D \mid \widehat{s}(^2) = p] \end{cases} . \tag{2}$$

- **G\9ee0jX0**: having enough observations with identical scores is difficult.
- **~ne«§_nj**: grouping obs. into B **@_j** , defined by the **x«9j§_e0** of predicted scores:
  - The average of observed values ($\bar{d}_b$ with $b \in \{1, \dots, B\}$), in each bin b can then be compared with the central value of the bin.

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures **h_ H9e_@}9§_nj** on predicted scores $\widehat{s}(^2)$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto ;(p) := \mathbb{E}[D \mid \widehat{s}(^2) = p] \end{cases} . \tag{2}$$

- **G\9ee0jX0**: having enough observations with identical scores is difficult.
- **~ne«§_nj**: grouping obs. into B **@_j** , defined by the **x«9j§_e0** of predicted scores:
  - The average of observed values ($\bar{d}_b$ with $b \in \{1, \dots, B\}$), in each bin b can then be compared with the central value of the bin.
  - **G9e_@}9§_nj H«}®0** (reliability diagram (Wilks, 1990): middle of each bin on the x-axis, averages of corresponding observations on the y-axis.

Visual approach: calibration curve

- Estimation of $;(\cdot)$ (which measures **h_ H9e_@}9§_nj** on predicted scores $\widehat{s}(^2)$):

$$;: \begin{cases} [0,1] \to [0,1] \\ p \mapsto ;(p) := \mathbb{E}[D \mid \widehat{s}(^2) = p] \end{cases} . \tag{2}$$

- **G\9ee0jXO**: having enough observations with identical scores is difficult.
- **~ne«§_nj**: grouping obs. into B **@_j** , defined by the **x«9j§_e0** of predicted scores:
  - The average of observed values ($\bar{d}_b$ with $b \in \{1, \dots, B\}$), in each bin b can then be compared with the central value of the bin.
  - **G9e_@}9§_nj H«}®0** (reliability diagram (Wilks, 1990): middle of each bin on the x-axis, averages of corresponding observations on the y-axis.
  - When the model is **°0ee]H9e_@}9§0L**, all B points lie on the **@_ 0H§n}**.

Metrics (1/2)

1 t T 2 + i 2 / * H B # ` i B Q M 1S ` FQ ` KQ ` M1L* 12UB M B 2V H X - k y R 8

$$1 * 1 = \sum_{b=1}^{B} \frac{n_b}{n} \left| \quad + (b) - + Q(M) \right|^{r}$$

r ? 2 ` n 2 B b i ? 2 b K T H 2 B b B i 2 2 M m K # 2 ` Q 7 Q # b 2 ` p { 1 , B . Q M } b X B M # B M

Metrics (1/2)

1 t T 2 + i 2 / * H B # ` i B Q M 1S ` FQ ` KQ ` M1 L* 12 UB M B 2V H X - k y R 8

$$1 * 1 = \sum_{b=1}^{B} \frac{n_b}{n} \left| \quad + (b) - + Q(M) \right|^{r}$$

r ? 2 ` n 2 B b i ? 2 b K T H b B i 2 2 M m K # 2 ` Q 7 Q # b 2 ` p { 1 , B . Q M } b X B M # B M

8 HH « } 9 H ´ +(b) , h ? 2 p 2 ` ; 2 Q 7 2 K T B ` B + H
T ` Q # # B H B i B 2 b Q ` 7 ` + T B 2 Q M + i 2 / 7 + Q ` ` 2 + i H v
+ H b b 2 b X

$$+(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \mathbb{1}_{\hat{d}_i = d_i} \qquad \text{U j V}$$

h ? 2 T ` 2 / B + i 2 $\hat{d}_i$ + H Q ` b Q # b 2 ` p B b B Q M
/ 2 i 2 ` K B M 2 / # b 2 / Q M + H b b B } + i B Q M i ? ` 2 b ? Q H /
$\tau \in [0, 1]$ r ? 2 ` $\hat{d}_i = 1$ B $s(t_i) \geq \tau$ M 9
Q i ? 2 ` r B b 2 X

Metrics (1/2)

1 t T 2 + i 2 / * H B # ` i B Q M 1S ` FQ ` K Q ` M 1 L * 12 UB M B 2 V H X - k y R 8

$$1 * 1 = \sum_{b=1}^{B} \frac{n_b}{n} \left| \quad + (b) - + Q(M) \right|^{r}$$

r ? 2 ` n 2 B b i ? 2 b K T H 2 BbB i 2 2 M m K # 2 ` Q 7 Q # b 2 ` p { 1, B . Q M } b X B M # B M

8 H H « } 9 H ´ +(b) , h ? 2 p 2 ` ; 2 Q 7 2 K T B ` B + H
T ` Q # # B H B i B 2 b Q ` 7 ` + T B 2 Q M + i 2 / 7 + Q ` ` 2 + i H v + Q(M) 7 A M / B + i 2 b i ? 2 K Q / 2 H
+ H b b 2 b X Gnj» L0 j H0 p 2 ` ; 2 + Q M } / 2 M + 2 b r B v ? p 2 M ` # ; B M ;

$$+(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \mathbb{1}_{\hat{d}_i = d_i} \qquad \text{UjV}$$

h ? 2 T ` 2 / B + i 2 / + H Q ` b Q # b 2 ` p B b B Q M + Q(M) 7 = $\frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \hat{s}(t_i)$
/ 2 i 2 ` K B M 2 / # b 2 / Q M + H b b B } + i B Q M i ? ` 2 b ? Q H /
$\tau \in [0,1]$ r ? 2 ` $\hat{d}_i = 1$ B $\hat{s}(t_i) \geq \tau$ M 0
Q i ? 2 ` r B b 2 X

Metrics (2/2)

Brier Score (Brier, 1950)

The ?}_0} ~Hn}0 does not depend on bins and is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{s}(^2{}_i))^2 \tag{4}$$

where d_i is the observed event and $\hat{s}(^2{}_i)$ the estimated score.

Metrics (2/2)

Brier Score (Brier, 1950)

The ?}_0} ~Hn}0 does not depend on bins and is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{s}(^2{}_i))^2 \tag{4}$$

where d_i is the observed event and $\hat{s}(^2{}_i)$ the estimated score.

Mean Squared Error (MSE)

By substituting the observed event d_i by the true probability p_i (which can only be observed in an experimental setup), the metric becomes the MSE:

$$\text{True MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{s}(^2{}_i))^2 \tag{5}$$

Our Approach: Smoother Visualization Technique

We propose an alternative approach to visualize model calibration, aiming for a hnn§\0} }0q}0 0j§9§_nj: enH9e }0X}0 _nj (Loader, 1999).

- Measuring calibration consists in estimating a HnjL_§_nj9e 0²q0H§9§_nj: a local regression seems appropriate.

Our Approach: Smoother Visualization Technique

We propose an alternative approach to visualize model calibration, aiming for a **hnn§\0} }0q}0 0j§9§_nj**: **enH9e }0X}0 _nj** (Loader, 1999).

- Measuring calibration consists in estimating a **HnjL_§_nj9e 0²q0H§9§_nj**: a local regression seems appropriate.
- Local Regression have been disregarded in high dimensions due to poor properties, but it is **_X\e´ 0 ¹ H_0j§ _j h 9ee L_h 0j _nj** , as in this case with only one predictive feature, $\hat{s}(x) \in [0, 1]$.

Our Approach: Smoother Visualization Technique

We propose an alternative approach to visualize model calibration, aiming for a hnn§\0} }0q}0 0j§9§_nj: enH9e }0X}0 _nj (Loader, 1999).

- Measuring calibration consists in estimating a HnjL_§_nj9e 0²q0H§9§_nj: a local regression seems appropriate.
- Local Regression have been disregarded in high dimensions due to poor properties, but it is _X\e´ 0 ¹ H_0j§ _j h 9ee L_h 0j _nj , as in this case with only one predictive feature, $\hat{s}(x) \in [0, 1]$.
- Given the number of data points, the precision of quantile binning can be suboptimal when determining the appropriate bin count.

Our Approach: Smoother Visualization Technique

We propose an alternative approach to visualize model calibration, aiming for a hnn§\0} }0q}0 0j§9§_nj: enH9e }0X}0 _nj (Loader, 1999).

- Measuring calibration consists in estimating a HnjL_§_nj9e 0²q0H§9§_nj: a local regression seems appropriate.
- Local Regression have been disregarded in high dimensions due to poor properties, but it is _X\e´ 0 ¹H_0j§ _j h9ee L_h0j _nj , as in this case with only one predictive feature, $\hat{s}(x) \in [0, 1]$.
- Given the number of data points, the precision of quantile binning can be suboptimal when determining the appropriate bin count.
- By contrast, with local regression, one can specify the percentage of nearest neighbors, providing greater flexibility.

Our new metric: LCS

Local Calibration Score (LCS)

A local regression of degree 0, denoted as $\hat{g}$, is fitted to the predicted scores $\hat{s}(^{2})$. This fit is then applied to a vector of e_j09}t´ q9HOL ®9e«0 within the interval $[0, 1]$. Each of these points is denoted by l_j, where $j \in \{1, \ldots, J\}$, with J being the target number of points on the visualization curve.
The LCS is defined as:

$$\text{LCS} = \sum_{j=1}^{J} w_j (\hat{g}(l_j) - l_j)^2, \tag{6}$$

where w_j is a weight defined as the density of the $score$ at l_j.

Our new metric: LCS

Local Calibration Score (LCS)

A local regression of degree 0, denoted as $\hat{g}$, is fitted to the predicted scores $\hat{s}(^{2})$. This fit is then applied to a vector of e_j09}t´ q9HOL ®9e«0 within the interval $[0, 1]$. Each of these points is denoted by l_j, where $j \in \{1, \ldots, J\}$, with J being the target number of points on the visualization curve.
The LCS is defined as:

$$\text{LCS} = \sum_{j=1}^{J} w_j (\hat{g}(l_j) - l_j)^2, \tag{6}$$

where w_j is a weight defined as the density of the $score$ at l_j.

Note: Austin and Steyerberg, 2019 defined a similar metric using a L1 norm.

Roadmap

Impact of Poor Calibration

Data Generating Process

We _h «e9§0 binary observations as in Gutman et al., 2022:

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

where individual probabilities are obtained using a logistic sigmoid function:

$$p_i = \frac{1}{1 + 2\,\mathrm{t}(\mathrm{T} - \eta_i)},$$
$$\eta_i = 9\,^2{}_i + \varepsilon_i$$

with $9 = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix} = \begin{bmatrix} 0.1 & 0.05 & 0.2 & -0.05 \end{bmatrix}$ and $^2{}_i = \begin{bmatrix} x_{1,i} & x_{2,i} & x_{3,i} & x_{4,i} \end{bmatrix}^\top$.

The observations $^2{}_i$ are drawn from a $\mathcal{U}(0,1)$ and $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$.

Forcing Poor Calibration

To simulate «jH9ɐ_@}9§_nj, we generate samples of $2,000$ observations and we apply (monotonous) transformations to the true probabilities, either on:

1. the latent probability p_i:

$$p_i^u = \left(\frac{1}{1 + \exp(-\eta_i)} \right)^{\alpha} \quad . \tag{7}$$

2. the linear predictor η_i:

$$\eta_i^u = \gamma \times ((-0.1)x_1 + 0.05x_2 + 0.2x_3 - 0.05x_4 + \varepsilon_i) \quad . \tag{8}$$

Forcing Poor Calibration

To simulate **«jH9ɐ_@}9§_nj**, we generate samples of $2,000$ observations and we apply (monotonous) transformations to the true probabilities, either on:

1. the latent probability p_i:

$$p_i^u = \left(\frac{1}{1 + \exp(-\eta_i)} \right)^{\alpha} . \tag{7}$$

2. the linear predictor η_i:

$$\eta_i^u = \gamma \times ((-0.1)x_1 + 0.05x_2 + 0.2x_3 - 0.05x_4 + \varepsilon_i) . \tag{8}$$

The resulting transformed probabilities are considered as the scores: $\widehat{s}(^2) := p_i^u$

Distortions

- We examine variations in $\{1/3, 1, 3\}$ for α and γ
- For each of the 6 scenarios, we generate 200 samples of $2,000$ obs.

Distortions

- We examine variations in $\{1/3, 1, 3\}$ for α and γ
- For each of the 6 scenarios, we generate 200 samples of $2,000$ obs.

Figure 2: Distorted Prob. as a Function of True Prob., Depending on the Value of α (left) or γ (right)

Impacts: Calibration Metrics

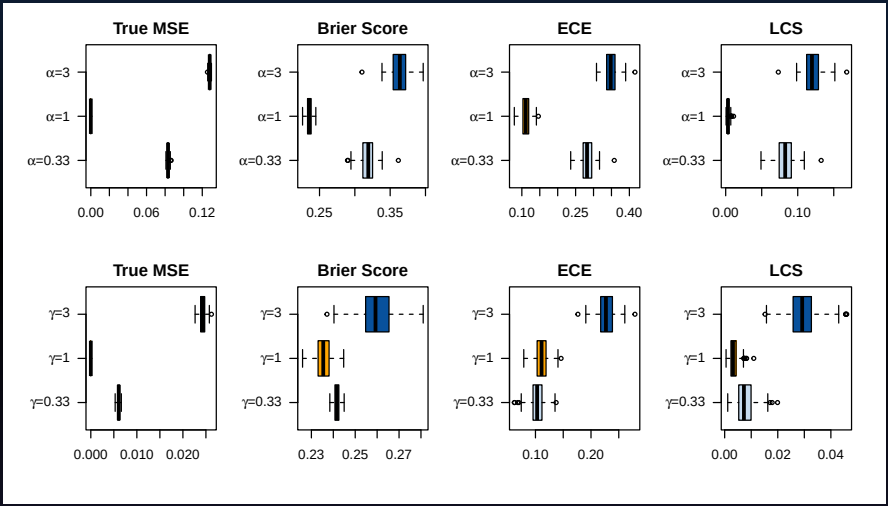


Figure 3: Calibration Metrics on 200 Simulations for each Value of α (top) or γ (bottom).

Impacts: Calibration Curves

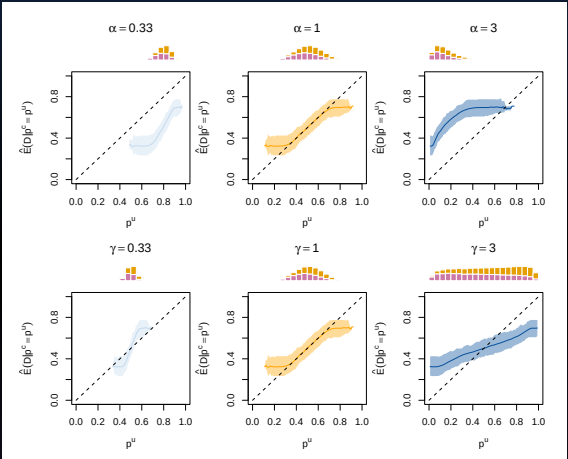


Figure 4: Calibration Curve Obtained with Local Regression, on 200 simulations for each Value of α (top) or γ (bottom). Distribution of the true probabilities are shown in the histograms (gold for $d = 1$, purple for $d = 0$).

(Mis-)Calibration and standard metrics

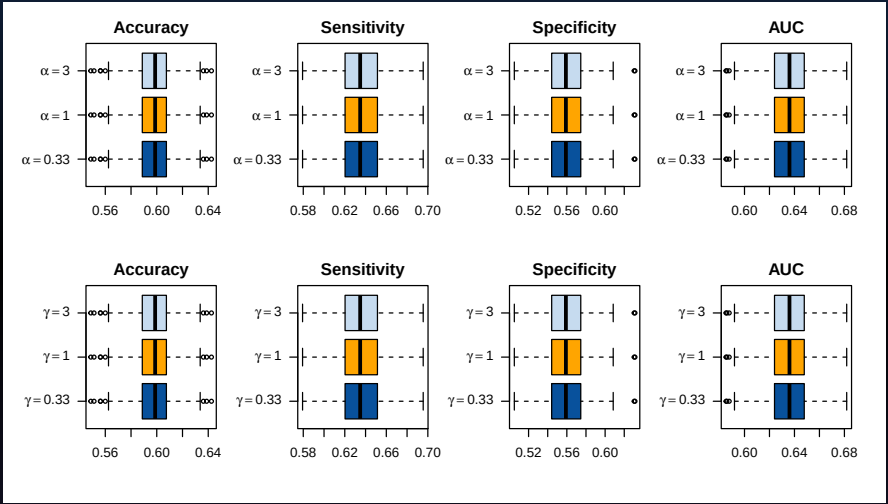


Figure 5: Standard Goodness of Fit Metrics on 200 Simulations for each Value of α (top) or γ (bottom). The probability threshold is set to $\tau = 0.5$.

Roadmap

Calibration and Tree-Based Methods

Calibration for Machine Learning Algorithms

- With the promise of better performance from machine learning models, it can be tempting to rely on these types of models, such as random forests or derivatives, to estimate binary events.

Calibration for Machine Learning Algorithms

- With the promise of better performance from machine learning models, it can be tempting to rely on these types of models, such as random forests or derivatives, to estimate binary events.
- However, the distribution of scores returned by these models can be far from the distribution of the underlying probabilities.

Calibration for Machine Learning Algorithms

- With the promise of better performance from machine learning models, it can be tempting to rely on these types of models, such as random forests or derivatives, to estimate binary events.
- However, the distribution of scores returned by these models can be far from the distribution of the underlying probabilities.
- Here we present an overview of the preliminary results we have obtained with regression trees.

Calibration for Machine Learning Algorithms

- With the promise of better performance from machine learning models, it can be tempting to rely on these types of models, such as random forests or derivatives, to estimate binary events.
- However, the distribution of scores returned by these models can be far from the distribution of the underlying probabilities.
- Here we present an overview of the preliminary results we have obtained with regression trees.
- More results in the next version of the paper...

Trees

Are trees well calibrated?

Trees

Are trees well calibrated?

- Some learning algorithms are designed to yield well-calibrated probabilities. These include LOH_ _nj §}OO , whose leaf probabilities are optimal on the training set (Kull et al., 2017)
- Earlier studies show that also classical methods such as LOH_ _nj §}OO , boosting, SVMs and naive Bayes classifiers tend to be miscalibrated (Wenger et al., 2020)

Preliminary Results: Calibration is not enough

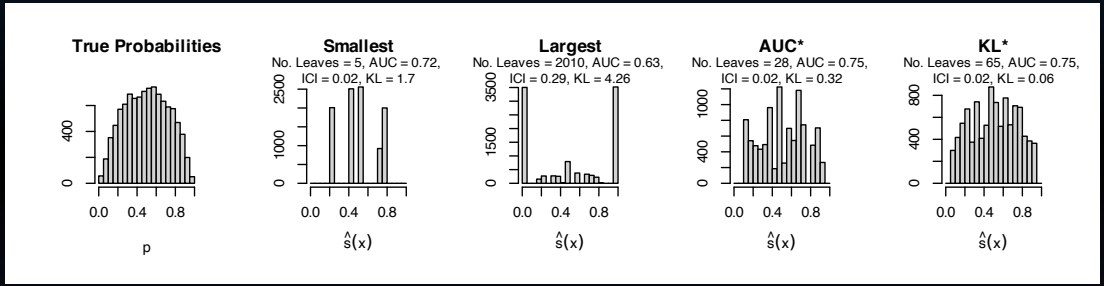


Figure 6: Distribution of true probabilities and estimated scores on validation set for trees of interest, $n = 10,000$

Wrap up

- Our new metric, the enH9e H9e_@}9§_nj Hn}O offers a more flexible way to visualise and measure calibration than methods based on empirical quantiles.
- G9e_@}9§_nj h9§§O} : when training classifiers, looking at calibration of models should not be disregarded.

Comments are welcome: ewen.gallic@univ-amu.fr

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.

a i iBbiB+b BM, 38(21):4051–4065.

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 AMi2`M iBQM H*QM72`2M+2 ,pages 566–575. PMLR.

Brier, G. W. (1950).

Verification of forecasts expressed in terms of probability.

JQMi?Hvq2i?2`, 78(1):1–3.

References II

Dawid, A. P. (1982).

The well-calibrated bayesian.

C Q m ` M H Q 7 i ? 2 K 2 ` B + M a i i B 7 B + Q M i B Q M, 77(379):605–610.

Gutman, R., Karavani, E., and Shimoni, Y. (2022).

Propensity score models are better when post-calibrated.

Kull, M., Filho, T. M. S., and Flach, P. (2017).

Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration.

1 H 2 + i ` Q M B + C Q m ` M H Q 7 a i i B b i B + b, 11(2):5052–5080.

References III

Loader, C. (1999).

6 B i i B M ; r B i ? G R * 6 A 3, chapter 3, pages 45–58.

Springer New York, New York, NY.

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015).

Obtaining well calibrated probabilities using bayesian binning.

S ` Q + 2 2 / B M ; b Q 7 i ? 2 A * Q M 7 2 ` 2 M + 2 Q M 29(1):2901–2907. A M i 2 H H B ; 2 M + 2

Schervish, M. J. (1989).

A General Method for Comparing Probability Assessors.

h ? 2 M M H b Q 7, 17(4):1856–1879.

References IV

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019).

Calibration: the achilles heel of predictive analytics.

" J * J 2 / B + B 1 7 (2).

Wenger, J., Kjellström, H., and Triebel, R. (2020).

Non-parametric calibration for classification.

In S ` Q + 2 2 / B M ; b Q 7 i ? 2 kj ` / A M i 2 ` M i B Q M H * Q M 7 2 ` 2 M + 2 Q M ` i B } + B H A M i
Proceedings of Machine Learning Research.

Wilks, D. S. (1990).

On the combination of forecast probabilities for consecutive precipitation periods.

q 2 i ? 2 ` M / 6 Q ` 2 , 5 (4) i 6 4 0 , 6 5 0.

(Mis-)Calibration and standard metrics

What are the impacts of miscalibration on standard metrics?

We will consider metrics based on the predictive performances calculated using a confusion table:

Table 1: Confusion Table

8H§«9e¤p}0L_H§0L	pn _§_®0	i 0X9§_®0
pn _§_®0	TP	FN
i 0X9§_®0	FP	TN

where

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

(Mis-)Calibration and standard metrics

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}}$$ Overall correctness of the model

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ Ability to correctly identify positive class

$$\text{Specificity} = TPR = \frac{\text{TN}}{\text{TN} + \text{FP}}$$ Ability to correctly identify negative class

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