Calibration of Probabilistic Scores of Classifiers

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

(discussant: Yang Lu)

Motivation

Econometrics $\longleftrightarrow$ **Machine Learning**

- Econometrics [Working, 1927], [Tinbergen, 1939] and, as in [Morgan, 1990]:

"*it has been considered legitimate to use some of the tools developed in statistical theory without accepting the very foundation upon which statistical theory is built* [...] *The reluctance among economists to accept probability models as a basis for economic research has, it seems, been founded upon a very narrow concept of probability and random variables,* " [Haavelmo, 1944]

- Machine Learning ("data mining" in [Friedman, 1998]), [Charpentier et al., 2018]:

"*the logistic regression can also be interpreted from a probabilistic perspective,* " [Watt et al., 2016]

- $(y_i, \boldsymbol{x}_i)$, realizations of $(Y_i, \boldsymbol{X}_i)$ on $(\Omega, \mathcal{F}, \mathbb{P})$, [Gourieroux and Monfort, 1995],

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}] = \frac{\exp[\boldsymbol{x}^\top \beta]}{1 + \exp[\boldsymbol{x}^\top \beta]} = s(\boldsymbol{x})$$

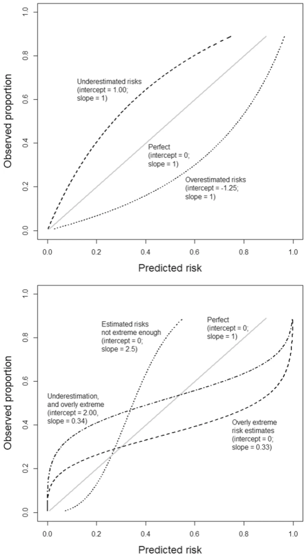
"probabilistic scores"

Motivation

In many applications of classification, there is a need for 'calibrated' probabilistic classifiers which reflect the likelihood of the positive class given the features of an instance in a frequentist statistical sense $\longrightarrow$ calibration

As explained in [Van Calster et al., 2019] "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*,"

- If 40 out of 100 in this group are found to have the disease, the risk is **underestimated**
- If we observe that in this group, 10 out of 100 have the disease, we have **overestimated** the risk.



Motivation

Interesting predictive power of Machine Learning algorithms (random forest, boosting, neural nets, etc), based on various "accuracy" metrics, e.g. $\text{AUC}_s = \int_{\mathbb{R}} D_s(p)\text{d}p$.

Definition (4.8 in [Gourieroux and Jasiak, 2015])

The discriminant curve of s is $D_s(p) = \overline{F}_1 \circ \overline{F}_0^{-1}(p)$, where

$$\overline{F}_0(p) = \mathbb{P}[s(\boldsymbol{X}) > p \mid Y = 0] = \text{FPR and } \overline{F}_1(p) = \mathbb{P}[s(\boldsymbol{X}) > p \mid Y = 1] = \text{TPR}.$$

If ψ is non-decreasing, $D_{\psi\circ s} = D_s$.

"*[Guo et al., 2017] have shown that modern neural networks are poorly calibrated and over-confident despite having better performance*," [Müller et al., 2019] or "*deep neural networks tend to be overconfident and poorly calibrated after training*," [Wang et al., 2021]

Scores

"*The individual characteristics are an essential part of any model for individual risk assessment. Their statistical summary is called a score*," [Gourieroux and Jasiak, 2015]

In the context of a logistic regression, the "*canonical score*" is $S : \mathbb{R}^p \to [0, 1]$,

$$S(\boldsymbol{x}) := \text{logit}(\boldsymbol{x}^\top \beta) = \frac{\exp[\boldsymbol{x}^\top \beta]}{1 + \exp[\boldsymbol{x}^\top \beta]}.$$

Following [Platt, 1999] (see also [Sollich, 1999], [Niculescu-Mizil and Caruana, 2005]), even non-probabilistic machine learning algorithm, such as SVM, could return some canonical score $S : \mathbb{R}^p \to [0, 1]$,
"*Standard SVMs do not provide such probabilities* [...] *we train an SVM, then train the parameters of an additional sigmoid function to map the SVM outputs into probabilities*," [Platt, 1999]

Probabilities... probabilities everywhere...

E.g., **structural probit method**, as defined in [Lee, 1979, Maddala, 1983], a two-stage method, used in [Robinson and Tomes, 1984] (on wages differentials in the public vs. private sectors) or [Kostiuk, 1990] (shift work vs. regular daytime)

$$\begin{cases} y_i \mid (d_i, \boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_d + \varepsilon_i \\ \mathbb{P}[D_i = 1 \mid \boldsymbol{x}_i] = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\alpha}) \end{cases}$$

probability to make a specific decision

E.g., **average treatment effects**, as computed in [Hirano et al., 2003], inspired by [Rosenbaum and Rubin, 1983, Rosenbaum and Rubin, 1984],

probability to be treated

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i t_i}{\widehat{s}(\boldsymbol{x}_i)} - \frac{y_i(1-t_i)}{1-\widehat{s}(\boldsymbol{x}_i)}\right) \text{ where } s(\boldsymbol{x}_i) = \mathbb{P}[T_i = 1 \mid \boldsymbol{x}_i]$$

[Fernandes Machado et al., 2024a], From uncertainty to precision: Enhancing binary classifier perf...

[Fernandes Machado et al., 2024c], Probabilistic scores of classifiers, calibration is not enough

[Fernandes Machado et al., 2024b], Post-calibration techniques: Balancing calibration and score dist....

Scores : Sufficiency

Following [Cook, 2007, Adragni and Cook, 2009], if Y is a random variable and $\mathbf{X}$ a $\mathbb{R}^p$-random vector, $S:\mathbb{R}^p\to\mathbb{R}^q$ with $q<p$ is a "sufficient dimension reduction" (SDR) if it satisfies one of the following three statements

$$\begin{cases}\text{inverse reduction: } (\boldsymbol{X}\mid Y,\boldsymbol{X})\overset{\mathcal{L}}{=}(\boldsymbol{X}\mid S(\boldsymbol{X}))\\ \text{forward reduction: } (Y\mid \boldsymbol{X})\overset{\mathcal{L}}{=}(Y\mid S(\boldsymbol{X}))\\ \text{joint reduction: } Y\perp\!\!\!\perp \boldsymbol{X}\mid S(\boldsymbol{X})\end{cases}$$

i.e. "*the reduction $S(\boldsymbol{X})$ carries all the information that $\boldsymbol{X}$ has about Y.*"

Definition

The **regression function**

$$\mu(\boldsymbol{x}) := \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$$

Scores : Sufficiency

Global balance,

$$\mathbb{E}[Y - \widehat{s}(\boldsymbol{X})] = \mathbb{E}[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X})] = 0.$$

Economically, if $\widehat{s}(\boldsymbol{x})$ is the price, the portfolio is self-financing (for random losses Y).

Marginal balance,

$$\begin{cases} \mathbb{E}[Y - \widehat{s}(\boldsymbol{X}) \mid X_j] = \mathbb{E}[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X}) \mid X_j] = 0 \\ \mathbb{E}[Y - \widehat{s}(\boldsymbol{X}) \mid \boldsymbol{X}] = \mathbb{E}[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X}) \mid \boldsymbol{X}] = 0 \end{cases}$$

Economically, subgroups $\boldsymbol{x}$ are self-financing (for random losses Y).

Well-calibration (or "marginal balance", w.r.t. $\widehat{s}(\boldsymbol{x})$)

$$\mathbb{E}[Y - \widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})] = \mathbb{E}[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})] = 0.$$

Economically, price-based subgroups $\widehat{s}(\boldsymbol{x})$ are self-financing (for random losses Y).

Calibration: Fairness

... "*on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs*," [Arrow, 1963].

i.e. in insurance, "**actuarially fair premiums**" = "**expected losses**" (global balance)

$$\mathbb{E}\big[Y - \widehat{\pi}(\boldsymbol{X})\big] = \mathbb{E}\big[\mu(\boldsymbol{x}) - \widehat{\pi}(\boldsymbol{X})\big] = 0$$

[Baumann and Loi, 2023] suggests that "**local actuarial fairness**" should be considered (well-calibration)

$$\mathbb{E}\big[Y - \widehat{\pi}(\boldsymbol{X}) \mid \widehat{\pi}(\boldsymbol{X})\big] = \mathbb{E}\big[\mu(\boldsymbol{x}) - \widehat{\pi}(\boldsymbol{X}) \mid \widehat{\pi}(\boldsymbol{X})\big] = 0$$

Calibration: Definition

Definition

Estimated score $\widehat{s}$ is **(well-)calibrated** (for Y w.r.t. $\boldsymbol{X}$) if

$$\widehat{s}(\boldsymbol{X}) = \mathbb{E}\big[Y \mid \widehat{s}(\boldsymbol{X})\big],\ \mathbb{P}\text{-a.s.}$$

For example, μ is well calibrated,

$$\mu(\boldsymbol{X}) := \mathbb{E}\big[Y \mid \boldsymbol{X}\big] = \mathbb{E}\big[Y \mid \mu(\boldsymbol{X})\big]$$

Calibration: Curve g, or "calibration curve"

[Schervish, 1989] defined well-calibrated as

$$\mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p] = p, \quad \forall p \in [0, 1].$$

Thus, based on that previous expression, consider the calibration curve, named "reliability diagrams" in [Sanders, 1963, Wilks, 1990]

Definition

The **calibration curve**

$$g : \begin{cases} [0,1] \to [0,1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p] \end{cases}$$

The g function for a well-calibrated model $\widehat{s}$ is the identity function $g(p) = p$.

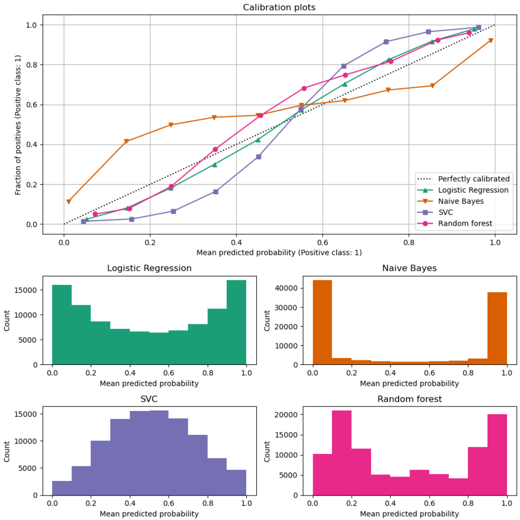
Calibration: Curve g, or "calibration curve"

[Wilks, 1990], [Pakdaman Naeini et al., 2015] and [Kumar et al., 2019] considered quantile-based bins : $\overline{g}$ is the continuous piecewise linear function, interpolating linearly between the points

$$\{(\overline{s}_k, \overline{y}_k)\} \text{ where } k = 1, \cdots, 10,$$

$$\overline{s}_k = \frac{10}{n} \sum_{i \in I_k} \widehat{s}(\boldsymbol{x}_i) \text{ and } \overline{y}_k = \frac{10}{n} \sum_{i \in I_k} y_i,$$

$$I_k = \left\{ i : \left\lceil \frac{k-1}{10} \cdot n \right\rceil \leq \text{rank}(\widehat{s}(\boldsymbol{x}_i)) \leq \left\lfloor \frac{k}{10} \cdot n \right\rfloor \right\}$$



Calibration: Curve g, or "calibration curve"

Given sample $\{(\boldsymbol{x}_i, y_i)\}$ and score $\widehat{s}$, consider a **local regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in [Loader, 2006], see [Austin and Steyerberg, 2019, Denuit et al., 2021]. E.g.

$$\widehat{g}(p) := \frac{\displaystyle\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i)) \cdot y_i}{\displaystyle\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i))}, \ \forall p \in [0,1],$$

based on [Nadaraya, 1964, Watson, 1964], for some kernel K and some bandwidth h.

Calibration: Curve g, or "calibration curve"

Since g should be increasing, quite naturally, we could consider an **isotonic regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in [Kruskal, 1964], see [Niculescu-Mizil and Caruana, 2005, Wüthrich and Ziegel, 2024], $\widetilde{g}$ is the continuous piecewise linear function, interpolating linearly between the points $(\widehat{s}(\boldsymbol{x}_i), \hat{y}_i)$, where $\widehat{s}(\boldsymbol{x}_i)$'s are sorted,

$$\widetilde{g}(p) := \begin{cases} \hat{y}_1 & \text{if } p \leq \widehat{s}(\boldsymbol{x}_1) \\ \hat{y}_i + \dfrac{p - \widehat{s}(\boldsymbol{x}_i)}{\widehat{s}(\boldsymbol{x}_{i+1}) - \widehat{s}(\boldsymbol{x}_i)}(\hat{y}_{i+1} - \hat{y}_i) & \text{if } \widehat{s}(\boldsymbol{x}_i) \leq x \leq \widehat{s}(\boldsymbol{x}_{i+1}) \\ \hat{y}_n & \text{if } x \geq \widehat{s}(\boldsymbol{x}_n) \end{cases}$$

where

$$\min_{\hat{y}_1,\cdots,\hat{y}_n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \text{ subject to } \hat{y}_i \leq \hat{y}_j \text{ for all } (i,j) \in E,$$

$E = \{(i,j) : \widehat{s}(\boldsymbol{x}_i) \leq \widehat{s}(\boldsymbol{x}_j)\}$ specifies the partial ordering of the observed inputs $\widehat{s}(\boldsymbol{x}_i)$.

Calibration : loss and curves

If $\widehat{s}$ is well-calibrated,

$$\widehat{s}(\boldsymbol{X}) \preceq_{cx} \mu(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}] \preceq_{cx} Y$$

Given a differentiable convex function φ, define Bregman loss function

$$\ell(y,s) = \varphi(y) - \varphi(s) - \varphi'(s)\cdot(y-s).$$

From [Krüger and Ziegel, 2021], Bregman dominance reduces to the convex order for autocalibrated predictors

$$\mathbb{E}\big[\ell(Y,\widehat{s}_1(\boldsymbol{X}))\big] \le \mathbb{E}\big[\ell(Y,\widehat{s}_2(\boldsymbol{X}))\big]\ \forall\varphi \iff \widehat{s}_1(\boldsymbol{X}) \preceq_{cx} \widehat{s}_2(\boldsymbol{X})$$

Calibration : loss and curves

Let $F_{\widehat{s}}(\cdot)$ denote the cumulative distribution function of $\widehat{s}(\boldsymbol{X})$, $F_{\widehat{s}}(t) = \mathbb{P}\big[\widehat{s}(\boldsymbol{X}) \leq t\big]$.
Lorenz curve (of y),

$$LC_Y(t) := \frac{\mathbb{E}[Y \cdot \mathbf{1}(Y \leq F_Y^{-1}(t))]}{\mathbb{E}[Y]}$$

while the **concentration curve** (of y w.r.t. $\widehat{s}$), [Yitzhaki and Schechtman, 2013],

$$CC_{Y|\widehat{s}}(t) := \frac{\mathbb{E}[Y \cdot \mathbf{1}(\widehat{s}(\boldsymbol{X}) \leq F_{\widehat{s}}^{-1}(t))]}{\mathbb{E}[Y]}$$

If $\widehat{s}$ is well-calibrated, then $CC_{\mu|\widehat{s}}(t) = LC_{\widehat{s}}(t)$, for every probability level $t \in [0,1]$.

Calibration: Metrics

A standard metric for assessing calibration is Brier score (see [Gupta et al., 2021, Kull et al., 2017, Platt, 1999, Rahimi et al., 2020]), from [Brier, 1950]:

$$\text{Brier score (MSE),}\quad \text{BS} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{s}(\mathbf{x}_i) - y_i\big)^2.$$

[Austin and Steyerberg, 2019] and [Zhang et al., 2020] proposes the **Integrated Calibration Index** (ICI) based on the calibration curve,

$$\text{Integrated Calibration Index,}\quad \text{ICI} = \frac{1}{n}\sum_{i=1}^{n}\big|\,\hat{s}(\mathbf{x}_i) - \hat{g}\big(\hat{s}(\mathbf{x}_i)\big)\,\big|\,.$$

$$\text{Local Calibration Score,}\quad \text{LCS} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{s}(\mathbf{x}_i) - \hat{g}\big(\hat{s}(\mathbf{x}_i)\big)\Big)^2.$$

Calibration: Decomposition

Let $\hat{s}$ denote a scoring classifier, $\mathcal{X} \to [0,1]$, then set $\widehat{S} := \hat{s}(\boldsymbol{X})$.

Let M denote the true regression function, $M := \mu(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$.

Let $C := \mathbb{E}[Y|\widehat{S}]$, corresponding to the true proportion of 1's among the instances for which the model has output the same estimate $\widehat{S}$.

[Murphy, 1972], [DeGroot and Fienberg, 1983] and [Bröcker, 2009] suggested the "calibration-refinement" decomposition,

Lemma (Adapted from [Bröcker, 2009])

The expected loss corresponding to any proper scoring rule is the sum of expected divergence of $\widehat{S}$ from C and the expected divergence of C from Y, denoted

$$\mathbb{E}[d(\widehat{S}, Y)] = \underbrace{\mathbb{E}[d(\widehat{S}, C)]}_{\text{calibration loss}} + \underbrace{\mathbb{E}[d(C, Y)]}_{\text{refinement loss}}.$$

Calibration: Decomposition

Here, the calibration loss is due to the difference between the model output score $\widehat{S}$ and the proportion of 1's among instances with the same output.

That's probably the easy one...

An alternative decomposition can be considered

Lemma (Adapted from [Kull and Flach, 2015])

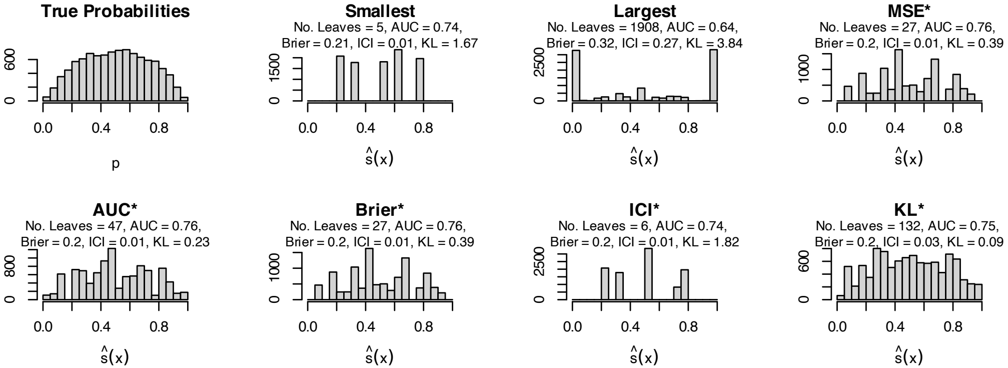
The expected loss corresponding to any proper scoring rule is the sum of expected divergence of $\widehat{S}$ from M and the expected divergence of M from Y,

$$\mathbb{E}[d(\widehat{S},Y)] = \underbrace{\mathbb{E}[d(\widehat{S},M)]}_{\text{epistemic loss}} + \underbrace{\mathbb{E}[d(M,Y)]}_{\text{irreducible loss}}.$$

More complicated. Need either expertise, or prior belief on the distribution of $\mu(\boldsymbol{X})$.

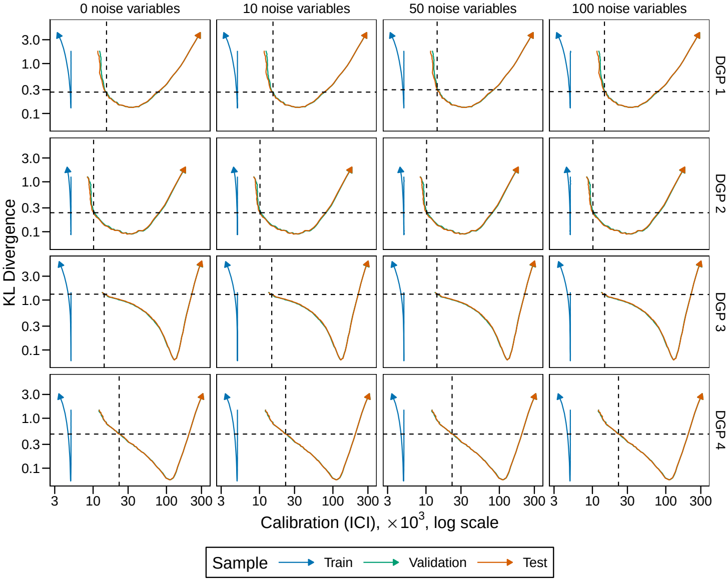
Calibration: Decomposition

Obviously, **calibration is not enough**, see trees,



from [Fernandes Machado et al., 2024c], on simulated data. See also evolution of ICI and KL as a function of the tree depth, (4 generating processes).

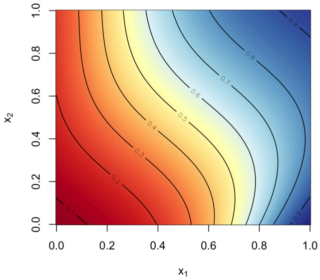
Calibration: Decomposition

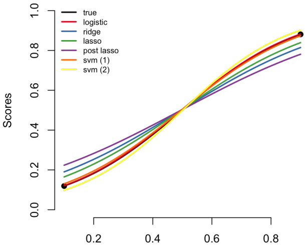


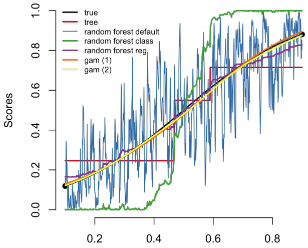
Simulations

plain linear — nonlinear component

$$(y_i, x_{1,i}, x_{2,i}), \text{ where } \mu(x_1,x_2) = \frac{\exp[\,x_1+x_2 + \psi(x_1,x_2)\,]}{1+\exp[\,x_1+x_2 + \psi(x_1,x_2)\,]},$$







Evolution of $x \mapsto \widehat{s}(x,x)$, on the diagonal.

Simulations

$$(y_i, x_{1,i}, x_{2,i}), \text{ where } \mu(x_1,x_2) = \frac{\exp[x_1+x_2+\psi(x_1,x_2)]}{1+\exp[x_1+x_2+\psi(x_1,x_2)]}, \text{ on the training dataset}$$

	AUC	Brier	ICI	KL	KS	
(plain) Logistic	0.761	0.199	**0.011**	**0.005**	**0.026**	➔
Logistic (ridge)	0.761	0.201	0.043	0.127	0.109	➔
Logistic (lasso)	0.761	0.200	0.025	0.057	0.074	➔
Logistic post lasso	0.733	0.209	**0.014**	0.054	0.084	➔
Linear discriminant analysis	0.761	0.199	**0.013**	**0.005**	**0.019**	➔
SVM (1)	0.760	0.199	**0.012**	**0.006**	**0.022**	➔
SVM (2)	0.760	0.200	0.029	0.033	0.053	
Logistic categorical	0.744	0.202	0.017	**0.592**	0.157	➔
Classification tree (1)	0.721	0.208	**0.005**	**0.708**	**0.250**	➔
Classification tree (2)	0.748	0.201	**0.006**	0.229	0.172	➔
Random Forest (default)	**1.000**	**0.031**	**0.145**	**1.144**	**0.247**	➔
Random Forest (classification)	0.769	0.243	**0.196**	**1.949**	**0.362**	➔
Random Forest (regression)	0.771	0.196	0.021	0.075	0.087	➔
GAM	0.763	0.198	**0.014**	**0.007**	**0.025**	➔
Bivariate spline	0.765	0.197	**0.007**	**0.002**	**0.017**	➔

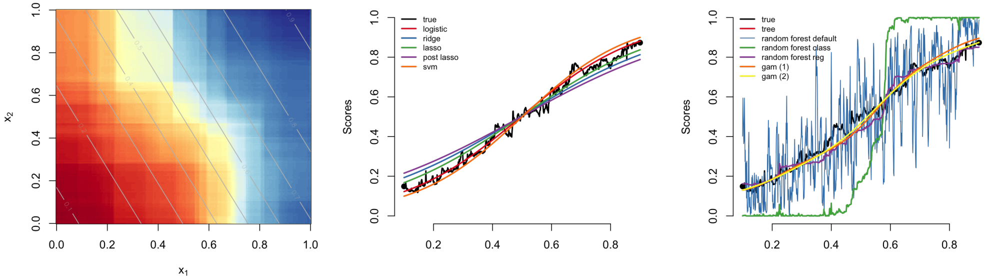
Simulations

$$(y_i, x_{1,i}, x_{2,i}), \text{ where } \mu(x_1, x_2) = \frac{\exp[x_1 + x_2 + \psi(x_1, x_2)]}{1 + \exp[x_1 + x_2 + \psi(x_1, x_2)]}, \text{ on the validation dataset}$$

	AUC	Brier	ICI	KL	KS	
(plain) Logistic	0.760	0.199	0.014	**0.004**	**0.026**	➔
Logistic (ridge)	0.760	0.202	0.042	0.125	0.110	➔
Logistic (lasso)	0.760	0.200	0.025	0.056	0.075	➔
Logistic post lasso	0.730	0.210	0.014	0.052	0.084	➔
Linear discriminant analysis	0.760	0.199	0.016	**0.006**	**0.024**	➔
SVM (1)	0.759	0.200	0.015	**0.005**	**0.024**	➔
SVM (2)	0.758	0.201	0.031	0.036	0.059	
Logistic categorical	0.739	0.204	0.016	**0.591**	0.158	➔
Classification tree (1)	0.716	0.210	**0.008**	**0.706**	**0.250**	➔
Classification tree (2)	0.741	0.204	0.013	0.233	0.172	➔
Random Forest (default)	**0.711**	0.228	**0.097**	0.307	0.100	➔
Random Forest (classification)	0.717	0.223	0.080	0.227	0.082	➔
Random Forest (regression)	0.762	0.199	0.019	0.081	0.088	➔
GAM	0.762	0.199	0.015	0.009	**0.025**	➔
Bivariate spline	**0.764**	0.198	**0.010**	**0.002**	**0.018**	➔

Simulations

$(y_i, x_{1,i}, x_{2,i})$, where $\mu(x_1, x_2) \leftarrow$ random forest , on the validation dataset



Evolution of $x \mapsto \widehat{s}(x, x)$, on the diagonal.

Simulations

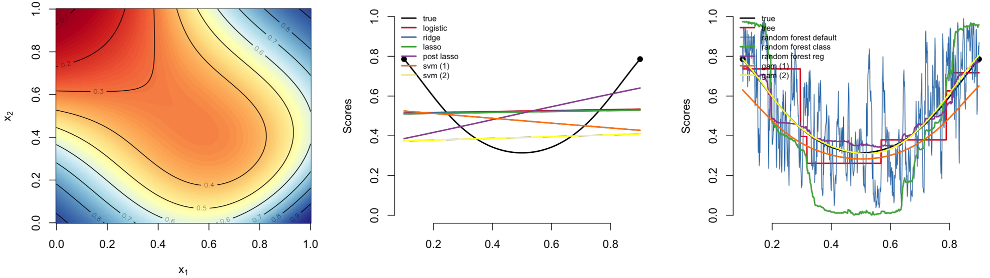
$(y_i, x_{1,i}, x_{2,i})$, where $\mu(x_1, x_2) \leftarrow$ random forest , on the validation dataset

	AUC	Brier	ICI	KL	KS
(plain) Logistic	**0.769**	0.196	**0.014**	**0.006**	**0.030**
Logistic (ridge)	**0.769**	0.199	0.048	0.119	0.111
Logistic (lasso)	0.768	0.197	0.029	0.052	0.076
Logistic post lasso	0.743	0.206	**0.013**	0.036	0.067
Linear discriminant analysis	0.768	0.196	0.015	**0.011**	**0.032**
SVM	0.767	0.196	**0.012**	**0.009**	**0.022**
Logistic categorical	0.750	0.200	0.021	0.384	0.139
Classification tree (1)	0.713	0.209	**0.008**	**0.814**	**0.267**
Classification tree (2)	0.746	0.200	**0.005**	0.449	**0.180**
Random Forest (default)	**0.708**	**0.271**	**0.232**	**1.193**	**0.248**
Random Forest (classification)	0.711	0.246	**0.202**	**0.678**	**0.182**
Random Forest (regression)	**0.771**	0.195	0.016	0.054	0.073
GAM	**0.770**	0.195	0.016	**0.011**	**0.025**
Bivariate spline	**0.772**	0.194	**0.009**	**0.008**	**0.014**

Simulations

quadratic (non monotonic)

$$(y_i, x_{1,i}, x_{2,i}),\text{ where } \mu(x_1,x_2) = \frac{\exp[\,(x_1+x_2-1)^2 + \psi(x_1,x_2)\,]}{1+\exp[\,(x_1+x_2-1)^2 + \psi(x_1,x_2)\,]},$$



Evolution of $x \mapsto \widehat{s}(x,x)$, on the diagonal.

Simulations

$$(y_i, x_{1,i}, x_{2,i}), \text{ where } \mu(x_1, x_2) = \frac{\exp[(x_1 + x_2 - 1)^2 + \psi(x_1, x_2)]}{1 + \exp[(x_1 + x_2 - 1)^2 + \psi(x_1, x_2)]}, \text{ validation dataset}$$

	AUC	Brier	ICI	KL	KS
(plain) Logistic	0.667	0.228	0.112	1.966	0.154
Logistic (ridge)	0.667	0.228	0.112	2.125	0.199
Logistic (lasso)	0.667	0.229	0.113	2.281	0.235
Logistic post lasso	**0.615**	0.239	0.077	2.673	**0.338**
Linear discriminant analysis	0.667	0.228	0.112	1.943	0.149
Logistic categorical	0.700	0.221	0.084	2.613	0.240
Classification tree (1)	0.703	0.212	**0.013**	**3.265**	0.256
Classification tree (2)	0.726	0.207	**0.009**	**3.348**	0.190
Random Forest (default)	0.720	0.224	0.090	-	0.106
Random Forest (classification)	0.754	0.242	**0.193**	-	**0.342**
Random Forest (regression)	**0.767**	0.198	0.028	2.118	0.141
GAM	0.716	0.214	0.053	1.927	0.163
Bivariate spline	**0.771**	0.195	**0.008**	1.919	**0.024**

Correction and Recalibration

Following [Denuit et al., 2019, Denuit et al., 2021], one can use the **local regression** estimate of g

$$\widehat{s}^{\star}_{\text{bc}}(\boldsymbol{X}) = \mathbb{E}\big[Y \mid \widehat{s}(\boldsymbol{X})\big] = g\big(\widehat{s}(\boldsymbol{X})\big) \text{ and } \widetilde{s}^{\star}_{\text{bc}}(\boldsymbol{X}) = \widehat{g}\big(\widehat{s}(\boldsymbol{X})\big).$$

This can be related to [Niculescu-Mizil and Caruana, 2005], that suggested to use an **isotonic regression**

$$\widetilde{s}^{\star}_{\text{iso}}(\boldsymbol{X}) = \widetilde{g}\big(\widehat{s}(\boldsymbol{X})\big).$$

A popular alternative is **Platt scaling**, from [Platt, 1999], obtained from a logistic regression

$$\widetilde{s}^{\star}_{\text{platt}}(\boldsymbol{X}) = \frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_1\widehat{s}(\boldsymbol{X})]}{1 + \exp[\widehat{\beta}_0 + \widehat{\beta}_1\widehat{s}(\boldsymbol{X})]} = \text{logit}\big(\widehat{\beta}_0 + \widehat{\beta}_1\widehat{s}(\boldsymbol{X})\big).$$

but one could also consider a **Beta-calibration** technique, as in [Kull et al.,].

Correction and Recalibration

In the a **Beta-calibration** technique, as in [Kull et al.,], suppose, $\widehat{s}(\boldsymbol{X})$ conditional on y is Beta distributed, $\mathcal{B}(a_y, b_y)$,

$$g(s)=\frac{\gamma s^{\alpha}(1-s)^{\beta}}{1+\gamma s^{\alpha}(1-s)^{\beta}},\text{ where }\gamma=\frac{B(a_0,b_0)}{B(a_1,b_1)},\ \begin{cases}\alpha=a_1-a_0\\ \beta=b_1-b_0\end{cases}$$

From [Denuit and Trufin, 2024], Proposition 4.5, for all Bregman loss functions ℓ, with φ convex, $\ell(y,\widehat{y})=\varphi(y)-\varphi(\widehat{y})-\varphi'(\widehat{y})\cdot[y-\widehat{y}]$, then

$$\mathbb{E}[\ell(Y,\mu(\boldsymbol{X}))]\leq\mathbb{E}[\ell(Y,\widehat{s}^{\star}_{\text{bc}}(\boldsymbol{X}))]\leq\mathbb{E}[\ell(Y,\widehat{s}(\boldsymbol{X}))],\ \forall\widehat{s}.$$

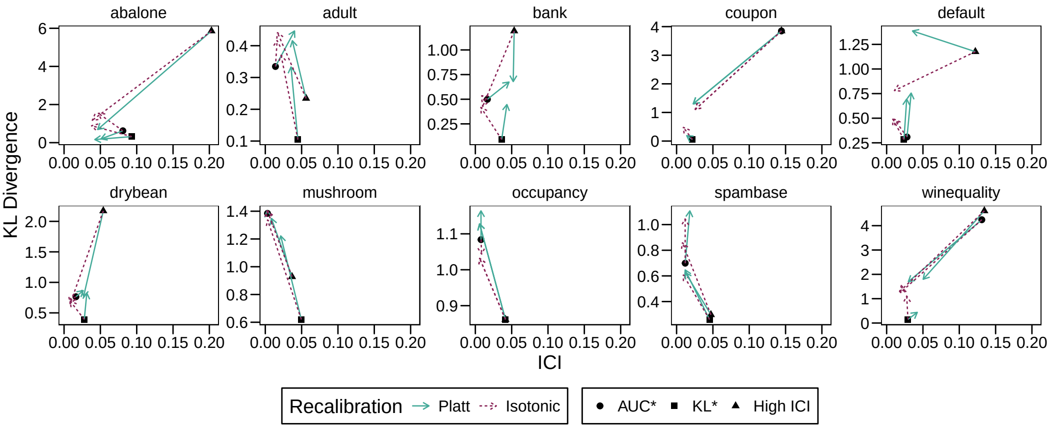
Data, real data

Table 1: Key characteristics of the datasets

Dataset	n	No. predictors	Prop. 1's	Reference	License
abalone	4,177	8	0.37	[Nash et al., 1995]	CC BY 4.0
adult	32,561	14	0.24	[Becker and Kohavi, 1996]	CC BY 4.0
bank	45,211	16	0.12	[Moro et al., 2012]	CC BY 4.0
default	30,000	23	0.22	[Yeh, 2016]	CC BY 4.0
drybean	13,611	16	0.26	[Koklu and Ali Ozkan, 2020]	CC BY 4.0
coupon	12,079	22	0.57	[Wang et al., 2020]	CC BY 4.0
mushroom	8,124	21	0.52	[Schlimmer, 1987]	CC BY 4.0
occupancy	20,560	5	0.23	[Candanedo, 2016]	CC BY 4.0
winequality	6,495	12	0.63	[Cortez et al., 2009]	CC BY 4.0
spambase	4,601	57	0.39	[Hopkins et al., 1999]	CC BY 4.0

Data, real data

From [Fernandes Machado et al., 2024c], Kullback-Liebler divergence ($d(\widehat{S}, M)$) again ICI ($d(\widehat{S}, C)$), with prior beliefs for the distribution of M,



Wrap-up

- In binary problems ($y \in \{0,1\}$), classical approach in econometrics is to use a **logistic/probit regression** to approximate $\mu(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$,
- **Machine learning** techniques are usually seen as better since they have high predictive power, e.g. random forests
  - **accuracy** (AUC) is usually higher
  - those models are usually **not well-calibrated** $\mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p] \neq p$,
  - the **distribution** of $\widehat{s}(\boldsymbol{X})$ is quite different from the one of $\mu(\boldsymbol{X})$.
- In many application **well-calibration** is a desirable property
- Sometimes **recalibration** can be achieved
- In most applications, a well fitted GAM regression model has better properties than advanced machine learning model

References

Adragni, K. P. and Cook, R. D. (2009).
Sufficient dimension reduction and prediction in regression.
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4385–4405.

Arrow, K. J. (1963).
Uncertainty and the welfare economics of medical care.
The American Economic Review, 53(5):941–973.

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.

Baumann, J. and Loi, M. (2023).
Fairness and risk: an ethical argument for a group fairness definition insurers can use.
Philosophy & Technology, 36(3):45.

References

Becker, B. and Kohavi, R. (1996).
Adult.
UCI Machine Learning Repository.
doi:10.24432/C5XW20.

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

Bröcker, J. (2009).
Reliability, sufficiency, and the decomposition of proper scores.
Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 135(643):1512–1519.

Candanedo, L. (2016).
Occupancy Detection .
UCI Machine Learning Repository.
doi:10.24432/C5X01N.

References

Charpentier, A., Flachaire, E., and Ly, A. (2018).
Econometrics and machine learning.
Economie et Statistique, 505(1):147–169.

Cook, R. D. (2007).
Fisher lecture: Dimension reduction in regression.
Statistical Science.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009).
Wine Quality.
UCI Machine Learning Repository.
doi:10.24432/C56S3T.

DeGroot, M. H. and Fienberg, S. E. (1983).
The comparison and evaluation of forecasters.
Journal of the Royal Statistical Society: Series D (The Statistician), 32(1-2):12–22.

References

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.

Denuit, M., Sznajder, D., and Trufin, J. (2019).
Model selection based on lorenz and concentration curves, gini indices and convex order.
Insurance: Mathematics and Economics, 89:128–139.

Denuit, M. and Trufin, J. (2024).
Convex and lorenz orders under balance correction in nonlife insurance pricing: Review and new developments.
Insurance: Mathematics and Economics.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a).
From uncertainty to precision: Enhancing binary classifier performance through calibration.
arXiv preprint arXiv:2402.07790.

References

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b).
Post-calibration techniques: Balancing calibration and score distribution alignment.
Thirty-Eighth Annual Conference on Neural Information Processing Systems.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024c).
Probabilistic scores of classifiers, calibration is not enough.
arXiv preprint arXiv:2408.03421.

Friedman, J. H. (1998).
Data mining and statistics: What's the connection?
Computing science and statistics, 29(1):3–9.

Gourieroux, C. and Jasiak, J. (2015).
The econometrics of individual risk: credit, insurance, and marketing.
Princeton University Press.

References

Gourieroux, C. and Monfort, A. (1995).
Statistics and econometric models, volume 1.
Cambridge University Press.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017).
On calibration of modern neural networks.
In *International conference on machine learning*, pages 1321–1330. PMLR.

Gupta, K., Rahimi, A., Ajanthan, T., Sminchisescu, C., Mensink, T., and Hartley, R. I. (2021).
Calibration of neural networks using splines.
In *International Conference on Learning Representations (ICLR)*.

Haavelmo, T. (1944).
The probability approach in econometrics.
Econometrica: Journal of the Econometric Society, pages iii–115.

References

Hirano, K., Imbens, G. W., and Ridder, G. (2003).
Efficient estimation of average treatment effects using the estimated propensity score.
Econometrica, 71(4):1161–1189.

Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. (1999).
Spambase.
UCI Machine Learning Repository.
doi:10.24432/C53G6X.

Koklu, M. and Ali Ozkan, I. (2020).
Dry Bean.
UCI Machine Learning Repository.
doi:10.24432/C50S4B.

Kostiuk, P. F. (1990).
Compensating differentials for shift work.
Journal of political Economy, 98(5, Part 1):1054–1075.

References

Krüger, F. and Ziegel, J. F. (2021).
Generic conditions for forecast dominance.
Journal of Business & Economic Statistics, 39(4):972–983.

Kruskal, J. B. (1964).
Nonmetric multidimensional scaling: a numerical method.
Psychometrika, 29(2):115–129.

Kull, M., Filho, T. M. S., and Flach, P. (2017).
Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration.
Electronic Journal of Statistics, 11(2):5052 – 5080.

Kull, M. and Flach, P. (2015).
Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration.
In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pages 68–85. Springer.

References

Kull, M., Silva Filho, T., and Flach, P.
Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers.
In *Artificial intelligence and statistics*. PMLR.

Kumar, A., Liang, P. S., and Ma, T. (2019).
Verified uncertainty calibration.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lee, L.-F. (1979).
Identification and estimation in binary choice models with limited (censored) dependent variables.
Econometrica, pages 977–996.

Loader, C. (2006).
Local regression and likelihood.
Springer.

References

Maddala, G. (1983).
Limited-dependent and qualitative variables in econometrics, volume 149.
Cambridge University Press.

Morgan, M. S. (1990).
The history of econometric ideas.
Cambridge University Press.

Moro, S., Rita, P., and Cortez, P. (2012).
Bank Marketing.
UCI Machine Learning Repository.
doi: https://doi.org/10.24432/C5K306.

Müller, R., Kornblith, S., and Hinton, G. E. (2019).
When does label smoothing help?
Advances in neural information processing systems, 32.

References

Murphy, A. H. (1972).
Scalar and vector partitions of the probability score: Part i. two-state situation.
Journal of Applied Meteorology and Climatology, 11(2):273–282.

Nadaraya, E. A. (1964).
On estimating regression.
Theory of Probability & Its Applications, 9(1):141–142.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., and Ford, W. (1995).
Abalone.
UCI Machine Learning Repository.
doi:10.24432/C55C7W.

Niculescu-Mizil, A. and Caruana, R. (2005).
Predicting good probabilities with supervised learning.
In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

References

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015).
Obtaining well calibrated probabilities using bayesian binning.
Proceedings of the AAAI Conference on Artificial Intelligence, 29(1):2901–2907.

Platt, J. (1999).
Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.
Advances in large margin classifiers, 10(3):61–74.

Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. (2020).
Intra order-preserving functions for calibration of multi-class neural networks.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13456–13467. Curran Associates, Inc.

Robinson, C. and Tomes, N. (1984).
Union wage differentials in the public and private sectors: A simultaneous equations specification.
Journal of Labor Economics, 2(1):106–127.

References

Rosenbaum, P. R. and Rubin, D. B. (1983).
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984).
Reducing bias in observational studies using subclassification on the propensity score.
Journal of the American statistical Association, 79(387):516–524.

Sanders, F. (1963).
On subjective probability forecasting.
Journal of Applied Meteorology and Climatology, 2(2):191–201.

Schervish, M. J. (1989).
A General Method for Comparing Probability Assessors.
The Annals of Statistics, 17(4):1856–1879.

References

Schlimmer, J. (1987).
Mushroom.
UCI Machine Learning Repository.
doi:10.24432/C5959T.

Sollich, P. (1999).
Probabilistic methods for support vector machines.
Advances in neural information processing systems, 12.

Tinbergen, J. (1939).
Statistical Testing of Business-Cycle Theories.
Oxford University Press.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019).
Calibration: the achilles heel of predictive analytics.
BMC medicine, 17(1):1–7.

References

Wang, D.-B., Feng, L., and Zhang, M.-L. (2021).
Rethinking calibration of deep neural networks: Do not be afraid of overconfidence.
Advances in Neural Information Processing Systems, 34:11809–11820.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. (2020).
In-Vehicle Coupon Recommendation.
UCI Machine Learning Repository.
doi:10.24432/C5GS4P.

Watson, G. S. (1964).
Smooth regression analysis.
Sankhyā: The Indian Journal of Statistics, Series A, pages 359–372.

Watt, J., Borhani, R., and Katsaggelos, A. K. (2016).
Machine learning refined: Foundations, algorithms, and applications.
Cambridge University Press.

References

Wilks, D. S. (1990).
On the combination of forecast probabilities for consecutive precipitation periods.
Weather and Forecasting, 5(4):640–650.

Working, E. J. (1927).
What do statistical "demand curves" show?
The Quarterly Journal of Economics, 41(2):212–235.

Wüthrich, M. V. and Ziegel, J. (2024).
Isotonic recalibration under a low signal-to-noise ratio.
Scandinavian Actuarial Journal, 2024(3):279–299.

Yeh, I.-C. (2016).
Default of Credit Card Clients.
UCI Machine Learning Repository.
doi:10.24432/C55S3H.

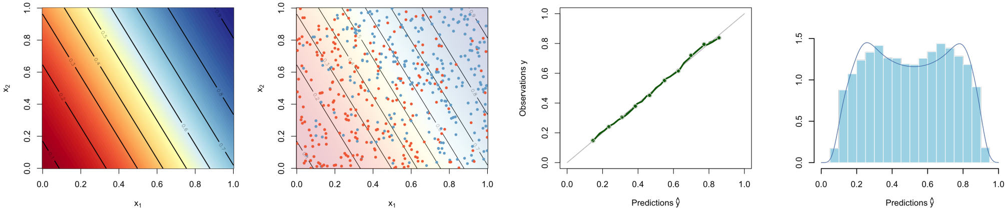
References

Yitzhaki, S. and Schechtman, E. (2013).
The Gini methodology: a primer on a statistical methodology, volume 272.
Springer.

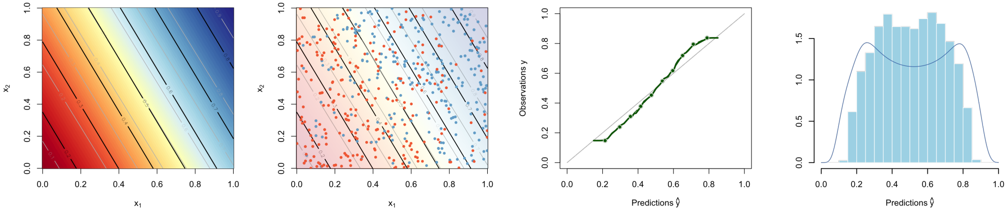
Zhang, J., Kailkhura, B., and Han, T. Y.-J. (2020).
Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning.
In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11117–11128. PMLR.

Simulated Nonlinear Logistic, training data

- (plain) Logistic

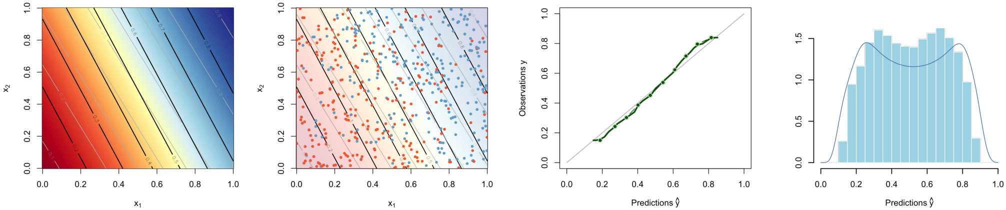


- Logistic with **Ridge** (ℓ_1 penalty)

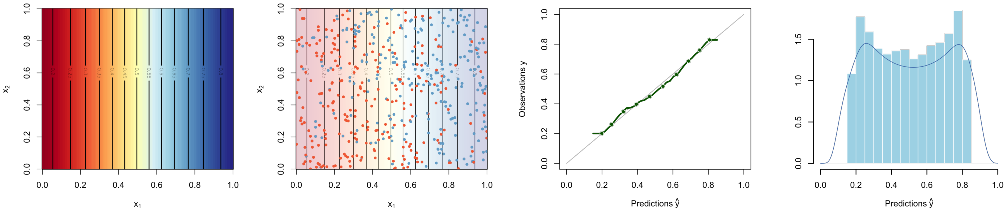


Simulated Nonlinear Logistic, training data

- Logistic with **lasso** (ℓ_2 penalty)

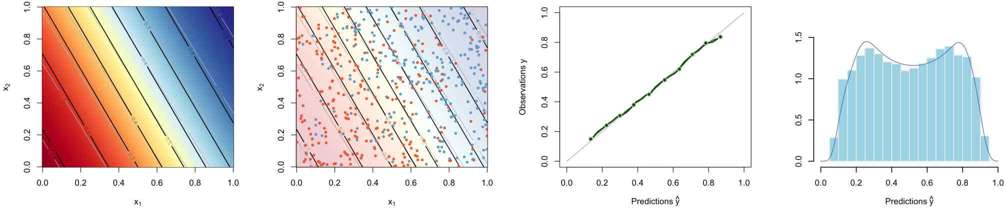


- Logistic with **post-lasso** (variable selection, here x_1)

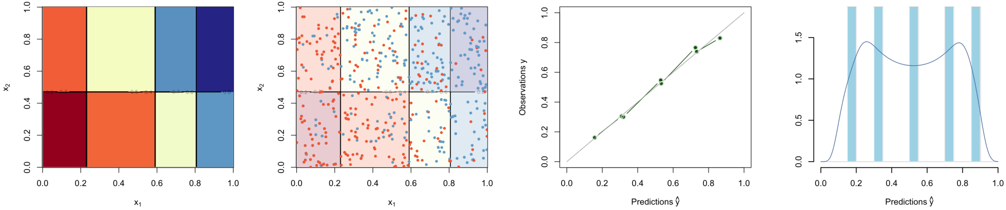


Simulated Nonlinear Logistic, training data

- Linear **discriminant analysis**

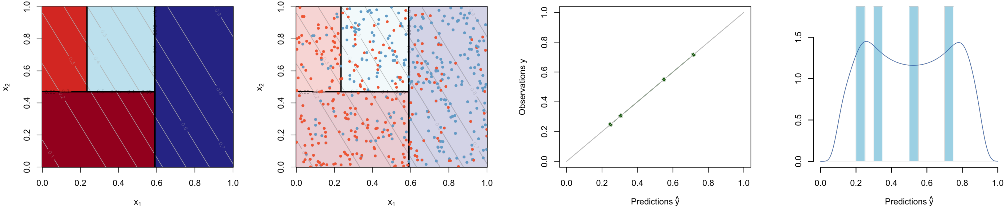


- Logistic with **categorical variables** (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)

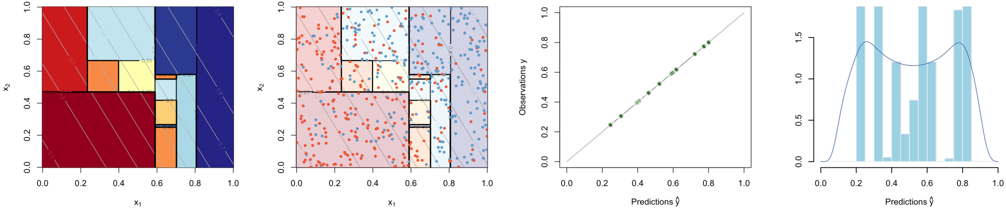


Simulated Nonlinear Logistic, training data

- **Classification Tree (1)**

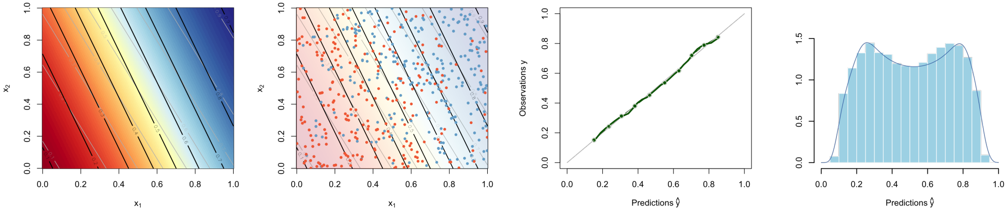


- **Classification Tree (2)**

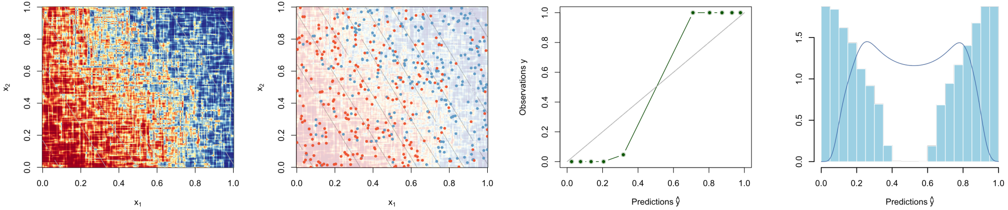


Simulated Nonlinear Logistic, training data

- **Support Vector Machine** (SVM) plain vanilla

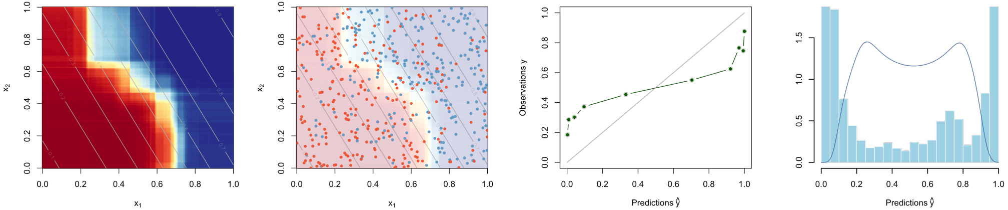


- **Classification Random Forest (default)**

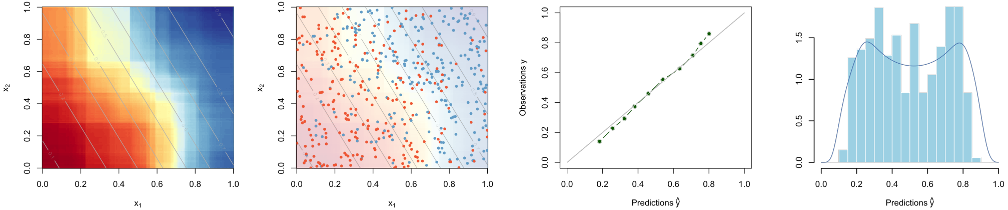


Simulated Nonlinear Logistic, training data

- **Classification Random Forest** with maximum nodes option

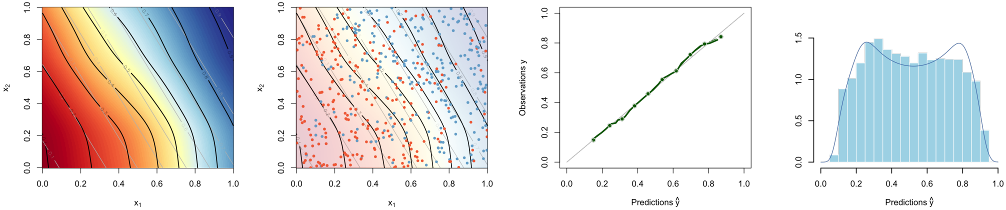


- **Regression Random Forest** with maximum nodes option

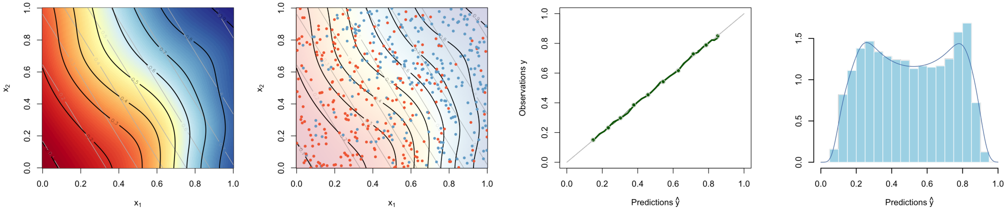


Simulated Nonlinear Logistic, training data

- Logistic **GAM** with additive splines

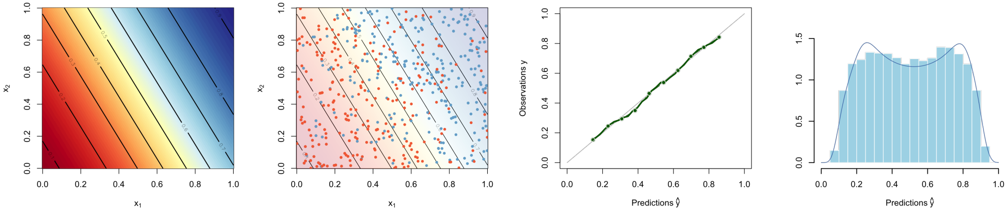


- Logistic **GAM** with bivariate splines

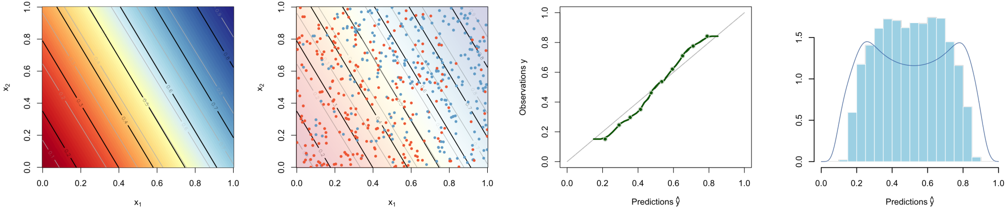


Simulated Nonlinear Logistic, validation data

- (plain) Logistic

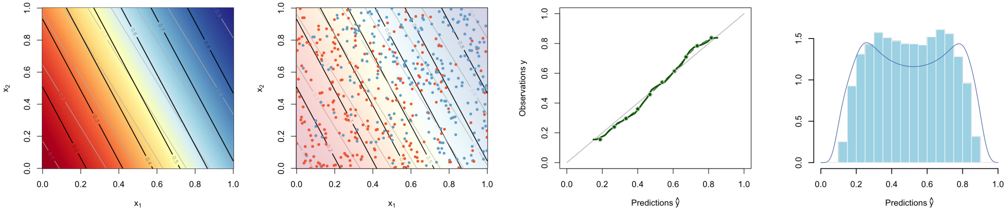


- Logistic with **Ridge** (ℓ_1 penalty)

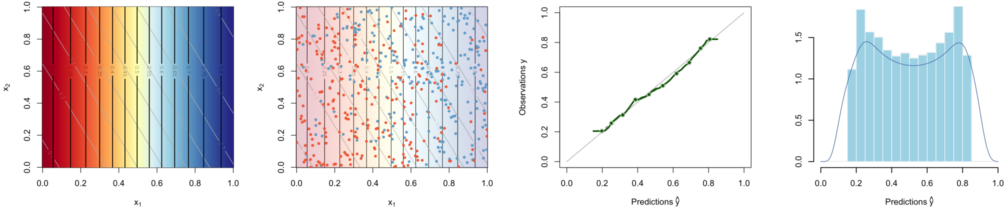


Simulated Nonlinear Logistic, validation data

- Logistic with **lasso** (ℓ_2 penalty)

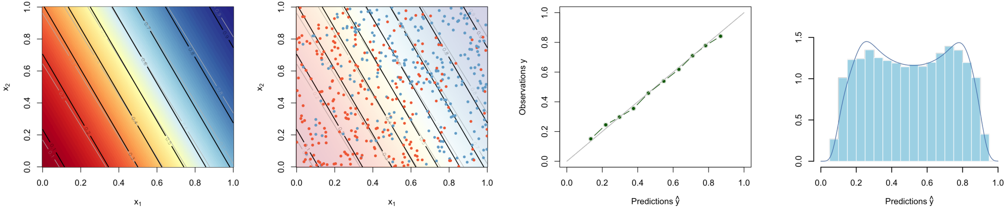


- Logistic with **post-lasso** (variable selection, here x_1)

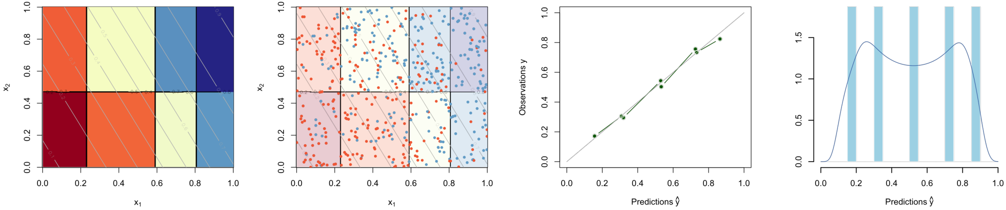


Simulated Nonlinear Logistic, validation data

- Linear **discriminant analysis**

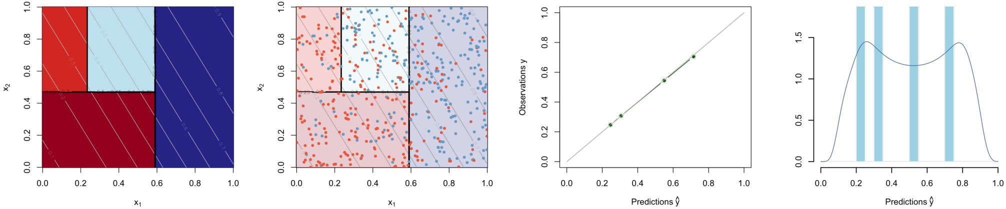


- Logistic with **categorical variables** (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)

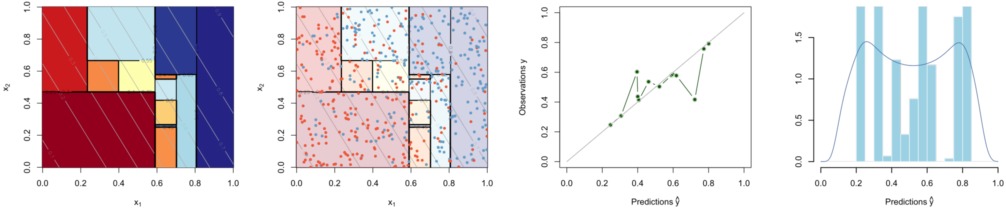


Simulated Nonlinear Logistic, validation data

- ## Classification Tree (1)

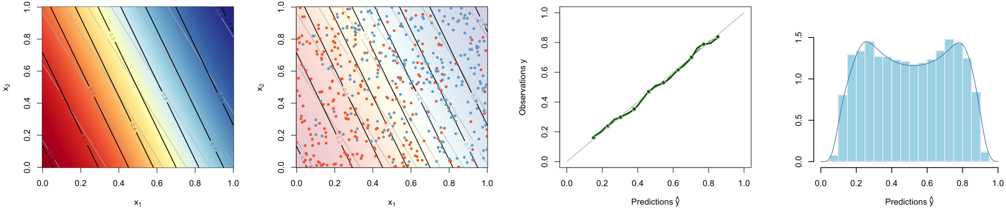


- ## Classification Tree (2)

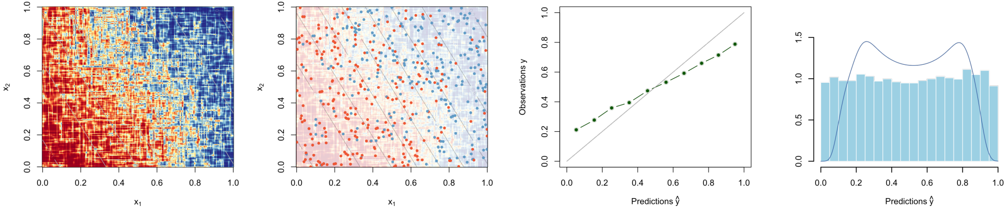


Simulated Nonlinear Logistic, validation data

- **Support Vector Machine** (SVM) plain vanilla

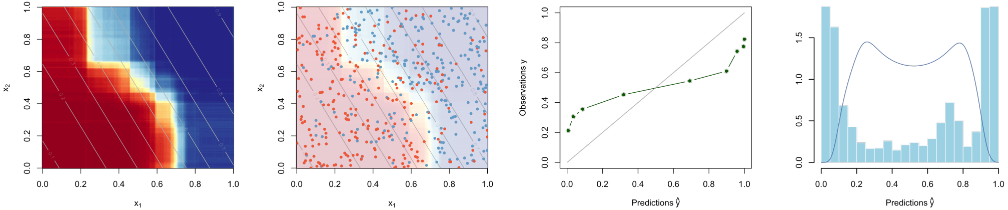


- **Classification Random Forest (default)**

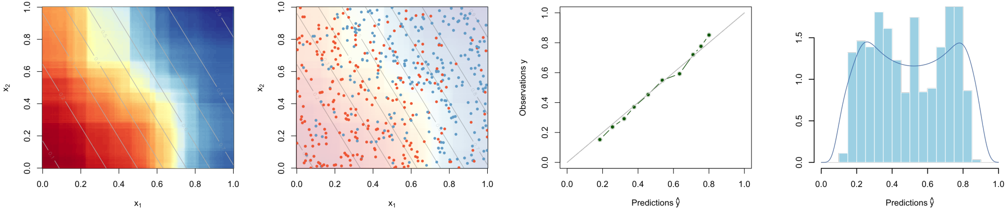


Simulated Nonlinear Logistic, validation data

- **Classification Random Forest** with maximum nodes option

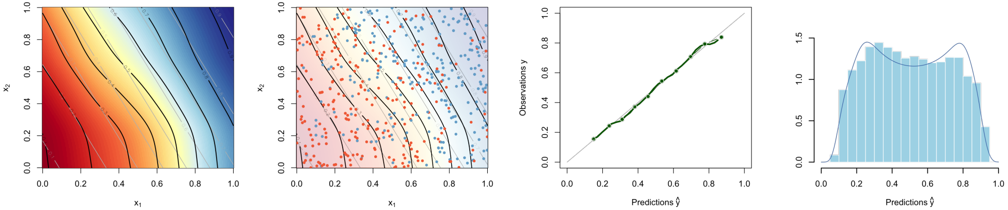


- **Regression Random Forest** with maximum nodes option

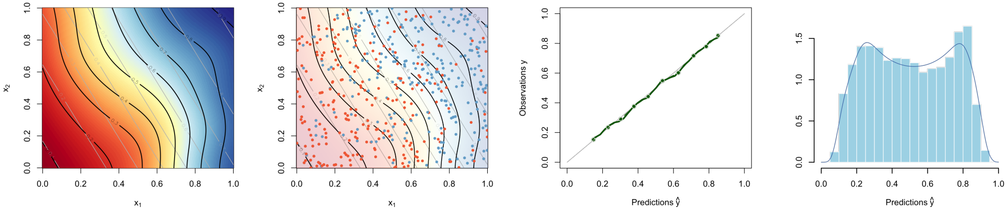


Simulated Nonlinear Logistic, validation data

- Logistic **GAM** with additive splines

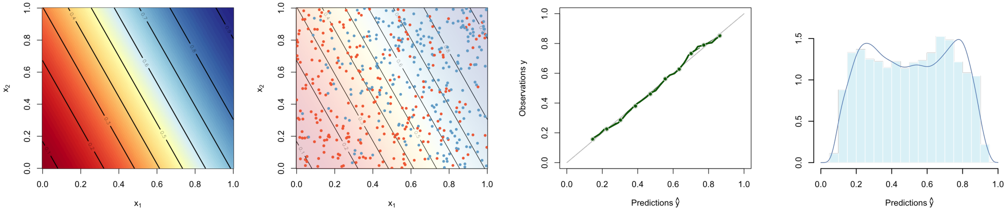


- Logistic **GAM** with bivariate splines

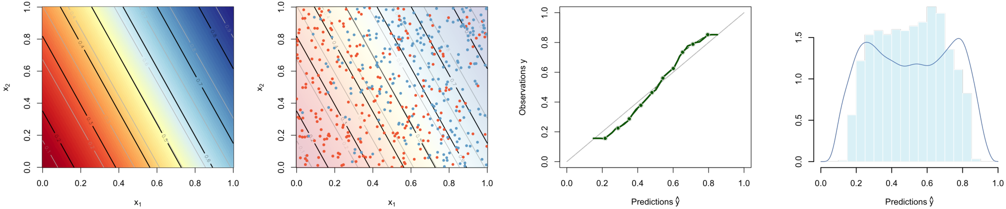


Simulated Random Forest, validation data

- (plain) Logistic

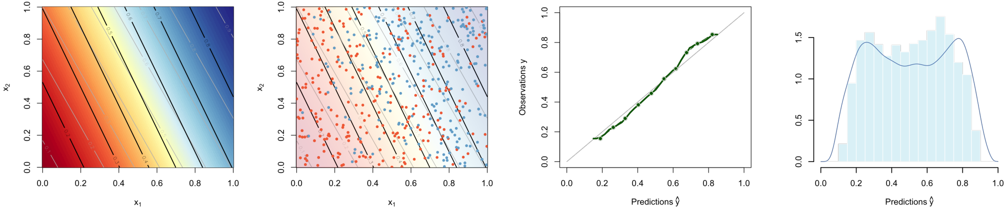


- Logistic with **Ridge** (ℓ_1 penalty)

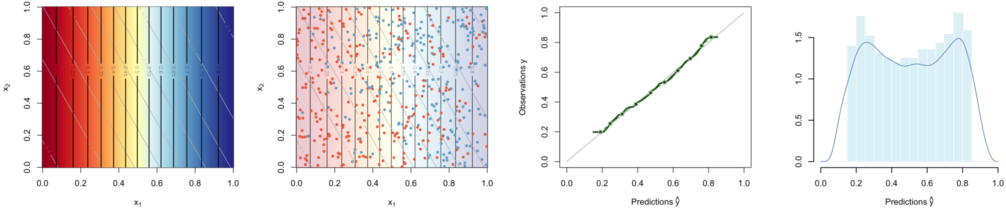


Simulated Random Forest, validation data

- Logistic with **lasso** (ℓ_2 penalty)

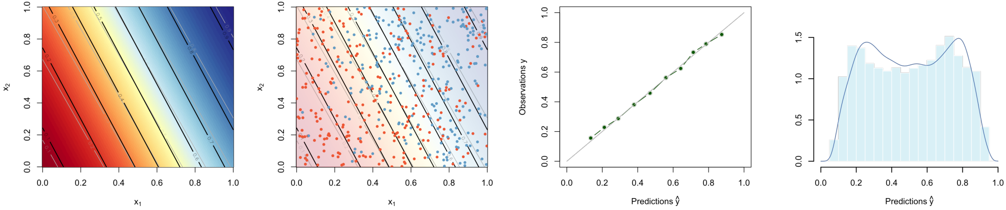


- Logistic with **post-lasso** (variable selection, here x_1)

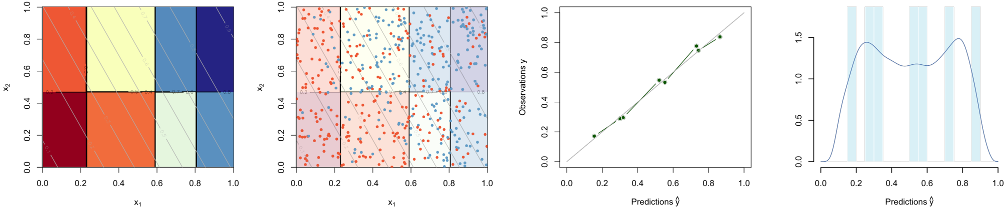


Simulated Random Forest, validation data

- Linear **discriminant analysis**

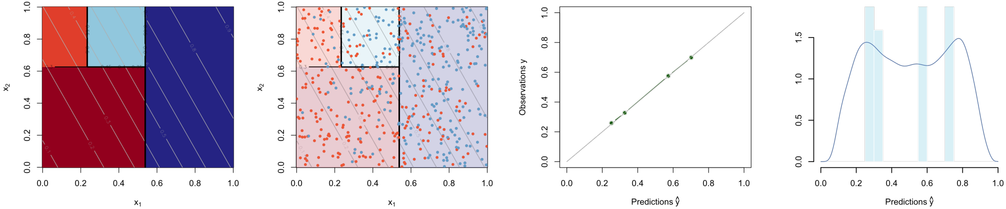


- Logistic with **categorical variables** (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)

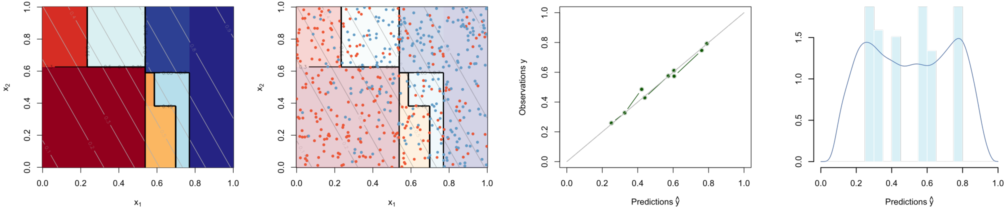


Simulated Random Forest, validation data

- **Classification Tree (1)**

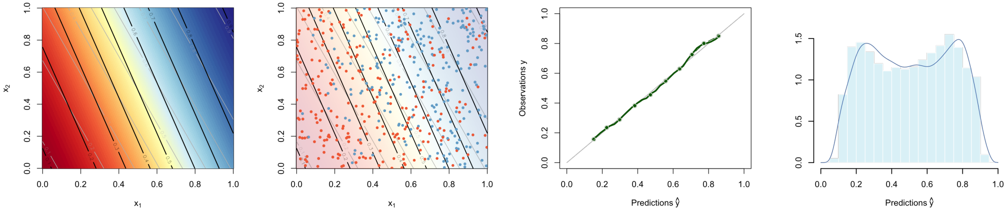


- **Classification Tree (2)**

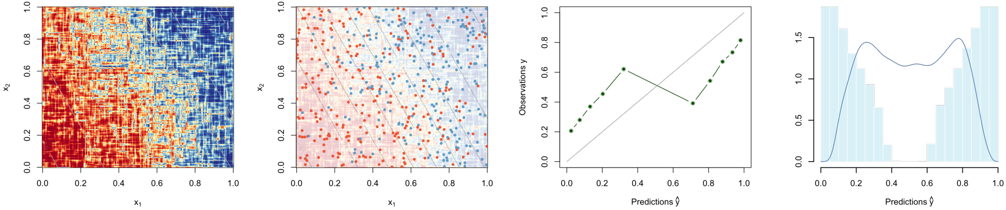


Simulated Random Forest, validation data

- **Support Vector Machine** (SVM) plain vanilla

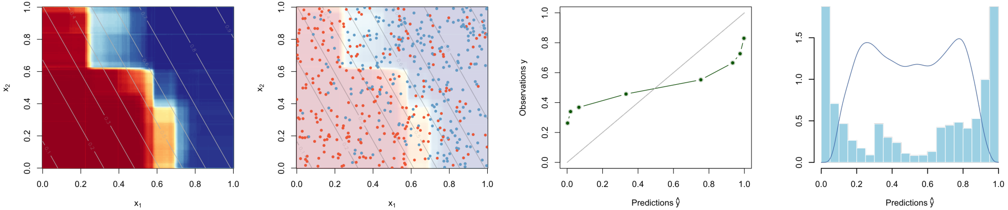


- **Classification Random Forest (default)**

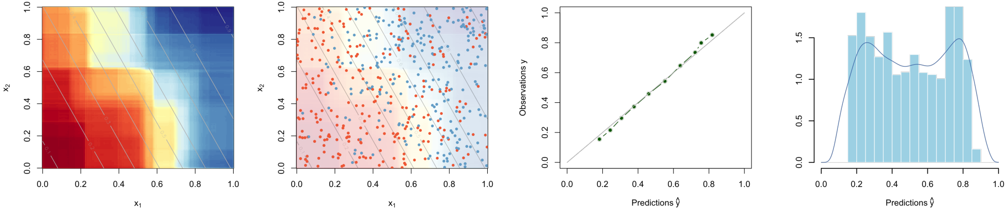


Simulated Random Forest, validation data

- **Classification Random Forest** with maximum nodes option

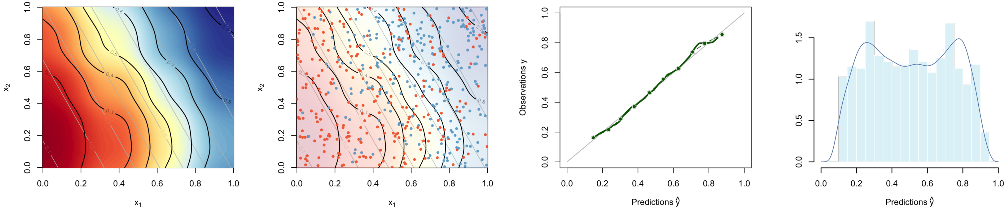


- **Regression Random Forest** with maximum nodes option

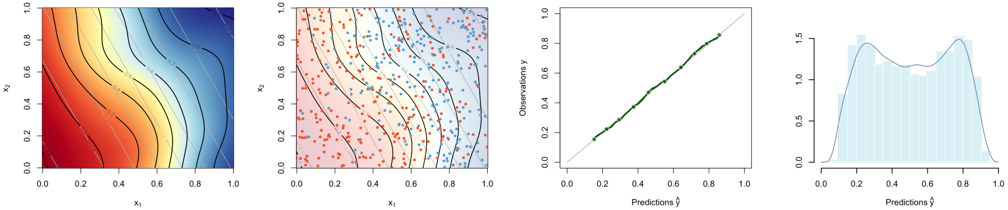


Simulated Random Forest, validation data

- Logistic **GAM** with additive splines

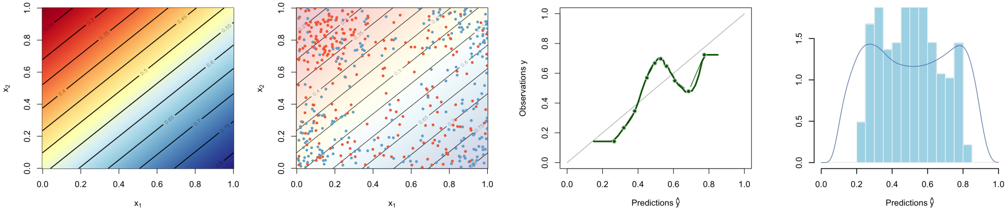


- Logistic **GAM** with bivariate splines

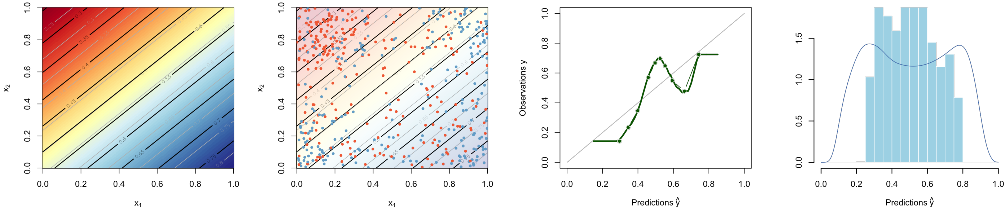


Simulated Non-monotonic Logistic, validation data

- (plain) Logistic

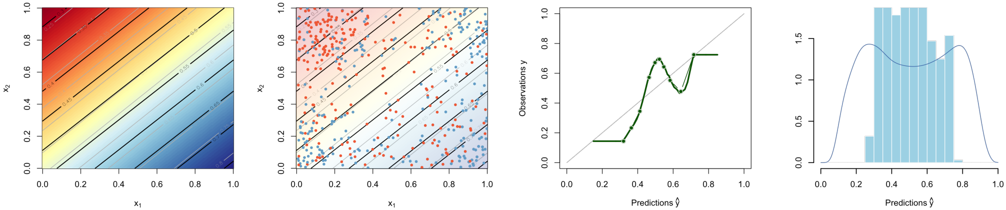


- Logistic with **Ridge** (ℓ_1 penalty)

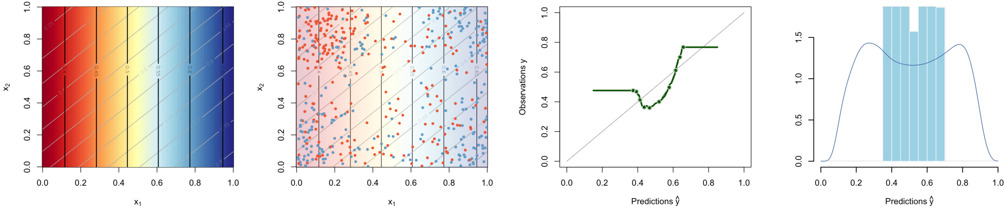


Simulated Non-monotonic Logistic, validation data

- Logistic with **lasso** (ℓ_2 penalty)

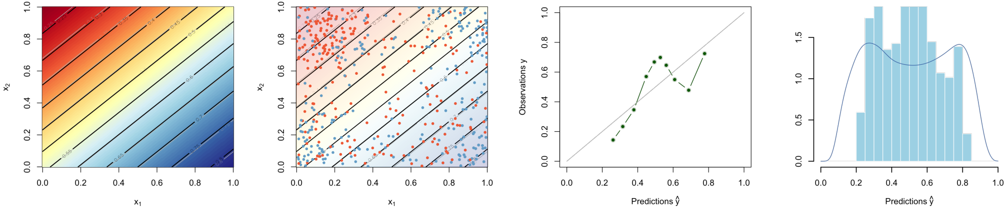


- Logistic with **post-lasso** (variable selection, here x_1)

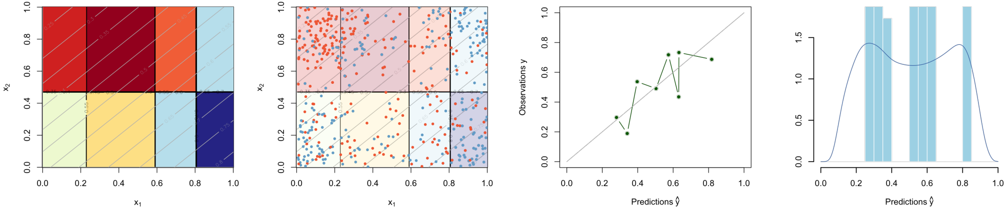


Simulated Non-monotonic Logistic, validation data

- Linear **discriminant analysis**

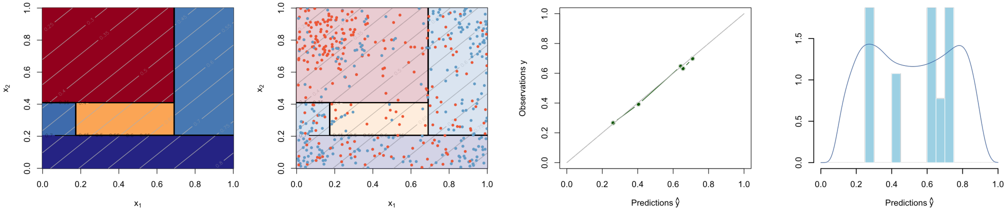


- Logistic with **categorical variables** (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)

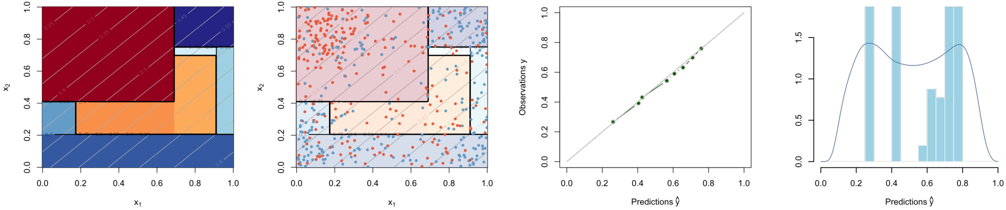


Simulated Non-monotonic Logistic, validation data

- **Classification Tree (1)**

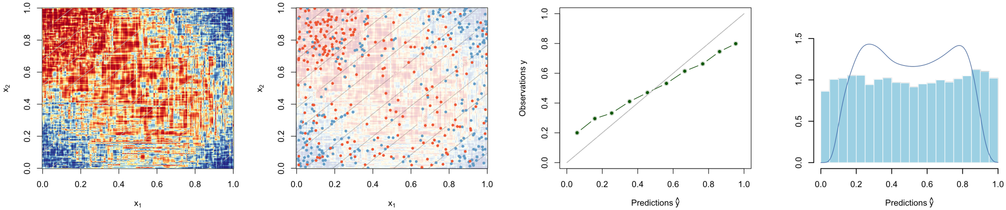


- **Classification Tree (2)**



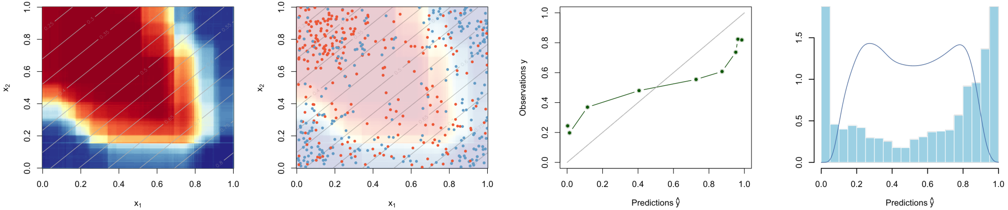
Simulated Non-monotonic Logistic, validation data

- **Classification Random Forest (default)**

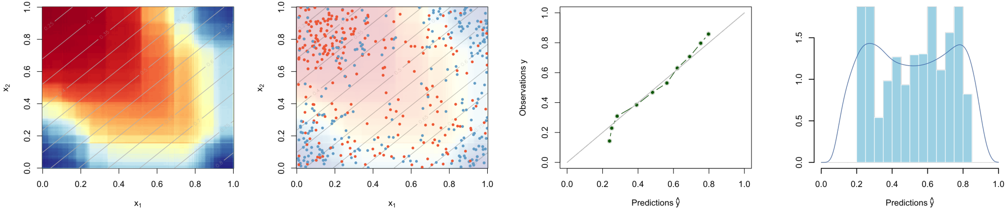


Simulated Non-monotonic Logistic, validation data

- **Classification Random Forest** with maximum nodes option

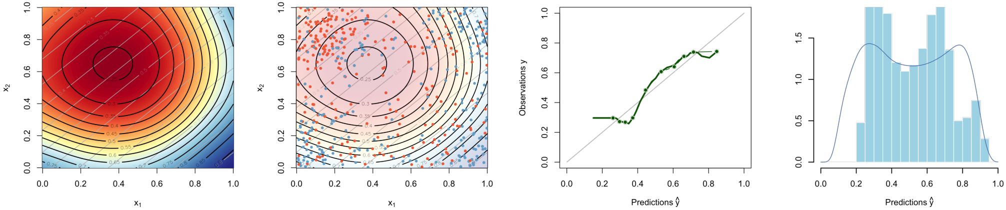


- **Regression Random Forest** with maximum nodes option



Simulated Non-monotonic Logistic, validation data

- Logistic **GAM** with additive splines



- Logistic **GAM** with bivariate splines