Solving the Measurement Problem in Machine Learning

Model Comparison and Calibration Assessment

Christian Lorentzen

26 August 2022

Measurement

Measure the right quantity / Ask the right question

Measurement

Measure the right quantity / Ask the right question

Time

Measurement

Measure the right quantity / Ask the right question

Time

Measurement

Measure the right quantity / Ask the right question

Time

Distance

Measurement

Measure the right quantity / Ask the right question

Time

Distance

Measurement

Measure the right quantity / Ask the right question

Time

Distance

Velocity

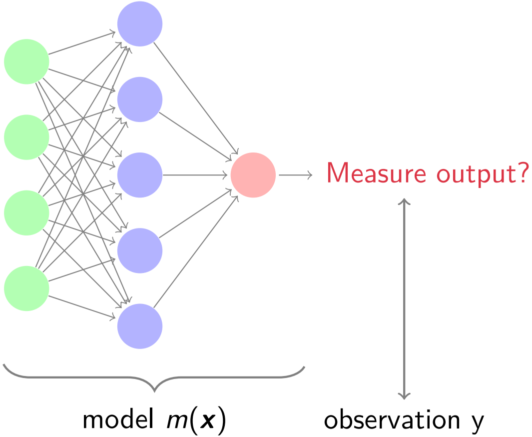
Measurement

Measure the right quantity / Ask the right question

Time

Distance

Velocity



Actuarial Models

Examples of widely applied actuarial models

- Pricing models for pure premium and profitability
- Reserving models for the ultimate claim costs
- Life tables
- NatCat models for annual losses
- Risk models for loss distribution of the company

Decisions are based on actuarial predictions.

Pursuit of Excellence

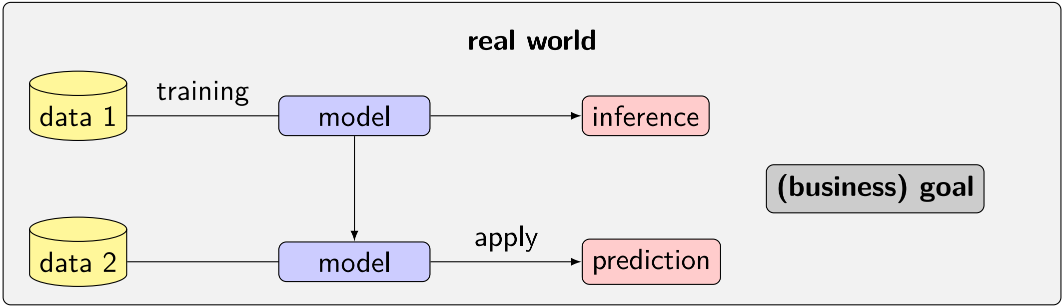
- **Find and use the *best* model among many.**
- **Assess if fit for production, e.g., bias under control.**
- Explain your model.

Possible use case

Use an XGBoost model instead of a GLM.

Outline

Picture of Machine Learning



Goal of a model

- inference—on observations/seen data
- **prediction**—on new, unseen data

Predictive Models

- features $\boldsymbol{X}$
- response variable Y

Remark

Y is random, there is no deterministic function $Y = g(\boldsymbol{X})$.

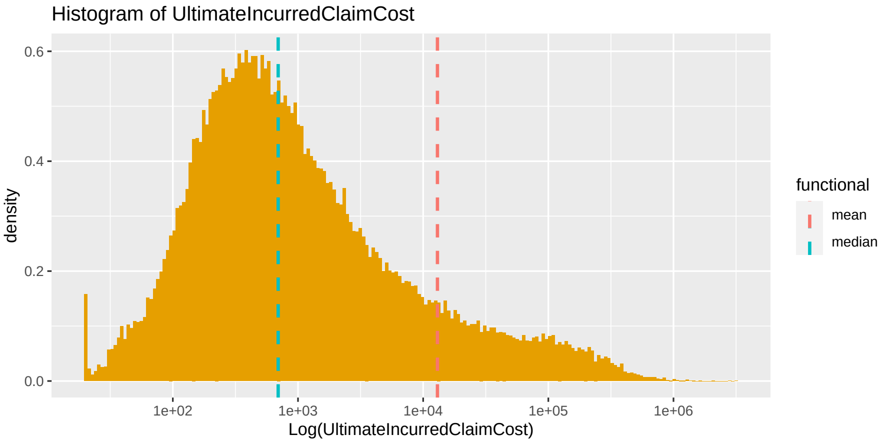
Prediction goals of model $m(\boldsymbol{X})$

- Probabilistic predictions aim for $F_{Y|\boldsymbol{X}}$.
- **Point predictions** aim for a property / (target) functional $T(F_{Y|\boldsymbol{X}})$.
  Convention: $T(F_{Y|\boldsymbol{X}}) = T(Y|\boldsymbol{X})$

Example

- expectation $T(Y|\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$
- median
- value at risk or α-quantile $T(Y|\boldsymbol{X}) = q_\alpha(Y|\boldsymbol{X}) = \inf\{t \in \mathbb{R} \mid F_{Y|\boldsymbol{X}}(t) \geq \alpha\}$
- expected shortfall

Workers Compensation Data Set



Model goal

Exectation

$\mathbb{E}[Y|\boldsymbol{X}]$

Workers Compensation data set `https://www.openml.org/d/42876`

y = UltimateIncurredClaimCost	InitialCaseEstimate	Age	Gender	WeeklyPay
102	9500	45	M	500
493	1000	18	F	373

Measuring Predictive Model Performance

Time

Measuring Predictive Model Performance

Time

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Measuring Predictive Model Performance

Time

Stricly Consistent Scoring Function S

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Measuring Predictive Performance

Scoring Functions

Measurement goal

Given a model $m(\boldsymbol{X})$ that predicts $T(Y|\boldsymbol{X})$ and observed input-output data $D = \{(\boldsymbol{x}_i, y_i), i = 1 \dots n\}$, how well does m perform?

Scoring (or loss) function

- A **scoring function** S measures the deviation of the model prediction $m(\boldsymbol{X})$ from T using observations Y by $S(m(\boldsymbol{X}), Y)$.
- Convention: The smaller S, the better.
- For model training as well as model comparison.

Example

- squared error $S(z, y) = (z - y)^2$
- absolute error $S(z, y) = |z - y|$

Iterative Optimisation (boosting, GD)

- $\overline{S}(m) = \frac{1}{n} \sum_i S(m(\boldsymbol{x}_i), y_i)$
- $m_{j+1} \approx \arg\min_{m \in \mathcal{M}} \underbrace{\overline{S}(m) - \overline{S}(m_j)}_{\text{model comparison}}$

Scores

Expected score / Statistical risk

We are interested in the expected score of model m (under distribution $F_{Y,\boldsymbol{X}}$):

$$R(m) = \mathbb{E}\left[S(m(\boldsymbol{X}), Y)\right] \tag{1}$$

Ideal model / Bayes rule

$$m^\star = \arg\min_{m \in \mathcal{M}} R(m) \tag{2}$$

Empirical score / risk

We estimate $R(m)$ by its empirical mean

$$\overline{R}(m; D) = \overline{S}(m; D) = \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} S(m(\boldsymbol{x}_i), y_i) \tag{3}$$

Data Split: Use a sound **train-validation-test data split** for reliable results.

Why Consistency Matters?

How to align the scoring function S with the model goal $T(Y|\boldsymbol{X})$?

Consistency

- It ensures that we get what we want: $m^\star = T(Y|\boldsymbol{X})$.
  (at least in the large sample limit by a Law of Large Numbers argument)
- Imagine a repeated game where each forecaster gets penalty / loss $S(z, y)$.

Counter example: Use absolute error $|z - y|$ when we aim for the expectation $T = \mathbb{E}$.

Elicitability

- Tells us if there exists a consistent scoring function for the functional T.
- Model comparison and (partially) backtesting is pointless for non-elicitable T.

Counter examples: Variance (alone) and expected shortfall (alone) are not elicitable.

Note: The pairs $(\mathrm{mean}, \mathrm{variance})$ and $(\alpha\text{-quantile}, \alpha\text{-ES})$ are elicitable!

Which One to Choose?

Use a strictly consistent scoring function!

Result: There are infinitely many ones (for elicitable T).

Example: deviances of exponential dispersion family (squared error, Poisson, Gamma and Tweedie deviance) for $T(Y|\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$.

Further criteria

- Domain / Range of target Y.
- Degree of homogeneity: $S(tz, ty) = t^h S(z, y)$ for all $t > 0$ and for all z, y
- Efficiency: How fast is the large sample convergence?
- Forecast dominance: Is one model dominating for many/all scoring functions? Assess with Murphy diagrams.

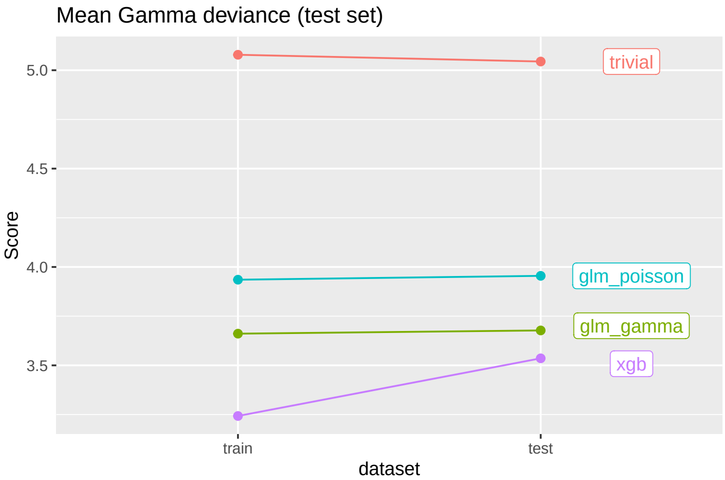
Squared error: $h = 2$

Gamma deviance:
Degree of homogeneity is $h = 0 \Rightarrow$ It only cares about relative differences:
$S(1, 10) = S(10, 100) = S(100, 1000) = 13.39$

Model Comparison

Compare empiricial mean scores: $\overline{S}(m) = \frac{1}{n}\sum_i S(m(\boldsymbol{x}_i), y_i)$

Gamma deviance for workers compensation

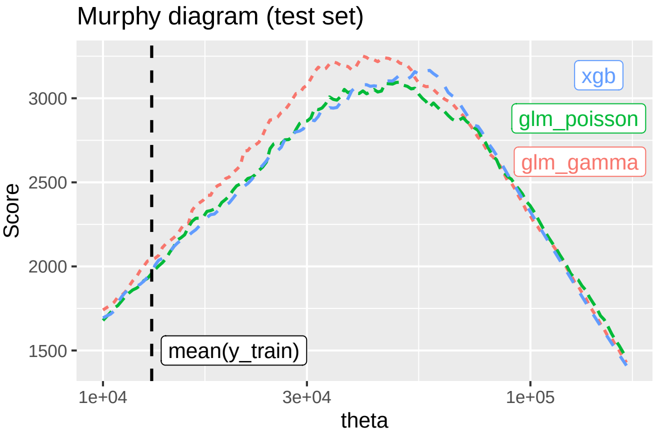


Models:

1. Trivial model always predicts mean(y) of the training set.
2. Poisson GLM with canonical log-link.
3. Gamma GLM with log-link.
4. XGBoost model with Gamma deviance and log-link.

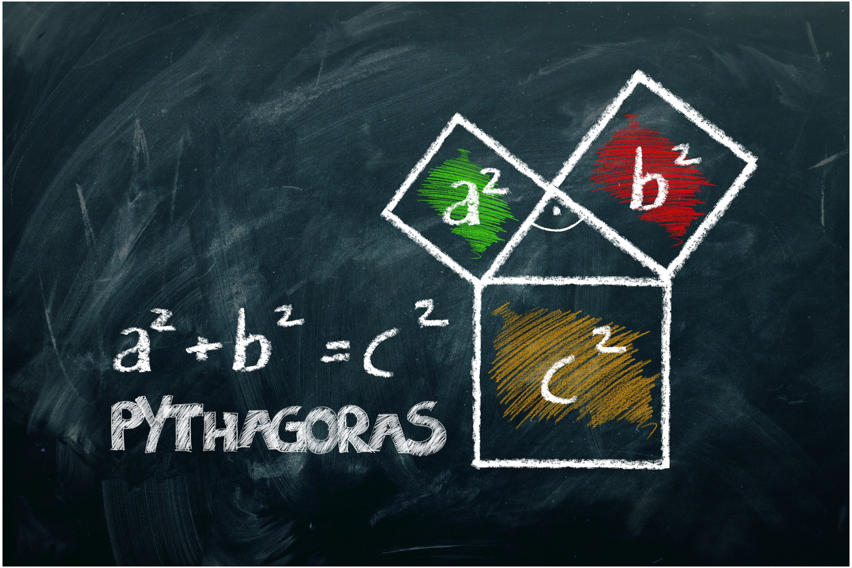
Murphy Diagram

Compare many scoring functions (sliding parameter θ) at once.
Forecast dominance: One model is better for all consistent scoring functions.



Elementary scoring function for $\mathbb{E}$: $S_\theta(z, y) = \frac{1}{2}|\theta - y|\mathbb{1}\{\min(z, y) \leq \theta < \max(z, y)\}$

Additive Score Decomposition



Score Decomposition

$$R(m) = \mathbb{E}[S(m(\boldsymbol{X}), Y)] = \text{miscalibration} - \text{resolution} + \text{uncertainty}$$

Score Decomposition

$$\mathbb{E}[S(m(\boldsymbol{X}),Y)] = \Big\{ \underbrace{\mathbb{E}[S(m(\boldsymbol{X}),Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})),Y)]}_{\text{auto-miscalibration}\geq 0} \Big\} - \Big\{ \underbrace{\mathbb{E}[S(T(Y),Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})),Y)]}_{\text{auto-resolution / auto-discrimination}\geq 0} \Big\} + \underbrace{\mathbb{E}[S(T(Y),Y)]}_{\text{uncertainty / entropy}} \quad (4)$$

Note:

Minimising consistent scores amounts to **jointly** minimising miscalibration and maximising resolution!

Squared Error / Brier Score

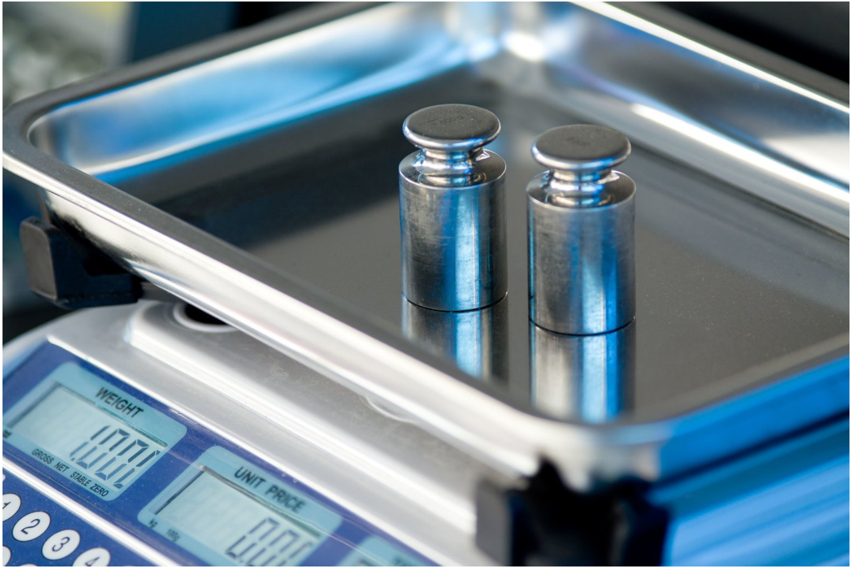
with $T(Y) = \mathbb{E}[Y]$ and $T(Y|m(\boldsymbol{X})) = \mathbb{E}[Y|m(\boldsymbol{X})]$

$$\mathbb{E}[(m(\boldsymbol{X}) - Y)^2] = \underbrace{\mathbb{E}[(m(\boldsymbol{X}) - \mathbb{E}[Y|m(\boldsymbol{X})])^2]}_{\text{auto-miscalibration}} - \underbrace{\mathrm{Var}[\mathbb{E}[Y|m(\boldsymbol{X})]]}_{\text{auto-resolution}} + \underbrace{\mathrm{Var}[Y]}_{\text{uncertainty}} \quad (5)$$

Score Decomposition of Gamma Deviance

Model	Mean deviance	Auto-miscalibration	Auto-resolution	Uncertainty
Trivial	5.04	0	0	5.04
GLM Gamma	3.68	0.190	1.56	5.04
GLM Poisson	3.95	0.482	1.57	5.04
XGB	**3.54**	**0.124**	**1.63**	5.04

Calibration



Motivation for Calibration

- Is the model fit for its prediction task?
- How well does the predictions align with observations?
- Detect bias and discrimination.

Bias can result in bad news.

Motivation for Calibration

- Is the model fit for its prediction task?
- How well does the predictions align with observations?
- Detect bias and discrimination.

Bias can result in bad news.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure: ProPublica article on COMPAS.

Calibration on Portfolio Level

Would we have made profit or loss (on test set)?

Note: Ideally neither loss nor profit, i.e. *balanced*.

$n$test $= 20504$

	$\frac{1}{n}\sum_i m(\boldsymbol{x}_i) - y_i$	p-value of t-test
Trivial	$\mathbf{-24}$	9.5×10^{-1}
GLM Gamma	-1207	8.8×10^{-4}
GLM Poisson	125	7.3×10^{-1}
XGBoost	-2044	1.4×10^{-8}

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y] \approx 0$

Calibration on Portfolio Level

Would we have made profit or loss (on test set)?

Note: Ideally neither loss nor profit, i.e. *balanced*.

$n\text{test} = 20504$

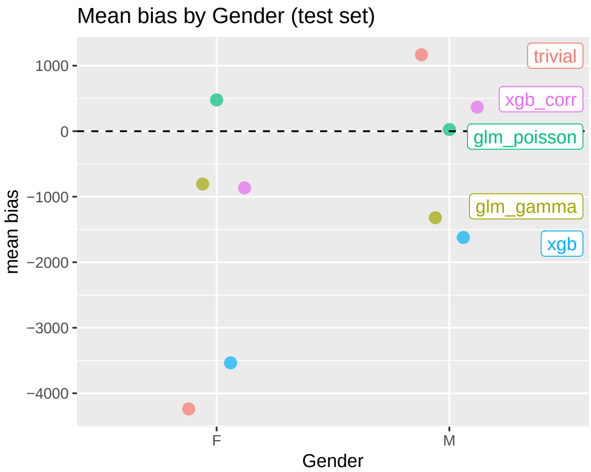
	$\frac{1}{n}\sum_i m(\boldsymbol{x}_i) - y_i$	p-value of t-test
Trivial	$\mathbf{-24}$	9.5×10^{-1}
GLM Gamma	-1207	8.8×10^{-4}
GLM Poisson	125	7.3×10^{-1}
XGBoost	-2044	1.4×10^{-8}
XGBoost corr	96	7.9×10^{-1}

Recalibrate XGBoost by a multiplicative constant (on training set).

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y] \approx 0$

Calibration Conditional on Gender

Is there a gender bias in the models?



model	$\frac{1}{n}\sum_{i\in subset} m(\boldsymbol{x})_i - y_i$ bias F	bias M
Trivial	−4240	1167
GLM Gamma	−807	−1320
GLM Poisson	**477**	**26**
XGBoost	−3536	−1623
XGBoost corr	−865	367

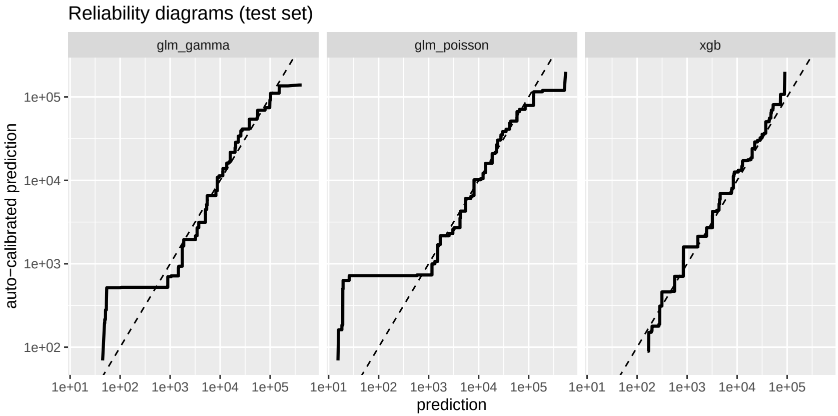
$\Rightarrow$ **conditional calibration:**

$$\mathbb{E}[m(\boldsymbol{X}) - Y|\boldsymbol{X}] \approx 0$$

Auto-Calibration

Are policies with same (actuarial) price self-financing?

Reliability diagram: Estimate $\mathbb{E}[Y|m(\boldsymbol{X})]$ via isotonic regression (PAV) and plot vs $m(\boldsymbol{X})$.



$\Rightarrow$ **auto-calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y|m(\boldsymbol{X})] \approx 0$

Measuring Model Calibration

Distance

Is $m(\boldsymbol{x})$ calibrated?

Measuring Model Calibration

Distance

Strict Identification Function V

Is $m(\boldsymbol{x})$ calibrated?

Assessing Calibration

Canonical identification function for the expectation: $V(z, y) = z - y$.

Notion	Definition	Check
conditional calibration	$m(\boldsymbol{X}) = T(Y \mid \boldsymbol{X})$	$\mathbb{E}[V(m(\boldsymbol{X}), Y) \mid \boldsymbol{X}] = 0 \quad a.s.$
auto-calibration	$m(\boldsymbol{X}) = T(Y \mid m(\boldsymbol{X}))$	$\mathbb{E}[V(m(\boldsymbol{X}), Y) \mid m(\boldsymbol{X})] = 0 \quad a.s.$
unconditional calibration	$\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$	$\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$

Table: Types of calibration for an identifiable functional T with strict identification function V.

- $V(m(\boldsymbol{x}_i), y_i)$ acts like a generalised residual.
- Conditional calibration is equivalent to $\mathbb{E}[\varphi(\boldsymbol{X}) V(m(\boldsymbol{X}), Y)] = 0$ for **all** (measurable) test functions $\varphi\colon \mathcal{X} \to \mathbb{R}$.
- Choose a φ and compute (and plot) $\overline{V}_\varphi(m) = \frac{1}{n} \sum_i \varphi(\boldsymbol{x}_i) V(m(\boldsymbol{x}_i), y_i)$.

Application

Transition from GLMs to modern ML models

- GLM acts as gold standard reference model.
- Ensure at least same predictive performance.
- Inspect calibration / bias.

Outlook

- Jointly model claim size below and above a threshold.[1]
- Think about long-tail claim reserves.

Personal insight

- Prefer good calibration over pure model performance.
- Don't be content with a single number/measure.
- Added value in bringing together multiple disciplines!

[1] T. Fissler, M. Merz, M. V. Wüthrich (2021). Deep Quantile and Deep Composite Model Regression. ArXiv:2112.03075.

Conclusion

> *A proper scoring rule is designed such that truth telling [...] is an optimal strategy in expectation. (Gneiting & Katzfuss, Annu. Rev. Stat. Appl. 2014. 1:125-51)*

- What is the model goal, what the prediction target?
- Strict identification functions assess model calibration (detect bias).
- Strictly consistent scoring (or loss) functions act as a "truth serum".

T. Fissler, C. Lorentzen & M. Mayer, (2022). Model Comparison and Calibration Assessment: User Guide for Consistent Scoring Functions in Machine Learning and Actuarial Practice. ArXiv:2202.12780.

Appendix

Binary Classification

$Y \in \{0, 1\}$

Probabilistic Classifier

- $p = \mathbb{P}(Y = 1|\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$
- Point prediction of the expectation is a fully probabilistic prediction.

Further consequences

- Prefer probabilistic classifiers (predict p) over deterministic ones (predict 0 or 1).
  $\Rightarrow$ More informative predictions, deliberate choice of a threshold t:
  $m(\boldsymbol{X}) \approx \mathbb{P}(Y = 1|\boldsymbol{X}) \geq t \Rightarrow$ decide for class $Y = 1$.
- Use a strictly consistent scoring function for the expectation.
  (and neither AUC nor accuracy)
- Scoring functions and scoring rules (for probabilistic predictions) coincide.

Reliability Diagram and Score Decomposition

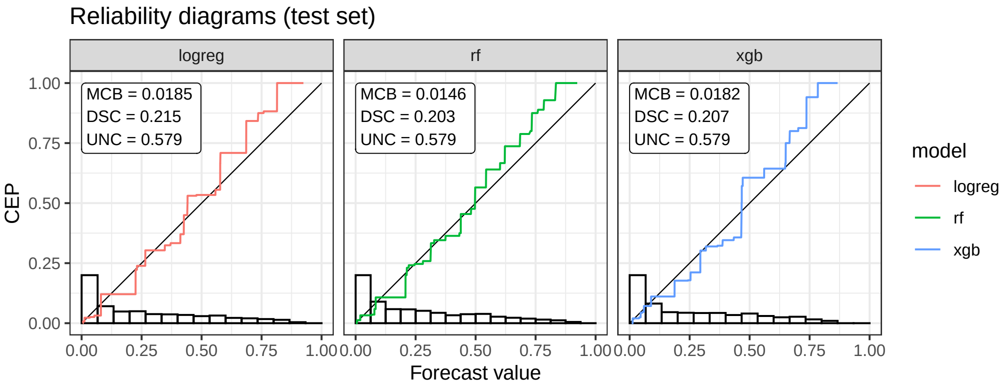


Figure: telco customer churn data set

Consistency & Elicatibility

Definition (Consistency)

Let $\mathcal{F}$ be a class of probability distributions where the functional T is defined on. A scoring function $S(z, y)$ is a function in a forecast z and an observation y. It is $\mathcal{F}$-**consistent** for T if

$$\int S(T(F), y)\,\mathrm{d}F(y) \leq \int S(z, y)\,\mathrm{d}F(y) \qquad \text{for all } z \in \mathbb{R},\ F \in \mathcal{F}. \tag{6}$$

The score is **strictly** $\mathcal{F}$-consistent for T if it is $\mathcal{F}$-consistent for T and if equality in (6) implies that $z = T(F)$.

Definition (Elicitability)

A functional T is **elicitable** on F if there is a strictly F-consistent scoring function for it.

Identification Functions

Definition

Let $\mathcal{F}$ be a class of probability distributions where the functional T is defined on. A **strict $\mathcal{F}$-identification function** for T is a function $V(z,y)$ in a forecast z and an observation y such that

$$\int V(z,y)\,\mathrm{d}F(y) = 0 \quad \Longleftrightarrow \quad z = T(F) \qquad \text{for all } z \in \mathbb{R},\ F \in \mathcal{F}\,. \tag{7}$$

If only the implication $\Longleftarrow$ in (7) holds, then V is just called an $\mathcal{F}$-identification function for T. If T admits a strict $\mathcal{F}$-identification function, it is **identifiable** on $\mathcal{F}$.

Identifiability $\Leftrightarrow$ elicitability (for 1-dim T and technical assumptions)

Canonical strict identification functions

Functional	Strict Identification Function	Domain of y, z
expectation $\mathbb{E}[Y]$	$V(z,y) = z - y$	$\mathbb{R}$
α-expectile	$V(z,y) = 2\lvert\mathbb{1}\{z \geq y\} - \alpha\rvert(z-y)$	$\mathbb{R}$
α-quantile $F_Y^{-1}(\alpha)$	$V(z,y) = \mathbb{1}\{z \geq y\} - \alpha$	$\mathbb{R}$

Identification Functions and Calibration

Let V be any strict $\mathcal{F}$-identification function for T.

Conditional calibration

Suppose that $\mathcal{F}$ contains the conditional distributions $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for almost all $\boldsymbol{x} \in \mathcal{X}$. Application of (7) to these conditional distributions yields that $m(\boldsymbol{x}) = T(Y|\boldsymbol{X} = \boldsymbol{x})$ if and only if $\int V(m(\boldsymbol{x}), y)\, \mathrm{d}F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = 0$. This shows that m is conditionally calibrated for T if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0 \qquad \text{almost surely.} \tag{8}$$

Auto-Calibration

Suppose the conditional distributions $F_{Y|m(\boldsymbol{X})=z}$ are in $\mathcal{F}$ for almost all $z \in \mathbb{R}$. Then m is auto-calibrated for T if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|m(\boldsymbol{X})] = 0 \qquad \text{almost surely.} \tag{9}$$

Note

By the tower property of the conditional calibration, conditional calibration implies auto-calibration for identifiable functionals with a sufficiently rich class $\mathcal{F}$.