Motivation: Underlying risk of weather forecasts

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

Likelihood of a loan default

“Investors assess the borrowers’ credit risk.”

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

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

Probability of accident

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

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

Mathematical formalism

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

\[
\mathbb{P}(Y = 1 | \hat{Y}(x) = \rho) = \rho, \quad \forall \rho \in [0,1],
\]

which is equivalent to:

\[
\mathbb{E}[Y | \hat{Y}(x) = \rho] = \rho, \quad \forall \rho \in [0,1].
\]

Visualization tools

Calibration curve of a binary classifier,

\[
g : \{0,1\} \rightarrow [0,1], \quad \rho \mapsto g(\rho) := \mathbb{E}[Y | \hat{Y}(x) = \rho].
\]

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

Reliability diagram: Based on bins defined by quantiles of \( \hat{Y}(x) \).

Calibration metrics

The Brier score is expressed as

\[
BS = \frac{1}{n} \sum_{i=1}^{n} (s(x_i) - y_i)^2.
\]

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

\[
ECE = \frac{1}{n} \sum_{b=1}^{n} |acc(b) - conf(b)|
\]

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

\[
ICI = \int f(p) \Phi(p) dp
\]

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

References


