Master Your Metrics with Calibration

Wissam Siblini(✉), Jordan Fréry, Liyun He-Guelton, Frédéric Oblé, and Yi-Qing Wang

Worldline, 53 avenue Paul Krüger, 69100 Villeurbanne, France
{wissam.siblini,jordan.frery,liyun.he-guelton,frederic.oble, yi-qing.wang}@worldline.com

Abstract. Machine learning models deployed in real-world applications are often evaluated with precision-based metrics such as F1-score or AUC-PR (Area Under the Curve of Precision Recall). Heavily dependent on the class prior, such metrics make it difficult to interpret the variation of a model's performance over different subpopulations/subperiods in a dataset. In this paper, we propose a way to calibrate the metrics so that they can be made invariant to the prior. We conduct a large number of experiments on balanced and imbalanced data to assess the behavior of calibrated metrics and show that they improve interpretability and provide a better control over what is really measured. We describe specific real-world use-cases where calibration is beneficial such as, for instance, model monitoring in production, reporting, or fairness evaluation.

Keywords: Performance metrics · Class imbalance · Precision-recall

1 Introduction

In real-world machine learning systems, the predictive performance of a model is often evaluated on multiple datasets, and comparisons are made. These datasets can correspond to sub-populations in the data, or different periods in time [15]. Choosing the best suited metrics is not a trivial task. Some metrics may prevent a proper interpretation of the performance differences between the sets [8,14], especially because different datasets generally not only have a different likelihood $\mathbb{P}(x|y)$ but also a different class prior $\mathbb{P}(y)$. A metric dependent on the prior (e.g. precision) will be affected by both differences indiscernibly [3] but a practitioner could be interested in isolating the variation of performance due to likelihood which reflects the intrinsic model's performance (see illustration in Fig. 1). Take the example of comparing the performance of a model across time periods: At time t, we receive data drawn from $\mathbb{P}_t(x,y) = \mathbb{P}_t(x|y)\mathbb{P}_t(y)$ where x are the features and y the label. Hence the optimal scoring function (i.e. model) for this dataset is the likelihood ratio [11]:

$$s_t(x) := \frac{\mathbb{P}_t(x|y=1)}{\mathbb{P}_t(x|y=0)} \tag{1}$$

In particular, if $\mathbb{P}_t(x|y)$ does not vary with time, neither will $s_t(x)$. In this case, even if the prior $\mathbb{P}_t(y)$ varies, it is desirable to have a performance metric $M(\cdot)$

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 457–469, 2020.
https://doi.org/10.1007/978-3-030-44584-3_36

satisfying $M(s_t, \mathbb{P}_t) = M(s_{t+1}, \mathbb{P}_{t+1}), \forall t$ so that the model maintains the same metric value over time. That being said, this does not mean that dependence to prior is an intrinsically bad behavior. Some applications seek this property as it reflects a part of the difficulty to classify on a given dataset (e.g. the performance of the random classifier evaluated with a prior-dependent metric is more or less high depending on the skew of the dataset).

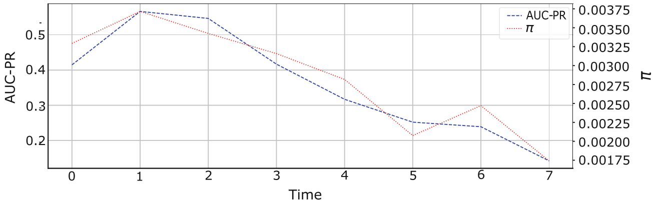


Fig. 1. Evolution of the AUC-PR of a fraud detection system and of the fraud ratio (π, i.e. the empirical $\mathbb{P}_t(y)$) over time. Both decrease, but, as the AUC-PR is dependent on the prior, it does not allow to tell if the performance variation is only due to the variation of π or if there was a drift in $\mathbb{P}_t(x|y)$

In binary classification, researchers often rely on the AUC-ROC (Area Under the Curve of Receiver Operating Characteristic) to measure a classifier's performance [6,9]. While this metric has the advantage of being invariant to the class prior, many real-world applications, especially when data are imbalanced, have recently begun to favor precision-based metrics such as AUC-PR and F-Score [12,13]. The reason is that AUC-ROC suffers from giving false positives too little importance [5] although the latter strongly deteriorate user experience and waste human efforts with false alerts. Indeed AUC-ROC considers a tradeoff between TPR and FPR whereas AUC-PR/F1-score consider a tradeoff between TPR (Recall) and Precision. With a closer look, the difference boils down to the fact that it normalizes the number of false positives with respect to the number of true negatives whereas precision-based metrics normalize it with respect to the number of true positives. In highly imbalanced scenarios (e.g. fraud/disease detection), the first is much more likely than the second because negative examples are in large majority.

Precision-based metrics give false positives more importance, but they are tied to the class prior [2,3]. A new definition of precision and recall into precision gain and recall gain has been recently proposed to correct several drawbacks of AUC-PR [7]. But, while the resulting AUC-PR Gain has some advantages of the AUC-ROC such as the validity of linear interpolation between points, it remains dependent on the class prior. Our study aims at providing metrics (i) that are precision-based to tackle problems where the class of interest is highly under-represented and (ii) that can be made independent of the prior for comparison purposes (e.g. monitoring the evolution of the performance of a classifier across several time periods). To reach this objective, this paper provides: (1) A formulation of calibration for precision-based metrics. It compute the value of precision as if the ratio π of the test set was equal to a reference class ratio π_0. We give theoretical arguments to explain why it allows invariance to the class prior. We also provide a calibrated version for precision gain and recall gain [7]. (2) An empirical analysis on both synthetic and real-world data to confirm our claims and show that new metrics are still able to assess the model's performance and are easier to interpret. (3) A large scale experiments on 614 datasets using openML [16] to (a) give more insights on correlations between popular metrics by analyzing how they rank models, (b) explore the links between the calibrated metrics and the regular ones.

Not only calibration solves the issue of dependence to the prior but also allows, with parameter π_0, anticipating a different ratio and controlling what the metric precisely reflects. This new property has several practical interests (e.g. for development, reporting, analysis) and we discuss them in realistic use-cases in Sect. 5.

2 Popular Metrics for Binary Classification: Advantages and Limits

We consider a usual binary classification setting where a model has been trained and its performance is evaluated on a test dataset of N instances. $y_i \in \{0, 1\}$ is the ground-truth label of the i^{th} instance and is equal to 1 (resp. 0) if the instance belongs to the positive (resp. negative) class. The model provides $s_i \in \mathbb{R}$, a score for the i^{th} instance to belong to the positive class. For a given threshold $\tau \in \mathbb{R}$, the predicted label is $\widehat{y}_i = 1$ if $s_i > \tau$ and 0 otherwise. Predictive performance is generally measured using the number of true positives ($\text{TP} = \sum_{i=1}^{N} \mathbb{1}(\widehat{y}_i = 1, y_i = 1)$), true negatives ($\text{TN} = \sum_{i=1}^{N} \mathbb{1}(\widehat{y}_i = 0, y_i = 0)$), false positives ($\text{FP} = \sum_{i=1}^{N} \mathbb{1}(\widehat{y}_i = 1, y_i = 0)$), false negatives ($\text{FN} = \sum_{i=1}^{N} \mathbb{1}(\widehat{y}_i = 0, y_i = 1)$). One can compute relevant ratios such as the True Positive Rate (TPR) also referred to as the Recall ($Rec = \frac{\text{TP}}{\text{TP}+\text{FN}}$), the False Positive Rate ($\text{FPR} = \frac{\text{FP}}{\text{TN}+\text{FP}}$) also referred to as the Fall-out and the Precision ($Prec = \frac{\text{TP}}{\text{TP}+\text{FP}}$). As these ratios are biased towards a specific type of error and can easily be manipulated with the threshold, more complex metrics have been proposed. In this paper, we discuss the most popular ones which have been widely adopted in binary classification: F1-Score, AUC-ROC, AUC-PR and AUC-PR Gain. F1-Score is the harmonic average between $Prec$ and Rec:

$$F_1 = \frac{2 * Prec * Rec}{Prec + Rec}. \tag{2}$$

The three other metrics consider every threshold τ from the highest s_i to the lowest. For each one, they compute TP, FP, TN and FN. Then, they plot one ratio against another and compute the Area Under the Curve (Fig. 2). AUC-ROC considers the Receiver Operating Characteristic curve where TPR is plotted against FPR. AUC-PR considers the Precision vs Recall curve. Finally, in AUC-PR Gain, the precision gain ($Prec_G$) is plotted against the recall gain (Rec_G).

They are defined in [7] as follows ($\pi = \frac{\sum_{i=1}^{N} y_i}{N}$ is the positive class ratio and we always consider that it is the minority class in this paper):

$$Prec_G = \frac{Prec - \pi}{(1-\pi)Prec} \tag{3}$$

$$Rec_G = \frac{Rec - \pi}{(1-\pi)Rec} \tag{4}$$

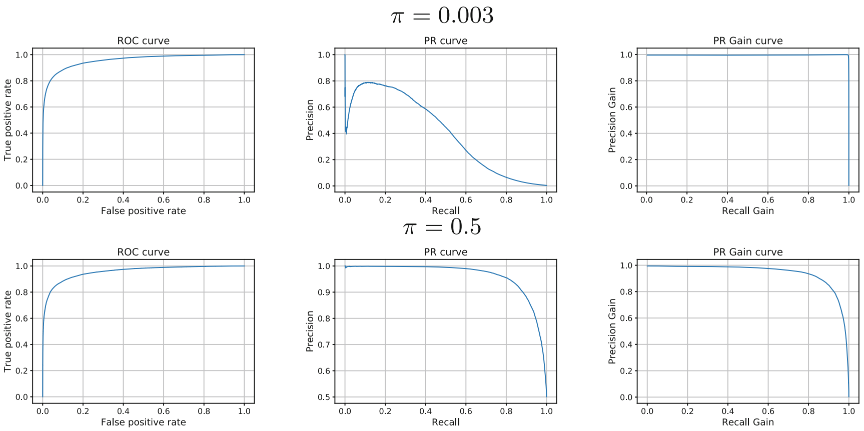


Fig. 2. ROC, PR and PR gain curves for the same model evaluated on an extremely imbalanced test set from a fraud detection application ($\pi = 0.003$, in the top row) and on a balanced sample ($\pi = 0.5$, in the bottom row).

PR Gain enjoys many properties of the ROC that the regular PR analysis does not (e.g. the validity of linear interpolations or the existence of universal baselines) [7]. However, AUC-PR Gain becomes hardly usable in extremely imbalanced settings. In particular, we can derive from (3) and (4) that $Prec_G/Rec_G$ will be mostly close to 1 if π is close to 0 (see top right chart in Fig. 2).

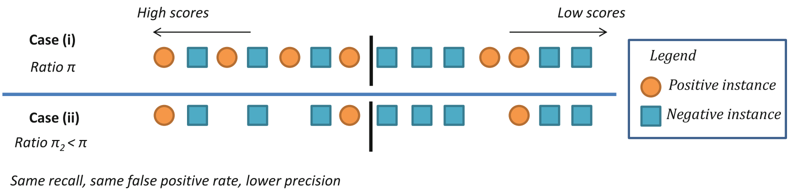


Fig. 3. Illustration of the impact of π on precision, recall, and the false positive rate. Instances are ordered from left to right according to their score given by the model. The threshold is illustrated as a vertical line between the instances: those on the left (resp. right) are classified as positive (resp. negative)

As explained in the introduction, precision-based metrics (F1, AUC-PR) are more adapted than AUC-ROC for problems with class imbalance. On the other hand, only AUC-ROC is invariant to the positive class ratio. Indeed, FPR and *Rec* are both unrelated to the class ratio because they only focus on one class but it is not the case for *Prec*. Its dependency on the positive class ratio π is illustrated in Fig. 3: when comparing a case (i) with a given ratio π and another case (ii) where a randomly selected half of the positive examples has been removed, one can visually understand that both recall and false positive rate are the same but the precision is lower in the second case.

3 Calibrated Metrics

We seek a metric that is based on *Prec* to tackle problems where data are imbalanced and the minority (positive) class is the one of interest but we want it to be invariant w.r.t. the class prior to be able to interpret its variation across different datasets (e.g. different time periods). To obtain such a metric, we will modify those based on *Prec* (AUC-PR, F1-Score and AUC-PR Gain) to make them independent of the positive class ratio π.

3.1 Calibration

The idea is to fix a reference ratio π_0 and to weigh the count of TP or FP in order to calibrate them to the value that they would have if π was equal to π_0. π_0 can be chosen arbitrarily (e.g. 0.5 for balanced) but it is preferable to fix it according to the task at hand (we analyze the impact of π_0 in Sect. 4 and describe simple guidelines to fix it in Sect. 5).

If the positive class ratio is π_0 instead of π, the ratio between negative examples and positive examples is multiplied by $\frac{\pi(1-\pi_0)}{\pi_0(1-\pi)}$. In this case, we expect the ratio between false positives and true positives to be multiplied by $\frac{\pi(1-\pi_0)}{\pi_0(1-\pi)}$. Therefore, we define the calibrated precision $Prec_c$ as follows:

$$Prec_c = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{\pi(1-\pi_0)}{\pi_0(1-\pi)}\mathrm{FP}} = \frac{1}{1 + \frac{\pi(1-\pi_0)}{\pi_0(1-\pi)}\frac{\mathrm{FP}}{\mathrm{TP}}} \tag{5}$$

Since $\frac{1-\pi}{\pi}$ is the imbalance ratio $\frac{N_-}{N_+}$ where N_+ (resp. N_-) is the number of positive (resp. negative) examples, we have: $\frac{\pi}{1-\pi}\frac{\mathrm{FP}}{\mathrm{TP}} = \frac{\mathrm{FP}/N_-}{\mathrm{TP}/N_+} = \frac{\mathrm{FPR}}{\mathrm{TPR}}$ which is independent of π.

Based on the calibrated precision, we can also define the calibrated F1-score, the calibrated $Prec_G$ and the calibrated Rec_G by replacing *Prec* by $Prec_c$ and π by π_0 in Eqs. (2), (3) and (4). Note that calibration does not change precision gain. Indeed, calibrated precision gain $\frac{Prec_c - \pi_0}{(1-\pi_0)Prec_c}$ can be rewritten as $\frac{Prec - \pi}{(1-\pi)Prec}$ which is equal to the regular precision gain. Also, the interesting properties of the recall gain were proved independently of the ratio π in [7] which means that calibration preserves them.

3.2 Robustness to Variations in π

In order to evaluate the robustness of the new metrics to variations in π, we create a synthetic dataset where the label is drawn from a Bernoulli distribution with parameter π and the feature is drawn from Normal distributions:

$$p(x|y=1;\mu_1) = \mathcal{N}(x;\mu_1,1), \qquad p(x|y=0;\mu_0) = \mathcal{N}(x;\mu_0,1) \tag{6}$$

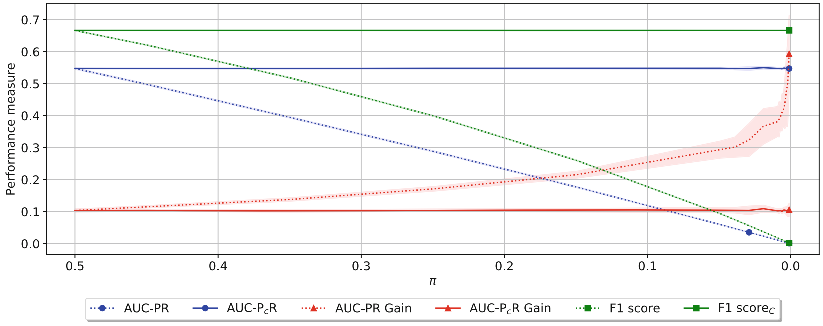


Fig. 4. Evolution of AUC-PR, AUC-PR Gain, F1-score and their calibrated version (AUC-P$_c$R, AUC-P$_c$R Gain, F1-score$_c$) as π decreases. We arbitrarily set $\pi_0 = 0.5$ for the calibrated metrics. The curves are obtained by averaging results over 30 runs and we show the confidence intervals.

For several values of π, data points are generated from (6) with $\mu_1 = 2$ and $\mu_0 = 1.8$. We consider a large number of points (10^6) so that the empirical class ratio π is approximately equal to the Bernouilli parameter π. We empirically study the evolution of several metrics (F_1-score, AUC-PR, AUC-PR Gain and their calibrated version) for the optimal model (as defined in (1)) as π decreases from $\pi = 0.5$ (balanced) to $\pi = 0.001$. We observe that the impact of the class prior on the regular metrics is important (Fig. 4). It can be a serious issue for applications where π sometimes vary by one order of magnitude from one day to another (see [4] for a real world example) as it leads to a significant variation of the measured performance (see the difference between AUC-PR when $\pi = 0.5$ and when $\pi = 0.05$) even if the optimal model remains the same. On the contrary, the calibrated versions remain very robust to changes in the class prior π even for extreme values. Note that we here experiment with synthetic data to have a full control over the distribution/prior and make the analysis easier but the conclusions are exactly the same on real world data.[1]

[1] See appendix in https://figshare.com/articles/Calibrated_metrics_IDA_Supplementary_material_pdf/11848146.

3.3 Assessment of the Model Quality

Besides the robustness of the calibrated metrics to changes in π, we also want them to be sensitive to the quality of the model. If this latter decreases regardless of the π value, we expect all metrics, calibrated ones included, to decrease in value. Let us consider an experiment where we use the same synthetic dataset as defined the previous section. However, instead of changing the value of π only, we change (μ_1, μ_0) to make the problem harder and harder and thus worsen the optimal model's performance. This can be done by reducing the distance between the two normal distributions in (6), because this would result in more overlapping between the classes and make it harder to discriminate between them. As a distance, we consider the KL-divergence that boils down to $\frac{1}{2}(\mu_1 - \mu_0)^2$.

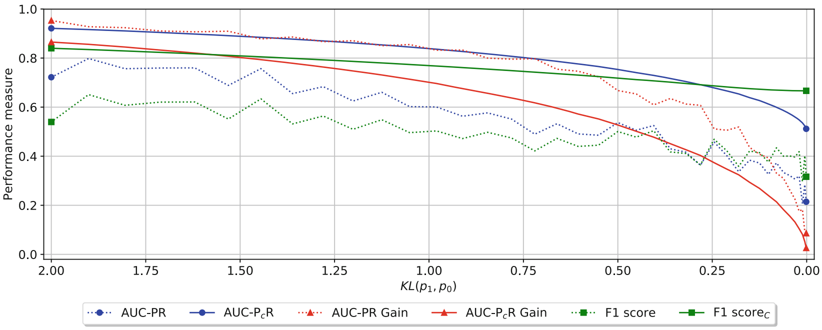


Fig. 5. Evolution of AUC-PR, AUC-PR Gain, F1-score and their calibrated version as $KL(p_1, p_0)$ tends to 0 and as π randomly varies. This curve was obtained by averaging results over 30 runs.

Figure 5 shows how the values of the metrics evolve as the KL-divergence gets closer to zero. For each run, we randomly chose the prior π in the interval $[0.001, 0.5]$. As expected, all metrics globally decrease as the problem gets harder. However, we can notice an important difference: the variation in the calibrated metrics are smooth and monotonic compared to those of the original metrics which are affected by the random changes in π. In that sense, variations of the calibrated metrics across the different generated datasets are much easier to interpret than the original metrics.

4 Link Between Calibrated and Original Metrics

4.1 Meaning of π_0

Let us first remark that for test datasets in which $\pi = \pi_0$, $Prec_c$ is equal to the regular precision $Prec$ since $\frac{\pi(1-\pi_0)}{\pi_0(1-\pi)} = 1$ (this is observable in Fig. 4 with the intersection of the metrics for $\pi = \pi_0 = 0.5$).

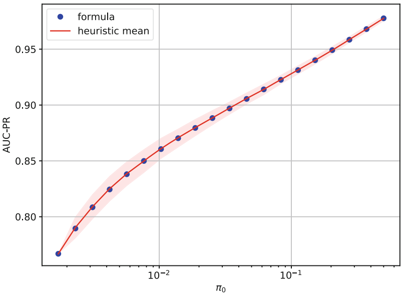


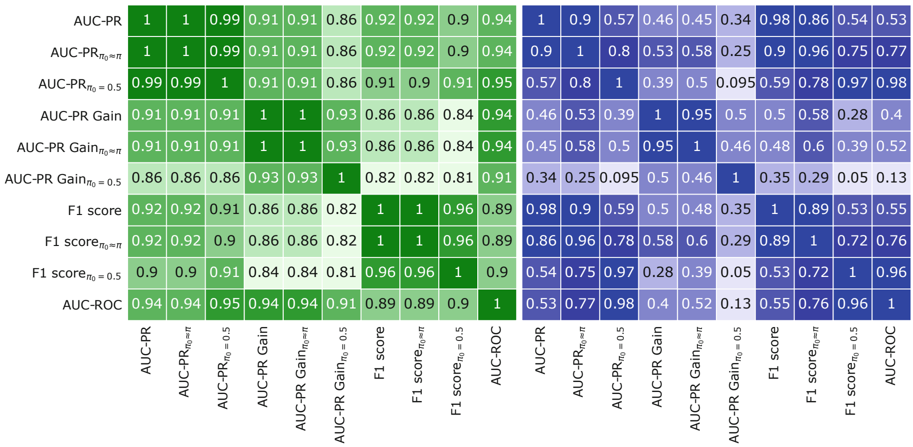
Fig. 6. Comparison between heuristic-based calibrated AUC-PR (red line) and our closed-form calibrated AUC-PR (blue dots). The red shadow represents the standard deviation of the heuristic-based calibrated AUC-PR over 1000 runs. (Color figure online)

If $\pi \neq \pi_0$, the calibrated metrics essentially have the value that the original ones would have if the positive class ratio π was equal to π_0. To further demonstrate that, we compare our proposal for calibration (5) with the only proposal from the past [10] that was designed for the same objective: a heuristic-based calibration. The approach from [10] consists in randomly undersampling the test set to make the positive class ratio π equal to a chosen ratio (let us refer to it as π_0 for the analogy) and then computing the regular metrics on the sampled set. Because of the randomness, sampling may remove more hard examples than easy examples so the performance can be over-estimated, and vice versa. To avoid that, the approach performs several runs and computes a mean estimation. In Fig. 6, we compare the results obtained with our formula and with their heuristic, for several reference ratio π_0, on a highly unbalanced ($\pi = 0.0017$) credit card fraud detection dataset available on Kaggle [4].

We can observe that our formula and the heuristic provide really close values. This can be theoretically explained (See Footnote 1) and confirms that our formula really computes the value that the original metric would have if the ratio π in the test set was π_0. Note that our closed-form calibration (5) can be seen as an improvement of the heuristic-based calibration from [10] as it directly provides the targeted value without running a costly Monte-Carlo simulation.

4.2 Do the Calibrated Metrics Rank Models in the Same Order as the Original Metrics?

Calibration results in evaluating the metric for a different prior. In this section, we analyze how this impacts the task of selectioning the best model for a given dataset. To do this, we empirically analyze the correlation of several metrics in terms of model ordering. We use OpenML [16] to select the 602 supervised binary classification datasets on which at least 30 models have been evaluated with a 10-fold cross-validation. For each one, we randomly choose 30 models, fetch their predictions, and evaluate their performance with the metrics. This leaves us with $614 \times 30 = 18,420$ different values for each metric. To analyze whether they rank the models in the same order, we compute the Spearman rank correlation coefficient between them for the 30 models for each of the 614 problems.[2] Most datasets roughly have balanced classes ($\pi > 0.2$ in more than 90% of the datasets). Therefore, to also specifically analyze the imbalance case, we run the same experiment with only the subset of 4 highly imbalanced datasets ($\pi < 0.01$). The compared metrics are AUC-ROC, AUC-PR, AUC-PR Gain and the best F1-score over all possible thresholds. We also add the calibrated version of the last three. In order to understand the impact of π_0, we use two different values: the arbitrary $\pi_0 = 0.5$ and another value $\pi_0 \approx \pi$ (for the first experiment with all datasets, $\pi_0 \approx \pi$ corresponds to $\pi_0 = 1.01\pi$ and for the second experiment where π is very small, we go further and $\pi_0 \approx \pi$ corresponds to $\pi_0 = 10\pi$ which remains closer to π than 0.5). The obtained correlation matrices are shown in Fig. 7. Each individual cell corresponds to the average Spearman correlation over all datasets between the row metric and the column metric.



	AUC-PR	AUC-PR$_{\pi_0\approx\pi}$	AUC-PR$_{\pi_0=0.5}$	AUC-PR Gain	AUC-PR Gain$_{\pi_0\approx\pi}$	AUC-PR Gain$_{\pi_0=0.5}$	F1 score	F1 score$_{\pi_0\approx\pi}$	F1 score$_{\pi_0=0.5}$	AUC-ROC
AUC-PR	1	1	0.99	0.91	0.91	0.86	0.92	0.92	0.9	0.94
AUC-PR$_{\pi_0\approx\pi}$	1	1	0.99	0.91	0.91	0.86	0.92	0.92	0.9	0.94
AUC-PR$_{\pi_0=0.5}$	0.99	0.99	1	0.91	0.91	0.86	0.91	0.9	0.91	0.95
AUC-PR Gain	0.91	0.91	0.91	1	1	0.93	0.86	0.86	0.84	0.94
AUC-PR Gain$_{\pi_0\approx\pi}$	0.91	0.91	0.91	1	1	0.93	0.86	0.86	0.84	0.94
AUC-PR Gain$_{\pi_0=0.5}$	0.86	0.86	0.86	0.93	0.93	1	0.82	0.82	0.81	0.91
F1 score	0.92	0.92	0.91	0.86	0.86	0.82	1	1	0.96	0.89
F1 score$_{\pi_0\approx\pi}$	0.92	0.92	0.9	0.86	0.86	0.82	1	1	0.96	0.89
F1 score$_{\pi_0=0.5}$	0.9	0.9	0.91	0.84	0.84	0.81	0.96	0.96	1	0.9
AUC-ROC	0.94	0.94	0.95	0.94	0.94	0.91	0.89	0.89	0.9	1

	AUC-PR	AUC-PR$_{\pi_0\approx\pi}$	AUC-PR$_{\pi_0=0.5}$	AUC-PR Gain	AUC-PR Gain$_{\pi_0\approx\pi}$	AUC-PR Gain$_{\pi_0=0.5}$	F1 score	F1 score$_{\pi_0\approx\pi}$	F1 score$_{\pi_0=0.5}$	AUC-ROC
AUC-PR	1	0.9	0.57	0.46	0.45	0.34	0.98	0.86	0.54	0.53
AUC-PR$_{\pi_0\approx\pi}$	0.9	1	0.8	0.53	0.58	0.25	0.9	0.96	0.75	0.77
AUC-PR$_{\pi_0=0.5}$	0.57	0.8	1	0.39	0.5	0.095	0.59	0.78	0.97	0.98
AUC-PR Gain	0.46	0.53	0.39	1	0.95	0.5	0.5	0.58	0.28	0.4
AUC-PR Gain$_{\pi_0\approx\pi}$	0.45	0.58	0.5	0.95	1	0.46	0.48	0.6	0.39	0.52
AUC-PR Gain$_{\pi_0=0.5}$	0.34	0.25	0.095	0.5	0.46	1	0.35	0.29	0.05	0.13
F1 score	0.98	0.9	0.59	0.5	0.48	0.35	1	0.89	0.53	0.55
F1 score$_{\pi_0\approx\pi}$	0.86	0.96	0.78	0.58	0.6	0.29	0.89	1	0.72	0.76
F1 score$_{\pi_0=0.5}$	0.54	0.75	0.97	0.28	0.39	0.05	0.53	0.72	1	0.96
AUC-ROC	0.53	0.77	0.98	0.4	0.52	0.13	0.55	0.76	0.96	1

Fig. 7. Spearman rank correlation matrices between 10 metrics over 614 datasets for the left figure and the 4 highly imbalanced datasets for the right figure.

A general observation is that most metrics are less correlated with each other when classes are unbalanced (right matrix in Fig. 7). We also note that the best F1-score is more correlated to AUC-PR than to AUC-ROC or AUC-PR Gain. In the balanced case (left matrix in Fig. 7), we can see that metrics defined as area under curves are generally more correlated with each other than with the threshold sensitive classification metric F1-score. Let us now analyze the impact of calibration. As expected, in general, when $\pi_0 \approx \pi$, calibrated metrics have a behavior really close to that of the original metrics because $\frac{\pi(1-\pi_0)}{\pi_0(1-\pi)} \approx 1$ and therefore

[2] The implementation of the paper experiments can be found at https://github.com/wissam-sib/calibrated_metrics.

$Prec_c \approx Prec$. In the balanced case (left), since π is close to 0.5, calibrated metrics with $\pi_0 = 0.5$ are also highly correlated with the original metrics. In the imbalanced case (on the right matrix of Fig. 7), when π_0 is arbitrarily set to 0.5 the calibrated metrics seem to have a low correlation with the original ones. In fact, they are less correlated with them than with AUC-ROC. And this makes sense given the relative weights that each of the metric applies to FP and TP. The original precision gives the same weight to TP and FP, although false positives are $\frac{1-\pi}{\pi}$ times more likely to occur ($\frac{1-\pi}{\pi} > 100$ if $\pi < 0.01$). The calibrated precision with the arbitrary value $\pi_0 = 0.5$ boils down to $\frac{\mathrm{TP}}{\mathrm{TP} + \frac{\pi}{(1-\pi)}\mathrm{FP}}$ and gives a weight $\frac{1-\pi}{\pi}$ times smaller to false positives which counterbalances their higher likelihood. ROC, like the calibrated metrics with $\pi_0 = 0.5$, gives $\frac{1-\pi}{\pi}$ less weight to FP because it is computed from FPR and TPR which are linked to TP and FP with the relationship $\frac{\pi}{1-\pi}\frac{\mathrm{FP}}{\mathrm{TP}} = \frac{\mathrm{FPR}}{\mathrm{TPR}}$.

To sum up the results, we first emphasize that the choice of the metrics to rank classifiers when datasets are rather balanced seems to be much less sensitive than in the extremely imbalanced case. In the balanced case the least correlated metrics have an average rank correlation of 0.81. For the imbalanced datasets, on the other hand, many metrics have low correlations which means that they often disagree on the best model. The choice of the metric is therefore very important here. Our experiment also seems to reflect that rank correlations are mainly a matter of how much weight is given to each type of error. Choosing these "weights" generally depends on the application at hand. An this should be remembered when using calibration. To preserve the nature of a given metrics, π_0 has to be fixed to a value close to π and not arbitrarily. The user still has the choice to fix it to another value if his purpose is to specifically place the results into a different reference with a different prior.

5 Guidelines and Use-Cases

Calibration could benefit ML practitioners when analyzing the performance of a model across different datasets/time periods. Without being exhaustive, we give four use-cases where it is beneficial (setting π_0 depends on the target use-case):

Comparing the Performance of a Model on Two Populations/Classes: Consider a practitioner who wants to predict patients with a disease and evaluate the performance of his model on subpopulations of the dataset (e.g. children, adults and elderly people). If the prior is different from one population to another (e.g. elderly people are more likely to have the disease), precision will be affected, i.e. population with a higher disease ratio will be more likely to have a higher precision. In this case, the calibrated precision can be used to obtain the precision of each population set to the same reference prior (for instance, π_0 can be chosen as the average prior over all populations). This would provide an additional balanced point of view and make the analysis richer to draw more precise conclusions and perhaps study fairness [1].

Model Performance Monitoring in an Industrial Context: In systems where a model's performance is monitored over time with precision-based metrics like F1-score, using calibration in addition to the regular metrics makes it easier to understand the evolution especially when the class prior can evolve (cf. application in Fig. 1). For instance, it can be useful to analyze the drift (i.e. distinguish between variations linked to π or $P(X|y)$) and design adapted solutions; either updating the threshold or completely retraining the model. To avoid denaturing too much the F1-score, here π_0 has to be fixed based on realistic values (e.g. average π in historical data).

Establishing Agreements with Clients: As shown in previous sections, π_0 can be interpreted as the ratio to which we refer to compute the metric. This can be useful to establish a guarantee, in an agreement, that will be robust to uncontrollable events. Indeed, if we take the case of fraud detection, the real positive class ratio π can vary extremely from one day to another and on particular events (e.g. fraudster attacks, holidays) which significantly affects the measured metrics (see Fig. 4). Here, after having both parties to agree beforehand on a reasonable value for π_0 (based on their business knowledge), calibration will always compute the performance relative to this ratio and not the real π and thus be easier to guarantee.

Anticipating the Deployment of a Model in Production: Imagine one collects a sample of data to develop an algorithm and reaches an acceptable AUC-PR for production. If the prior in the collected data is different from reality, the non-calibrated metric might have given either a pessimistic or optimistic estimation of the post-deployment performance. This can be extremely harmful if the production has strict constraints. Here, if the practitioner uses calibration with π_0 equal to the minimal prior envisioned for the application at hand, he/she would be able to anticipate the worst case scenario.

6 Conclusion

In this paper, we provided a formula of calibration, empirical results, and guidelines to make the values of metrics across different datasets more interpretable. Calibrated metrics are a generalization of the original ones. They rely on a reference π_0 and compute the value that we would obtain if the positive class ratio π in the evaluated test set was equal to π_0. If the user chooses $\pi_0 = \pi$, this does not change anything and he retrieves the regular metrics. But, with different choices, the metrics can serve several purposes such as obtaining robustness to variation in the class prior across datasets, or anticipation. They are useful in both academic and industrial applications as explained in the previous section: they help drawing more accurate comparisons between subpopulations, or study incremental learning on streams by providing a point of view agnostic to virtual concept drift [17]. They can be used to provide more controllable performance indicators (easier to guarantee and report), help preparing deployment in production, and prevent false conclusions about the evolution of a deployed model. However, π_0 has to be chosen with caution as it controls the relative weights given to FP and TP and, consequently, can affect the selection of the best classifier.

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. NIPS Tutorial (2017)
2. Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modeling on imbalanced domains. ACM Comput. Surv. (CSUR) **49**(2), 31 (2016)
3. Brzezinski, D., Stefanowski, J., Susmaga, R., Szczech, I.: On the dynamics of classification measures for imbalanced and streaming data. IEEE Trans. Neural Netw. Learn. Syst. (2019)
4. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Trans. Neural Netw. Learn. Syst. **29**(8), 3784–3797 (2018)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)
6. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
7. Flach, P., Kull, M.: Precision-recall-gain curves: PR analysis done right. In: Advances in Neural Information Processing Systems, pp. 838–846 (2015)
8. Garcıa, V., Sánchez, J.S., Mollineda, R.A.: On the suitability of numerical performance measures for class imbalance problems. In: International Conference in Pattern Recognition Applications and Methods, pp. 310–313 (2012)
9. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)
10. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data-recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245–251. IEEE (2013)
11. Neyman, J., Pearson, E.S.: IX. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character **231**(694–706), 289–337 (1933)
12. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3), e0118432 (2015)
13. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. In: Advances in Neural Information Processing Systems, pp. 5228–5237 (2018)
14. Santafe, G., Inza, I., Lozano, J.A.: Dealing with the evaluation of supervised classification algorithms. Artif. Intell. Rev. **44**(4), 467–508 (2015). https://doi.org/10.1007/s10462-015-9433-y

15. Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. In: Advances in Neural Information Processing Systems, pp. 1920–1930 (2018)
16. Vanschoren, J., Van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: networked science in machine learning. ACM SIGKDD Explor. Newslett. **15**(2), 49–60 (2014)
17. Widmer, G., Kubat, M.: Effective learning in dynamic environments by explicit context tracking. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 227–243. Springer, Heidelberg (1993). https://doi.org/10.1007/3-540-56602-3_139