Precision-Recall-Gain Curves: PR Analysis Done Right

Peter.Flach, Meelis.Kull @ bristol.ac.uk

University of Bristol
Intelligent Systems Lab

chist-era
EPSRC Engineering and Physical Sciences Research Council
REFRAME
Rethinking the Essence, Flexibility and Reusability of Advanced Model Exploitation
www.reframe-d2k.org

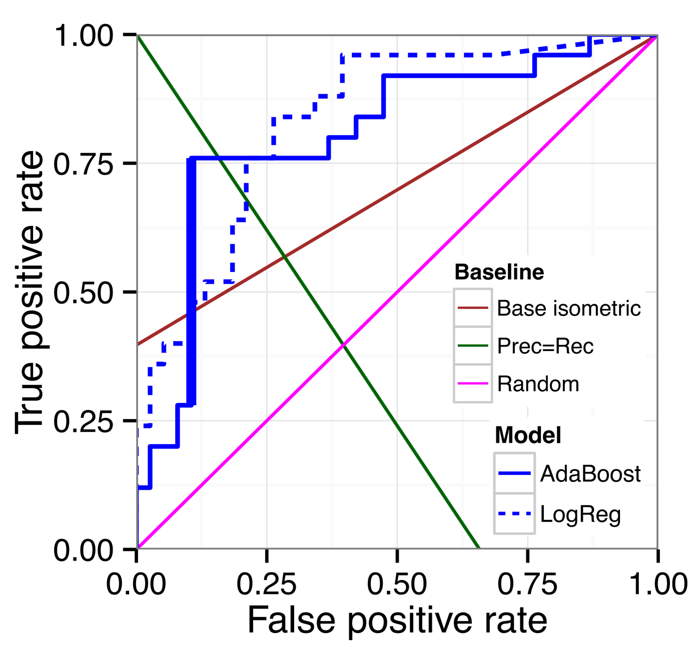
ROC Curve

		Predicted ⊕	Predicted ⊖	
Actual	⊕	TP	FN	Pos
	⊖	FP	TN	Neg
		$PPos$	$PNeg$	N

$$tpr \triangleq \frac{TP}{Pos} = \frac{TP}{TP+FN}$$

$$fpr \triangleq \frac{FP}{Neg} = \frac{FP}{FP+TN}$$

$$\pi \triangleq \frac{Pos}{N}$$



$$acc = \pi\, tpr + (1-\pi)(1-fpr)$$

$$acc_c \triangleq 2c\pi\, tpr + 2(1-c)(1-\pi)(1-fpr)$$

1. Baselines are universal
2. Interpolation is linear
3. Accuracy-isometrics are linear
4. Pareto-front is convex
5. Area-under-curve relates to expected accuracy

Theorem 0. Let the operating points of a model with area AUC under the ROC-curve be chosen such that $rate = \pi\, tpr + (1-\pi) fpr$ is uniformly distributed within $[0,1]$. Then the expected accuracy is equal to

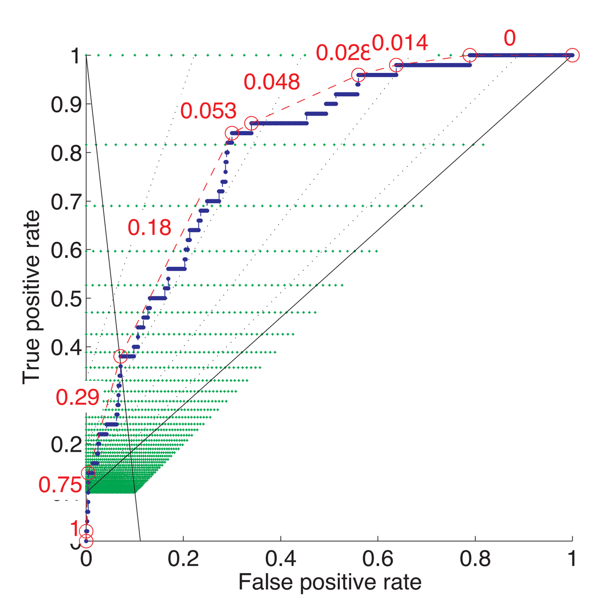
$$\mathbb{E}[acc] = \frac{1}{2} + \pi(1-\pi)(2AUC-1).$$

Corollary. *For balanced classes $\pi = 1/2$ and hence*

$$\mathbb{E}[acc] = AUC/2 + 1/4.$$

6. Accuracy-calibrated scores via slope

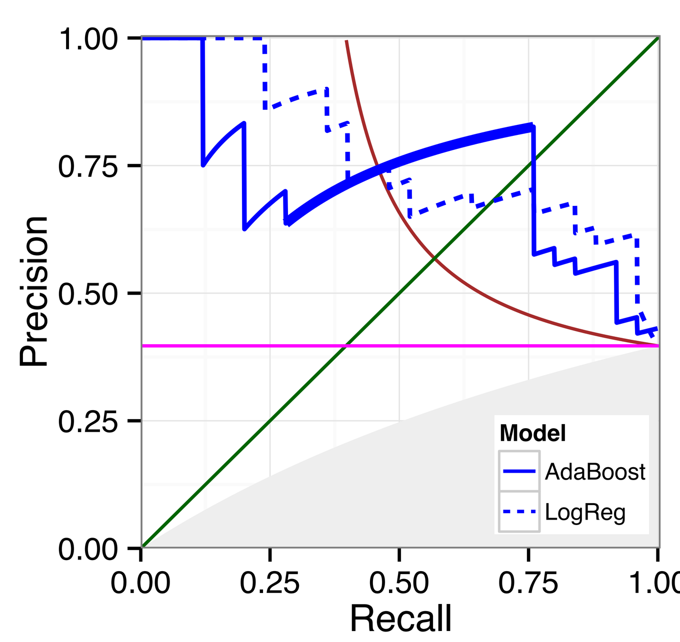
$$c_{ROC} = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{1}{slope_{ROC}}}$$



Precision-Recall Curve

$$prec \triangleq \frac{TP}{TP+FP}$$

$$rec \triangleq \frac{TP}{TP+FN}$$

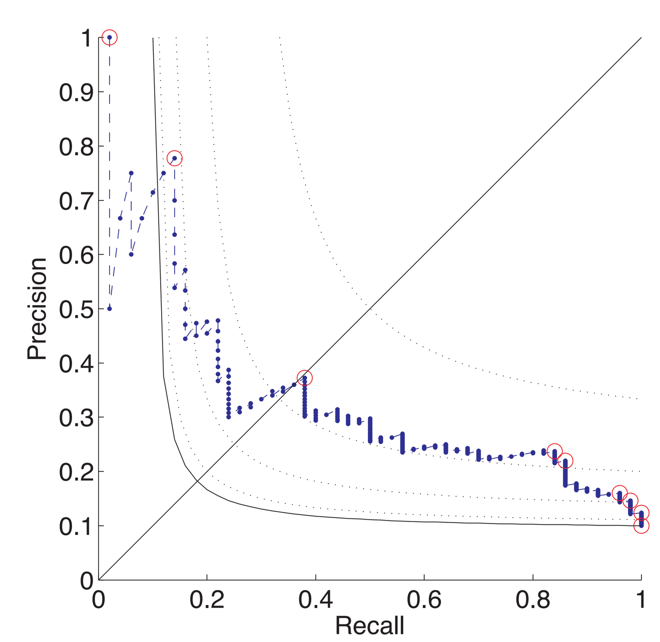


$$F_1 \triangleq \frac{2}{1/prec + 1/rec} = \frac{TP}{TP + (FP+FN)/2}$$

$$F_\beta \triangleq \frac{1}{\frac{1}{1+\beta^2}/prec + \frac{\beta^2}{1+\beta^2}/rec} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + FP + \beta^2 FN}$$

1. Baselines are non-universal
2. Interpolation is non-linear
3. F-isometrics are non-linear
4. Pareto-front is non-convex
5. Area-under-curve does not relate to expected F-measure and there is an unachievable region

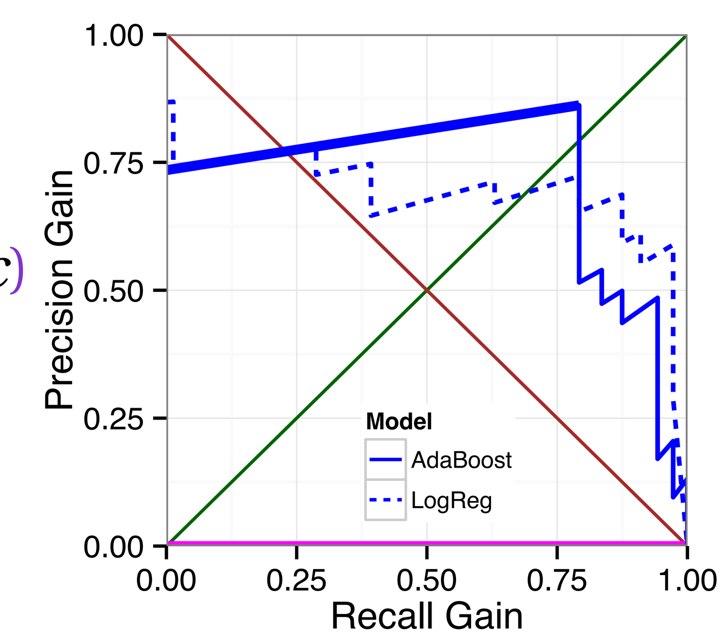
6. F-calibrated scores not via slope



Precision-Recall-Gain Curve

$$precGain \triangleq \varphi(prec)$$

$$recGain \triangleq \varphi(rec)$$



$$FGain_1 \triangleq \varphi(F_1) = \frac{1}{2}precGain + \frac{1}{2}recGain$$

$$FGain_\beta \triangleq \varphi(F_\beta) = \frac{1}{1+\beta^2}precGain + \frac{\beta^2}{1+\beta^2}recGain$$

1. Baselines are universal
2. Interpolation is linear **Theorem 1**
3. FGain- & F-isometrics are linear
4. Pareto-front is convex **Theorem 2**
5. Area-under-curve relates to expected F-Gain **Theorem 3**

Theorem 3. Let the operating points of a model with area $AUPRG$ under the PRG-curve be chosen such that $\Delta = recGain/\pi - precGain/(1-\pi)$ is uniformly distributed within $[-y_0/(1-\pi), 1/\pi]$, where y_0 is the value of $recGain$ at $precGain = 0$. Then the expected $FGain_1$ score is equal to

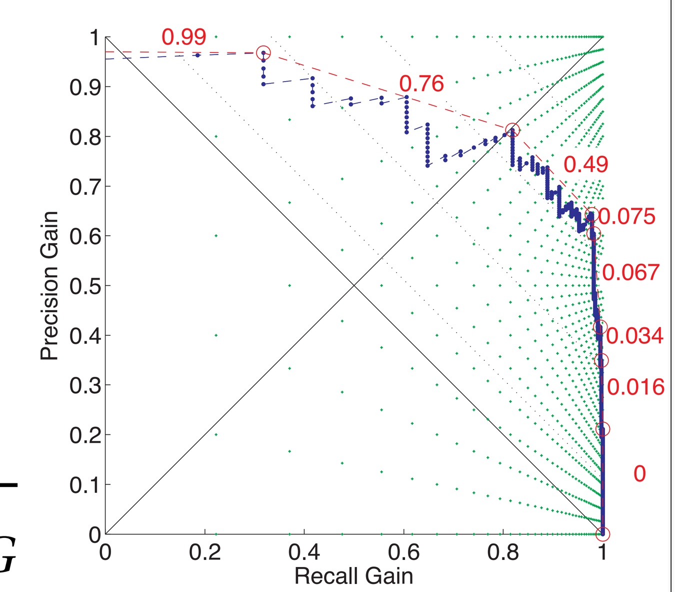
$$\mathbb{E}[FGain_1] = \frac{AUPRG/2 + 1/4 - \pi(1-y_0^2)/4}{1-\pi(1-y_0)}$$

Corollary. *If the model has $prec = 1$ at $rec = \pi$ then $y_0 = 1$ and hence*

$$\mathbb{E}[FGain_1] = AUPRG/2 + 1/4.$$

6. FGain-calibrated scores via slope

$$c_{PRG} = \frac{1}{1 - slope_{PRG}}$$



Linearising a Harmonic Scale

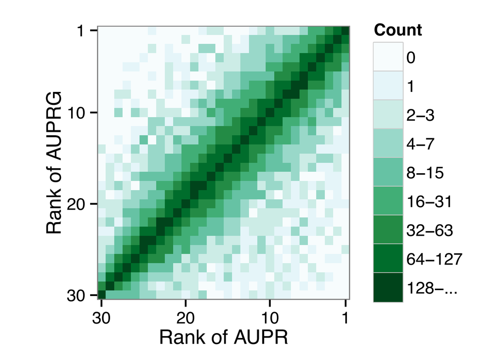
$$\text{Linear } [min,max] \mapsto \text{Linear } [0,1] : \quad x \mapsto \frac{x - min}{max - min}$$

$$\text{Harmonic } [min,max] \mapsto \text{Linear } [0,1] : \quad x \mapsto \frac{1/min - 1/x}{1/min - 1/max}$$

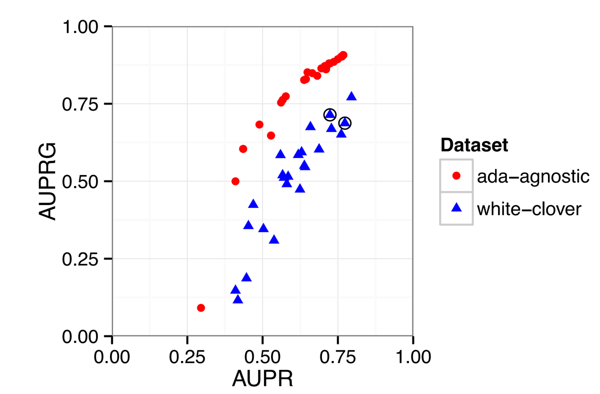
$$\text{Harmonic } [\pi,1] \mapsto \text{Linear } [0,1] : \quad \varphi(x) \triangleq \frac{1/\pi - 1/x}{1/\pi - 1} = \frac{1 - \pi/x}{1-\pi}$$

Experiments on 426 datasets

In each of 886 OpenML binary classification tasks covering 426 datasets we fetched predictions from 30 models and ranked these models according to AUPR and AUPRG.



Results for two tasks are shown on the right. AUPR and AUPRG rank the two encircled blue models differently. These are AdaBoost and LogReg, with ROC-, PR-, and PRG-curves shown left above.



Take Home Messages

PR analysis requires proper treatment of the harmonic scale — arithmetic averages or linear expectations of F-scores etc are incoherent.

Precision-Recall-Gain curves properly linearise the quantities involved and their area is meaningful as an aggregate performance score.

These things matter in practice as AUPR can easily favour worse-performing models.

Using PRG curves we can identify all F_β-optimal thresholds for any β in a single calibration procedure.

Code is available at http://www.cs.bris.ac.uk/~flach/PRGcurves/.