

ROC Curves / Profit Curves

Clayton W. Schupp, Galvanize

Spring 2015

Confusion Matrix

		Actual class	
		p	n
Predicted class	Y	True Positives	False Positives
	N	False negatives	True negatives

TP rate = TP / P recall (hit rate)

FP rate = FP / N (false alarm rate)

Accuracy = $(TP + TN) / (P + N)$

Precision = $TP / (TP + FP)$

Other Common Terms ...

Sensitivity = recall

Specificity = $TN / (FP + TN)$
 $= 1 - FP \text{ rate}$

Problems with Unbalanced Classes

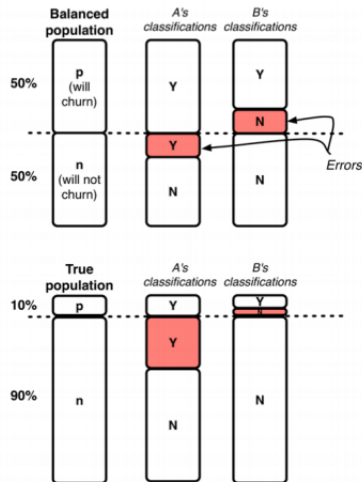
- Two different cases

Table 7-2. Confusion matrix of A

	churn	not churn
Y	500	200
N	0	300

Table 7-3. Confusion matrix of B

	churn	not churn
Y	300	0
N	200	500



Building the ROC Curve

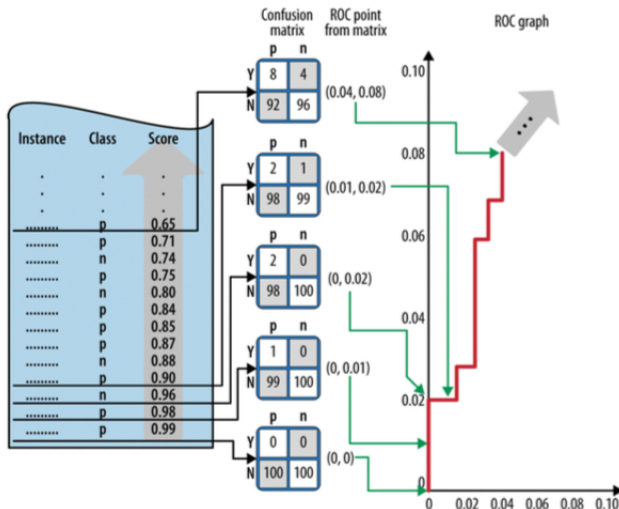
For a given model f , each threshold value T gives a point on the ROC Curve

Model score is the probability of class membership

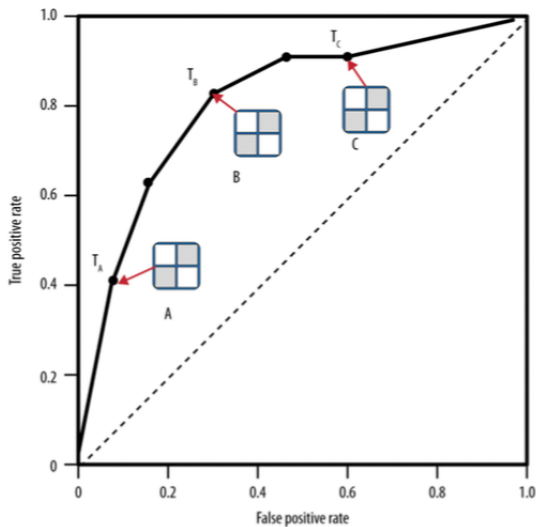
- 1 $T = \text{minimum score}$
- 2 $TP=0, FP=0$
- 3 For each observation, i :
 - If $i > T \longrightarrow \text{increment TP}$
 - else $\longrightarrow \text{increment FP}$
- 4 Add point $(FP/N, TP/P)$ to ROC Graph

Increment T from min-score to max-score, repeating steps 1-4

Building the ROC Curve



Sample ROC Curve

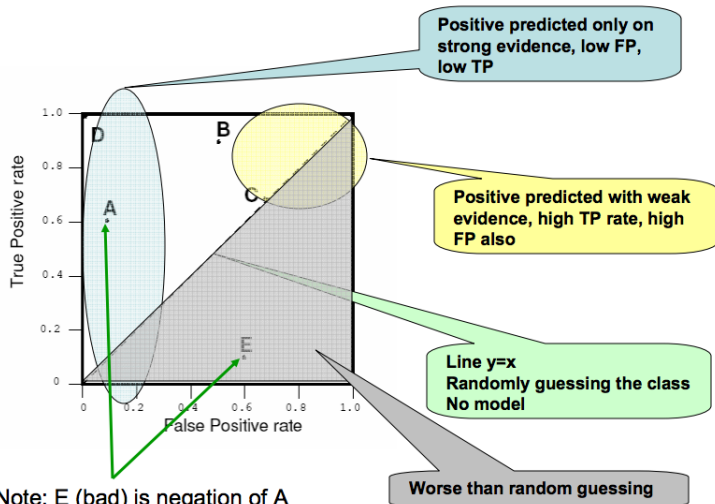


Choosing Between Models

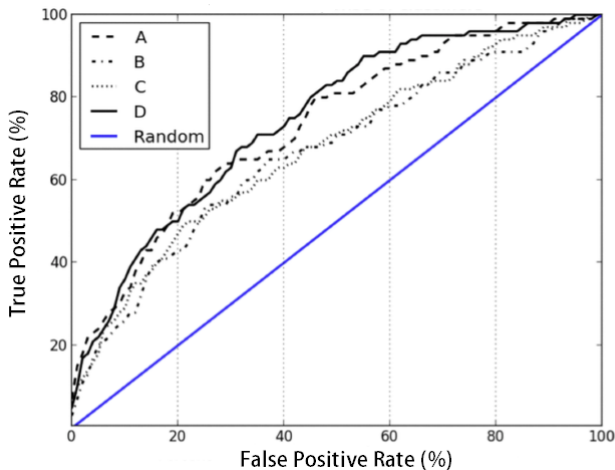
How do we go about choosing a model based on the ROC curve?

- Depends on the goal of the model
 - Screening Test vs. Diagnostic Test
- We can examine the regions of the ROC curve based on desired result

Regions of the ROC Curve



ROC Curve for Multiple Classifiers



Cost-Benefit Information

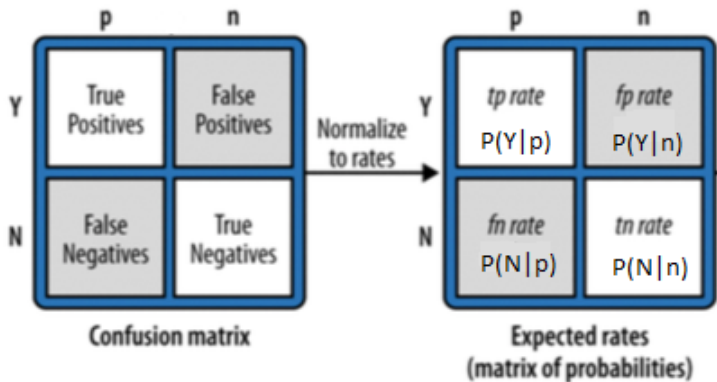
- ROC Curves alone assume equal cost of misclassification
- Different kinds of errors have different costs associated
- Correct classifications could also have different benefits

Profit Curves allow us to compare models and select the one that will maximize profit for a specified cost-benefit

Cost-Benefit Matrix

		Actual	
		p	n
Predicted	Y	$b(Y,p)$	$c(Y,n)$
	N	$c(N,p)$	$b(N,n)$

Normalize Confusion Matrix to Rates



Expected Profit

By combining information from the Confusion Matrix and the Cost-Benefit Matrix, we can calculate the Expected Profit:

$$\begin{aligned} E[\textit{Profit}] &= P(Y, p) \cdot b(Y, p) + P(Y, n) \cdot c(Y, n) + \\ &= P(N, p) \cdot c(N, p) + P(N, n) \cdot b(N, n) \\ &= P(Y|p) \cdot P(p) \cdot b(Y, p) + P(Y|n) \cdot P(n) \cdot c(Y, n) + \\ &= P(N|p) \cdot P(p) \cdot c(N, p) + P(N|n) \cdot P(n) \cdot b(N, n) \\ &= P(p) \cdot [P(Y|p) \cdot b(Y, p) + P(N|p) \cdot c(N, p)] + \\ &= P(n) \cdot [P(Y|n) \cdot c(Y, n) + P(N|n) \cdot b(N, n)] \end{aligned}$$

Building the Profit Curve

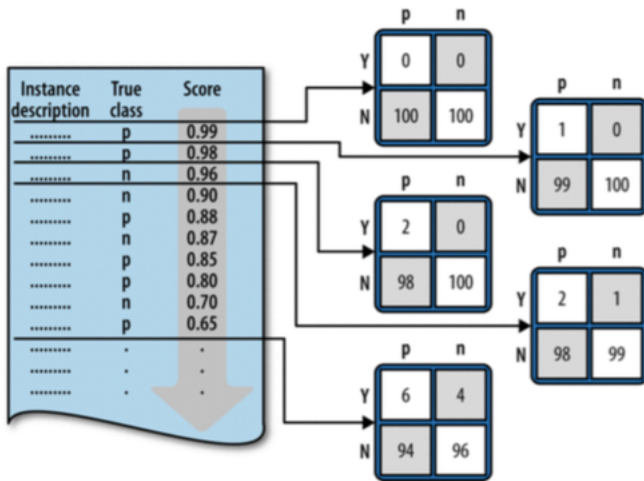
Similar to building ROC Curve. For a given model f , each threshold value T gives a point on the Profit Curve

Model score is the probability of class membership

- 1 $T = \text{maximum score}$
- 2 Using the confusion matrix and cost-benefit matrix, calculate $E[Profit]$
- 3 For each observation, i :
 - If $i > T \rightarrow \text{increment TP}$
 - else $\rightarrow \text{increment FP}$
- 4 Add point $(E[Profit], \% \text{ Test Instances})$ to Profit Graph

Increment T from max-score to min-score, repeating steps 1-4

Building the Profit Curve



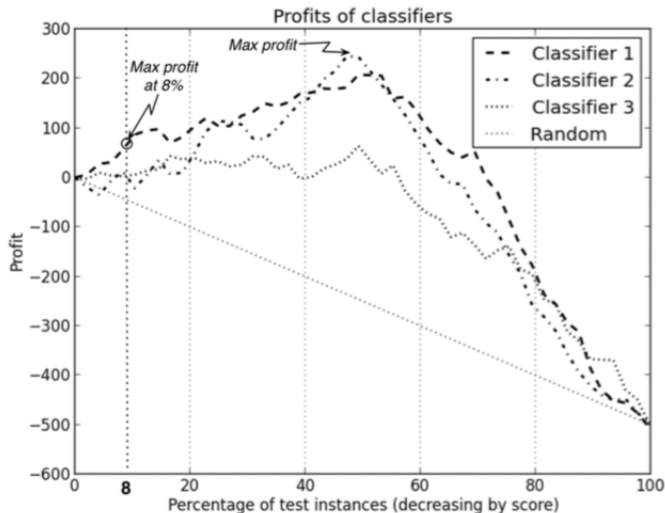
Example: Building the Profit Curve

Let's assume our profit margin is small: each offer costs \$5 to make and market and each accepted offer earns \$9, for a profit of \$4.

The cost matrix is:

	p	n
Y	\$4	-\$5
N	\$0	\$0

Profit Curves for Multiple Classifiers



Profit Curves Conditions

Critical Condition underlying the profit calculations

- Class Priors: the proportion of positive and negative instances in the target population
- Costs/Benefits: sensitive to relative levels of costs and benefits

sklearn models

Want to point out the following:

- Because of 0/1 classification, confusion matrix lists negatives then positives
- The productive probabilities has two columns, one for each class
 - The first column is for '0' class
 - The second column is for '1' class