Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

INTRODUCTION TO BIG DATA

Juan D. Velásquez
Felipe E. Vildoso

Ingeniería Industrial

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

CHAPTER 2

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

LECTURE 11

Chapter 2

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

LAST CLASS WE SAW...

Decision Analytic Thinking

EVALUATING CLASSIFIERS

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$= 1 - \text{error rate}$$

$$F - measure = 2\left(\frac{Precision\ x\ Recall}{Precision + Recall}\right)$$

EVALUATING CLASSIFIERS: THE CONFUSION MATRIX

	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

A KEY ANALYTICAL FRAMEWORK

Other Evaluation Metric

EXPECTED VALUE

The expected value computation provides a **framework** that is useful in **organizing thinking about data-analytic problems**.

It decomposes data-analytic thinking into:

- the structure of the problema.
- the elements of the analysis that can be extracted from the data.
- the elements of the analysis that need to be acquired from other sources.

$$EV = P(o_1)v(o_1) + P(o_2)v(o_2) + P(o_3)v(o_3) + \cdots$$

EXAMPLE: EXPECTED VALUE FRAMEWORK IN USE PHASE

Online marketing:

- Expected Benefit of Targeting

$$EV = p_R(x)v_R + (1 - p_R(x))v_{NR}$$

- Product Price $200
- Product Cost $100
- Targeting Cost $1

$$EV = p_R(x)\$(100 - 1) - (1 - p_R(x))\$1 > 0$$

EXAMPLE: EXPECTED VALUE FRAMEWORK IN USE PHASE

Online marketing:

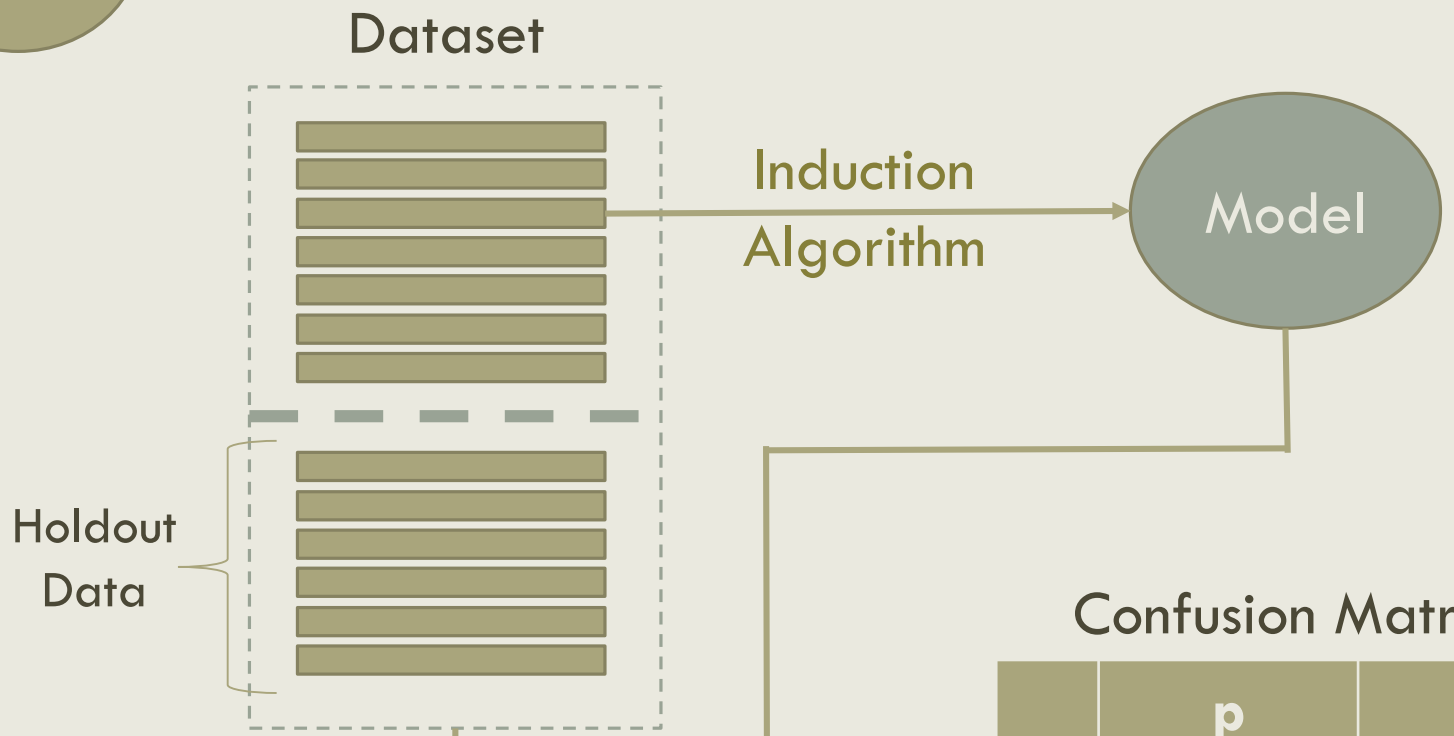
$$EV = p_R(x)\$(100 - 1) - (1 - p_R(x))\$1 > 0$$

$$p_R\,(x) > 0{,}01$$

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

USING EXPECTED VALUE TO FRAME CLASSIFIER EVALUATION

1



Confusion Matrix

	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives

2

Confusion Matrix

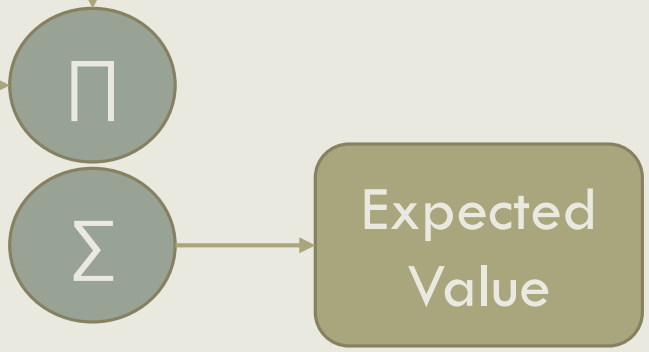
	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives

Expected Rates
(Matrix of Probabilities)

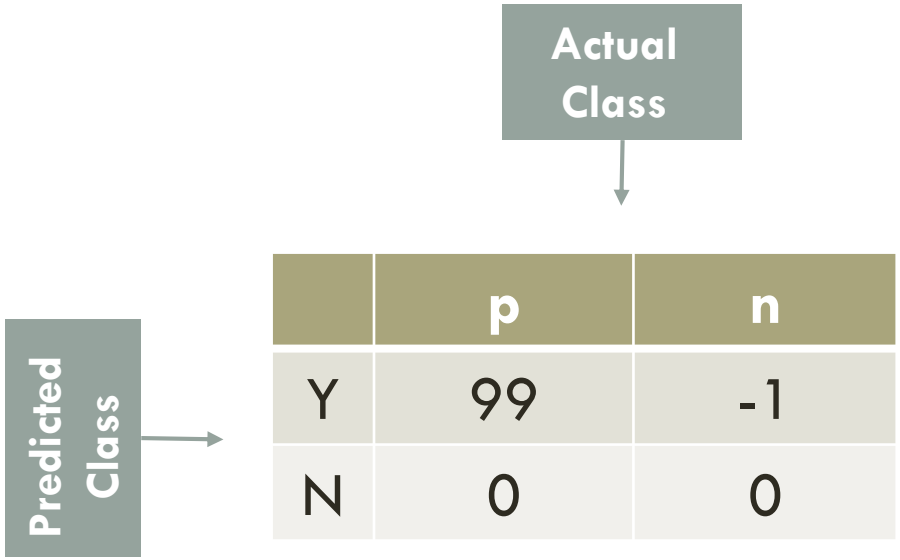
	p	n
Y	TP rate P(Y,p)	FP rate P(Y,n)
N	FN P(N,p)	TN P(N,n)

Cost/Benefit Information

	p	n
Y	b(Y,p)	c(Y,n)
N	c(N,p)	b(N,n)



A COST-BENEFIT MATRIX FOR THE MARKETING EXAMPLE



EXPECTED VALUE

We can get something like this:

$$Expected\ Profit = p(Y,p)b(Y,p) + p(N,p)b(N,p) + p(N,n)b(N,n) + p(Y,n)b(Y,n)$$

CONDITIONAL PROBABILITY

A rule of basic probability is:

$$p(Y, p) = p(y)p(x|y)$$

USING CONDITIONAL PROBABILITY…

$$Expected\ Profit = p(Y,p)b(Y,p) + p(N,p)b(N,p) + p(N,n)b(N,n) + p(Y,n)b(Y,n)$$

$$Expected\ Profit = p(Y|p)p(p)b(Y,p) + p(N|p)p(p)b(N,p) + p(N|n)p(n)b(N,n) + p(Y|n)p(n)b(Y,n)$$

$$Expected\ Profit = p(p)\,[p(Y|p)b(Y,p) + p(N|p)b(N,p)] + p(n)[p(N|n)b(N,n) + p(Y|n)b(Y,n)]$$

USING EXPECTED VALUE TO FRAME CLASSIFIER EVALUATION

$T = 110$

$P = 0.61$

$p(p) = 0.55$

$p(Y|p) = \frac{56}{61} = 0.92$

$p(N|p) = \frac{5}{61} = 0.8$

$N = 49$

$p(n) = 0.45$

$p(Y|n) = \frac{7}{49} = 0.14$

$p(N|n) = \frac{42}{49} = 0.86$

$$Expected\ Profit = p(p)\,[p(Y|p)b(Y,p) + p(N|p)b(N,p)] + p(n)[p(N|n)b(N,n) + p(Y|n)b(Y,n)]$$

USING EXPECTED VALUE TO FRAME CLASSIFIER EVALUATION

$T = 110$

$P = 0.61$

$p(p) = 0.55$

$p(Y|p) = \frac{56}{61} = 0.92$

$p(N|p) = \frac{5}{61} = 0.8$

$N = 49$

$p(n) = 0.45$

$p(Y|n) = \frac{7}{49} = 0.14$

$p(N|n) = \frac{42}{49} = 0.86$

$$Expected\ Profit = 0.55\ [0.92\ x\ 99 + 0.8\ x\ 0] + 0.45\ [0.86\ x\ 0 + 0.14\ x(-1)] \approx \$50.04$$

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

OVERFITTING

OVERFITTING

Finding chance occurrences in data that look like interesting patterns, but which **do not generalize**, is called over-fitting the data.

> *"If you torture the data long enough, it will confess"*

We want models to apply not just to the exact training set but to the general population from which the training data came.

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

VISUALIZING MODEL PERFORMANCE

RANKING INSTEAD OF CLASSIFYING

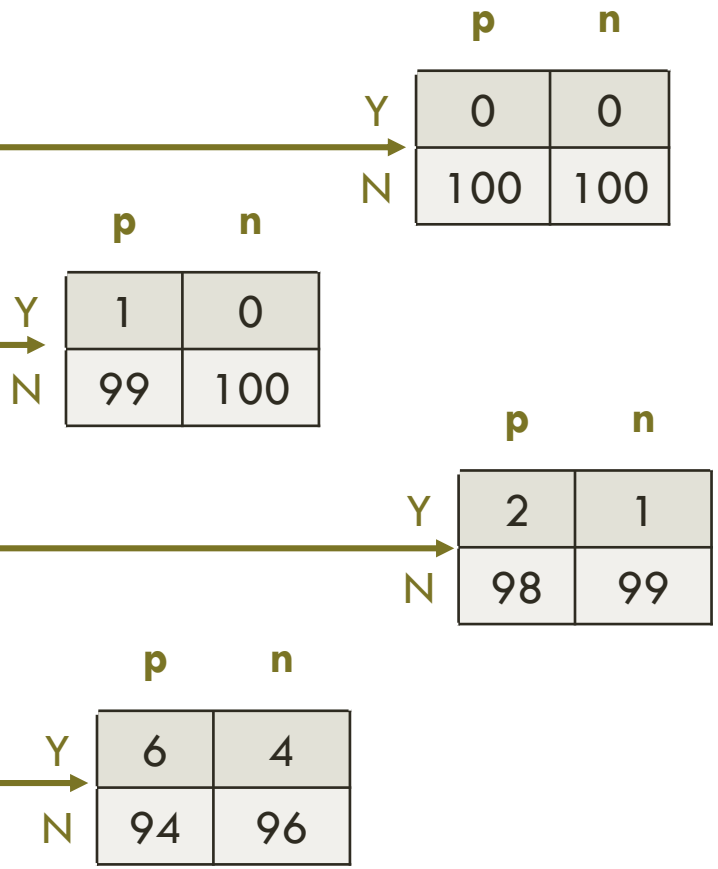
Recall the score assigned by a model can be used to compute a decision for each individual case based on its expected value: Classify.

A different strategy for making decisions is to *rank* a set of cases by these scores, and then take actions on the cases at the top of the ranked list.

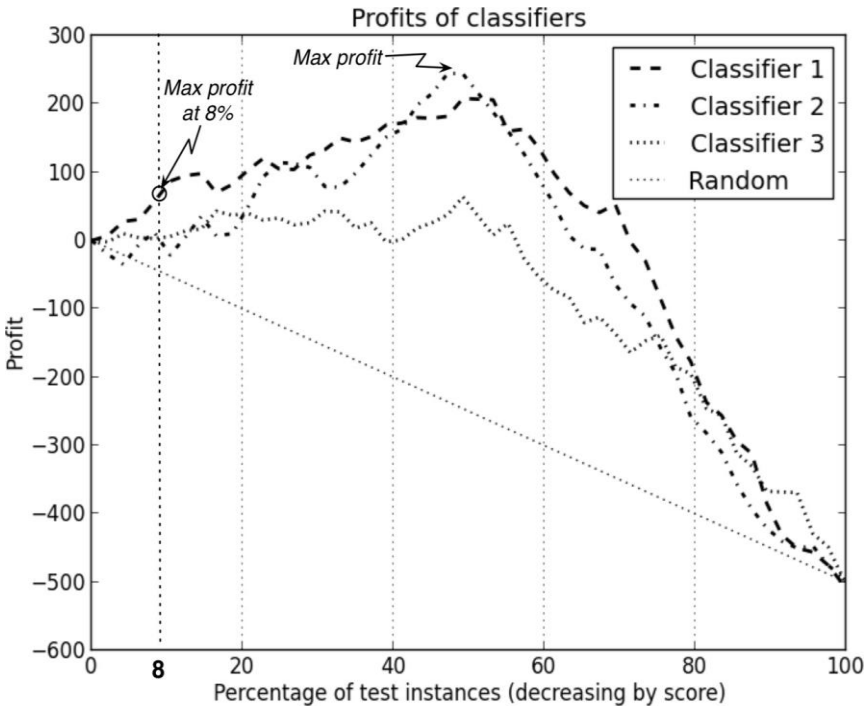
Instead of deciding each case separately, we may decide to **take the top *n* cases**.

RANKING INSTEAD OF CLASSIFYING

Instance Description	True Class	Score
…	p	0.99
…	p	0.98
…	n	0.96
…	n	0.90
…	p	0.88
…	n	0.87
…	.	.
…	.	.
…	p	0.65
…	.	.
…	.	.
…	.	.



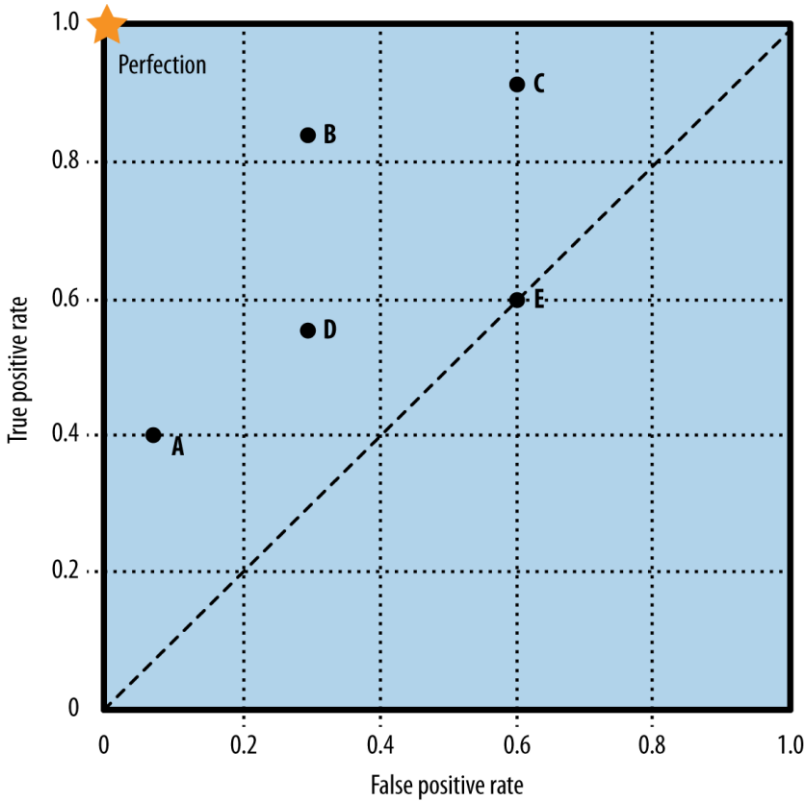
PROFIT CURVES



Each threshold, i.e., each set of predicted positives and negatives, will have a corresponding confusion matrix.

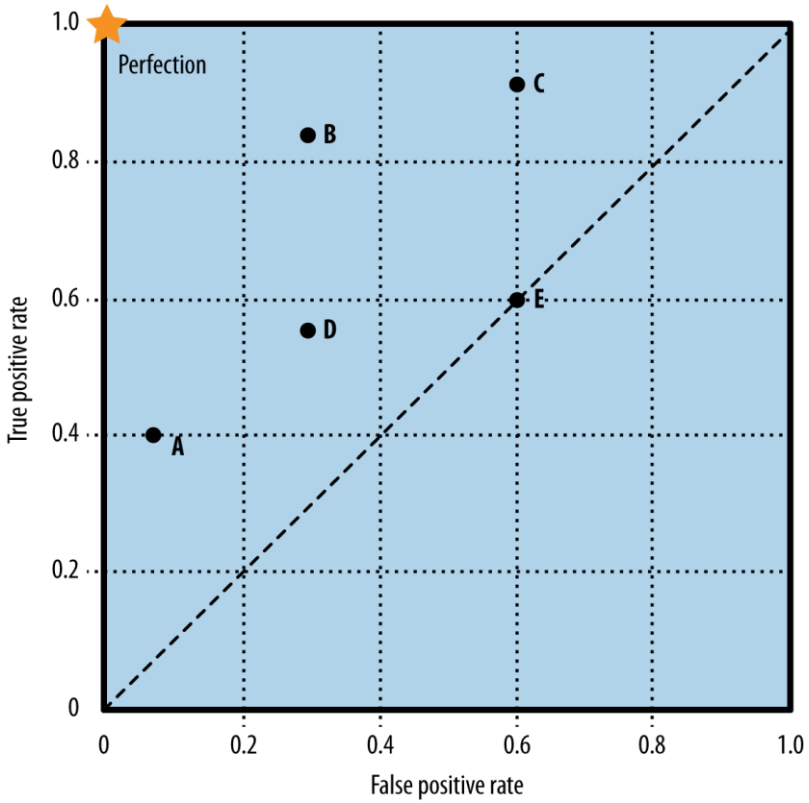
At each cut-point we record the percentage of the list predicted as positive and the corresponding **estimated profit**. Graphing these values gives us a profit curve.

ROC GRAPHS



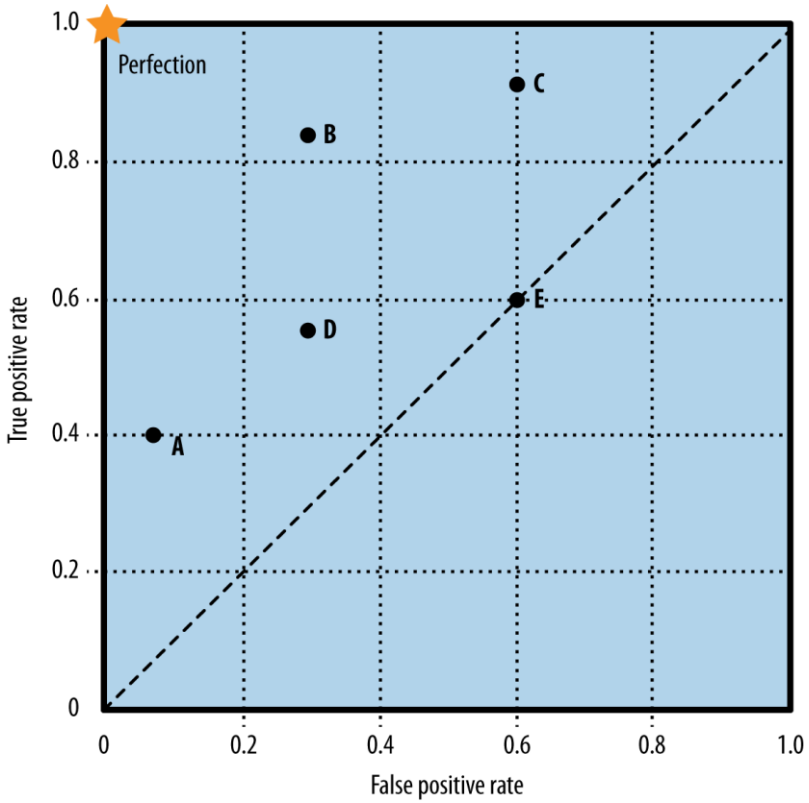
- (0, 0) : Never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives.
- (1, 1): Unconditionally issuing positive classifications.
- (0, 1): Perfect classification.

ROC GRAPHS



- Diagonal line: Policy of guessing a class.
- **E**'s performance at (0.6, 0.6) is **virtually random**.
- Note that no classifier should be in the lower right triangle of a ROC graph. This represents performance that is worse than random guessing.

ROC GRAPHS



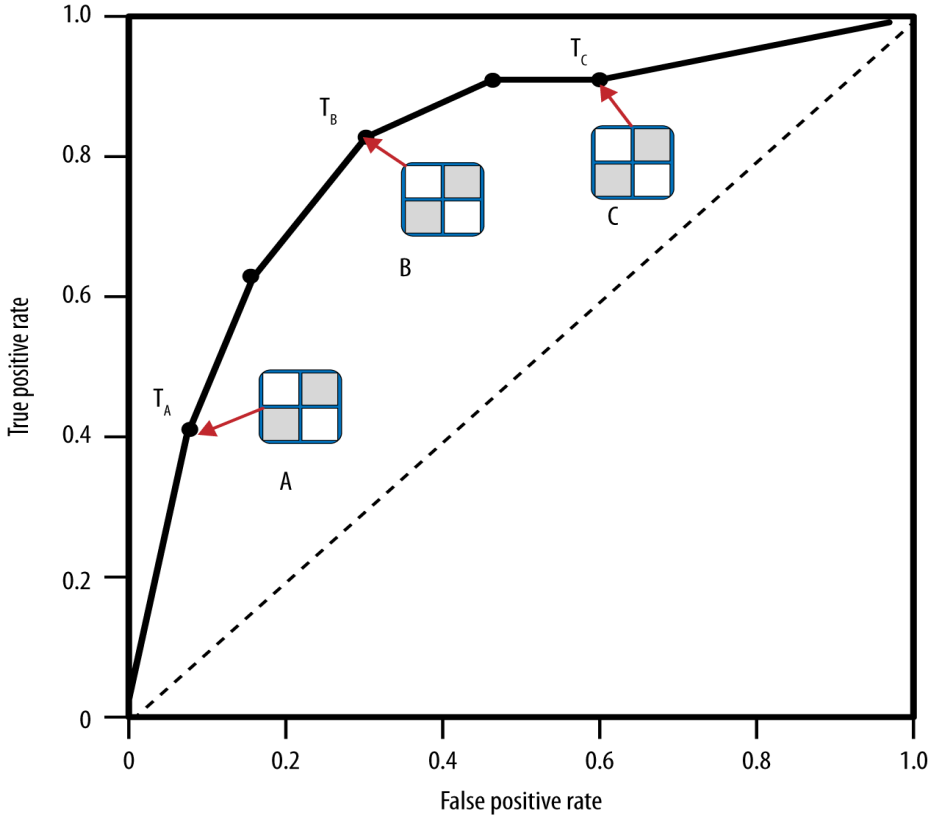
- One point in ROC space is superior to another if it is to the northwest of the first.
- Classifiers appearing on the lefthand side of a ROC graph, near the *x* axis, may be thought of as **"conservative"**.
- Classifiers on the upper righthand side of a ROC graph may be thought of as **"permissive"**.

ROC CURVES

As discussed previously, a ranking model can be used with a threshold to produce a discrete (binary) classifier.

If the classifier output is above the threshold, the classifier produces a **Y**, else an **N**.

Each threshold value produces a different point in ROC space.



AREA UNDER THE ROC CURVE (AUC)

Probability that a randomly chosen positive instance will be ranked ahead of a randomly chosen negative instance

The area under a classifier's curve expressed as a **fraction of the unit square**. Its value ranges from zero to one.

AREA UNDER THE ROC CURVE (AUC)

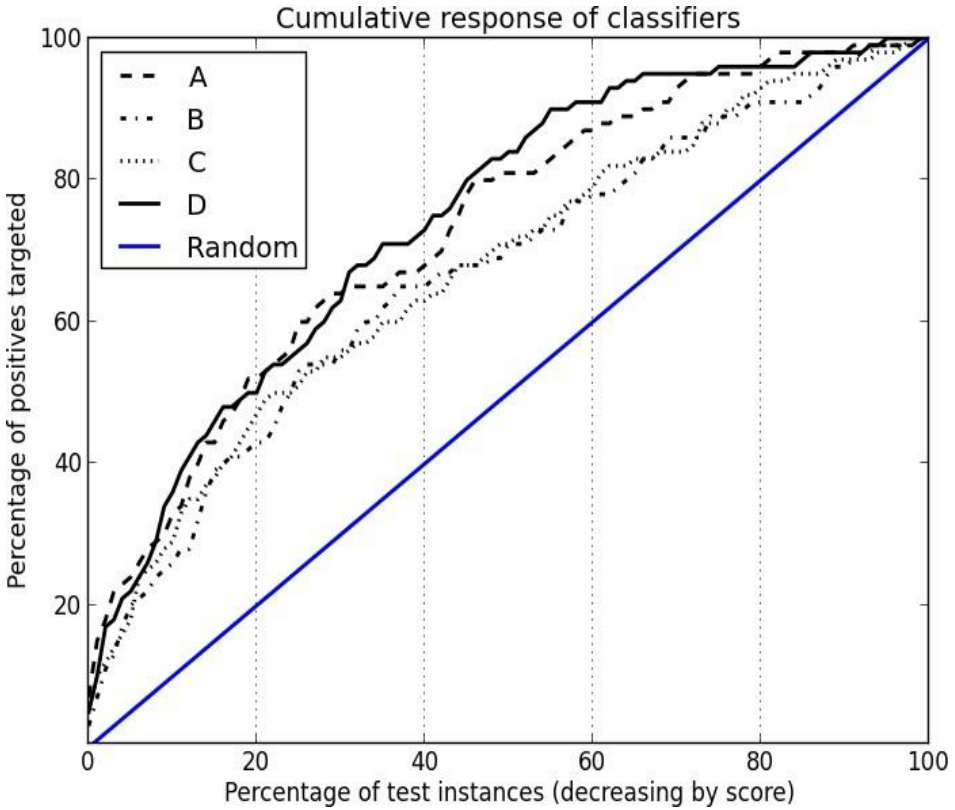
When is it useful?

- When a single number is needed to summarize performance.
- When nothing is known about the operating conditions.

But a **ROC curve** provides **more information** than its area.

CUMULATIVE RESPONSE CURVE

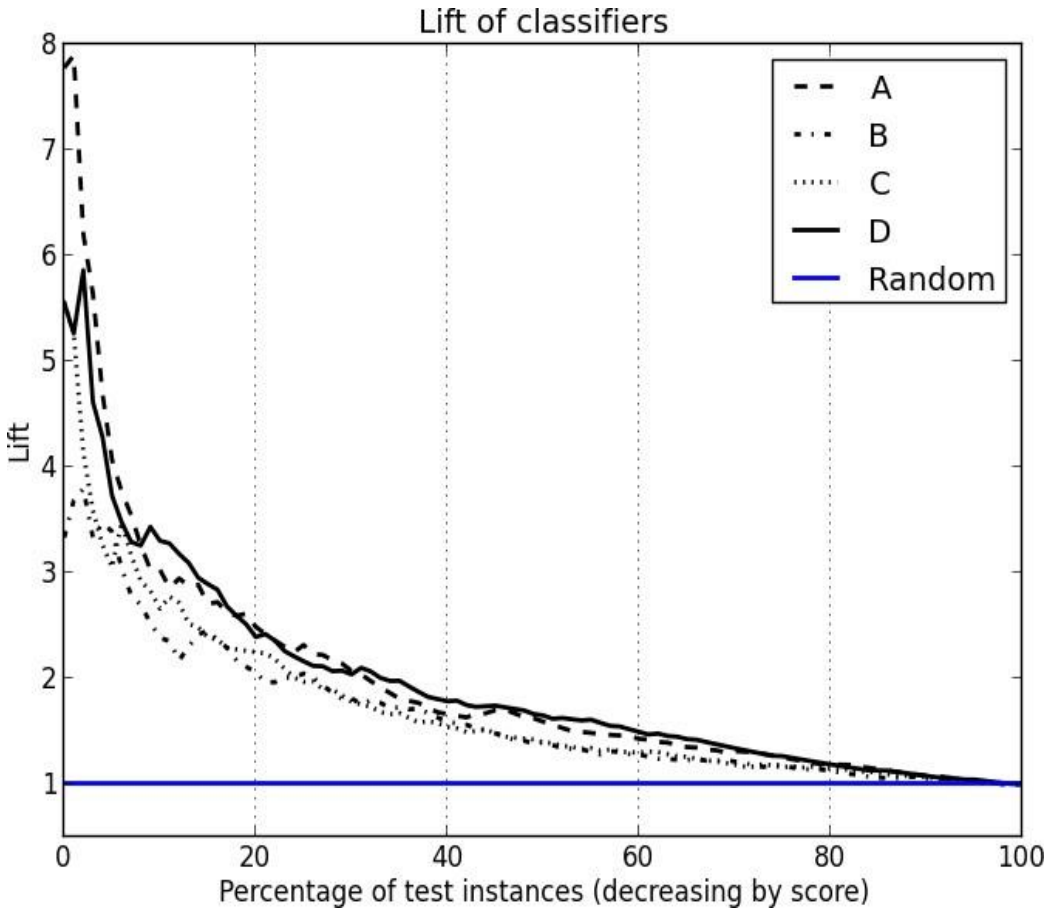
- Percentage of positives correctly classified (tp rate; y axis) vs. the percentage of the population that is targeted (x axis).
- Diagonal line *x*=*y*: *R*andom performance.
- Any classifier above the diagonal is providing some advantage.



LIFT CURVE

- The **lift** of a classifier represents the **advantage it provides over random guessing**.

$$lift = \frac{TP\ rate\ (x)}{x}$$



VISUALIZING MODEL PERFORMANCE

	Requirements		Intuitive?
	Class Priors	Costs and Benefits	
Profit Curves	YES	YES	YES
ROC Curves	NO	NO	NO
Cumulative Response	YES	NO	YES
Lift Curves	YES	NO	YES

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

QUESTIONS?

SEE YOU ON THE NEXT CLASS!

EXAMPLE

EXAMPLE: PERFORMANCE EVALUATION

Training Set:

Model	Accuracy
Classification Tree	95%
Logistic Regression	93%
k-Nearest Neighbors	100%
Naïve Bays	76%

Test Set:

Model	Accuracy	AUC
Classification Tree	91.8%±0.0	0.614±0.014
Logistic Regression	93.0%±0.1	0.574±0.023
k-Nearest Neighbors	93.0%±0.0	0.537±0.015
Naïve Bays	76.5%±0.6	0.632±0.019

EXAMPLE: PERFORMANCE EVALUATION

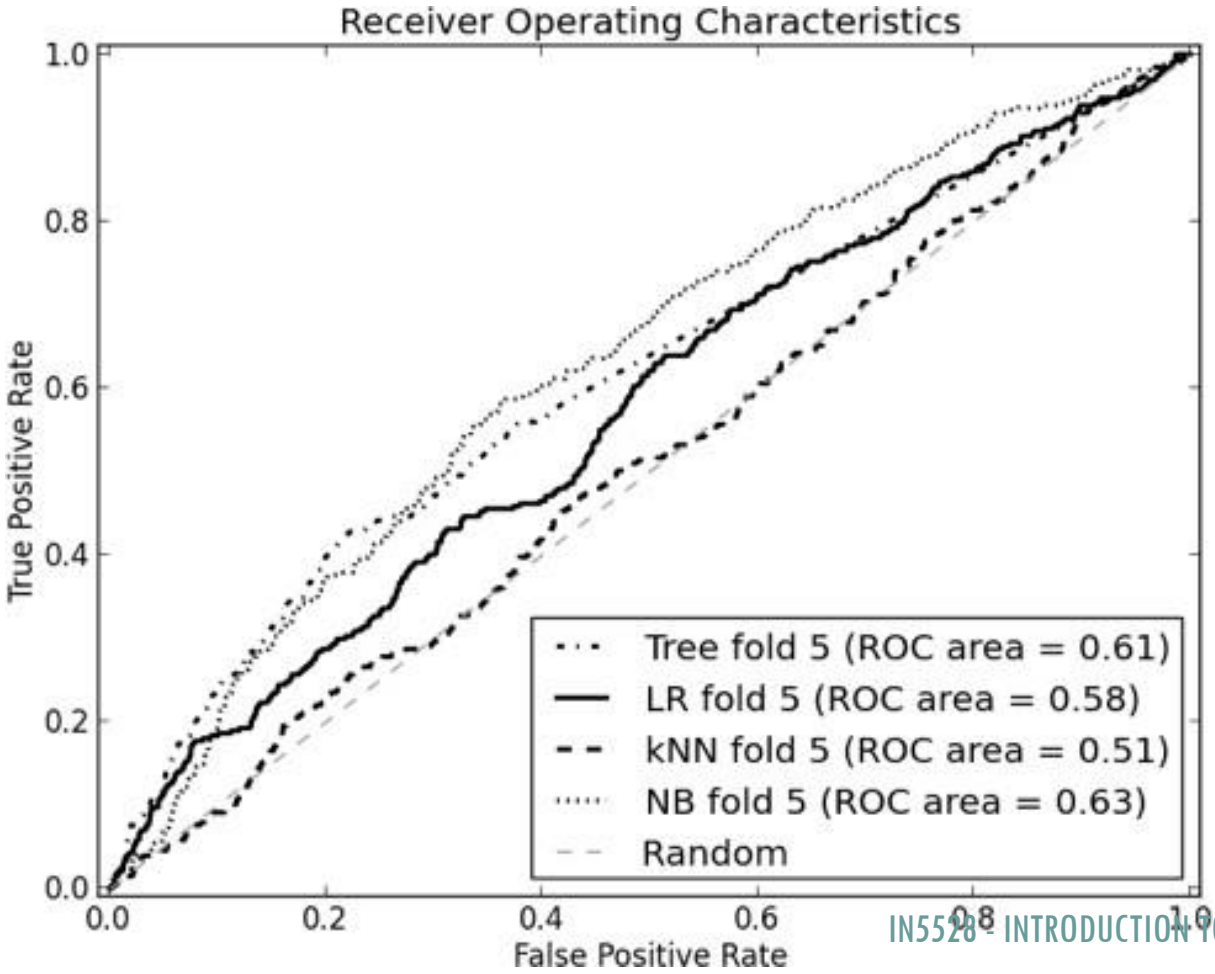
Naïve Bayes confusion matrix:

	p	n
Y	127 (3%)	848 (18%)
N	200 (4%)	3518 (75%)

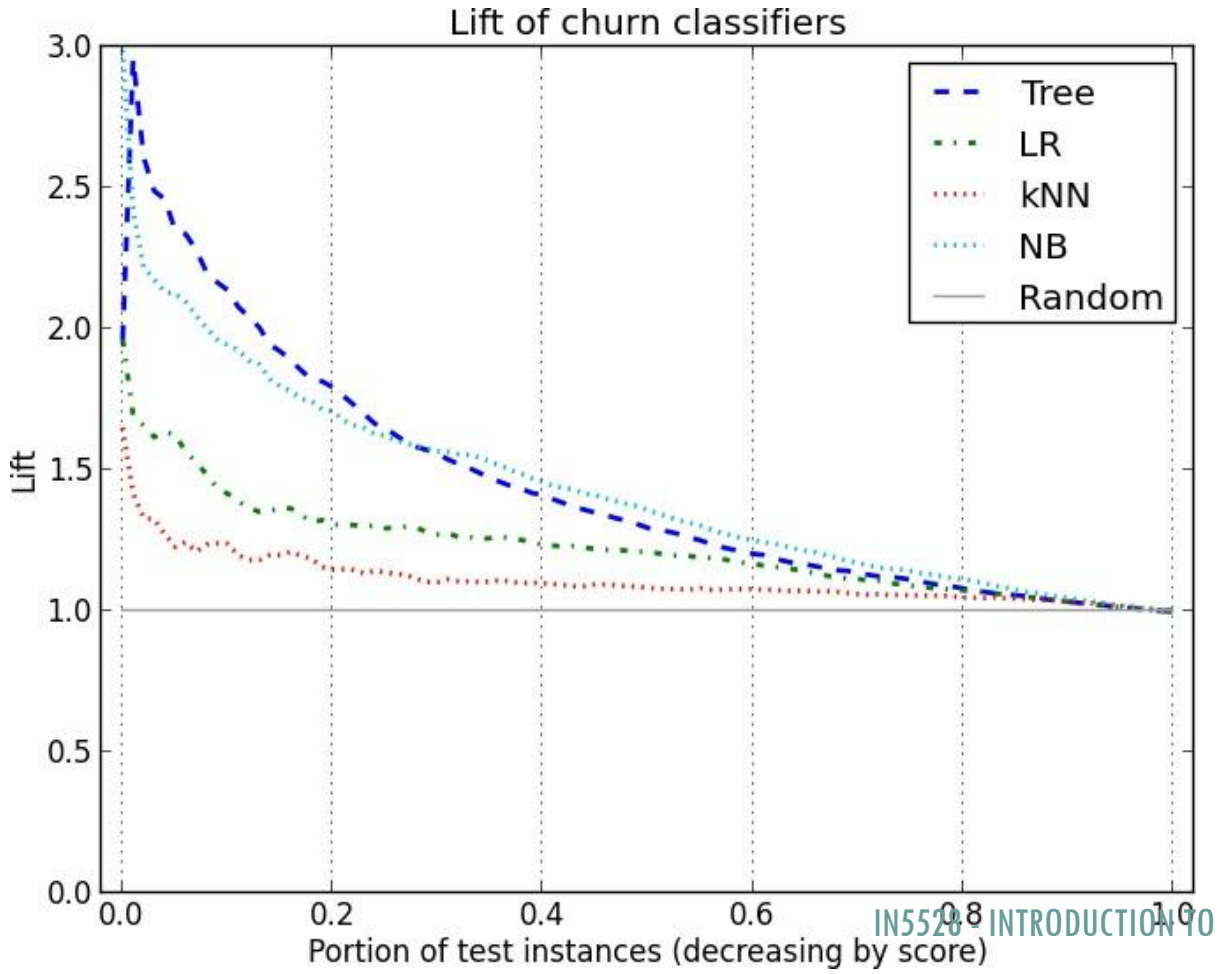
k-Nearest Neighbors confusion matrix:

	p	n
Y	3 (0%)	15 (0%)
N	324 (7%)	4351 (93%)

EXAMPLE: ROC CURVE



EXAMPLE: LIFT CURVE



EXAMPLE: PROFIT CURVES