Data Science for Business
Lecture #5
Evaluating Models

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

Outline

Lift Charts

ROC Curves and AUC

Confusion Matrices and Cost

Computing Lift Charts

Source: Greg Piatetsky-Shapiro,

KDnuggets Data Mining Course

Direct Marketing Evaluation

Accuracy on the entire dataset is not the right measure

Approach

- develop a target model
- score all prospects and rank them by decreasing score
- select top P% of prospects for action

How to decide what is the best selection?

Direct Marketing Paradigm

Find most likely prospects to contact

Not everybody needs to be contacted

Number of targets is usually much smaller than number of prospects

Typical Applications

- retailers, catalogues, direct mail (and e-mail)
- customer acquisition, cross-sell, attrition prediction

Lift Charts

In practice, decisions are usually made by comparing possible scenarios taking into account different costs.

Example:

- Promotional mailout to 1,000,000 households. If we mail to all households, we get 0.1% respond (1000).
- Data mining tool identifies (a) subset of 100,000 households with 0.4% respond (400); or (b) subset of 400,000 households with 0.2% respond (800);
- Depending on the costs we can make final decision using lift charts!
- A lift chart allows a visual comparison.

Model-Sorted List

No	Score	Target	CustID	Age
1	0.97	Y	1746	…
2	0.95	N	1024	…
3	0.94	Y	2478	…
4	0.93	Y	3820	…
5	0.92	N	4897	…
…	…		…	…
99	0.11	N	2734	…
100	0.06	N	2422	

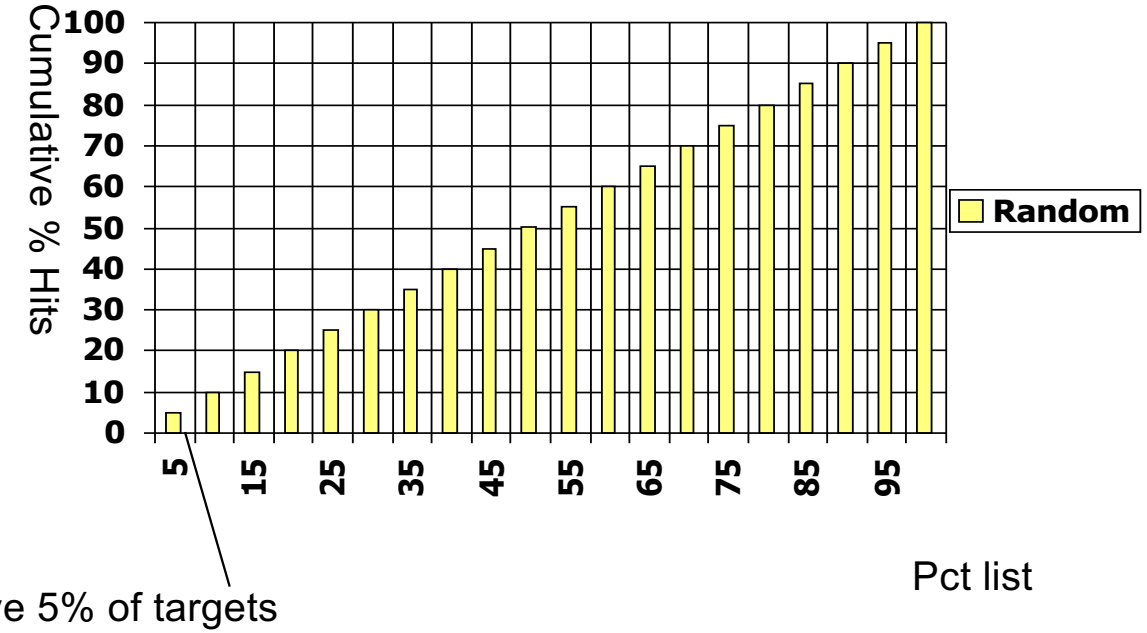
3 hits in top 5% of the list

If there 15 targets overall, then top 5 has 3/15=20% of targets

Use a model to assign score to each customer
Sort customers by decreasing score
Expect more targets (hits) near the top of the list

CPH (Cumulative Pct Hits)

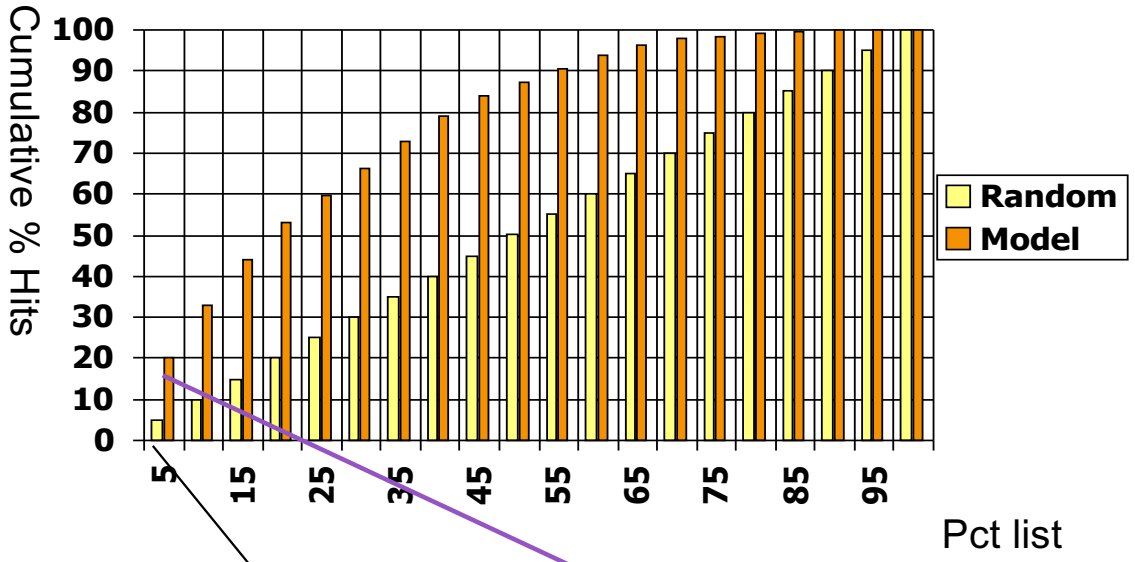
Definition:
CPH(P,M)
= % of all targets in the first P% of the list scored by model M
CPH frequently called Gains



Q: What is expected value for CPH(P,Random) ?

A: Expected value for CPH(P,Random) = P

CPH: Random List vs Model-ranked list



5% of random list have 5% of targets,

but 5% of model ranked list have 21% of targets CPH(5%,model)=21%.

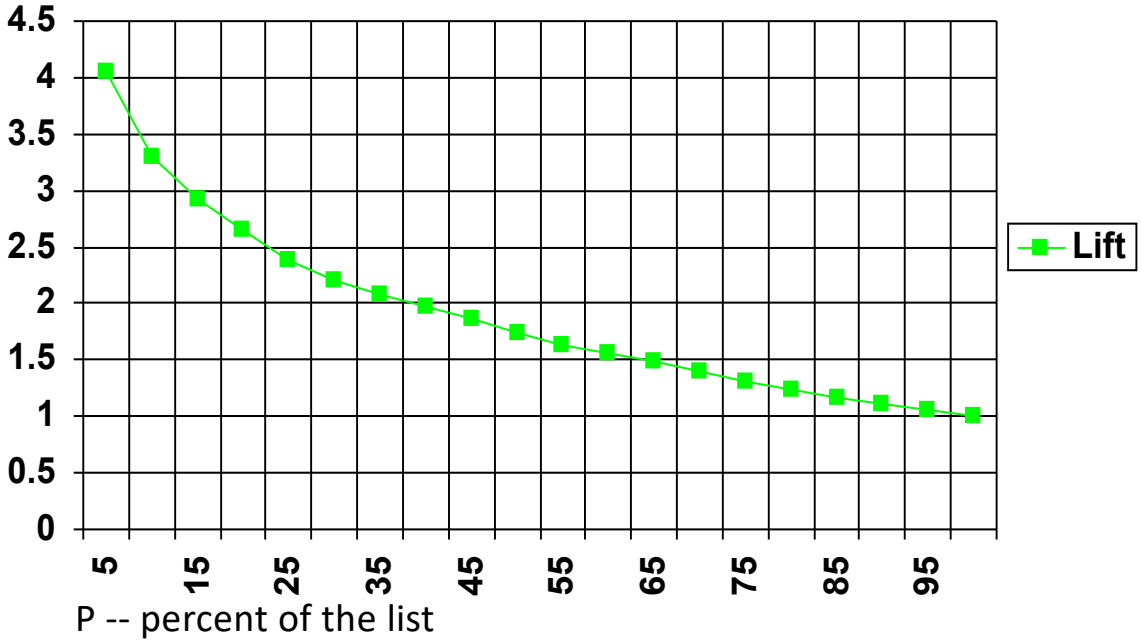
Lift

Lift(*P,M*) = CPH(*P,M*) / *P*

Lift (at 5%)

= 21% / 5%

= 4.2
better
than random



Note: Some (including Witten & Eibe) use "Lift" for what we call CPH.

Lift Properties

Q: Lift(P,Random) =

- ***A:*** 1 (expected value, can vary)

Q: Lift(100%, M) =

- ***A:*** 1 (for any model M)

Q: Can lift be less than 1?

- ***A***: yes, if the model is inverted (all the non-targets precede targets in the list)

Generally, a better model has higher lift

Example: Lift-Table for Intuit Example

The average probability is 8.1%.

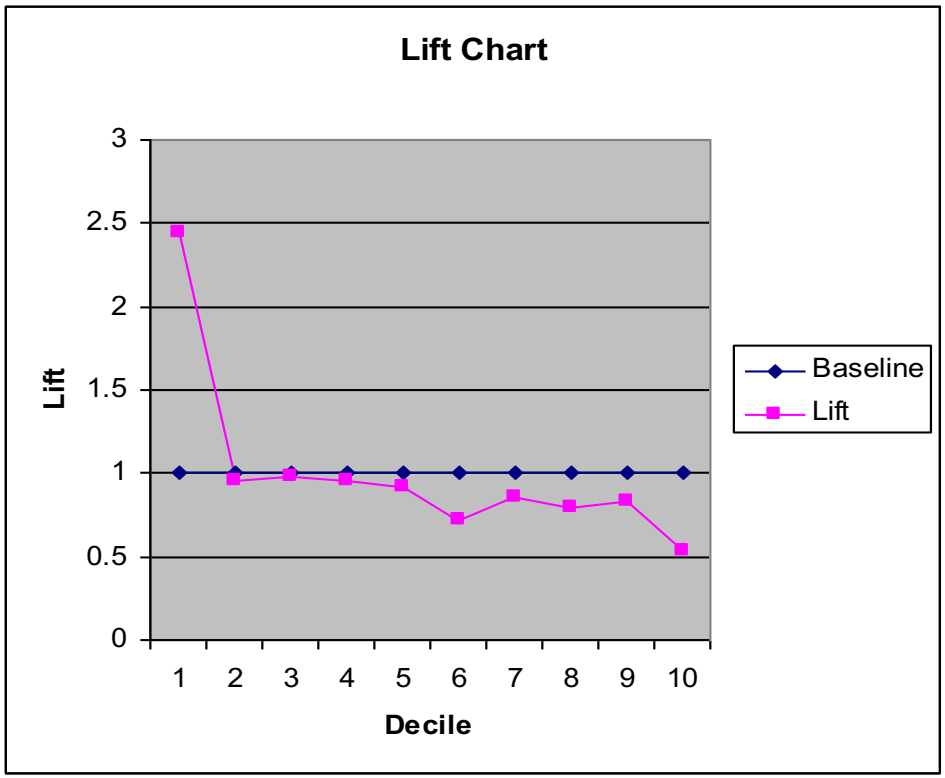
Decile	Prob	Lift
1	19.8%	2.44
2	7.8%	.96
3	8.0%	.98
4	7.7%	.95
5	7.5%	.92
6	5.8%	.72
7	7.0%	.86
8	6.4%	.79
9	6.7%	.83
10	4.3%	.53

Lift for Decile 1 = 19.8% / 8.1% = 2.44

Lift Chart

If we plot lift against the decile we get a lift chart.

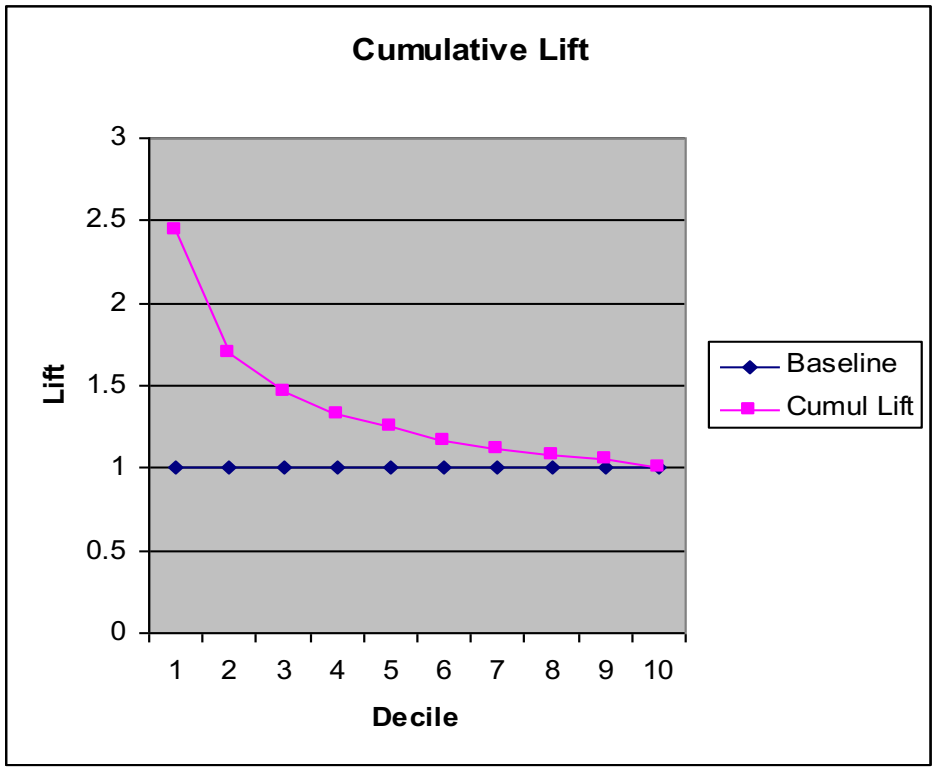
Notice lift drops below chance after the first decile.



Cumulative Lift Chart

Alternatively plot the cumulative lift against the decile we get a lift chart.

Notice that the cumulative lift is always greater than 1.

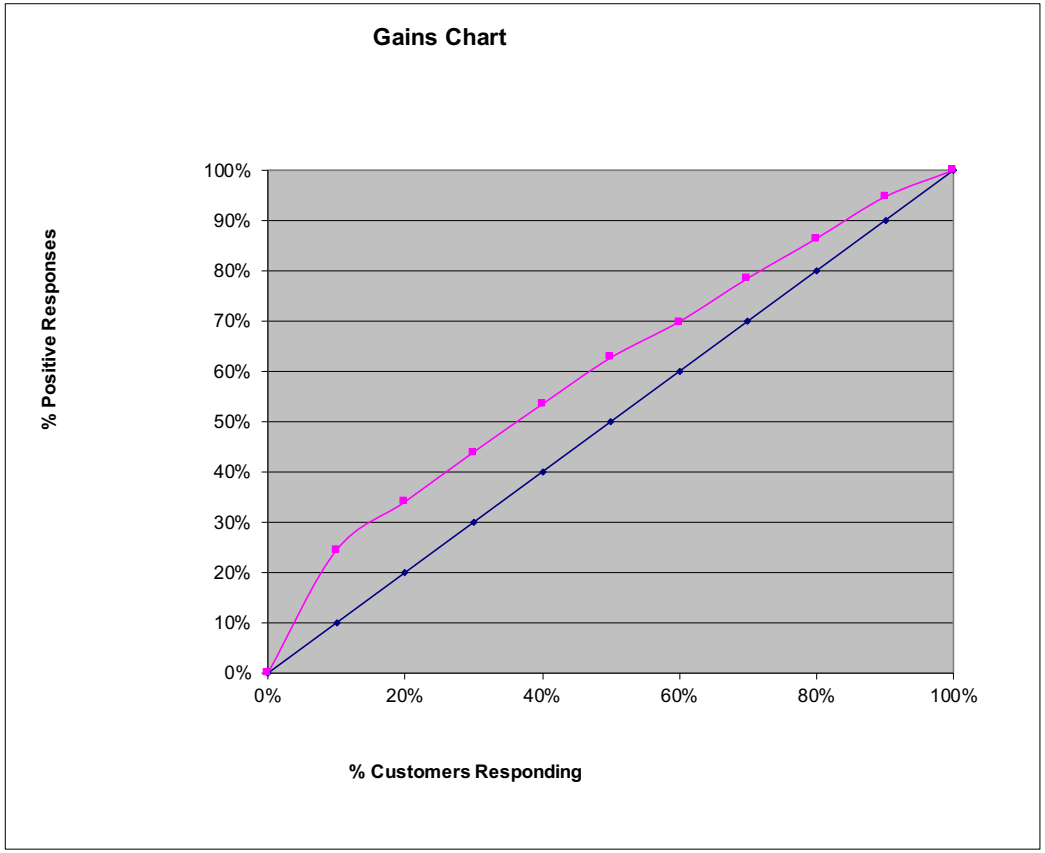


Gains Chart

Compute the cumulative probability of positive responses and plot it against the percent of customers responding.

For example, in the previous slide we know decile 1 accounts has 19.8% positive responses, which is 19.8% / 81% = 24% of the positive responses.

Notice the upper right corner is always (100% , 100%)



Receiver Operating Characteristic (ROC) Curve

Measuring the diagnostic ability of a classifier

ROC curves

ROC curves are similar to gains charts

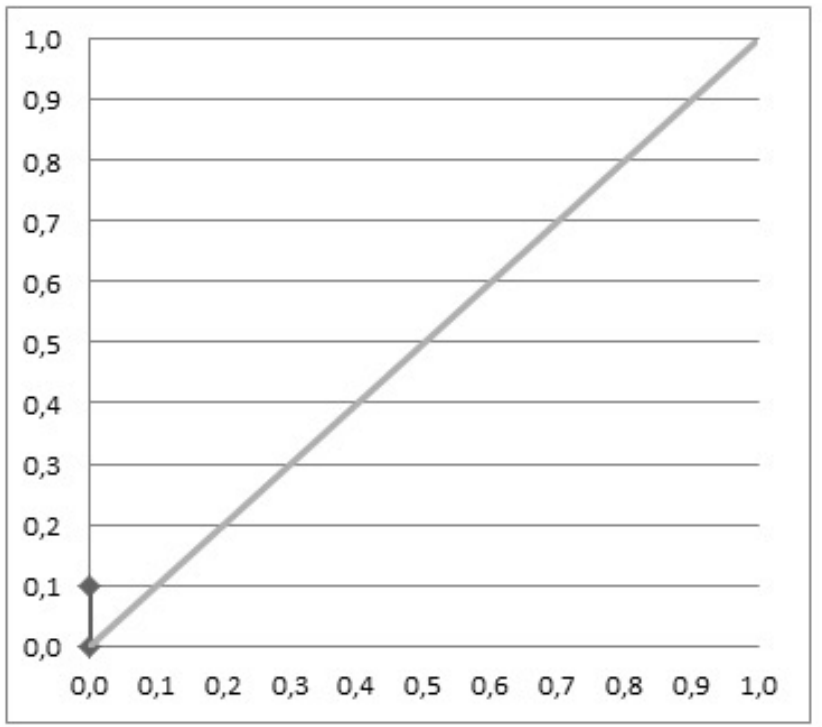
- Stands for “receiver operating characteristic”
- Used in signal detection to show tradeoff between hit rate and false alarm rate over noisy channel

Differences from gains chart:

- y axis shows percentage of true positives in sample *rather than absolute number*
- x axis shows percentage of false positives in samples *rather than sample size*

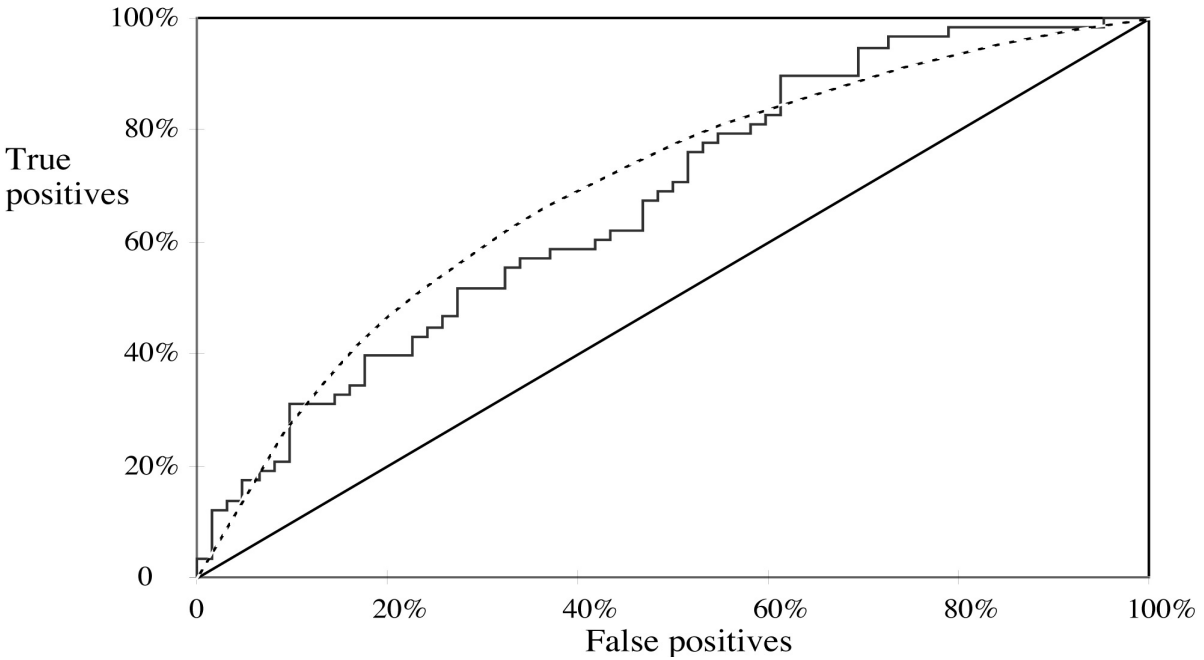
Illustrating ROC computation

#	C	Score
1	P	0,9
2	P	0,8
3	N	0,7
4	P	0,6
5	P	0,55
6	P	0,54
7	N	0,53
8	N	0,52
9	P	0,51
10	N	0,505
11	P	0,4
12	N	0,39
13	P	0,38
14	N	0,37
15	N	0,36
16	N	0,35
17	P	0,34
18	N	0,33
19	P	0,3
20	N	0,1



http://mlwiki.org/index.php/ROC_Analysis

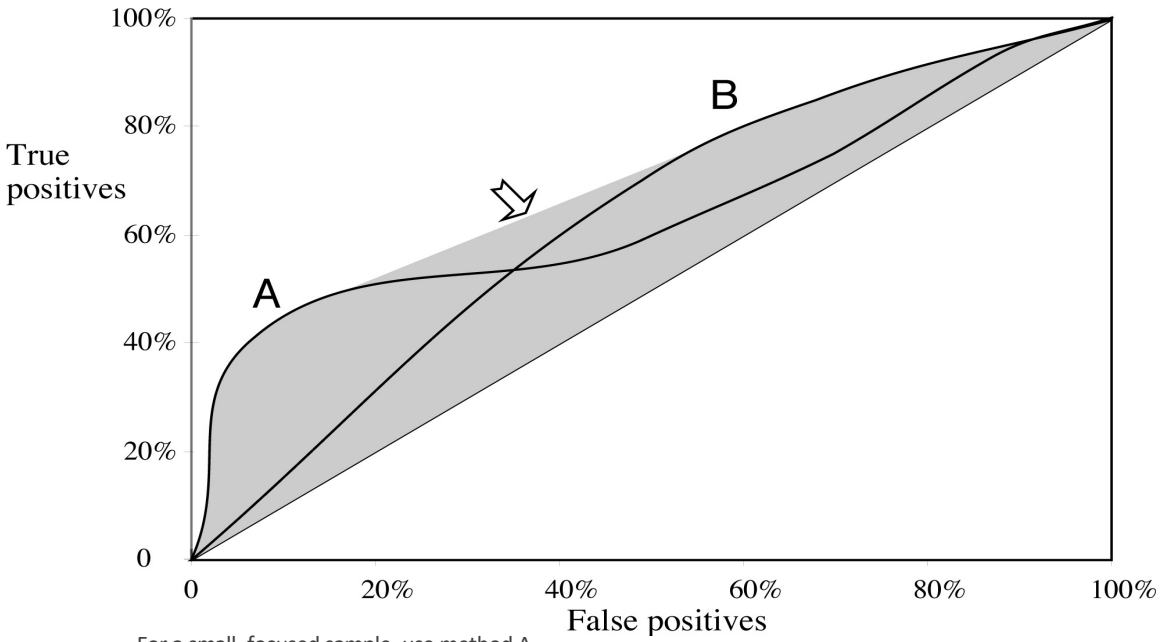
A sample ROC curve



Jagged curve—one set of test data

Smooth curve—use cross-validation

ROC curves for two schemes



For a small, focused sample, use method A

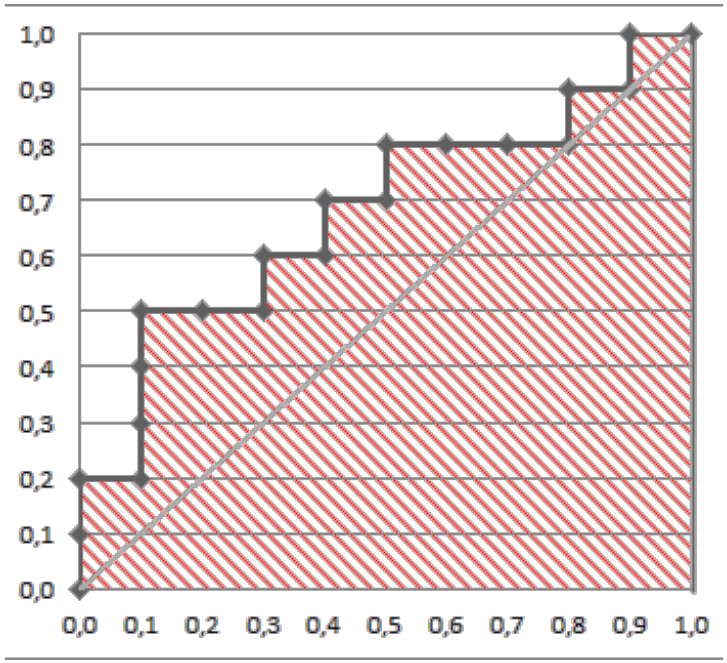
For a larger one, use method B

In between, choose between A and B with appropriate probabilities

Area Under Curve (AUC)

Area Under ROC Curve

- Measure for evaluating the performance of a classifier
- It's the area under the ROC Curve
- Total area is 100% so AUC = 1 is for a perfect classifier for which all positive come after all negatives
- AUC = 0.5, then classifier is randomly ordered
- AUC = 0, then all negative come before all positive
- Typically we don't have classifiers with AUC < 0.5, since we usually will not do worse than random guessing



Computing ROC curves and AUC in R

```
# install the ROCR library
if (!require(ROCR)) {install.packages("ROCR"); library(ROCR)}

# three commands for drawing ROC curve using ROCR
pred = prediction(lpredict,disease$cd)  # compute predictions using "prediction"
perf = performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf, col=rainbow(10))
abline(a=0, b= 1)

# command to compute AUROC using ROCR
auc.tmp = performance(pred,"auc")
(auc = as.numeric(auc.tmp@y.values))
```

Evaluating our Model from a Business Viewpoint

Measuring predictive ability

Can count number (or percent) of correct predictions or errors

But in business applications different errors (different decisions) have different *costs* and *benefits* associated with them

Usually need either

- to rank cases or
- to compute probability of the target

(class probability estimation rather than just classification)

Evaluating Classifiers

Assume that we test a classifier on some test set and we derive at the end the following *confusion matrix*:

Actual class		*Predicted class* Pos	*Predicted class* Neg	
	Pos	*TP*	*FN*	*P*
	Neg	*FP*	*TN*	*N*

Metrics for Classifier's Evaluation

Accuracy = (*TP*+*TN*)/(*P*+*N*)
Error = (*FP*+*FN*)/(*P*+*N*)
Precision = *TP*/(*TP*+*FP*)
Recall/TP rate = *TP*/*P*
FP Rate = *FP*/*N*

		Predicted class		
		Pos	Neg	
Actual class	Pos	*TP*	*FN*	*P*
	Neg	*FP*	*TN*	*N*

Counting the Costs

In practice, different types of classification errors often incur different costs

Examples:

- Terrorist profiling
  - "Not a terrorist" correct 99.99% of the time
- Loan decisions
- Fault diagnosis
- Promotional mailing

Confusion Matrices

Validation data can show you are, say, 90% accurate. What 90% are you getting?

Predicted

Actual		Big Gain	Zero	Big Loss
	Big Gain	50	0	10
	Zero	5	130	5
	Big Loss	10	0	90

Evaluating Confusion Matrices

Previous Example

Actual \ Predicted	Big Gain	Zero	Big Loss
Big Gain	50	10	0
Zero	5	130	5
Big Loss	0	10	90

New Example

Actual \ Predicted	Big Gain	Zero	Big Loss
Big Gain	30	0	30
Zero	0	140	0
Big Loss	0	0	100

Associating Costs (or Benefits) with our Confusion Matrix

Given multiple confusion matrices, you can choose the best by creating a confusion cost matrix:

c(i,j) = cost of assigning category j for true value i

Confusion Cost Matrix

Example:

You are going to send a $5 mailer to all those who are “Big Gain” and a $1 letter to those who are “Zero”. Here are the costs (or benefits) with each outcome:

- Every “Big Gain” you reach with a $5 mailer is worth $20 ($15 net)
- Every “Big Gain” you reach with a $1 letter is worth $10
- Every “Big Loss” reached in any way is worth -$10
- Every “Zero” is worth 0

Predicted

Actual	Big Gain	Zero	Big Loss
Big Gain	15	9	0
Zero	-5	-1	0
Big Loss	-15	-11	0

Calculating Value
by taking the expected cost (or gain)

Predicted

Actual	Big Gain	Zero	Big Loss
Big Gain	50	0	10
Zero	5	130	5
Big Loss	10	0	90

Confusion Matrix

Predicted

Actual	Big Gain	Zero	Big Loss
Big Gain	15	9	0
Zero	-5	-1	0
Big Loss	-15	-11	0

Cost Matrix

SUMPRODUCT with
Cost matrix to get value

Value
= 50 x 15
+ 5 x -5
+ 10 x -15
+ 130 x -1
= 445

Returning to our previous examples

Previous Example

Predicted

Actual	Big Gain	Zero	Big Loss
Big Gain	50	10	0
Zero	5	130	5
Big Loss	0	10	90

Value=575

New Example

Predicted

Actual	Big Gain	Zero	Big Loss
Big Gain	30	0	30
Zero	0	140	0
Big Loss	0	0	100

Value=310

Value of Perfect Information

Can also determine value of “best possible” result by assuming perfect predictions:

In this case: 760 = $\underbrace{15 \times 60}_{\text{Big Gain}} - \underbrace{1 \times 140}_{\text{Zero}}$

Note that the decision rule is not optimal. It would be better to not send to the “zeros”. Perfect Information is generally relative to some decision rule.