NYU

Introduction to Data Science

Center for Data Science
Iddo Drori, Spring 2019

Agenda

- Performance evaluation (40 minutes)
- ROC, AUC, cumulative response, lift curves (40 minutes)
- Collaborative filtering (20 minutes)

Evaluating Performance

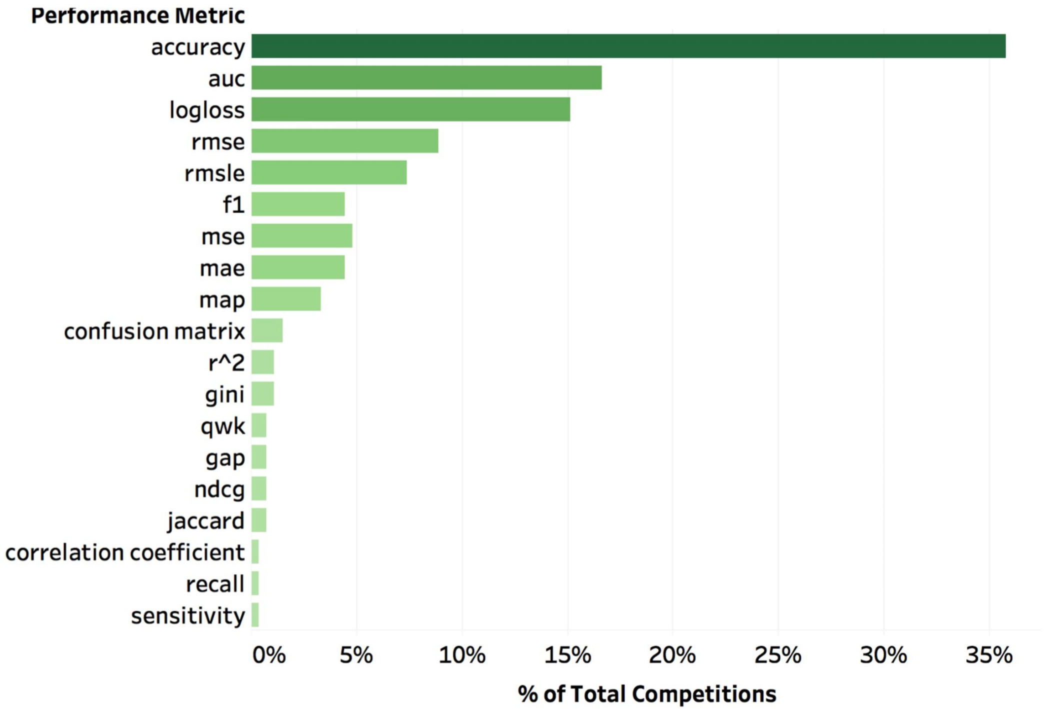
Performance Metrics

Measure performance in a meaningful way

Performance measures differ for various applications

Common themes, issues, and solutions which apply broadly

Performance Metrics



Distribution of Kaggle competitions according to performance metrics

Source: Drori et al, 2018

Accuracy

Fraction of correct predictions

$$accuracy = \frac{\#\, correct\ decisions}{total\ \#\ of\ decisions} = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$error\ rate = 1 - accuracy$$

may be very misleading

Binary Classification

Positive and negative classes

True positive (TP): correctly predicted as positive

True negative (TN): correctly predicted as negative

False positive (FP): incorrectly predicted as positive (type 1 error)

False negative (FN): incorrectly predicted as negative (type 2 error)

Confusion Matrix and Probabilities

actual

predicted

	p	n
Y	True positive (56)	False positive (7)
N	False negative (5)	True negative (42)

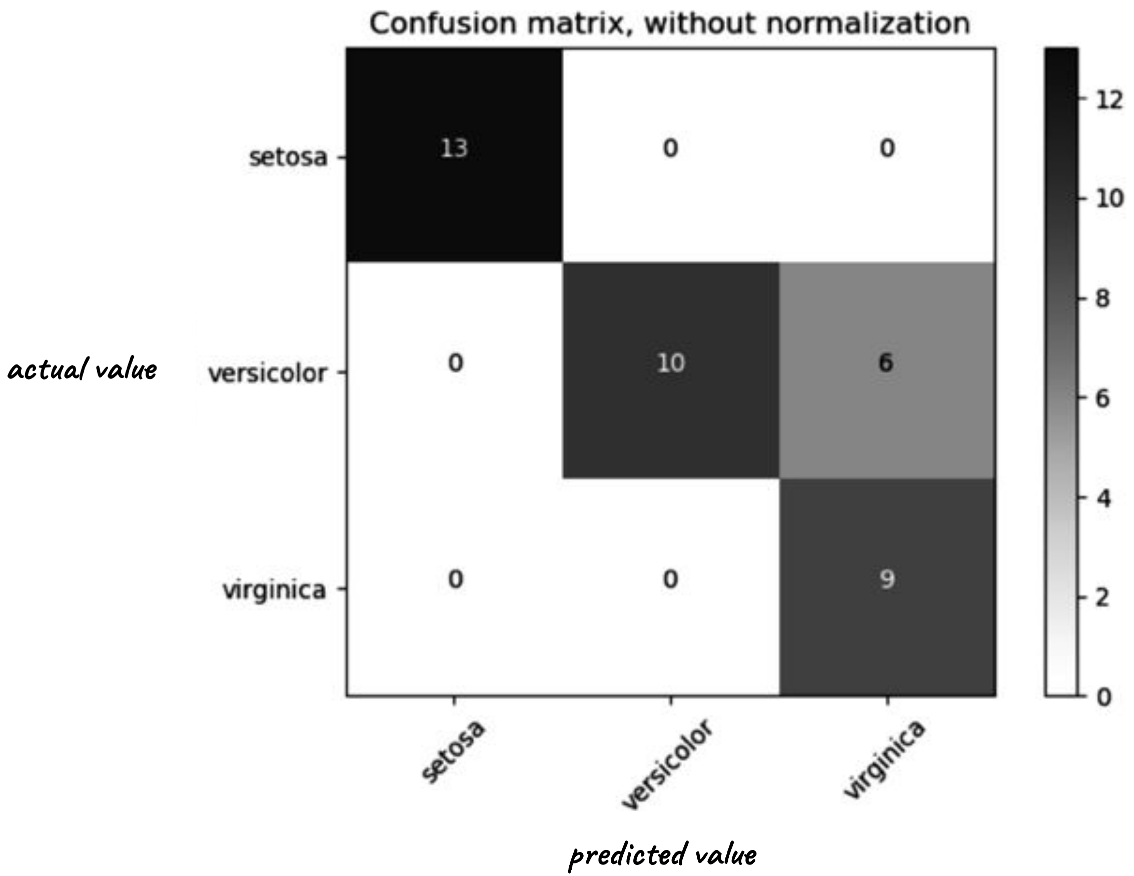
$$T = 110$$

$$p(\mathbf{Y},\mathbf{p}) = 56/110 = 0.51 \quad p(\mathbf{Y},\mathbf{n}) = 7/110 = 0.06$$

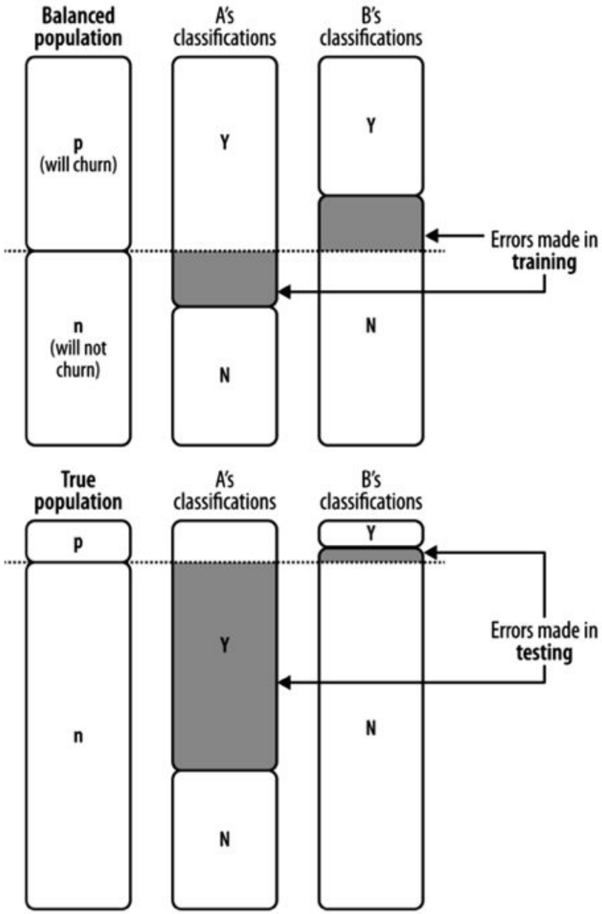
$$p(\mathbf{N},\mathbf{p}) = 5/110 = 0.05 \quad p(\mathbf{N},\mathbf{n}) = 42/110 = 0.38$$

Source: DSB

Multi-class Confusion Matrix



Training and Testing Populations



Source: DSB

Confusion Matrix and Rates

actual

predicted

	p	n
Y	True positive (56)	False positive (7)
N	False negative (5)	True negative (42)

$T = 110$

$P = 61$ | $N = 49$

$p(\mathbf{p}) = 0.55$ | $p(\mathbf{n}) = 0.45$

$tp\ rate = 56/61 = 0.92$ | $fp\ rate = 7/49 = 0.14$

$fn\ rate = 5/61 = 0.08$ | $tn\ rate = 42/49 = 0.86$

Source: DSB

Expected Value Example

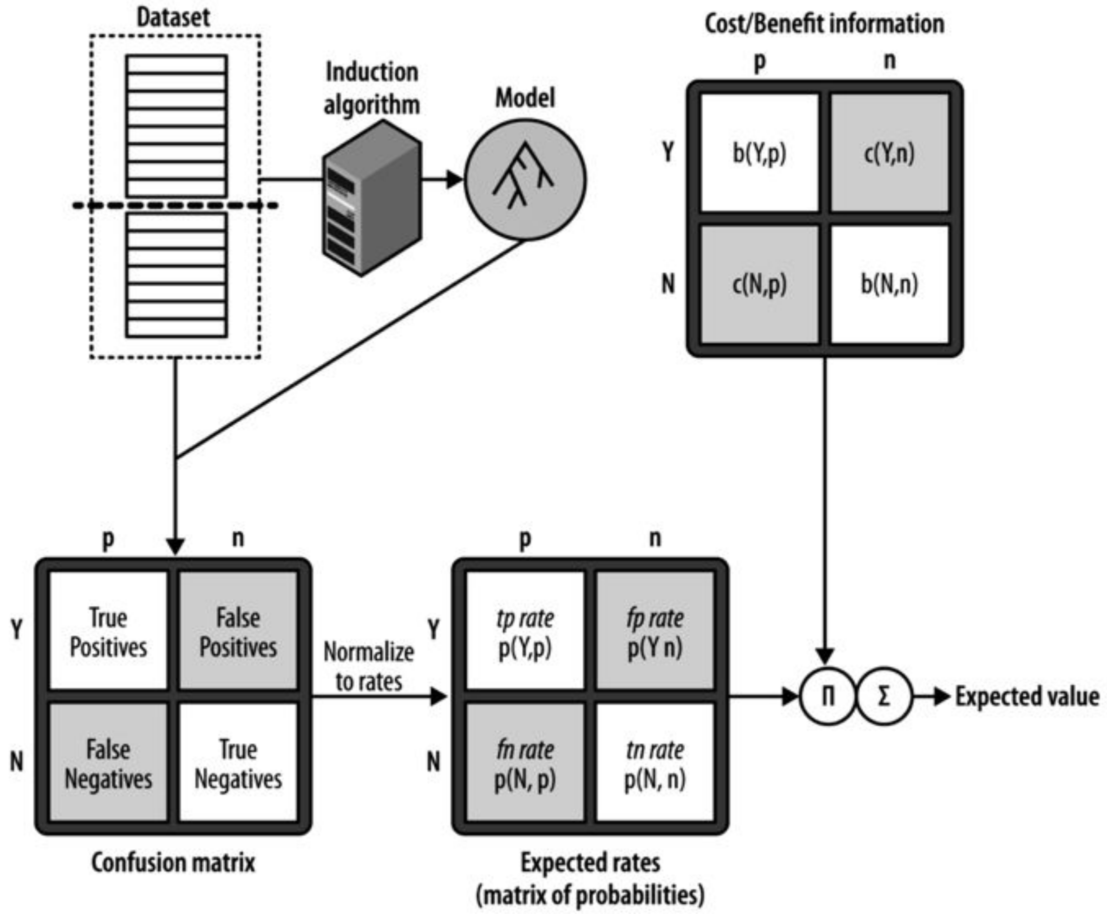
$$P_{response}(x)V_{response} + (1-P_{response}(x))V_{no\text{-}response}$$

$$P_{response}(x)\$99 + (1-P_{response}(x))(-\$1) > 0$$

$$P_{response}(x)\$99 > (1-P_{response}(x))\$1$$

$$P_{response}(x) > 0.01$$

Expected Value



Source: DSB

Confusion Probabilities

	p	n
Y	56	7
N	5	42

$T = 110$

$p(\mathbf{Y},\mathbf{p}) = 56/110 = 0.51 \quad p(\mathbf{Y},\mathbf{n}) = 7/110 = 0.06$

$p(\mathbf{N},\mathbf{p}) = 5/110 = 0.05 \quad p(\mathbf{N},\mathbf{n}) = 42/110 = 0.38$

Source: DSB

Cost-Benefit Matrix

actual

predicted

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

	p	n
Y	\$99	-\$1
N	0	0

Expected Profit

$$\textit{Expected profit} = p(\mathbf{Y}, \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N}, \mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + p(\mathbf{N}, \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y}, \mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n})$$

$$p(x, y) = p(y) \cdot p(x \mid y)$$

$$\text{Expected profit} = p(\mathbf{Y} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n})$$

$$\text{Expected profit} = p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot c(\mathbf{Y}, \mathbf{n})]$$

Source: DSB

Expected Profit

	p	n
Y	56	7
N	5	42

$T = 110$

$P = 61$ $\quad$ $N = 49$

$p(\mathbf{p}) = 0.55$ $\quad$ $p(\mathbf{n}) = 0.45$

$tp\ rate = 56/61 = 0.92$ $\quad$ $fp\ rate = 7/49 = 0.14$

$fn\ rate = 5/61 = 0.08$ $\quad$ $tn\ rate = 42/49 = 0.86$

$$
\begin{aligned}
\text{expected profit} &= p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\
&\quad p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{p}) \cdot c(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot b(\mathbf{Y}, \mathbf{p}) + 0.08 \cdot b(\mathbf{N}, \mathbf{p})] + \\
&\quad 0.45 \cdot [0.86 \cdot b(\mathbf{N}, \mathbf{n}) + 0.14 \cdot p(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot 99 + 0.08 \cdot 0] + \\
&\quad 0.45 \cdot [0.86 \cdot 0 + 0.14 \cdot \ - 1] \\
&= 50.1 - 0.063 \\
&\approx \mathbf{\$50.04}
\end{aligned}
$$

Source: DSB

Precision

Accuracy of positive predictions

TP / (TP+FP)

High precision: if you test positive you're probably positive

% of documents offered as relevant that are actually relevant

Recall = Sensitivity

True positive rate

Accuracy of positive class

TP / (TP + FN)

High recall is not missing many positives

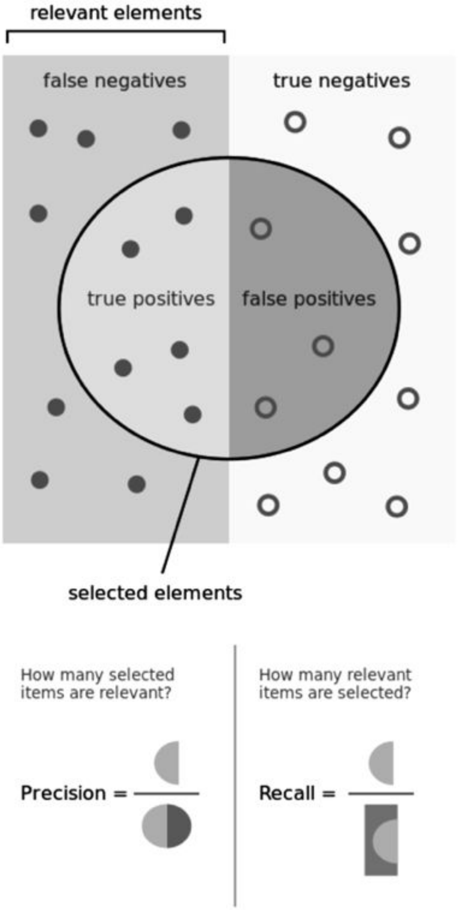
% of relevant documents that were found

What fraction of people with disease are identified

How sensitive is the test to indicators of disease

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Precision and Recall

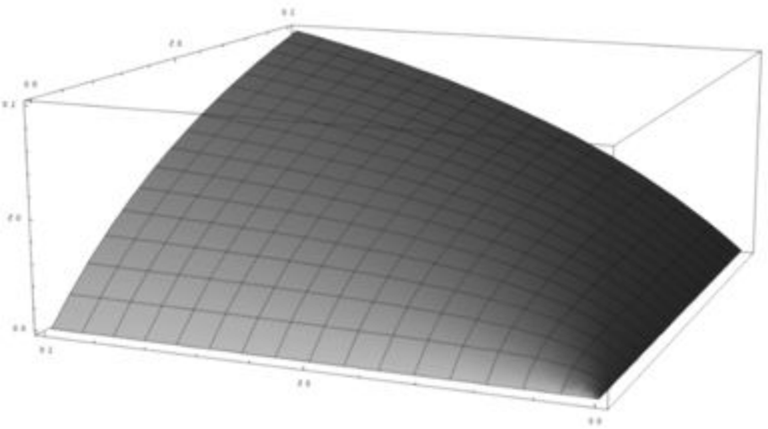


F1 Score

$$\frac{2 \cdot precision \cdot recall}{precision + recall}$$

Harmonic mean of precision and recall in [0,1]

Goal is high precision and high recall



F-Beta Score

$$\frac{(1+\beta^2)\cdot precision\cdot recall}{\beta^2\cdot precision+recall}$$

Weight precision and recall

Specificity

True negative rate

Accuracy on negative class

TN / (FP + TN)

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

What fraction of people without disease are identified

High specificity: few false alarms

Sensitivity and Specificity: Airport Security

Sensitivity: quantifies avoiding of false negatives

Specificity: quantifies avoiding false positives

For a test there is usually a trade-off between them.

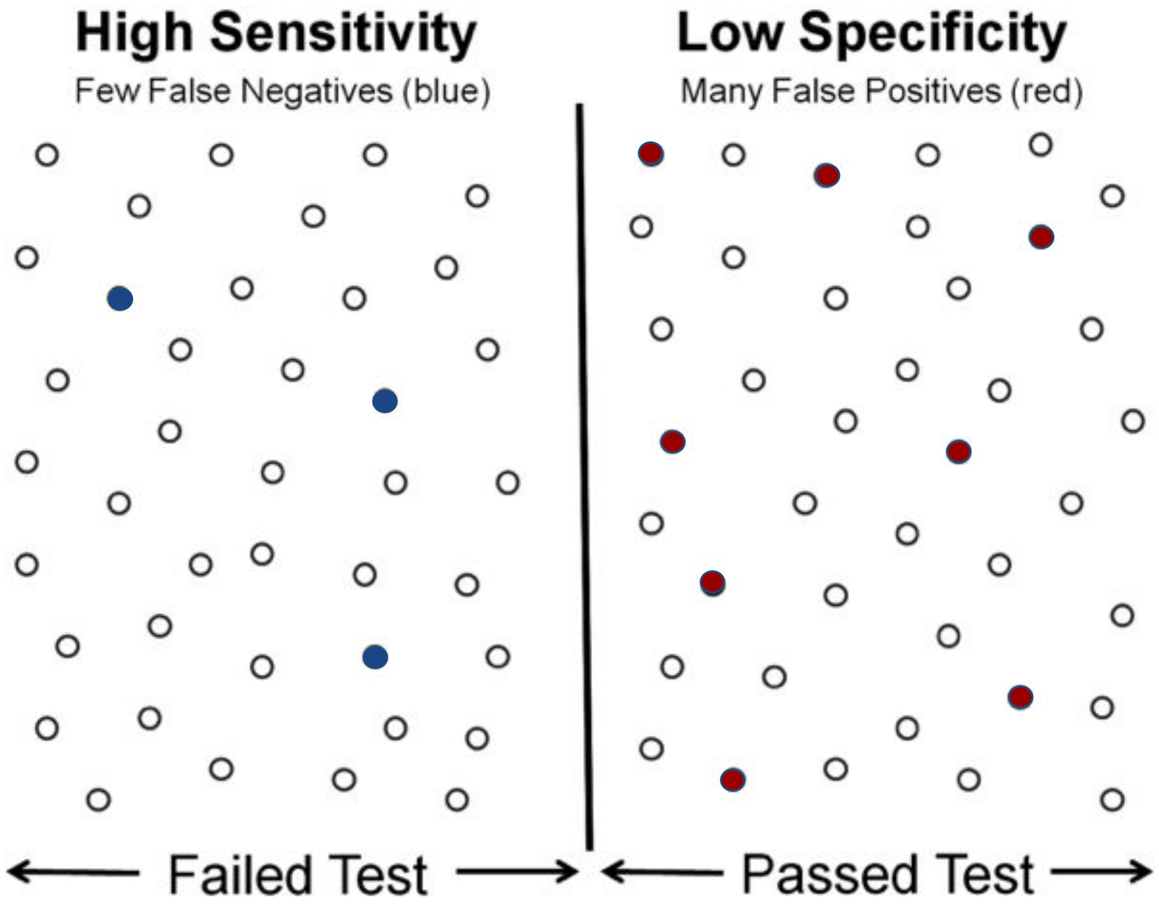
Low specificity and high sensitivity:
Testing passengers for potential safety threats
Scanners may be set to trigger alarms on low-risk items like buckles, keys (low specificity) to increase probability of identifying dangerous objects and minimize risk of missing objects that pose a threat (high sensitivity).

Sensitivity and Specificity: Medical Diagnosis

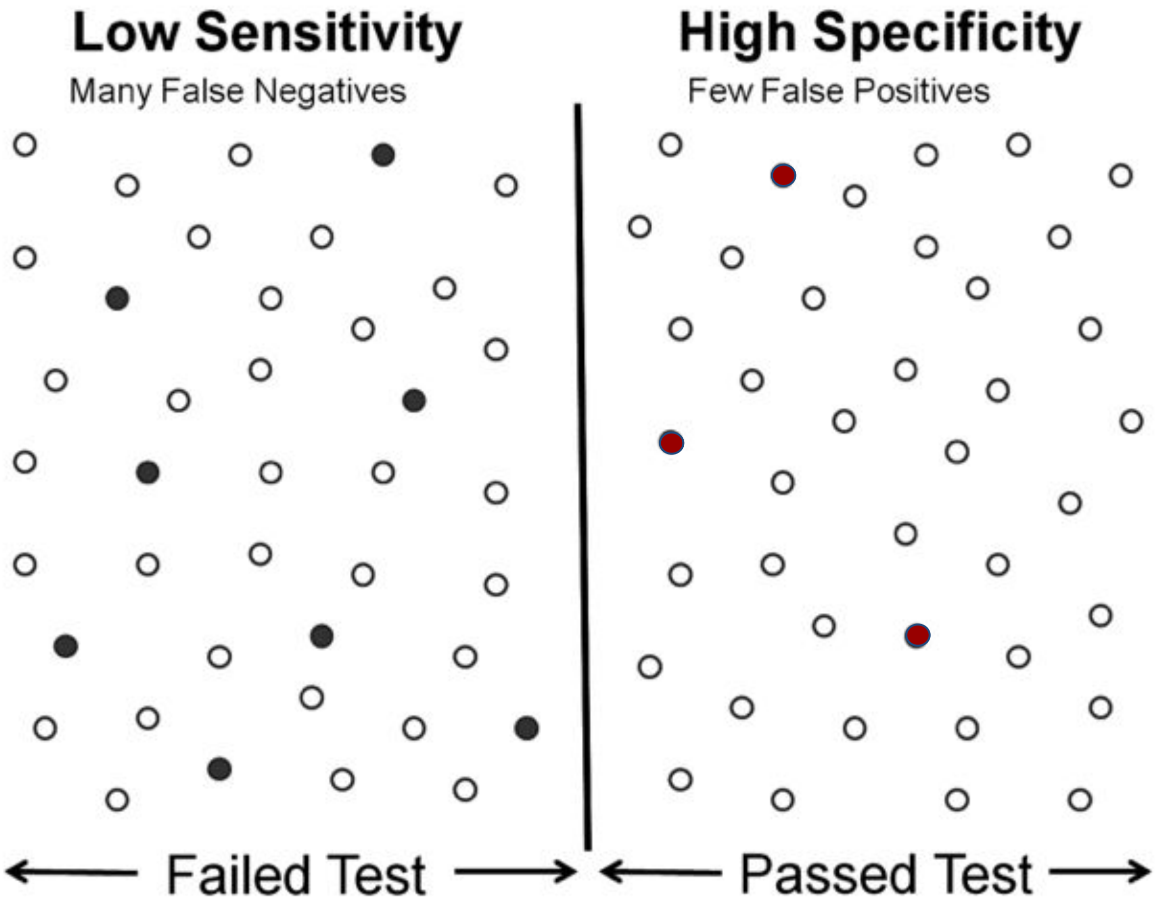
Goal to have high sensitivity and high specificity

Perfect predictor is both:
100% sensitive: all sick people are correctly identified as sick
100% specific: no healthy individuals are incorrectly identified as sick

Sensitivity and Specificity Trade-off



Sensitivity and Specificity Trade-off



False Positive Rate = Fall Out = False Alarm Rate

Error rate on negative class

FP / (FP + TN)

False Negative Rate = Miss Rate

FN / (FN + TP)

False Discovery Rate

FP / (FP + TP)

Positive Predictive Value

TP / (TP + FP)

PPV = 1 - FDR

	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Small positive predictive value of 10% indicates many of positive results from testing procedure are false positives.

Follow up any positive result with more reliable test to obtain more accurate assessment. Test may be useful if it is inexpensive and convenient.

Summary of Performance Measures

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive,** Power	**False positive,** Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR}-}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Comparison with Base Rate

Random

Majority classifier

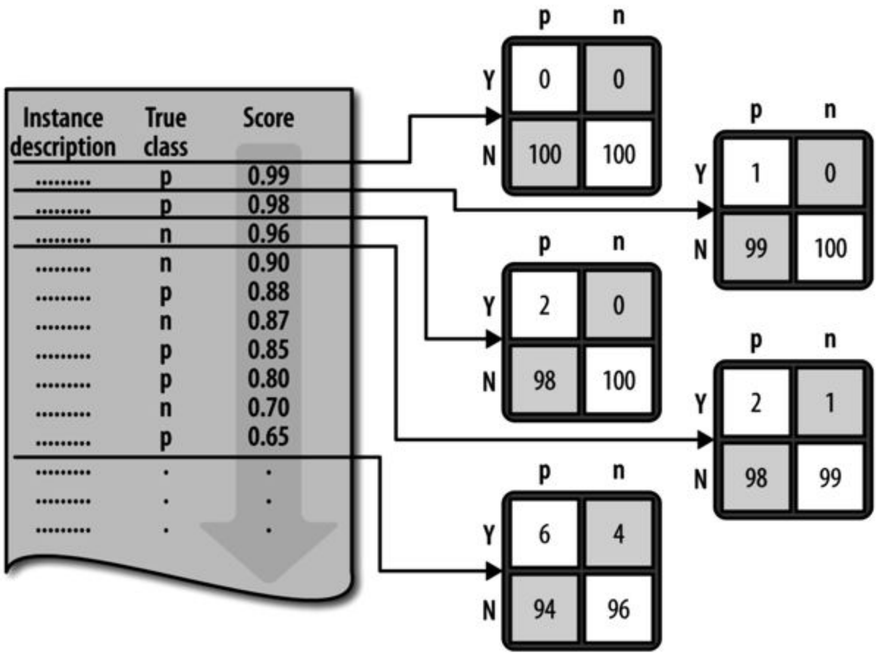
Weather forecasting: persistence, climatology.

Decision tree stump

Evaluating Performance

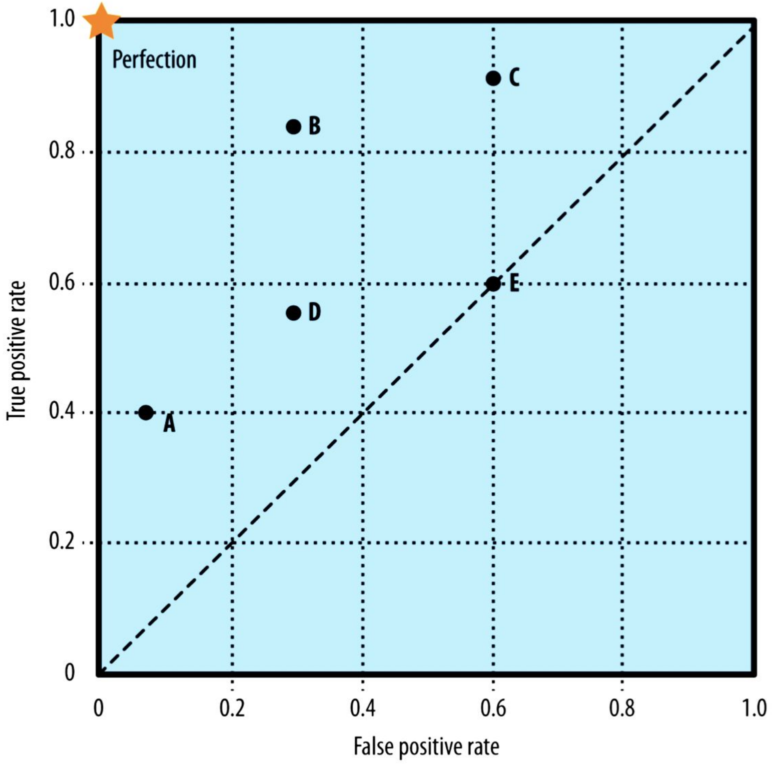
Probability Threshold

Different thresholds result in different confusion matrices
Decreasing threshold increases recall: TP / (TP + FN)
Increasing threshold may increase precision: TP / (TP + FP)



Source: DSB

Receiver Operating Characteristic (ROC) Curve

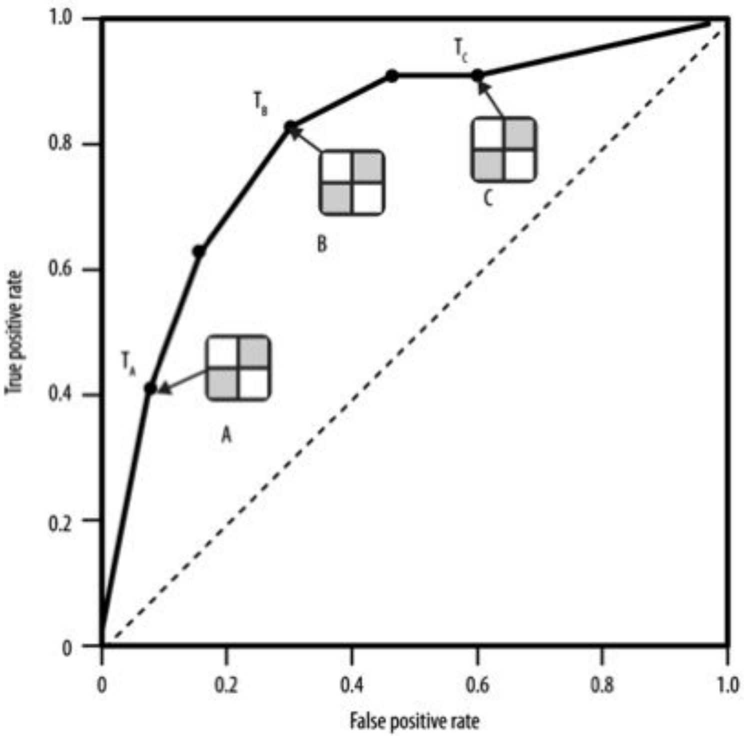


Source: DSB

Receiver Operating Characteristic (ROC) Curve

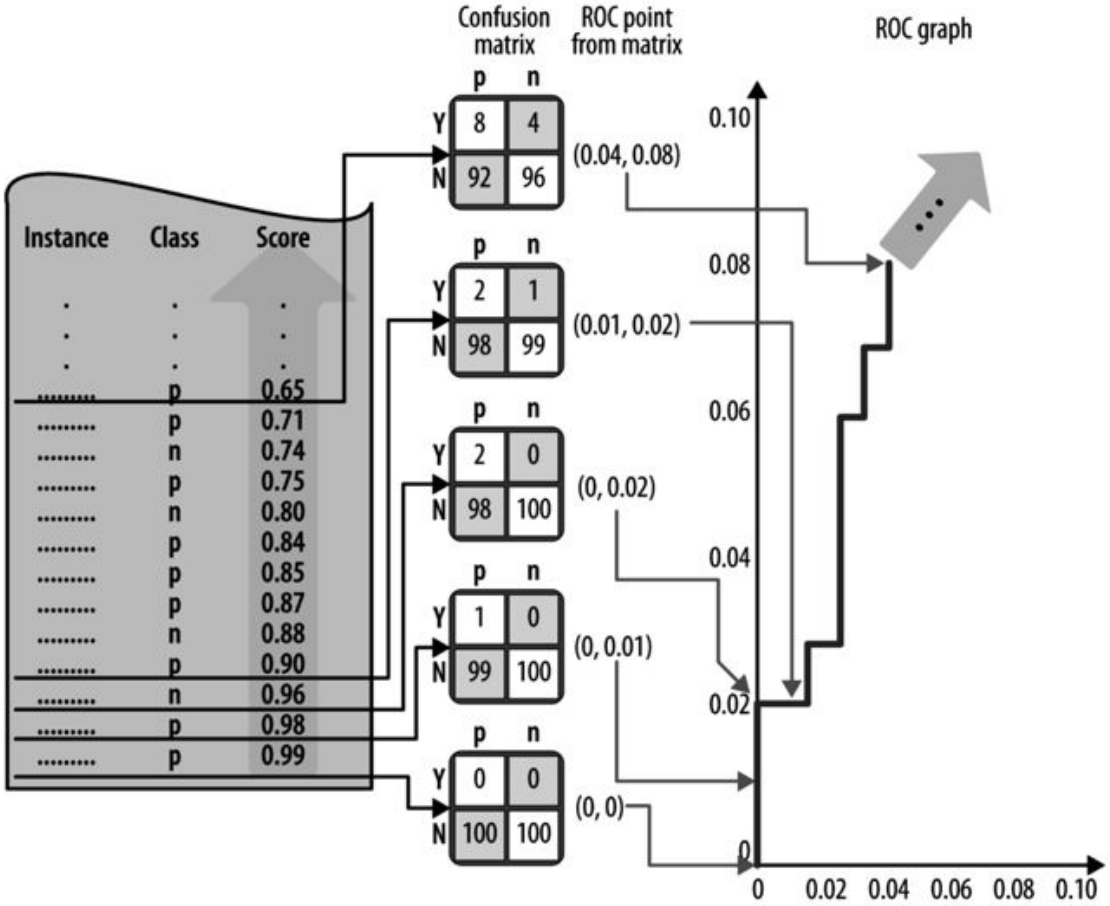
Best curve would go straight up and left
Non-increasing slope

(0,0)
(1,1)
(0,1)
diagonal
above diagonal
below diagonal



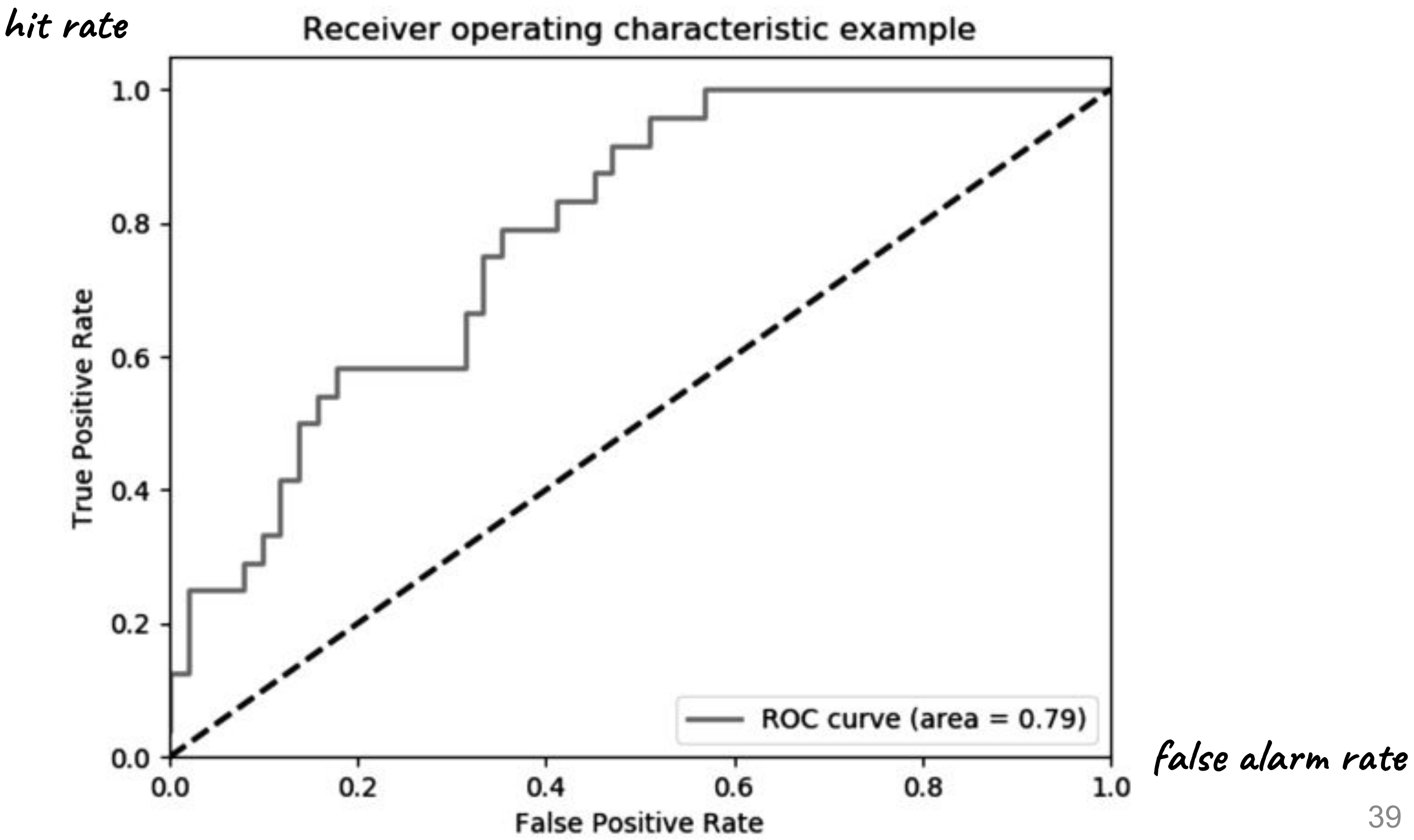
Source: DSB

Receiver Operating Characteristic (ROC) Curve



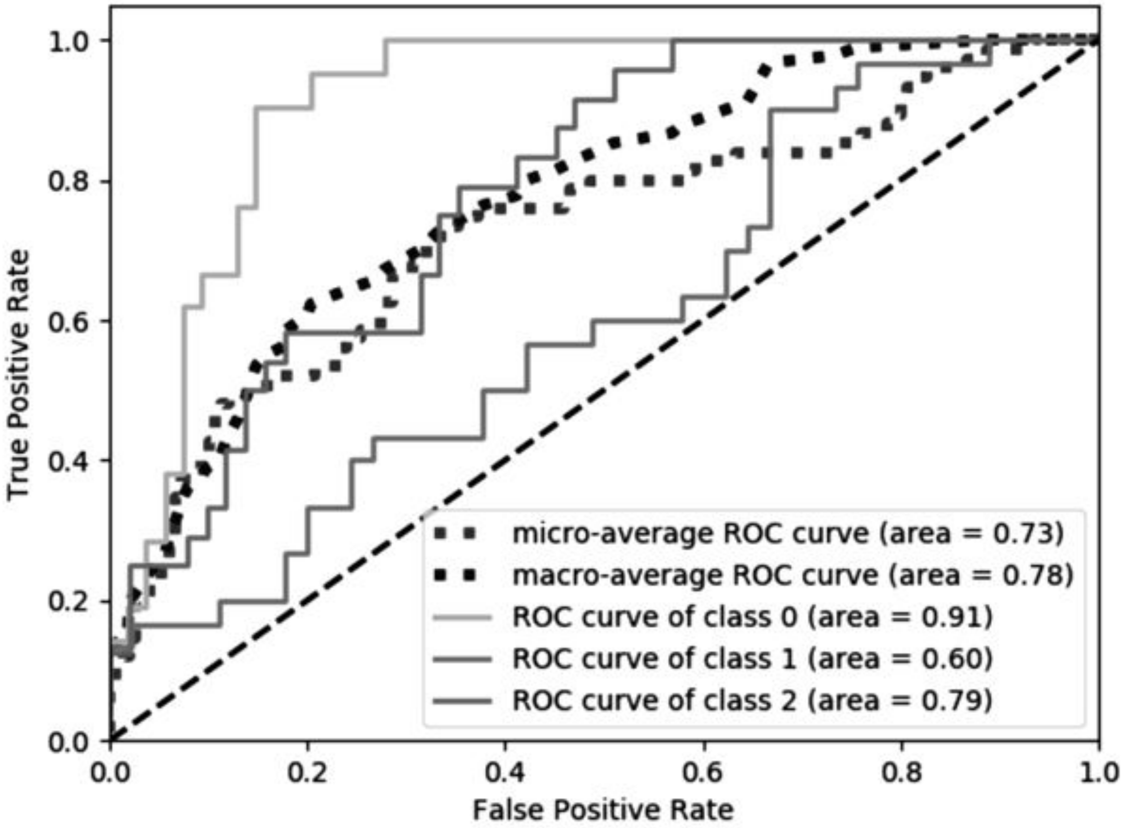
Source: DSB

Receiver Operating Characteristic (ROC) Curve



Source: DSB

Receiver Operating Characteristic (ROC) Curve



Source: DSB

Area Under ROC Curve (AUC)

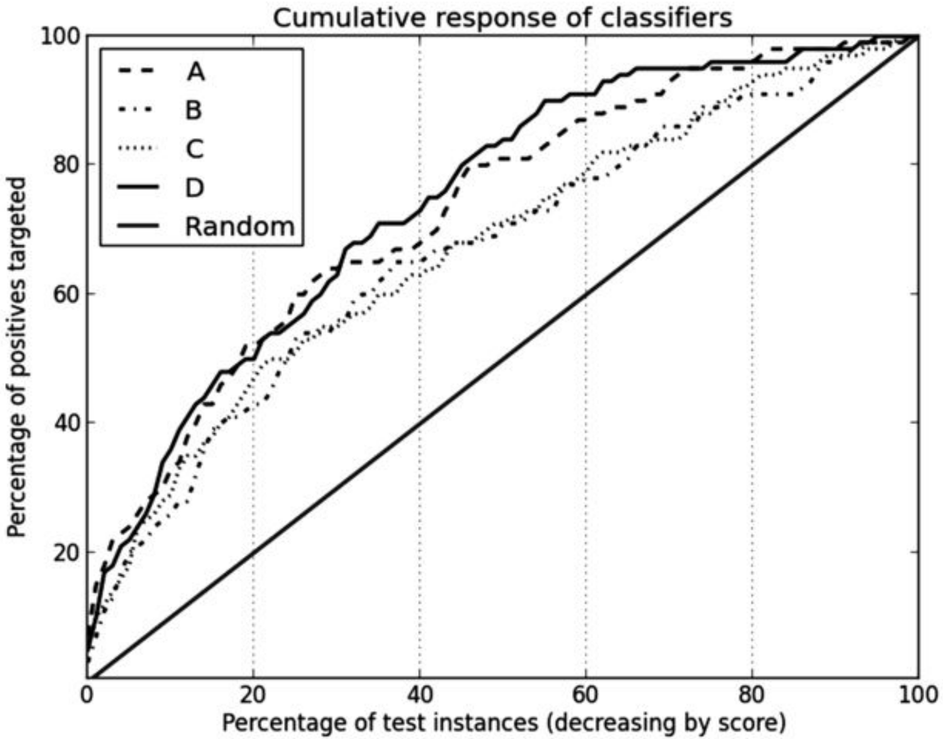
Single number used to summarize classifier performance
Perfect prediction is 1
Random prediction is 1/2, ROC along diagonal

Distribution of Kaggle competitions solutions by AUC performance metric

Source: Drori et al, 2018

Cumulative Response Curve

hit rate = tp rate
% of positives correctly classified

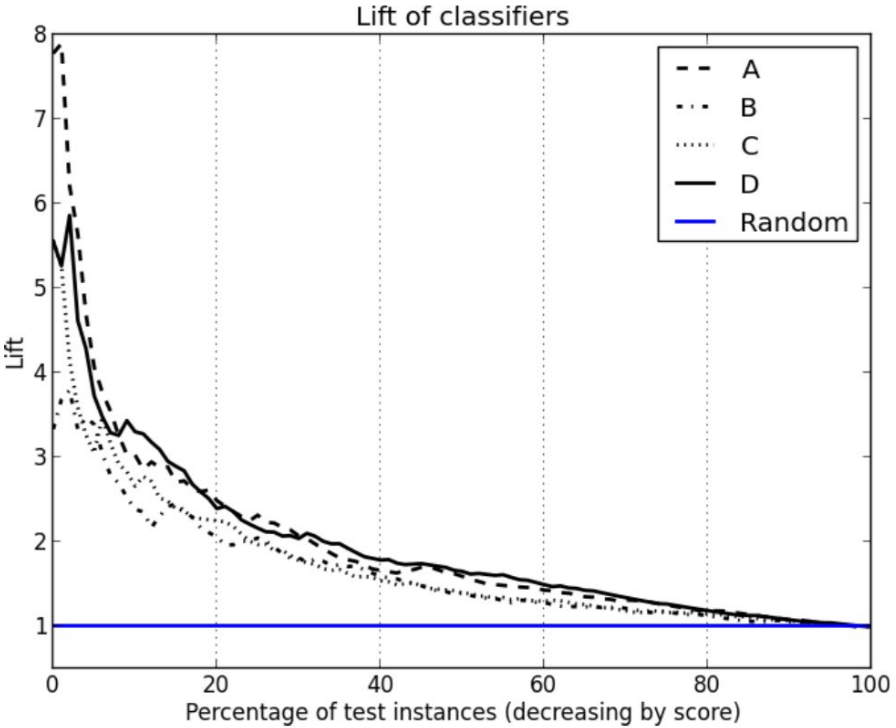


% of population targeted

Source: DSB

Lift

cumulative response / diagonal
how much better than random



% of population targeted

Source: DSB

Churn Example

- Customers switch carriers: churn
- Cheaper to retain customer than acquire new customer
- Retain customer by promotion or discount
- Giving discount to customer who is about to churn saves resources
- Giving discount to customer who is not going to churn wastes resources
- Train classifiers to predict churn.
- Select score threshold for giving discount.

Churn Example

Training accuracy

Model	Accuracy
Classification tree	95%
Logistic regression	93%
k-Nearest Neighbor	100%
Naive Bayes	76%

Test accuracy and AUC

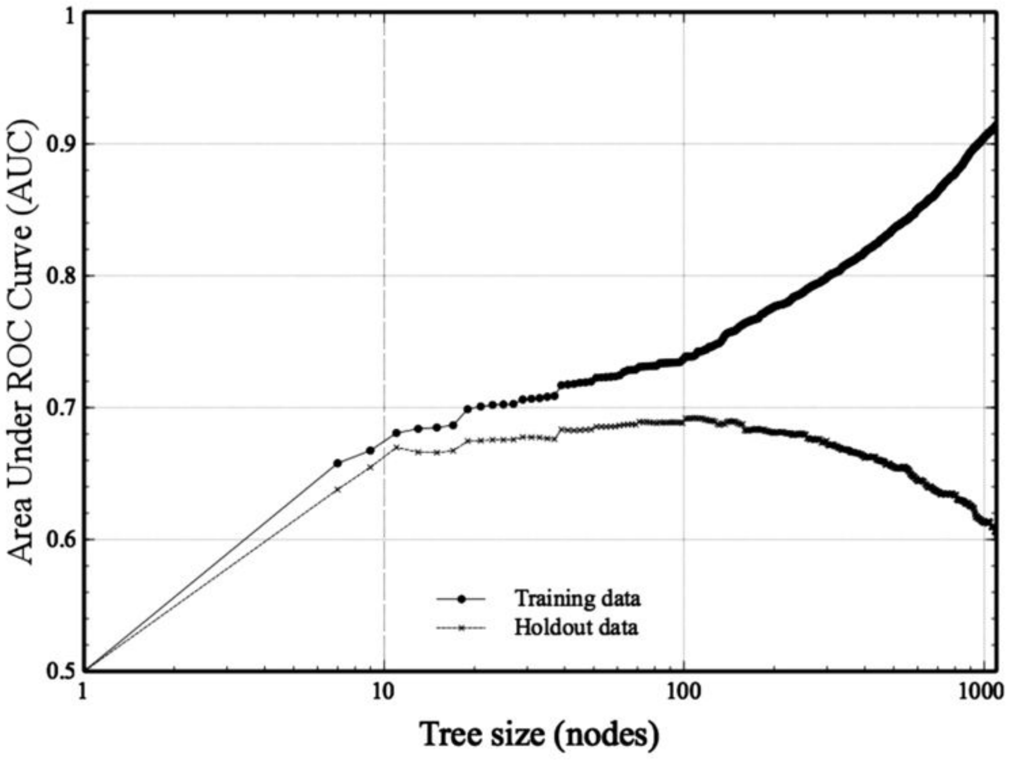
Model	Accuracy (%)	AUC
Classification Tree	91.8 ± 0.0	0.614 ± 0.014
Logistic Regression	93.0 ± 0.1	0.574 ± 0.023
k-Nearest Neighbor	93.0 ± 0.0	0.537 ± 0.015
Naive Bayes	76.5 ± 0.6	0.632 ± 0.019

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

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

Source: DSB

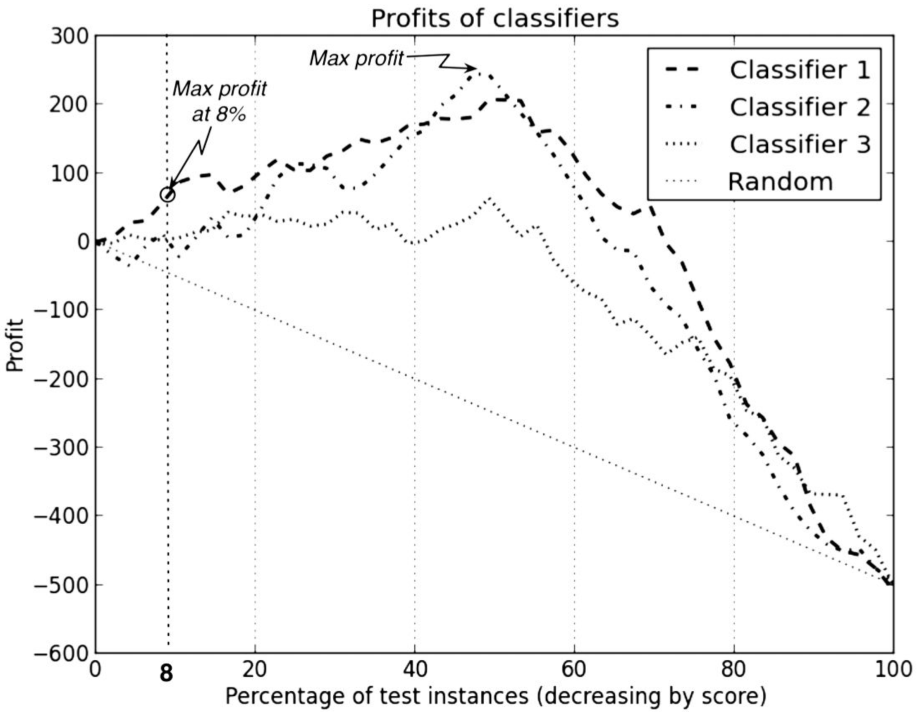
Fitting Curves



Source: DSB

Profit Curve

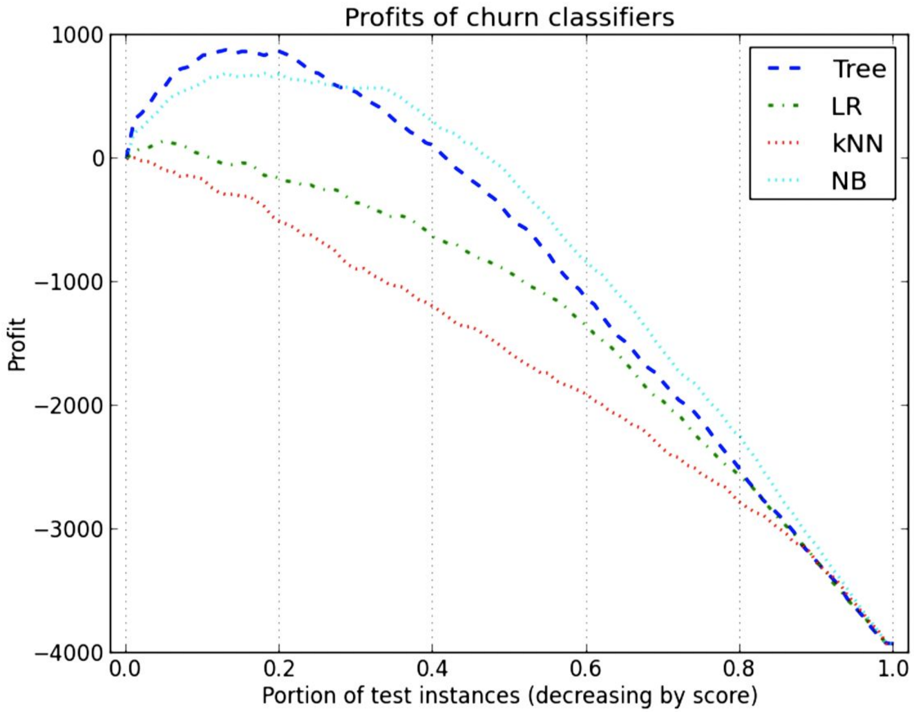
cumulative profit



% of population targeted

Source: DSB

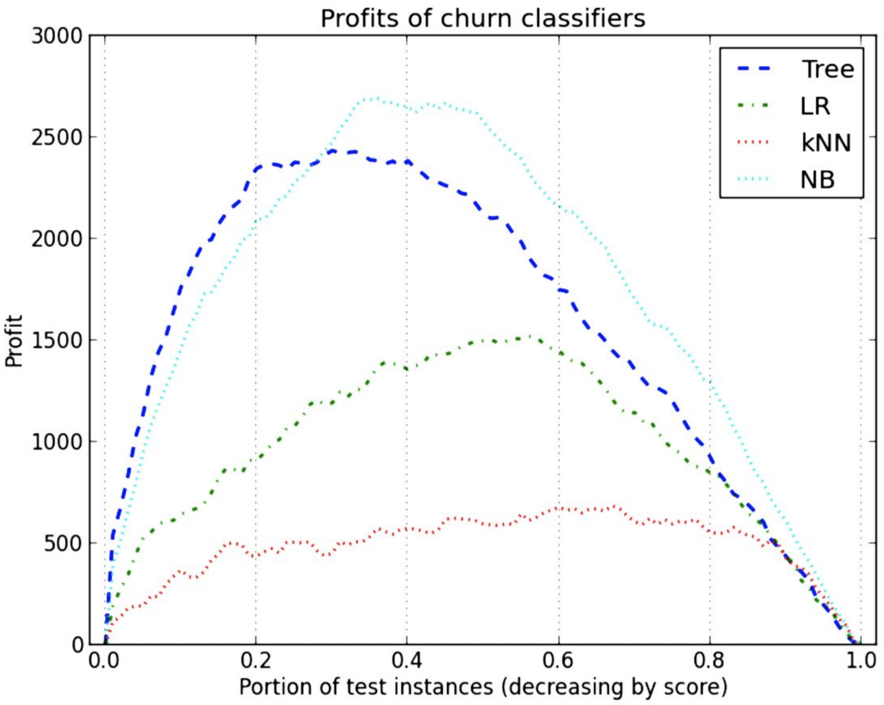
Profit Curve



9:1 benefit to cost ratio

Source: DSB

Profit Curve



12:1 benefit to cost ratio

Source: DSB

Collaborative Filtering

Example Problem

- Items $i=1..n$
- Users $j=1..p$
- Ratings Y_{ij}
- Binary mask if rating is available M_{ij}

- Problem: matrix completion

Naive Solution

- If k dimensional feature vectors $x^{(i)}$ are known for each item $i=1..n$
- Learn parameter vectors $\theta^{(j)}$ for each user $j=1..p$
- Predict user j rating item i by $\theta^{(j)^T} x^{(i)}$

$$\underset{\theta^{(j)}}{\text{minimize}} \frac{1}{2} \sum_{i:M_{ij}=1} \left(\theta^{(j)^T} x^{(i)} - Y_{ij}\right)^2 + \frac{\lambda}{2} \sum_k \theta^{(j)^2}$$

Naive Solution

- If k dimensional feature vectors $x^{(i)}$ are known for each item $i=1..n$
- Learn parameter vectors $\theta^{(j)}$ for **all users** $j=1..p$ using gradient descent
- Predict user j rating item i by $\theta^{(j)^T} x^{(i)}$

$$\underset{\theta^{(1)}\ldots\theta^{(p)}}{\text{minimize}} \frac{1}{2} \sum_{i,j:M_{ij}=1} \left(\theta^{(j)^T} x^{(i)} - Y_{ij}\right)^2 + \frac{\lambda}{2} \sum_j \sum_k \theta^{(j)^2}$$

Problem

- If k dimensional feature vectors $x^{(i)}$ are unknown
- Given parameter vectors $\theta^{(j)}$ for all users $j=1..p$ learn feature vectors $x^{(i)}$

$$\underset{x^{(1)}\ldots x^{(n)}}{\text{minimize}} \frac{1}{2} \sum_{i,j:M_{ij}=1} \left({\theta^{(j)}}^T x^{(i)} - Y_{ij} \right)^2 + \frac{\lambda}{2} \sum_{i} \sum_{k} {x_k^{(i)}}^2$$

Iterative Solution

- Given $\{x^{(1)}, \dots, x^{(n)}\}$ learn $\{\theta^{(1)}, \dots, \theta^{(p)}\}$

- Given $\{\theta^{(1)}, \dots, \theta^{(p)}\}$ learn $\{x^{(1)}, \dots, x^{(n)}\}$

Collaborative Filtering Solution

- Learn $\{x^{(1)}, \dots, x^{(n)}\}$ and $\{\theta^{(1)}, \dots, \theta^{(p)}\}$ together

$$\underset{x^{(1)}\dots x^{(n)},\theta^{(1)}\dots\theta^{(p)}}{\text{minimize}} \frac{1}{2} \sum_{i,j:M_{ij}=1} \left(\theta^{(j)^T} x^{(i)} - Y_{ij}\right)^2 + \frac{\lambda}{2} \sum_i \sum_k x_k{}^{(i)^2} + \frac{\lambda}{2} \sum_j \sum_k \theta_k{}^{(j)^2}$$

- Predictions $\theta^{(j)^T} x^{(i)}$

Collaborative Filtering Solution

- Low rank factorization of rating into product of feature matrix and parameter matrix

$$\begin{bmatrix} {\theta^{(1)}}^T x^{(1)} & \cdots & {\theta^{(p)}}^T x^{(1)} \\ \vdots & \ddots & \vdots \\ {\theta^{(1)}}^T x^{(n)} & \cdots & {\theta^{(p)}}^T x^{(n)} \end{bmatrix}$$