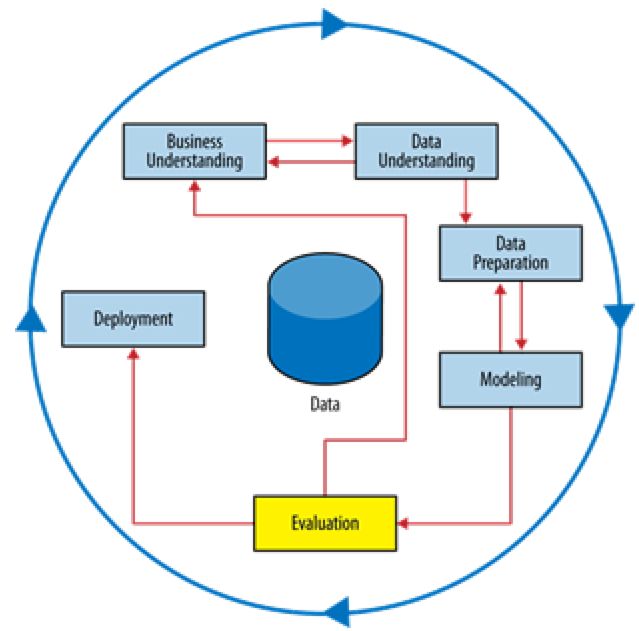
Model Evaluation: Metrics

PF7

Learning Goals

- What is your **desirable business goal** from a data project?
  - Baseline models
    - majority model, random model, conditional model

- How to evaluate model performance?
  - Confusion matrix
  - Expected value
  - Evaluation metrics
    - Sensitivity/Recall
    - Specificity
    - Precision
    - F1 score



Class Labels in Business: Binary Classification

- **Positives** are unusual outcomes worthy of attention (*Alarm*)
  - Medical test with positives = *Disease*
  - Loan defaulting case with positives = *Default*
  - Marketing campaign with positives = *Respond (or Leave)*

- **Negatives** are common outcomes that are unworthy of attention.

- Real-word Data is usually **unbalanced**.
  - More harmless negatives than positives.

Prediction Errors

- **False positive** errors (*false alarms*)
  - A healthy person (Actual: -) predicted as sick (Pred: +)
  - A client clearing debts timely (Actual: -) predicted as default (Pred: +)
  - An inactive customer (Actual: -) predicted as active responder (Pred: +)

- **False negative** errors (*missed alarms*)
  - A patient with cancer (Actual: +) predicted as healthy (Pred: -)
  - A bad client (Actual: +) predicted as good borrower (Pred: -)
  - An active responder (Actual: +) predicted as inactive (Pred: -)

Baseline Performance

- **Majority model**: always predict the majority class in training set.
  - For unbalanced datasets, the majority model yields a very high accuracy.
  - However, maximizing accuracy might not be an appropriate goal.

- **Random model**: same probability for both classes.
  - If predict x% of instances as positive (Y), then x% of positive (p) and negative (n) will be predicted Y.

- **Conditional model**: a model based on the most informative feature.
  - Implement a simple model based on domain knowledge:
    - A sudden jump in account usage can be a useful predictor for credit card fraud.
  - Reconsider the cost and benefit of extra data sources.

Evaluating Classifiers: Confusion Matrix

- **Confusion matrix** for a binary classification is a 2×2 matrix.
  - Rows: actual classes – Positive (p) or Negative (n)
  - Columns: predicted classes – Positive (**Y**) or Negative (**N**)

		Predicted Class	
		Y	**N**
Actual Class	p	True Positives (TP)	False Negatives (FN)
	n	False Positives (FP)	True Negatives (TN)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Example: Confusion Matrix

Actual Class (1 = p; 0 = n)	Predicted Class (1 = Y; 0 = N)
0	0
1	1
0	1
0	1
0	0
1	1
0	0
0	0
1	1
1	0

False Negative

Predicted ⇒ / Actual ⇓	Y	N
p	3	1
n	2	4

False Positive

Problems with Accuracy

- **Plain accuracy** $$Accuracy = \frac{Number\ of\ correct\ predictions\ made}{Total\ number\ of\ predictions\ made}$$

- **Error rate** $$Error\ rate = 1 - Accuracy$$

- **Problems of plain accuracy**
  - Data structure is ignored: real-word class distribution are often unbalanced.
    - In a marketing data with 99% non-responders (n), a majority model which always predict the majority class (in train set) yields 99% accuracy (base rate).
  - Prediction errors are treated equally.
    - Different class/error often has different importance in real world.

Errors are Not Equally Important

- Medical diagnosis:
  - False positive: a healthy person was informed he has disease (*another test*).
  - False negative: a patient with cancer was informed healthy(*miss early detection*)
- Marketing campaign:
  - False positive: target a customer but he won't respond (*waste money*).
  - False negative: fail to target a potential customer (*lose a customer*).
- Loan application:
  - False positive: reject a good customer (*lose a customer*).
  - False negative: approve a bad customer who will not pay back (*lose money*).
- What is your business goal? Are we assessing the data mining results according to that goal?

An Analytical Framework: Expected Value

- **Expected value**: the weighted sum of business values for all possible outcomes.

$$EV = p(o_1) \times v(o_1) + p(o_2) \times v(o_2) \ldots$$

  - o_i: possible decision outcomes
    - For example: TP, FP, TN, FN
  - $p(o_i)$: the probability of occurrence for each outcome.
    - Usually estimated from the data (e.g., confusion matrix).
  - $v(o_i)$: the business value of each outcome.
    - Usually estimated from domain knowledge: can be either profit or cost.

Expected Value in Model Use

- A classifier was trained to predict customers' response in a marketing campaign, which threshold shall be used in classification?
  - Only customers predicted as *respond* (Y) will be targeted.
  - Cost of targeting = $\$1$ per person, product price $= \$200$, product cost $= \$100$.

- Business value of a positive prediction (i.e., targeting a customer) is:
  - True positives (*R*): $\$200 - \$100 - \$1 = \$99$
  - False positives (*NR*): $-\$1$

- **Expected value** of targeting a consumer:

$$EV = p(+) \times v_R + p(-) \times v_{NR} > 0$$

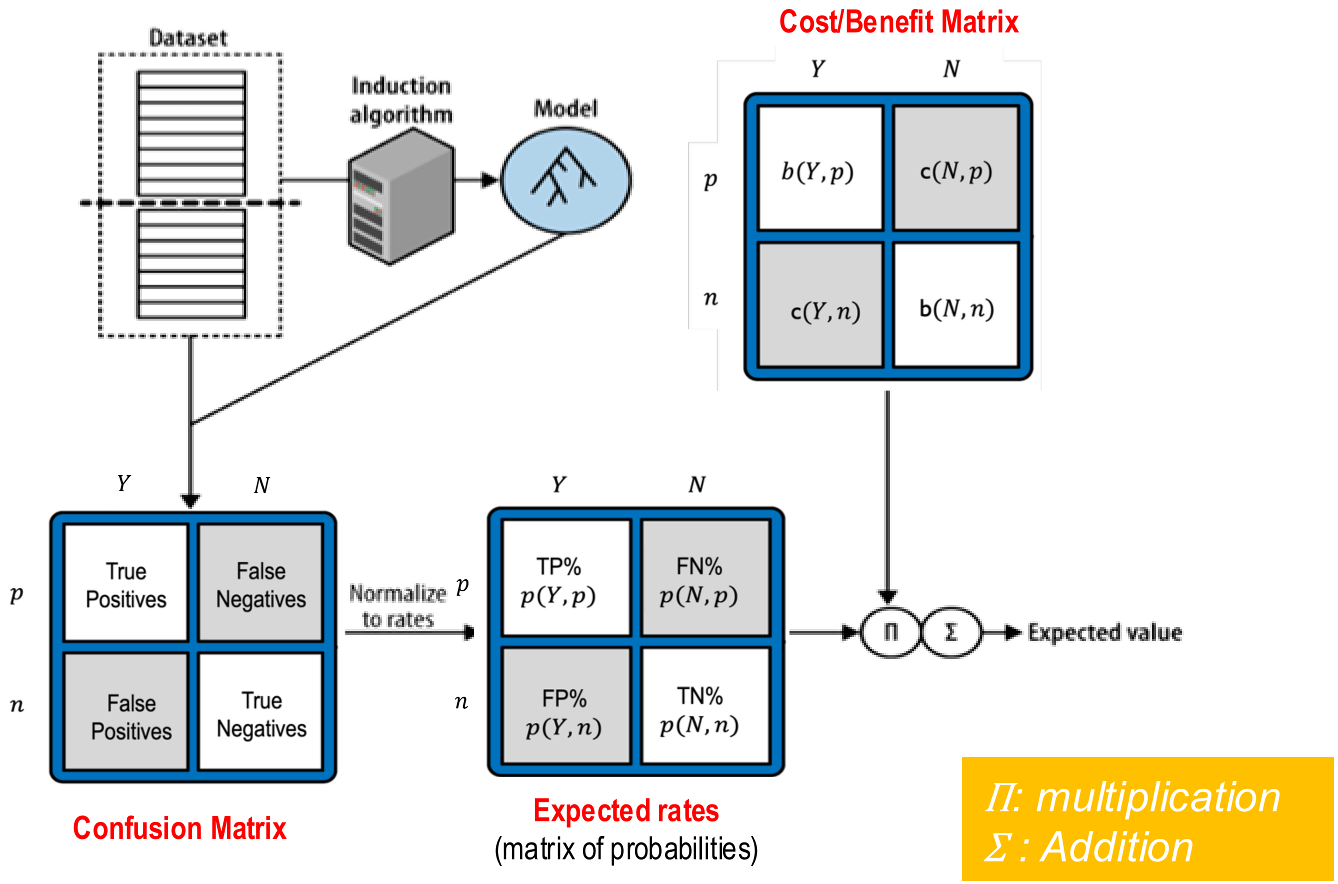
$$p(+) \times \$99 + [1 - p(+)] \times (-\$1) > 0$$

$$p(+) > 0.01$$

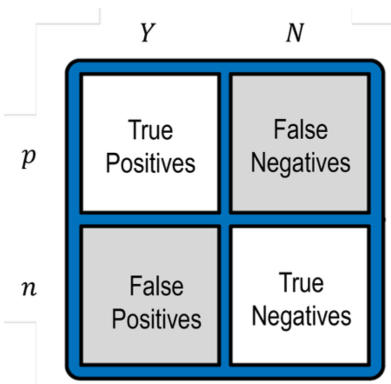
- **Business action**: target a customer if the estimated $p(+) > 1\%$.

Expected Value in Model Evaluation

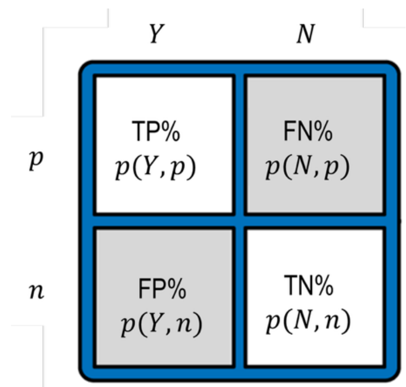
- How to compare multiple classifiers, each yielding a confusion matrix?



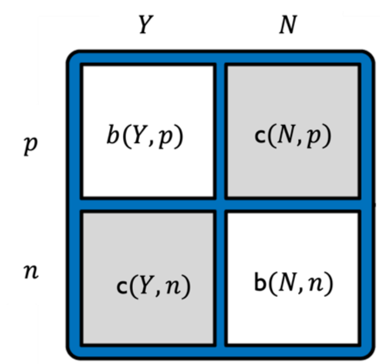
Example: Expected Value



Normalize to rates



Cost/Benefit Matrix



	Y	N
p	51	5
n	6	38

	Y	N
p	0.51	0.05
n	0.06	0.38

	Y	N
p	99	0
n	-1	0

$$\textbf{EV} = p(Y,p) * b(Y,p) + p(Y,n) * c(Y,n) + p(N,p) * c(N,p) + p(N,n) * b(N,n)$$

$$= 0.51 \times 99 + 0.06 \times (-1) + 0.05 \times 0 + 0.38 \times 0$$

$$= 50.43$$

Expected value per person

Evaluation Metrics

	Y	N
p	True Positives (TP)	False Negatives (FN)
n	False Positives (FP)	True Negatives (TN)

- $\text{TPR} = \frac{TP}{TP+FN} = \frac{TP}{p}$ **(Sensitivity/Recall)**

  "If a person has disease (p), what is the likelihood he will be tested positive (Y)?"

- $\text{FNR} = \frac{FN}{TP+FN} = \frac{FN}{p}$ **(Miss rate)**

  FNR = 1 - TPR

- $\text{TNR} = \frac{TN}{FP+TN} = \frac{TN}{n}$ **(Specificity)**

  "If a person is healthy (n), what is the likelihood he will be tested negative (N)?"

- $\text{FPR} = \frac{FP}{FP+TN} = \frac{FP}{n}$ **(False alarm rate)**

  FPR = 1 – TNR

- **Precision** $= \frac{TP}{TP+FP} = \frac{TP}{Y}$

  "If a person is tested positive (Y), what is the likelihood that he actually has disease (p)?"

Precision for N class $\frac{\text{TN}}{\text{TN+FN}} = \frac{\text{TN}}{\text{N}}$

typos in textbook p.204

Evaluation Metrics

Predicted ⇒ Actual ⇓	Y	N
p	a	c
n	b	d

Sensitivity $TPR = \frac{a}{a+c}$

Specificity $TNR = \frac{d}{b+d}$

Precision (Y) $\frac{a}{a+b}$

Precision (N) $\frac{d}{c+d}$

Tradeoff between Precision and Recall

- **Sensitivity/Recall**-oriented machine learning tasks:
  - Medical Diagnosis (e.g., cancer detection)
    - Models with low recall (high miss rate) fail to raise alarm for ill patients.
  - Fraud detection
    - Models with low recall (high miss rate) fail to identify fraud transactions.

Don't miss anything important ! (minimize FN)

- **Precision**-oriented machine learning tasks:
  - Customer-facing tasks: recommendation systems, spam email detection
    - Models with low precision make lots of irrelevant recommendations.

Ensure predicted positives are correct ! (minimize FP)

- **F1 score** is a balance between ***Recall*** and ***Precision***.

$$\textbf{F1 score} = 2 * \frac{precision * recall}{precision + recall}$$

Evaluation Metrics: Exercise

- An unbalanced data with actual labels as below:
  - Respond (p) = 100 Not Respond (n) = 900
- Compare models by choosing a proper evaluation metric.
  - Which metric if we aim to target as many responders (p) as possible?
    - Only target a customer if s/he is predicted as *Respond* (i.e., Y).
  - Which metric if we aim to have most targeted customers (Y) actually respond to our offer (p) ?

Model A

	Y	N
p	80	20
n	120	780

Random model

	Y	N
p	20	80
n	180	720

Majority model

	Y	N
p	0	100
n	0	900

A random classifier predicts 20% instances as Y: 20% p and 20% n will be predicted as Y

Model Evaluation Metrics: Exercise

	Model A	Random Model	Majority Model
Accuracy $\frac{TP+TN}{TP+FP+TN+FN}$			
Sensitivity $\text{TPR} = \frac{TP}{TP+FN\ (p)}$			
Precision $\frac{TP}{TP+FP\ (\text{Y})}$ or $\frac{TN}{\text{TN}+FN\ (\text{N})}$			
Specificity $\text{TNR} = \frac{TN}{FP+TN\ (n)}$			

	Y	N
p	80	20
n	120	780

Model A

	Y	N
p	20	80
n	180	720

Random Model

	Y	N
p	0	100
n	0	900

Majority Model

Model Evaluation: Exercise

- Calculate the **expected value** for each model:
  - Which model will generate the most profit?

Cost/Benefit Matrix

	Y	N
p	99	0
n	-1	0

	Y	N
p	80	20
n	120	780

Model A

	Y	N
p	20	80
n	180	720

Random Model

	Y	N
p	0	100
n	0	900

Majority Model