COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Machine Learning I

- Predictive performance. Classification

comillas.edu

What is classification?

- **Classification** is the task of *learning* ***a target*** *function* **f** that maps attribute set **x** to one of the predefined class labels **y**

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

One of the attributes is the class attribute
In this case: Cheat

Two class labels (or classes): Yes (1), No (0)

Why classification?

- The target function f is known as a classification model
- Descriptive modeling: Explanatory tool to distinguish between objects of different classes (e.g., understand why people cheat on their taxes)
- Predictive modeling: Predict a class of a previously unseen record

Classification vs Prediction

- Classification:
  - predicts categorical class labels
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction:
  - models continuous-valued functions, i.e., market prices, traffic, temperature, …
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

Examples of classification tasks

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying spam email, spam web pages, adult content
- Understanding if a web query has commercial intent or not

Binary or multi-class classification

- The attribute to classify can have two classes (binary classification) or more (multi-class classification)

- The methods used for binary classification are easily extended to multi-class classification problems

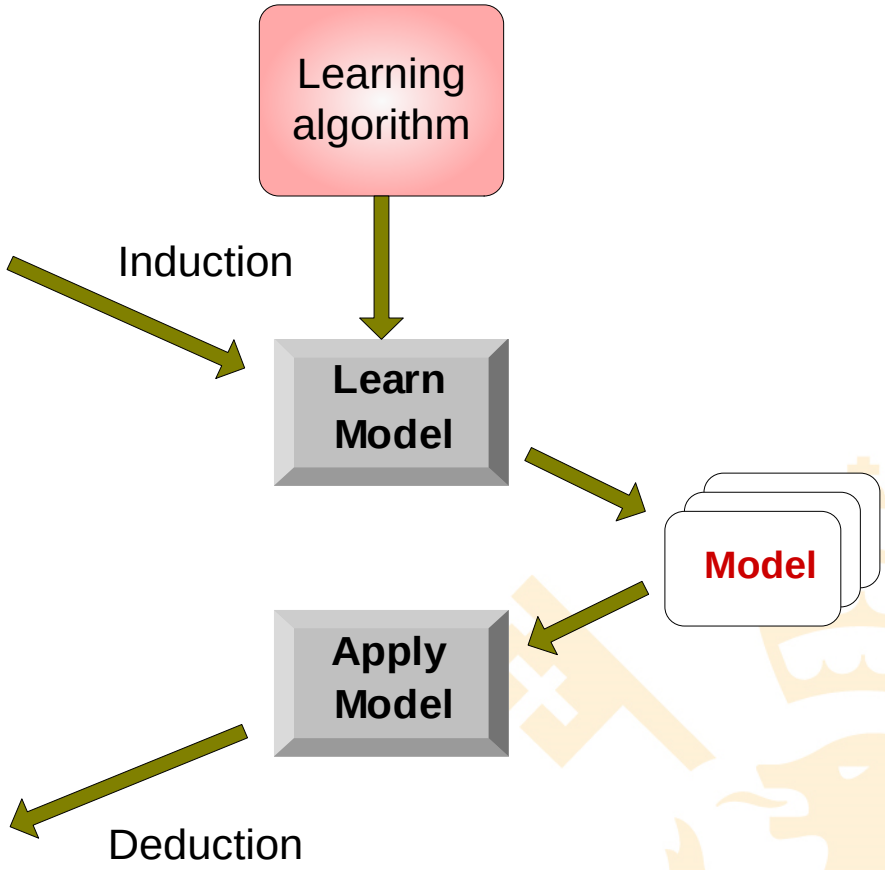
General approach to classification

- Training set consists of records with known class labels
- Training set is used to build a classification model
- A labeled test set *with known class labels* includes data records not used in the training set. This is used to evaluate the quality of the model.
- If the test is passed successfully, the classification model is applied to new records with unknown class labels

Training and Test data sets

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Training and Test data sets

- *Test set*: independent instances that have played no part in formation of classifier
- Assumption: *both training data and test data* are representative samples of the underlying problem
- Test and training data may differ in nature
- Example: classifiers built using customer data from two different towns A and B
- To estimate performance of classifier from town A in completely new town, test it on data from B

Evaluation of binary classification models. Metrics

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- **Confusion Matrix:** Counts of test records that are correctly (or incorrectly) predicted by the classification model

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	f_{11}	f_{10}
	Class=No	f_{01}	f_{00}

f_{11}: TP (true positive)

f_{10}: FN (false negative)

f_{01}: FP (false positive)

f_{00} : TN (true negative)

Evaluation of binary classification models

- f_{00} and f_{11} are called respectively the number of cases *True Positives* (predicted 1 and real 1) and *True Negatives* (predicted 0 and real 0)
- f_{01} and f_{10} are called respectively the number of cases *False Positives* (predicted 1 and real 0) and *False Negatives* (predicted 0 and real 1)
- *Sensibility or Recall*

  TP/(TP+FN)=TPR (ratio of TP)
- *Specificity*

  TN/(TN+FP)=TNR(ratio of TN)

Actual Class \ Predicted Class	Class = 1	Class = 0
Class = 1	f_{11}=TP	f_{10}=FN
Class = 0	f_{01}=FP	f_{00}=TN

This is for binary classification but the confusion matrix can be extended to multi-class problems

Evaluation of binary classification models. Metrics

- *Confusion matrix*

Actual Class \ Predicted Class	Class = 1	Class = 0
Class = 1	f_{11}	f_{10}
Class = 0	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\#\,\text{correct predictions}}{\text{total}\,\#\,\text{of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Error rate} = \frac{\#\,\text{wrong predictions}}{\text{total}\,\#\,\text{of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

ROC Curve(Receiver Operating Characteristic)

- Developed in 1950s for signal detection theory to analyze noisy signals
  - Characterize the trade-off between positive hits and false alarms
- **ROC** curve plots **TPR** (on the **y**-axis) against **FPR** (on the **x**-axis)

Sensibility

$$TPR = \frac{TP}{TP + FN}$$

Fraction of positive instances predicted correctly

1- Specificity

$$FPR = \frac{FP}{FP + TN}$$

Fraction of negative instances predicted incorrectly

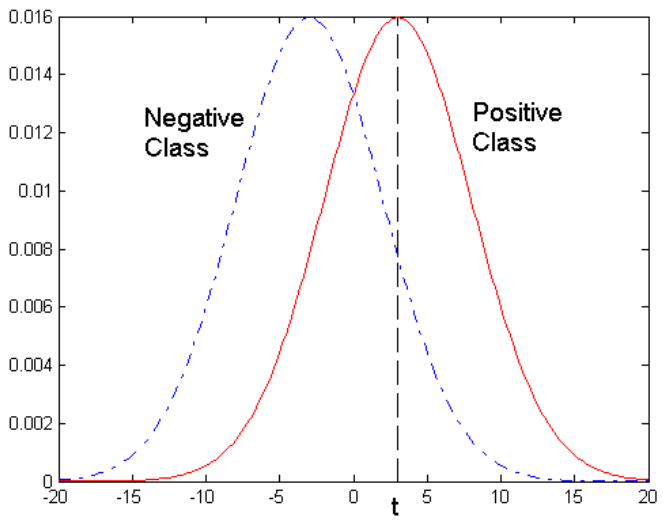
		PREDICTED CLASS	
		Yes	No
Actual	Yes	a (TP)	b (FN)
	No	c (FP)	d (TN)

ROC Curve (Receiver Operating Characteristic)

- The performance of a classifier is represented as a point on the **ROC** curve according to a **t** threshold

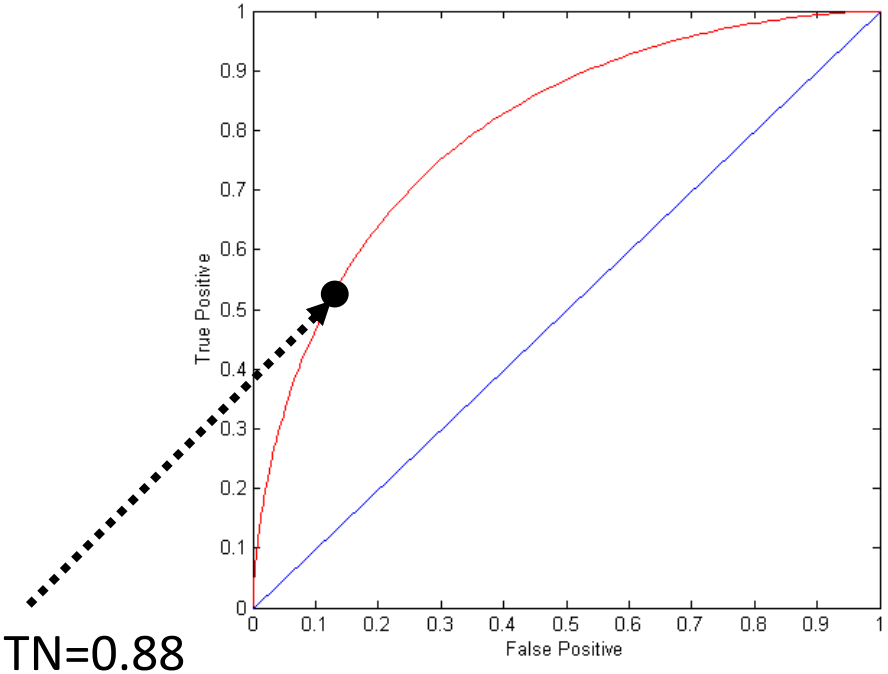
- Considering a data set containing 2 classes (***positive*** and ***negative***)

- any point located at x > t is classified as ***positive***



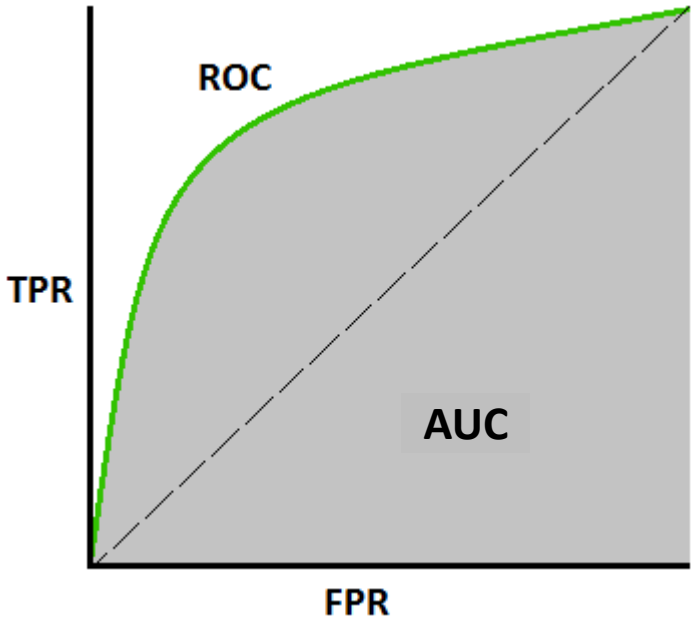
At threshold **t**:

TP=0.5, FN=0.5, FP=0.12, TN=0.88



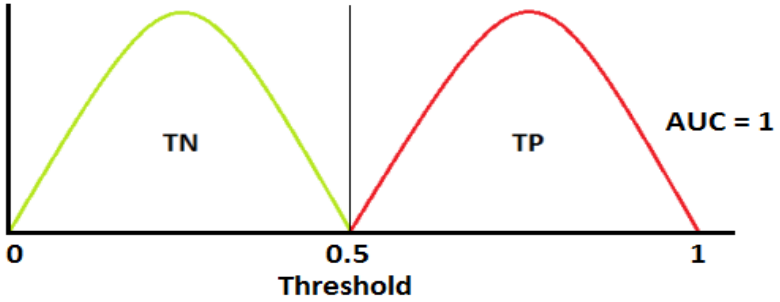
ROC Curve (Receiver Operating Characteristic)

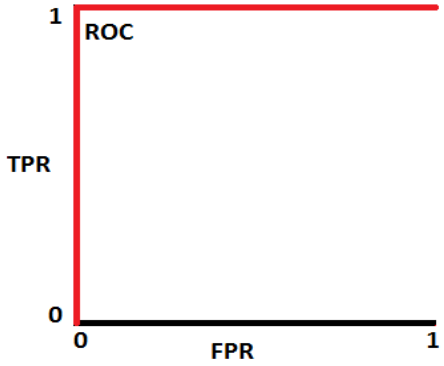
- AUROC (Area Under the Receiver Operating Characteristics) Developed in 1950s for signal detection theory to analyze noisy signals
- AUROC or AUC is a performance measurement for classification problem at various thresholds settings
- ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.



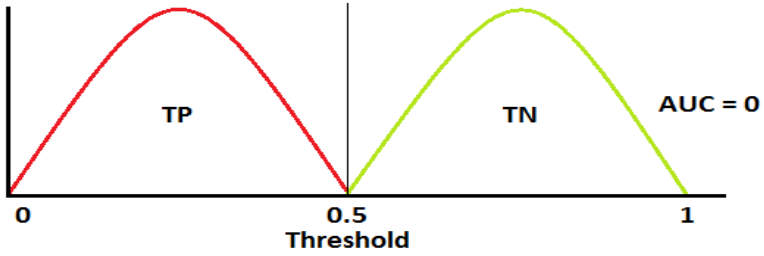
ROC Curve(Receiver Operating Characteristic)

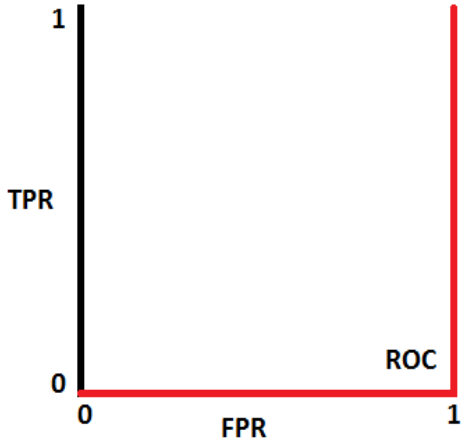
- Ideal scenario: not overlapping





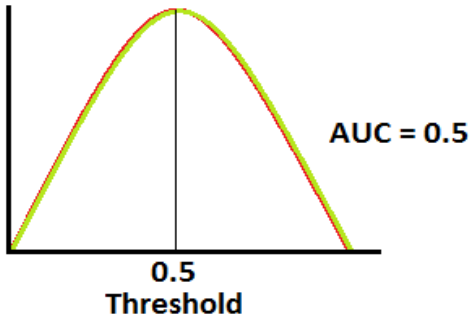
- Worst scenario

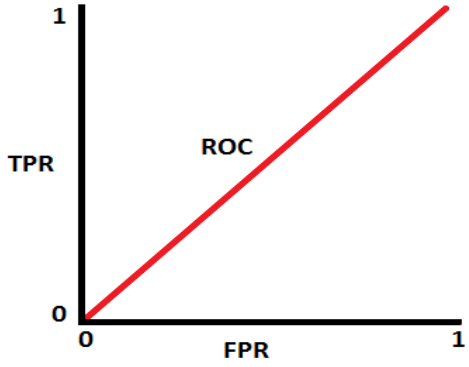




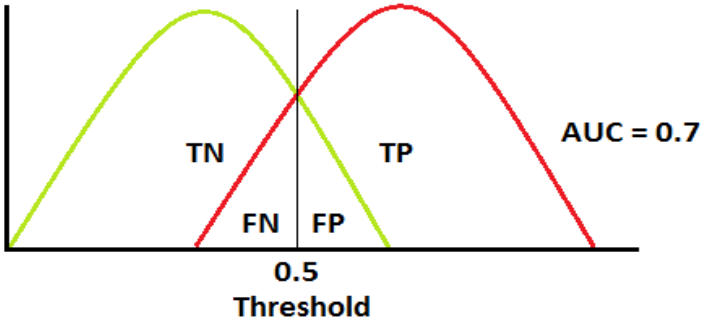
ROC Curve(Receiver Operating Characteristic)

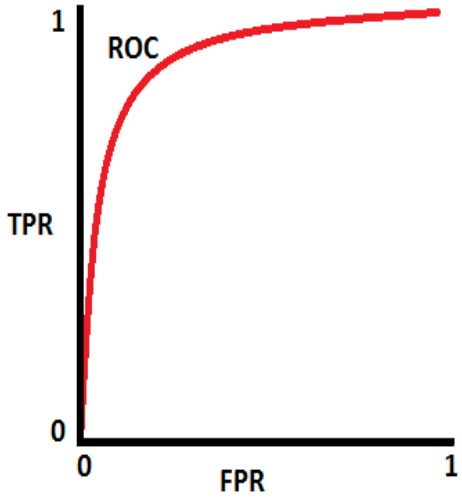
- Limit scenario





- Usual scenario

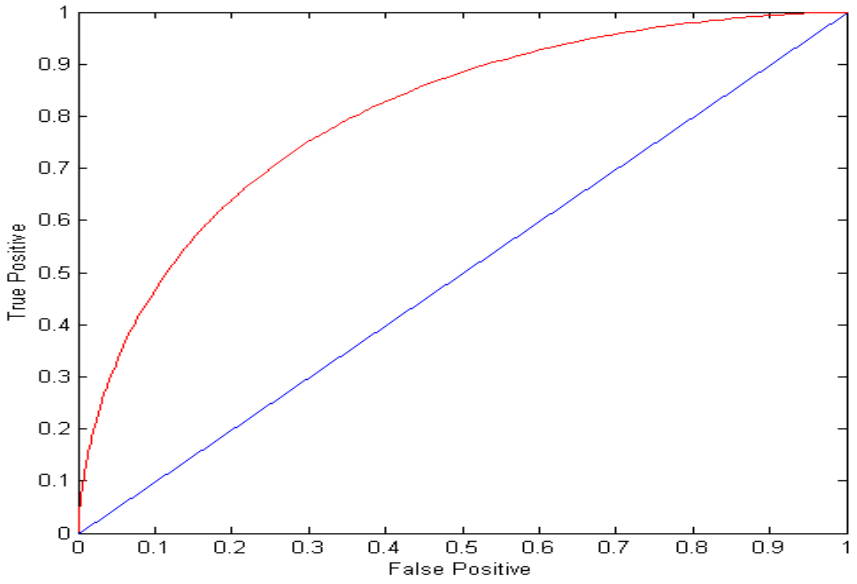




ROC Curve(Receiver Operating Characteristic)

(TP,FP):

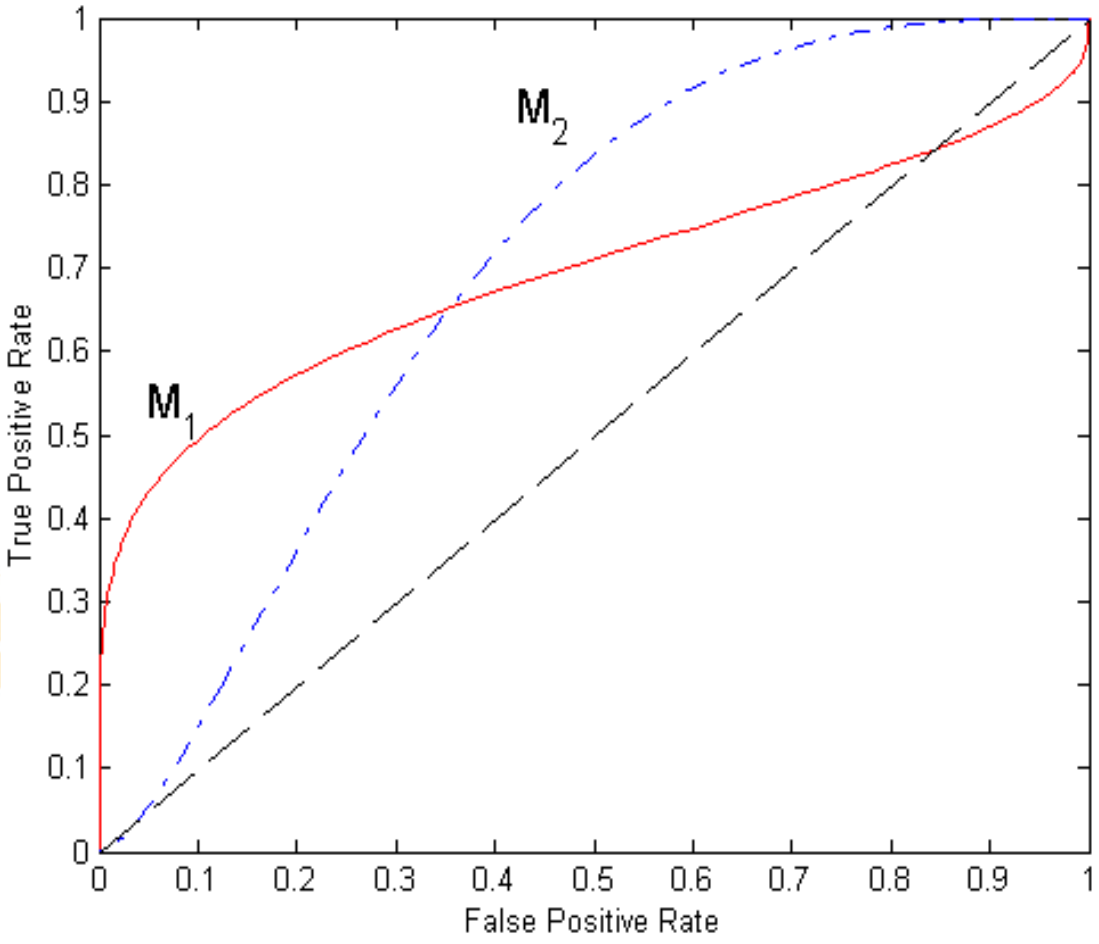
- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal



- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - prediction is opposite of the true class

		PREDICTED CLASS	
		Yes	No
Actual	Yes	a (TP)	b (FN)
	No	c (FP)	d (TN)

Using ROC for Model Comparison



- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve (AUC)
  - Ideal: Area = 1
  - Random guess:
    - Area = 0.5

Cumulative gain and Lift Charts

- Cumulative gains and lift charts are visual aids for measuring model performance. Both charts consist of a lift curve and a baseline

- Lift is a measure of the effectiveness of a predictive model calculated as *the ratio between the results obtained with and without the predictive model*.

- The greater the area between the lift curve and the baseline, the better the model

Generating Cumulative Gain and Lift Charts

http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html

- A company wants to do a mail marketing campaign. It costs the company $1 for each item mailed. They have information on 100,000 customers. Create a cumulative gains and a lift chart from the following data.
- Overall Response Rate - No model exists

If we assume we have no model other than the prediction of the overall response rate, then we can predict the number of positive responses as a fraction of the total customers contacted. Suppose the response rate is 20% from previous experience of the company. If all 100,000 customers are contacted we will receive around 20,000 positive responses.

Cost ($)	Total Customers Contacted	Positive Reponse
100,000	100,000	20,000

Example for Generating Cumulative Gain and Lift Charts

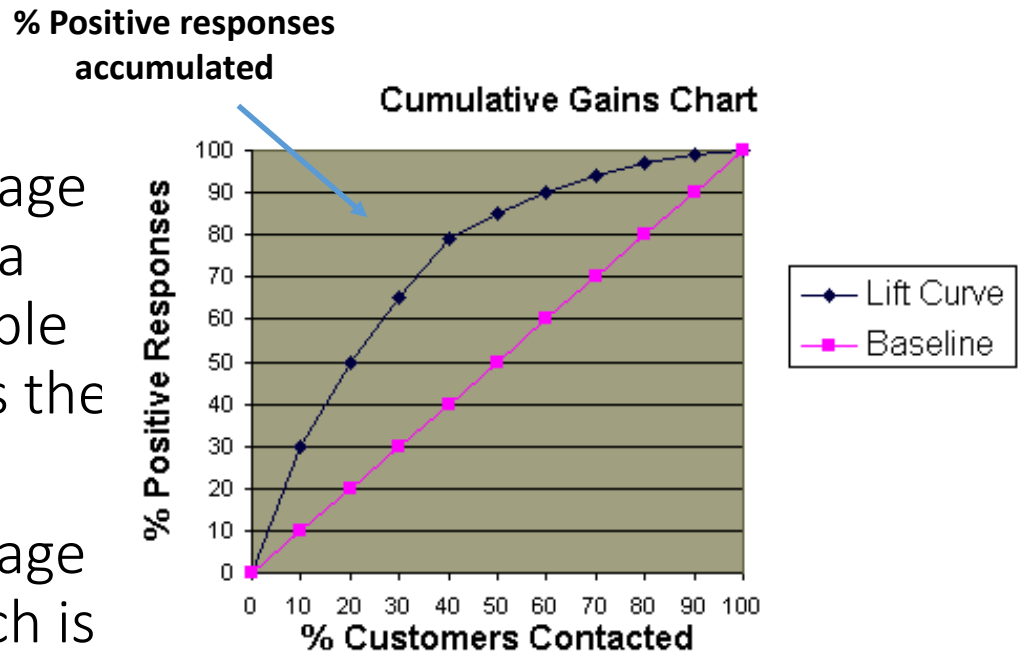
- Prediction of Response Model – Model exists

A response model will predict who will respond to a marketing campaign. If we have a response model, we can make more detailed predictions. For example, we use the response model to assign a score to all 100,000 customers and predict the results of contacting only the top 10,000 customers, the top 20,000 customers, etc.

Cost ($)	Total Customers Contacted	Positive Responses	% Positive responses accumulated
10000	10000	6000	30
20000	20000	10000	50
30000	30000	13000	65
40000	40000	15800	79
50000	50000	17000	85
60000	60000	18000	90
70000	70000	18800	94
80000	80000	19400	97
90000	90000	19800	99
100000	100000	20000	

Over 20000 positive responses

Cumulative Gains Chart

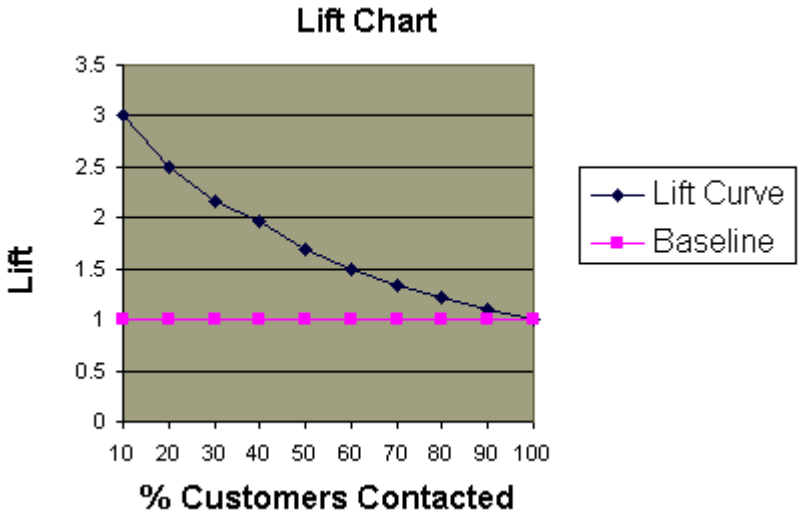


- The *y*-axis shows the percentage of positive responses. This is a percentage of the total possible positive responses (20,000 as the overall response rate shows).
- The *x*-axis shows the percentage of customers contacted, which is a fraction of the 100,000 total customers.
- **Baseline (overall response rate):** If we contact *X*% of customers then we will receive *X*% of the total positive responses.

Lift Curve: Using the predictions of the response model, calculate the percentage of positive responses for the percent of customers contacted and map these points to create the lift curve

Lift Chart

- To plot the chart calculate the points on the lift curve by determining the ratio between the result predicted by our model (gain chart) and the result using no model.
- Example: For contacting 10% of customers, using no model we should get 10% of responders and using the given model we should get 30% of responders. The y-value of the lift curve at 10% is 30 / 10 = 3.



Another example gain and lift chart

- Using the response model P(x)=100-AGE(x) for customer x and the data table presented, construct the cumulative gains and lift charts for responses observed.
- Ties in ranking should be arbitrarily broken by assigning a higher rank to who appears first in the table.

Customer Name	*Height*	*Age*	*Actual Response*
Alan	70	39	N
Bob	72	21	Y
Jessica	65	25	Y
Elizabeth	62	30	Y
Hilary	67	19	Y
Fred	69	48	N
Alex	65	12	Y
Margot	63	51	N
Sean	71	65	Y
Chris	73	42	N
Philip	75	20	Y
Catherine	70	23	N
Amy	69	13	N
Erin	68	35	Y
Trent	72	55	N
Preston	68	25	N
John	64	76	N
Nancy	64	24	Y
Kim	72	31	N
Laura	62	29	Y

Another example gain and lift chart

- 1. Calculate P(x) for each person x
- 2. Order the people according to rank P(x)

Customer Name	*P(x)*	*Actual Response*
Alex	88	Y
Amy	87	N
Hilary	81	Y
Philip	80	Y
Bob	79	Y
Catherine	77	N
Nancy	76	Y
Jessica	75	Y
Preston	75	N
Laura	71	Y
Elizabeth	70	Y
Kim	69	N
Erin	65	Y
Alan	61	N
Chris	58	N
Fred	52	N
Margot	49	N
Trent	45	N
Sean	35	Y
John	24	N

Another example gain and lift chart

- 3. Calculate the percentage of total responses for each cutoff point

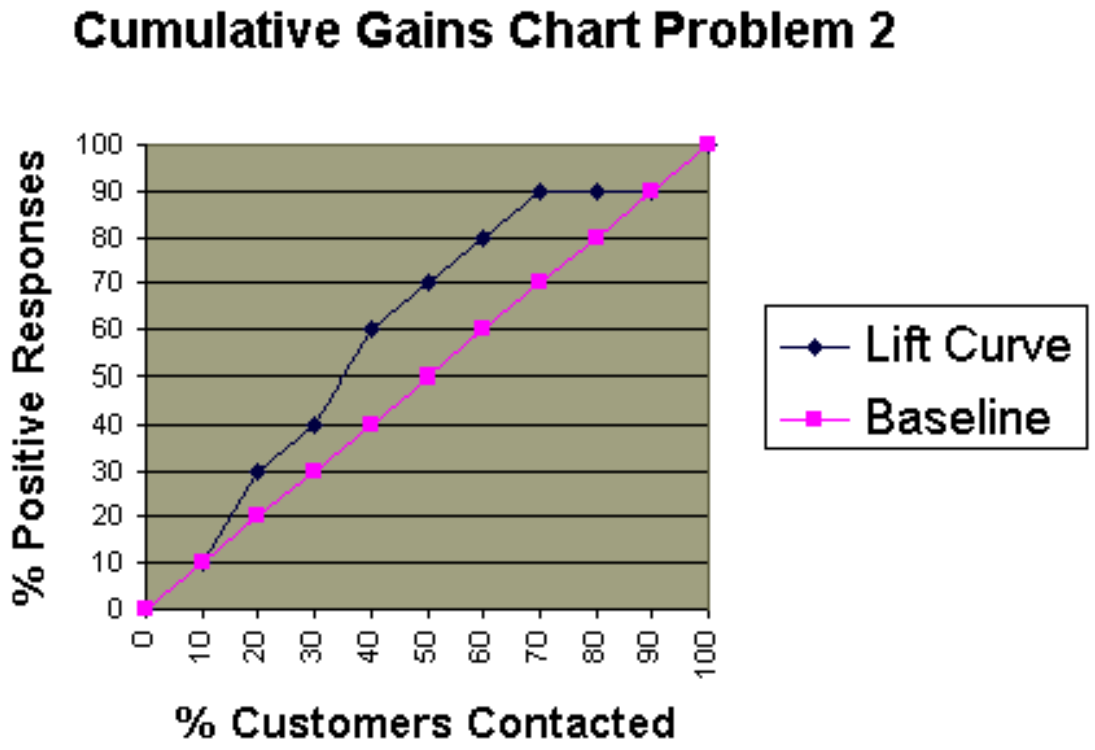
Response Rate = Number of Responses / Total Number of Responses (10 - Y)

Total Customers Contacted	Number of Responses	Response Rate
2	1	10%
4	3	30%
6	4	40%
8	6	60%
10	7	70%
12	8	80%
14	9	90%
16	9	90%
18	9	90%
20	10	100%

Another example gain and lift chart

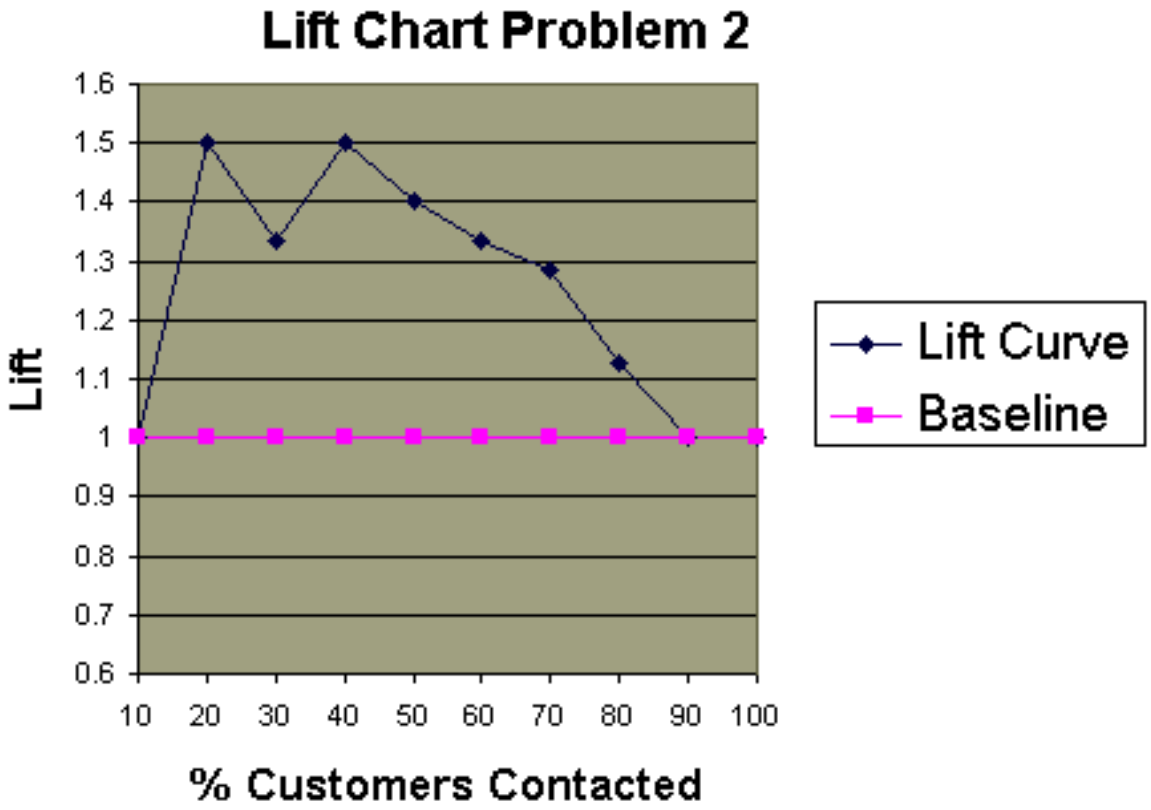
- 4. Create the cumulative gains chart:

The lift curve and the baseline have the same values for 0%-10% and 90%-100%.



Another example gain and lift chart

- **5. Create the lift chart:**



COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

comillas.edu