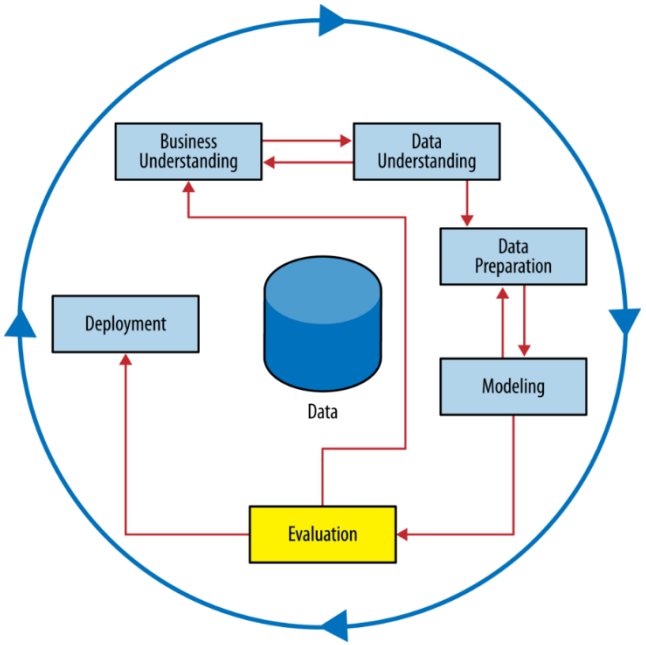
Visualize Model Performance

PF8

Learning Goals

- Ranking instead of Classifying
  - Which is the best threshold in class prediction?

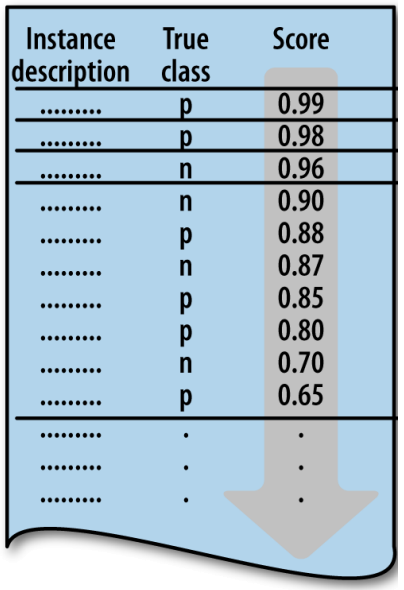
- Visualizing Model Performance
  - Profit Curve
  - ROC Curve
  - Cumulative Response Curve
  - Lift Curve
  - …



Ranking Instead of Classifying

- Score ranking
  - Rank customers by **their scores** in descending order.
    - **Score**: the estimated responding probability (i.e., positive probability).
  - Target customers on the top of the list.
    - That is, predict those with top scores as *Responder* (Y).

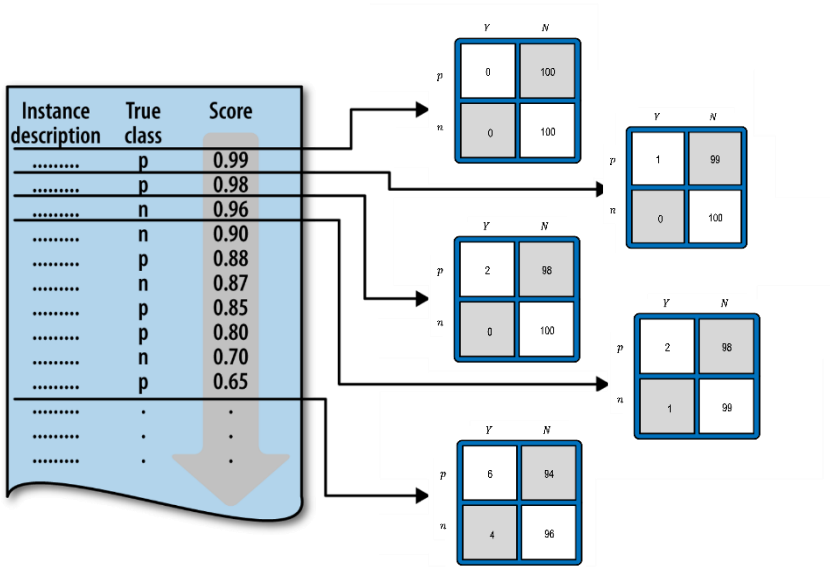
- What is the **threshold** (i.e., the cut-off) in class prediction?
  - How many customers should be targeted (i.e., predicted as Y)?



Ranking Instead of Classifying

- Suppose $(p, n) = (100, 100)$, a customer is predicted as
  - *Responder* (Y) if his/her score is above a predefined threshold.
  - *Non-responder* (N) his/her score is below the threshold.

- **Different thresholds** produces different classification results, each represented as a confusion matrix.
  - As the threshold decreases, more instances are predicted as Y.



- **How do we choose a proper threshold?**
  - How many customers shall be targeted (predicted as Y)?

Profit Curve

- **Profit curve** plots the estimated profit a model yields on a dataset against the percentage of positive predictions (i.e., %Y).
  - Class priors: the number of actual positive & negative instances in the dataset.

	p	n
#	500	500

  - Confusion matrix: produced by different thresholds given a model.

	Y	N
p	?	?
n	?	?

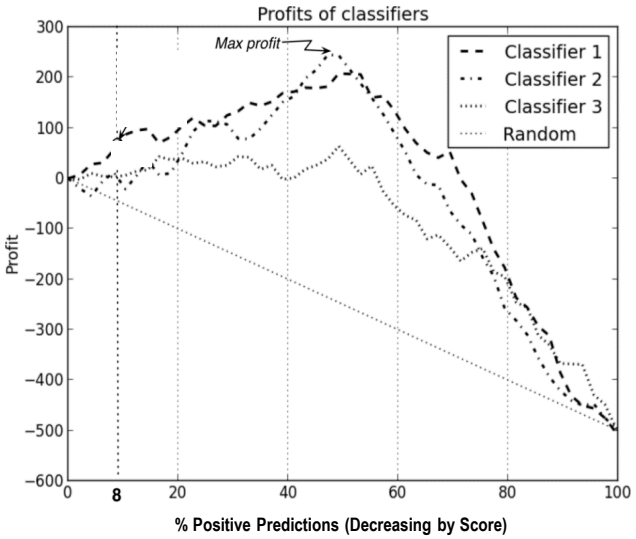
  - Cost & benefit matrix: business value for each prediction outcome
    - revenue = 9, cost = 5, profit = 4

	Y	N
p	\$4	\$0
n	-\$5	\$0

Profit Curve

- Some good models (Classifier 1, 2 or 3)
  - Step 1: Rank the 1000 customers by their score in decreasing order.
  - Step 2: Make class predictions at different thresholds (in decreasing order).
    - When the threshold decreases, more customers are predicted as Y (targeted).
  - Step 3: Plot total profit (y-axis) against the percentage of positive predictions (x-axis) made at different thresholds.

    - $\boldsymbol{x = 0\%}$ **(All predicted as N, target no one)**
      $Profit = \$0$
    - $\boldsymbol{x = 100\%}$ **(All predicted as Y)**
      $Profit = 500 \times \$4 + 500 \times (-\$5) = -\$500$
    - $\boldsymbol{x = 40\%}$ **(Top 400 predicted as Y)**
      $(Assume\ 240\ p\ and\ 160\ n)$
      $Profit = 240 \times \$4 + 160 \times (-\$5) = \$160$



If budget allows, max profit is achieved with classifier 2, when 50% instances are targeted.

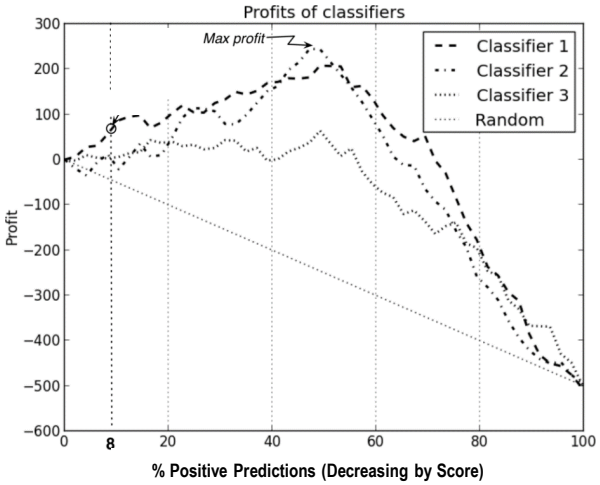
Profit Curve: Random Model

- **Random classifier** randomly predict an instance as Y or N.
  - That is, 50% probability for each class.

- When a random classifier predicts 40% instances as Y, 40% of p and n are predicted as positive.
  - In a balanced dataset $(p, n) = (500,500)$

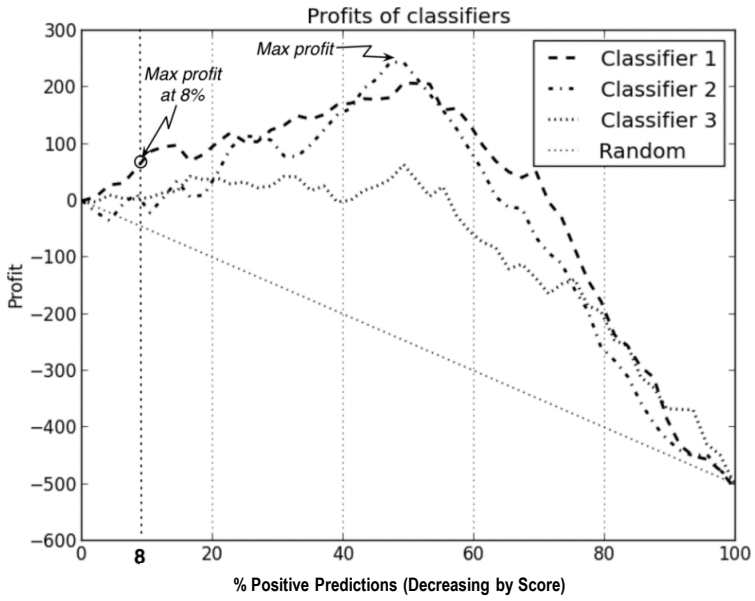
	Y	N
p	200	300
n	200	300

  - What is the profit?
    - $200 \times \$4 + 200 \times (-\$5) = -\$200$



Exercise: when budget is limited…

- A marketing team has $4,000 budget for a campaign. There are 10,000 customers in total, and cost is $5 per offer (i.e., targeting).
  - How many customers can the team target at most?
    - How many customers can be predicted as positive (Y) in maximum?
  - Compare classifier 1,2,3, which one shall be adopted?

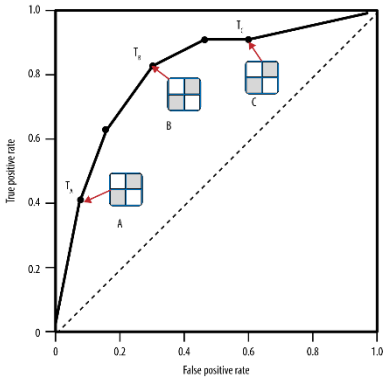


ROC Curve

- **Receiver Operating Characteristics** (ROC) curve shows trade-off between benefits (TP) and costs (FP).
  - y-axis: TP Rate (*Sensitivity/Recall*)
    - $TP\ rate = \frac{TP}{TP+FN} = \frac{TP}{p}$
  - x-axis: FP Rate (*False alarm rate, 1 – Specificity*)
    - $FP\ rate = \frac{FP}{FP+TN} = \frac{FP}{n}$

	Y	N
p	TP	FN
n	FP	TN

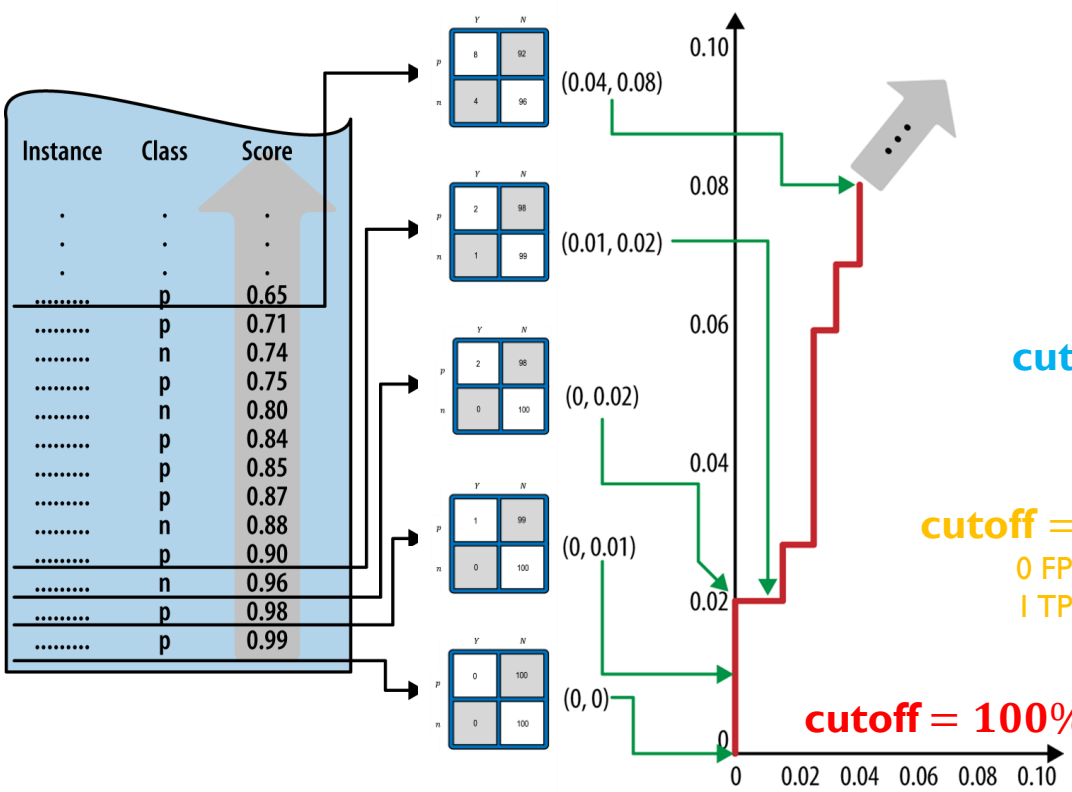
- ROC curve is independent of the cost & benefit matrix and class distribution.
  - Each classifier produces a curve in ROC space.
  - Each point on the curve represents a confusion matrix, produced by different thresholds.



- ROC curve shows trade-off between sensitivity and specificity *(i.e., TNR).*
  - TPR ↑ (sensitivity ↑) ➔ FPR ↑ (False alarm rate ↑, specificity ↓)

Plot a ROC Curve

- Sort instances by score in ascending order, start from the bottom:
  - The threshold at the bottom is 100%, which yields all N predictions.
  - As the threshold decreases, both TP and FP increase (more Y predictions).



cutoff = 65%
4 FP (FPR = 4/100)
8 TP (TPR = 8/100)

cutoff = 90%
1 FP (FPR = 1/100)
2 TP (TPR = 2/100)

cutoff = 96%
0 FP (FPR = 0/100)
2 TP (TPR= 2/100)

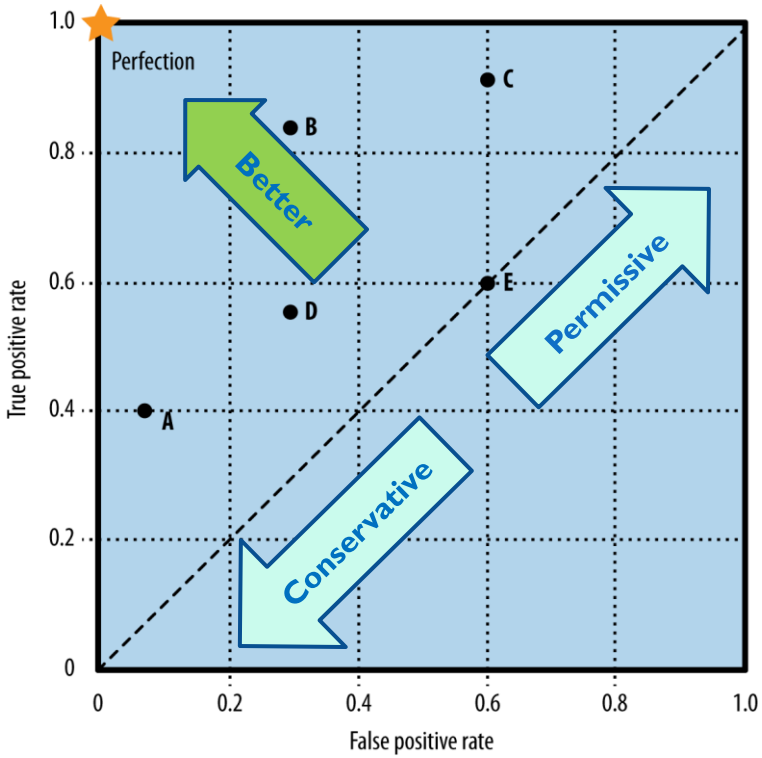
cutoff = 98%
0 FP (FPR= 0/100)
1 TP (TPR= 1/100)

cutoff = 100% *(FPR= 0, TPR= 0)*

Important Points at ROC Space

- **(0,0)**: all instances predicted as N (i.e., FPR = 0, TPR = 0).
- **(1,1)**: all instances predicted as Y (I.e., FPR = 100%, TPR = 100%).
- **(0,1)**: perfect classification (i.e., FPR = 0, TPR = 100%).
- The diagonal line: random model.
  - %Y = TPR = FPR (e.g., point E).

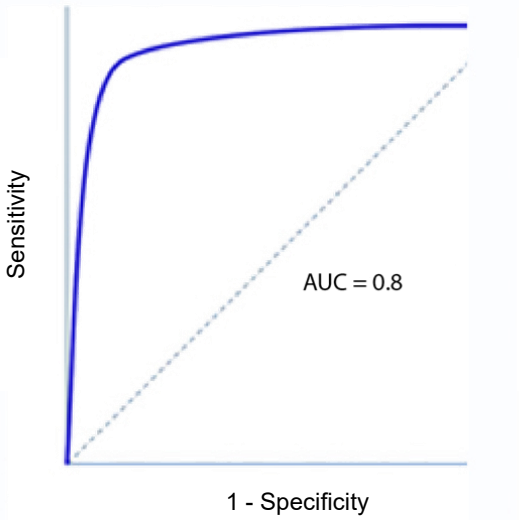
- Degree of *conservativeness*
  - Alarm only with strong evidence (high threshold)
  - $A \succ D \succ B \succ C$
- Degree of *permissiveness*
  - Alarm even without much evidence (low threshold)
  - $C \succ B \succ D \succ A$

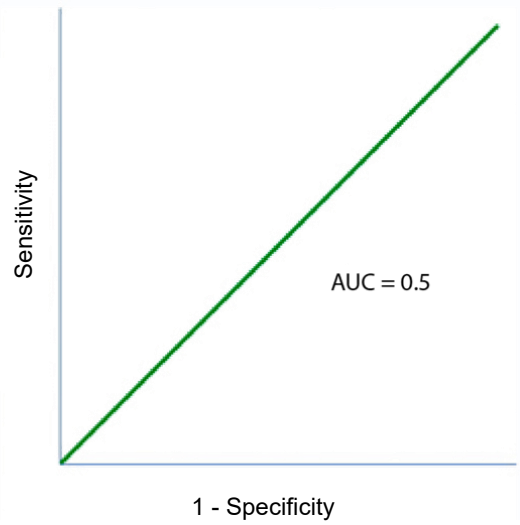


With unbalanced data, even a small FPR is unmanageable!
Imagine for a customer data $(p, n) = (1000, 100{,}000)$, 1% FPR yields 1000 FP predictions!

Area Under Curve

- **Area under curve** (AUC) is the proportion of area under a ROC curve against the entire ROC space.
  - Area of the whole ROC space = 1
  - AUC = 1: perfect classification
  - AUC = 0.5: randomness (TPR = FPR)

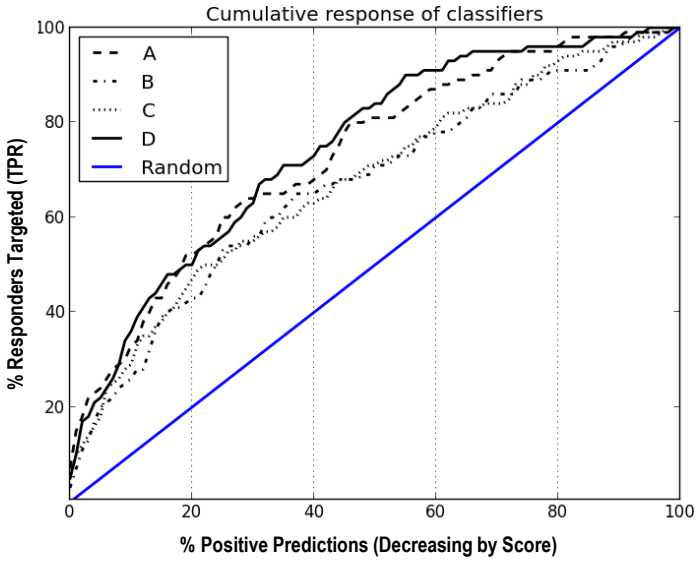




AUC can be used for hyperparameter tuning when threshold is unknown!

Cumulative Response Curve

- **Cumulative response curve** (CRC) plots TP rate (i.e., sensitivity) against the percentage of positive predictions made (i.e., %Y).
  - y-axis: percentage of actual responders (p) predicted as Responders (Y)
  - x-axis: percentage of customers predicted as Responders (Y)



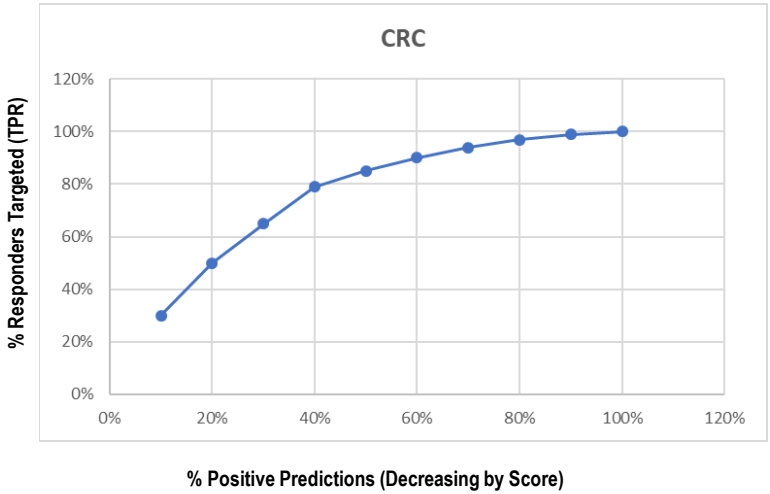
CRC is more intuitive for business stakeholder.

- Diagonal line $y = x$ represents the random model (%Y=TPR=FPR).
  - Any classifiers above the diagonal line are better than *random classifier.*

Example: Plot CRC

- For an unbalanced data: $(p, n) = (20{,}000, 80{,}000)$
  - Total number of instances: 100,000

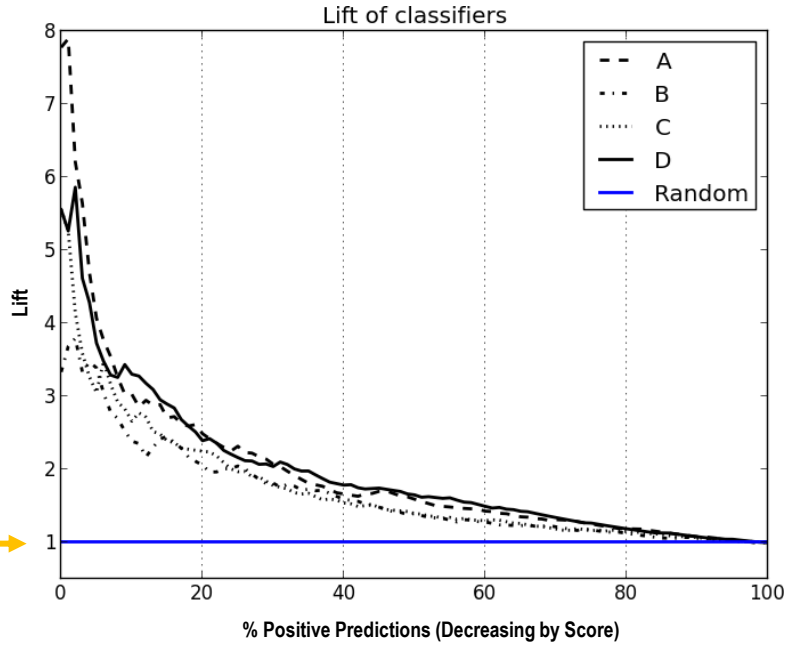
Total Customers Targeted (% 'Y')	Cumulative Responses (TPR = $\frac{TP}{p}$)
10,000 (10%)	6,000 (30%)
20,000 (20%)	10,000 (50%)
30,000 (30%)	13,000 (65%)
40,000 (40%)	15,800 (79%)
50,000 (50%)	17,000 (85%)
60,000 (60%)	18,000 (90%)
70,000 (70%)	18,800 (94%)
80,000 (80%)	19,400 (97%)
90,000 (90%)	19,800 (99%)
100,000 (100%)	**20,000 (100%)**



Lift Curve

- **Lift Curve** (LC) shows the advantage a model provides, compared to the random model.
  - i.e., the degree to which a model "pushes up" positive instances (p) above negative instances (n) in a score rank list.

$$Lift = \frac{TPR\ (y\ axis\ of\ CRC)}{\%Y\ (x\ axis\ of\ CRC\ )}$$

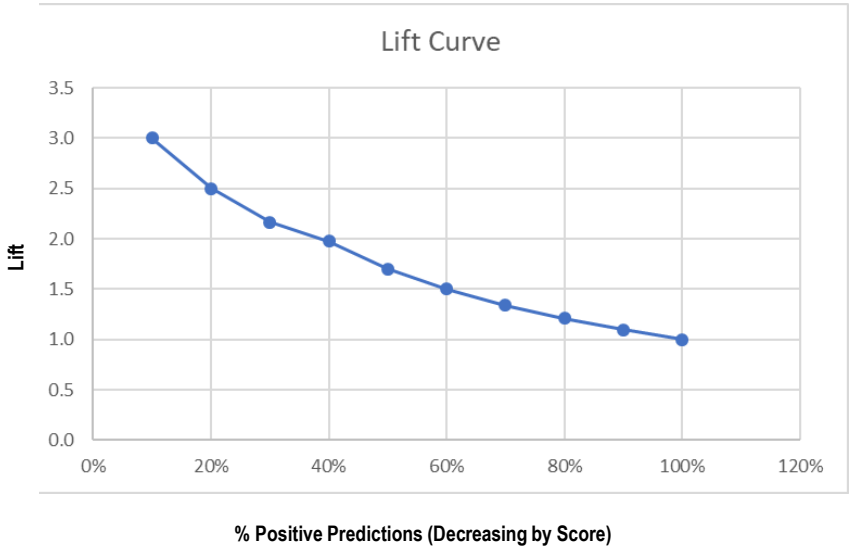


Random classifier: lift = 1

- lift = 2 means:
  - At the chosen threshold, the model is twice as good as the random model.

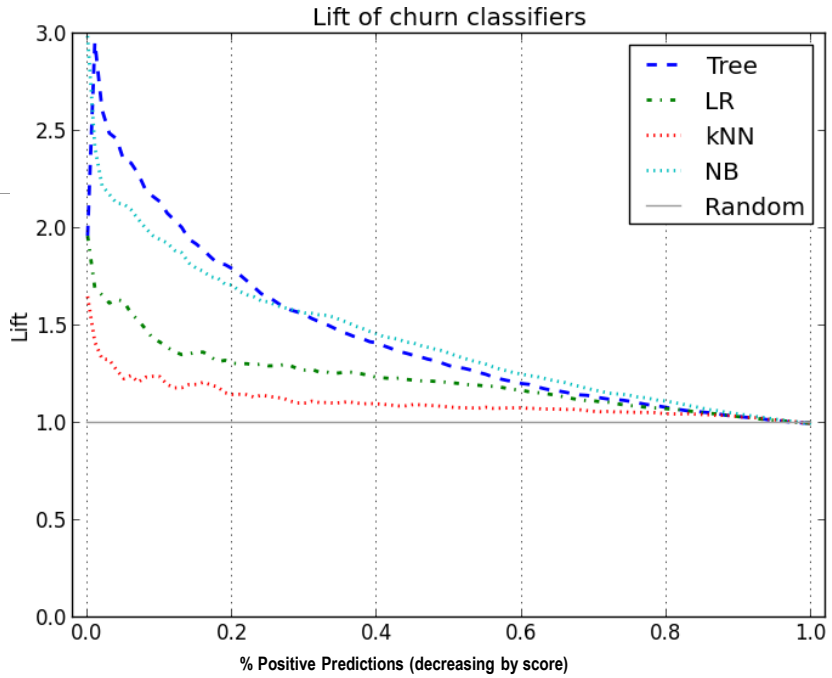
Example: Plot Lift Curve

Total Customers Targeted (% 'Y')	Cumulative Responses (TP rate)	Lift
10,000 (10%)	6,000 (30%)	3
20,000 (20%)	10,000 (50%)	2.5
30,000 (30%)	13,000 (65%)	2.17
40,000 (40%)	15,800 (79%)	1.98
50,000 (50%)	17,000 (85%)	1.7
60,000 (60%)	18,000 (90%)	1.5
70,000 (70%)	18,800 (94%)	1.34
80,000 (80%)	19,400 (97%)	1.21
90,000 (90%)	19,800 (99%)	1.1
100,000 (100%)	**20,000 (100%)**	**1**



Model Evaluation with Lift Curve

- Compare four models based with lift curves.



- If targeting top 25% customers, the tree model is the best.
- If targeting more than 25% customers, the Naive Bayes model performs the best.