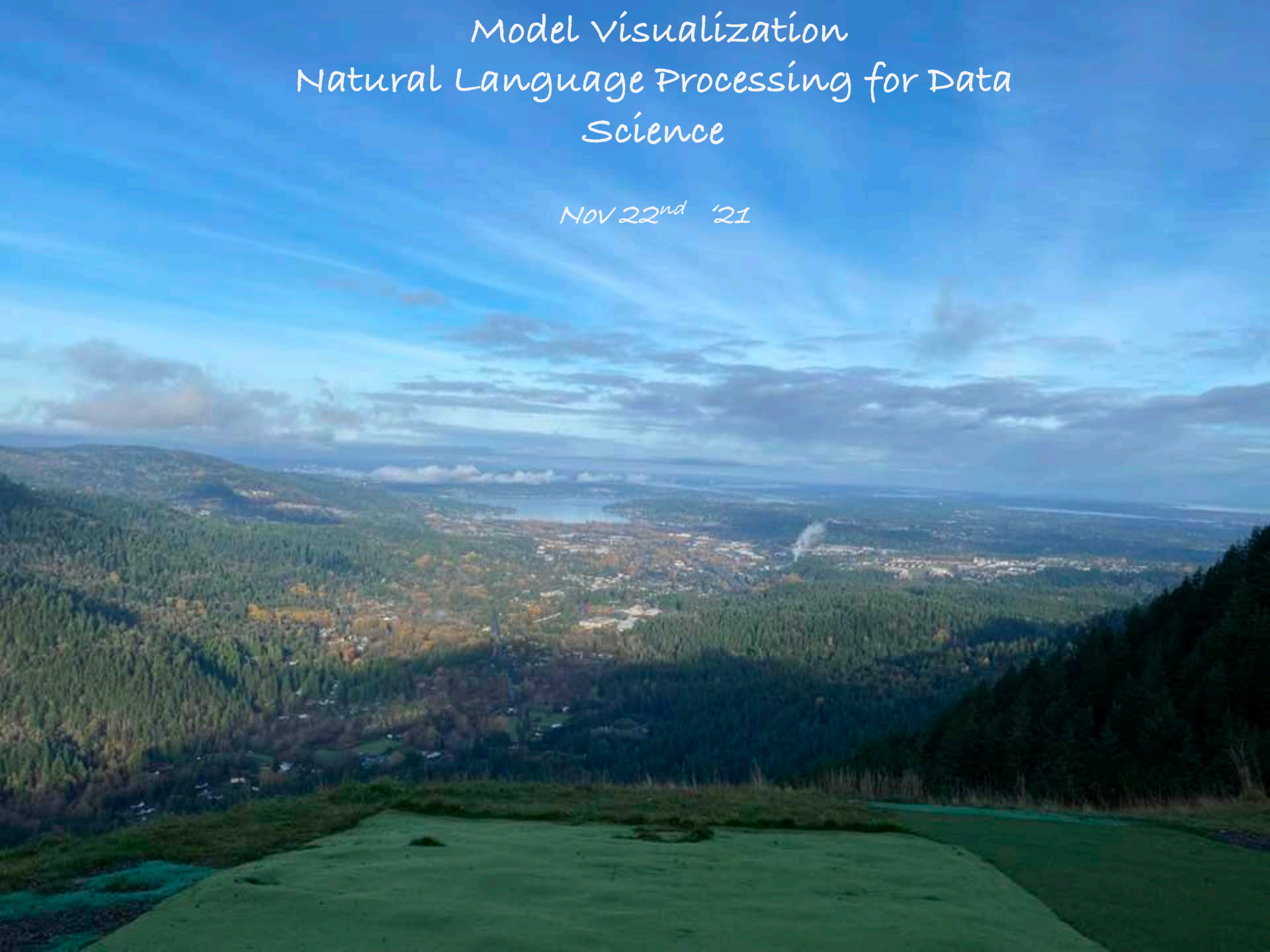
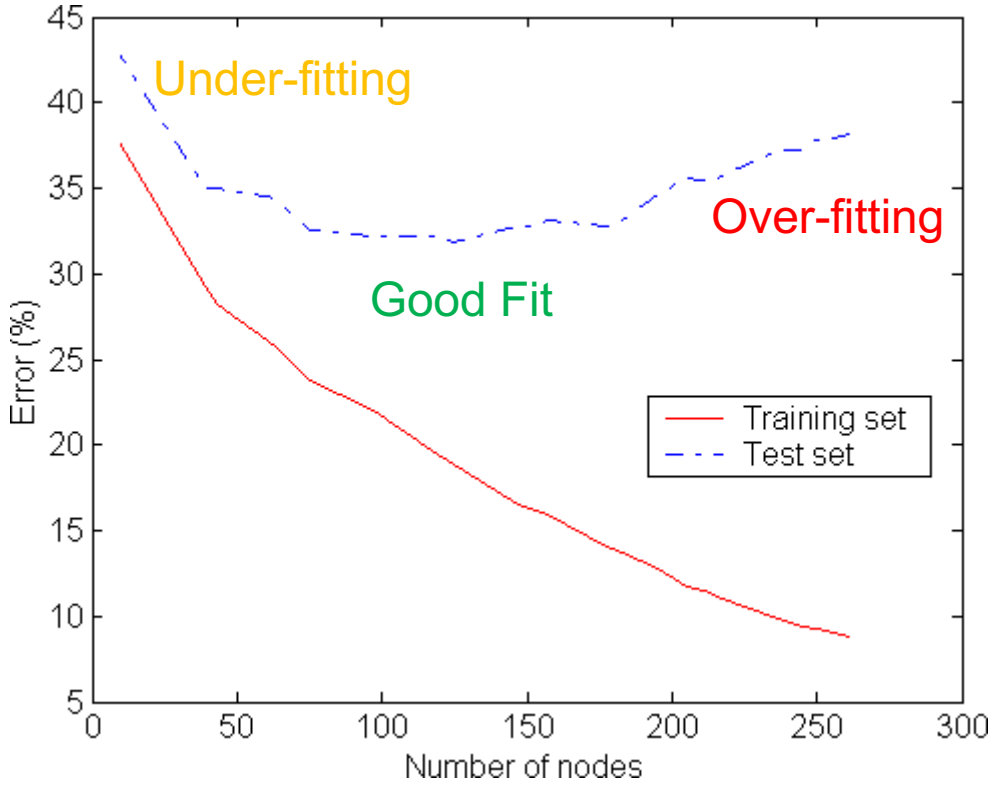
Model Visualization
Natural Language Processing for Data Science

Nov 22nd '21



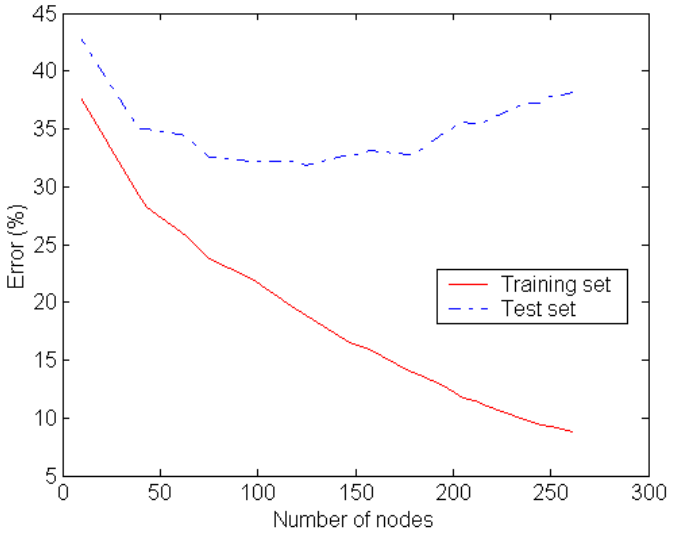
Over-fitting



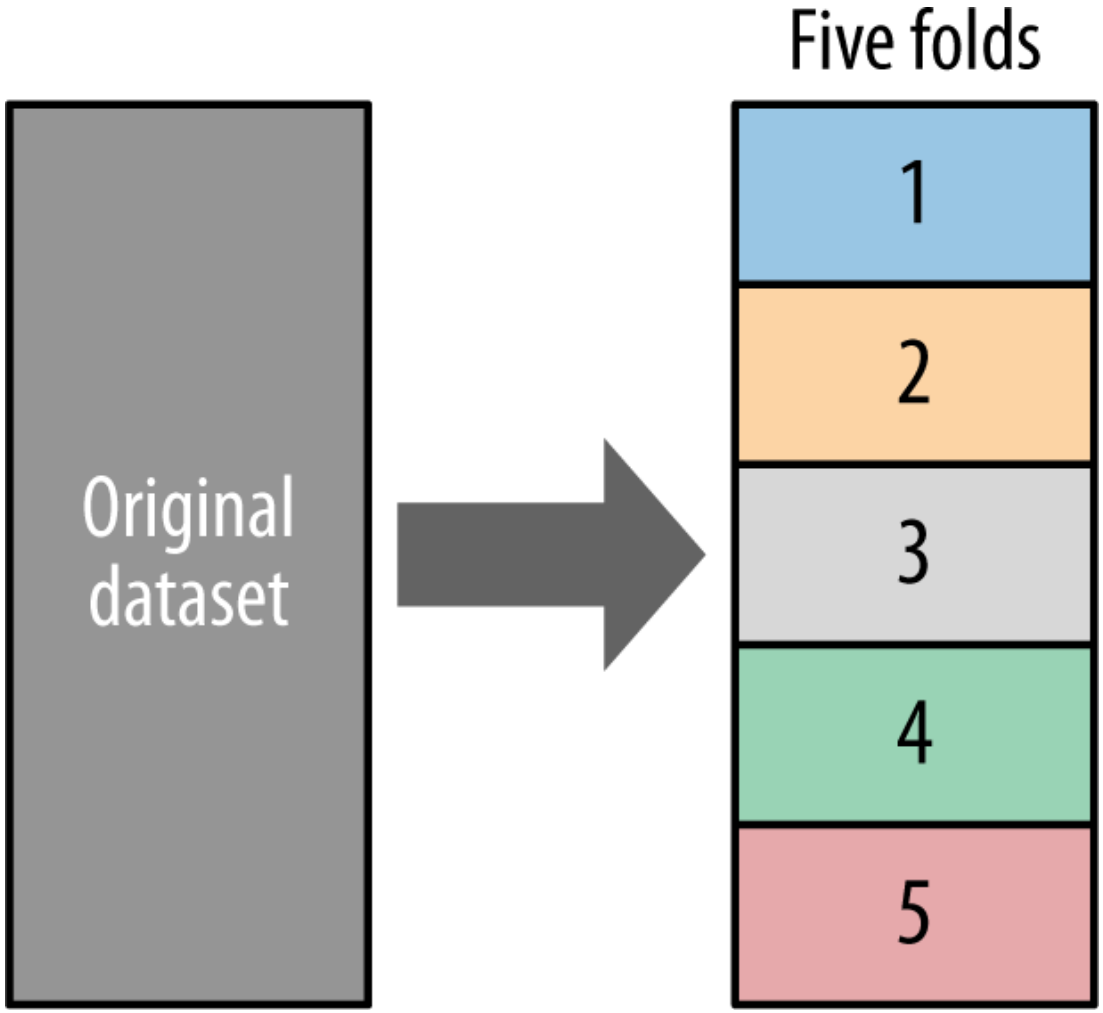
- Over-fitting: Model "memorizes" the properties of the particular training set rather than learning the underlying concept or phenomenon

Holdout validation

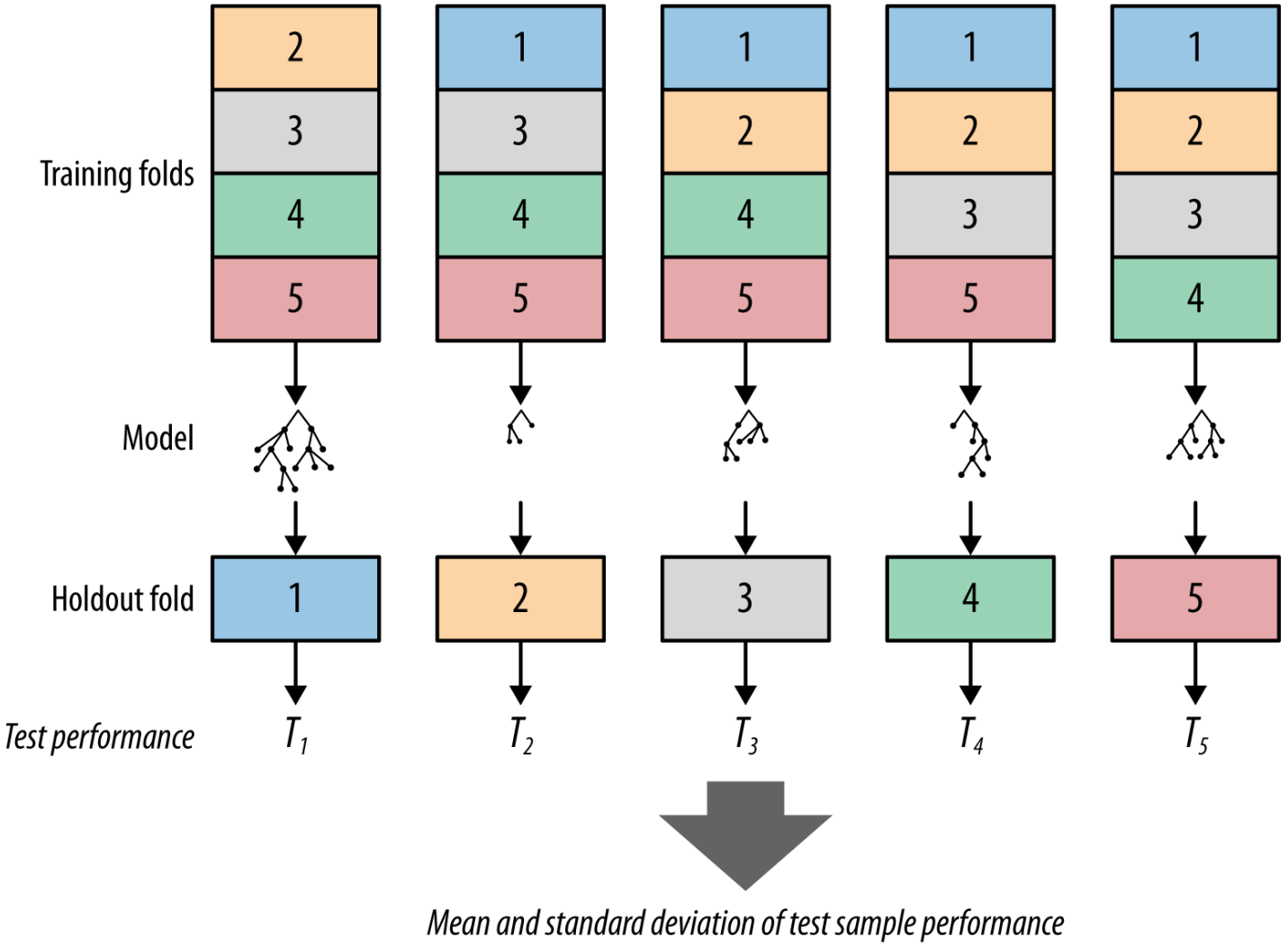
- We are interested in generalization
  - The performance on data not used for training
- Given only one data set, we hold out some data for evaluation
  - Holdout set for final evaluation is called the test set
- Accuracy on training data is sometimes called **"in-sample" accuracy**, vs. **"out-of-sample" accuracy** on test data



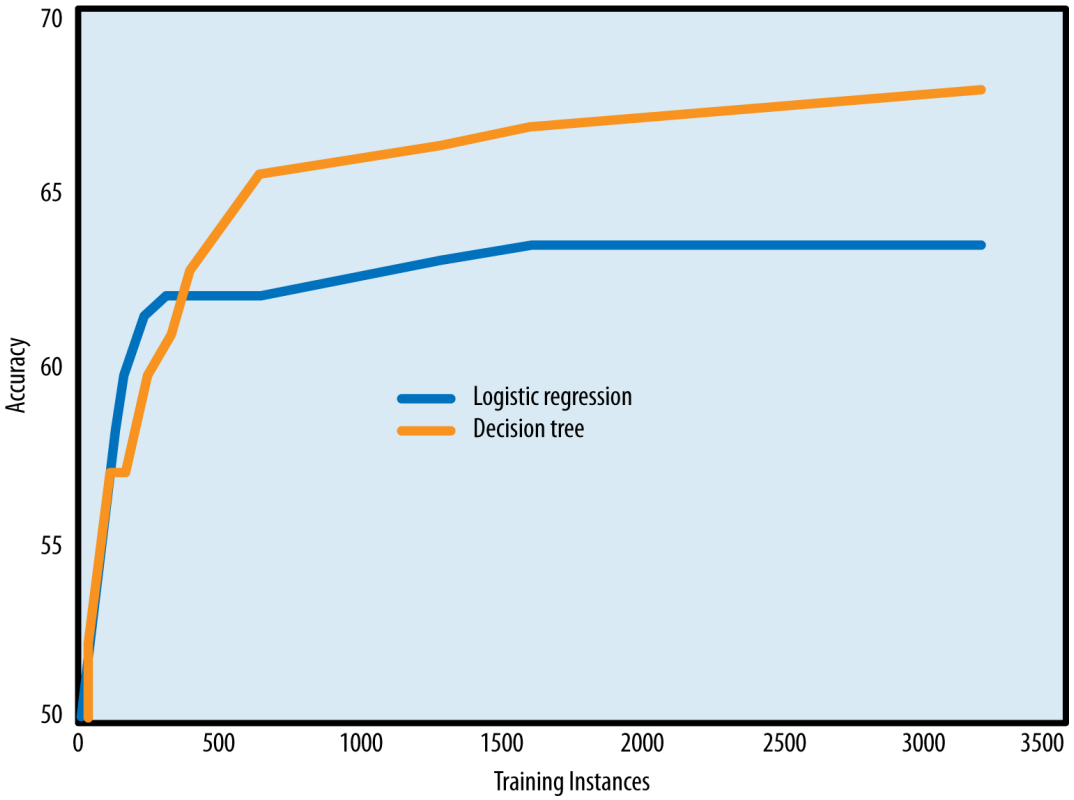
Cross-Validation



Cross-Validation

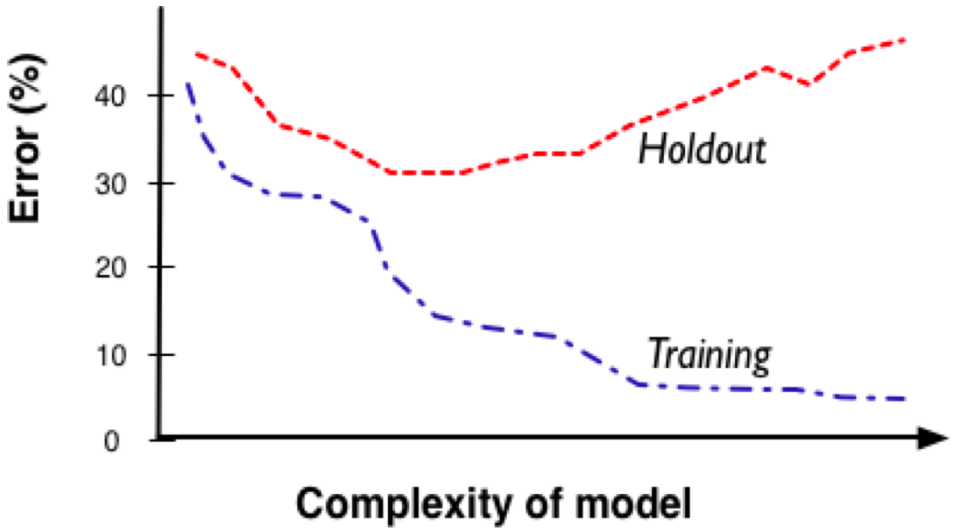


Learning Curves



Learning curves vs Fitting graphs

- A learning curve shows the generalization performance plotted against the amount of training data used
- A fitting graph shows the generalization performance as well as the performance on the training data, but plotted against model complexity
- Fitting graphs generally are shown for a fixed amount of training data



Avoiding Over-fitting

Tree Induction:

- Post-pruning
  - takes a fully-grown decision tree and discards unreliable parts
- Pre-pruning
  - stops growing a branch when information becomes unreliable

Linear Models:

- Feature Selection
- Regularization
  - Optimize some combination of fit and simplicity

Regularization

Regularized linear model:

$$\arg\max_{\mathbf{w}} \; [\text{fit}(\mathbf{x}, \mathbf{w}) - \lambda \cdot \text{penalty}(\mathbf{w})]$$

- "L2-norm"
  - The sum of the *squares* of the weights
  - L2-norm + standard least-squares linear regression = ridge regression
- "L1-norm"
  - The sum of the *absolute values* of the weights
  - L1-norm + standard least-squares linear regression = lasso
  - Automatic feature selection

Evaluating Classifiers: Plain Accuracy

$$\text{Accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

$$= 1 - \text{error rate}$$

- *Too* simplistic..

Evaluating Classifiers: The Confusion Matrix

- A confusion matrix for a problem involving $n$classes is an $n\times n$ matrix,
  - with the columns labeled with actual classes and the rows labeled with predicted classes
- It separates out the decisions made by the classifier,
  - making explicit how one class is being confused for another

	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives

- The errors of the classifier are the false positives and false negatives

Building a Confusion Matrix

Default Truth	Model Prediction
0	0
1	1
0	1
0	1
0	0
1	1
0	0
0	0
1	1
1	0

Predicted class / Actual class	Default	No Default	Total
Default	3	1	4
No Default	2	4	6
Total	5	5	10

Other Evaluation Metrics

- $\text{Precision} = \frac{TP}{TP + FP}$
- $\text{Recall} = \frac{TP}{TP + FN}$
- $\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

A Key Analytical Framework: Expected Value

- The expected value computation provides a framework that is useful in organizing thinking about data-analytic problems
- It decomposes data-analytic thinking into:
  - the structure of the problem,
  - the elements of the analysis that can be extracted from the data, and
  - the elements of the analysis that need to be acquired from other sources
- The general form of an expected value calculation:

$$EV = p(o_1) \cdot v(o_1) + p(o_2) \cdot v(o_2) + p(o_3) \cdot v(o_3) \ldots$$

$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

Expected Value Framework in Use Phase

- Online marketing:

- Expected benefit of targeting = p_R

$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

- Product Price: $200

- Product Cost: $100

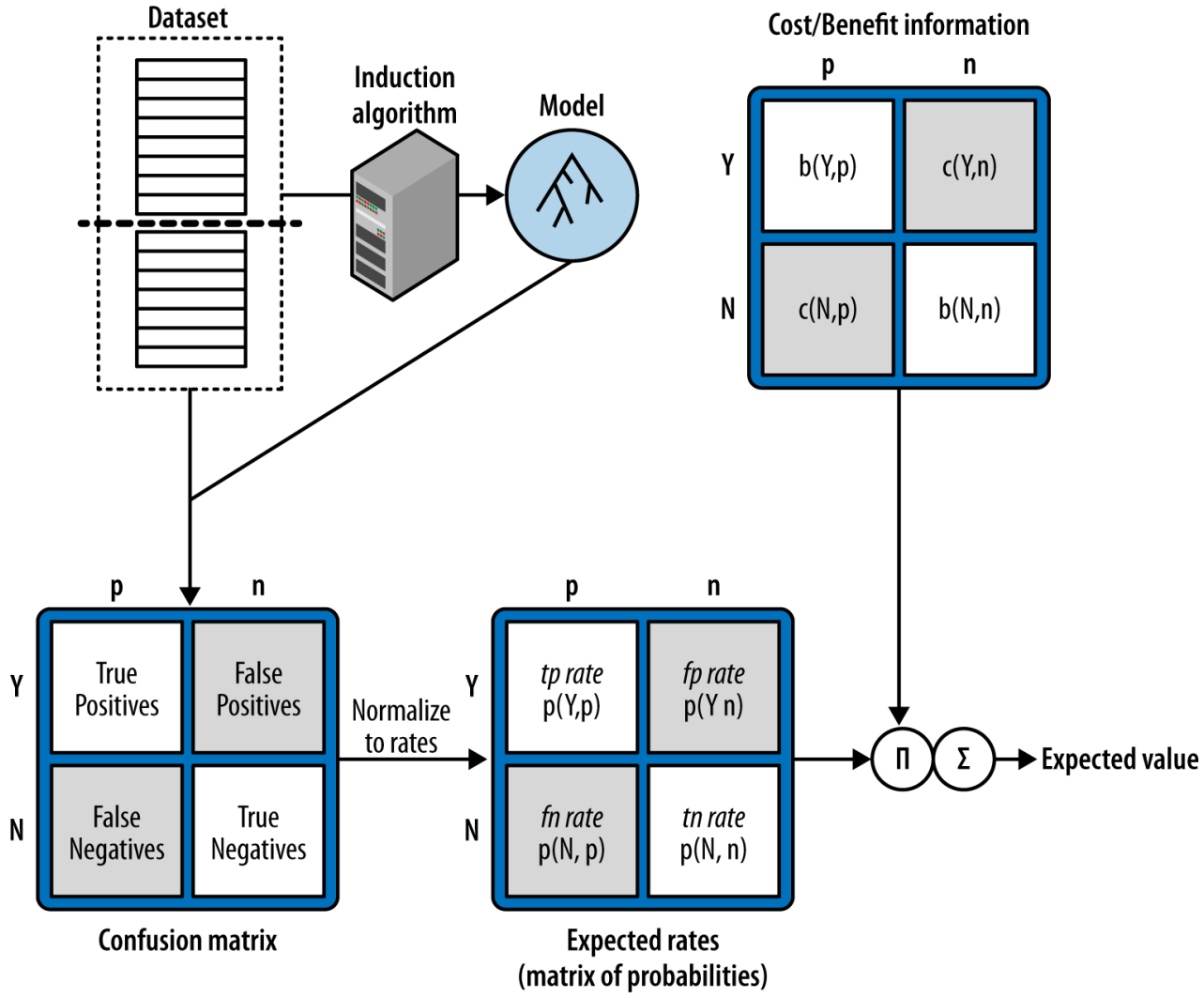
- Targeting Cost: $1

$$p_R(\mathbf{x}) \cdot \$99 - [1 - p_R(\mathbf{x})] \cdot \$1 > 0$$

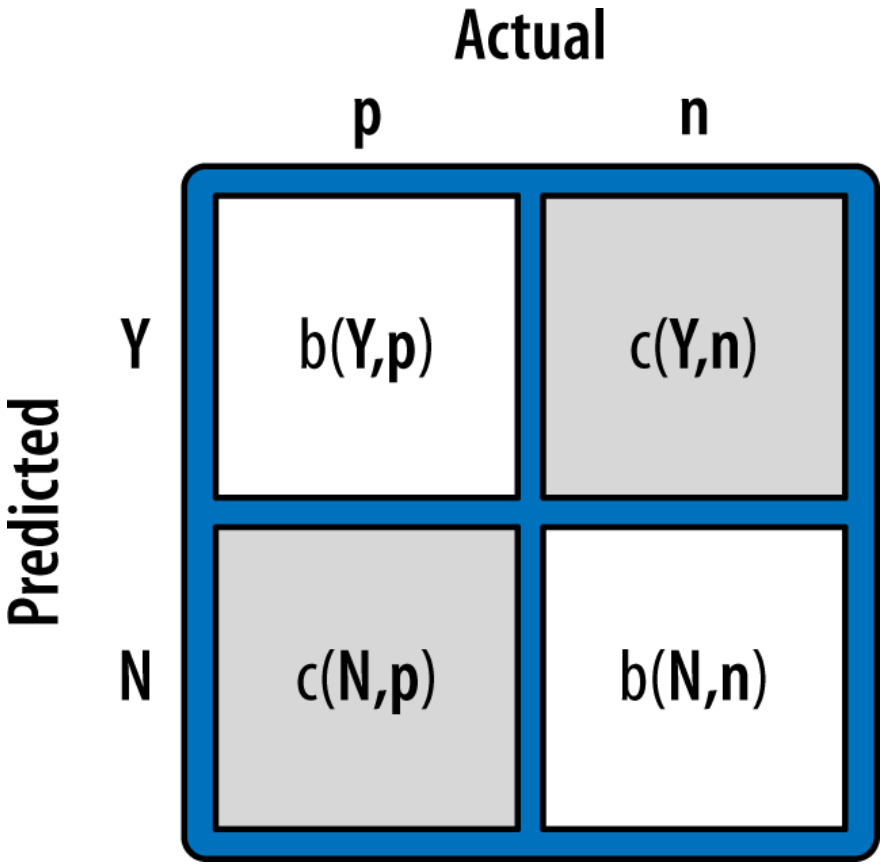
$$p_R(\mathbf{x}) \cdot \$99 > [1 - p_R(\mathbf{x})] \cdot \$1$$

$$p_R(\mathbf{x}) > 0.01$$

Using Expected Value to Frame Classifier Evaluation



A cost-benefit matrix



A cost-benefit matrix for the marketing example

		Actual	
		p	**n**
Predicted	**Y**	99	-1
	N	0	0

Using Expected Value to Frame Classifier Evaluation

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

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

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

	p	n
Y	56	7
N	5	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

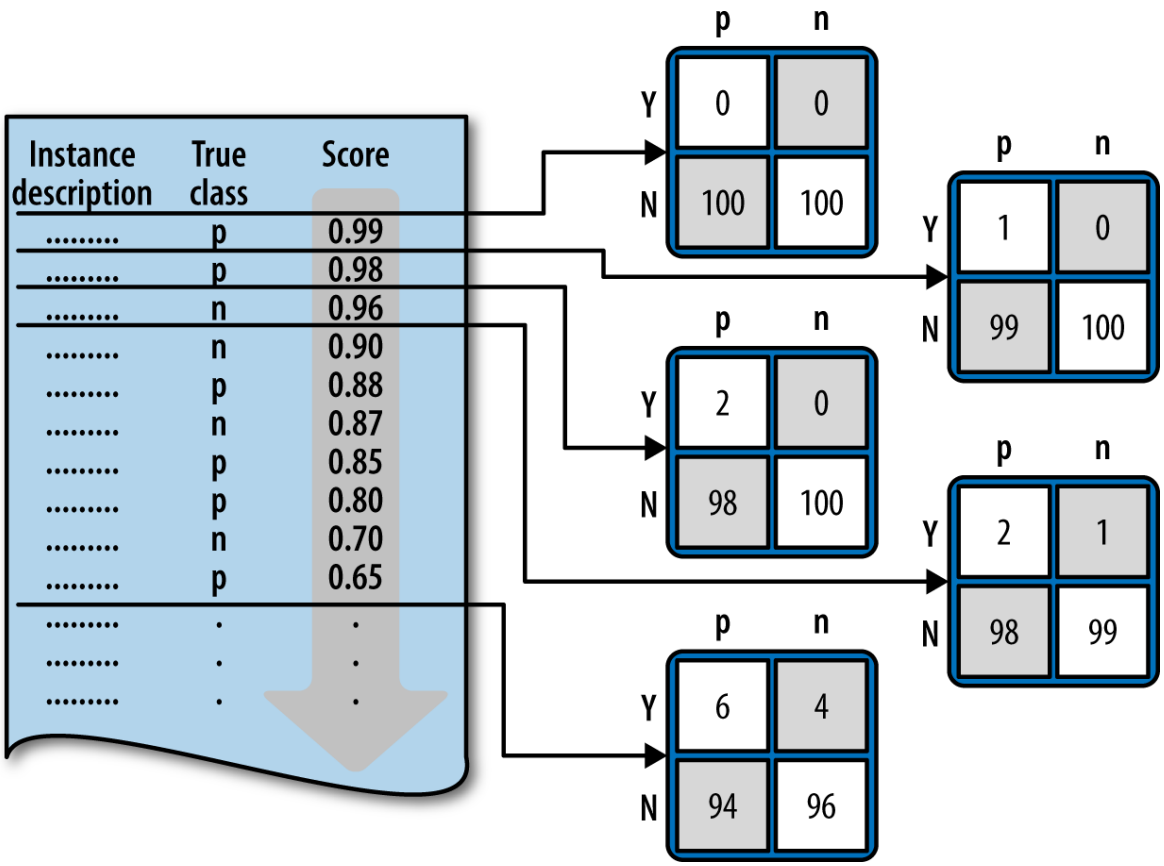
$$
\begin{aligned}
\text{expected profit} &= p(\mathbf{p}) \cdot [\, p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\
&\quad\; p(\mathbf{n}) \cdot [\, p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y}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

Other terms

Precision = TP / (TP+FP)
Recall = TP / (TP + FN)

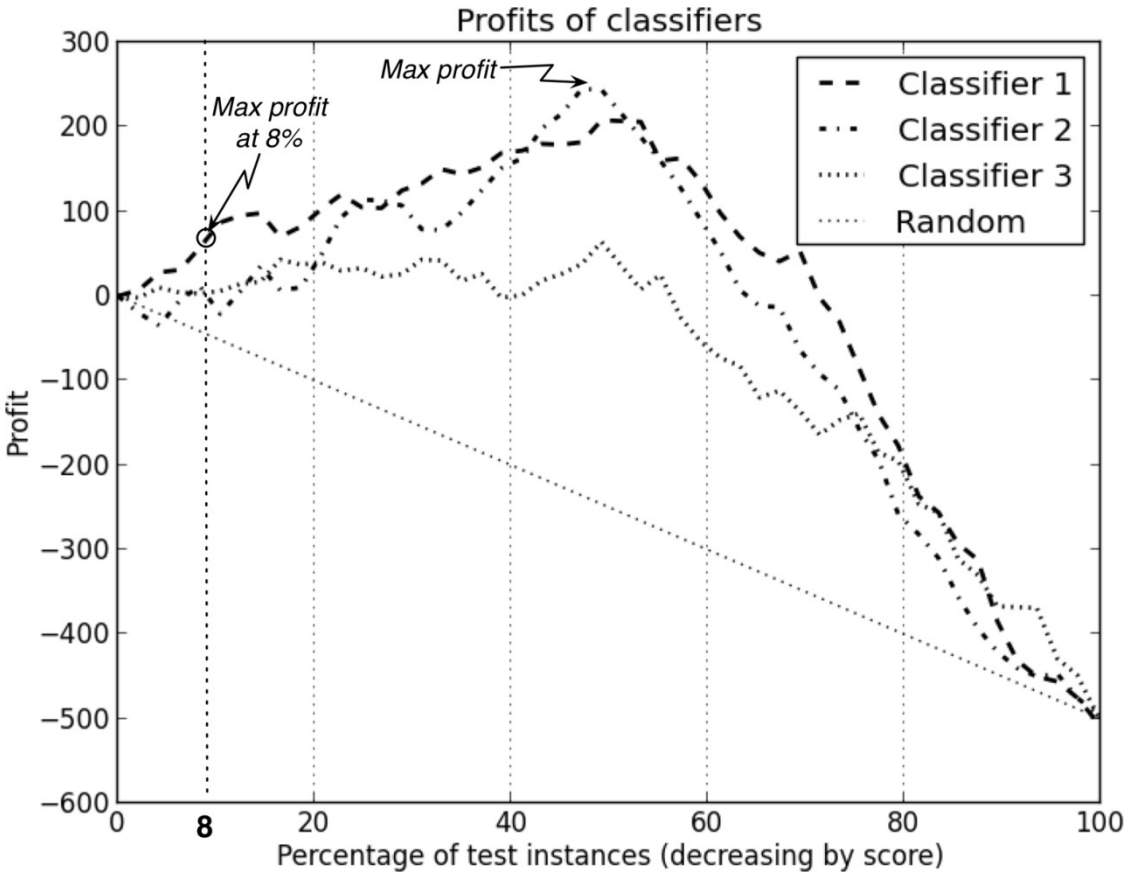
Sensitivity = $TN / (TN + FP)$ = True negative rate = 1 - False positive rate
Specificity = $TP / (TP + FN)$ = True positive rate

Ranking Instead of Classifying



Profit Curves

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



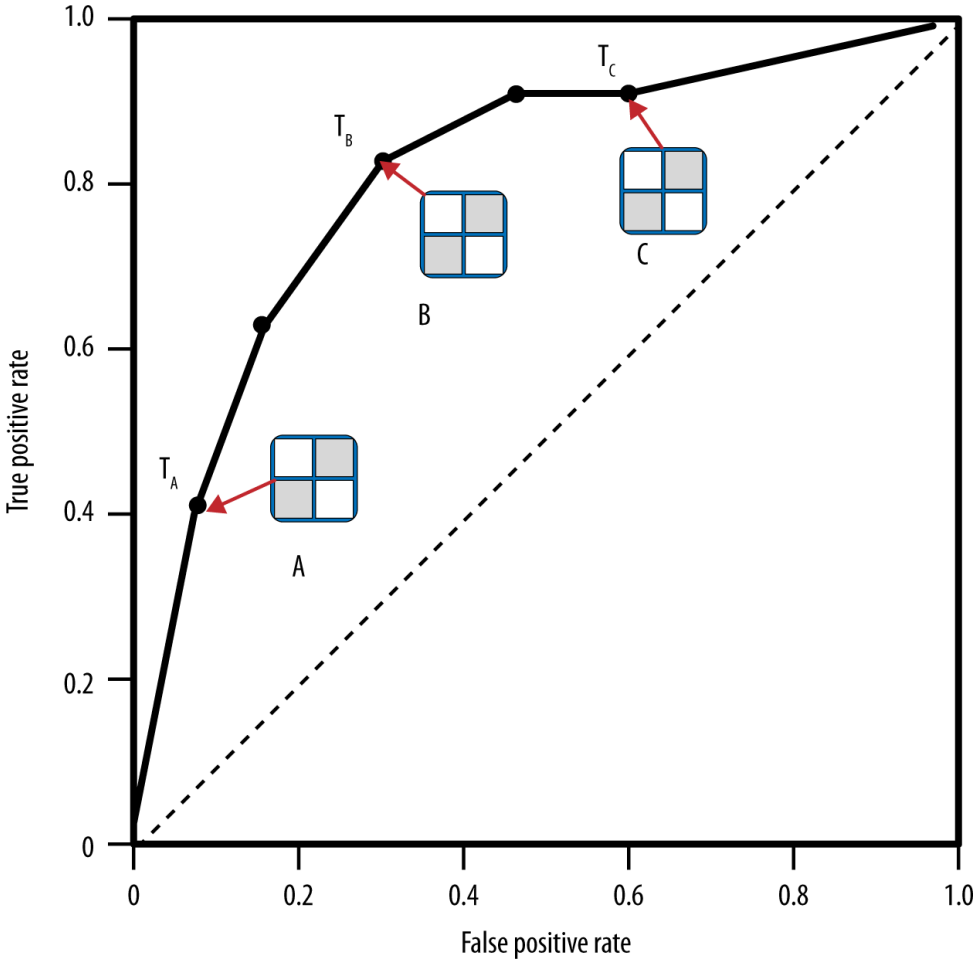
Profit Curves

There are two critical conditions underlying the profit calculation:

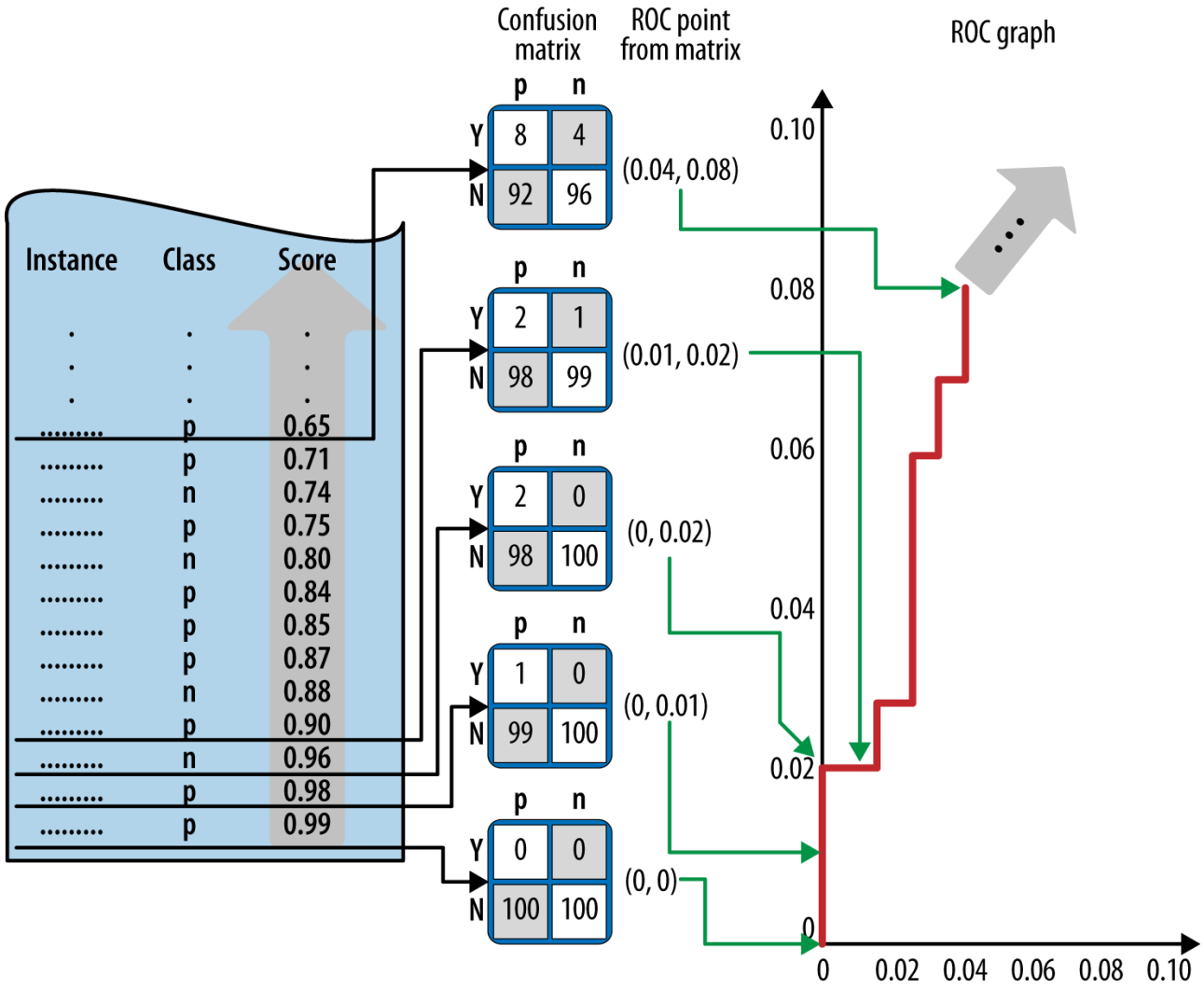
- The class priors
  - The proportion of positive and negative instances in the target population
- The costs and benefits
  - The expected profit is specifically sensitive to the relative levels of costs and benefits for the different cells of the cost-benefit matrix

Reality..???

ROC Graphs and Curves



ROC Graphs and Curves



Generating ROC curve: Algorithm

- Sort the test set by the model predictions
- Start with cutoff = max (prediction)
- Decrease cutoff, after each step count the number of true positives TP (positives with prediction above the cutoff) and false positives FP (negatives above the cutoff)
- Calculate TP rate (TP/P) and FP (FP/N) rate
- Plot current number of TP/P as a function of current FP/N

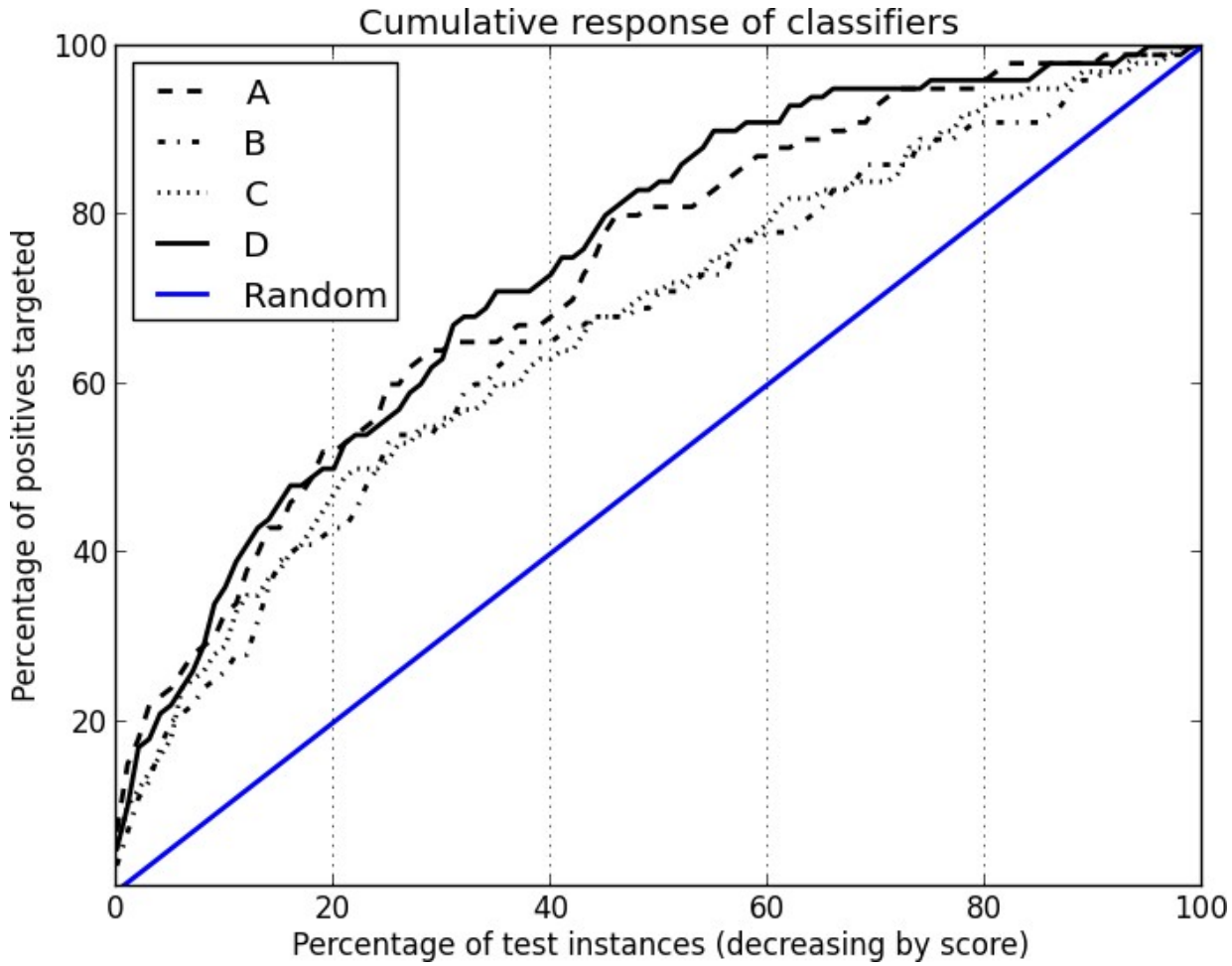
ROC Graphs and Curves

- ROC graphs decouple classifier performance from the conditions under which the classifiers will be used
- ROC graphs are independent of the class proportions as well as the costs and benefits
- Not the most intuitive visualization for many business stakeholders

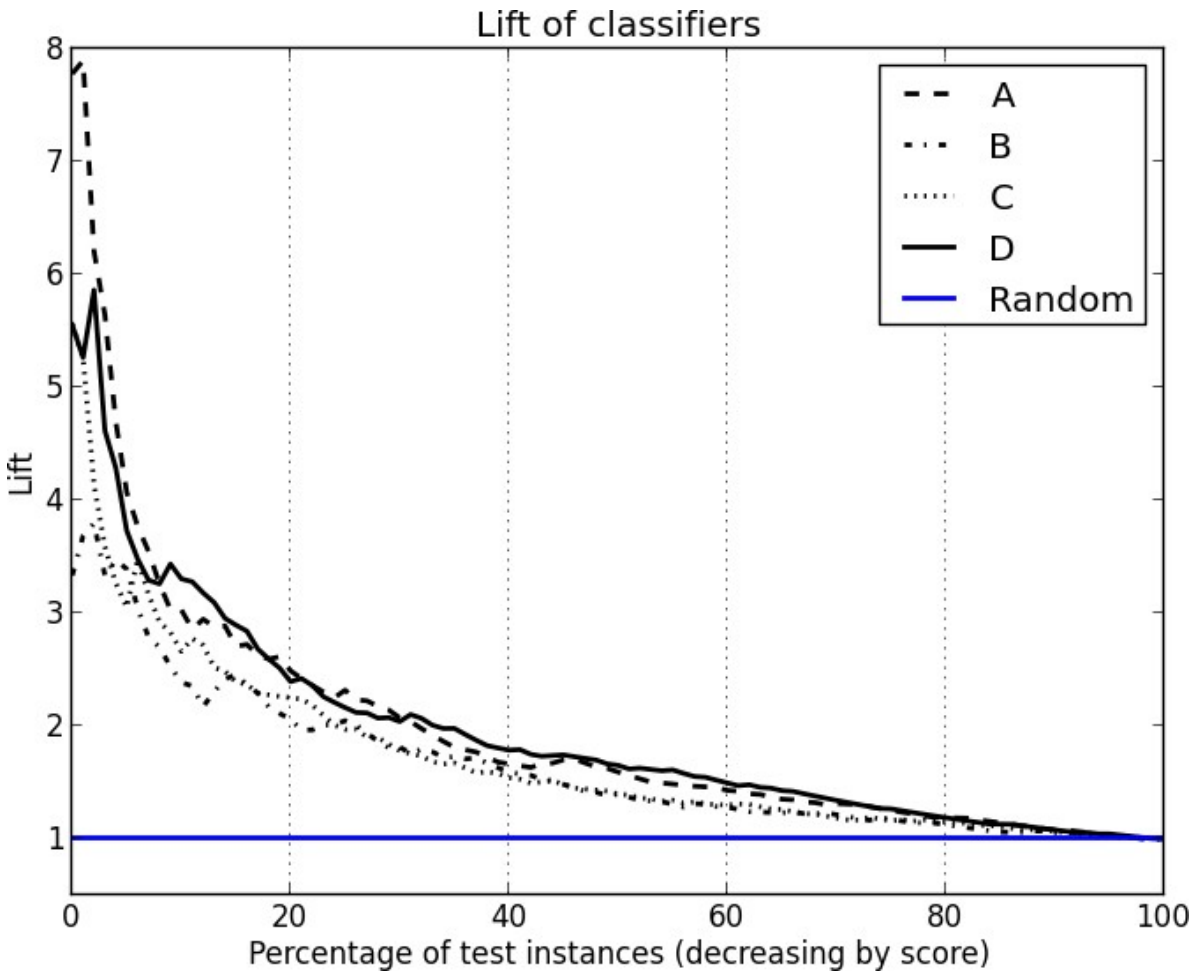
Area Under the ROC Curve (AUC)

- The area under a classifier's curve expressed as a fraction of the unit square
  - Its value ranges from zero to one
- The AUC is useful when a single number is needed to summarize performance, or when nothing is known about the operating conditions
  - A ROC curve provides more information than its area
- Equivalent to the Mann-Whitney-Wilcoxon measure
  - Also equivalent to the Gini Coefficient (with a minor algebraic transformation)
  - Both are equivalent to the probability that a randomly chosen positive instance will be ranked ahead of a randomly chosen negative instance

Cumulative Response curve



Lift Curve


Lift of classifiers
A
B
C
D
Random
Lift
Percentage of test instances (decreasing by score)
0
20
40
60
80
100
1
2
3
4
5
6
7
8

Performance Evaluation

Training Set:

Model	Accuracy
Classification Tree	95%
Logistic Regression	93%
k-Nearest Neighbors	100%
Naïve Bays	76%

Test Set:

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 Neighbors	93.0%±0.0	0.537±0.015
Naïve Bays	76.5%±0.6	0.632±0.019

Performance Evaluation

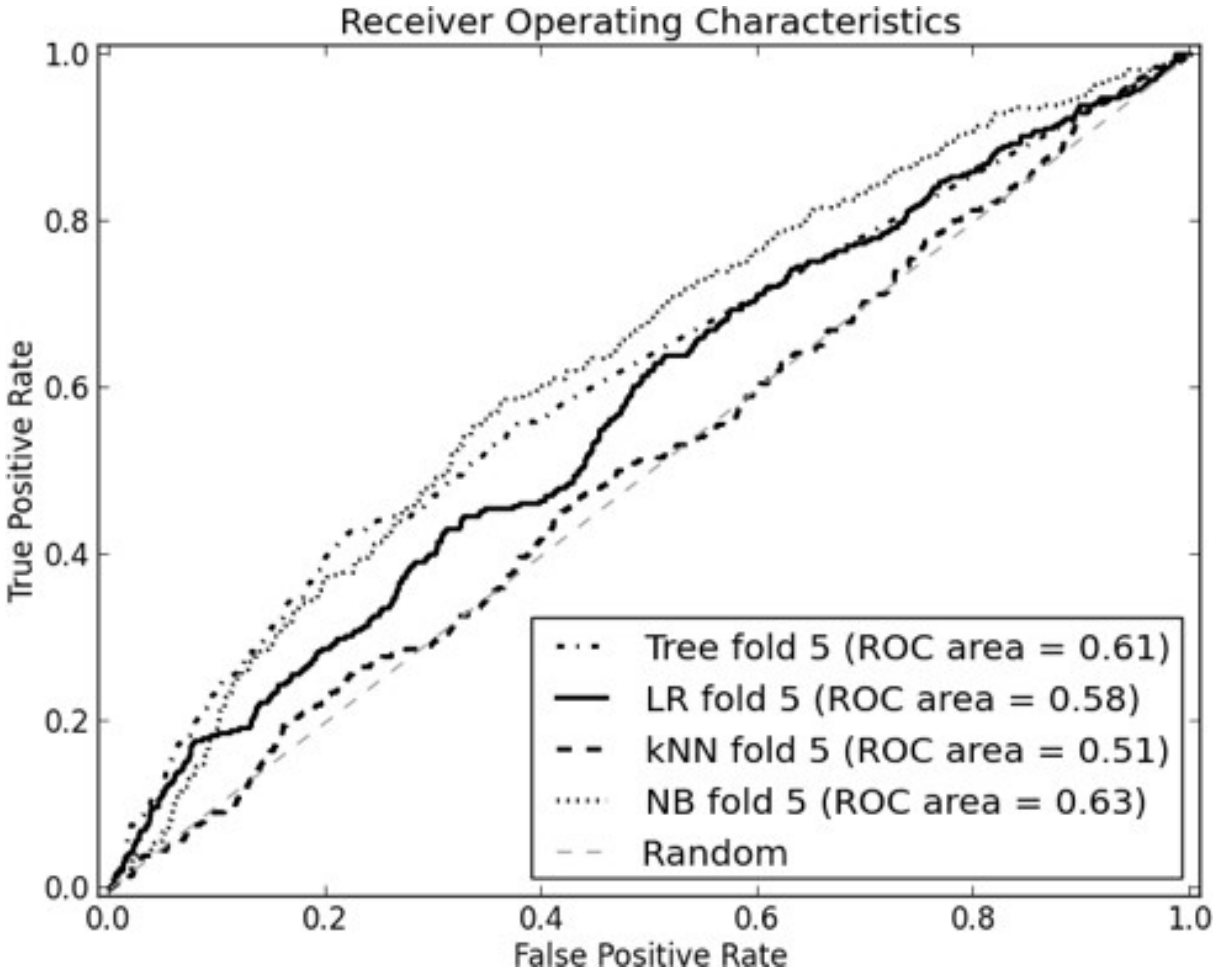
Naïve Bayes confusion matrix:

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

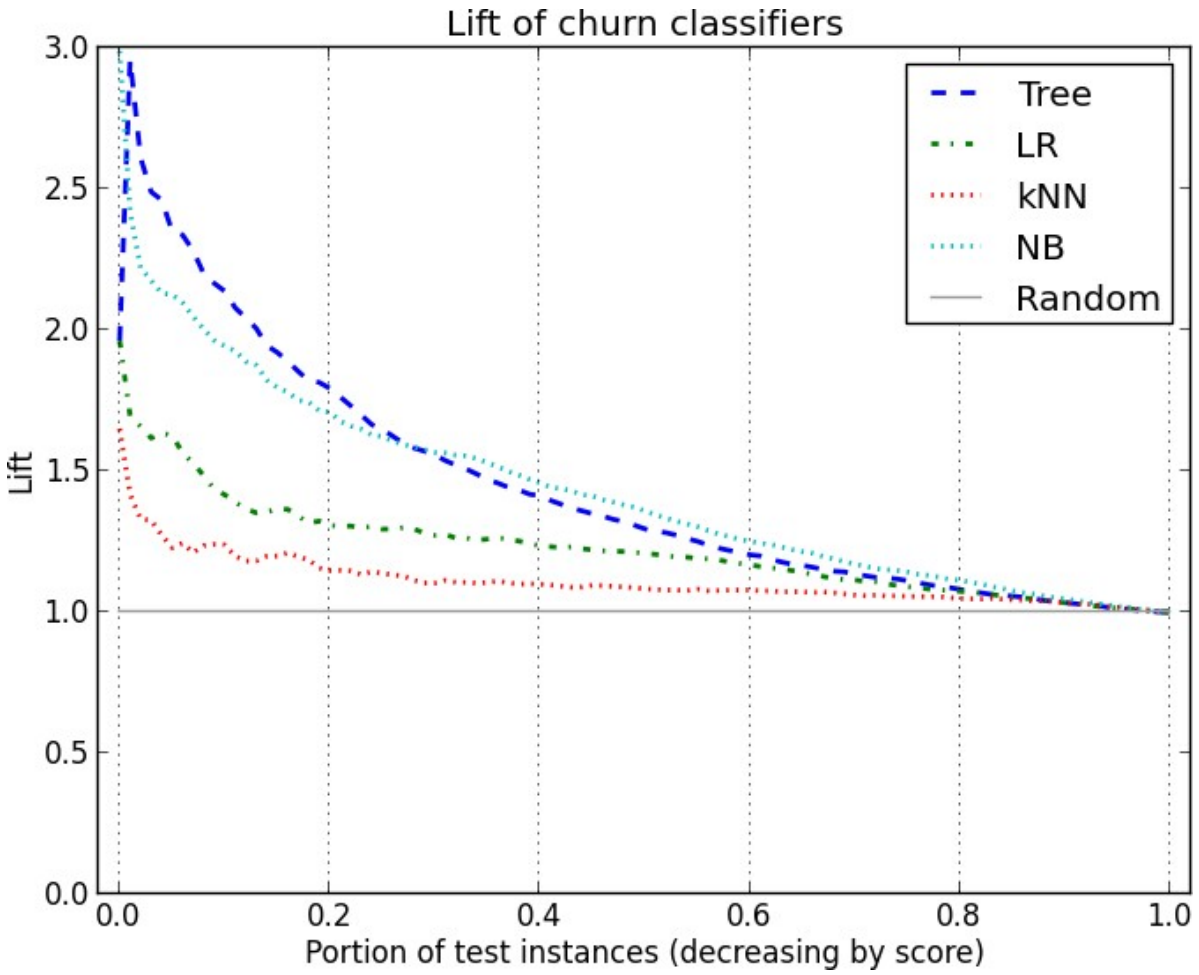
k-Nearest Neighbors confusion matrix:

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

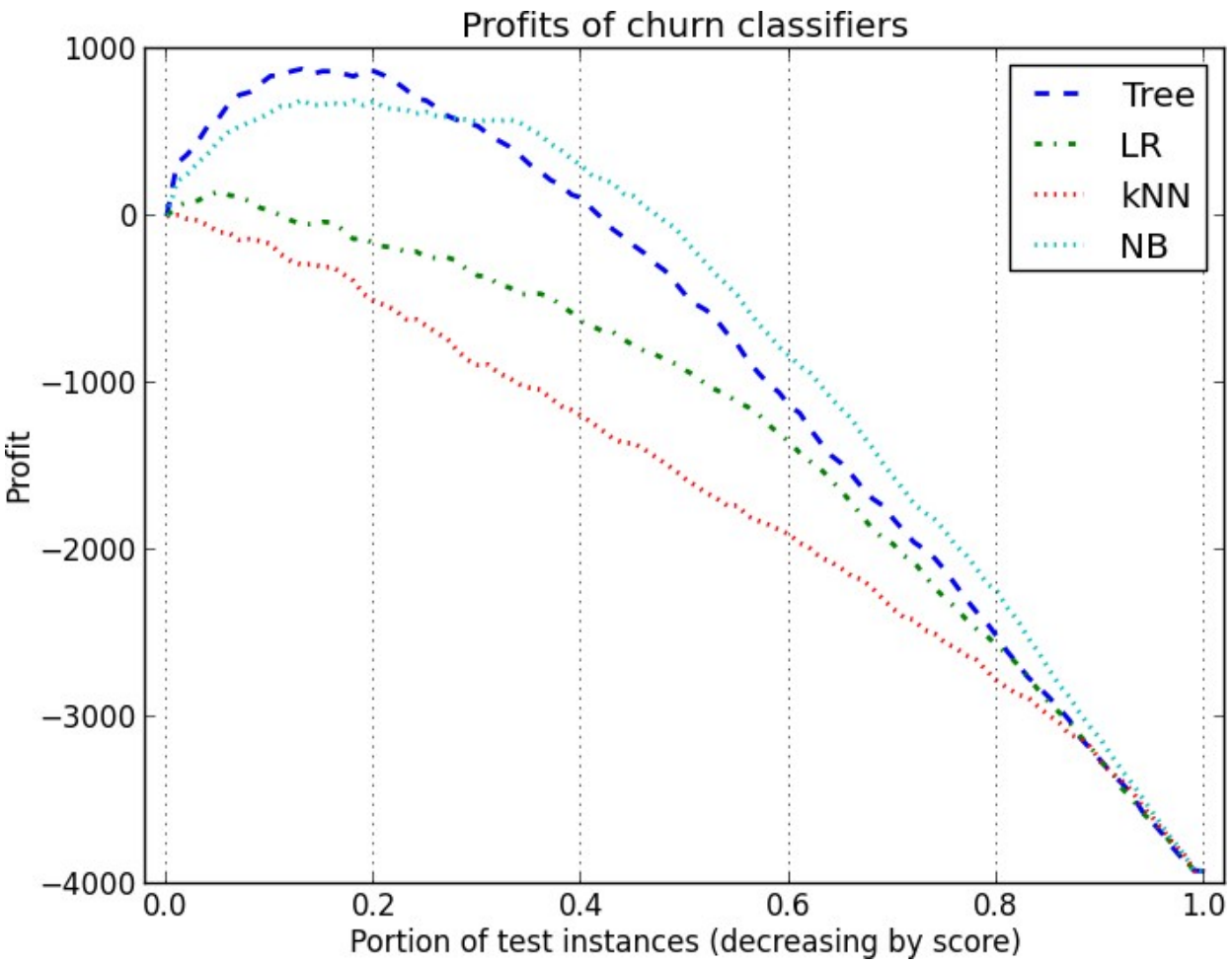
ROC Curve


Receiver Operating Characteristics
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Tree fold 5 (ROC area = 0.61)
LR fold 5 (ROC area = 0.58)
kNN fold 5 (ROC area = 0.51)
NB fold 5 (ROC area = 0.63)
Random

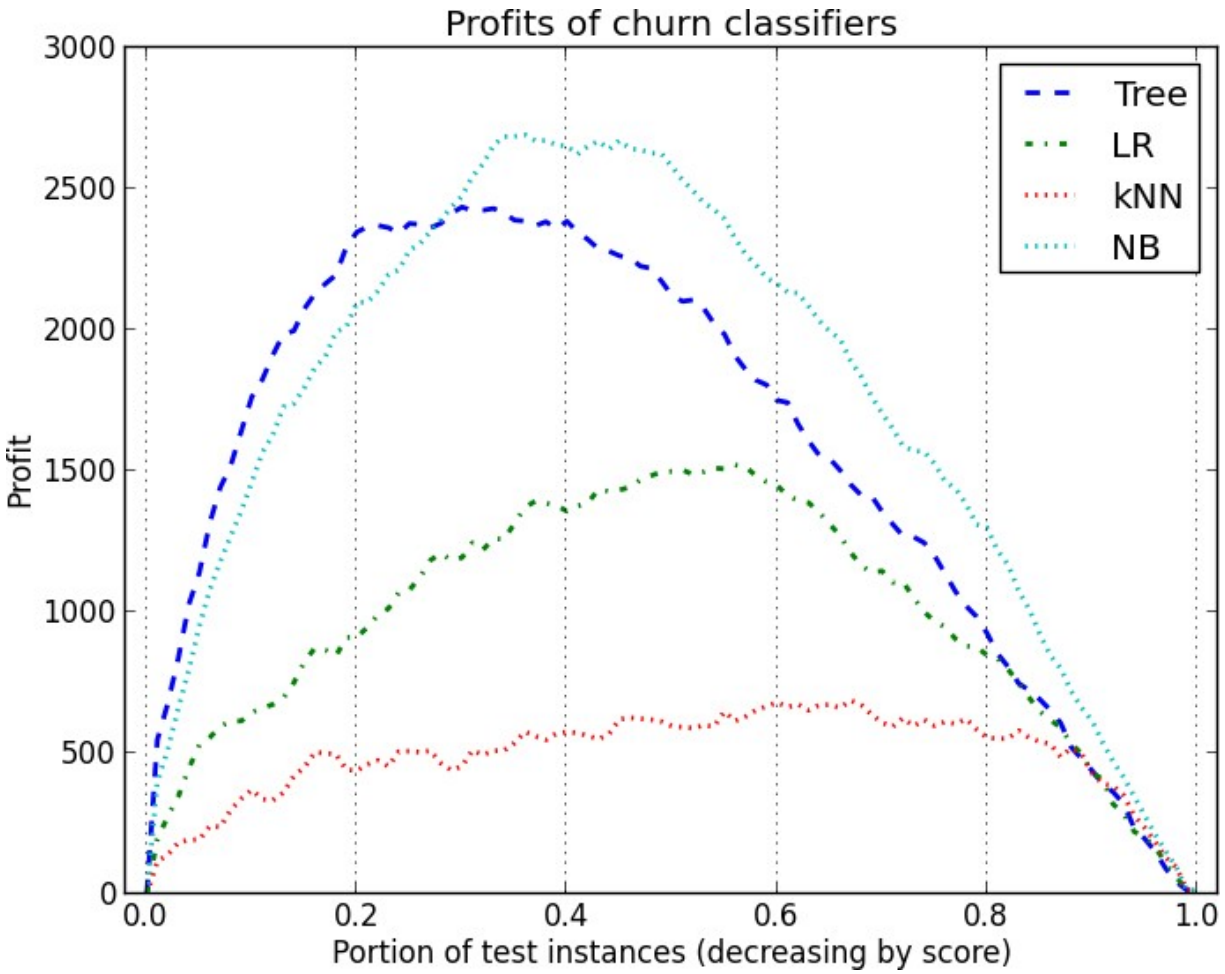
Lift Curve


Lift of churn classifiers
Tree
LR
kNN
NB
Random
Lift
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Portion of test instances (decreasing by score)

Profit Curves



Profit Curves



NATURAL LANGUAGE PROCESSING(NLP)
UNSTRUCTURED DATA

Scottie Pippen says it's 'hard' to call Buccaneers QB Tom Brady the NFL's best player

The Bulls legend cited Brady's inability to play other positions as why

By Wajih AlBaroudi Nov 19, 2021 at 7:59 pm ET • 1 min read

Hall of Famer Scottie Pippen has thrown shade at his longtime Chicago Bulls teammate Michael Jordan, widely considered the greatest basketball player of all time, since the 2020 debut of the Jordan-centric docuseries "The Last Dance." Now, Pippen is directing his criticism toward the player widely known as football's greatest of all time.

In an interview with GQ Sports, Pippen said it's "hard" to anoint Tampa Bay Buccaneers quarterback Tom Brady the NFL's best player ever because football is a team sport.

"It's hard to place Tom Brady at the top of the NFL," Pippen said. "Even though he's won a lot of championships. There are almost 70 players on each team, so is he playing on every side of the football? Is he on kickoff? Is he on the punt returns? Because if he ain't playing all them roles, then he got to give credit to his team."

While Brady won't be lining up as a gunner or kick returner any time soon, his impact on the Buccaneers and New England Patriots has been immeasurable.

Sports

More Sports

Florida fires coach Dan Mullen in fourth season as Gators implode one year after winning SEC East

CBSSports.com · 2 hours ago

- **Mullen and Stricklin need more discussions**
  247Sports · 12 hours ago · Opinion
- **Dan Mullen out as Florida football coach**
  Fox News · 1 hour ago
- **Gators should keep Mullen — unless they can make home-run hire like Bob Stoops | Commentary**
  Orlando Sentinel · 6 hours ago · Opinion
- **BREAKING: Florida fires Head Coach Dan Mullen | CBS Sports HQ**
  CBS Sports HQ · 1 hour ago

View Full Coverage

Business

More Business

No stranger to turmoil, Dutch dealmaker Wynaendts set for Deutsche chair

Reuters · 4 hours ago

More

The U.S. emergency oil stash is in the spotlight as gas price surge. What to know

NPR · 10 hours ago

Prosecutors in Elizabeth Holmes Trial Revealed Untruths, but Did They Prove Intent?

The Wall Street Journal · 7 hours ago

- **Elizabeth Holmes takes the stand in her criminal fraud trial**

Item 6: Performance-Based Fees and Side-By-Side Management

We receive compensation from a number of pooled investment vehicles that include fees or similar charges assessed on account performance, which is based on capital appreciation over certain periods. The portfolio managers of the pooled investment vehicles also advise other client accounts that are charged standard fees. As a result, conflicts of interest may arise because the portfolio managers may have an incentive to favor the pooled investment vehicles over other client accounts. We have in place policies and procedures designed to reduce the likelihood of such conflicts, which include monitoring accounts as appropriate, and, if deemed necessary, imposing trading restrictions on certain securities.

Item 7: Types of Clients

We provide investment advice to individuals, pension and profit sharing plans, trusts, estates and charitable organizations, corporations and other types of business entities, and institutional clients. We also provide investment advice to educational institutions, private investment partnerships and other entities. We generally require a minimum starting portfolio asset value of $200,000 but retain the discretion to set other minimums. A Portfolio Manager may have a higher minimum starting portfolio asset value for accounts under his or her management.

Item 8: Methods of Analysis, Investment Strategies and Risk of Loss

A substantial majority of the assets we manage are invested in equity securities. Equity securities include publicly and privately issued equity securities, common and preferred stocks, warrants, rights to subscribe to common stock and convertible securities, exchange-listed securities, over-the-counter securities, as well as instruments that attempt to track the price movement of equity indices. In making equity investments, our portfolio managers endeavor to use a risk-averse, value-oriented approach. We seek to identify companies with good businesses, proven profitability, strong balance sheets, a consistent record, conservative accounting, and managements that are devoted to increasing values for their shareholders.

University Health Network

LABORATORY MEDICINE PROGRAM

DEPARTMENT OF PATHOLOGY
200 Elizabeth Street
Toronto, Ontario, M5G 2C4
TEL: 416-340-3325
FAX: 416-586-9901

Surgical Pathology Consultation Report

* Addended *

Patient Name:	**Patient, USCAP**			Accession #:	S16-12345
MRN:	9876543	Service:	TGH Thoracic	Collected:	May-05-2016
DOB:	11/22/1947 (Age: 68)	Visit #:	23412312345	Received:	May-05-2016
Gender:	F	Location:	2C Pre Operative Care Unit	Reported:	Jun-01-2016
HCN:	123456775CH	Facility:	TGH/PMH		
Ordering MD:	Deep Cutter, MD				
Copy To:	Good P Friend, MD Stat Response, MD				

Specimen(s) Received

1. Lymph-Node: ST10R TB Angle
2. Right middle lobe
3. Station 11R
4. Station 4R
5. Station 7
6. Interlobar ST11
7. Right middle and upper bilobectomy

Consolidated Theranostic Report

Interpretation

Invasive moderately differentiated adenocarcinoma, acinar-predominant, pT2aN1
- POSITIVE for EGFR L858R mutation (see Molecular Diagnostics report)
- NEGATIVE for ALK by immunohistochemistry (performed using the 5A4 antibody with a protocol optimized for detection of ALK gene rearrangement)
- See Diagnosis, Comment, and Synoptic Report below for further details

Signed out by: Lung Path, MD
Date Reported: Jun-01-2016

Diagnosis

1,3-6. Lymph nodes (ST10R right tracheobronchial, ST11R right interlobar, ST4R right lower paratracheal, ST7 subcarinal, ST11 interlobar):
- At least one lymph node per station, negative for malignancy (x5) (0/5)

2. Lung, resection (right middle lobectomy):
   a. Invasive moderately differentiated adenocarcinoma, acinar-predominant, pT2aN1, with:
      i. Greatest tumor dimension: 1.2 cm (see Comment)
      ii. Visceral pleural and lympho-vascular invasion present
      iii. Stapled parenchymal resection margin positive for carcinoma (see Comment)
   b. One of five lymph nodes focally positive for adenocarcinoma by direct invasion (1/5) (see Comment)

NLP operations

- Named Entity Extraction
- Classification
- Sentiment
- Topic modeling
- Clustering

Pointers

- Theory
  - Introduction to Natural Language Processing: *Daniel Jurafsky*
    - Course Slides !
- Practice (Python)
  - Natural Language Processing Recipes: *Kulkarni & Shivananda*