METIS

Advanced Model Evaluation

INTRODUCTION TO DATA SCIENCE – FALL 2018

SESSION 8

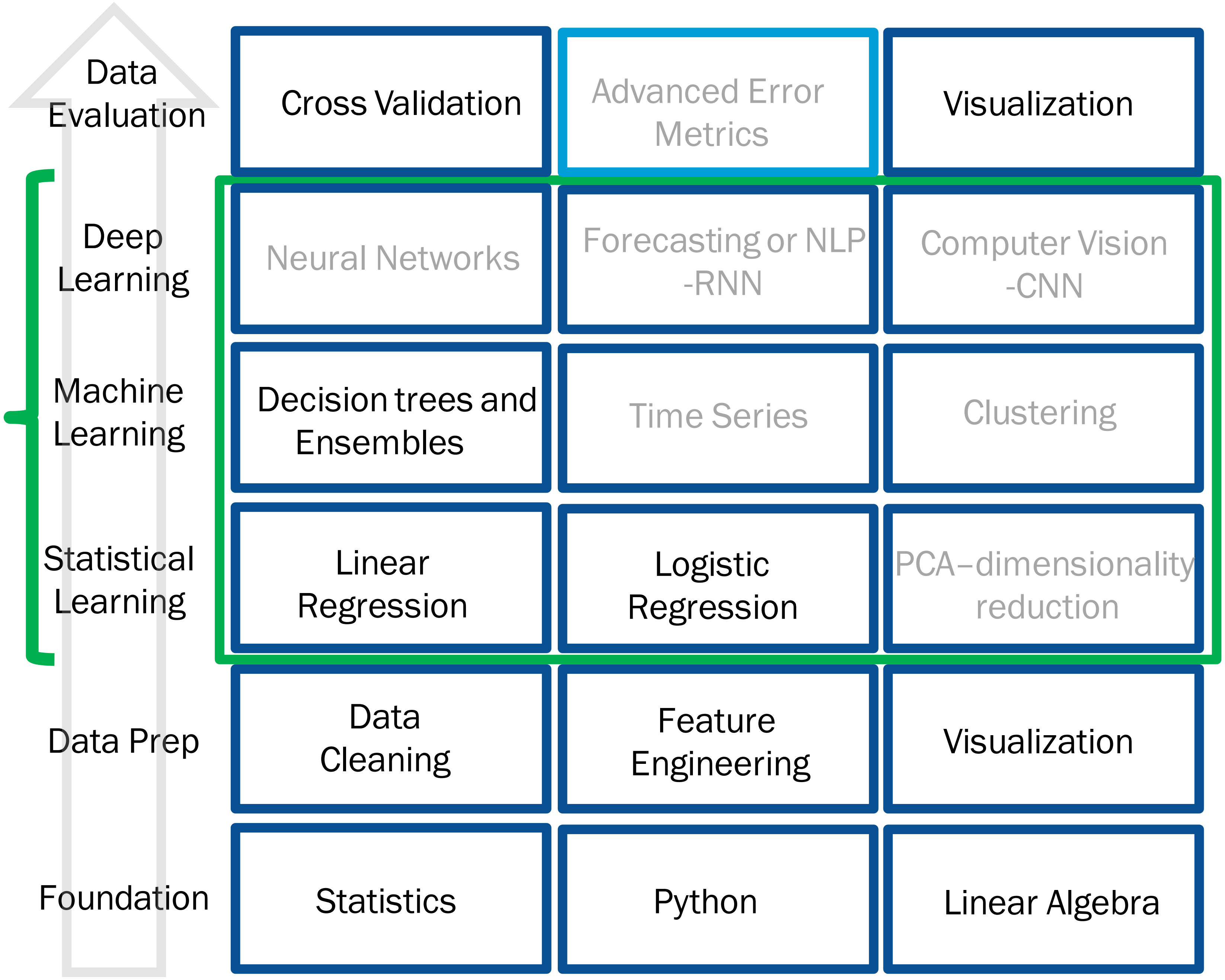
AGENDA

Session 8

1. Evaluating classifiers
2. ROC graph
3. AUC

Introduction to Data Science

- Learning the steps in the Data Science Process
- Learning multiple model methodologies



Data Evaluation	Cross Validation	Advanced Error Metrics	Visualization
Deep Learning	Neural Networks	Forecasting or NLP -RNN	Computer Vision -CNN
Machine Learning	Decision trees and Ensembles	Time Series	Clustering
Statistical Learning	Linear Regression	Logistic Regression	PCA–dimensionality reduction
Data Prep	Data Cleaning	Feature Engineering	Visualization
Foundation	Statistics	Python	Linear Algebra

Evaluating classifiers

INTRODUCING THE CONFUSION MATRIX

Machine learning classifiers

- Classification is the process of predicting the class of given data points
- Classes are sometimes called targets, labels or categories
- What have we learned?
  - K-nearest neighbors
  - Logistic regression
  - Decision Trees
  - Random Forest
  - Boosted Trees
  - Neural Networks

Binary classification

- There are 2 outcome classes
- Usually refer to as positive and negative
- Think of a classifier as sifting through a large population consisting of mostly negative, uninteresting cases, while looking for a small number of rare, positive instances
- Positive: one worthy of attention or alarm
- Negative: uninteresting or benign

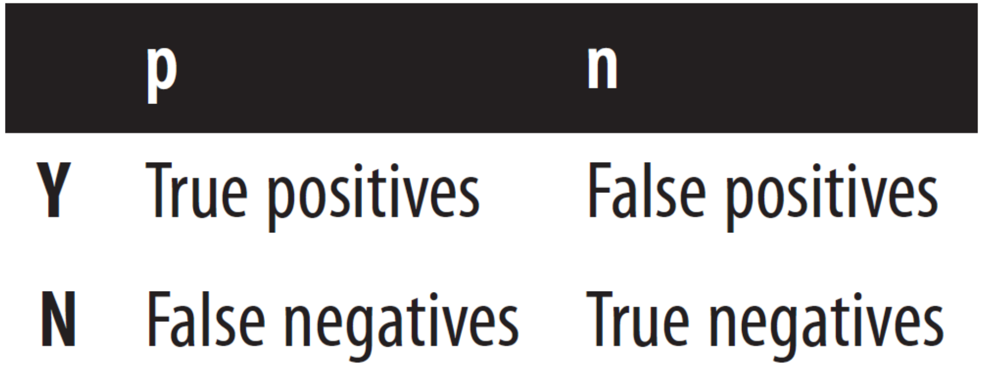
- Biological sample, we are testing for disease
  - Test comes back positive – disease is present
  - Test is negative – there is no cause for alarm

Confusion Matrix

- Columns are labeled with actuals
  - Positive or negative
- Rows are labeled with predictions
  - Yes or No
- Main diagonal are correct counts
- False positives are negative instances classified as Yes
- False negatives are positive instances classified as No

Actuals

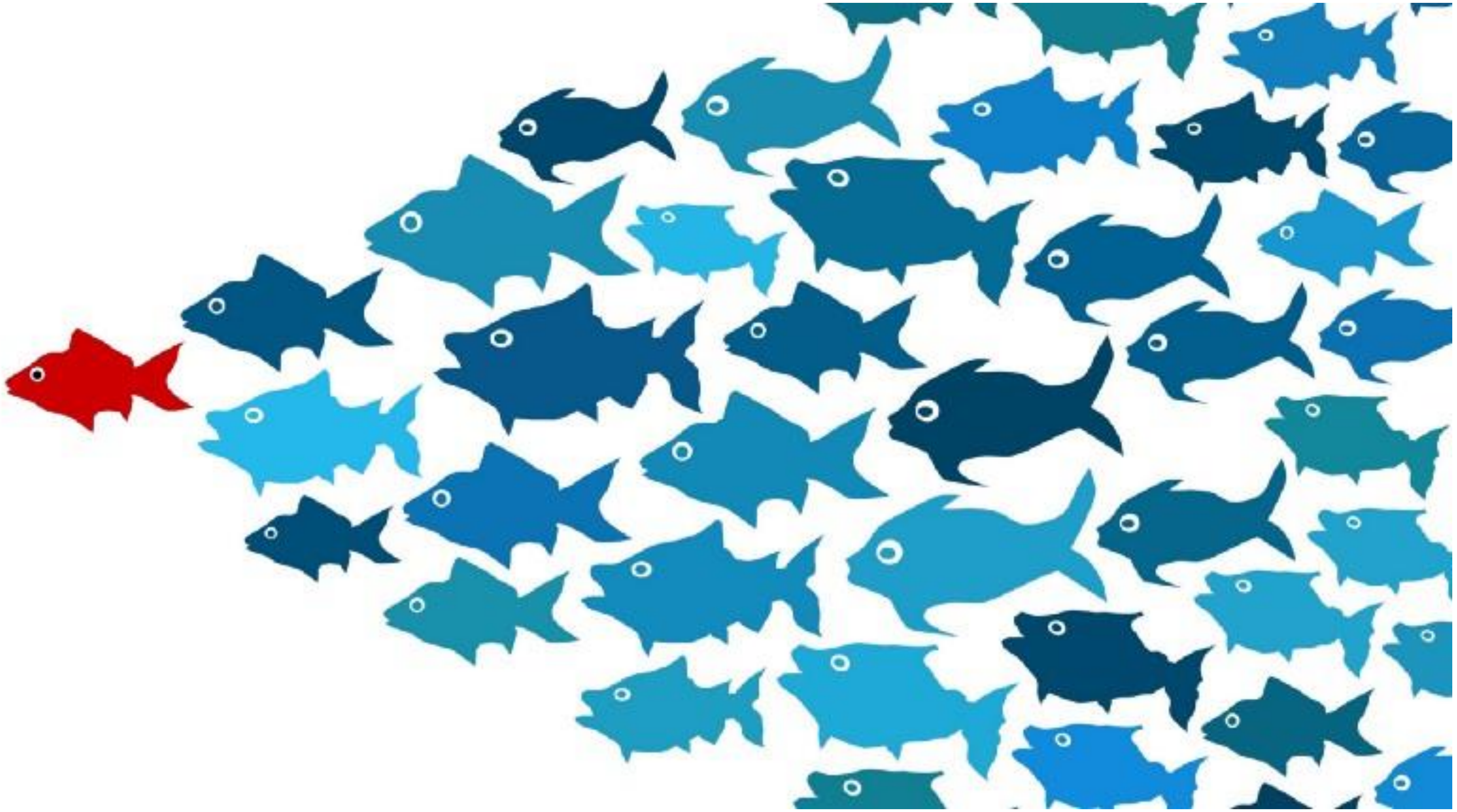
Predictions



	p	n
Y	True positives	False positives
N	False negatives	True negatives

Problems with unbalanced classes

- One class is rare
- Unbalanced or skewed
- Evaluation of accuracy breaks down

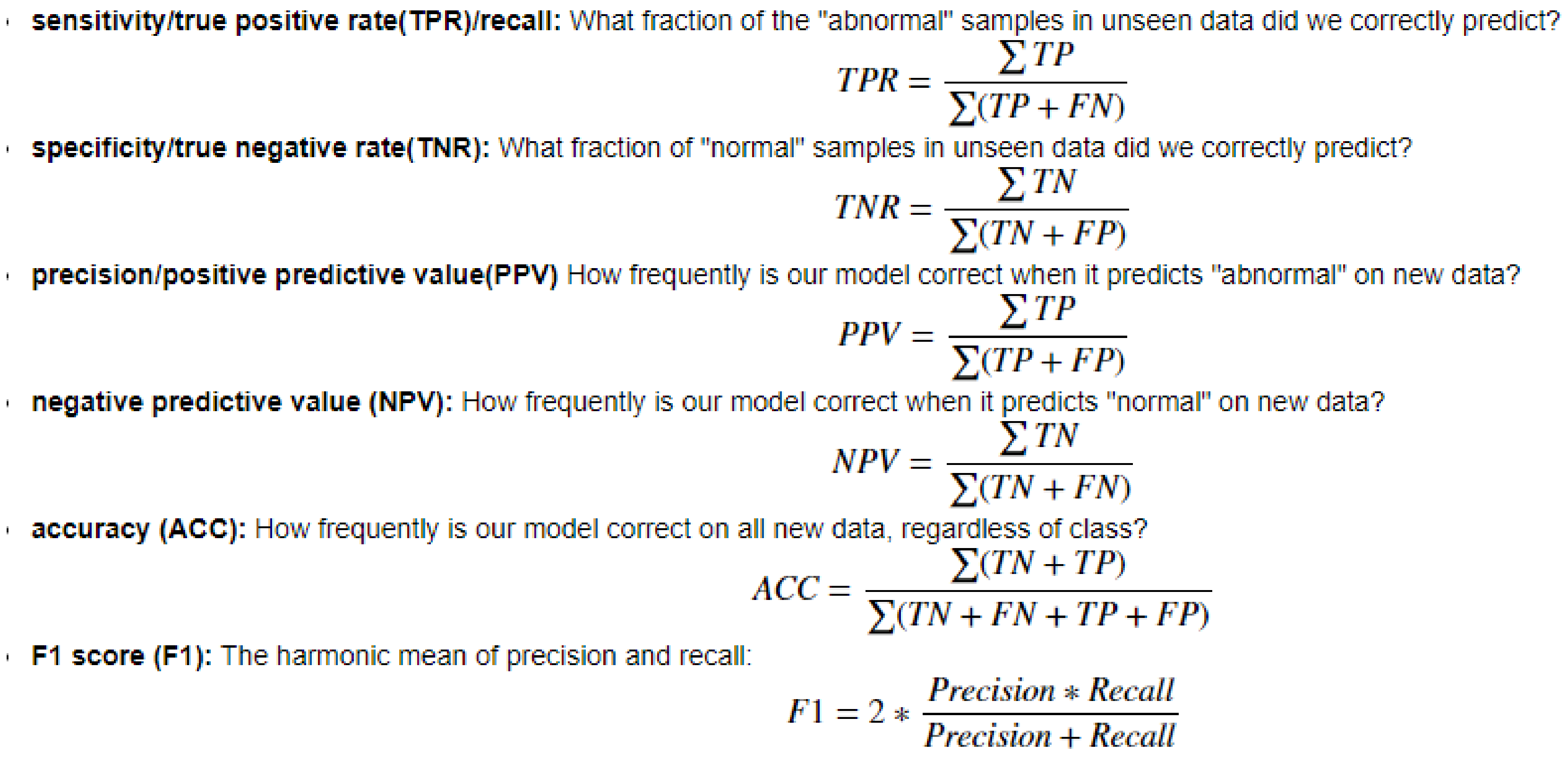


Predicting blue fish every time will give you a 99% accuracy

Metrics

- Different metrics based on the confusion matrix

- We will focus on two



- **sensitivity/true positive rate(TPR)/recall:** What fraction of the "abnormal" samples in unseen data did we correctly predict?

$$TPR = \frac{\sum TP}{\sum (TP + FN)}$$

- **specificity/true negative rate(TNR):** What fraction of "normal" samples in unseen data did we correctly predict?

$$TNR = \frac{\sum TN}{\sum (TN + FP)}$$

- **precision/positive predictive value(PPV)** How frequently is our model correct when it predicts "abnormal" on new data?

$$PPV = \frac{\sum TP}{\sum (TP + FP)}$$

- **negative predictive value (NPV):** How frequently is our model correct when it predicts "normal" on new data?

$$NPV = \frac{\sum TN}{\sum (TN + FN)}$$

- **accuracy (ACC):** How frequently is our model correct on all new data, regardless of class?

$$ACC = \frac{\sum (TN + TP)}{\sum (TN + FN + TP + FP)}$$

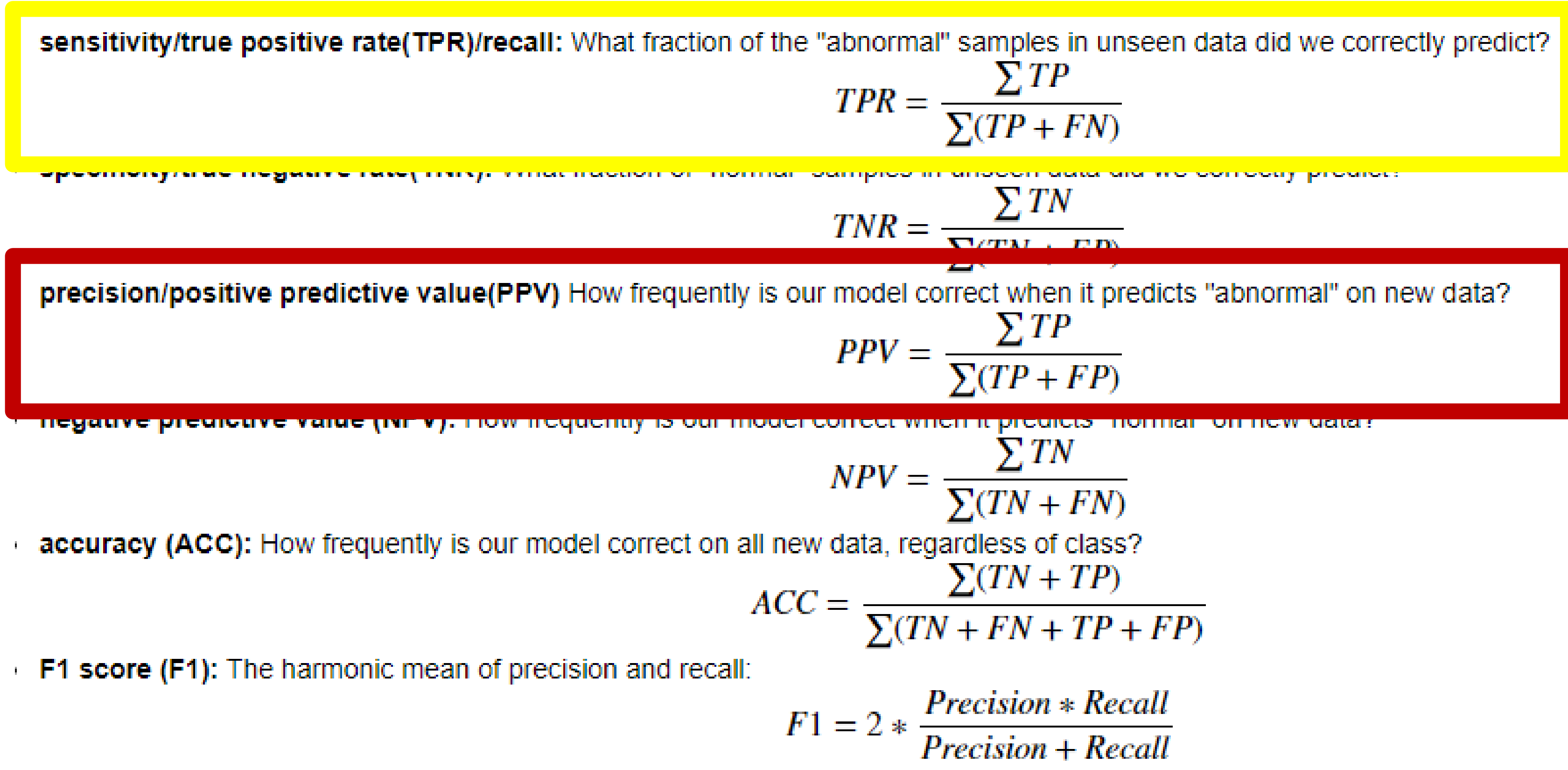
- **F1 score (F1):** The harmonic mean of precision and recall:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Metrics

- Different metrics based on the confusion matrix

- We will focus on two



- **sensitivity/true positive rate(TPR)/recall:** What fraction of the "abnormal" samples in unseen data did we correctly predict?

$$TPR = \frac{\sum TP}{\sum(TP + FN)}$$

- **specificity/true negative rate(TNR):** What fraction of "normal" samples in unseen data did we correctly predict?

$$TNR = \frac{\sum TN}{\sum(TN + FP)}$$

- **precision/positive predictive value(PPV)** How frequently is our model correct when it predicts "abnormal" on new data?

$$PPV = \frac{\sum TP}{\sum(TP + FP)}$$

- **negative predictive value (NPV):** How frequently is our model correct when it predicts "normal" on new data?

$$NPV = \frac{\sum TN}{\sum(TN + FN)}$$

- **accuracy (ACC):** How frequently is our model correct on all new data, regardless of class?

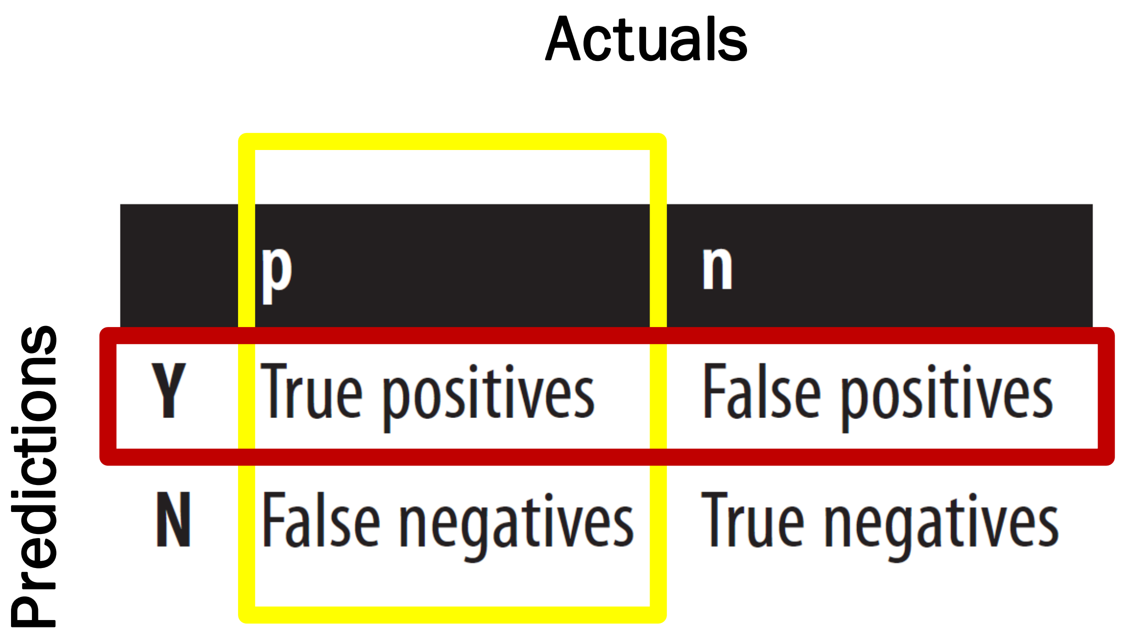
$$ACC = \frac{\sum(TN + TP)}{\sum(TN + FN + TP + FP)}$$

- **F1 score (F1):** The harmonic mean of precision and recall:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Metrics

- Recall: percent of actual positive instances that were classified as such
- Precision: percent of Yes predictions that are truly positive



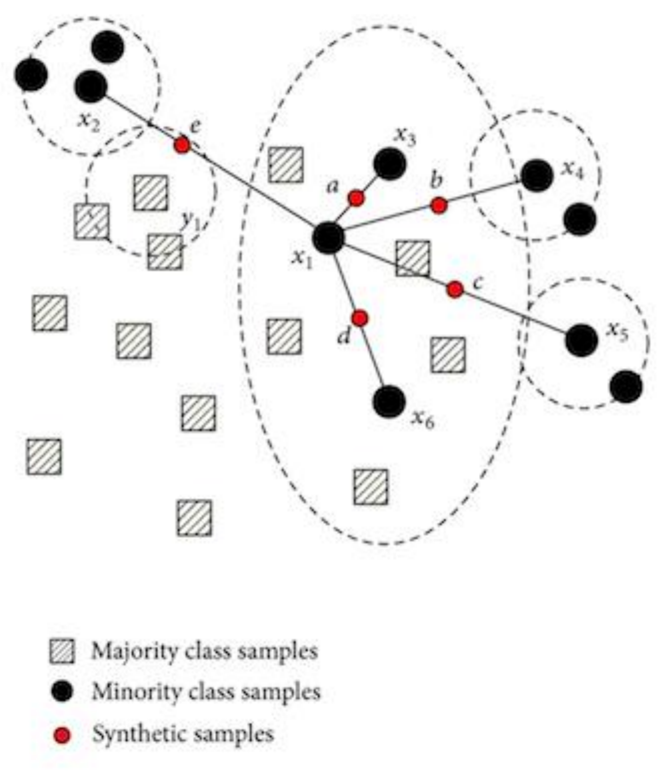
Actuals

Predictions

	p	n
Y	True positives	False positives
N	False negatives	True negatives

Minority class problem, we are concerned more with recall
In medical detection, it is usually more costly to miss a positive instance, than to falsely label a negative instance

Methods to overcome imbalance

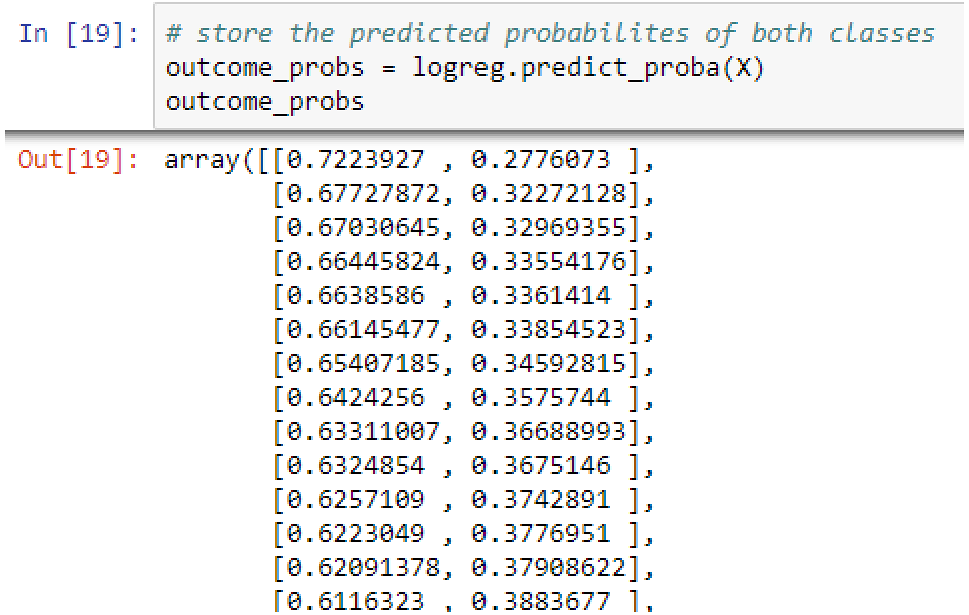


Visualization of SMOTE

- Cost- sensitive Learning
  - Create a function that specifies the cost of misclassifying the minority class more heavily than the majority class
- Sampling
  - Oversample the minority class
  - Undersample the majority class
- SMOTE – Synthetic Minority Over-sampling Technique
  - Create new instances of minority class by forming combinations of neighboring instances

Visualizing model evaluation

- Remember in session 5, Logistic regression
- We are able to predict class probability for the binary classification
- Take the probability of belonging to one of the classes and rank the observations

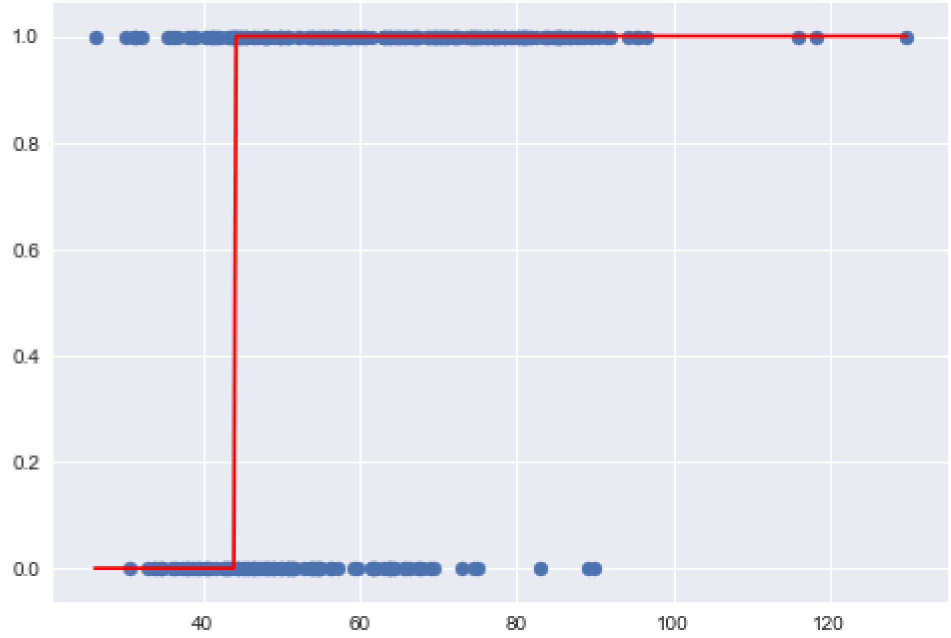


```
In [19]: # store the predicted probabilites of both classes
         outcome_probs = logreg.predict_proba(X)
         outcome_probs

Out[19]: array([[0.7223927 , 0.2776073 ],
                [0.67727872, 0.32272128],
                [0.67030645, 0.32969355],
                [0.66445824, 0.33554176],
                [0.6638586 , 0.3361414 ],
                [0.66145477, 0.33854523],
                [0.65407185, 0.34592815],
                [0.6424256 , 0.3575744 ],
                [0.63311007, 0.36688993],
                [0.6324854 , 0.3675146 ],
                [0.6257109 , 0.3742891 ],
                [0.6223049 , 0.3776951 ],
                [0.62091378, 0.37908622],
                [0.6116323 , 0.3883677 ],
```

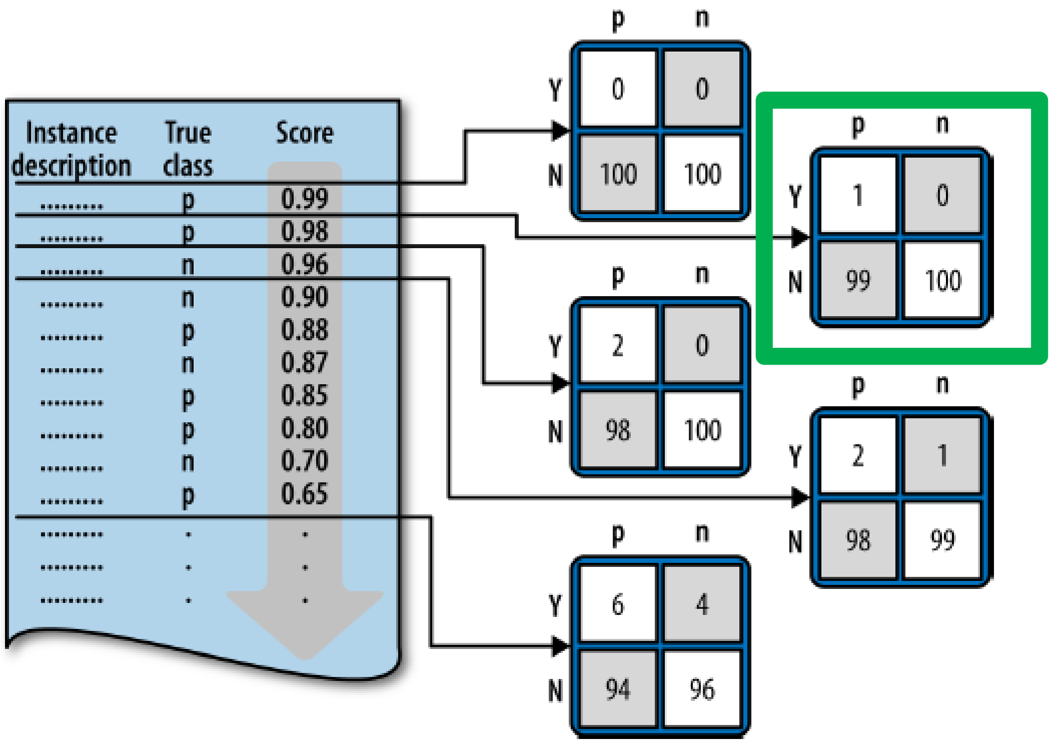
Visualizing model evaluation

- Determine a threshold and label predictions as 0 or 1



Visualizing model evaluation

- A classifier plus a threshold produces a single confusion matrix



Instance description	True class	Score
....	p	0.99
....	p	0.98
....	n	0.96
....	n	0.90
....	p	0.88
....	n	0.87
....	p	0.85
....	p	0.80
....	n	0.70
....	p	0.65
....	.	.
....	.	.
....	.	.

	p	n
Y	0	0
N	100	100

	p	n
Y	1	0
N	99	100

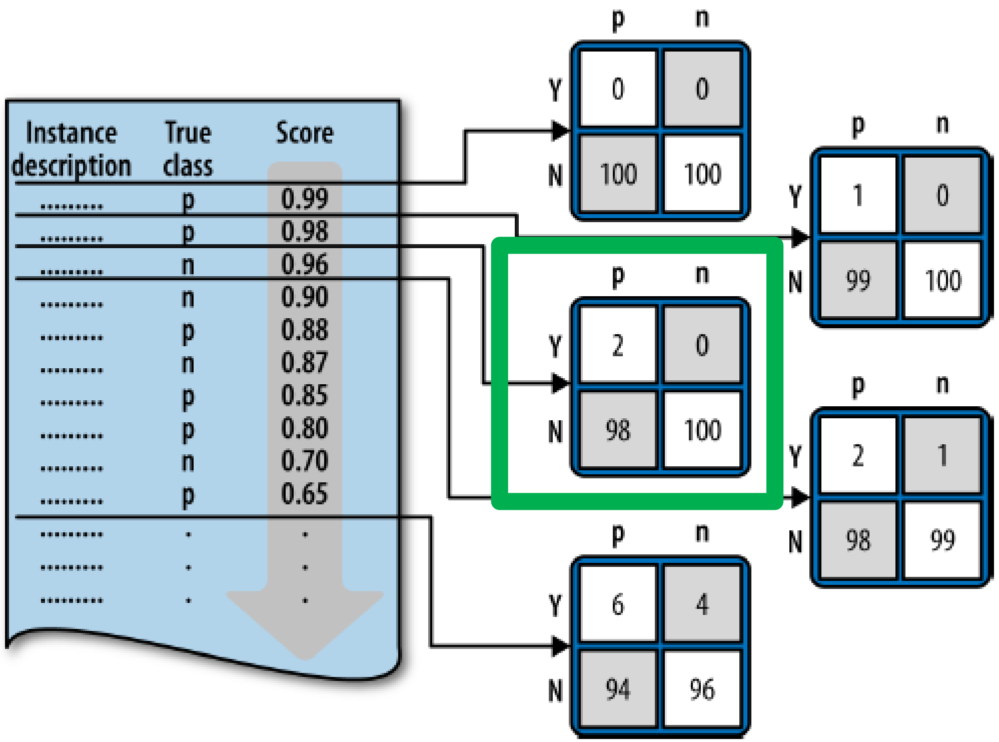
	p	n
Y	2	0
N	98	100

	p	n
Y	2	1
N	98	99

	p	n
Y	6	4
N	94	96

Visualizing model evaluation

- Whenever the threshold changes, the confusion matrix may change as well
- The number of true positives and false positives change



Instance description	True class	Score
....	p	0.99
....	p	0.98
....	n	0.96
....	n	0.90
....	p	0.88
....	n	0.87
....	p	0.85
....	p	0.80
....	n	0.70
....	p	0.65
....	.	.
....	.	.
....	.	.

	p	n
Y	0	0
N	100	100

	p	n
Y	2	0
N	98	100

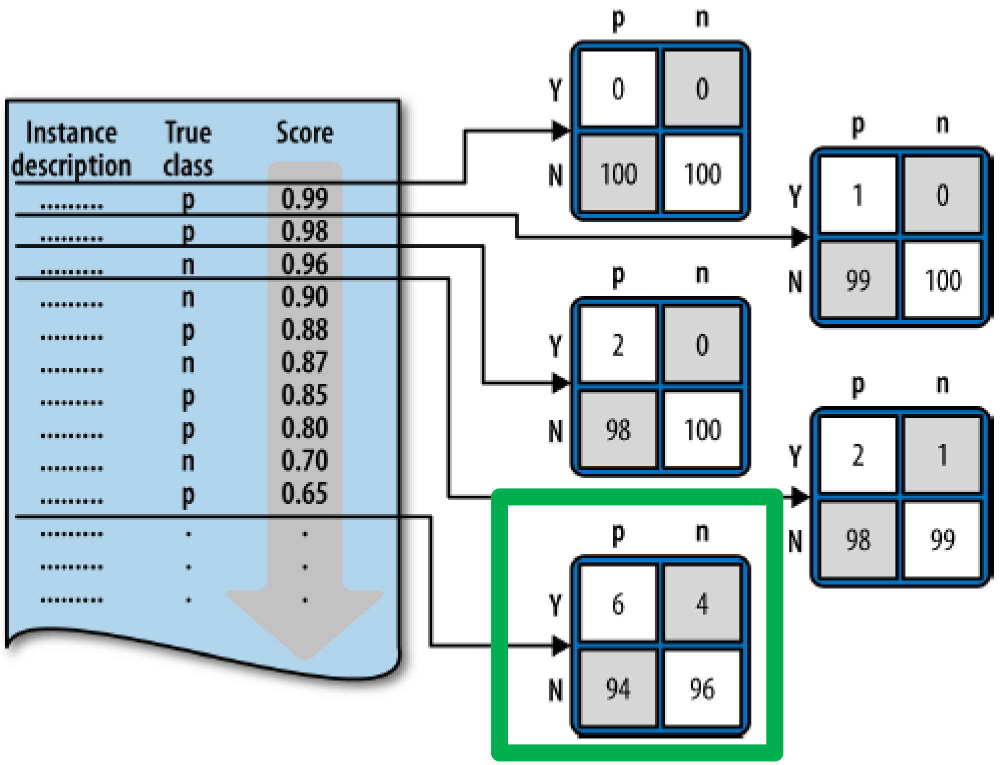
	p	n
Y	6	4
N	94	96

	p	n
Y	1	0
N	99	100

	p	n
Y	2	1
N	98	99

Visualizing model evaluation

- As the threshold is lowered, instances move up from the N row into the Y row of the confusion matrix
  - An instance that was considered negative is now classified as positive, so the counts change
- It can now become a true positive (Y,p) or a false positive (Y,n)



Instance description	True class	Score
....	p	0.99
....	p	0.98
....	n	0.96
....	n	0.90
....	p	0.88
....	n	0.87
....	p	0.85
....	p	0.80
....	n	0.70
....	p	0.65
....	.	.
....	.	.
....	.	.

	p	n
Y	0	0
N	100	100

	p	n
Y	1	0
N	99	100

	p	n
Y	2	0
N	98	100

	p	n
Y	2	1
N	98	99

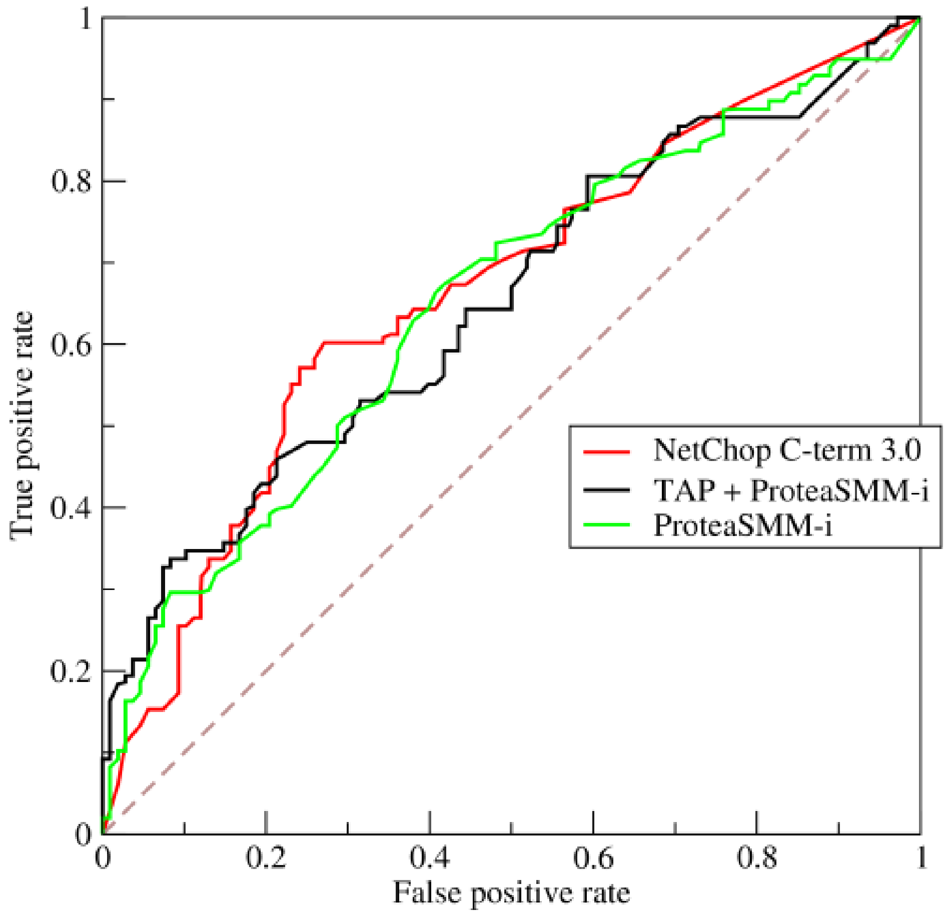
	p	n
Y	6	4
N	94	96

ROC graphs

TRUE POSITIVE RATE VS FALSE POSITIVE RATE

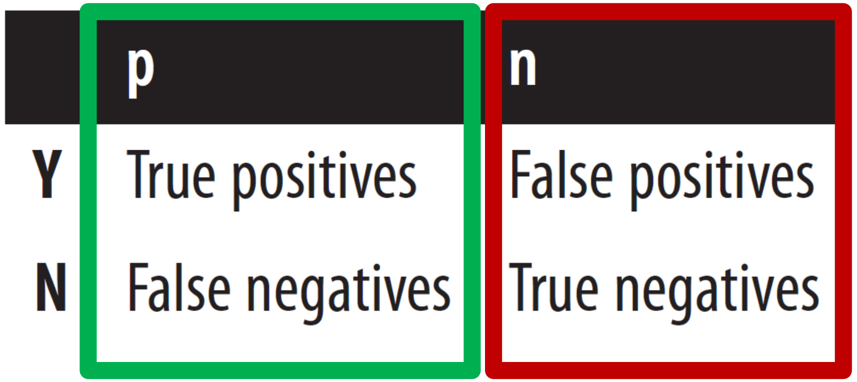
ROC graph

- Used to show the entire space of performance possibilities
- 2 dimensional plot of a classifier
- X axis: False positive rate
- Y axis: True positive rate
- ROC depicts relative tradeoffs that a classifier makes

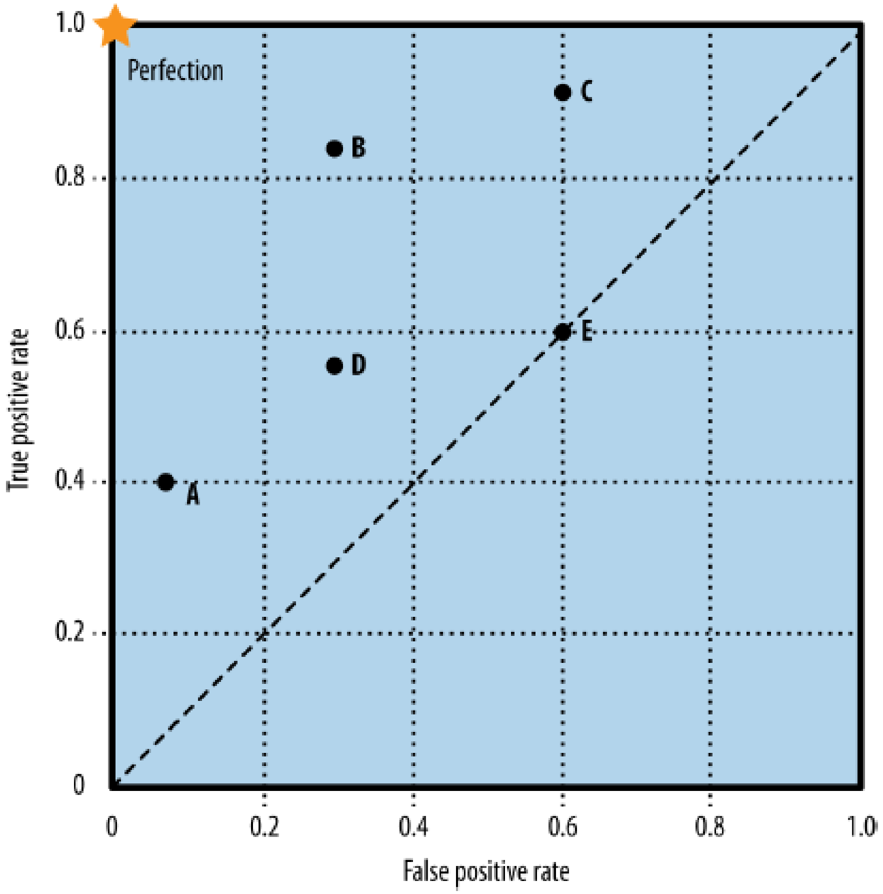


TPR vs FPR

- Five classifiers A-E with their performance shown
- Each confusion matrix gives us the 2 rates for each point
- TPR : actual positive examples
- FPR: actual negative examples
  - Also is seen as (1- specificity)

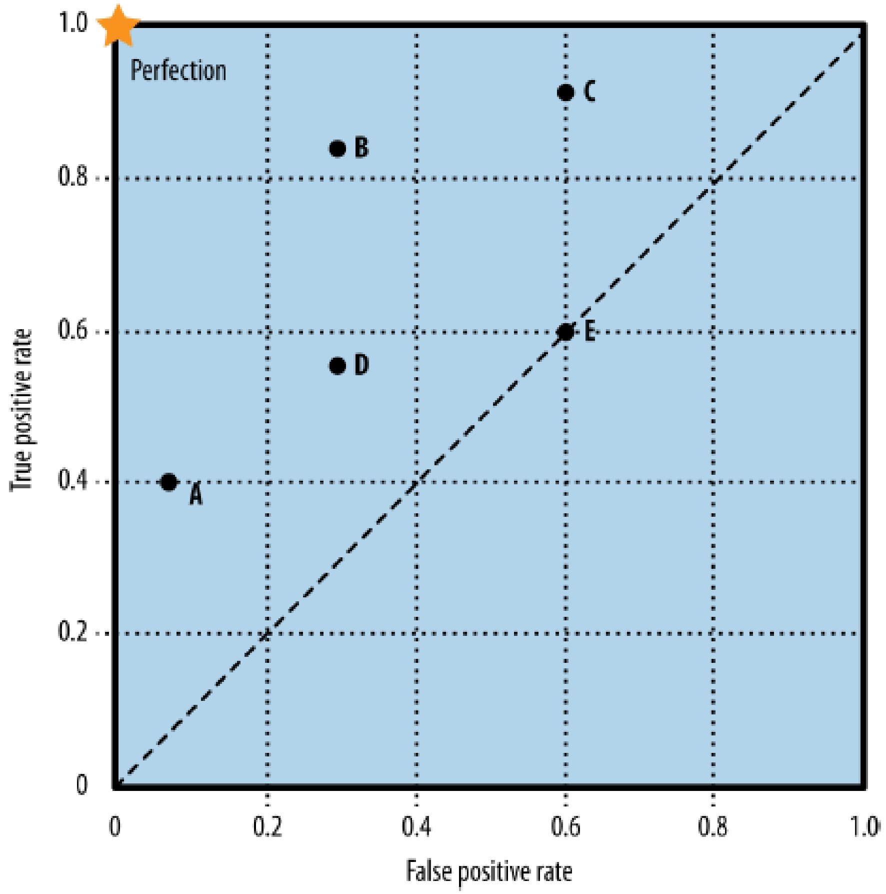


	p	n
Y	True positives	False positives
N	False negatives	True negatives



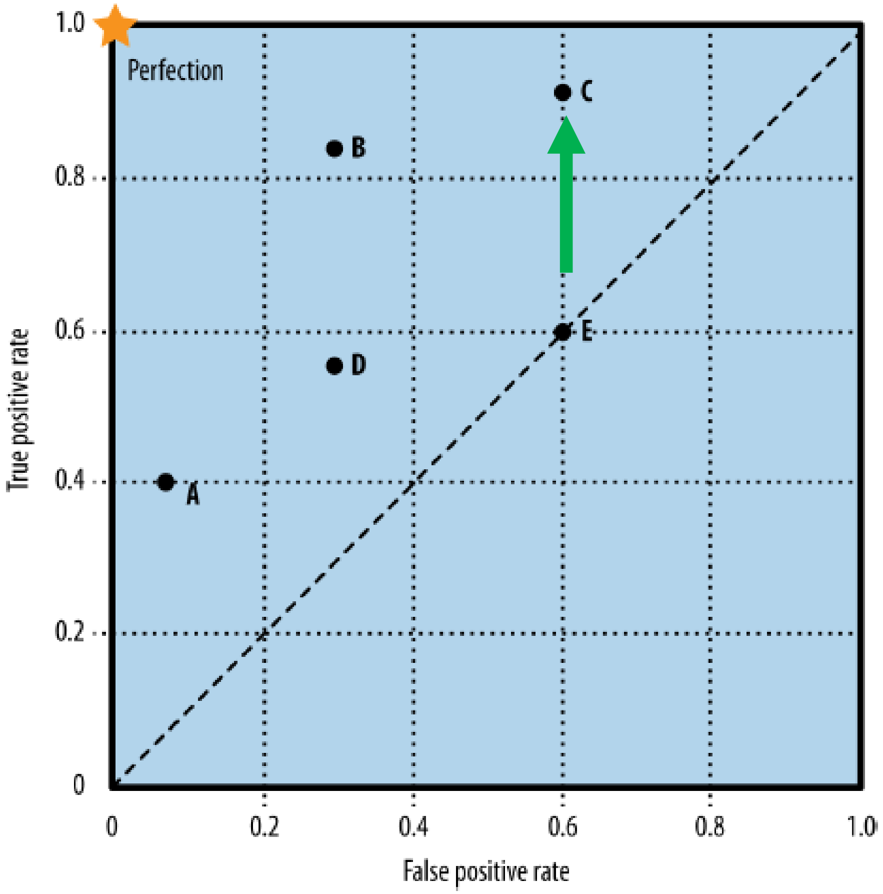
TPR vs FPR

- (0,0) – never issuing a positive classification
  - Commits no False positives, but also gains no True positives
- (1,1) – unconditionally issuing positive classifications
- (0,1) – perfect classification – Gold star
- (0.5, 0.5) - random guessing
- In order to move away from the diagonal, the classifier must exploit some info in the data



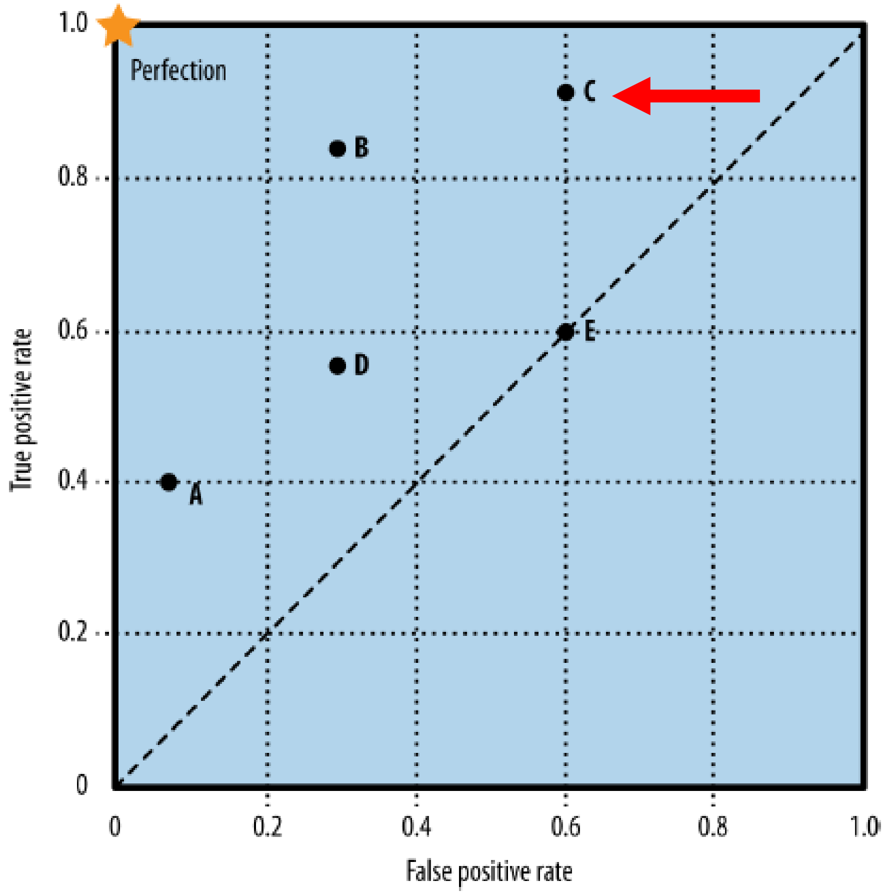
What do the sections mean?

- Upper Left (C)
  - TPR is higher, FPR is no worse
  - FPR is lower and TPR is no worse



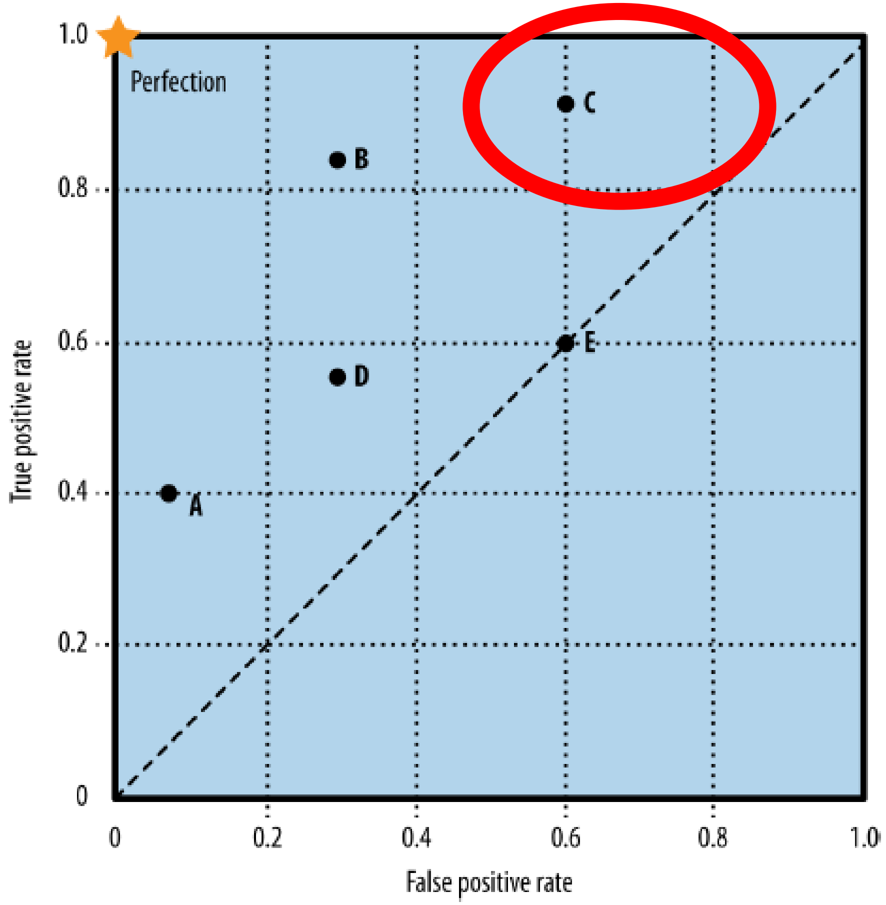
What do the sections mean?

- Upper Left (C)
  - TPR is higher, FPR is no worse
  - FPR is lower and TPR is no worse



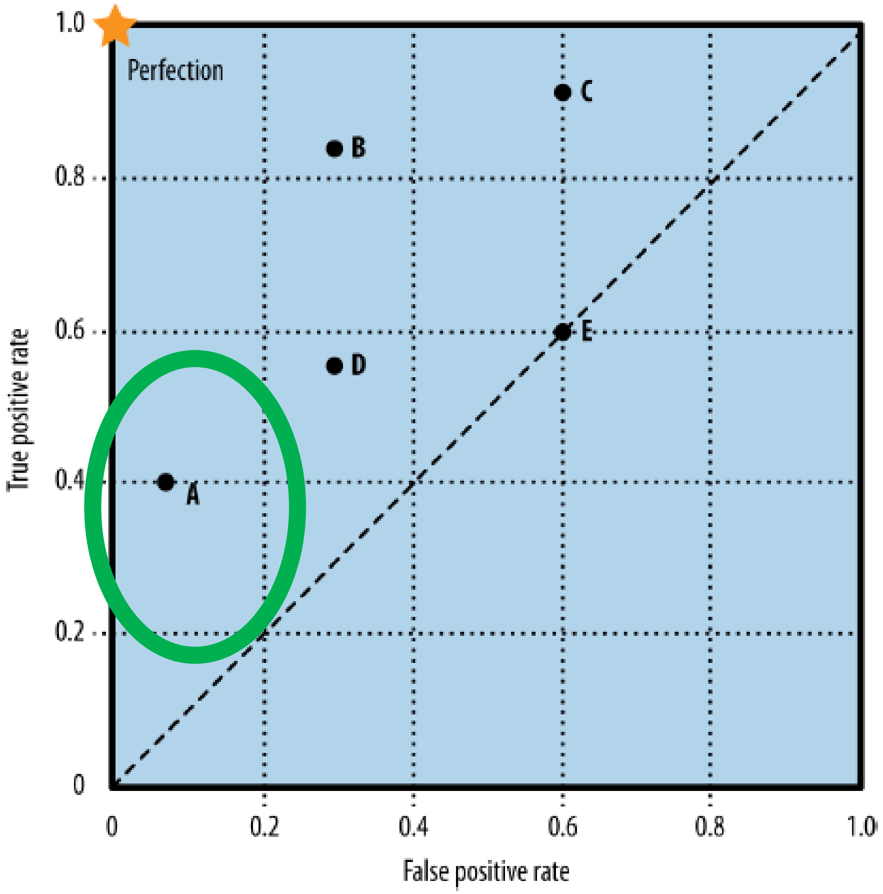
What do the sections mean?

- Considered permissive classifiers
- They make positive classifications with weak evidence
- Classify nearly all positives correctly
- High FPR



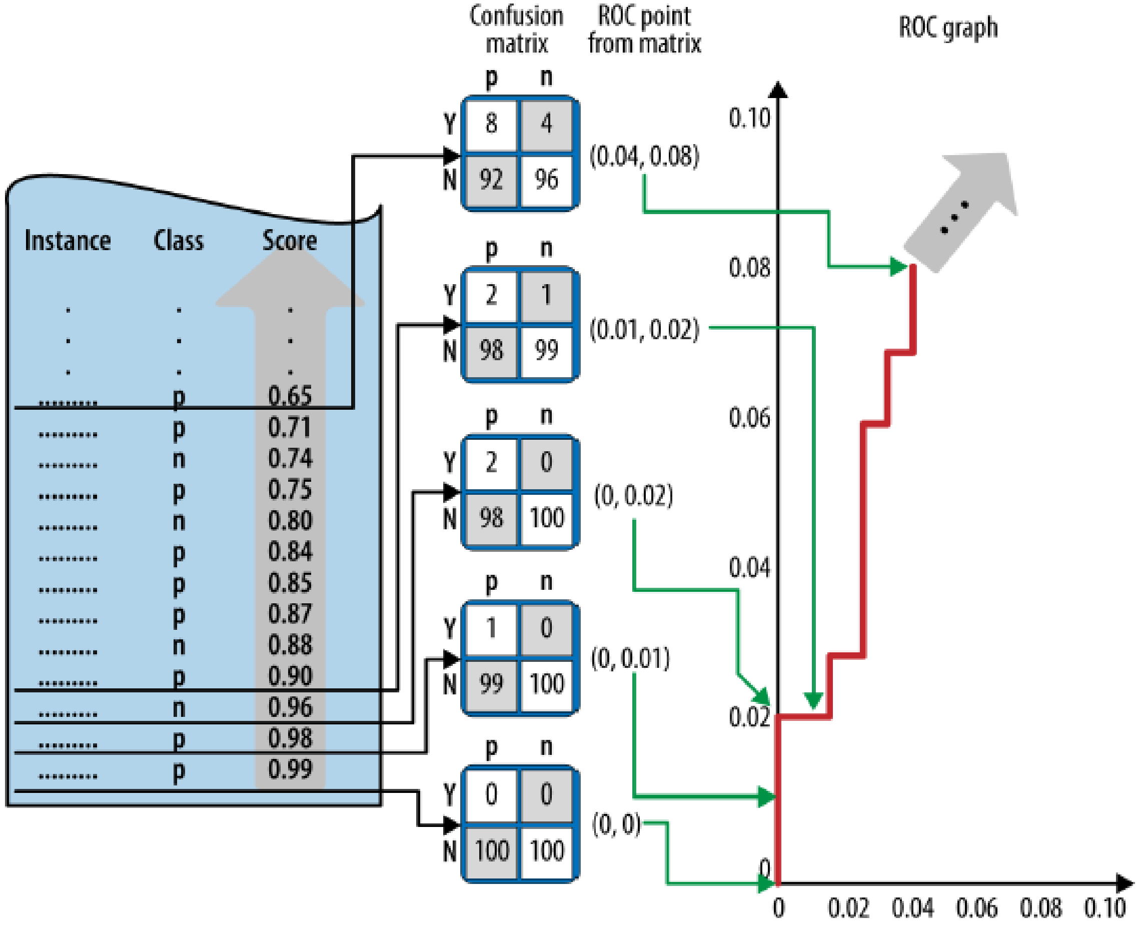
What do the sections mean?

- Considered conservative classifiers
- Raise alarms only with strong evidence
- Make few false positives
- Have low TPR as well
- Imagine if there are many negative examples
  - Then even a moderate false alarm



ROC is a stepwise graph

- The model assigns a score to each instance
- Instances are ordered decreasing from bottom to top
- Start at the bottom where initial confusion matrix where everything is classified as N
- Moving upward every instance moves a count of 1 from the N row to the Y row, resulting in a new confusion matrix.
- Each confusion matrix maps to a FPR, TPR pair in the ROC space



Instance	Class	Score
.	.	.
.	.	.
.	.	.
....	p	0.65
....	p	0.71
....	n	0.74
....	p	0.75
....	n	0.80
....	p	0.84
....	p	0.85
....	p	0.87
....	n	0.88
....	p	0.90
....	n	0.96
....	p	0.98
....	p	0.99

Confusion matrix / ROC point from matrix

	p	n
Y	8	4
N	92	96

(0.04, 0.08)

	p	n
Y	2	1
N	98	99

(0.01, 0.02)

	p	n
Y	2	0
N	98	100

(0, 0.02)

	p	n
Y	1	0
N	99	100

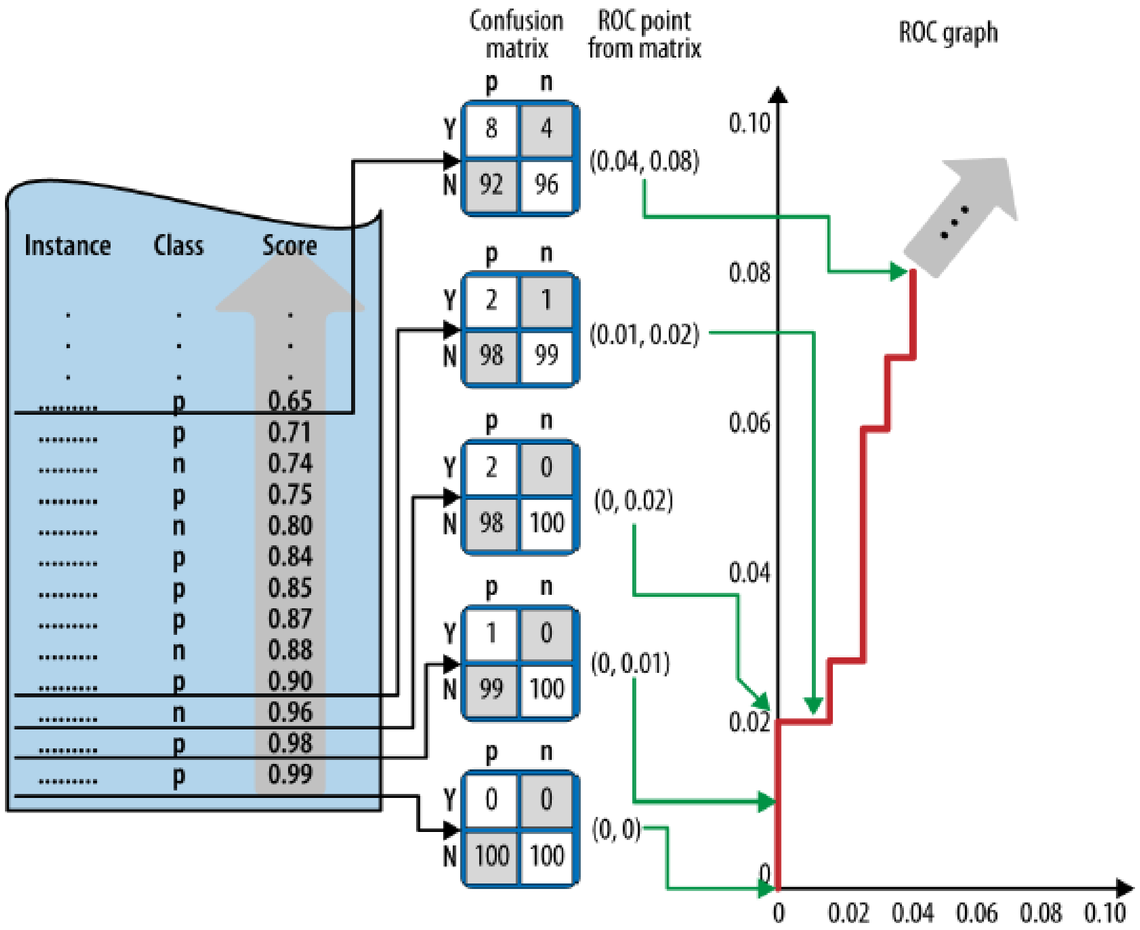
(0, 0.01)

	p	n
Y	0	0
N	100	100

(0, 0)

ROC is a stepwise graph

- Whenever we pass a positive instance, we take a step upward, increasing TPR
- Whenever we pass a negative instance, we take a step rightward, increasing FPR
- This curve is a step function for a single test set, but it appears smooth
- ROC graph is independent of the class proportions as well as the costs and benefits



Instance	Class	Score
.	.	.
.	.	.
.	.	.
....	p	0.65
....	p	0.71
....	n	0.74
....	p	0.75
....	n	0.80
....	p	0.84
....	p	0.85
....	p	0.87
....	n	0.88
....	p	0.90
....	n	0.96
....	p	0.98
....	p	0.99

Confusion matrix / ROC point from matrix

	p	n
Y	8	4
N	92	96

(0.04, 0.08)

	p	n
Y	2	1
N	98	99

(0.01, 0.02)

	p	n
Y	2	0
N	98	100

(0, 0.02)

	p	n
Y	1	0
N	99	100

(0, 0.01)

	p	n
Y	0	0
N	100	100

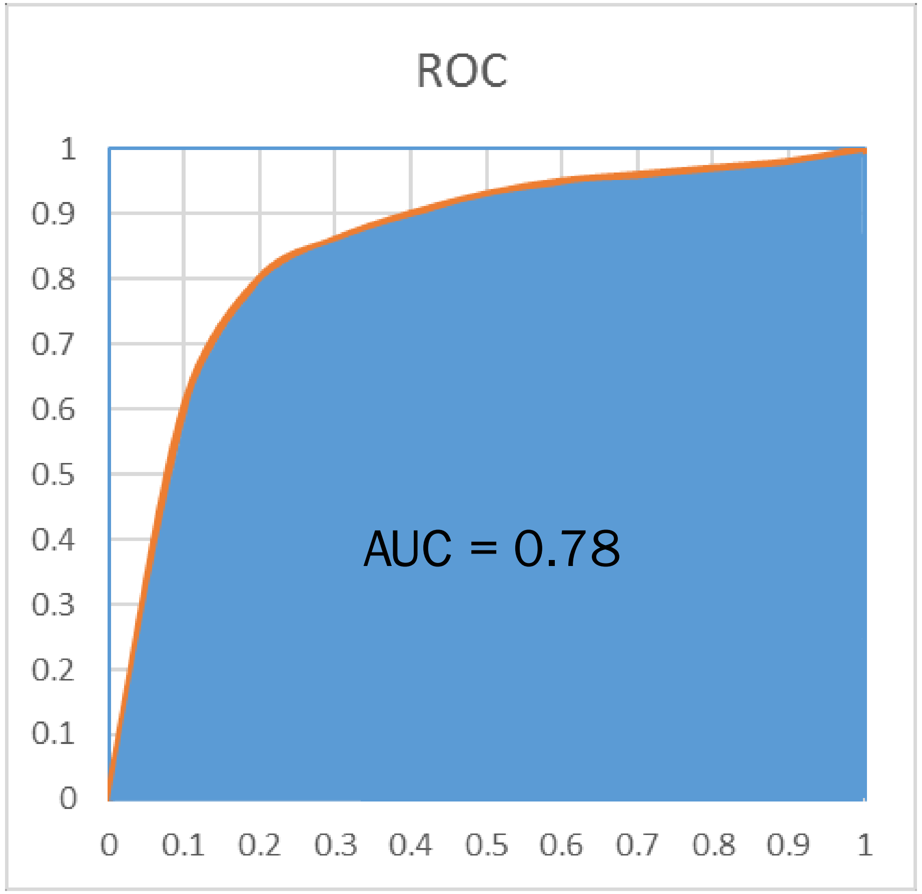
(0, 0)

Area Under the Curve

A MEASURE OF CLASSIFICATION ACCURACY

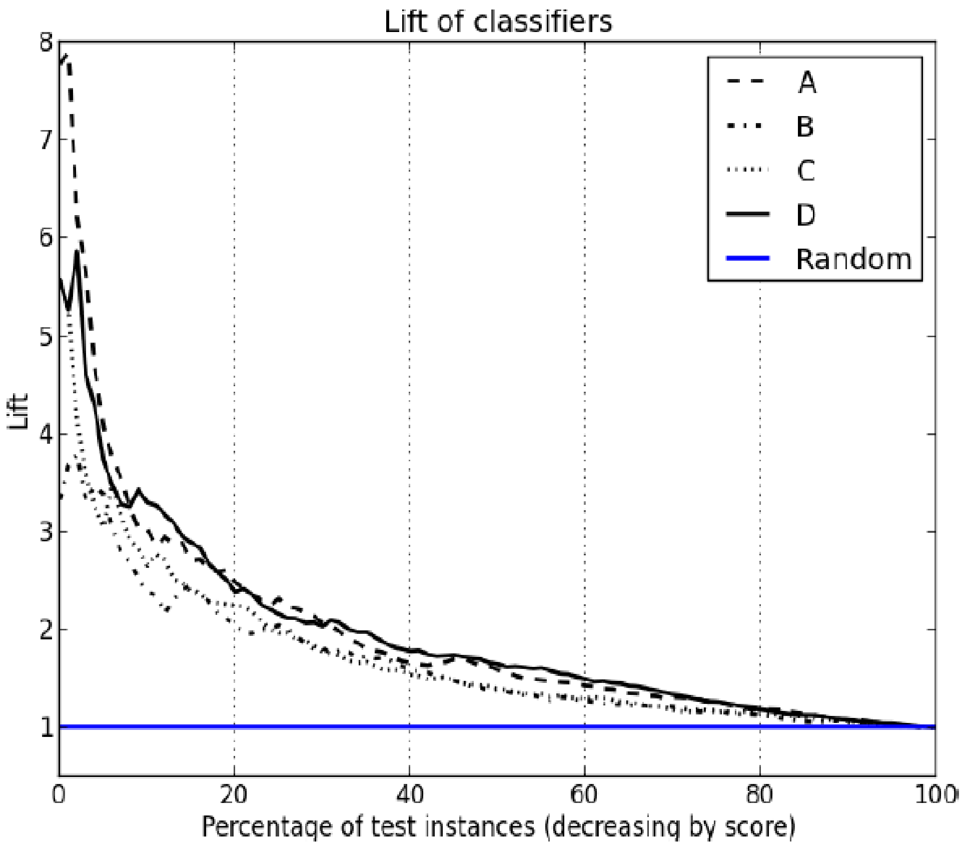
AUC is a summary statistic

- AUC ranges from zero to one
- Useful as a single number to summarize performance
- Common tool for visualizing model performance for classification
- However, this is not intuitive for business stakeholders



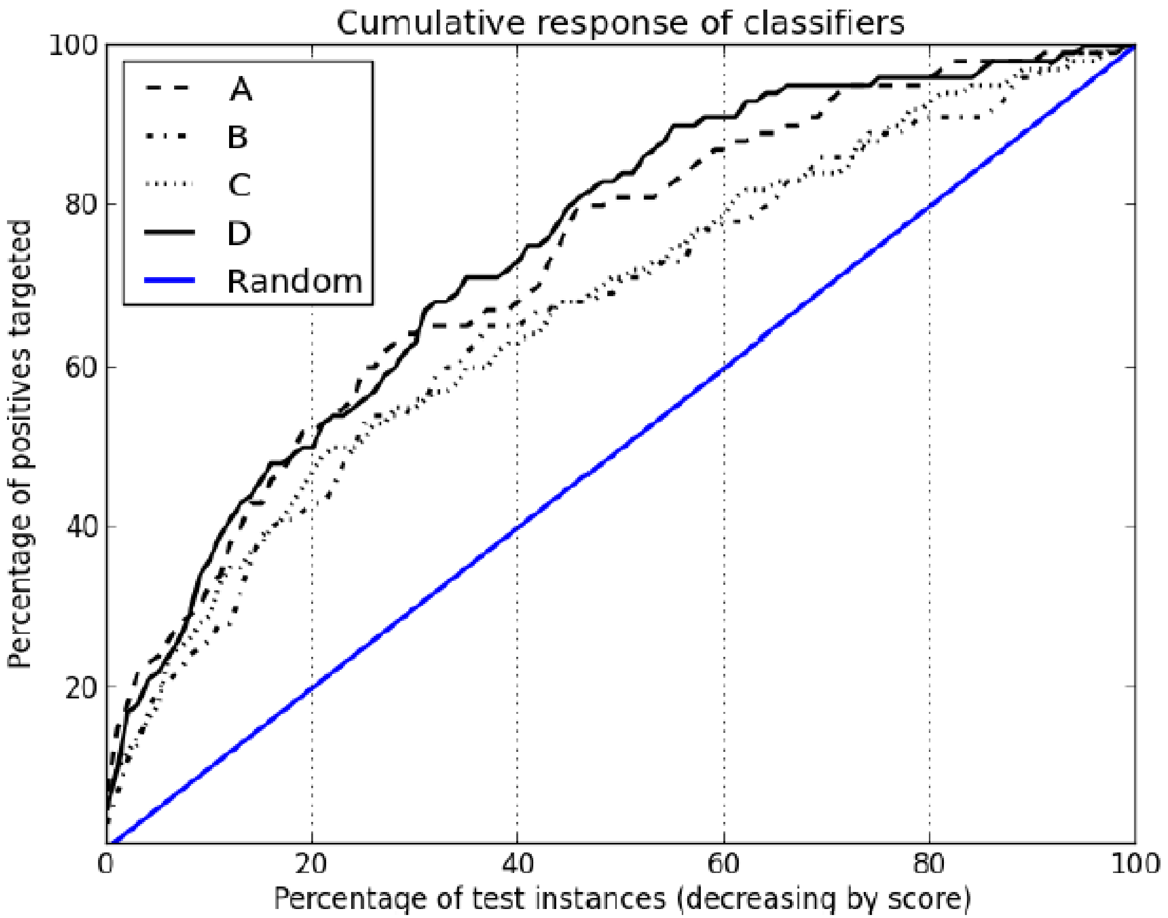
Cumulative response curve

- Plot the hit rate, TPR
- The percentages of positives correctly classified as a function of the percentage of population that is targeted
- The lift of a classifier represents the advantage it provides over random guessing
- The lift is the degree to which it pushes up the positive instances in a list above the negative instances



Lift Curve

- Both lift curves and cumulative response curves must be used with care if the exact proportion of positives in the population is unknown or is not represented accurately in the test data
- These curves assume that the test set has the same target class priors as the population to which the model will be applied
- Simplifying assumptions that need to mention, that allow for a more intuitive visualization



Summary of ROC curves

Valuable visualization for data scientists

Display the trade-offs that each model is making

AUC is a summary statistic for comparison

✓

Business stakeholders don’t understand

Value in Lift curves or cumulative return curves

✗

Appendix

Acknowledgments

Sources for this lecture include but not limited to:

Provost, Data Science for Business: What you need to know

https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

https://arxiv.org/pdf/1106.1813.pdf