Metrics for Classification

Classification models predict class and/or probability

- Classification model outcomes:
  - HARD: Class prediction (e.g. whether a patient has a disease or not)
  - SOFT: Probability of being a given class

- Not all classification models have a meaningful definition of probability
  - Some have a pseudo-probability (e.g. tree models and SVMs) – generally costly for SVM

- **We judge our models based on the class and probability predictions they make**

- Sometimes we care only about the class predictions, other times we use probabilities as inputs into other downstream models

The most naive metric: Accuracy

What percent did we get right?

Accuracy is % of observations classified correctly

- Accuracy for any classification model is defined as:

$$\frac{\text{Observations Correctly Classified}}{\text{All Observations}}$$

Accuracy is % of observations classified correctly

- Accuracy: (observations correctly classified / all observations)
- Accuracy is useful as a first heuristic, but it has shortcomings

Student exercise

(a) Is 95 percent accuracy a good score?
(b) Can you name some shortcomings of accuracy as a metric? Think about cases where we're trying to predict highly imbalanced classes.

Accuracy Example

- Say we’re trying to predict whether a patient has a disease
- In our sample, 99.5% of patients do not have the disease

- **What naive model could we use to get high accuracy?**

Accuracy Example

- Say we’re trying to predict whether a patient has a disease
- In our sample, 99.5% of patients do not have the disease

- **What naive model could we use to get high accuracy?**
- **If we naively predict “no disease” for every observation, we get 99.5% accuracy!**

Accuracy Example

- Say we’re trying to predict whether a patient has a disease
- In our sample, 99.5% of patients do not have the disease

- What naive model could we use to get high accuracy?
- If we naively predict “no disease” for every observation, we get 99.5% accuracy!

- **So... maybe we should look at other metrics as well**

Demystifying the confusion matrix

A confusion matrix is accuracy by class

Predicted

Actual

	# days it was sunny	# days it rained
# days it was sunny	75	5
# days it rained	10	25

A confusion matrix is accuracy by class

Predicted

Actual

	# days it was sunny	# days it rained
# days it was sunny	75	5
# days it rained	10	25

A confusion matrix is used to detail the performance of a classification model.

A confusion matrix is accuracy by class

Predicted

Actual

	Negative class (0)	Positive class (1)
Negative class (0)	True negatives	False positives
Positive class (1)	False negatives	True positives

For those quadrants where our model was correct, we call them true positive/negative. Where our model was wrong, we call them false positive/negative.

Confusion Matrix Example

From the previous naïve model where we predicted "no disease" for every observation, what does the confusion matrix look like for 1000 people?
(Recall: 99.5% do not have disease)

Predicted

Actual

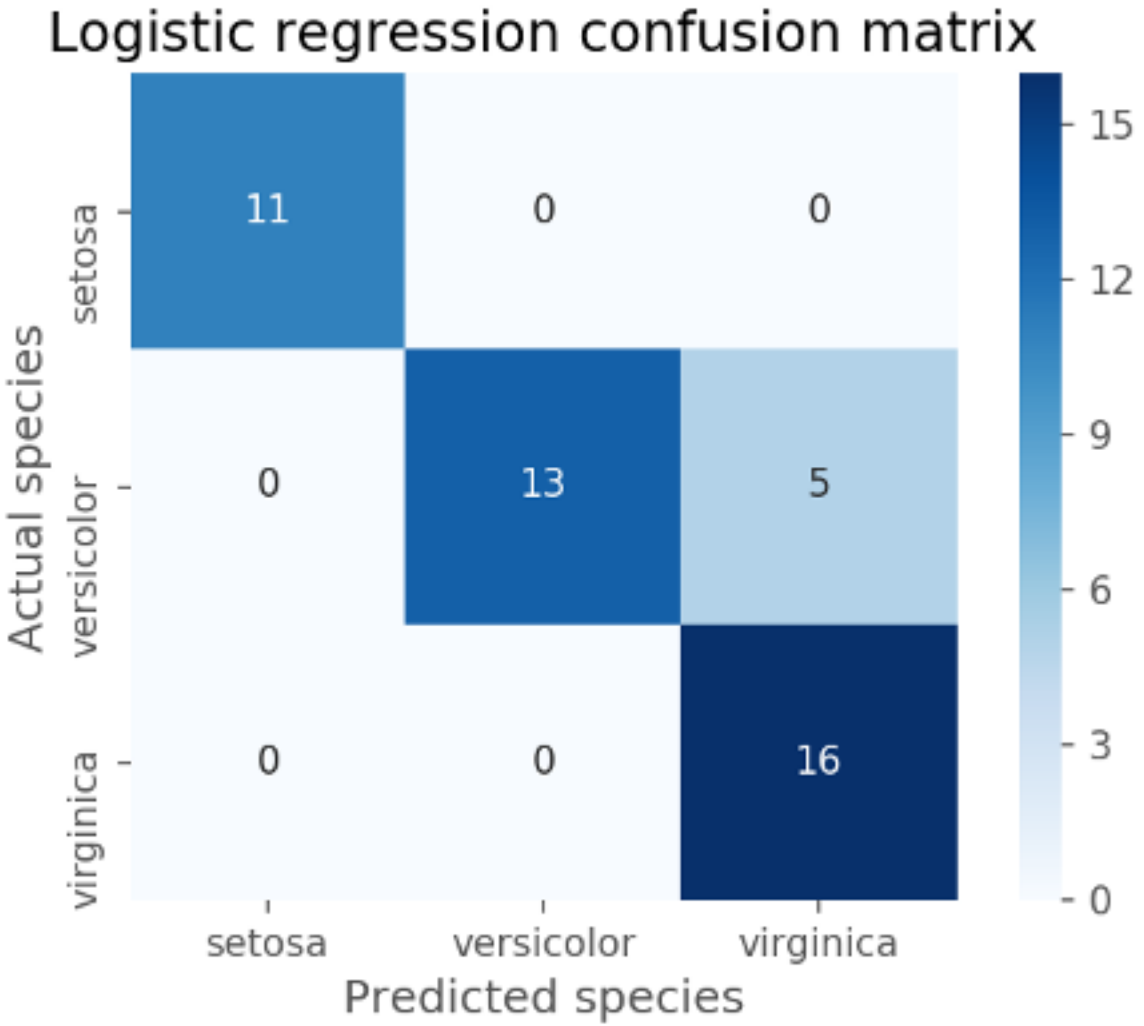
	Negative class	Positive class
Negative class		
Positive class		

Confusion Matrix Example

From the previous naïve model where we predicted "no disease" for every observation, what does the confusion matrix look like for 1000 people?
(Recall: 99.5% do not have disease)

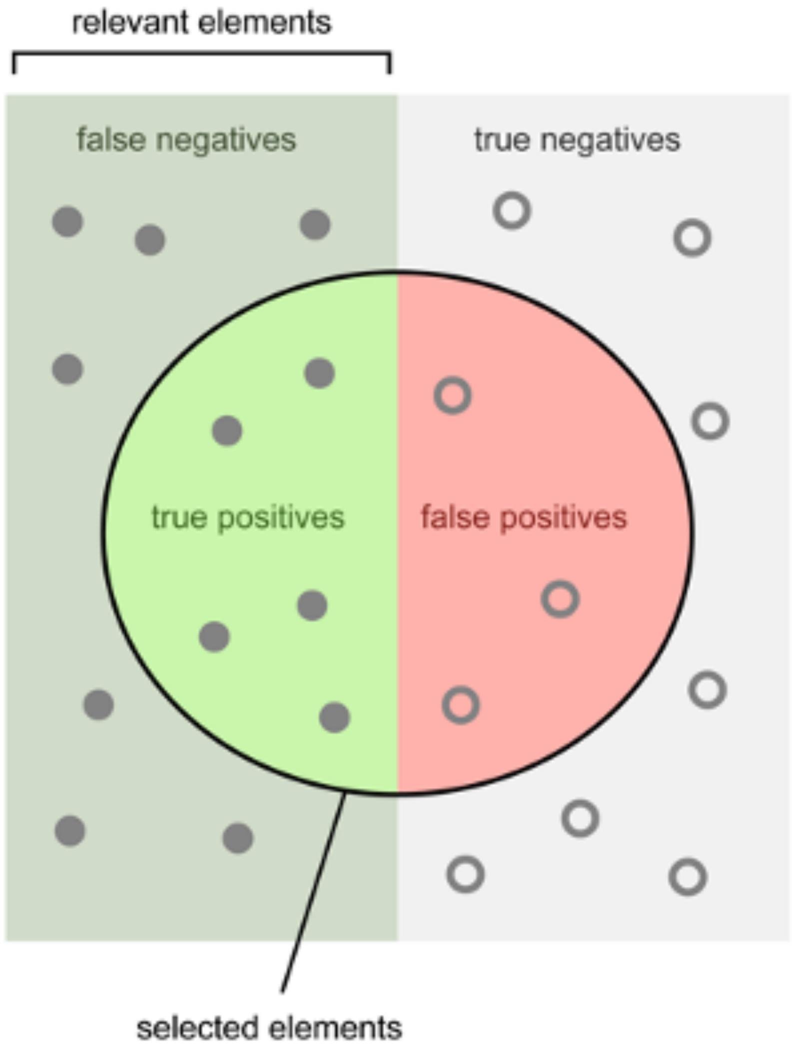
		Predicted	
		Negative class	Positive class
Actual	Negative class	995	0
	Positive class	5	0

A confusion matrix is useful in multiclass problems



Other accuracy-based metrics: Precision and recall

When getting one class correct is more important

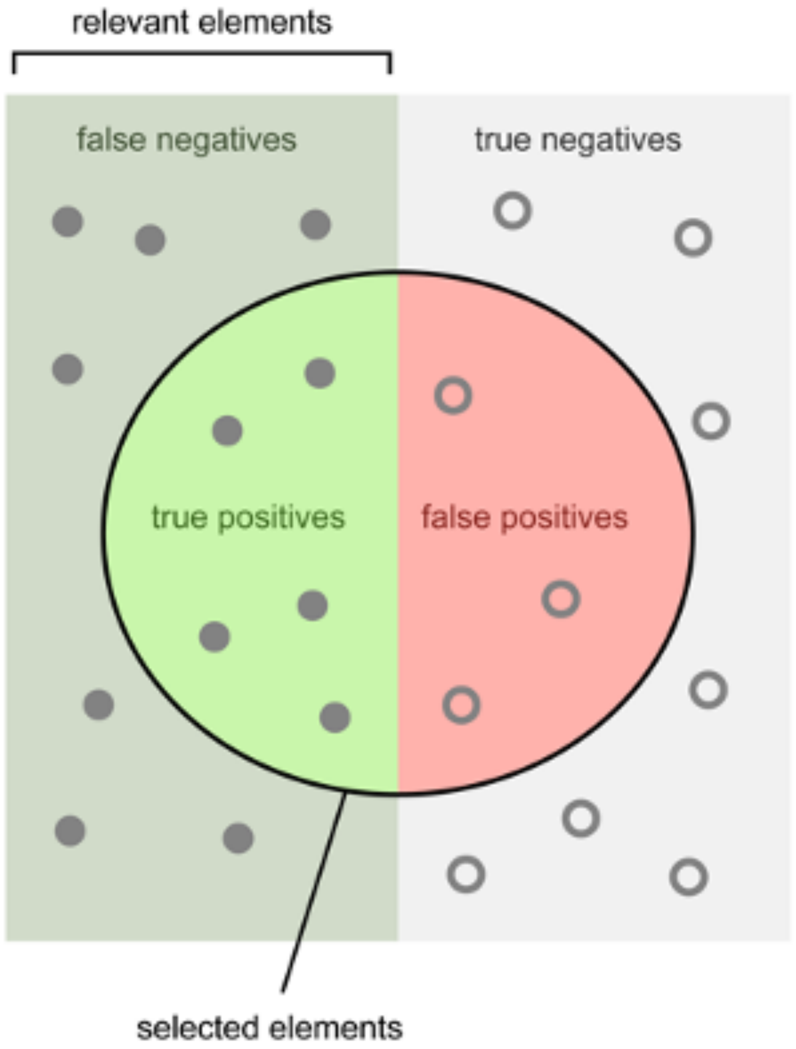


Sometimes we don't care about the accuracy of all classes equally

- e.g. Detecting credit card fraud or the presence of a rare disease
- Sometimes we're willing to trade misclassifying one class to get better accuracy in a different class

Images from Wikipedia.

Precision and recall



Student exercise:

- Can you name cases where we may care more about precision?
- What about cases where we care more about recall?

Images from Wikipedia.

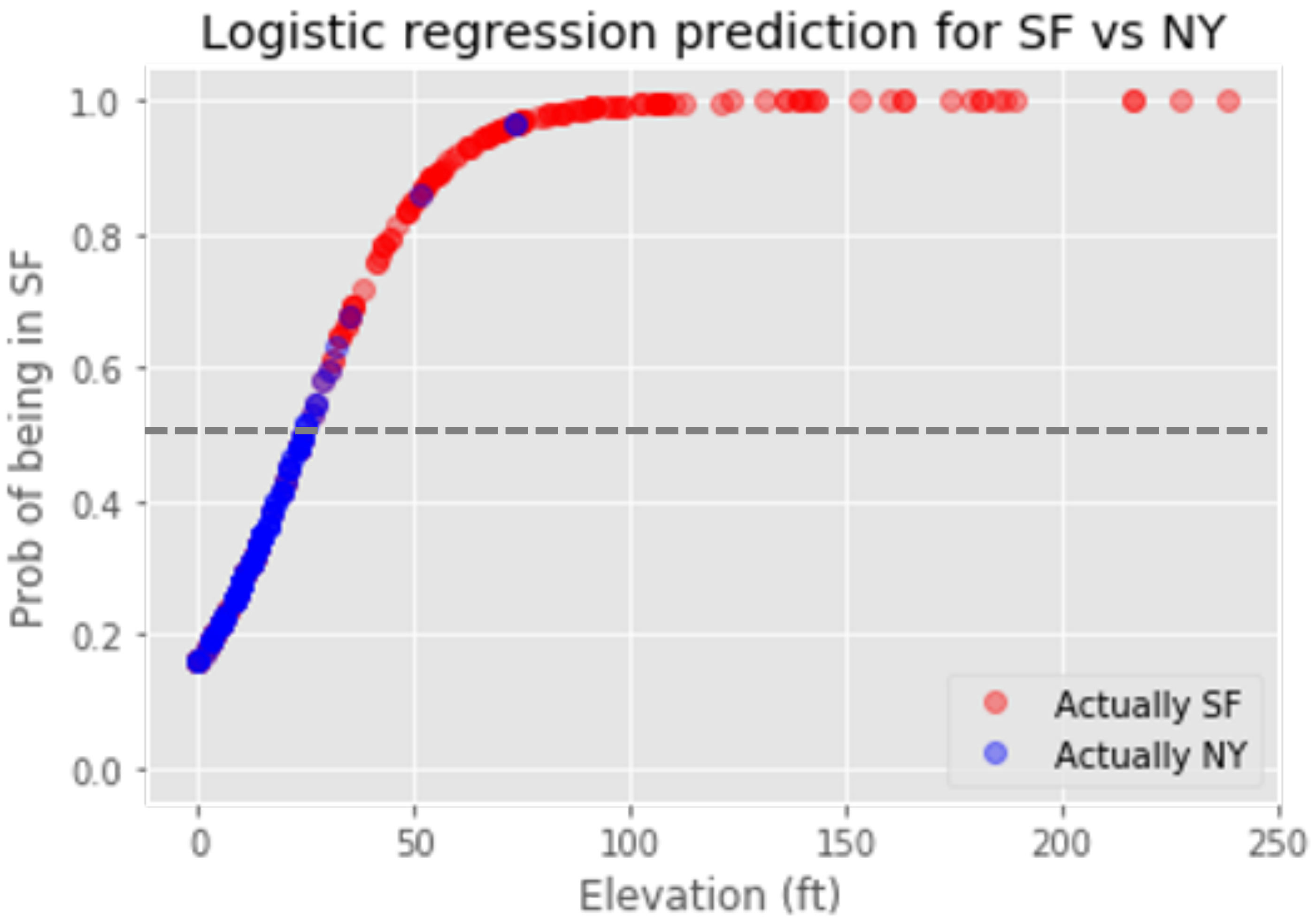
Using class probability predictions

Choosing a probability threshold

- **Some models give us probability predictions and not just class predictions**

- So far we've looked at accuracy, precision and recall derived from a 50% probability threshold (sklearn default)
  - But, we don't have to take the 50% cutoff, we can choose our own!
  - Setting this cutoff means choosing our probability threshold

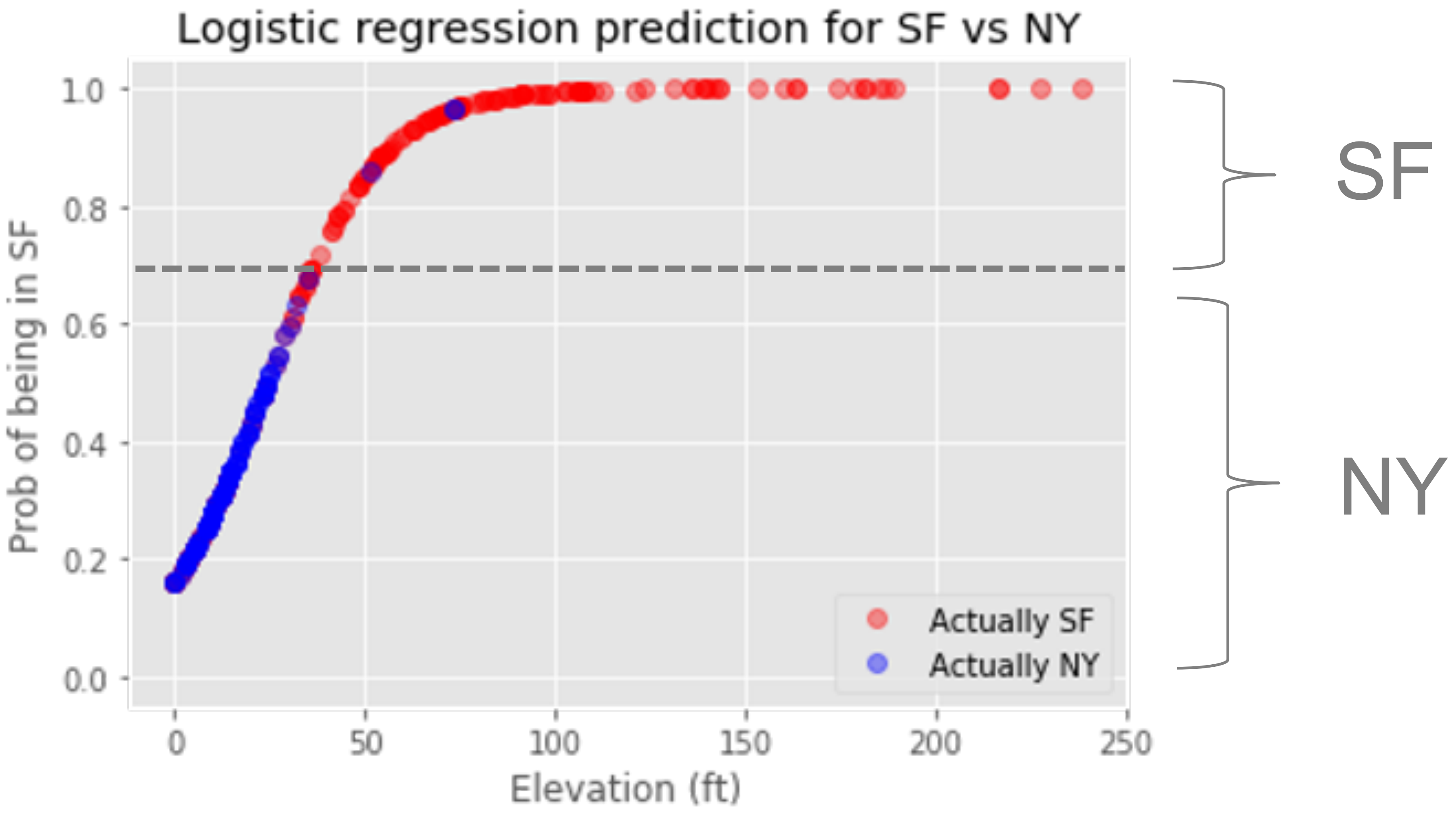
Probability threshold – 50%



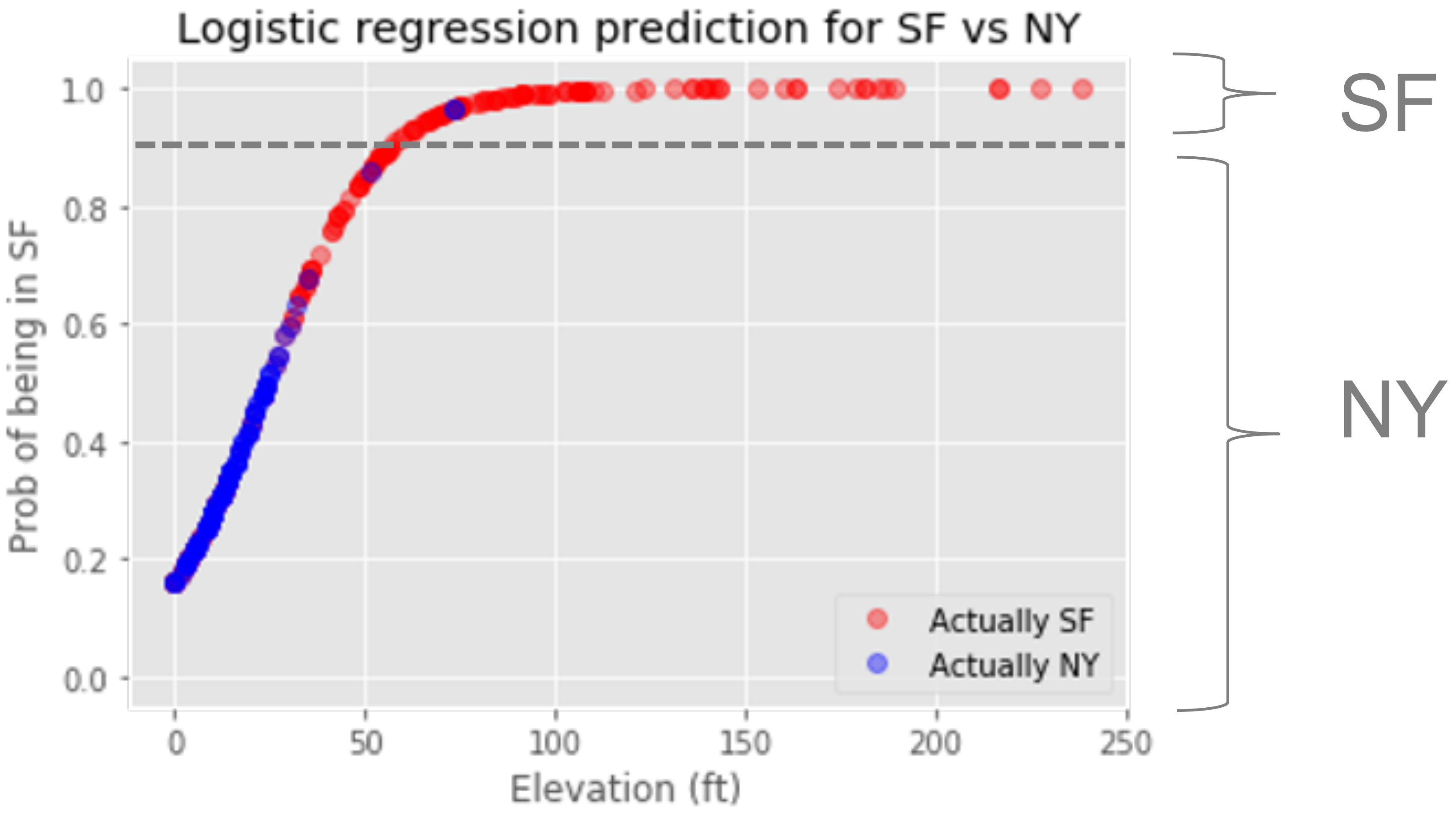
SF

NY

Probability threshold – 70%



Probability threshold – 90%

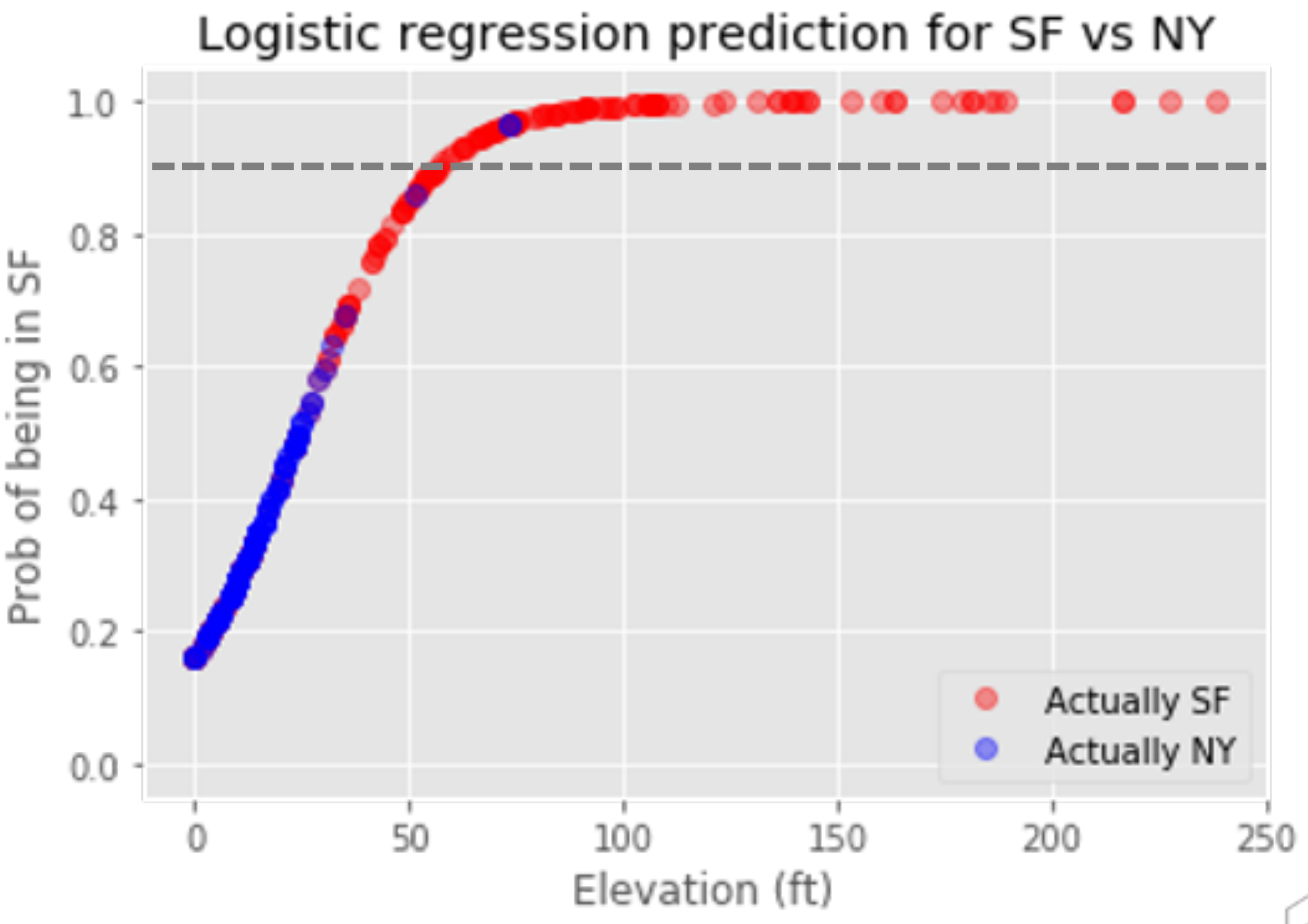


Probability threshold – 90%

Note: SF = positive class

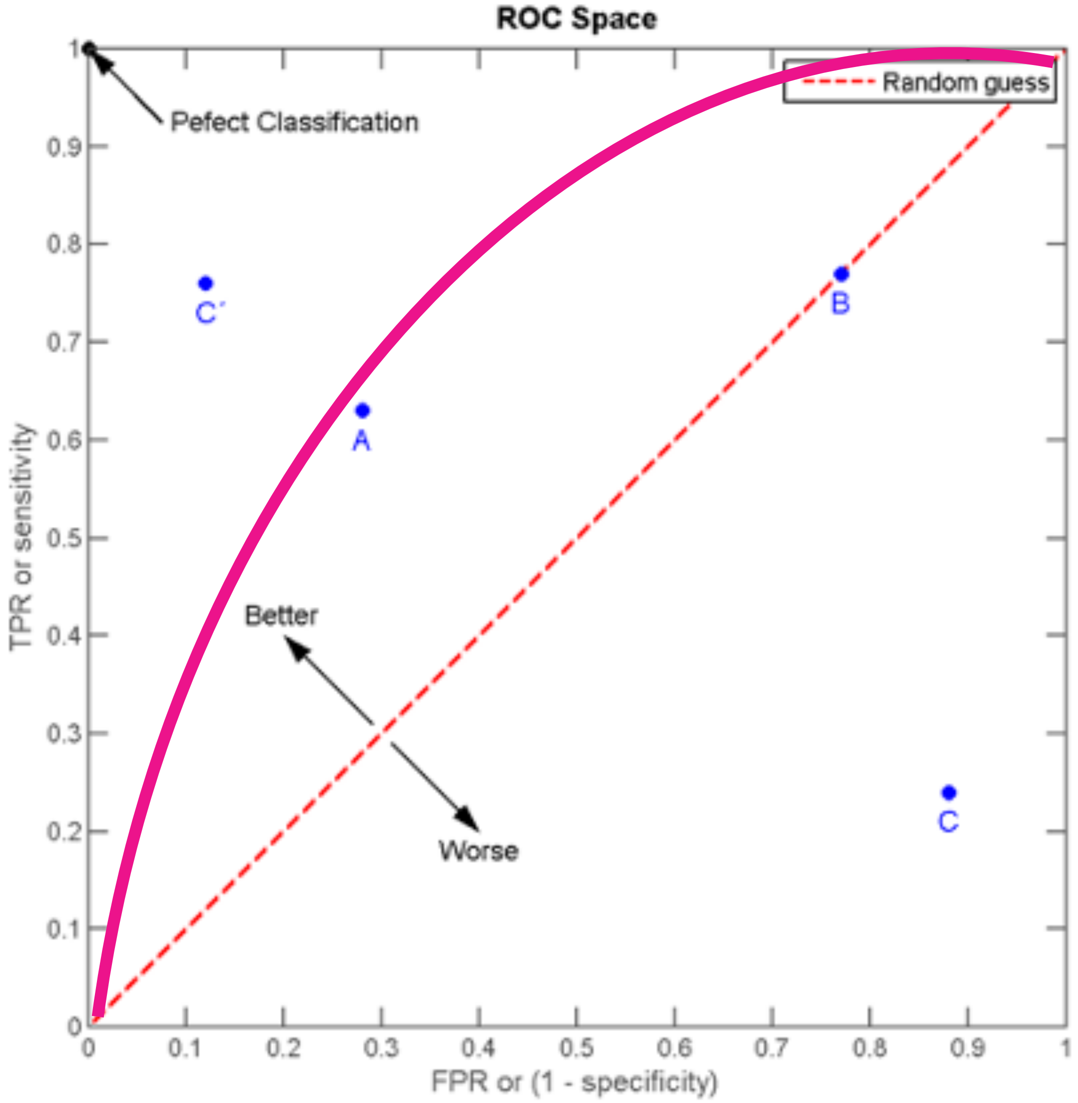
Student exercise:

- As we increase the threshold, do we have:
  - Lower/higher recall?
  - Lower/higher precision?
  - Lower/higher true positive rate? (TP/AP)
  - Lower/higher false positive rate? (FP/AN)



Using a ROC curve to determine probability thresholds

- Drawing a ROC curve: change the probability threshold and plot how true positive rate and false positive rate change
- Each threshold gives us a new model!
- **We can plot the ROC curve only for binary cases**

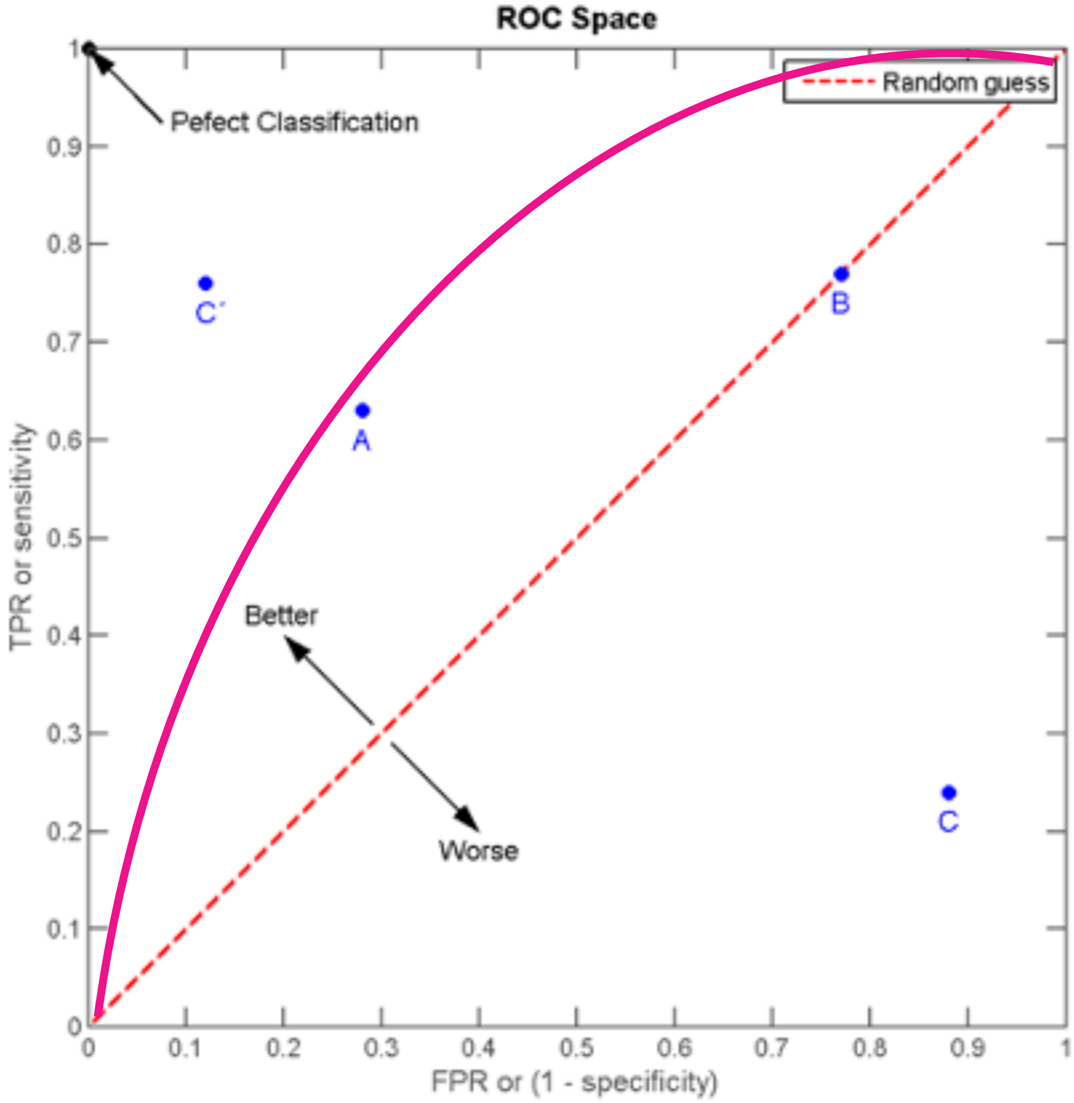


Images from Wikipedia.

Using a ROC curve to determine probability thresholds

Check for understanding:

Which corner represents a higher threshold? Lower threshold?



NOTE:

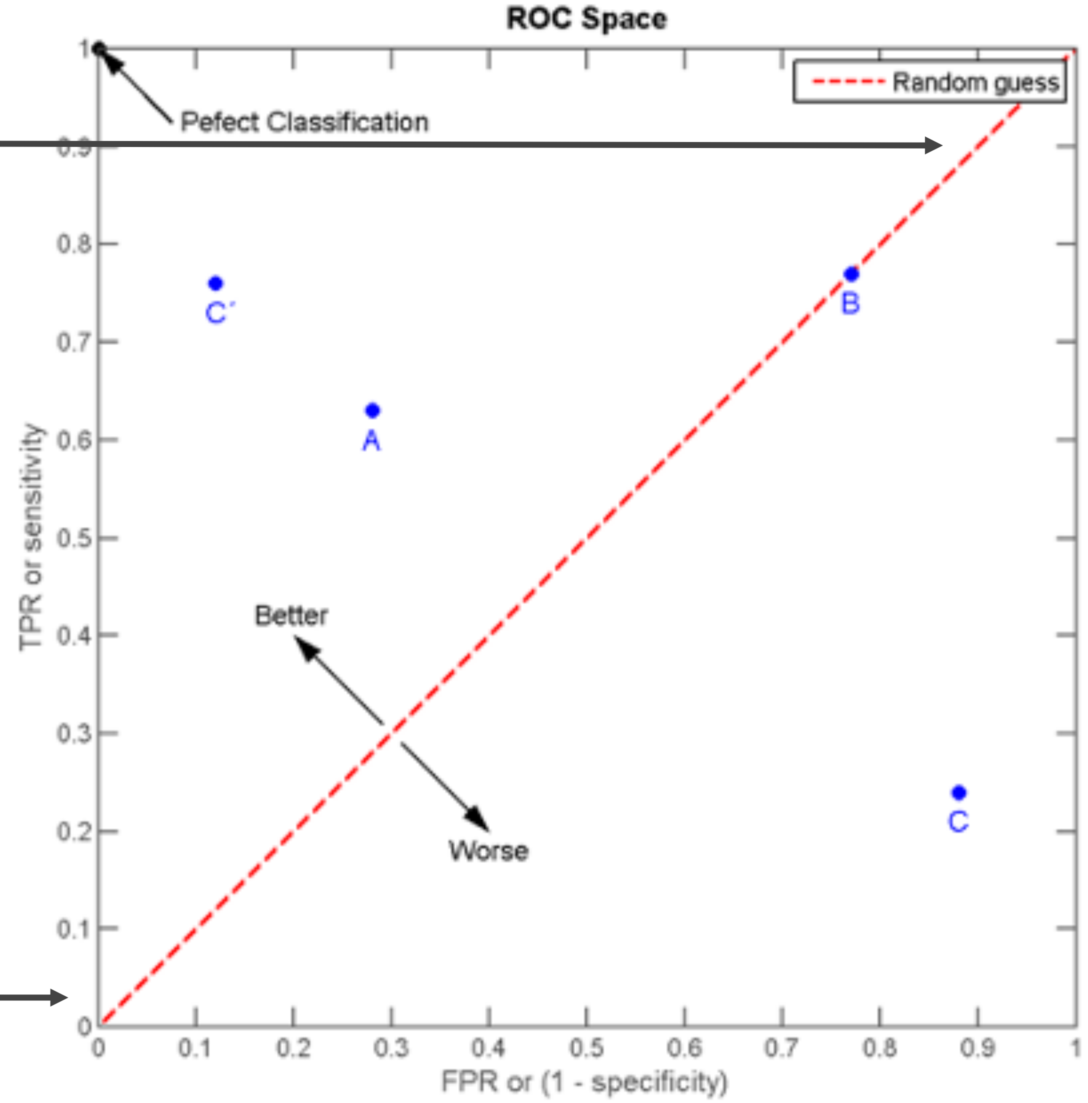
- TPR = True Positive Rate
  (True Positives/Actual Positives)
- FPR = False Positive Rate
  (False Positives/Actual Negatives)

Images from Wikipedia.

Using a ROC curve to determine probability thresholds

- Lower threshold: Better at catching positives. Higher recall, lower precision. Higher true positive rate, higher false positive rate

- Higher threshold: Better at catching negatives. Lower recall, higher precision. Lower true positive rate, lower false positive rate.

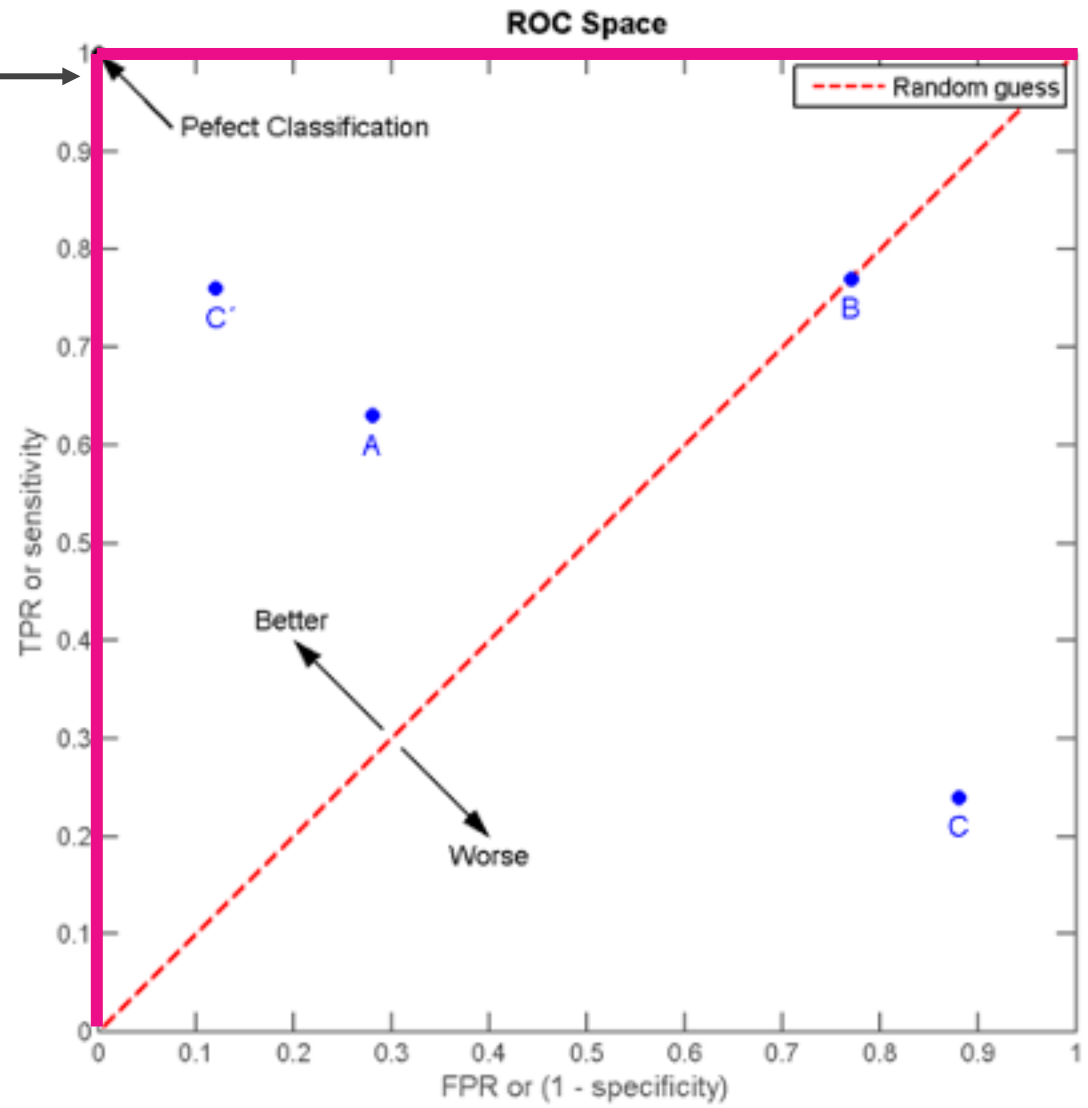


Images from Wikipedia.

The perfect classifier

- The perfect classifier (pink line) would be a curve that reaches the northwest corner
- This represents a zero false positive rate and a 100% true positive rate

- A metric related to the ROC curve is the **area under the curve (AUC)**
- Notice that for the perfect classifier, the AUC would equal 1
- An AUC closer to 1 is better, and it ranges from 0 (0.5) to 1



Images from Wikipedia.

ROC-AUC Interpretation

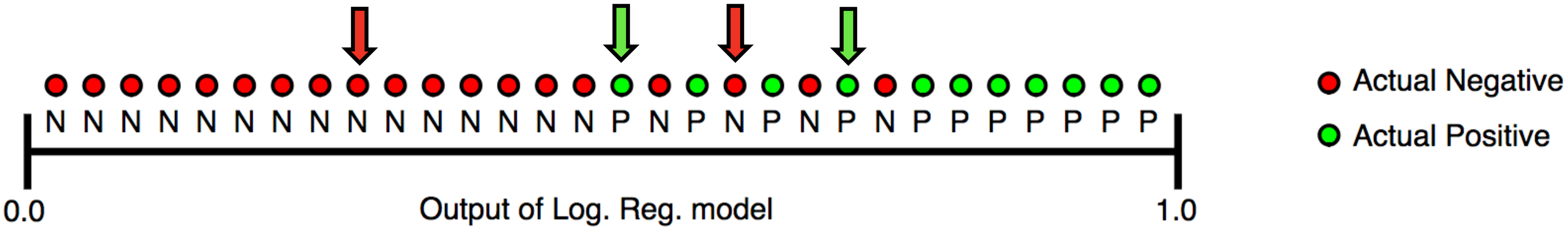


Figure 6. Predictions ranked in ascending order of logistic regression score.

One way of interpreting AUC:

Probability that random actual positive is ranked higher than random actual negative

ROC-AUC Interpretation

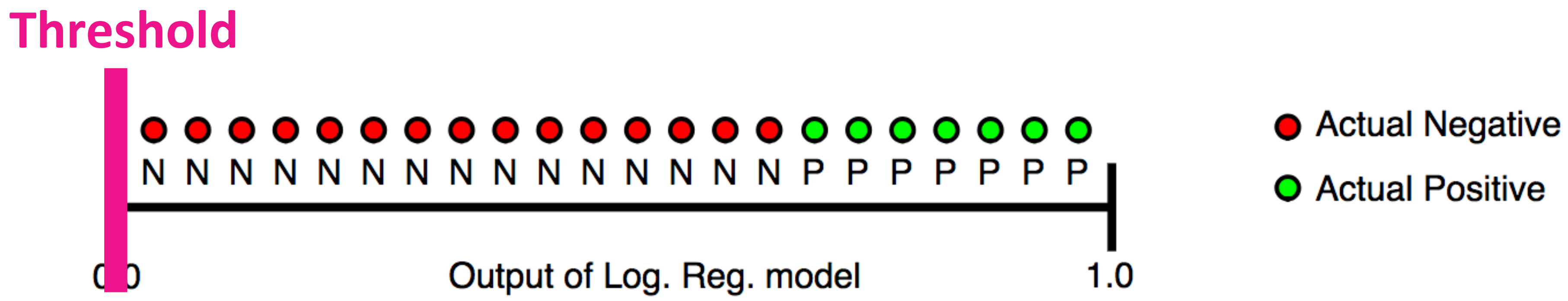


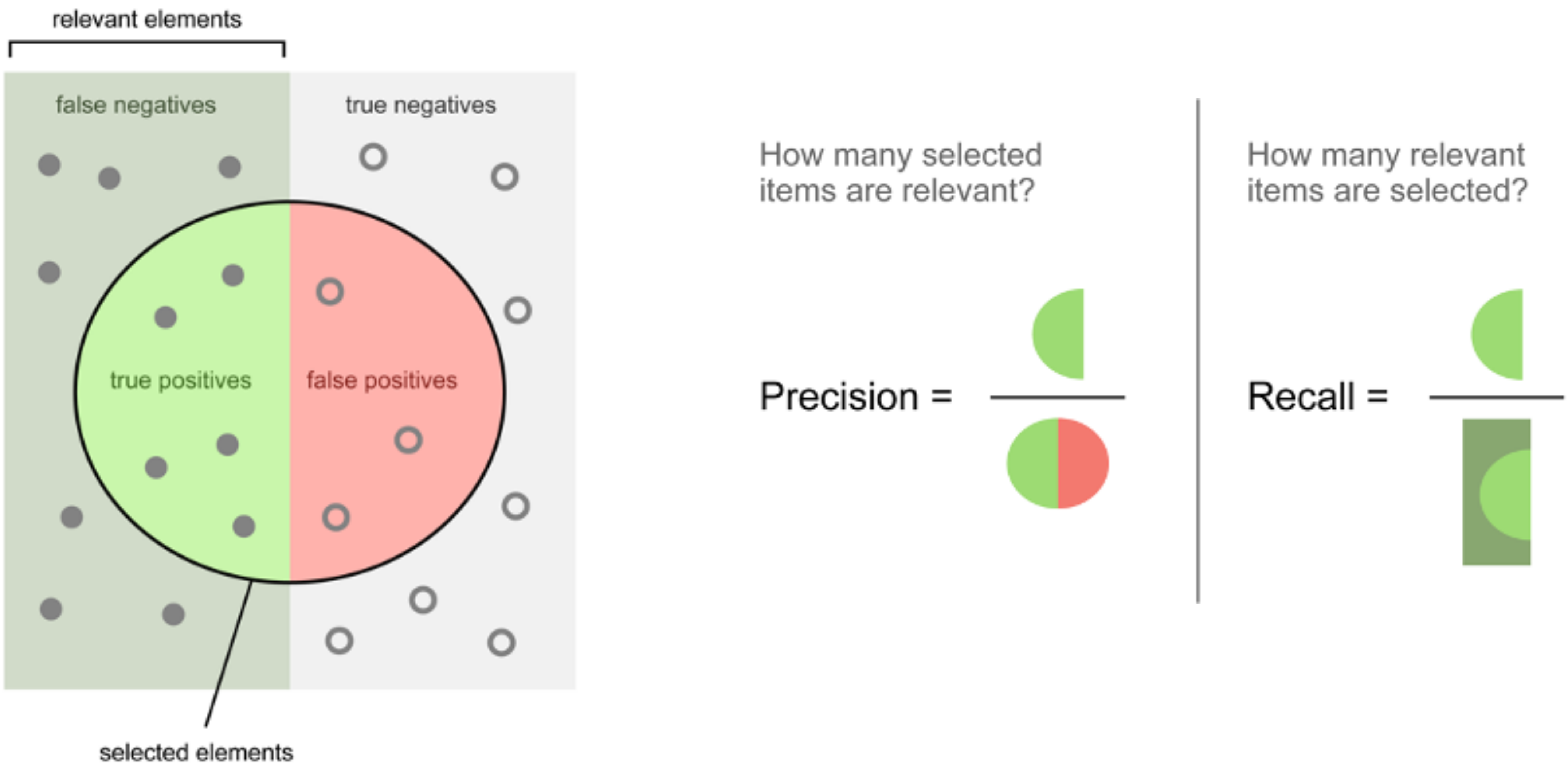
Figure 6. Predictions ranked in ascending order of logistic regression score.

Perfect Classifier:

All actual negatives have lower probability than all actual positives.

ROC AUC = 1

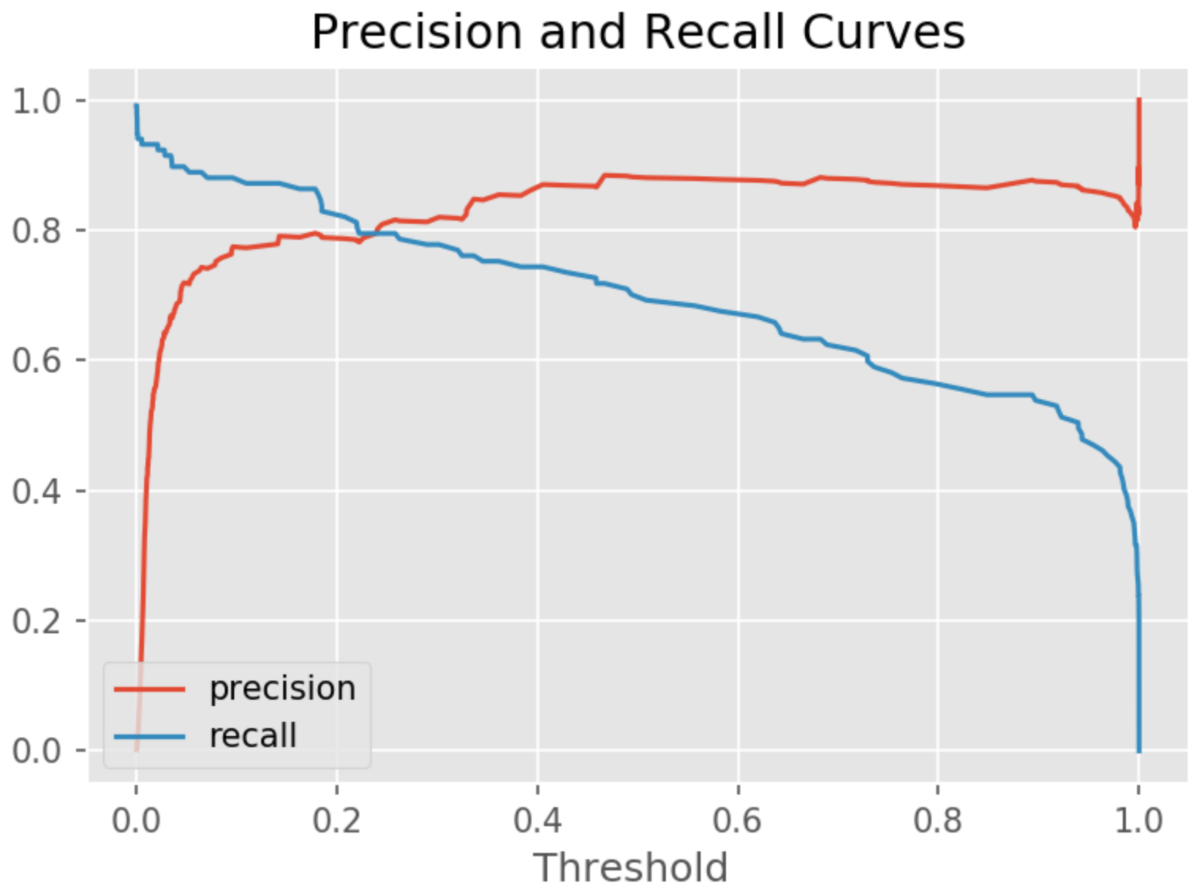
Precision and recall -- encore!



Images from Wikipedia.

The precision-recall curve

- So we can change our probability threshold, and thus the hard classifications of our model
- As we move the threshold, we will also change our precision and recall
- Want to find every positive class in the data?
    - Then decrease the threshold to almost nothing: Voilà! Almost all observations will be classified as positive
- Want to make sure what you classify as positive is truly positive?
    - Increase the threshold and make it harder for the model to classify observations as positive



Applications to model development

Fit to training data, evaluate on test (or cross-val)

FIT MODEL TO TRAINING DATA

EVALUATE ERROR METRICS ON VAL (or K-FOLD CV)

- **Accuracy, Precision, Recall**
- **Confusion Matrix**
- **ROC Curve**
- **F1-Score**

…AND FINALLY…

REPORT METRICS FROM HOLDOUT TEST DATA

Appendix:
Objectives & Even more classification metrics

Learning Objectives

- Understand the difference between model class predictions vs. probability predictions
- Learn about the most common error metrics for classification:
  - Accuracy and accuracy-based metrics:
    - Confusion matrix
    - Precision and recall
    - Log-loss as a measure that takes the magnitude of uncertainty into account
  - Others:
    - ROC curve and maximizing the area under the curve (AUC)
- Understand when to apply each metric, particularly the difference between two-class and multiclass problems

More classification metrics

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = (Σ True positive + Σ True negative) / Σ Total population	
	Negative likelihood ratio (**LR–**) = FNR/TNR	False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR–				

Image from Wikipedia.

More classification metrics

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR, Fall-out) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
	Negative likelihood ratio (**LR−**) = FNR/TNR	False negative rate (FNR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR, Specificity, SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$		
	Diagnostic odds ratio (**DOR**) = LR+/LR−				

Image from Wikipedia.

More classification metrics

		Condition **(as determined by "Gold standard")**			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive Σ Test outcome positive	False discovery rate (FDR) = Σ False positive Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative Σ Total population	
	Negative likelihood ratio (**LR–**) = FNR/TNR	False negative rate (FNR) = Σ False negative Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR–				

Image from Wikipedia.

More classification metrics

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population	
	Negative likelihood ratio (**LR–**) = FNR/TNR	False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR–				

Image from Wikipedia.

Using a ROC curve to compare algorithms