Intro to Machine Learning
Classification: Performance Measures

Slide Credit: Prof. Ben Leong

Correctness

Classification is correct when prediction $\hat{y}$ is the same as actual label y, i.e.,

$$\text{Correct} = [\hat{y} = y]$$

where
$\hat{y} = M(\boldsymbol{x})$ is the predicted value from model M
instance $\boldsymbol{x}$
y is the ground truth value

Accuracy

"Average correctness" across test dataset with m instances:

$$A = \frac{1}{m}\sum_{j=1}^{m}\left[\hat{y}_j = y_j\right]$$

where

$\hat{y}_j = M(\boldsymbol{x}_j)$ is the predicted value from model M of the jth instance $\boldsymbol{x}_j$

y_j is the ground truth value of the jth instance

Confusion Matrix

Inst.	Predicted $\hat{y}$	Actual y	
1	Alert	Alert	TP
2	Alert	Alert	
3	Sleepy	Alert	FN
4	Sleepy	Alert	
5	Sleepy	Alert	
6	Sleepy	Sleepy	TN
7	Sleepy	Sleepy	
8	Sleepy	Sleepy	
9	Sleepy	Sleepy	
10	Alert	Sleepy	FP

student alertness prediction

		Actual Label	
		Alert	Sleepy
Predicted Label	Alert	2	1
	Sleepy	3	4

		Actual Label	
		+ve	-ve
Predicted Label	+ve	TP True Positive	FP False Positive
	-ve	FN False Negative	TN True Negative

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}$$

Precision vs Recall

Inst.	Predicted $\hat{y}$	Actual y
1	Alert	Alert
2	Alert	Alert
3	Sleepy	Alert
4	Sleepy	Alert
5	Sleepy	Alert
6	Sleepy	Sleepy
7	Sleepy	Sleepy
8	Sleepy	Sleepy
9	Sleepy	Sleepy
10	Alert	Sleepy

student alertness prediction

Actual Label

Predicted Label	Alert	¬Alert	
Alert	2 True Positive	1 False Positive	3 ∑ Pred. Pos.
¬Alert	3 False Negative	4 True Negative	7 ∑ Pred. Neg.
	5 ∑ Actual Pos.	5 ∑ Actual Neg.	

Precision
P = TP / (TP+FP)

Recall
R = TP / (TP+FN)

FP: Type I error

FN: Type II error

Precision vs Recall

Consider a “Minority Report”-like program that is used in the Courtroom of the Future. Our program predicts whether an accused person is guilty.

What is precision?	Correctly convict a guilty person
What is recall?	Percentage of correct convictions

Precision vs Recall

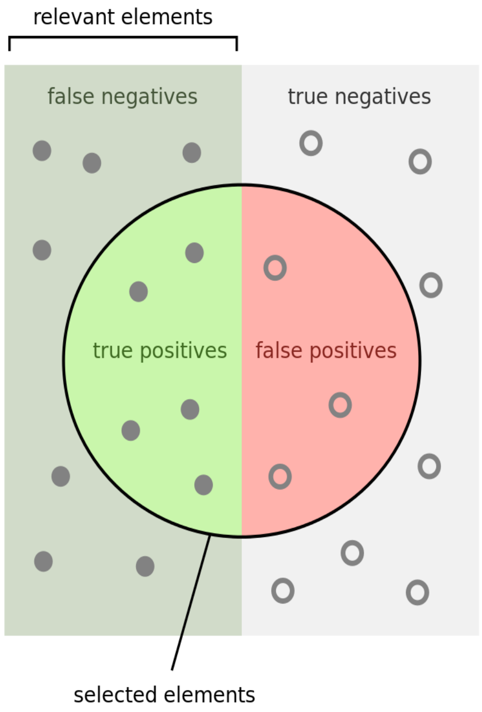
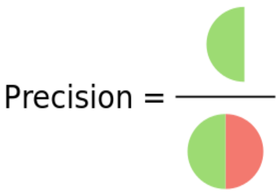


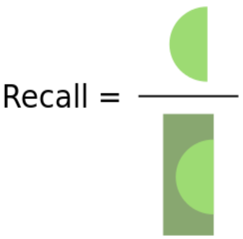
Image credit: https://en.wikipedia.org/wiki/Precision_and_recall



How many **selected** items are **relevant**?

How **precise** were the positive predicted instances?

Maximize this if false positive (FP) is very costly.
E.g., email spam, satellite launch date prediction



How many **relevant** items are **selected**?

How many positive instances can be **recalled** (predicted)?

Maximize this if false negative (FN) is very dangerous.
E.g., cancer prediction but not music recommendation

Good discussion:
https://datascience.stackexchange.com/questions/30881/when-is-precision-more-important-over-recall

F1 Score: $\left(\frac{P^{-1}+R^{-1}}{2}\right)^{-1}$

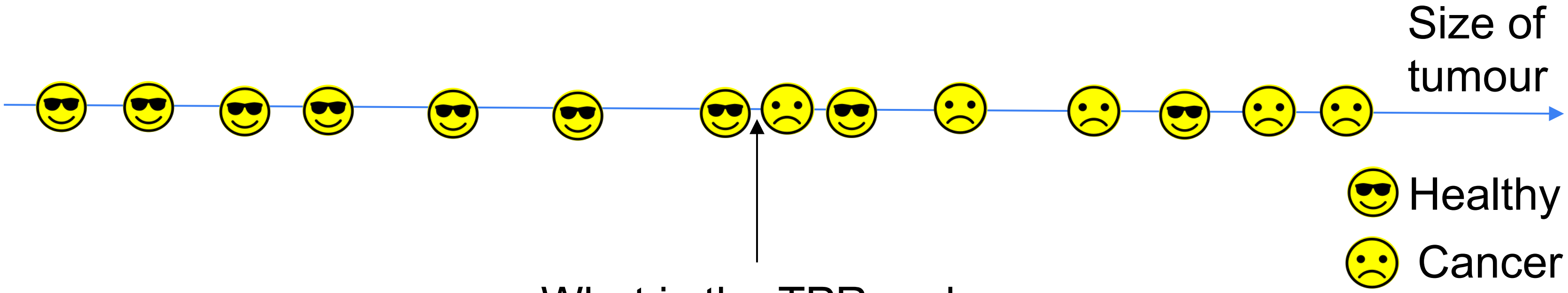
1. The measure is more **robust** (less sensitive to extreme values)
   Ref: https://stackoverflow.com/a/26360501

2. It considers that the numerators of P and R are the same, so it compares their denominators

$$F_1 = \left(\frac{P^{-1} + R^{-1}}{2}\right)^{-1} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2\,TP}{(TP + FP) + (TP + FN)} = \frac{2\,TP}{2TP + FP + FN}$$

Other "fairer" metrics that consider true negatives (TN):
Matthews correlation coefficient, Youden's index, Cohen's kappa

Cancer Prediction



What is the TPR and FPR rates here?

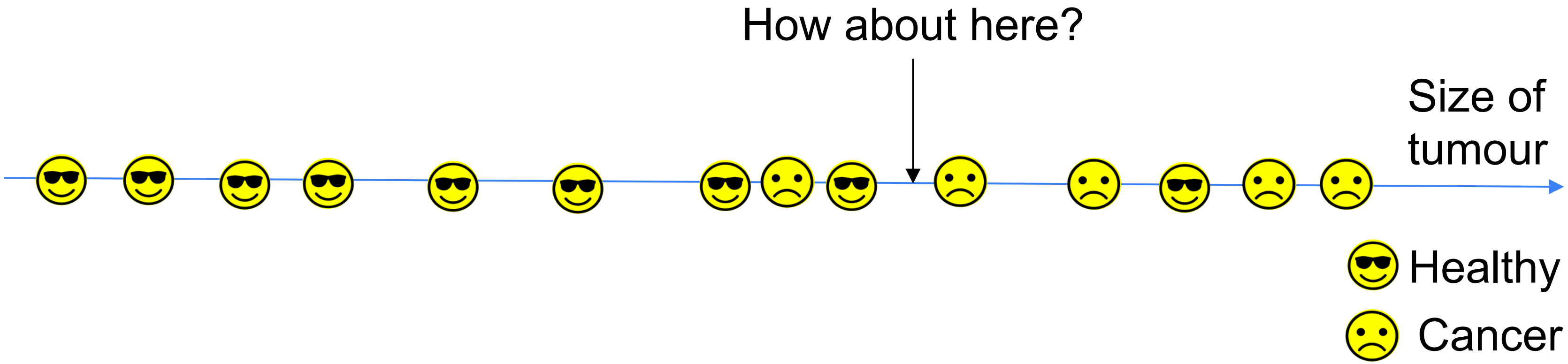
TPR =5/5 = 1

FPR =2/9

FPR = FP / (FP+TN)

TPR = TP / (TP+FN)

Cancer Prediction



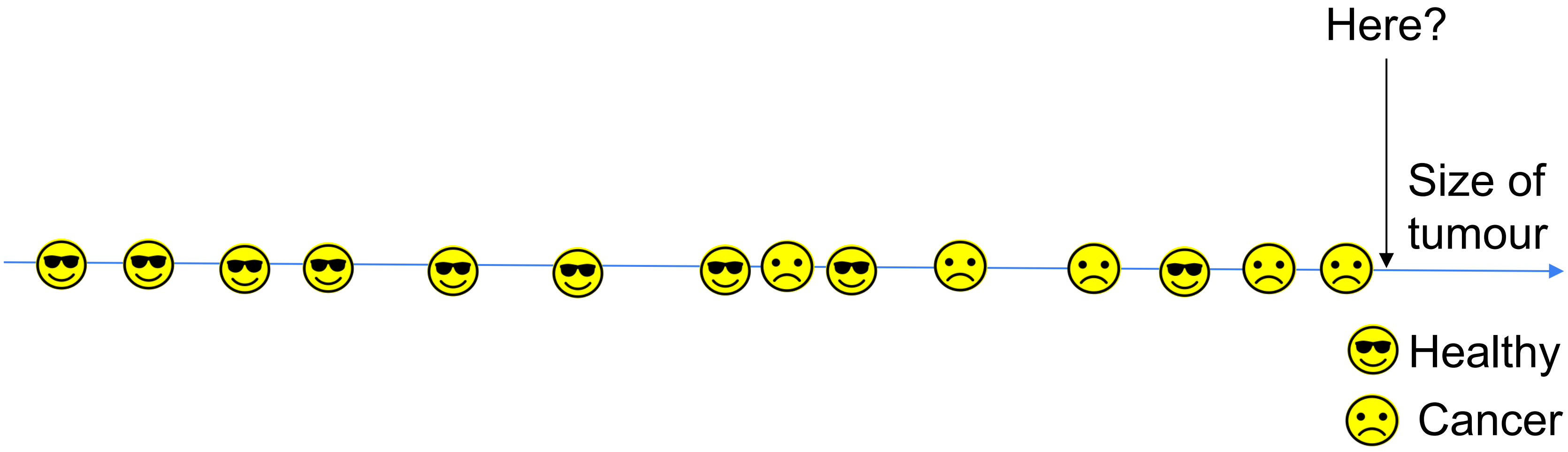
FPR =
FP / (FP+TN)

TPR =4/5 = 0.8

FPR =1/9

TPR =
TP / (TP+FN)

Cancer Prediction



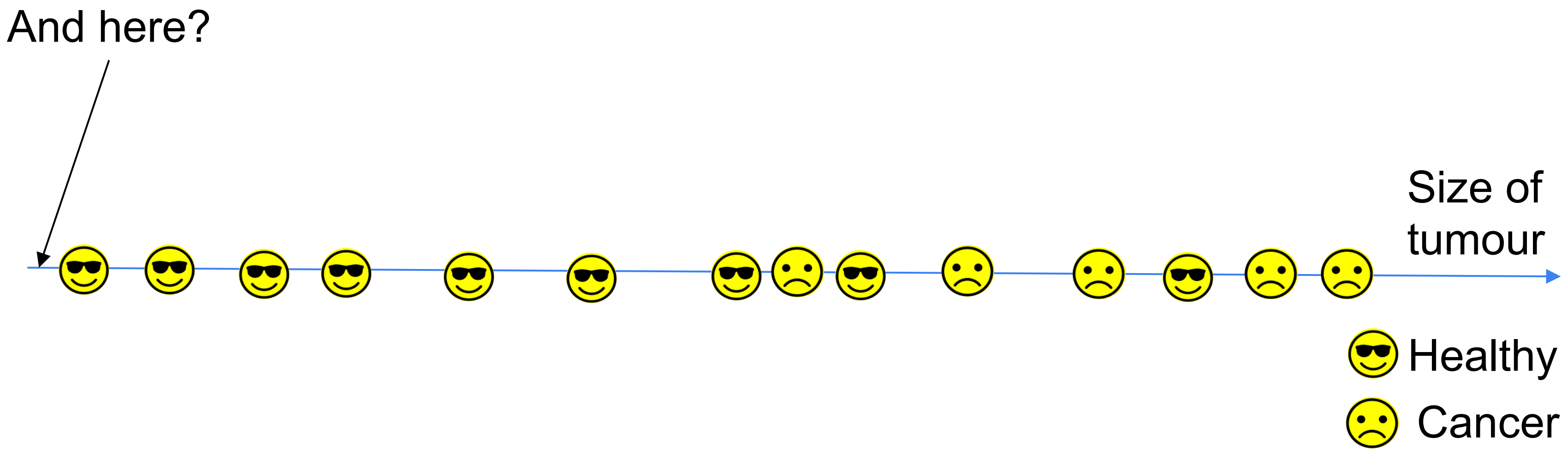
FPR =
FP / (FP+TN)

TPR =
TP / (TP+FN)

TPR =0/5 = 0

FPR =0/9 = 0

Cancer Prediction



FPR =
FP / (FP+TN)

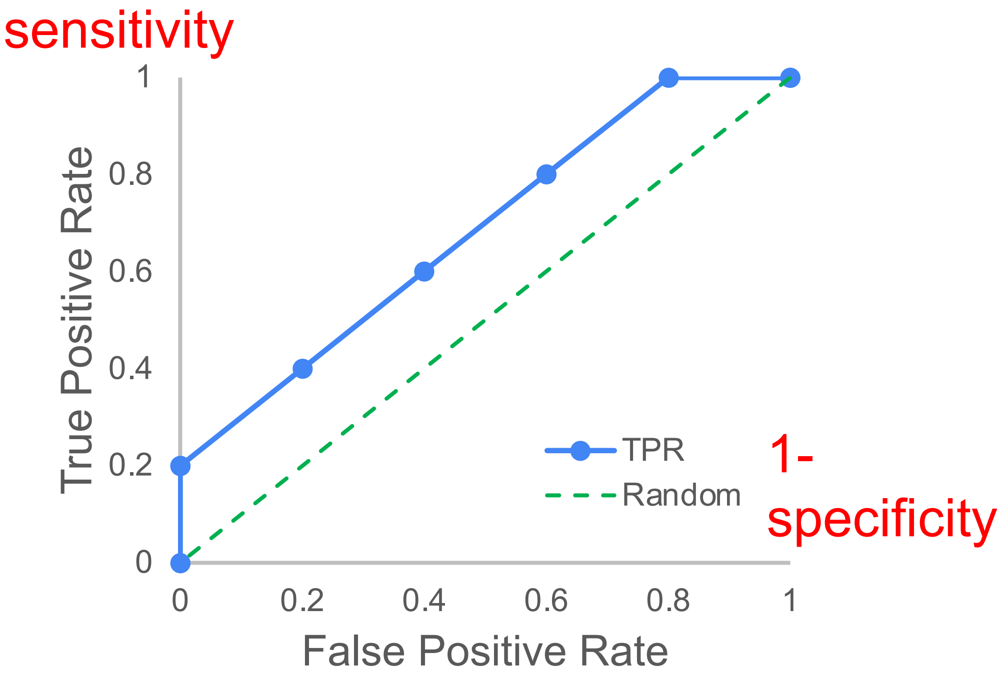
TPR =
TP / (TP+FN)

TPR =5/5 = 1

FPR =9/9 = 1

Receiver Operator Characteristic (ROC) Curve

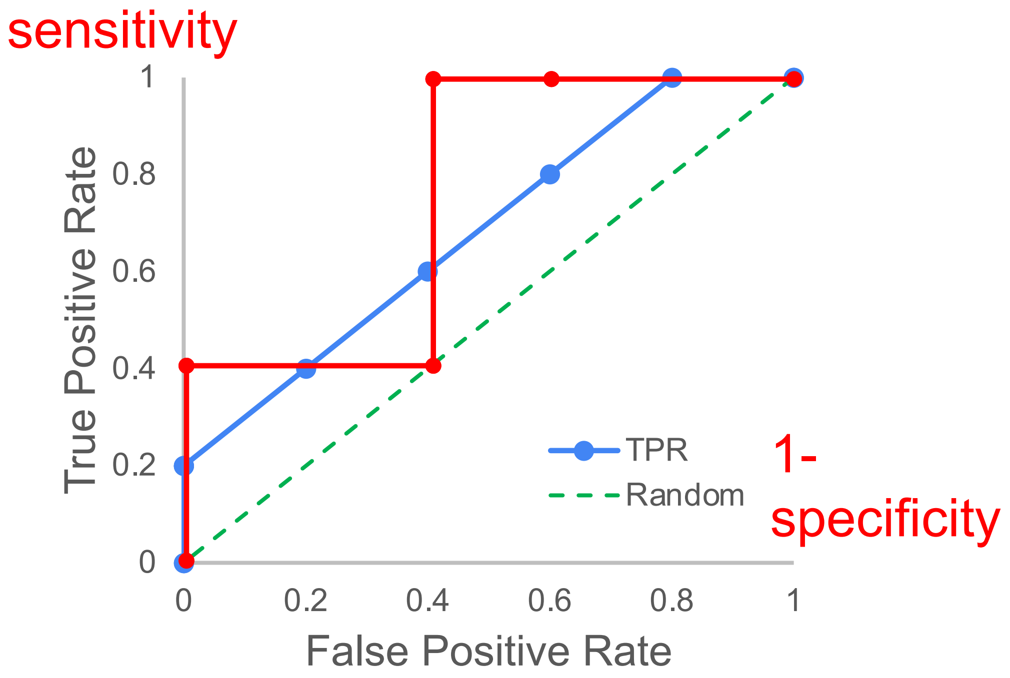
Threshold π	TPR	FPR
0	1	1
0.1	1.0	1.0
0.2	1.0	0.8
0.3	0.8	0.6
0.4	0.6	0.4
0.5	0.4	0.2
0.6	0.2	0.0
0.7	0.2	0.0
0.8	0.2	0.0
0.9	0.0	0.0
1	0	0



Model is more accurate than random chance
If its **ROC curve** is above the diagonal **random** line.

Receiver Operator Characteristic (ROC) Curve

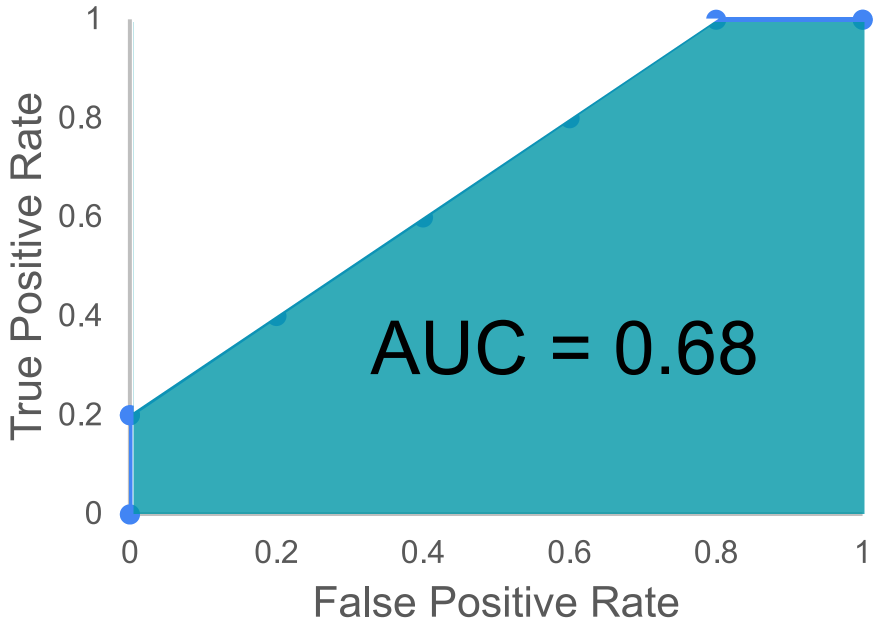
Threshold π	TPR	FPR
0	1	1
0.1	1.0	1.0
0.2	1.0	0.8
0.3	0.8	0.6
0.4	0.6	0.4
0.5	0.4	0.2
0.6	0.2	0.0
0.7	0.2	0.0
0.8	0.2	0.0
0.9	0.0	0.0
1	0	0



Model is more accurate than random chance
If its **ROC curve** is above the diagonal **random** line.

Area Under Curve (AUC) of ROC

Threshold π	TPR	FPR
0	1	1
0.1	1.0	1.0
0.2	1.0	0.8
0.3	0.8	0.6
0.4	0.6	0.4
0.5	0.4	0.2
0.6	0.2	0.0
0.7	0.2	0.0
0.8	0.2	0.0
0.9	0.0	0.0
1	0	0



AUC is **concise metric** instead of a full figure.
Concise metrics enable *clearer comparisons*.
AUC > 0.5 means the model is better than chance.
AUC ≈ 1 means model is very accurate.

Area Under Curve (AUC) of ROC (example)

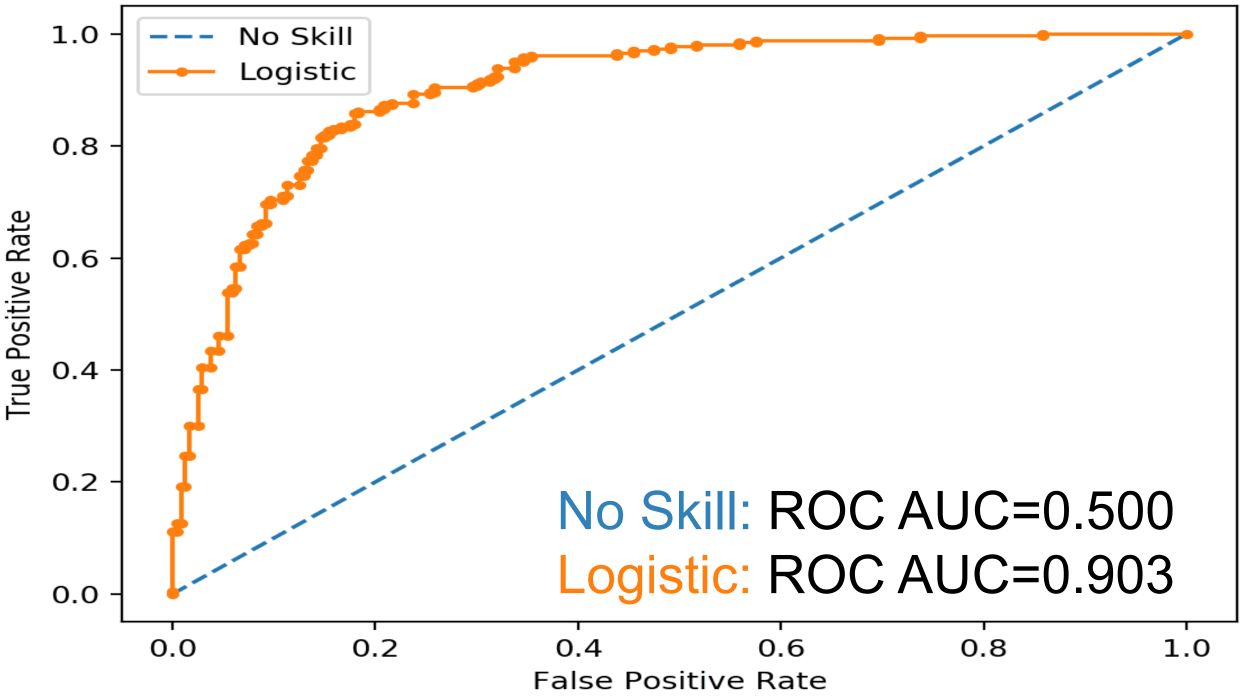


Image credit: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/