Evaluation metrics for `single` threshold

Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

- True positive (TP): correctly predicted positive
- True negative (TN): correctly predicted negative
- False positive (FP): incorrectly predicted positive
- False negative (FN): incorrectly predicted negative

As thresold ⬆️, positive predictions (TP, FP) ⬇️ and negative predictions (TN, FN) ⬆️

Visualizing the confusion matrix

https://developers-dot-devsite-v2-prod.appspot.com/machine-learning/crash-course/classification/thresholding_cd2cec3b3711b6befffa498911d9a6be0fa233b9b7238880d23cdb7593116511.frame

Which mistake is more costly?

Spam detection:

- FP: non-spam email is classified as spam
- FN: spam email is classified as non-spam

Cancer detection:

- FP: non-cancerous tumor is classified as cancerous
- FN: cancerous tumor is classified as non-cancerous

Credit card fraud detection:

- FP: non-fraudulent transaction is classified as fraudulent
- FN: fraudulent transaction is classified as non-fraudulent

Accuracy

$$\frac{\text{correct predictions}}{\text{total predictions}} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Use when the classes are balanced

Avoid for imbalanced datasets

- 99% of the data is negative, and 1% is positive. A model that predicts all negative will have 99% accuracy.

Recall (True positive rate)

$$\frac{\text{correctly predicted positive}}{\text{actual positive}} = \frac{TP}{(TP + FN)}$$

Use when false negatives (FN) are more expensive than false positives (FP).

- spam email is classified as non-spam
- cancerous tumor is classified as non-cancerous
- fraudulent transaction is classified as non-fraudulent

False positive rate

$$\frac{\text{incorrectly predicted negative}}{\text{actual negative}} = \frac{FP}{(FP + TN)}$$

Use when false positives (FP) are more expensive than false negatives (FN).

- non-spam email is classified as spam
- non-cancerous tumor is classified as cancerous
- non-fraudulent transaction is classified as fraudulent

Precision

$$\frac{\text{correctly predicted positive}}{\text{predicted positive}} = \frac{TP}{(TP + FP)}$$

Use when it's very important for positive predictions to be accurate.

- spam email is classified as spam
- cancerous tumor is classified as cancerous
- fraudulent transaction is classified as fraudulent

F1 score

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Use when you want a single metric that balances precision and recall.

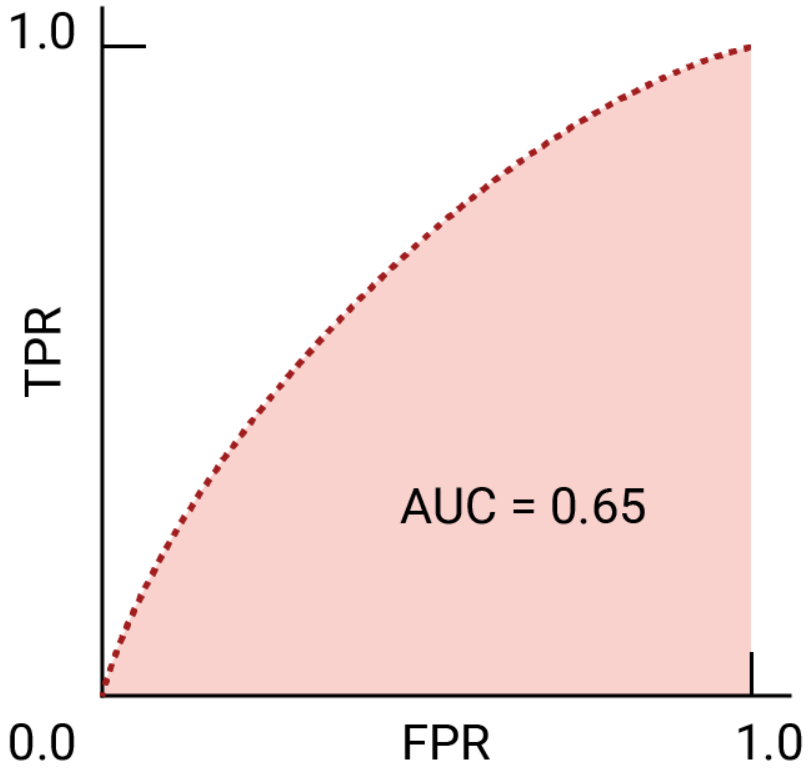
Evaluation metrics for `all` possible thresholds

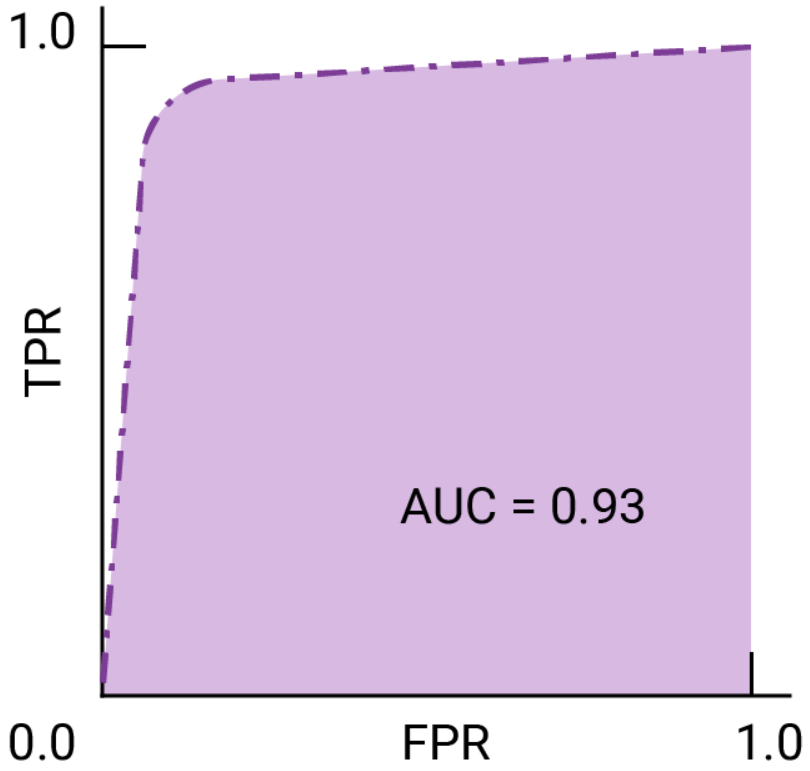
ROC: Receiver-operating characteristic curve

- False positive rate (FPR) vs. True positive rate (TPR) across all thresholds

AUC: Area under the ROC curve

- Probability that the model will rank the actual positive higher than the actual negative.
- e.g., a spam classifier with AUC of 1.0 always assigns a random spam email a higher probability of being spam than a random legitimate email.


1.0
TPR
AUC = 0.65
0.0
FPR
1.0

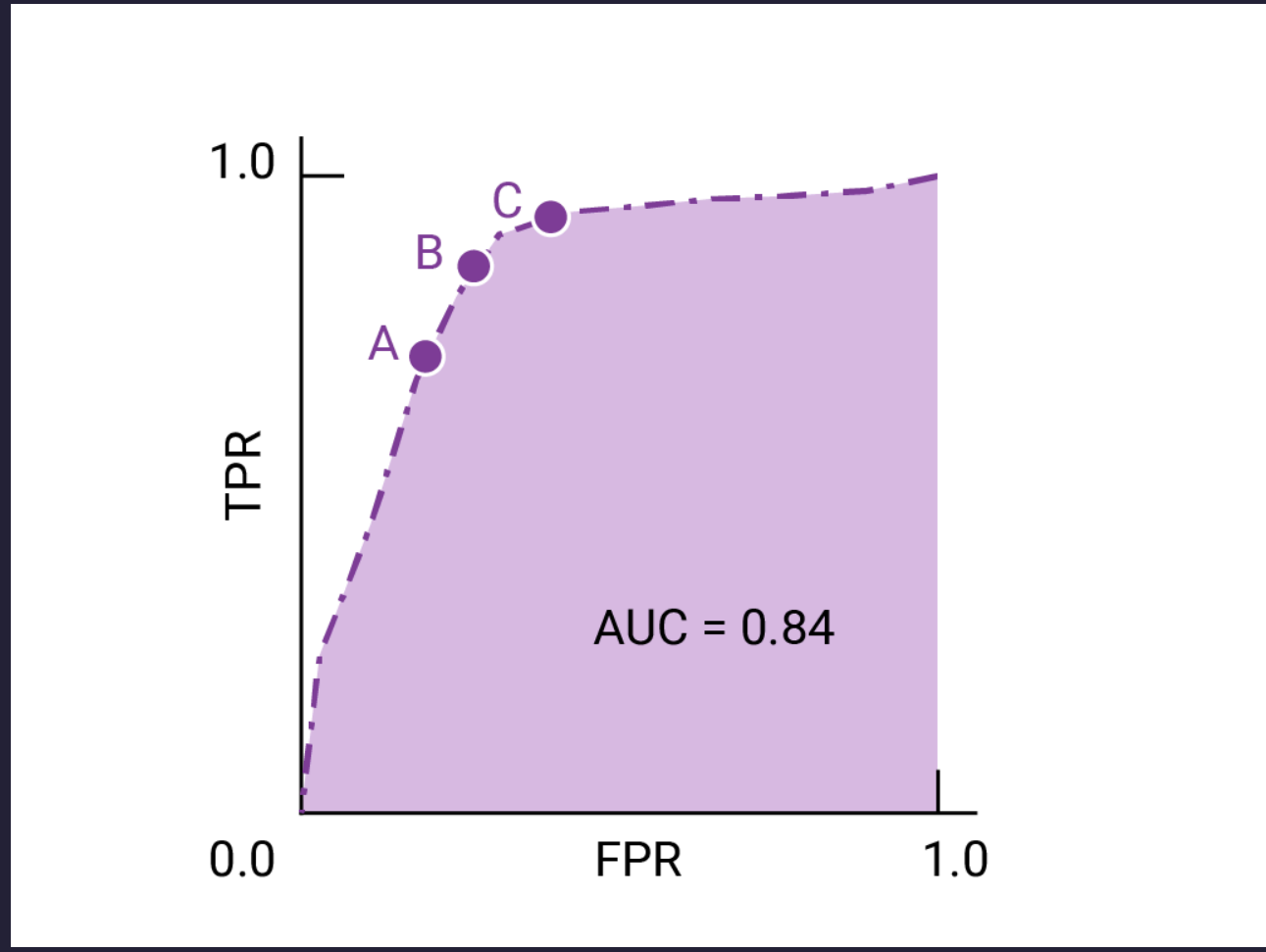

1.0
TPR
AUC = 0.93
0.0
FPR
1.0

Which threshold to choose?

`B` : highest TPR for a given FPR (closest to the top-left corner)

`A` : lowest FPR (when FP is costly)

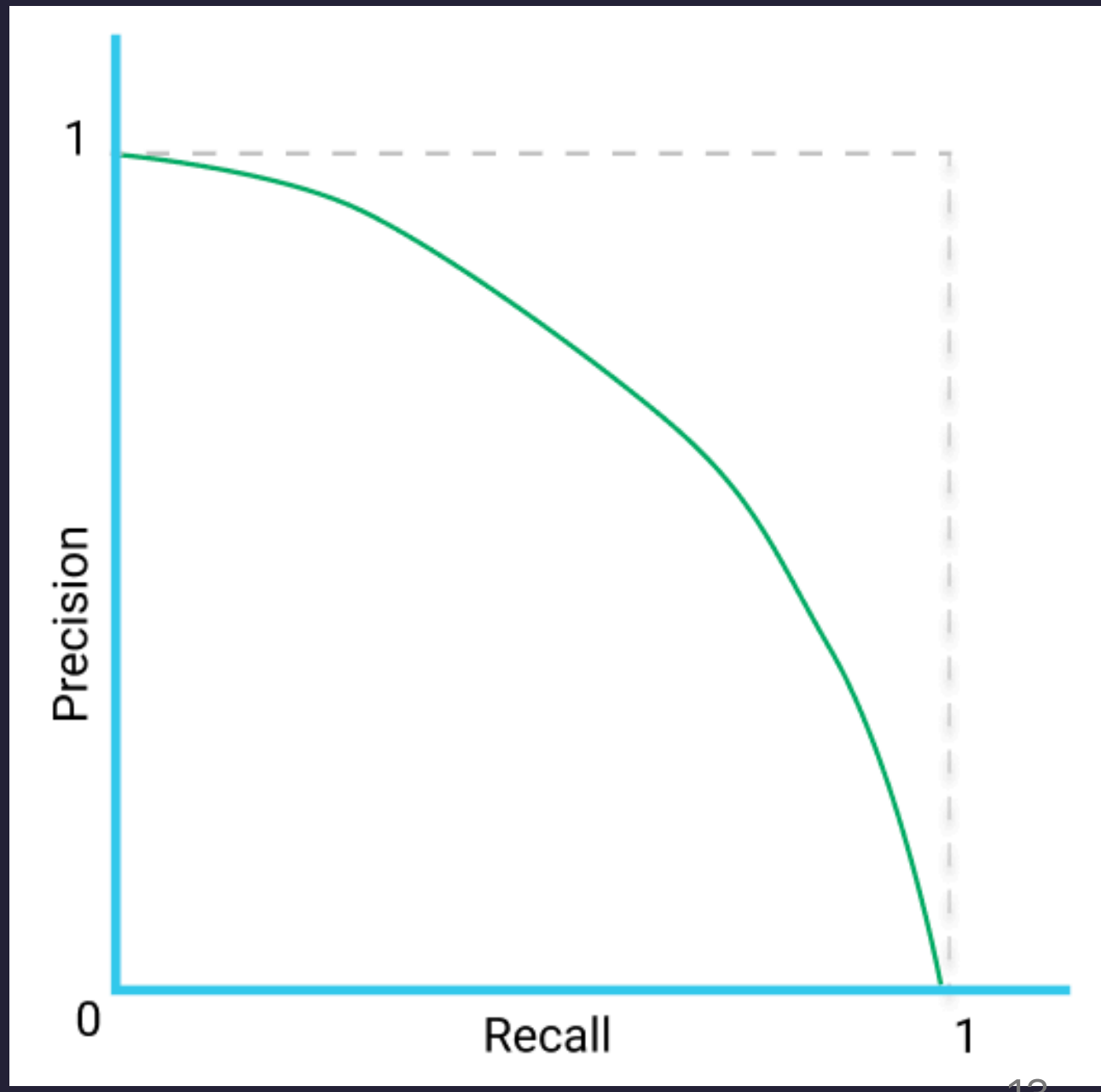
`C` : highest TPR (when FN is costly)



Precision-recall curve (PRC)

For imbalanced datasets, use PRC instead of ROC.

Precision vs. Recall across all thresholds.



Visualizing evaluation metrics

https://developers-dot-devsite-v2-prod.appspot.com/machine-learning/crash-course/classification/roc-and-auc_3689cac9917eb19cc4a8c29c3140b8e30ffacdd8fcfc99df2ec5a1879dbef187.frame