Evaluation metrics for classification problems

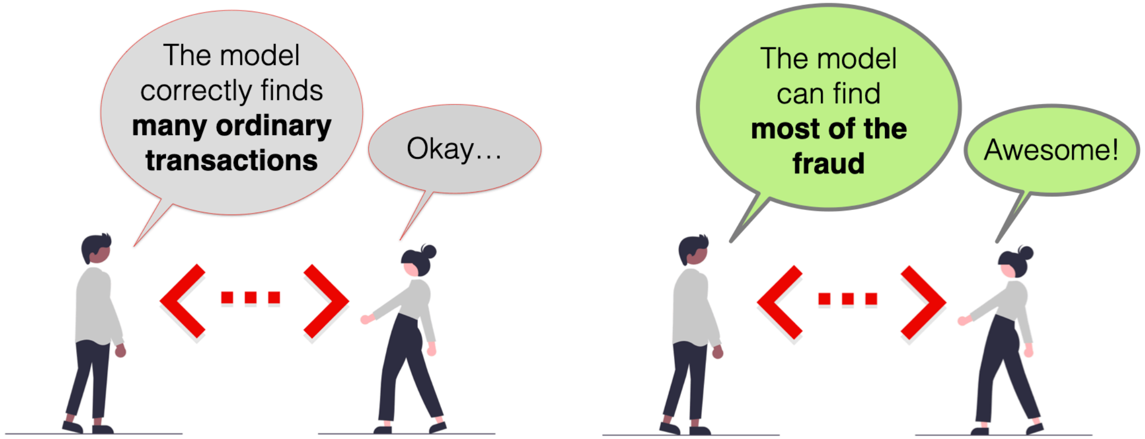
Fraud detection: Is it a fraudulent transaction?

	Predicted	Actual	Correct?
1.	Not fraud	Not fraud	✓
2.	Not fraud	Not fraud	✓
3.	Not fraud	Fraud	X
4.	Fraud	Fraud	✓
. . .			
n.	Fraud	Not fraud	X

	Predicted	Actual
1.	Not fraud	Not fraud
2.	Not fraud	Not fraud
3.	Not fraud	Fraud
4.	Fraud	Fraud
n.	Fraud	Not fraud

True Negative

True Positive



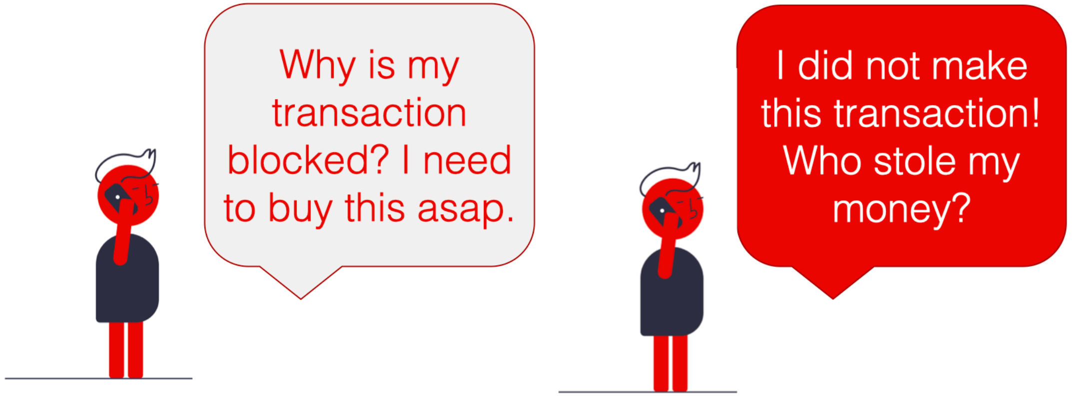
True negative

True positive

	Predicted	Actual
1.	Not fraud	Not fraud
2.	Not fraud	Not fraud
3.	Not fraud	Fraud
4.	Fraud	Fraud
n.	Fraud	Not fraud

False Negative

False Positive

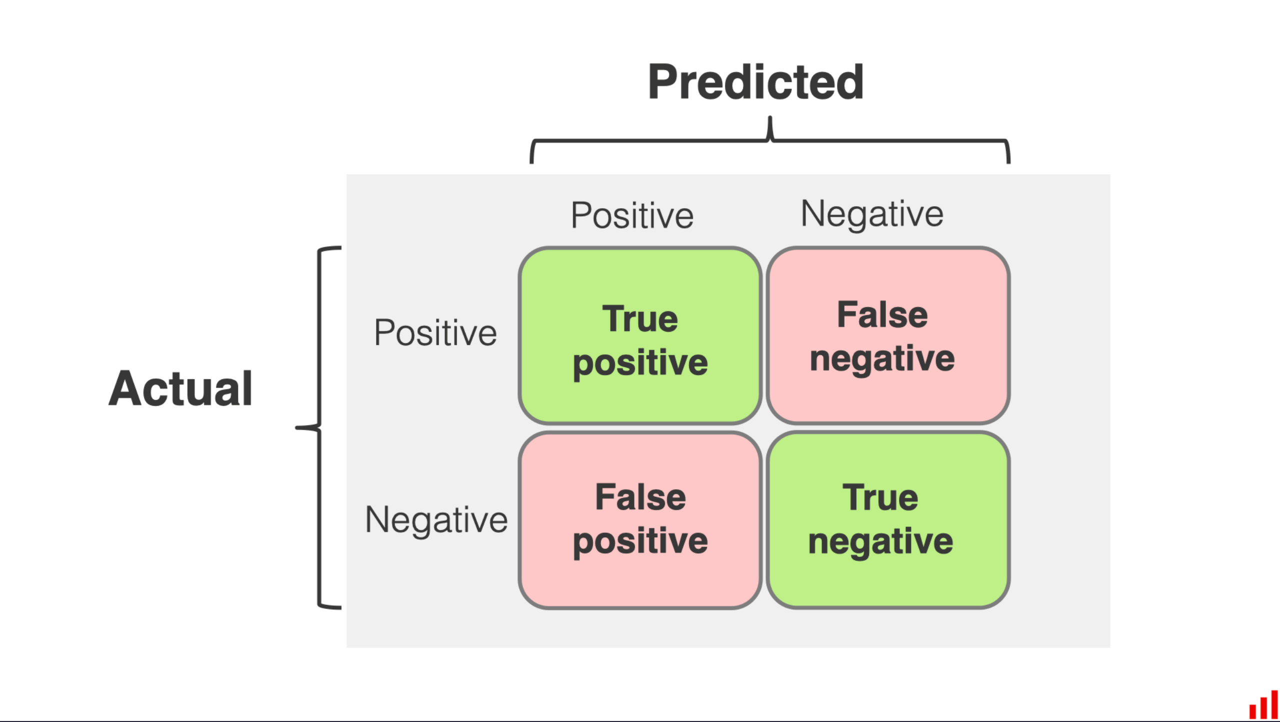


False positive

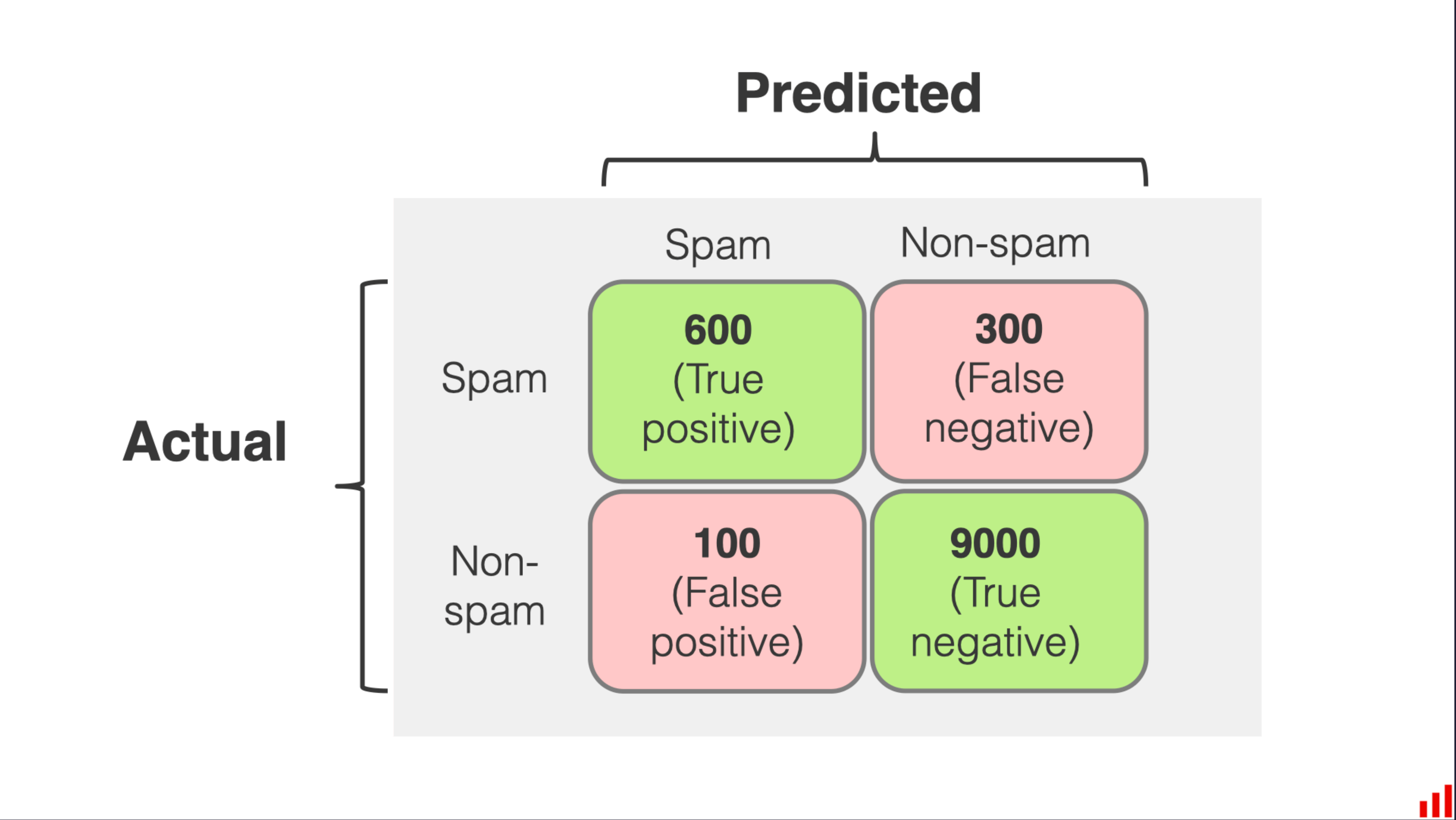
False negative

Confusion matrix

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

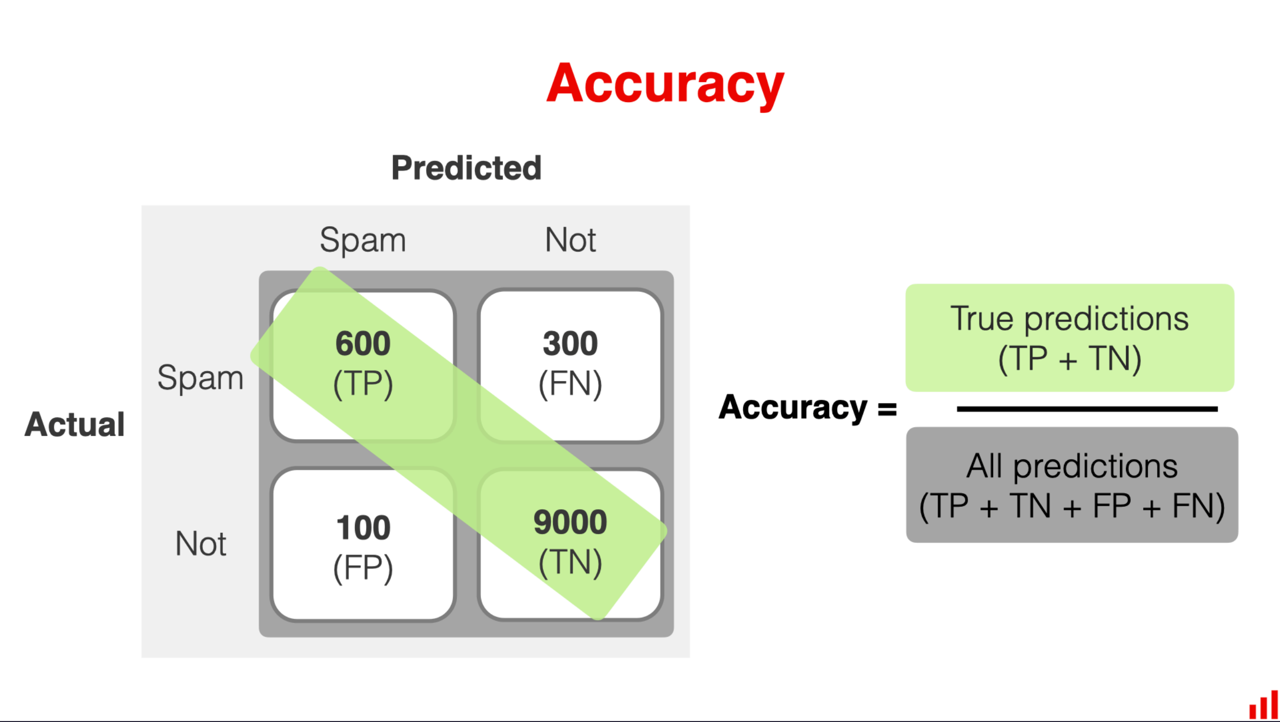


Spam detection: Is it a spam email?



Accuracy

$$\frac{\text{correct predictions}}{\text{total predictions}}$$

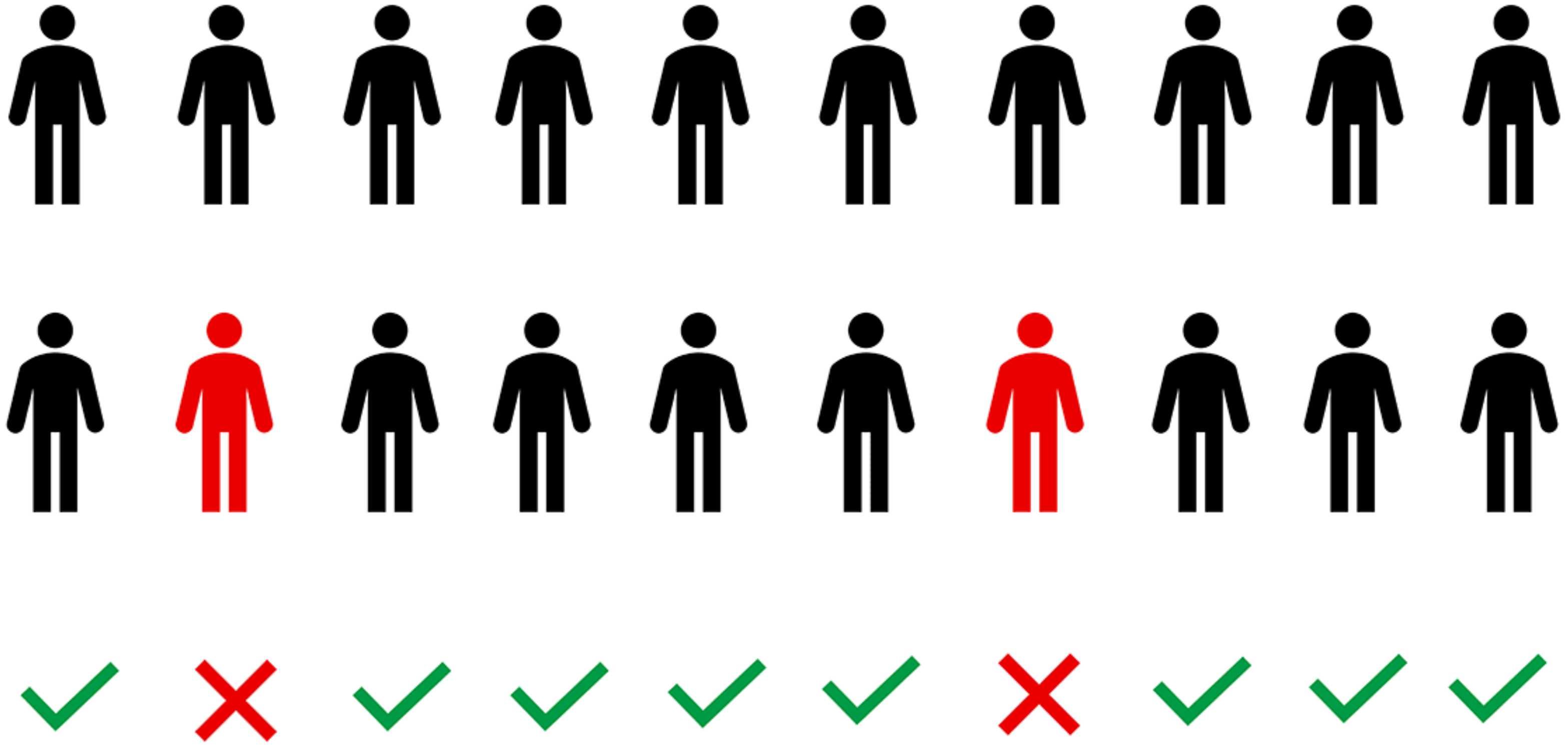


A constant prediction: right in 8 out of 10 cases

Predicted Class

Actual Class

Classification Quality



Accuracy paradox

A model that predicts all negative will have high accuracy on an imbalanced dataset.

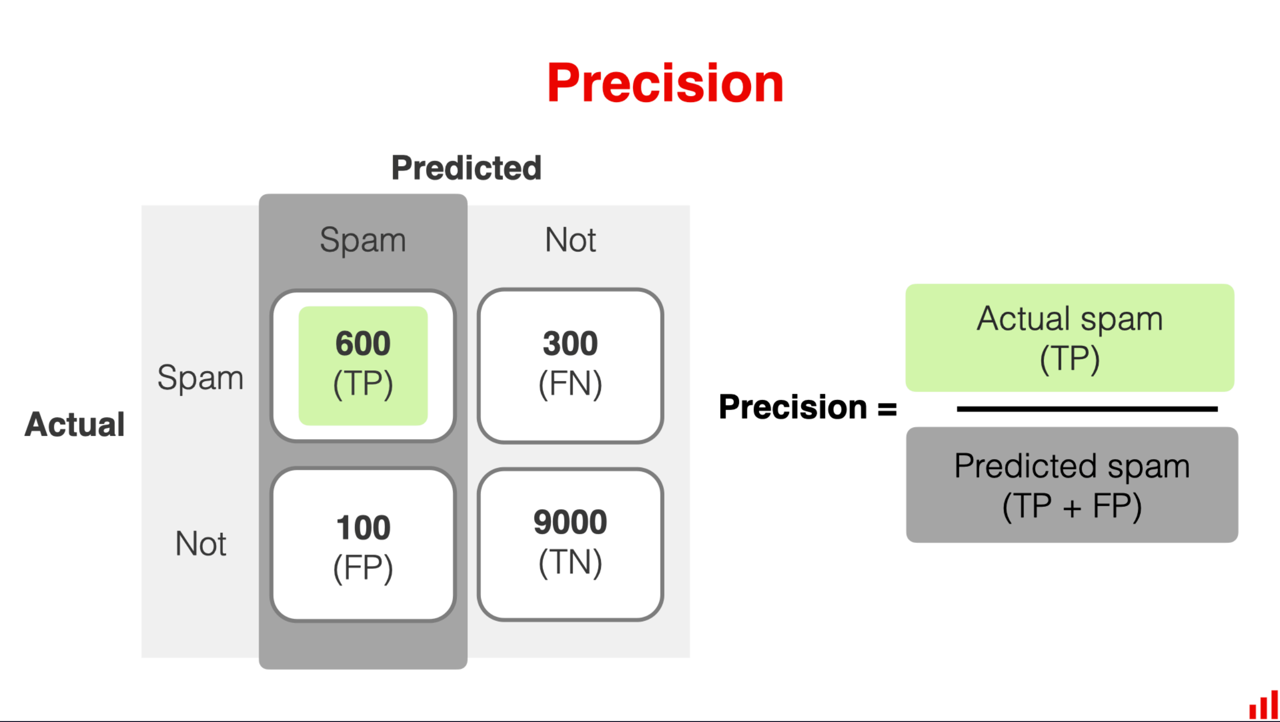
Predicting frequent class (negative samples) adds little value

Precision

$$\frac{\text{correctly predicted positive}}{\text{predicted positive}}$$

Use when FP is costly

- FP ⬆️, Precision ⬇️

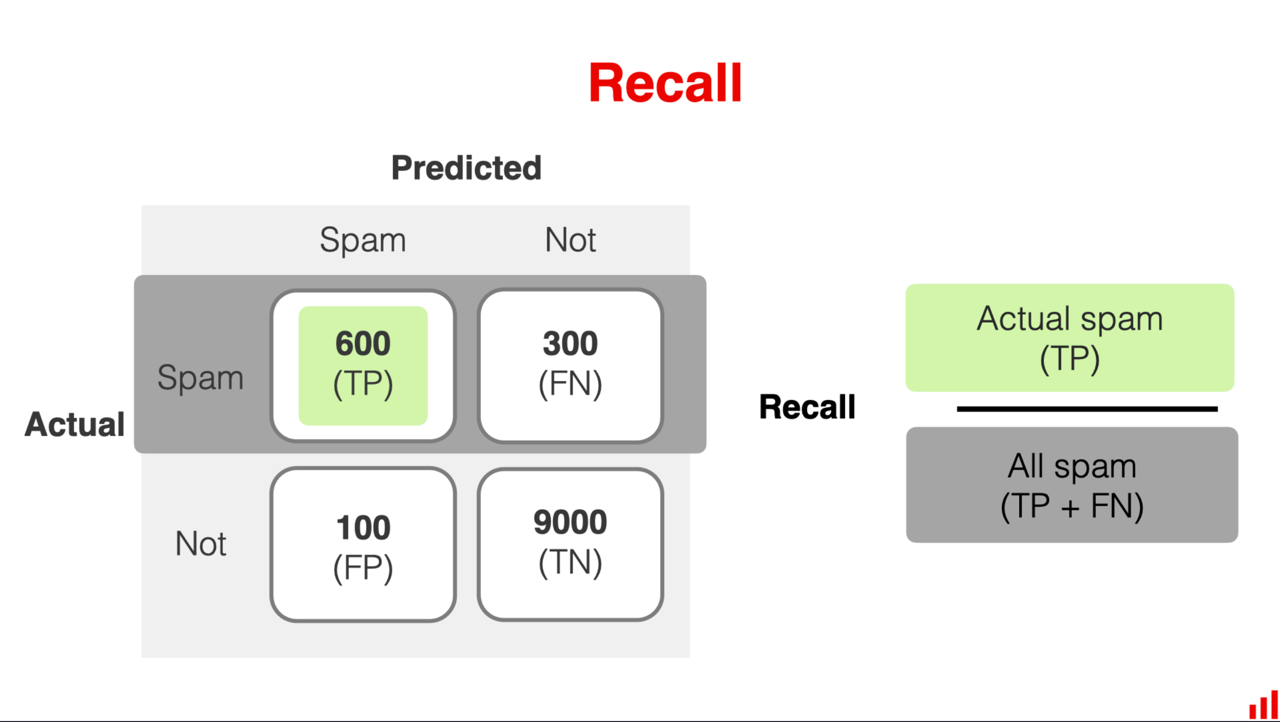


Recall

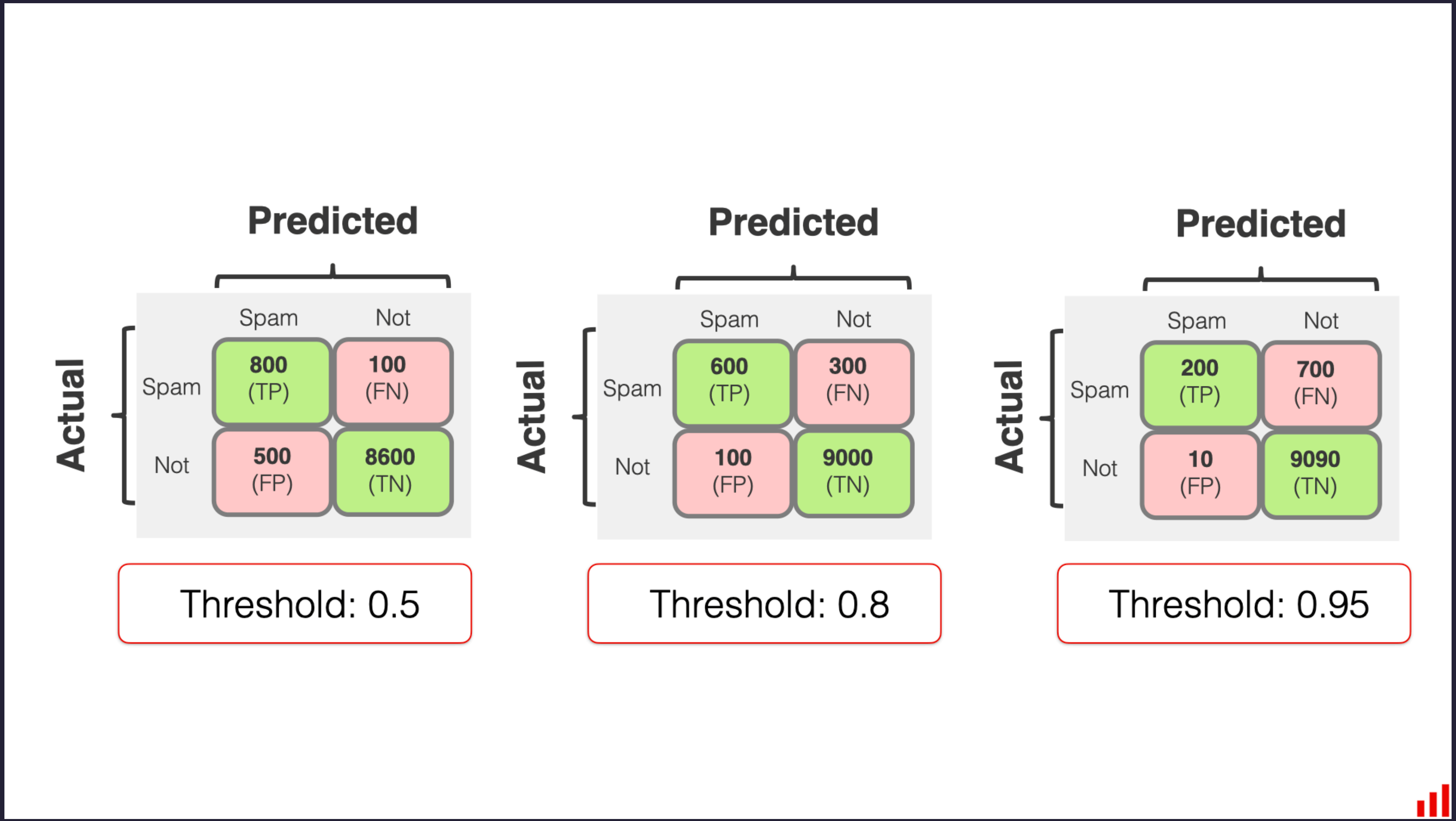
$$\frac{\text{correctly predicted positive}}{\text{actual positive}}$$

Use when FN is costly

- FN ⬆️, Recall ⬇️



Classification threshold changes the confusion matrix



Precision-recall tradeoff

Increase threshold:

- More conservative: fewer positive predictions, but mostly right
- When in doubt, predict negative (y=0)
- Precision ⬆️, Recall ⬇️

Decrease threshold:

- More aggressive: more positive predictions, but more mistakes
- When in doubt, predict positive (y=1)
- Precision ⬇️, Recall ⬆️

Visualizing the confusion matrix

https://developers-dot-devsite-v2-prod.appspot.com/machine-learning/crash-course/classification/accuracy-precision-recall_48d642036b6a12f05752bd92fcdf132b3f73d1d61a0897727e060c27ed347370.frame

Optimize for precision or recall

FP is costly: optimize for precision

FN is costly: optimize for recall

Fraud detection:

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

Propensity to buy:

- FP: non-buyer is classified as buyer
- FN: buyer is classified as non-buyer

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.

💻 Validate on different evaluation metrics

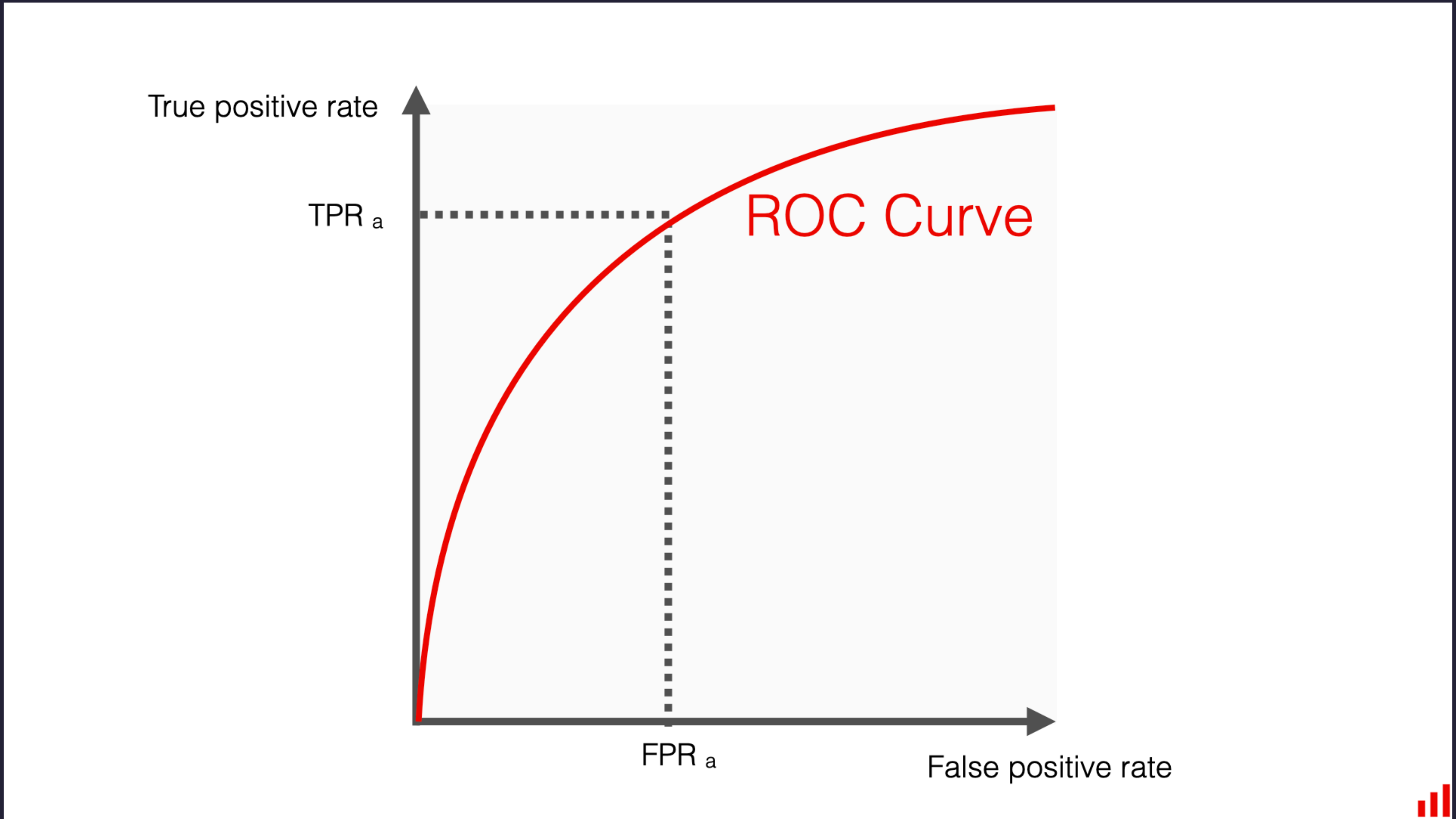
ROC and ROC-AUC

ROC: Receiver-operating characteristic curve

- shows the performance of a binary classifier with ***different thresholds***

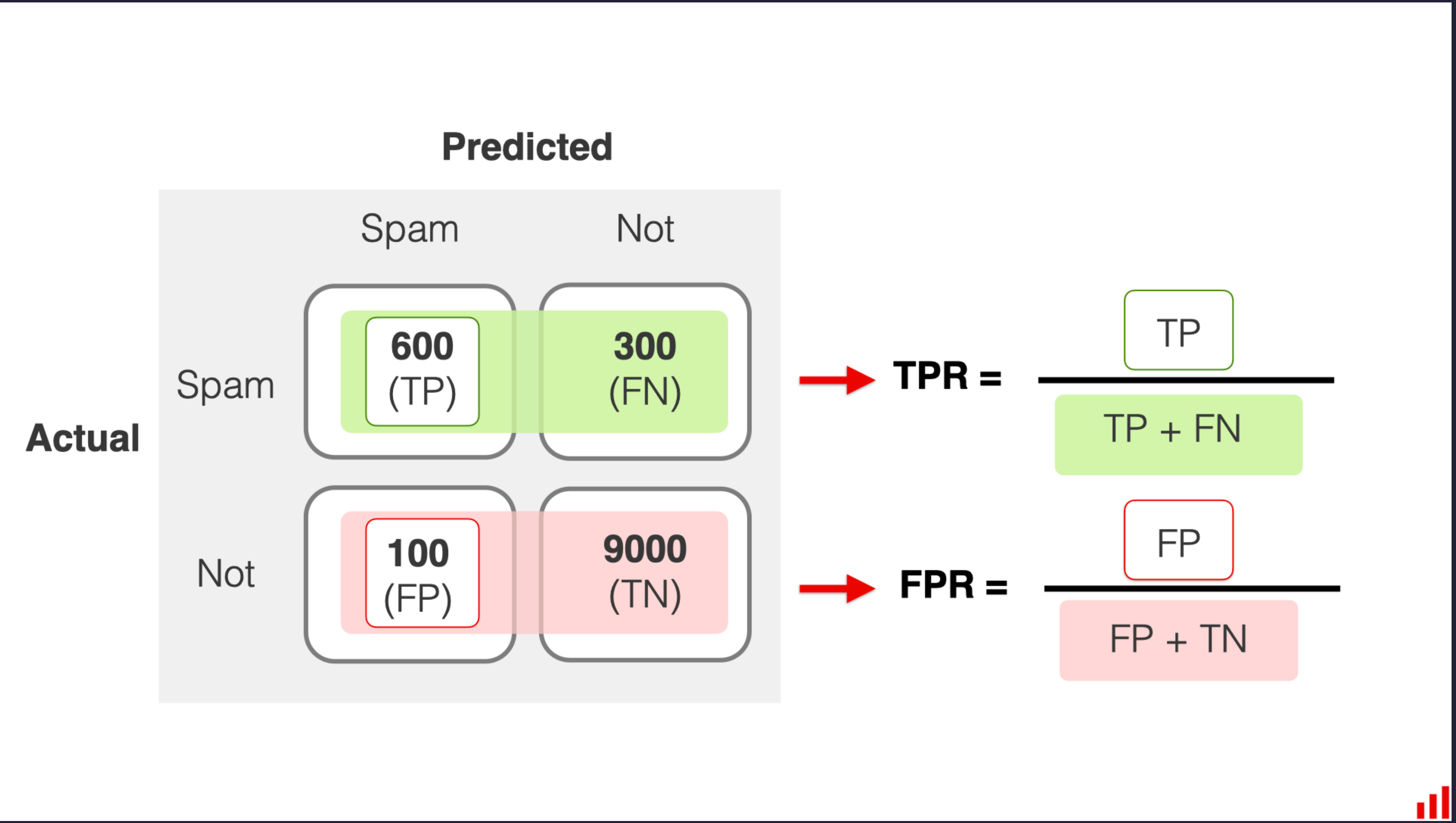
ROC-AUC: Area under the ROC curve

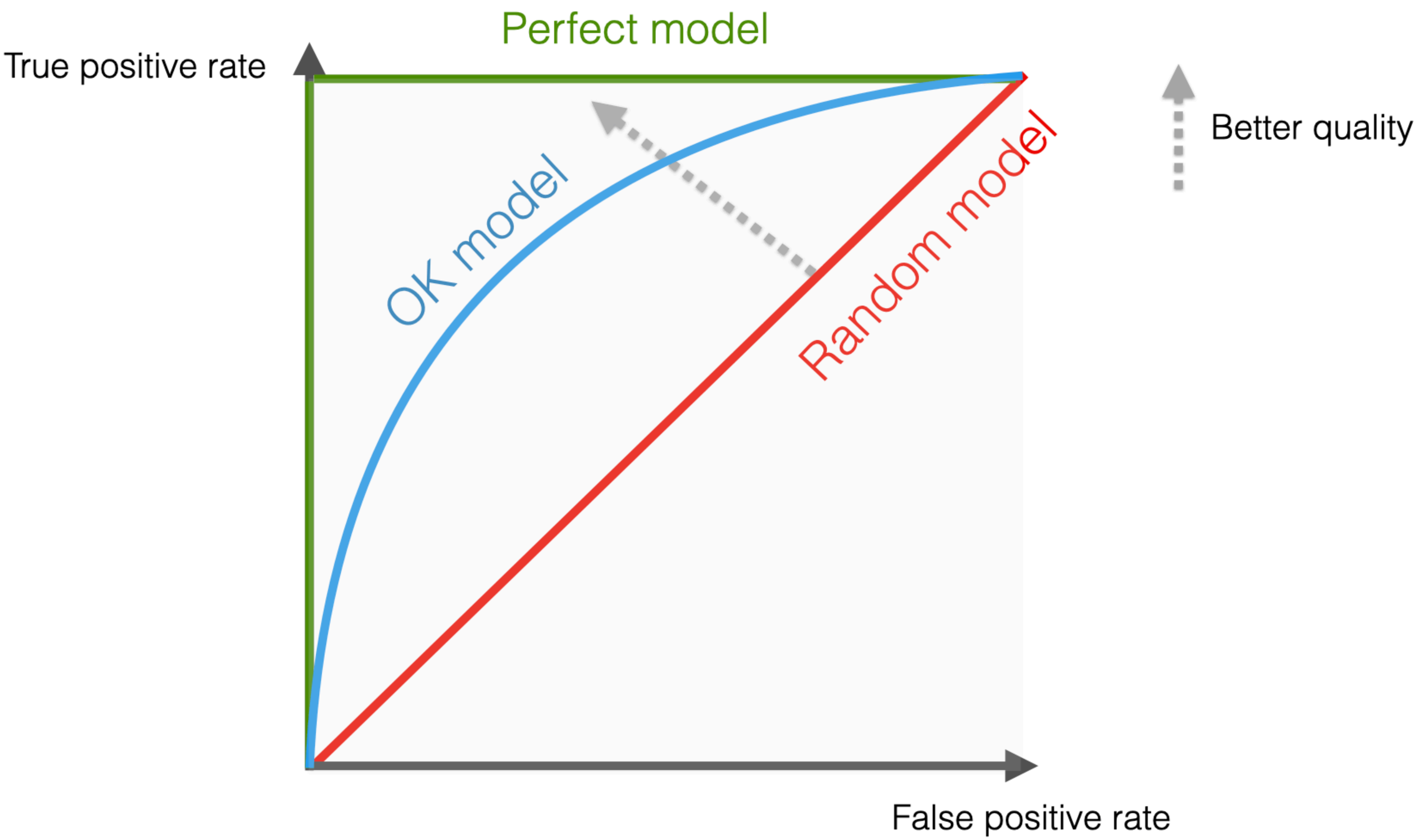
- Relative scores to discriminate between positive or negative instances across all classification thresholds.

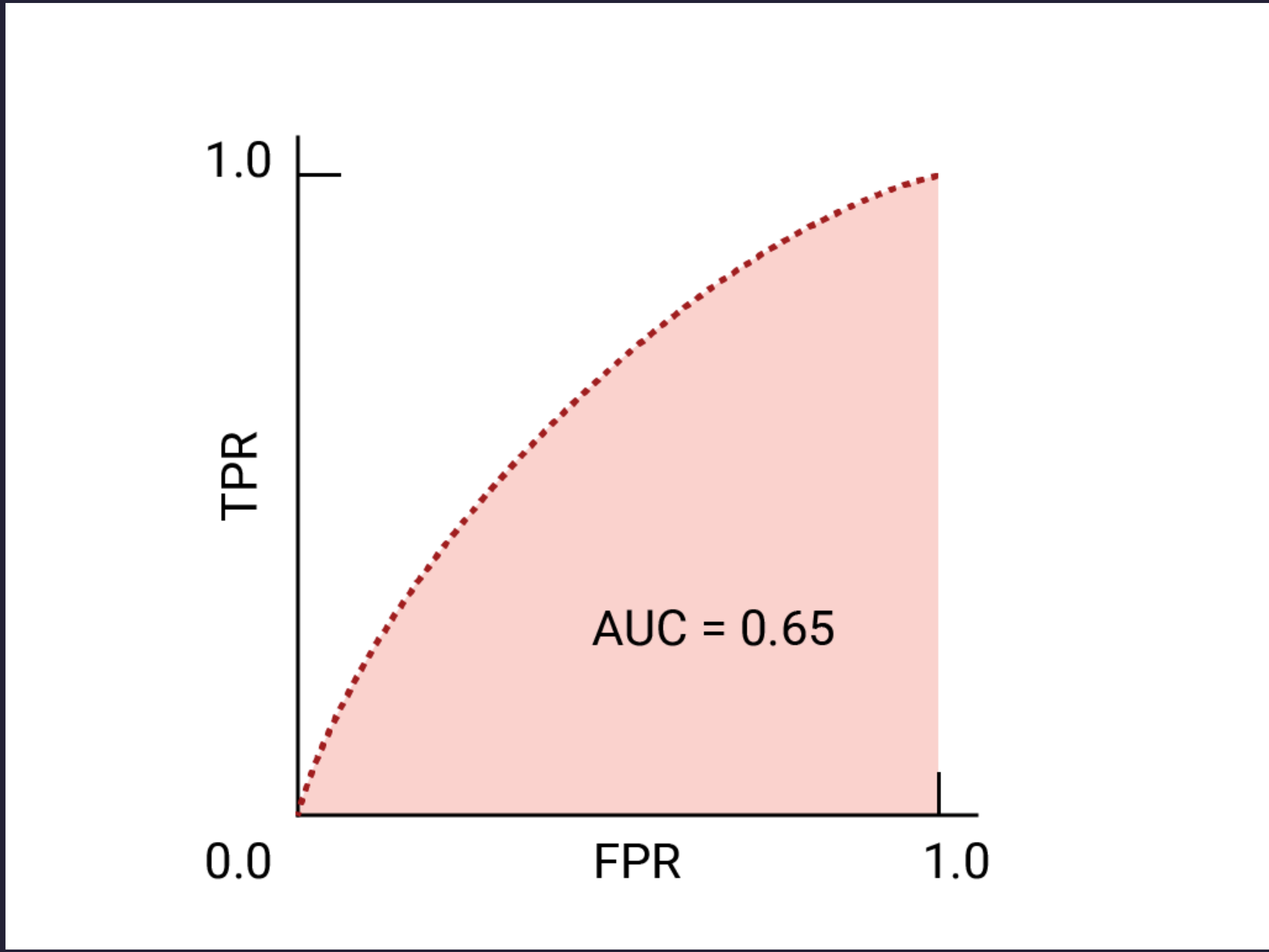


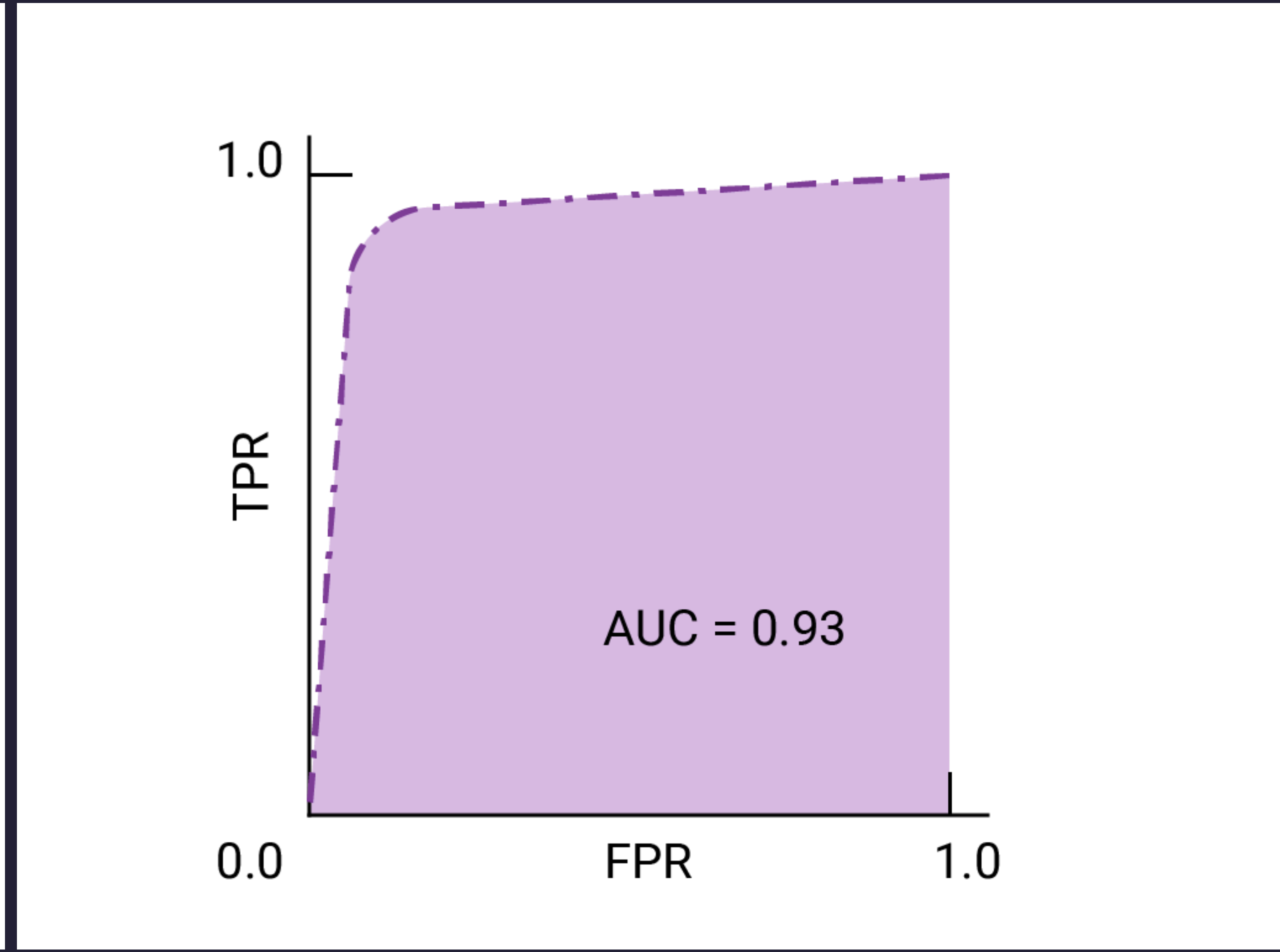
https://developers-dot-devsite-v2-prod.appspot.com/machine-learning/crash-course/classification/roc-and-auc_3689cac9917eb19cc4a8c29c3140b8e30ffacdd8fcfc99df2ec5a1879dbef187.frame

TPR and FPR


Predicted
Spam
Not
Actual
Spam
600
(TP)
300
(FN)
Not
100
(FP)
9000
(TN)
TPR =
TP
TP + FN
FPR =
FP
FP + TN


Perfect model
True positive rate
OK model
Random model
Better quality
False positive rate


1.0
TPR
AUC = 0.65
0.0
FPR
1.0


1.0
TPR
AUC = 0.93
0.0
FPR
1.0

When to use ROC-AUC

ROC-AUC: a single metric that summarizes the model performance across all thresholds

Useful:

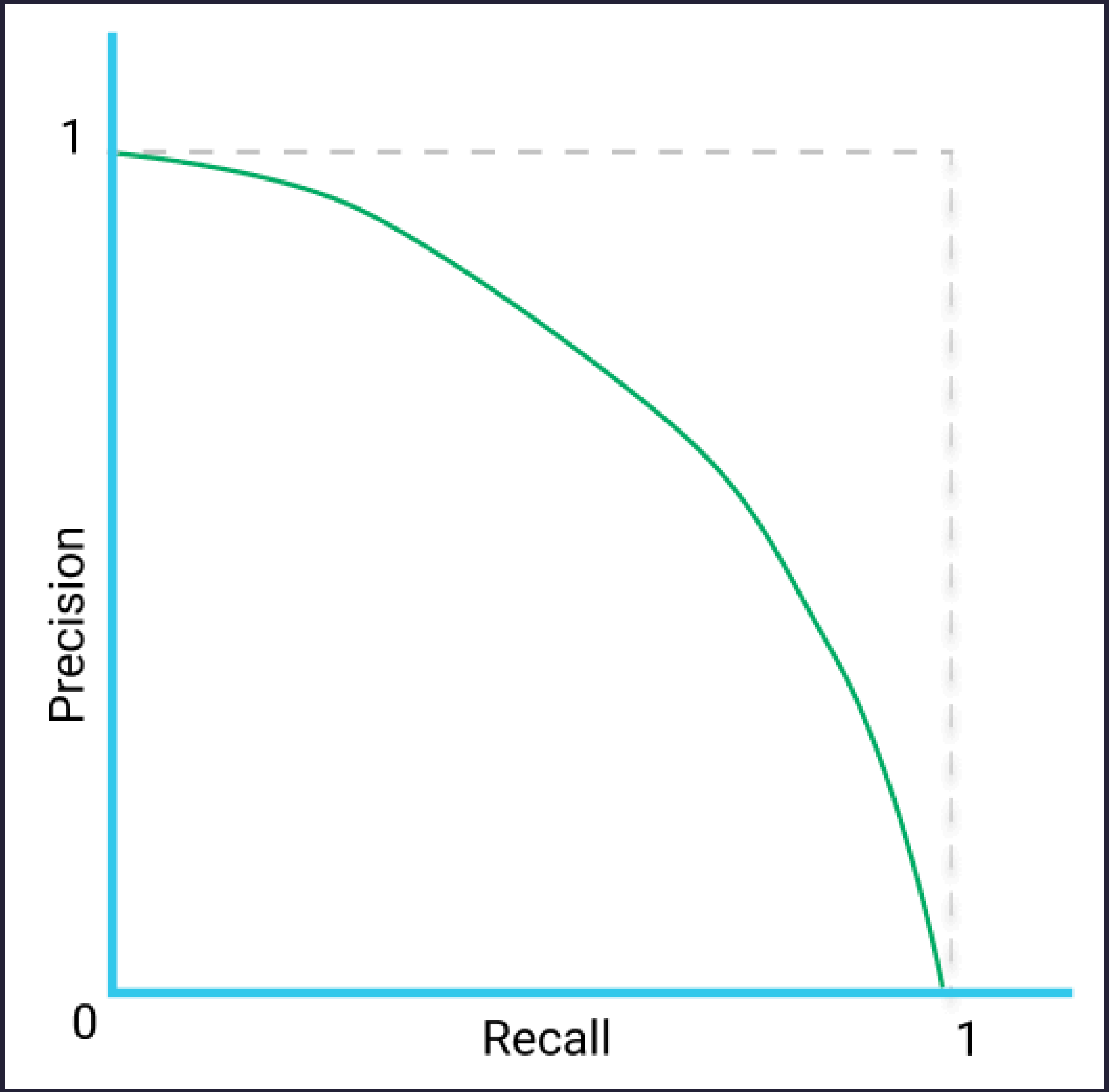
- for model comparison
- when the costs of errors are similar
- when the data is balanced

Less useful :

- when you care about different costs of error
- when the data is heavily imbalanced

Precision-recall (PR) curve

- For imbalanced datasets, use PR curve instead of ROC curve
- Precision vs. Recall across all thresholds
- Precision-recall tradeoff
- PR AUC: Area under the PR curve



💻 ROC and PR curves