CSCI 467: Machine Learning

USC

Discussion – Cross-Validation and Evaluation Metrics

Wang (Bill) Zhu

Sep 2023

Materials from: *https://developers.google.com/machine-learning/crash-course*

Cross-Validation Overview

- Training and Test Sets
- Validation Set
- Cross-Validation

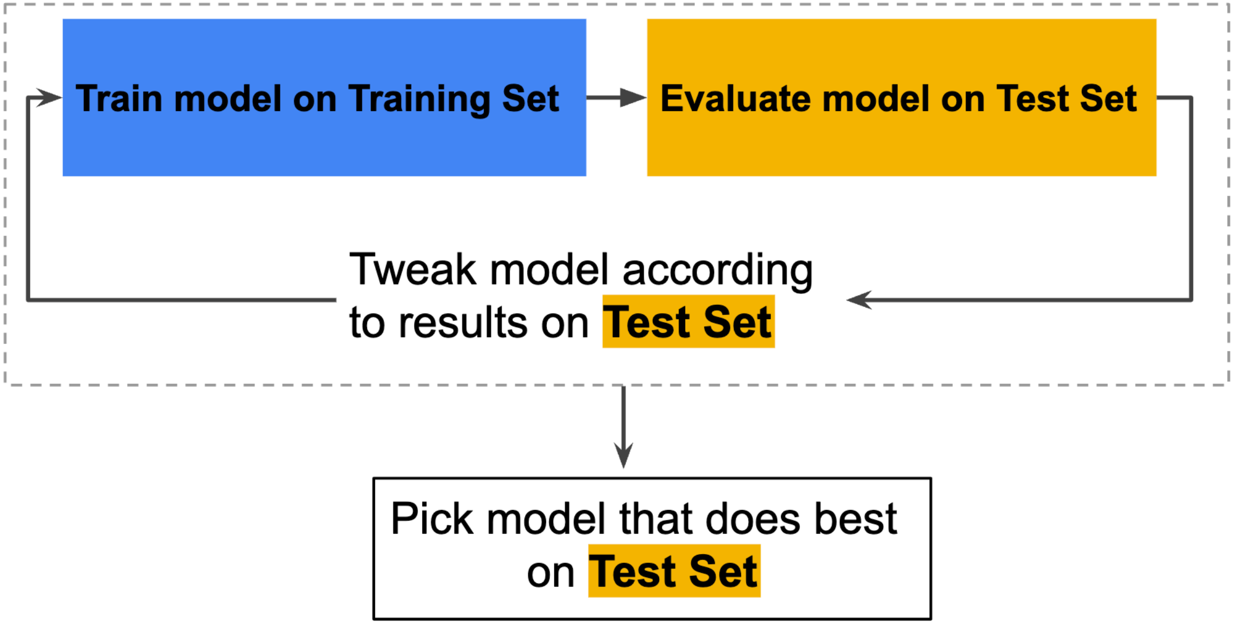
Training and Test Sets

- Training set - a subset to train a model.
- Test set – a subset to test a trained model

You could imagine slicing the single data set as follows (80%/20%):

Training and Test Sets

- With two partitions, the workflow could look as follows (may overfit the test set)

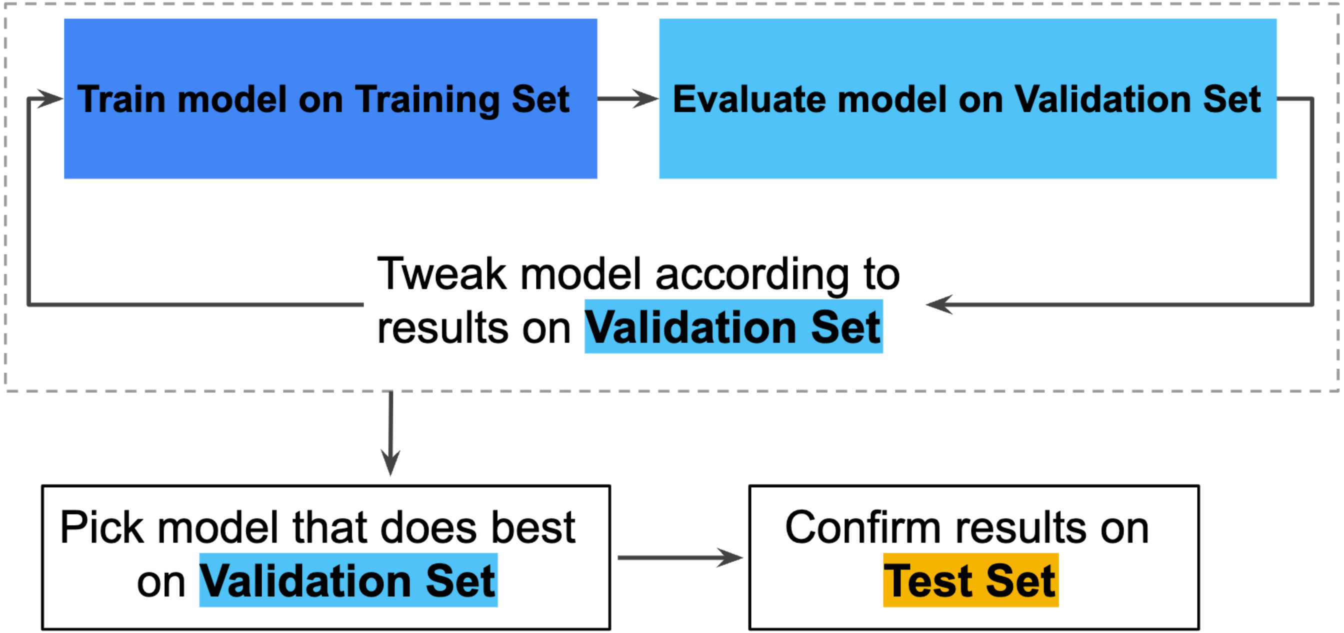


Validation Set

- You can greatly reduce your chances of overfitting by partitioning the data set into the three subsets shown in the following figure.
- Use the **validation set to evaluate results** from the training set. Then, use the **test set to double-check your evaluation** after the model has "passed" the validation set. (exam analogy: Lectures, HWs, Finals)

Validation Set

- Tune hyper-parameters (batch size, learning rate, etc.) on the validation set



Cross-Validation

- You need the validation set to be large (avoid overfitting)
- You need the validation set to be small (to have enough training data)

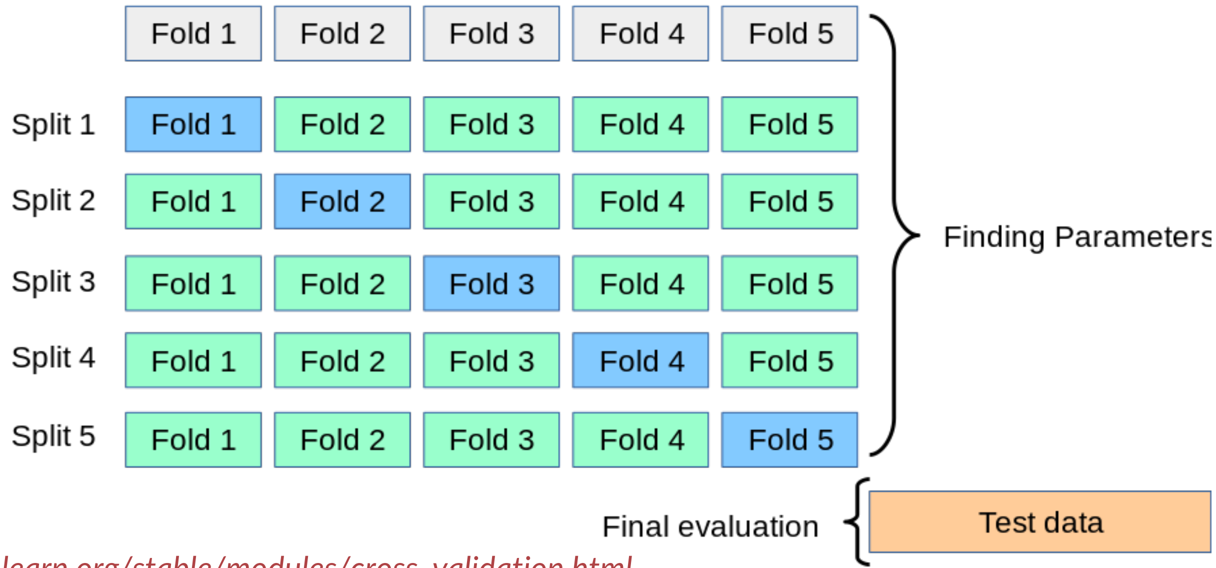


Image from: https://scikit-learn.org/stable/modules/cross_validation.html

Cross-Validation

- Split the data into k fold, use (k-1) fold for training and 1 fold for validation
- After finalizing hyper-parameters, use the entire training+validation to train the model

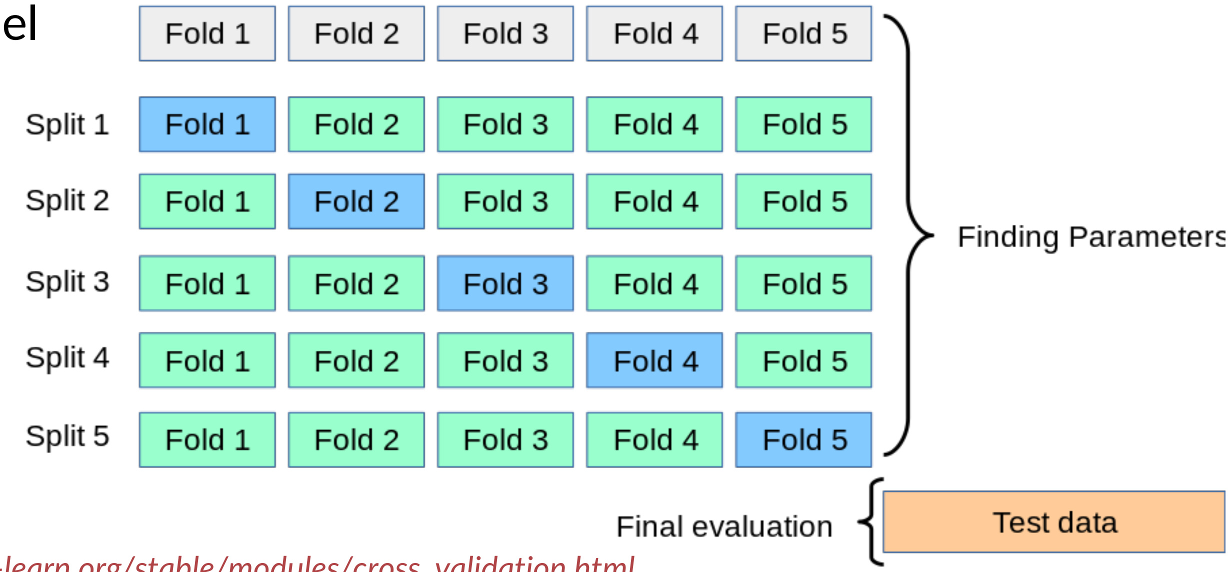


Image from: https://scikit-learn.org/stable/modules/cross_validation.html

Evaluation Metrics Overview

- Thresholding
- Confusion matrix
- Accuracy
- Precision and Recall
- ROC and AUC
- Calibration

Thresholding

- Binary classification: $y = f(x), y \in \{0, 1\}$
- A logistic regression model outputs a probability in $(0, 1)$
- Choose a threshold to convert it to a binary value
- 0.5 is not always the best
- *Why? Depends on the evaluation metrics.*

Confusion Matrix – Tumor Prediction

- Use 2x2 confusion matrix to separate out different kinds of errors
- Class-imbalanced setup: 9% of examined tumors are malignant, 91% benign

True Positives (TP) Reality: Malignant ML predicted: Malignant	**False Positives (FP)** Reality: Benign ML predicted: Malignant Type-1 Error
False Negatives (FN) Reality: Malignant ML predicted: Benign Type-2 Error	**True Negatives (TN)** Reality: Benign ML predicted: Benign

Evaluation Metrics: Accuracy - Can Be Misleading

- Accuracy is the fraction of predictions our model got right

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Evaluation Metrics: Accuracy - Can Be Misleading

- Accuracy is the fraction of predictions our model got right
- Accuracy $= \frac{TP+TN}{TP+FP+FN+TN}$

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Evaluation Metrics: Accuracy - Can Be Misleading

- Accuracy is the fraction of predictions our model got right
- $\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$
- How about a model that predicts negative all the time?

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Exercise (2 mins)

In which of the following scenarios would suggest that the ML model is doing a good job?

A. A deadly, but curable, medical condition afflicts .01% of the population. An ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%.

B. An expensive robotic chicken crosses a very busy road a thousand times per day. An ML model evaluates traffic patterns and predicts when this chicken can safely cross the street with an accuracy of 99.99%.

C. In the game of roulette, a ball is dropped on a spinning wheel and eventually lands in one of 38 slots. Using visual features (the spin of the ball, the position of the wheel when the ball was dropped, the height of the ball over the wheel), an ML model can predict the slot that the ball will land in with an accuracy of 50%.

Evaluation Metrics: Precision and Recall

- What proportion of positive identifications was actually correct?
- Precision $= \frac{TP}{TP+FP}$

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Evaluation Metrics: Precision and Recall

- What proportion of positive identifications was actually correct?
- Precision $= \frac{TP}{TP+FP}$
- 0.5

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Evaluation Metrics: Precision and Recall

- What proportion of actual positives was identified correctly?
- Recall $= \frac{TP}{TP+FN}$

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

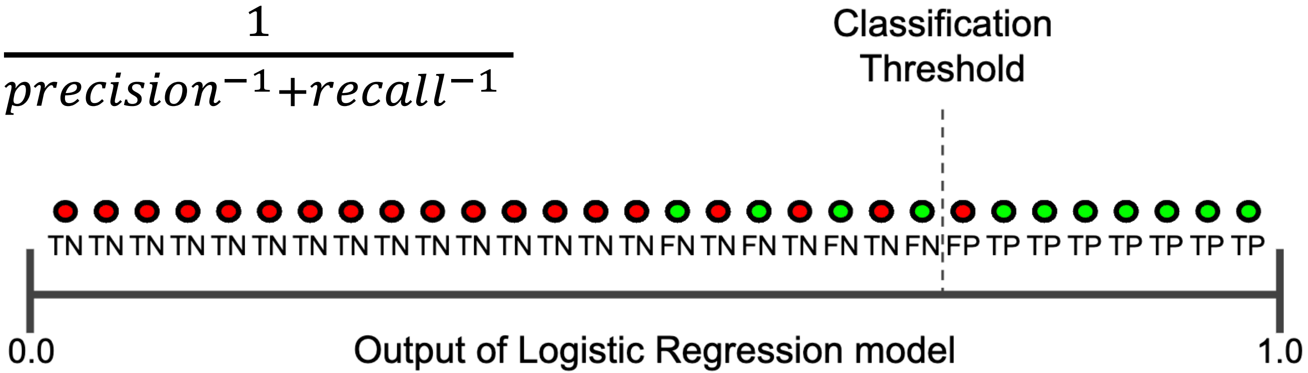
Evaluation Metrics: Precision and Recall

- What proportion of actual positives was identified correctly?
- Recall $= \frac{TP}{TP+FN}$
- 0.11

True Positives (TP) Reality: Malignant ML predicted: Malignant Number of TP results: 1	**False Positives (FP)** Reality: Benign ML predicted: Malignant Number of FP results: 1
False Negatives (FN) Reality: Malignant ML predicted: Benign Number of FN results: 8	**True Negatives (TN)** Reality: Benign ML predicted: Benign Number of TN results: 90

Precision and Recall: A Tug of War

- Cannot improve both at the same time by changing threshold
- Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$
- F1 $= \frac{1}{precision^{-1}+recall^{-1}}$



Exercise (2 min)

Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to precision?

A. Probably increase.

B. Probably decrease.

C. Definitely increase.

D. Definitely decrease.

Consider two models—A and B—that each evaluate the same dataset. Which one of the following statements is true?

A. If model A has better recall than model B, then model A is better.

B. If model A has better precision and better recall than model B, then model A is probably better.

C. If Model A has better precision than model B, then model A is better.

Recap: HW1 Q2.3

- What if FP has a cost of 1 but FN has a cost of 3? Decision boundary still 0.5? (Hint: insert the costs into the loss function)

2.3 (2pts) When $F(x; A, k, b)$ is a CDF, we can interpret $F(x; A, k, b)$ as the probability of x belonging to the class 1 (in a binary classification problem). Suppose we know $F(x; A, k, b)$ and want to predict the label of a datapoint x. We need to decide on a threshold value for $F(x; A, k, b)$, above which we will predict the label 1 and below which we will predict -1. Show that setting the threshold to be $F(x; A, k, b) \geq 1/2$ minimizes the classification error.

Ans. Expected error for a given threshold (T) for a given x is

$$\begin{aligned}\text{Expected loss given } x &= p(y=0|x)I[F(x) \geq T] + p(y=1|x)I[F(x) < T] \\ &= [1 - F(x)]I[F(x) \geq T] + [F(x)]I[F(x) < T]\end{aligned}$$

The above is minimized at $T = 0.5$.

Recap: HW1 Q2.3

- What if FP has a cost of 1 but FN has a cost of 3? Decision boundary still 0.5? (Hint: insert the costs into the loss function)
- $1 - T = 3T, T = 0.25$

2.3 (2pts) When $F(x; A, k, b)$ is a CDF, we can interpret $F(x; A, k, b)$ as the probability of x belonging to the class 1 (in a binary classification problem). Suppose we know $F(x; A, k, b)$ and want to predict the label of a datapoint x. We need to decide on a threshold value for $F(x; A, k, b)$, above which we will predict the label 1 and below which we will predict -1. Show that setting the threshold to be $F(x; A, k, b) \geq 1/2$ minimizes the classification error.

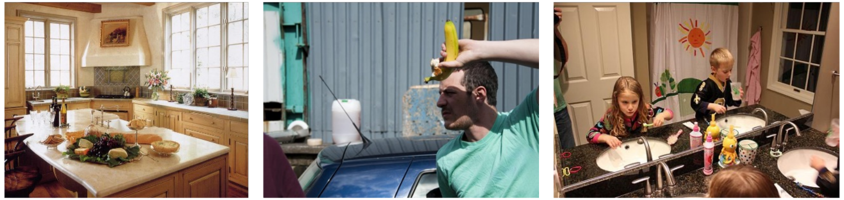
Ans. Expected error for a given threshold (T) for a given x is

$$\begin{aligned}\text{Expected loss given } x &= p(y=0|x)I[F(x) \geq T] + p(y=1|x)I[F(x) < T] \\ &= [1 - F(x)]I[F(x) \geq T] + [F(x)]I[F(x) < T]\end{aligned}$$

The above is minimized at $T = 0.5$.

Application: Contrast set

*Q: Is there **at least** 1 image with exactly 2 dark bottles on a counter.*



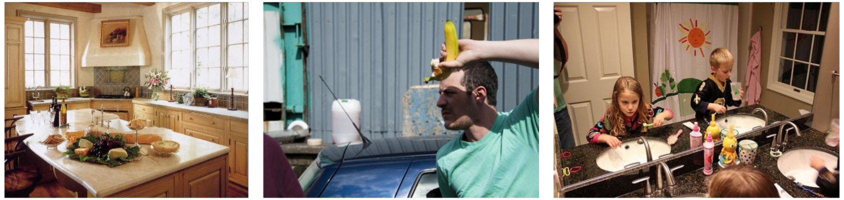
Expected A: True

Vision-Language Model

Acc: 88

Application: Contrast set

Q: Is there ***at least*** *1 image with exactly 2 dark bottles on a counter.*



Contrast Q: Is there ***less than*** *1 image with exactly 2 dark bottles on a counter.*

Expected A: True

Vision-Language Model

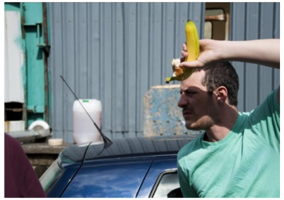
Acc: 88

Acc: 21

Expected A: False

Application: Contrast set

*Q: Is there **at least** 1 image with exactly 2 dark bottles on a counter.*


*Contrast Q: Is there **less than** 1 image with exactly 2 dark bottles on a counter.*

Expected A: True

Vision-Language Model

Acc: 88

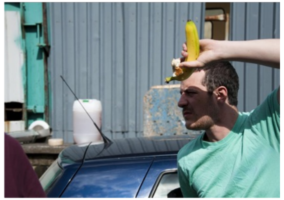
Acc: 21

Expected A: False

What does this tell us? Contrast Qs are hard? They have low correlation/grounding on images? The VL model is bad?

Application: Contrast set

*Q: Is there **at least** 1 image with exactly 2 dark bottles on a counter.*

*Contrast Q: Is there **less than** 1 image with exactly 2 dark bottles on a counter.*

Expected A: True

Vision-Language Model

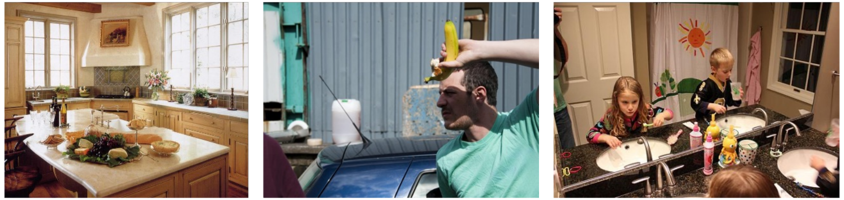
Expected A: False

TP: 80	**FP:** 11
FN: 25	**TN:** 188

TP: 20	**FP:** 83
FN: 157	**TN:** 44

Application: Contrast set

*Q: Is there **at least** 1 image with exactly 2 dark bottles on a counter.*



*Contrast Q: Is there **less than** 1 image with exactly 2 dark bottles on a counter.*

Expected A: True

Vision-Language Model

Expected A: False

TP: 80	**FP:** 11
FN: 25	**TN:** 188

TP: 20	**FP:** 83
FN: 157	**TN:** 44

What does this tell us? (Probably) the model is over-stable on its prediction.

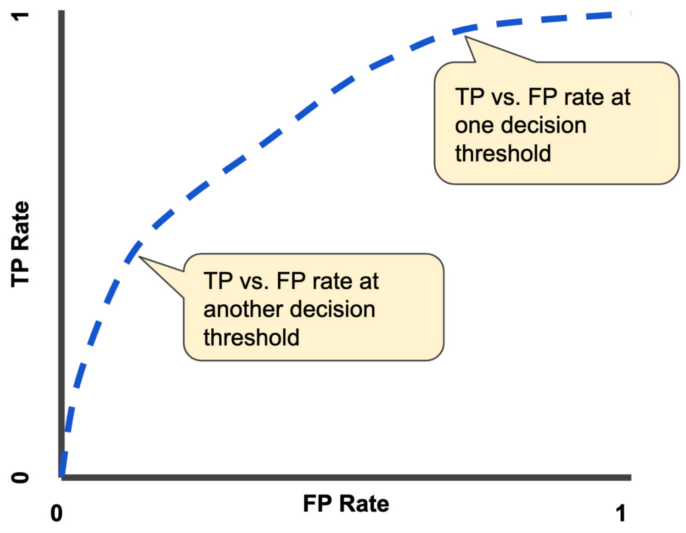
Example from: https://arxiv.org/abs/2210.15037

A ROC Curve

- Each point is the TP and FP rate at one decision threshold

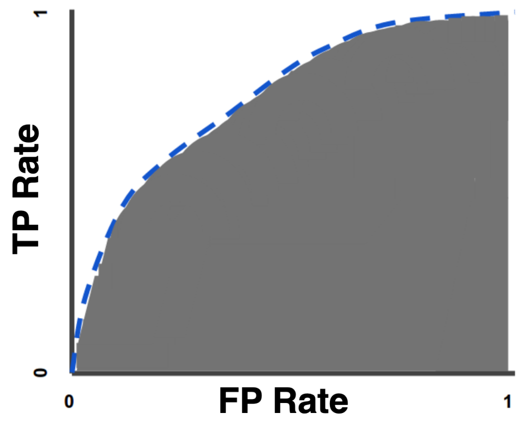
- TPR (Recall) $= \dfrac{TP}{TP+FN}$

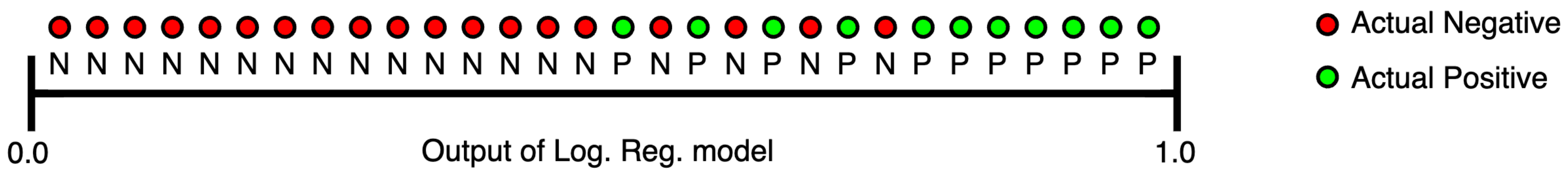
- FPR $= \dfrac{FP}{FP+TN}$



Evaluation Metrics: AUC (AUROC)

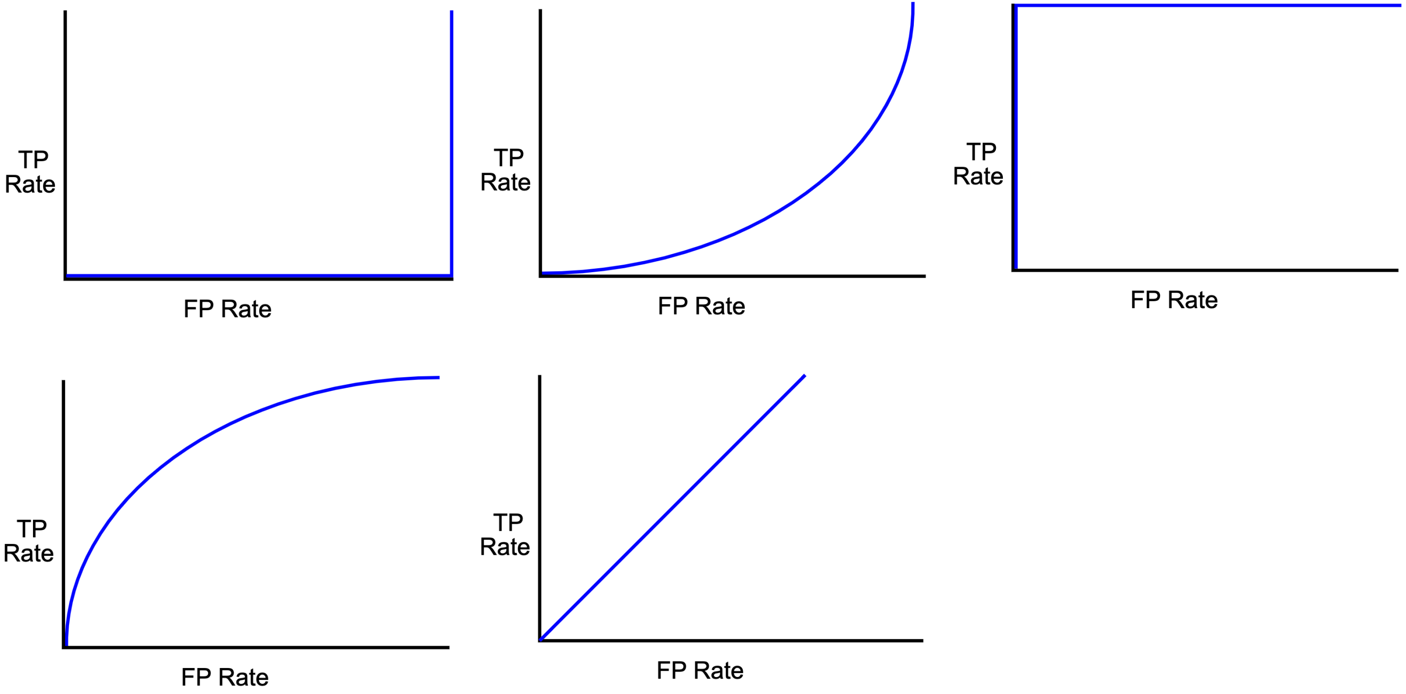
- AUC: “Area under the ROC Curve”
- The probability that the model ranks a random positive example more highly than a random negative example
- Independent of the threshold





Exercise (2 mins)

Which of the following ROC curves produce AUC values greater than 0.5?



Calibration

- Prediction bias = average of prediction - average of labels
- Zero bias alone does not mean everything is perfect
- It's a great sanity check: incomplete features? noisy data? buggy pipeline?
- Don't fix bias with a calibration layer, fix it in the model