AI and Deep Learning

2-6 Evaluation of Binary Classification Models

Zhonglei Wang

WISE, SOE, CHOW

XMU

(Version v1)

Contents

Recall

1. Binary classification models
   - Logistic regression model
   - Support vector machine
   - Random forest
   - Neural network models
2. For most models,
   - A certain score is estimated
   - A certain threshold is used to determine the estimated class

Recall

1. We use a training dataset to learn a model
2. Test dataset is used to evaluate the performance of the learnt model
3. We consider the case that a model has estimated labels for each example in the test dataset

Confusion matrix

1. Among the true binary responses in the test dataset,
   - True positive number: Number of true positive responses
   - True negative number: Number of true negative responses
2. A certain binary classification model predicts responses for the test dataset,
   - Predicted positive number: Number of predicted positive responses
   - Predicted negative number: Number of predicted negative responses

Confusion matrix

1. There are four possibilities

- TP (True Positive): Number of correctly estimated positives
- FN (False Negative): Number of falsely estimated negatives

  (should be positives but estimated as negatives)
- TN (True Negative): Number of correctly estimated negatives
- FP (False Positive): Number of falsely estimated positives

  (should be negative but estimated as positives)

Confusion matrix

1. Confusion matrix

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

Measures--Accuracy

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Accuracy

$$\begin{aligned}\text{Accuracy} &= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN}\end{aligned}$$

2. Remarks

- Accuracy is an overall performance metric, and it does not differentiate between positives and negatives
- It is a good metric for balanced datasets
- However, it is almost useless or misleading for unbalanced datasets

Measures--Precision

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Precision

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{TP}{TP + FP}$$

2. Remark

- Focus on the reliability for positive estimation results, and try to control false positives (type I error)
- When the cost of false positive alarms is high, precision is a key measure
- A more suitable measure for unbalanced datasets

Measures--Precision

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

3. Disadvantages

- It ignores the impact of false negative alarms, so it is less sensitive to that
- Not suitable for unbalanced datasets and our focus is negatives
- It is not a comprehensive measure and should be used with others, such as recall

Measures--Recall

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Recall

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{TP}{TP + FN}$$

2. Remark
   - Recall is also called True Positive Rate (TPR) or sensitivity
   - When the cost of false negatives is high, recall is a key measure
   - A more suitable measure for unbalanced datasets
   - Focus on false negaitves (type II error)
   - Combined with precision, it can provide a more comprehensive model evaluation

Measures--Recall

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Disadvantages
   - It ignores the false positives
   - Sacrifice model efficiency
   - It is not a comprehensive measure and should be used with others, such as precision

Measures--False positive rate

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. False Positive Rate (FPR)

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = \frac{FP}{FP + TN}$$

2. Remark
   - FPR is also called false alarm rate
   - Directly measures "social cost" or "resource waste"
   - A more suitable measure for unbalanced datasets
   - Focus on false positives (type I error)
   - Combined with recall, it can provide a more comprehensive model evaluation

Measures--False positive rate

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Disadvantages

- The conclusion may be misleading for unbalanced datasets
- Ignores actual cost and consequences
- It is not a comprehensive measure and should be used with others, such as recall
- Confusion with precision —though this in itself is not a flaw

FPR vs precision

	FPR	1-Precision
Definition	The proportion of false positives among actual negatives	The proportion of false negative among predicted positives
Formula	FP / (FP + TN)	FP / (FP+TP)
Perspective	Measure false positives alarms for the actual negatives based on actual cases	Measure reliability of positive alarms based on predicted cases
Implications	What is the probability of harming an innocent person? It relates to the broad-based costs of screening.	Upon an alarm, what are the odds it's a false alert? It reflects the trustworthiness of the alarm
Remark	When actual negatives dominate, the model may be useless, although FPR is small	When actual negatives dominate, it may cause precision to collapse even when FP increases slightly, reflecting the severity of the ``cry wolf'' effect

Measures--F1 score

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. F1 score

$$\text{F1 score} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2. Remark
   - F1 score directly depends on TP, FP and FN
   - It takes values from $[0, 1]$
     - F1 score equals to 1, indicating perfect precision and recall
     - F1 score equals to 0, indicating the worst model
   - It balances precision and recall
   - Pay more attention to the performance balance of positive prediction

Measures--F1 score

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

1. Disadvantages
   - Ignores true negatives (TN)
   - Sensitive to very small precision or recall
   - A single value is not sufficient to evaluate a model

Remarks

1. All five measures are based on a set of estimated labels
2. For a certain model,
   - It only has one set of these measures, if it estimates the classes directly
   - It corresponds to different set of measures, if it estimates the "scores" first
     - Different thresholds correspond to different estimated labels

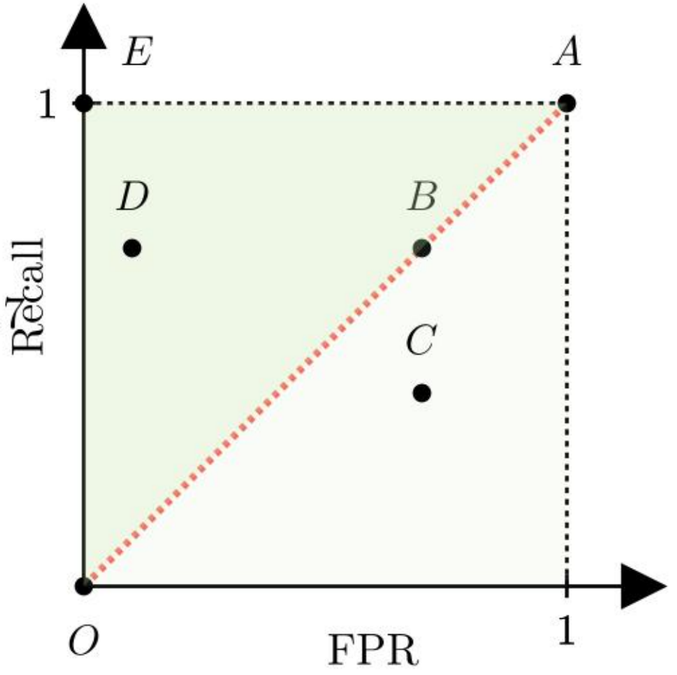
ROC (Receiver Operating Characteristic) curve

1. ROC curve plots Recall against FPR
2. Assume the availability of estimated classes for the test dataset,
   - Obtain the Recall (sensitivity)
   - Obtain the FPR (1-specificity)
3. Two scenarios for the ROC curve:
   - For a model predicting LABELS directly, it corresponds to a POINT
   - For a model predicting SCORES, it corresponds to a CURVE

ROC (Receiver Operating Characteristic) curve

1. Different estimation corresponds to different points

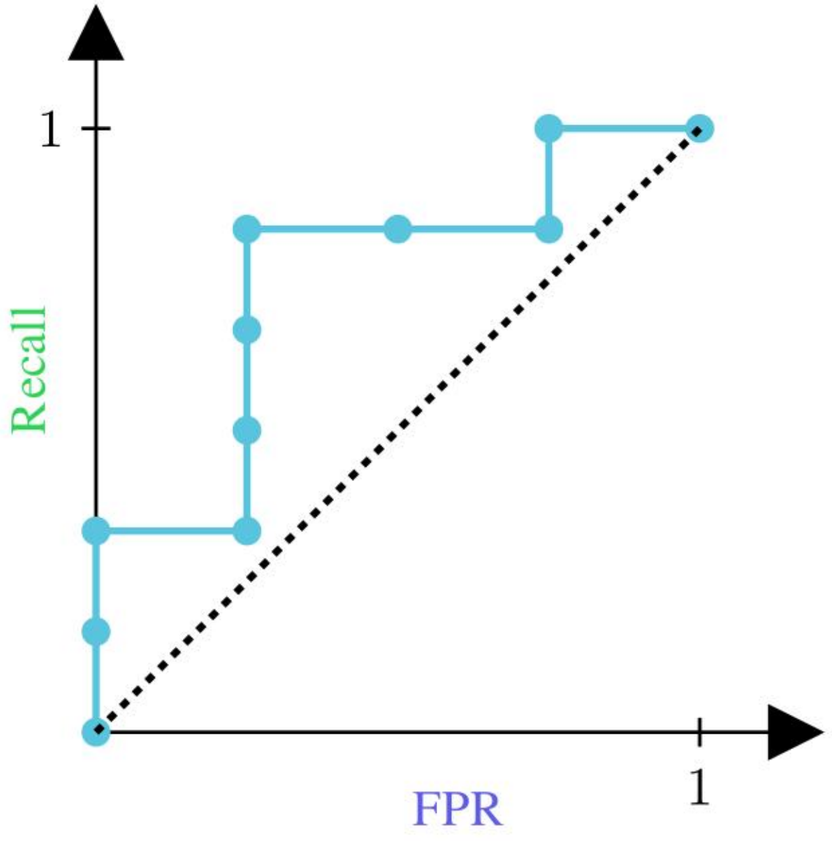
- Points on the diagonal: Random guess with different probabilities
  - $O\,(0,0)$: Assign all test examples to the negative class
  - $B\,(0.7,0.7)$: Assign a test example to the positive class with probability 0.7
  - $A\,(1,1)$: Assign all test examples to the positive class
- Points above the diagonal: Good models
  - $D\,(0.1,0.7)$: It can correctly identify 70% positives with low FPR
  - $E\,(0,1)$: Perfect model
  - The more up-left, the better
- Points below the diagonal: Actually, not that bad
  - $C\,(0.7,0.4)$: Although recall is low but FPR is high, the model messes up the two labels



Draw an ROC curve

1. For models predicting “scores”,
   - Different thresholds correspond to different points
   - Join those points produces an ROC curve
2. We use a toy example to show how to obtain an ROC curve

FPR	Recall
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1



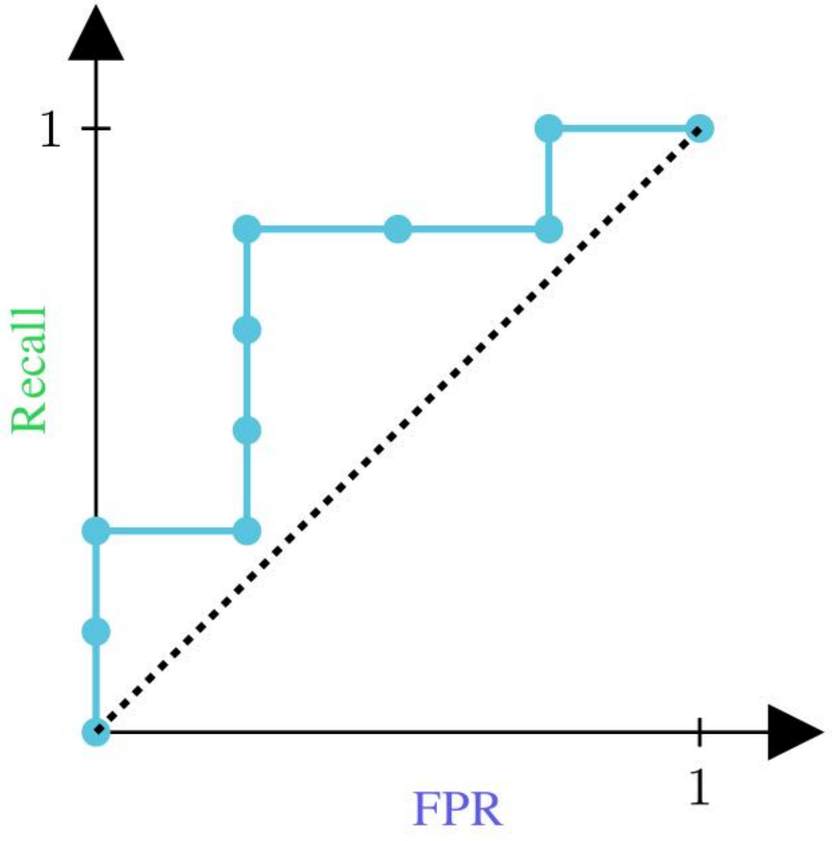
Sam. Ind.	True Lab.	Est. Lab.	Est. Sco.
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

Question 1: What is the step size vertically and horizontally?

Answer: The vertical step size is one over true positive instances, 1/6 for this example

The horizontal step size is one over true negative instances, 1/4 for this example

FPR	Recall
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1

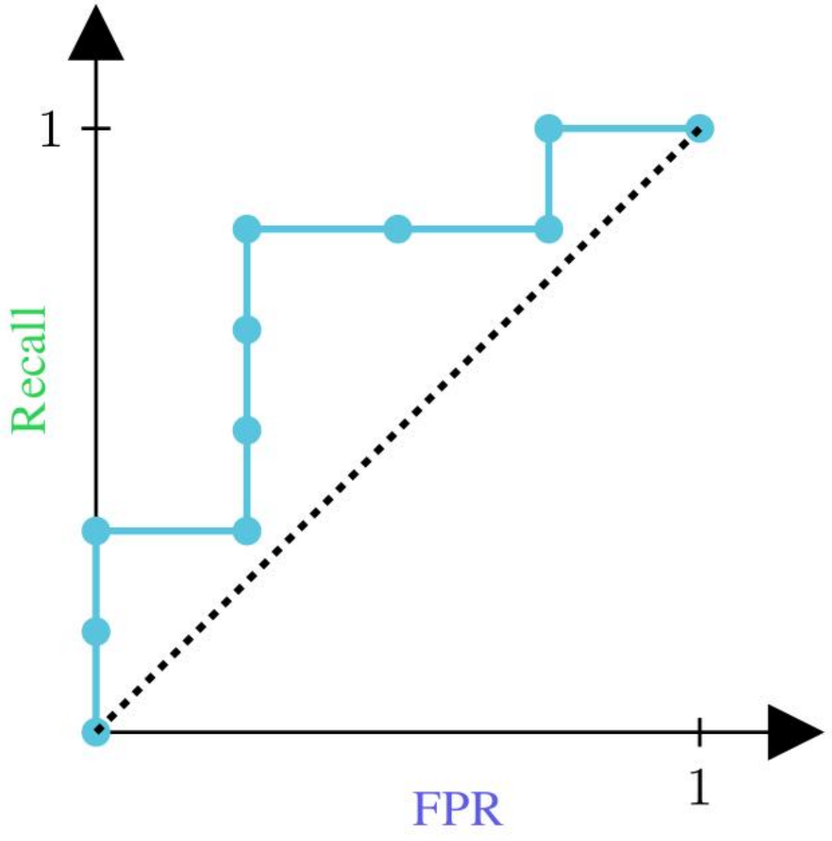


Sam. Ind.	True Lab.	Est. Lab.	Est. Sco.
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

Question 2: When does it moves vertically?

Answer: Based on the ranked results, if the next true label is positive, we move vertically

FPR	Recall
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1

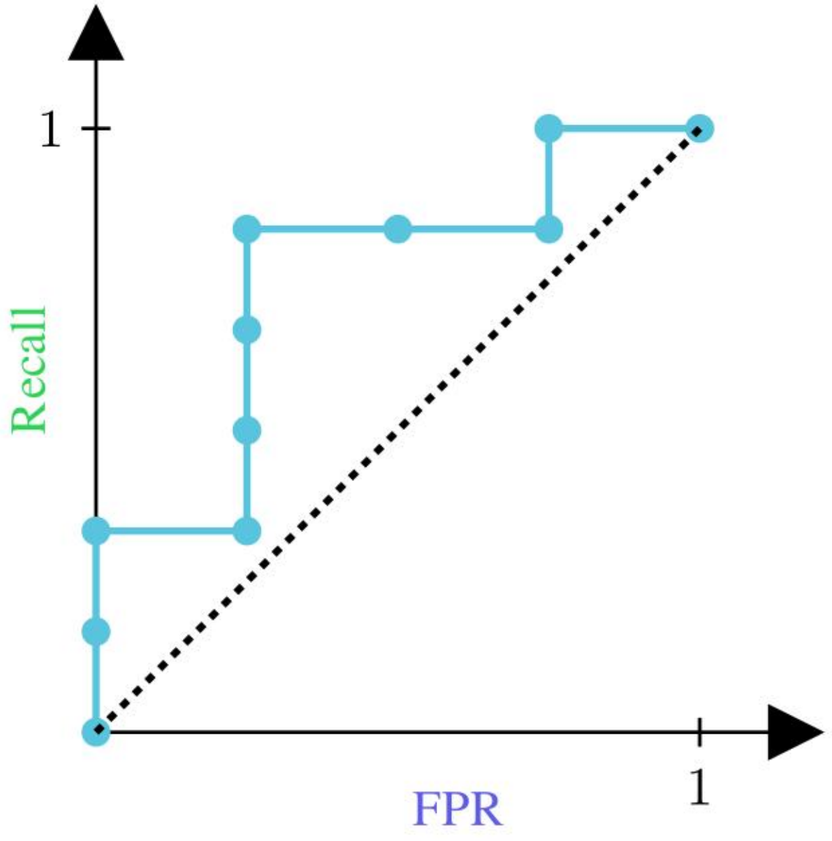


Sam. Ind.	True Lab.	Est. Lab.	Est. Sco.
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

Question 3: When does it moves horizontally?

Answer: Based on the ranked results, if the next true label is negative, we move horizontally

FPR	Recall
0	0
0	1/6
0	2/6
1/4	2/6
1/4	3/6
1/4	4/6
1/4	5/6
2/4	5/6
3/4	5/6
3/4	1
1	1

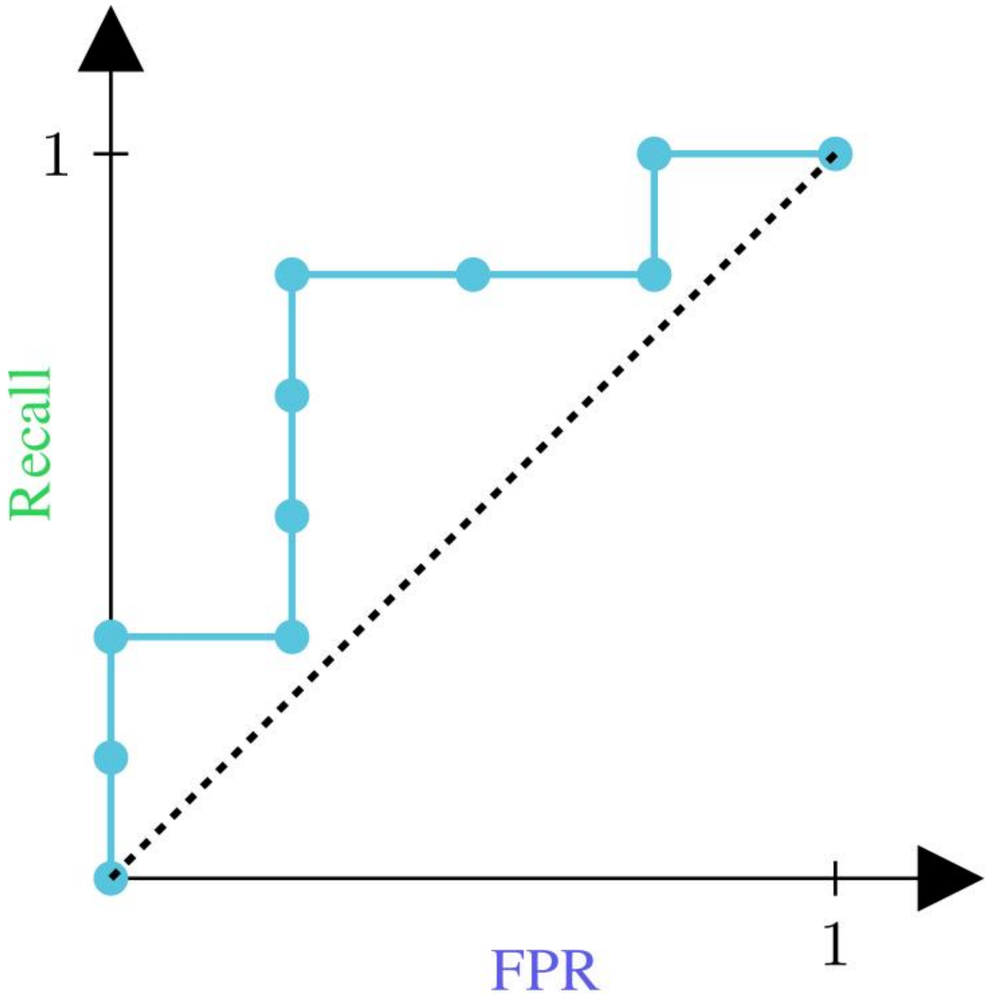


Sam. Ind.	True Lab.	Est. Lab.	Est. Sco.
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

Question 4: What does the ROC curve look like for a perfect model?

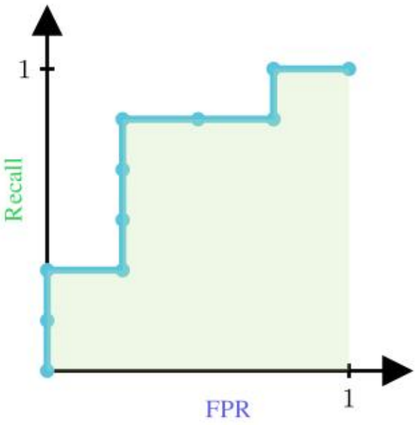
Answer: A perfect model should rank all positives before negatives.

Thus, its ROC curve should move vertically to $(0, 1)$, then horizontally to $(1, 1)$



Sam. Ind.	True Lab.	Est. Lab.	Est. Sco.
1	1		0.9
2	1		0.8
3	0		0.7
4	1		0.6
5	1		0.55
6	1		0.54
7	0		0.53
8	0		0.52
9	1		0.51
10	0		0.505

AUC



1. AUC is short for "area under the ROC curve"
   - It calculates the area under this curve
2. A good classification model corresponds to a high AUC
   - If a model "ranks" positive examples higher than negative ones, its AUC is high
   - A perfect model is the one with AUC being 1, where the scores for positives are uniformly larger than those for the negatives.
   - AUC = 0.5 indicates that the model has no ability to distinguish between positives and negatives

Remarks

1. AUC provides a single and comparable measure
2. It balances overall performance across different thresholds
3. It is robust to unbalanced datasets
4. Generally, it takes values from $[0.5, 1]$
5. If the AUC is very small, the associated model is not so bad, since it messes up positives and negatives
6. We can also use a precision-recall curve for model evaluation

Summary

1. Different metrics play different roles
   - Accuracy: Overall performance
   - Precision: Focus on the reliability of predicted positives
   - Recall: Focus on the positive samples that the model missed
   - FPR: Focus on false alarm rate among the actual negatives
   - F1 score: Balance precision and recall
   - ROC and AUC: Focus on performance under different thresholds