Lecture 9: Classification Metrics

Firas Moosvi (Slides adapted from Varada Kolhatkar)

UBC Computer Science

Announcements

- Reminder: Test 2 is this week!
- Practice Test 2 should be released soon
- See the announcement on Ed Discussion for the details of resources you'll be provided on the exam

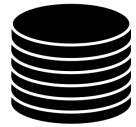
ML workflow

What question do I want to answer?

↓

Formulation to supervised machine learning problem

Supervised ML process

1. Data collection 

2. Data cleaning, splitting 

3. Data exploration (EDA) 

4. Preprocessing, Feature engineering 

5. Feature selection, Model building 

6. Evaluation, model selection 

7. Test data predictions, visualizations, possible deployment 

Accuracy

- So far, we've been measuring model performance using **Accuracy**.
- **Accuracy** is the proportion of all predictions that were correct — whether *positive* or *negative*.

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}}$$

- But is **accuracy** always the right metric to evaluate a model? 🤔

A fraud classification example

```
(139554, 29)
```

	Class	Time	Amount	V1	V2	V3	
64454	0	51150.0	1.00	-3.538816	3.481893	-1.827130	-0.5
37906	0	39163.0	18.49	-0.363913	0.853399	1.648195	1.11
79378	0	57994.0	23.74	1.193021	-0.136714	0.622612	0.78
245686	0	152859.0	156.52	1.604032	-0.808208	-1.594982	0.20
60943	0	49575.0	57.50	-2.669614	-2.734385	0.662450	-0.0

5 rows × 31 columns

DummyClassifier

Let's try a DummyClassifier, which makes predictions without learning any patterns.

```
1 dummy = DummyClassifier()
2 cross_val_score(dummy, X_train, y_train).mean()
```

```
np.float64(0.9983017327649726)
```

- The accuracy looks surprisingly high!
- Should we be happy with this model and deploy it?

Problem: Class imbalance

```
1 y_train.value_counts()
```

```
Class
0    139317
1       237
Name: count, dtype: int64
```

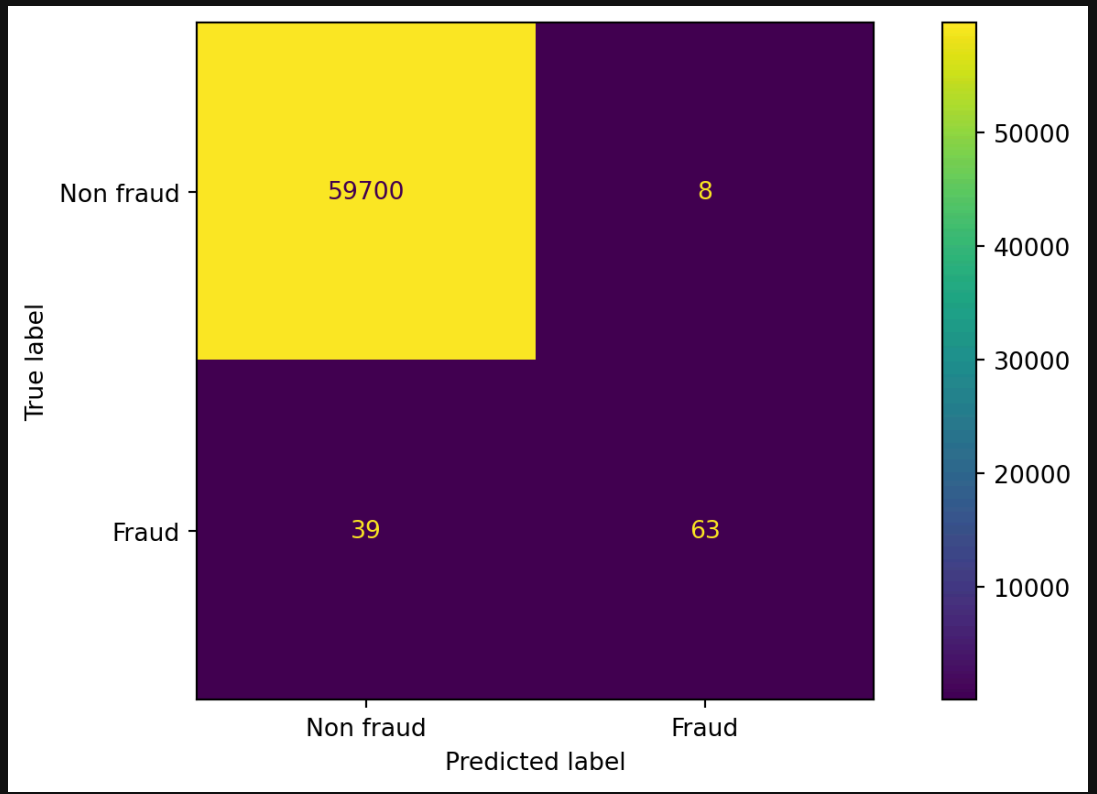
- In many real-world problems, some classes are much rarer than others.
- A model that always predicts “no fraud” could still achieve >99% accuracy!
- This is why accuracy can be misleading in imbalanced datasets.
- We need metrics that differentiate types of errors.

`DummyClassifier`: Confusion matrix

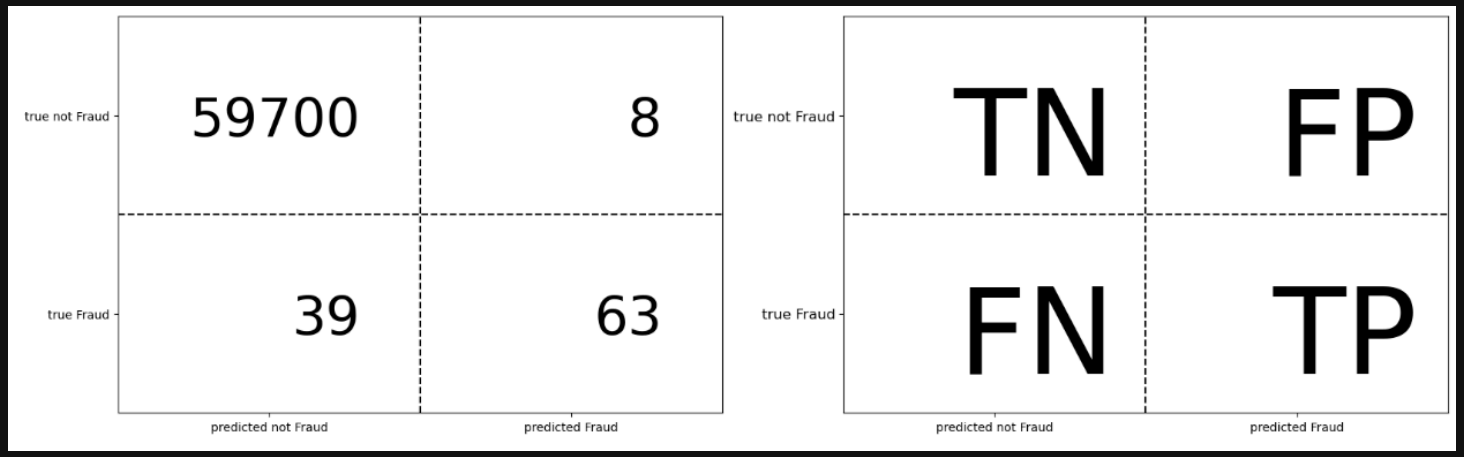
Which types of errors would be most critical for the bank to address? Missing a fraud case or flagging a legitimate transaction as fraud?

`LogisticRegression`: Confusion matrix

Are we doing better with logistic regression?



Understanding the confusion matrix



- TN ⟶ True negatives
- FP ⟶ False positives
- FN ⟶ False negatives
- TP ⟶ True positives

Practice: confusion matrix terminology

Confusion matrix questions

Imagine a spam filter model where emails labeled **1 = spam, 0 = not spam**.

If a spam email is incorrectly classified as not spam, what kind of error is this?

- a. A false positive
- b. A true positive
- c. A false negative
- d. A true negative

Confusion matrix questions

In an intrusion detection system, **1 = intrusion, 0 = safe**.

If the system misses an actual intrusion and classifies it as safe, this is a:

- a. A false positive
- b. A true positive
- c. A false negative
- d. A true negative

Confusion matrix questions

In a medical test for a disease, **1 = diseased, 0 = healthy**.

If a healthy patient is incorrectly diagnosed as diseased, that's a:

- a. A false positive
- b. A true positive
- c. A false negative
- d. A true negative

Metrics other than accuracy

Now that we understand the different types of errors, we can explore metrics that better capture model performance when **accuracy falls short**, especially for **imbalanced datasets**.

We'll start with three key ones:

- **Precision**
- **Recall**
- **F1-score**

Precision and recall

Let's revisit our fraud detection scenario. The circle below represents **all transactions predicted as fraud** by an **imaginary toy model** designed to detect fraudulent activity.

Intuition behind the two metrics

- **Precision**: *Of all the transactions predicted as fraud, how many were actually fraud?*
  - High precision ⟶ few false alarms (low false positives).
- **Recall**: *Of all the actual fraud cases, how many did the model catch?*
  - High recall ⟶ few missed frauds (low false negatives).

Trade-off between precision and recall

- Increasing **recall** often decreases **precision**, and vice versa.
- Example:
  - Predict *"fraud"* for every transaction ⟶ perfect recall, terrible precision.
  - Predict *"fraud"* only when 100% sure ⟶ high precision, low recall.

The right balance depends on the application and cost of errors.

F1-score

- Sometimes, we want a **single metric** that balances precision and recall.
- The **F1-score** is the **harmonic mean** of the two:

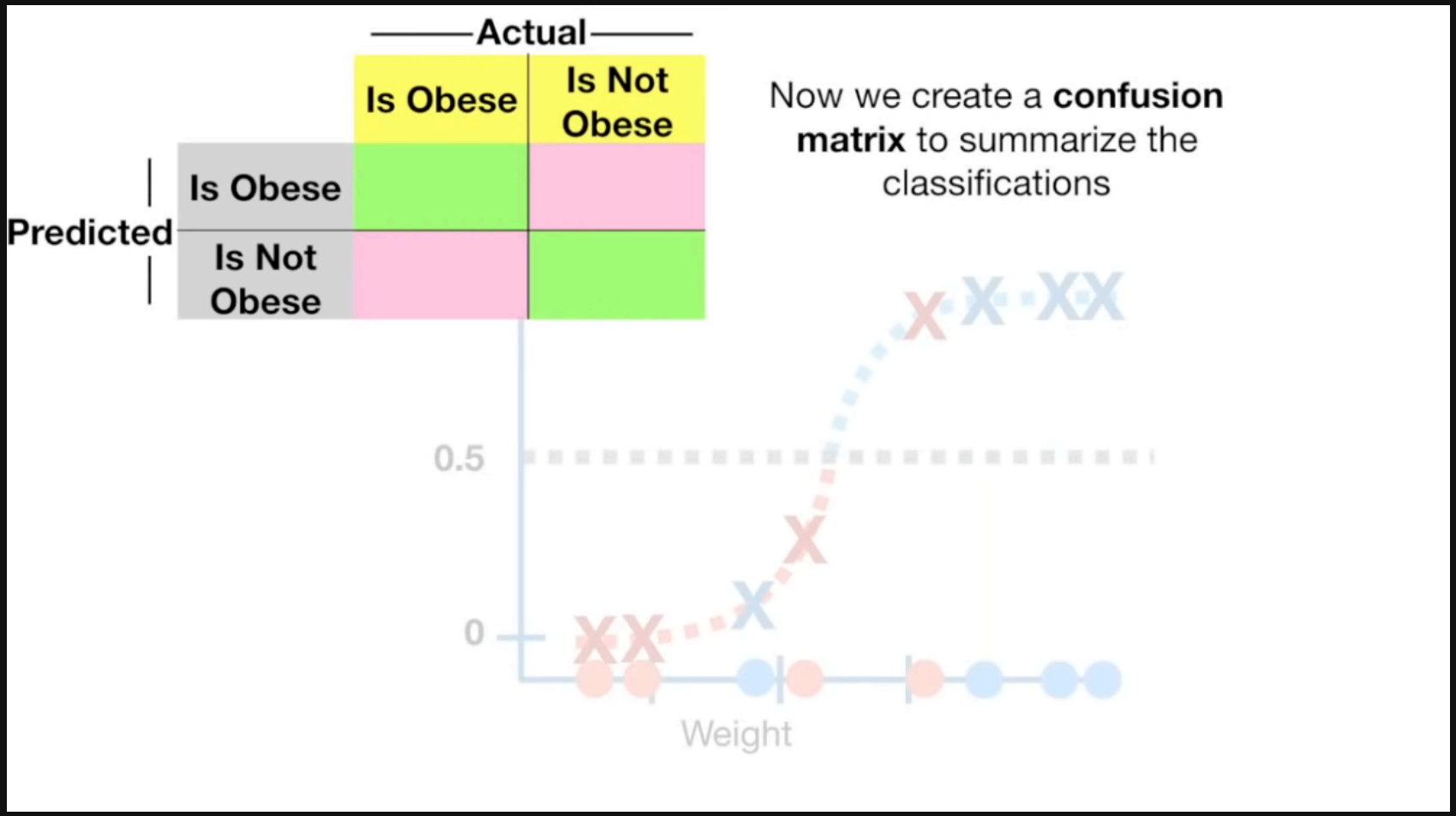
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- High **F1** means both precision and recall are strong.
- Useful when we care about both false positives **and** false negatives.

Summary

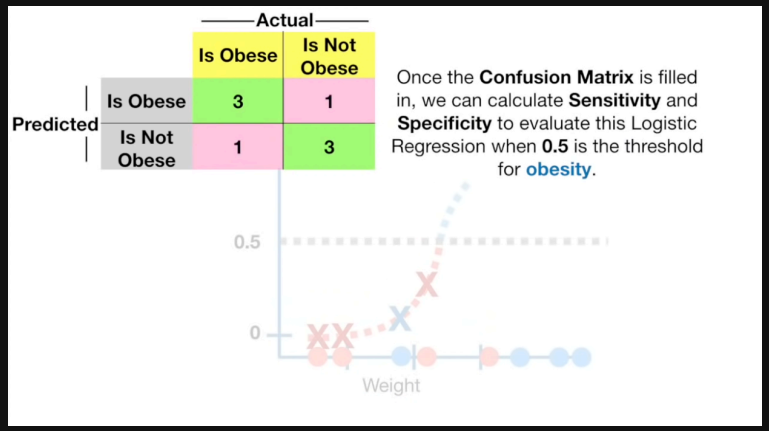
Metric	What it measures	High value means
Accuracy	Overall correctness	Model gets most predictions right
Precision	Quality of positive predictions	Few false alarms
Recall	Quantity of true positives caught	Few missed positives
F1-score	Balance of precision & recall	Both precision and recall are high

Activity 1: Create Confusion Matrix



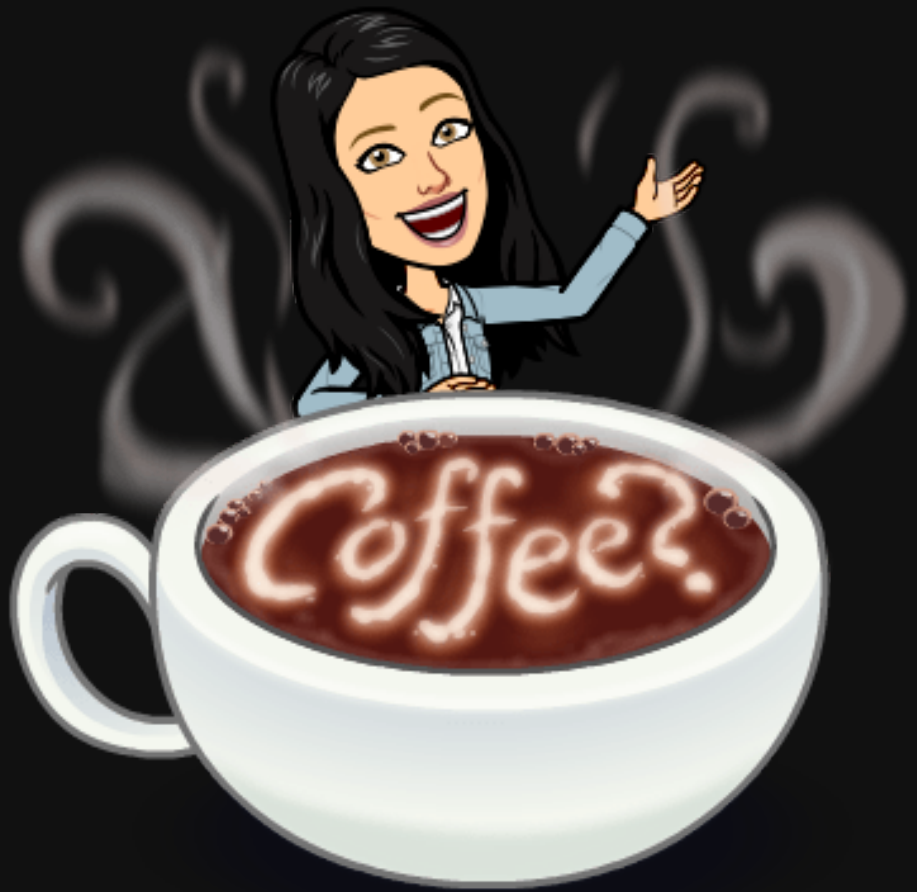
Activity 2: Calculate Precision, Recall, Specificity

- Recall (aka Sensitivity in biomedical literature)
  - TP/(TP+FN)
- Precision
  - TP/(TP+FP)
- Specificity
  - TN/(TN+FP)



Break!

Let’s take a break!



Clicker Exercise 9.1

Select all of the following statements which are TRUE.

- a. In medical diagnosis, false positives are more damaging than false negatives (assume “positive” means the person has a disease, “negative” means they don’t).
- b. In spam classification, false positives are more damaging than false negatives (assume “positive” means the email is spam, “negative” means they it’s not).
- c. If method A gets a higher accuracy than method B, that means its precision is also higher.
- d. If method A gets a higher accuracy than method B, that means its recall is also higher.

Counter examples

Method A - higher accuracy but lower precision

Negative	Positive
90	5
5	0

Method B - lower accuracy but higher precision

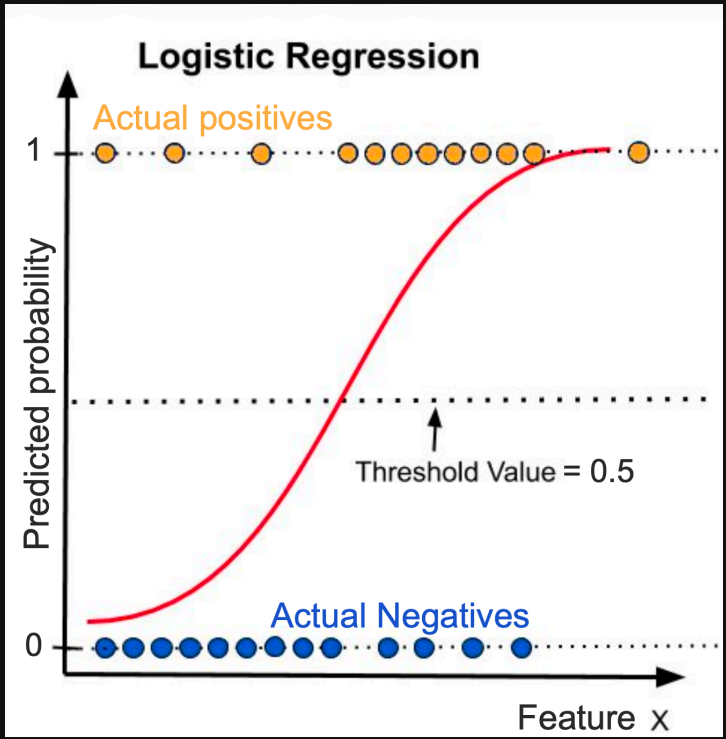
Negative	Positive
80	15
0	5

Takeaway

- **Accuracy** summarizes overall correctness but hides class-specific behaviour.
- You can have **high accuracy but poor precision or recall**, especially in **imbalanced datasets**.
- Always check **multiple metrics** before deciding which model is better.

Threshold-based classification

Predicting with logistic regression



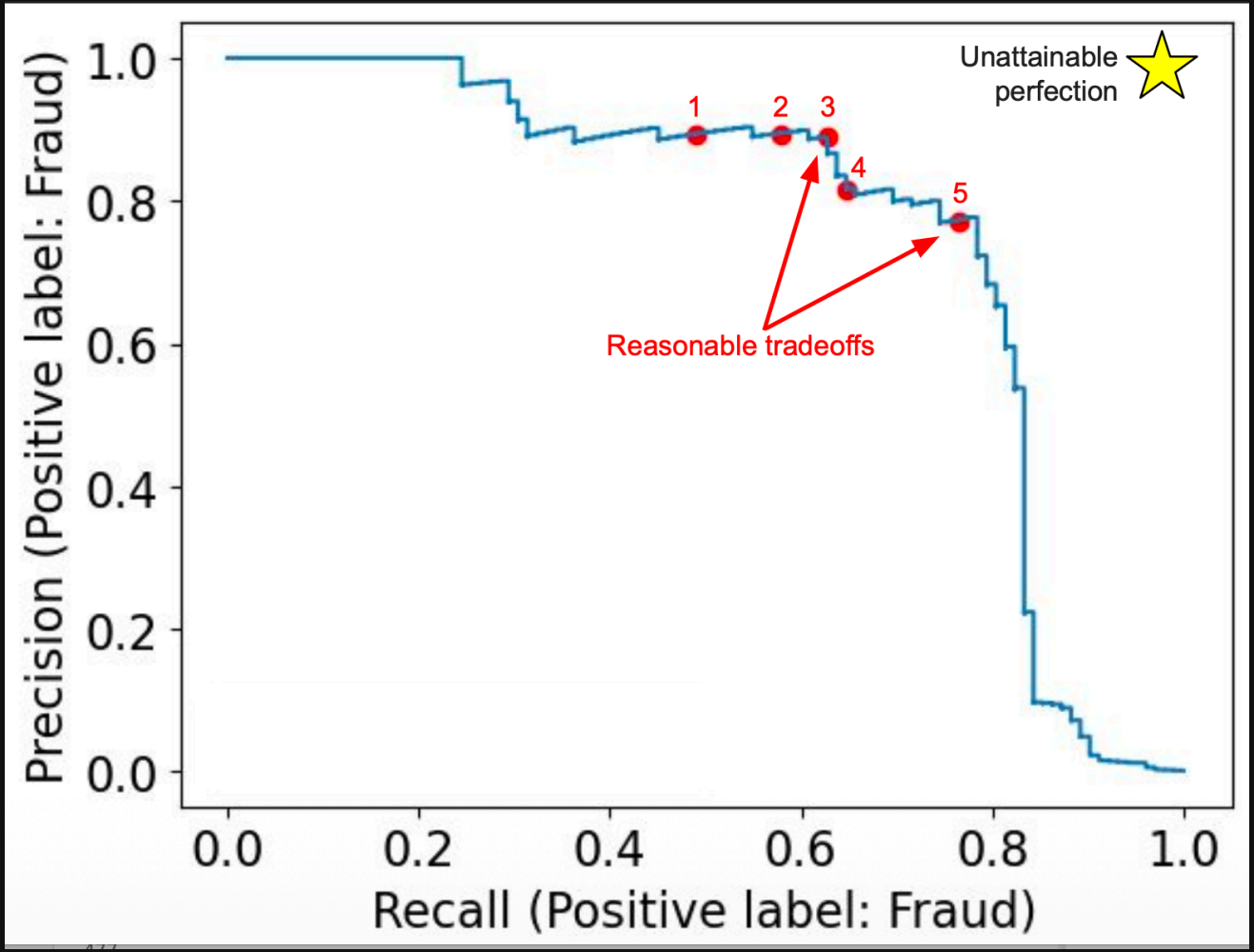
- Most classification models don't directly predict labels. They predict scores or probabilities.
- To get a label (e.g., "fraud" or "non fraud"), we choose a threshold (often 0.5). If the threshold changes, predictions change, and so do the errors.
- What happens to precision and recall if we change the probability threshold?
- Play with classification thresholds

PR curve

- Calculate precision and recall (TPR) at every possible threshold and graph them.
- Top left ⟶ Very high threshold (strict model = high precision)
- Bottom right ⟶ Very low threshold (linient model = high recall)

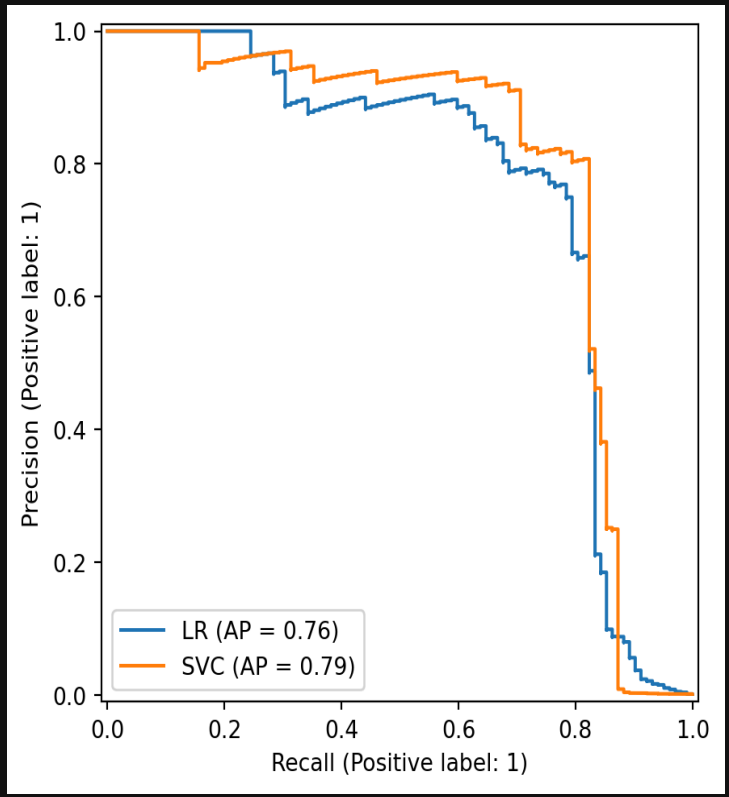
PR curve different thresholds

- Which of the red dots are reasonable trade offs?



Average Precision (AP) Score

- AP score summarizes the PR curve by calculating the area under the curve
- It measures the ranking ability of a model; how well it assigns higher probabilities to positive examples than to negative ones, regardless of the specific threshold.



Clicker Exercise 9.2

Choose the appropriate evaluation metric for the following scenarios:

Scenario 1: Balance between precision and recall for a threshold.

Scenario 2: Assess performance across all thresholds.

- a. F1 for 1, AP for 2
- b. AP for 1, F1 Score for 2
- c. AP for both
- d. F1 for both

Clicker Exercise 9.3

Select all of the following statements which are TRUE.

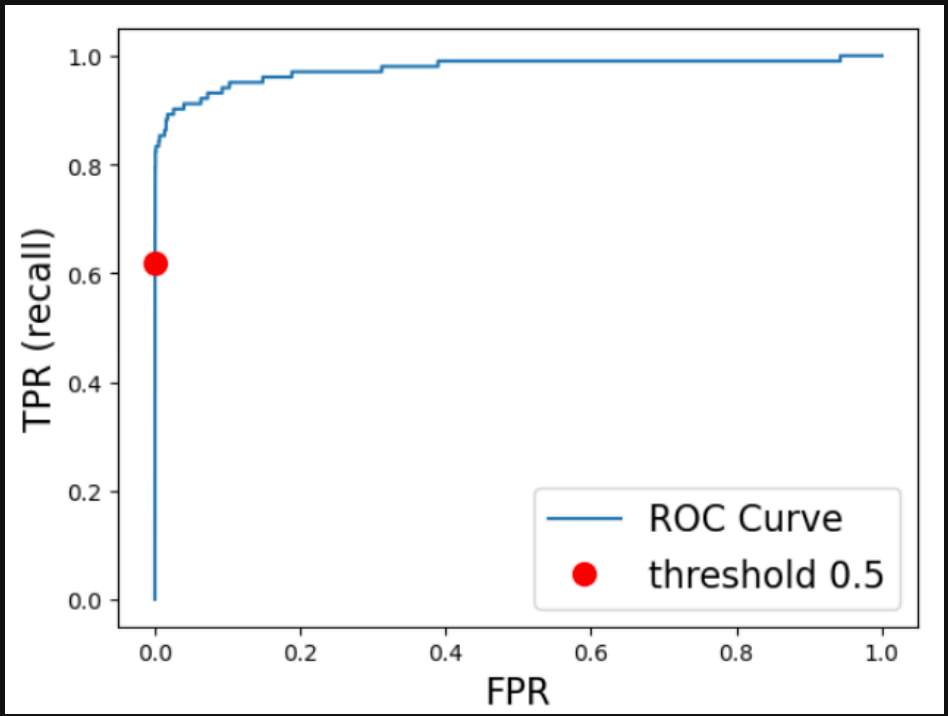
- a. If we increase the classification threshold, both true and false positives are likely to decrease.
- b. If we increase the classification threshold, both true and false negatives are likely to decrease.
- c. Lowering the classification threshold generally increases the model's recall.
- d. Raising the classification threshold can improve the precision of the model if it effectively reduces the number of false positives without significantly affecting true positives.

ROC Curve

- Compute the **True Positive Rate (TPR)** and **False Positive Rate (FPR)** at **every possible threshold**, and plot **TPR** vs **FPR**.
- How well does the model separate positive and negative classes in terms of predicted probability?
- A good choice when the dataset is **reasonably balanced** or **not extremely imbalanced** (e.g., fraud detection, disease diagnosis).

ROC Curve example

- **Bottom-left** ⟶ very high threshold (almost everything predicted negative: low recall, low FPR).
- **Top-right** ⟶ very low threshold (almost everything predicted positive: high recall, high FPR).

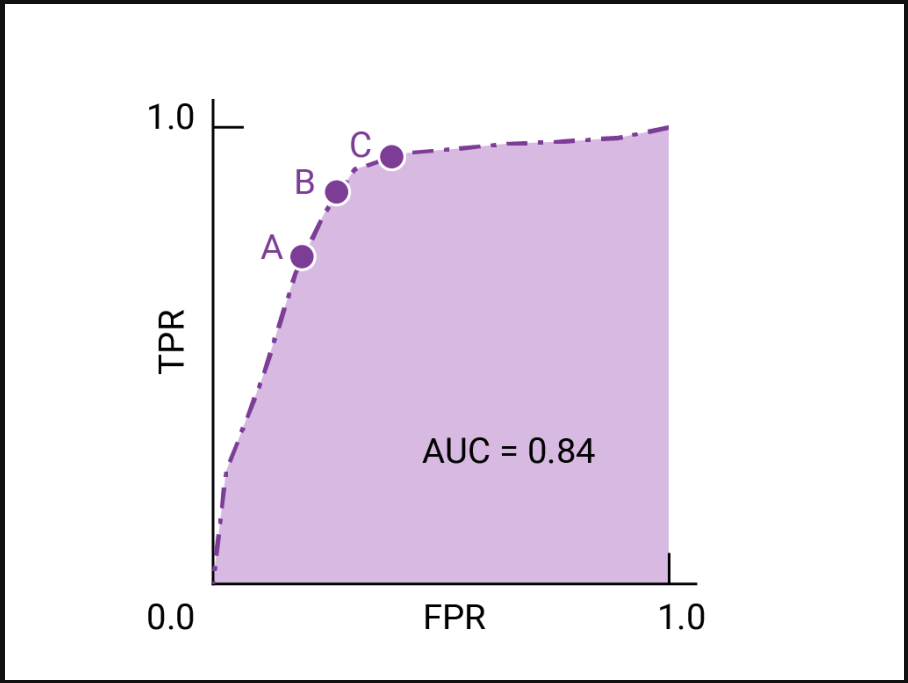


AUC

- The area under the ROC curve (AUC) represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative.

ROC AUC questions

Consider the points A, B, and C in the following diagram, each representing a threshold. Which threshold would you pick in each scenario?



- a. If false positives (false alarms) are highly costly
- b. If false positives are cheap and false negatives (missed true positives) highly costly
- c. If the costs are roughly equivalent

Source

Group Work: Class Demo & Live Coding

For this demo, each student should click this link to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.

What did we learn?

- Why accuracy is not always a good metric?
- Confusion matrix
- Precision, recall, & f1-score
- Precision-recall curves & average precision
- Receiver Operator Characteristic (ROC) curves & AUC