CPSC 330 Lecture 9: Classification Metrics

Firas Moosvi (Slides adapted from Varada Kolhatkar)

Announcements

- Learning Log 02 is now out!
- Reminder: HW4 is due Monday night
- Reminder: Midterm 1 is this week!

Recap of Hyperparameter optimization (Demo)

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.

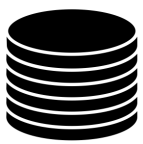
ML workflow

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

Classification Metrics

At the end of last class we talked about some of the problems with “accuracy”, and we brainstormed some possible alternatives, and saw that there are tonnes of options.

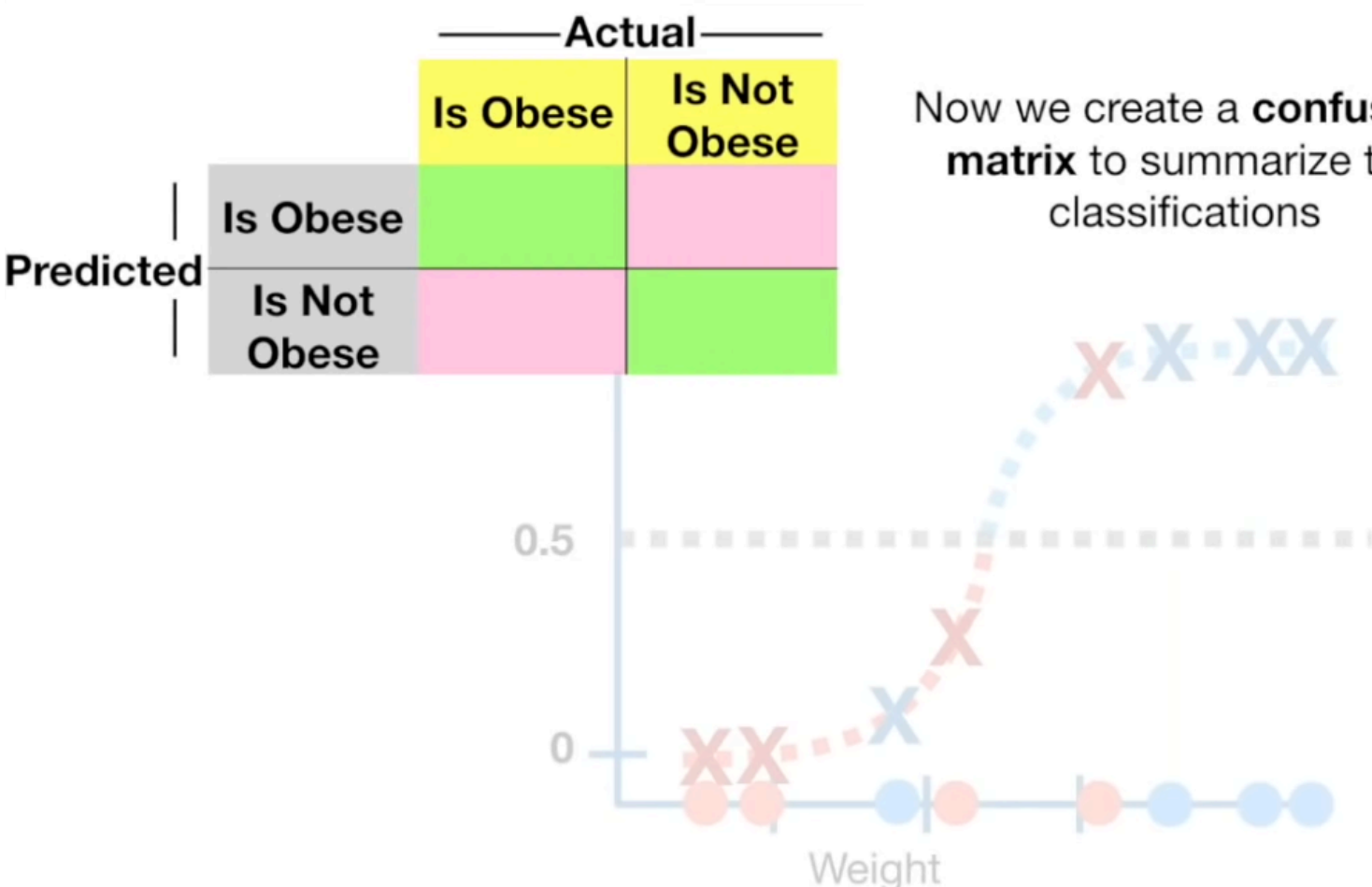
Today, let’s sift through the noise and develop some intuition about **why** we need classification metrics, and **how** some of them are used.

Example from StatQuest!

Let's first walk through this example through StatQuest with obese mice and classifying them using Logistic Regression:

Source: StatQuest

Activity 1: Create Confusion Matrix



Source: StatQuest

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!

Confusion matrix questions

Imagine a spam filter model where emails classified as spam are labeled 1 and non-spam emails are labeled 0. If a spam email is incorrectly classified as non-spam, what is this error called?

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

Confusion matrix questions

In an intrusion detection system, intrusions are identified as 1 and non-intrusive activities as 0. If the system fails to identify an actual intrusion, wrongly categorizing it as non-intrusive, what is this type of error called?

- 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, diseased states are labeled as 1 and healthy states as 0. If a healthy patient is incorrectly diagnosed with the disease, what is this error known as?

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

iClicker Exercise 9.1

iClicker cloud join link: https://join.iclicker.com/YJHS

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

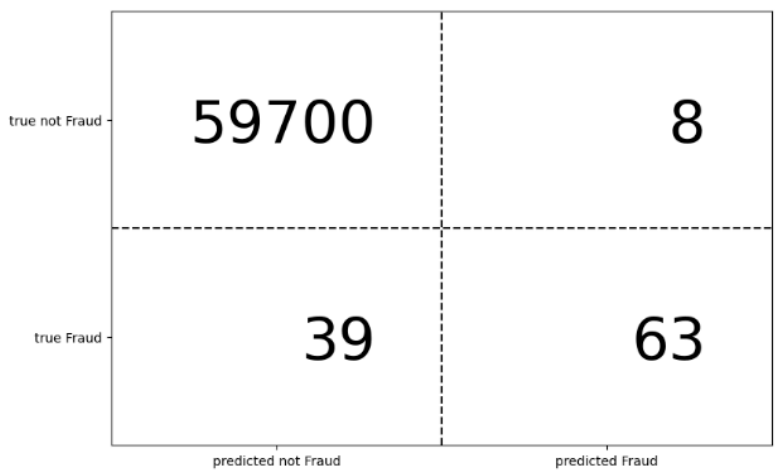
Method A - higher accuracy but lower precision

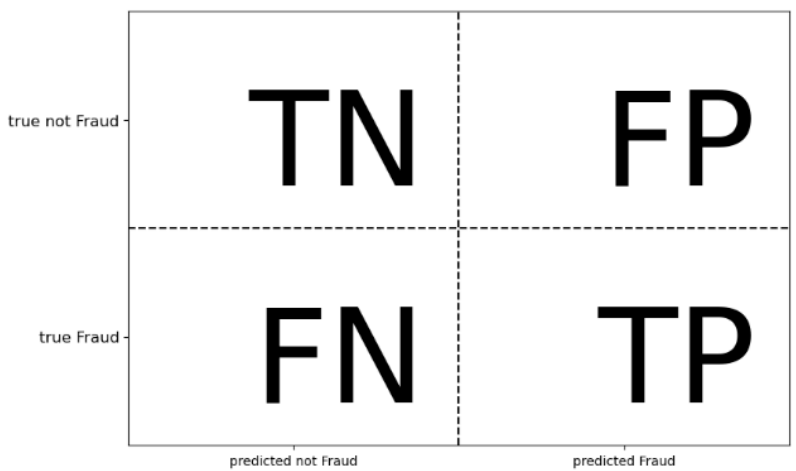
Negative	Positive
90	5
5	0

Method B - lower accuracy but higher precision

Negative	Positive
80	15
0	5

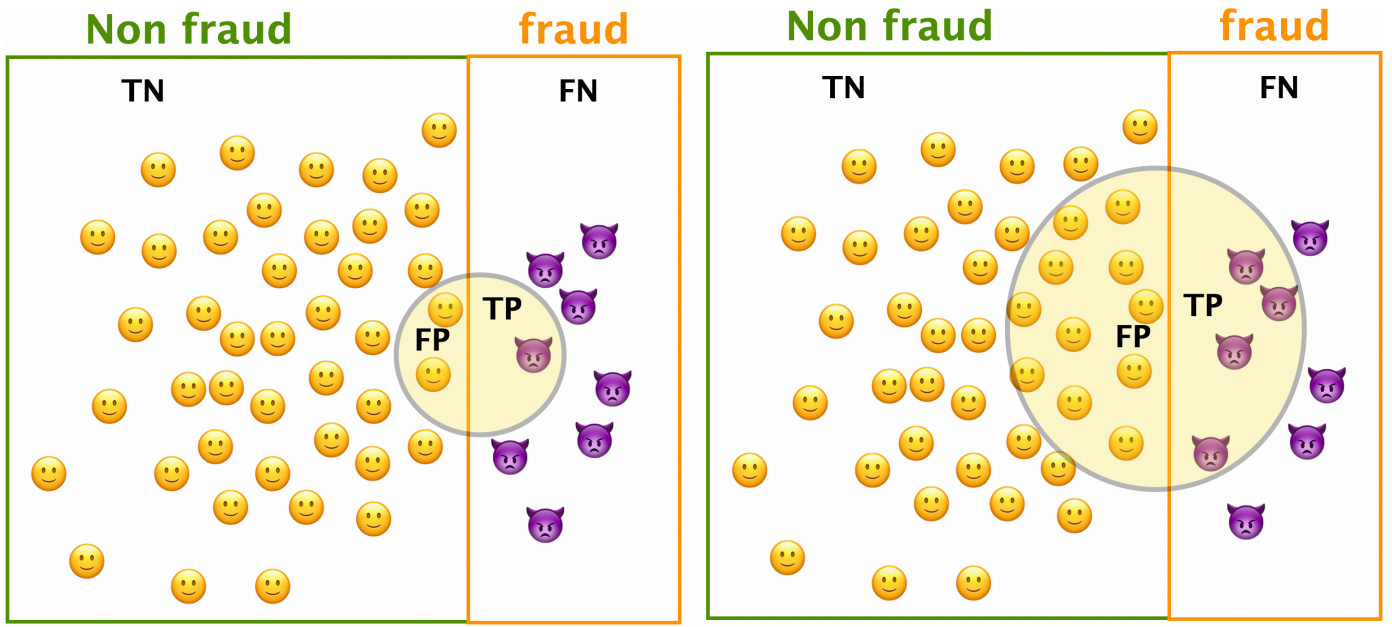
Recap: Confusion matrix





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

Recap: Precision, Recall, F1-Score



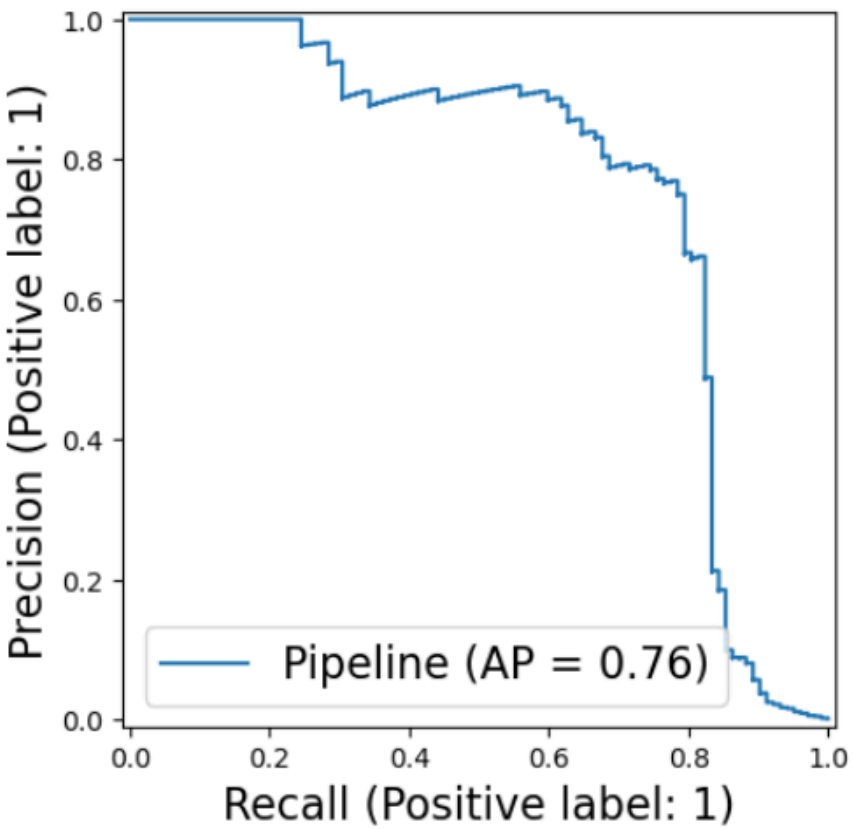
$$f1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$$Precision = \frac{1}{3} \qquad Precision = \frac{TP}{TP + FP} \qquad \downarrow Precision = \frac{4}{15}$$

$$Recall = \frac{1}{8} \qquad Recall = \frac{TP}{TP + FN} \qquad \uparrow Recall = \frac{4}{8}$$

Recap: PR curve

- Calculate precision and recall (TPR) at every possible threshold and graph them.
- Better choice for highly imbalanced datasets because it focuses on the performance of the positive class.

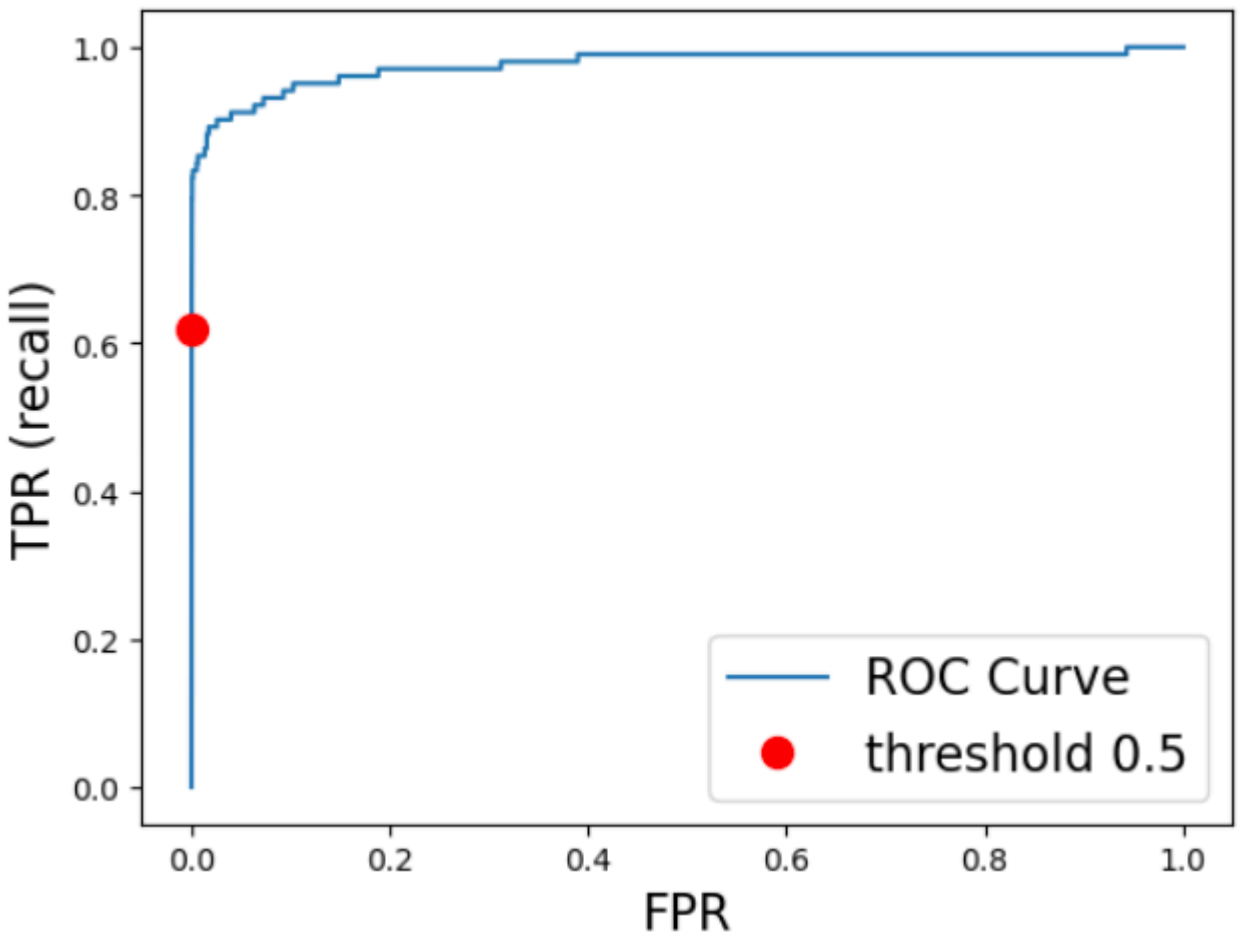


Demo: PR curve

Google's Machine Learning Modules

Recap: ROC curve

- Calculate the true positive rate (TPR) and false positive rate (FPR) ($\frac{FP}{FP+TN}$) at every possible thresholding and graph TPR over FPR.
- Good choice when the datasets are roughly balanced.



Recap: ROC Curve

- Not a great choice when there is an extreme imbalance because FPR can remain relatively low even if the number of false positives is high, simply because the number of negatives is very large.

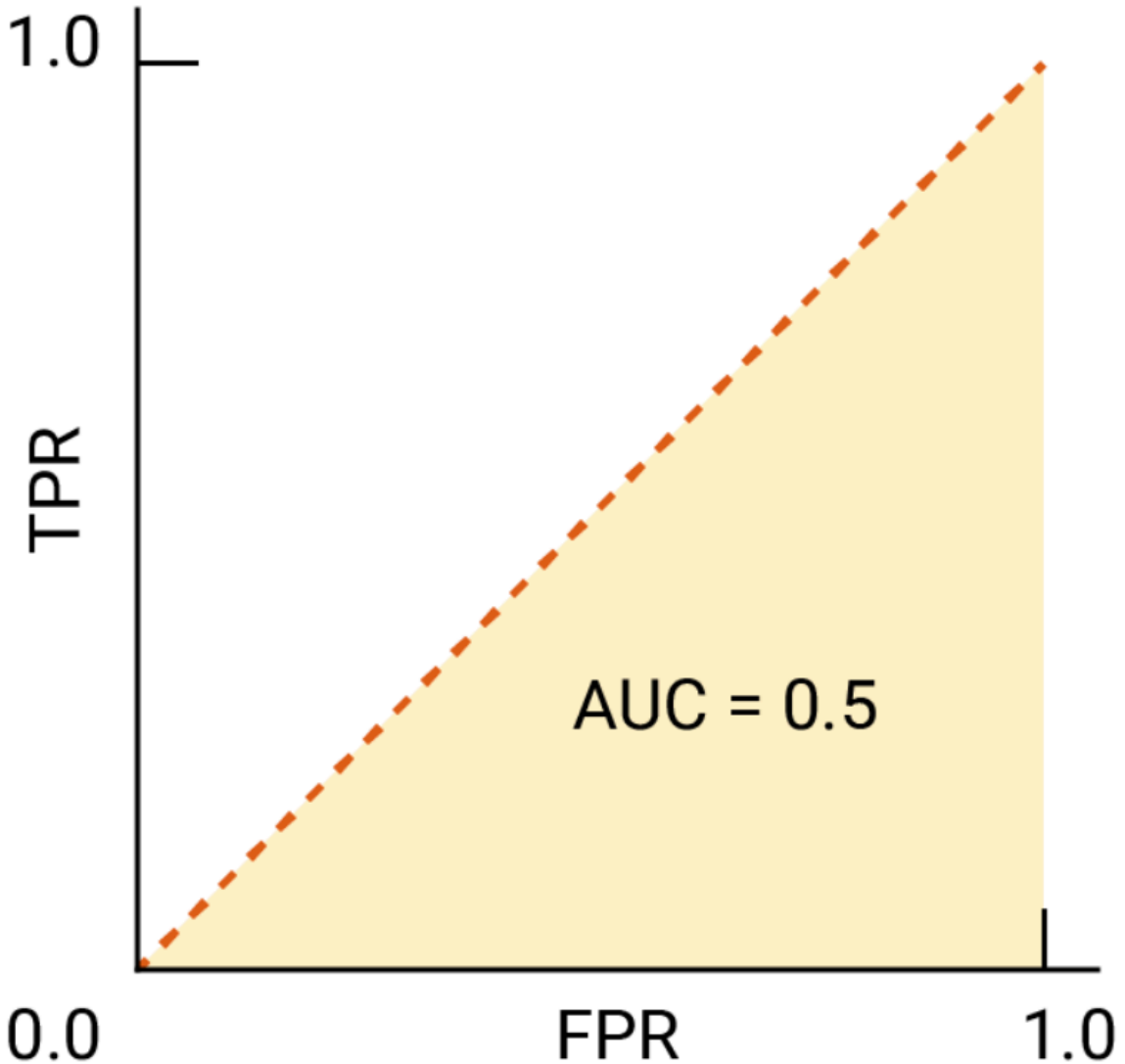
$$FPR = \frac{FP}{FP + TN}$$

- 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.

Questions for you

- What's the difference between the average precision (AP) score and F1-score?
- Which model would you pick?

Questions for you



- What's the AUC of a baseline model?

Source

Questions for you

- Which model would you pick?

iClicker Exercise 9.2

iClicker cloud join link: https://join.iclicker.com/YJHS

Select all of the following statements which are TRUE.

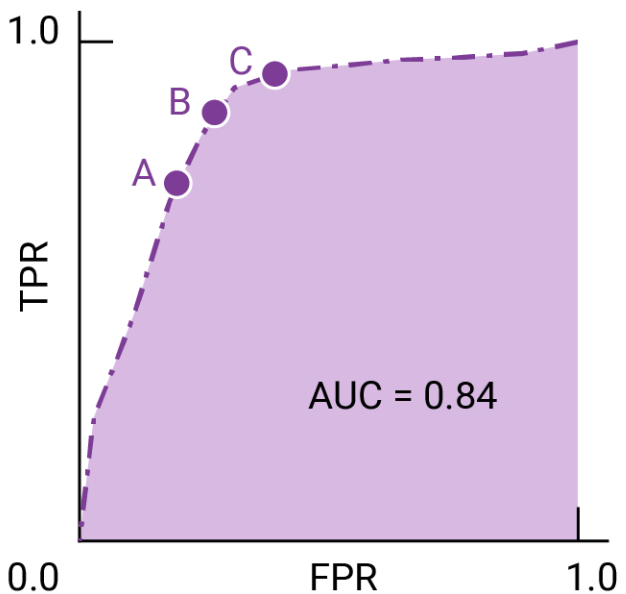
- a. If we increase the classification threshold, both true and false positives are likely to decrease.
- b. If we increase the classification threshold, both true and false negatives are likely to decrease.
- c. Lowering the classification threshold generally increases the model's recall.
- d. Raising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

Dealing with class imbalance

- Under sampling
- Oversampling
- `class weight="balanced"` (preferred method for this course)
- SMOTE

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