Welcome to Data Science Online Bootcamp

Week#4_Day#1

Democratizing Data Science Learning

Learning Objectives

Multi- Logistic Classification

Why not Accuracy?

Evaluating the Performance of Logistic Regression model

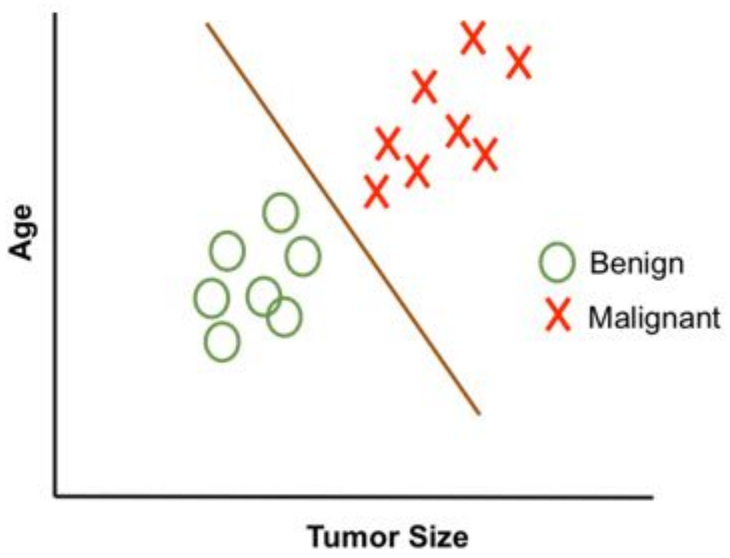
Confusion Matrix

What is Classification?

Let's learn with some examples:

- In **Classification we** classify the outcome
- **Examples:**
  - Predict whether a transaction is fraud or not fraud
  - Predict whether to give loan or not
  - Predict whether to give college admission or not
  - Predict the grade (Grade A, B, C, D)
  - Note: Classification can be more than two

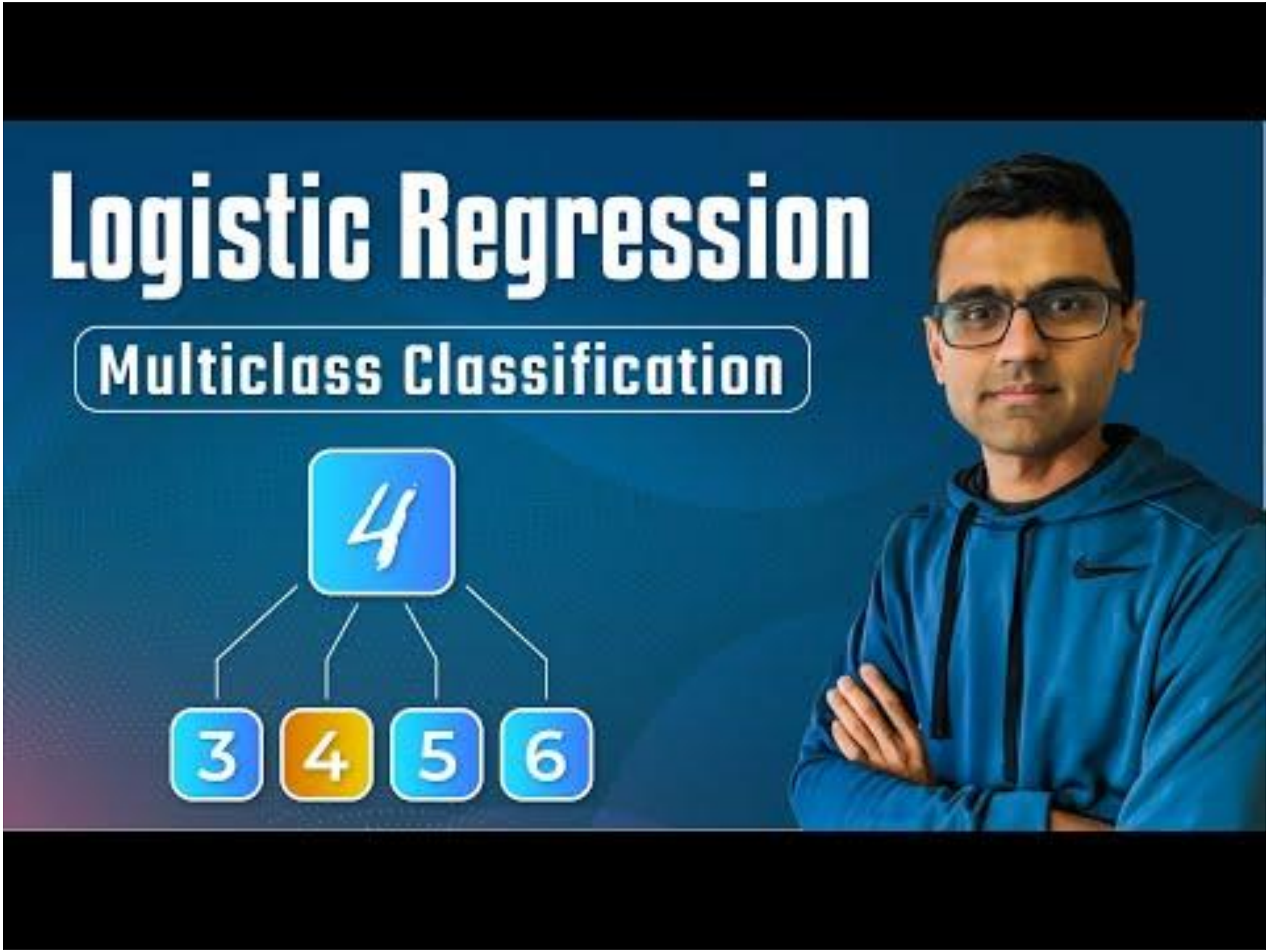
Feature	Tumor Age and Tumor Size
Label	Tumor (Benign or Malignant)
Goal/ Aim	We want to predict whether a tumor is benign or malignant from the given age and tumor size



What is Multi-Classification?

It is as simple as dividing waste into 4 categories - plastic, glass, metal, paper

Understanding Multi-Class Logistic Regression



Evaluating the Performance of Logistic Regression model

- Model Evaluation is a very important part in any analysis to answer the following questions:

  How well does the model fit the data?, Which predictors are most important?, Are the predictions accurate?

- Guess what, evaluating a Classification model is not as simple as Linear Regression.

- But why?

- You must be wondering 'Can't we just use accuracy of the model as the holy grail metric?'

Notebook

Link to notebook:

https://github.com/dphi-official/ML_Models/blob/master/Logistic_Regression/multiclass_logistic_regression.ipynb

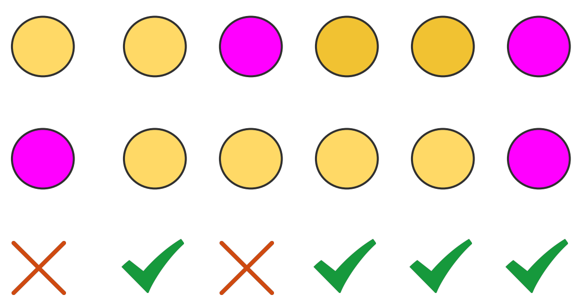
Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Actual Labels:

Predicted Labels:



Task at hand - Separate Yellow and pink balls

Colors predicted by our model

Correctly predicted = 4
Total Predictions = 6
Accuracy = 4/6 = 66.7%

Why not Accuracy?

- Accuracy is very important, but it might not be the best metric all the time. Let's look at why with an example -:

- Let's say we are building a model which predicts if a transaction is fraudulent or not

- Let's imagine, we build a basic model which always predicts that a transaction is not fraudulent. Guess what would be the accuracy of this model? **~99% !!** (You may ask why? Well, less than 1% transactions are usually fraudulent and there is a huge class imbalance. So even if you fit a wrong model that always predicts a transaction to be not fraudulent, the accuracy will remain 99% owing to class imbalance)

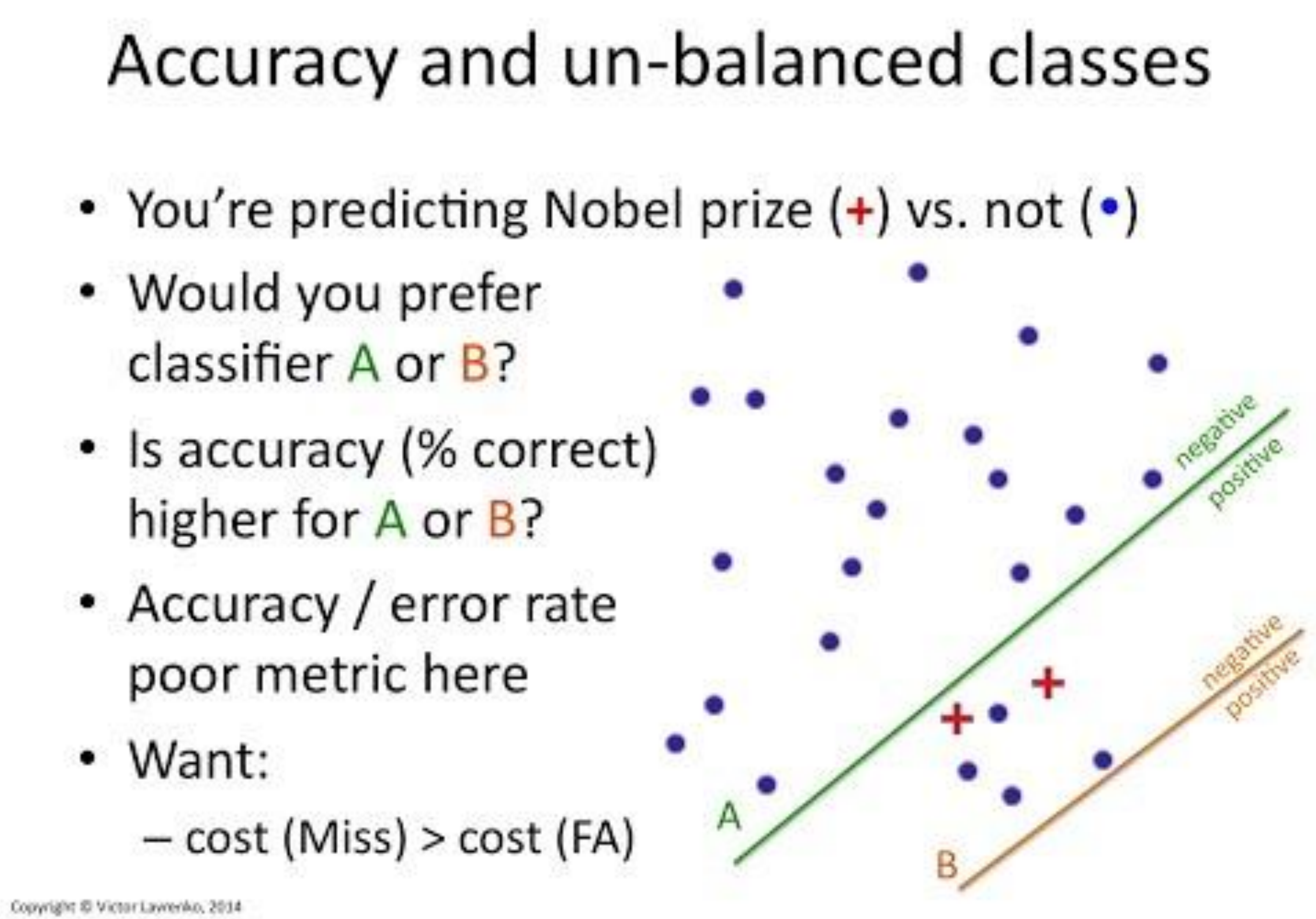
- Impressive, right? Well, the probability of a bank buying this model is absolute zero. 😆

- In a problem where there is a large class imbalance, a model can predict the value of the **majority class for all predictions** and achieve a high classification accuracy.

- **While our model has a stunning accuracy, this is an apt example where accuracy is definitely not the right metric.**

Why not Accuracy?

Watch till 1 min 14 secs to understand why accuracy is bad metric for model performance



Evaluating the Performance of Logistic Regression model

Logistic Regression employs different sets of metrics than Linear Regression. Here, we deal with probabilities and categorical values.

In the following slides, we describe a few of the evaluation metrics used for Logistic Regression:

Is confusion matrix confusing or it resolves the confusion?

You decide!

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Let's start with an example confusion matrix for a binary classifier for disease prediction (though it can easily be extended to the case of more than two classes):

Predicted by our model

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Actual value

Confusion Matrix

Let's now define the most basic terms, which are whole numbers (not rates):

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.

- **true negatives (TN):** We predicted no, and they don't have the disease.

- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

I know these seem hard to memorise. One thing that has helped me remember these are by putting it in a better way:

false positives = falsely classified as being positive.

Confusion Matrix

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

- **Precision:** Correctly predicted as positives compared to total predicted as positives
  Precision = **TP/(TP+FP)** = 100/110 = 0.91

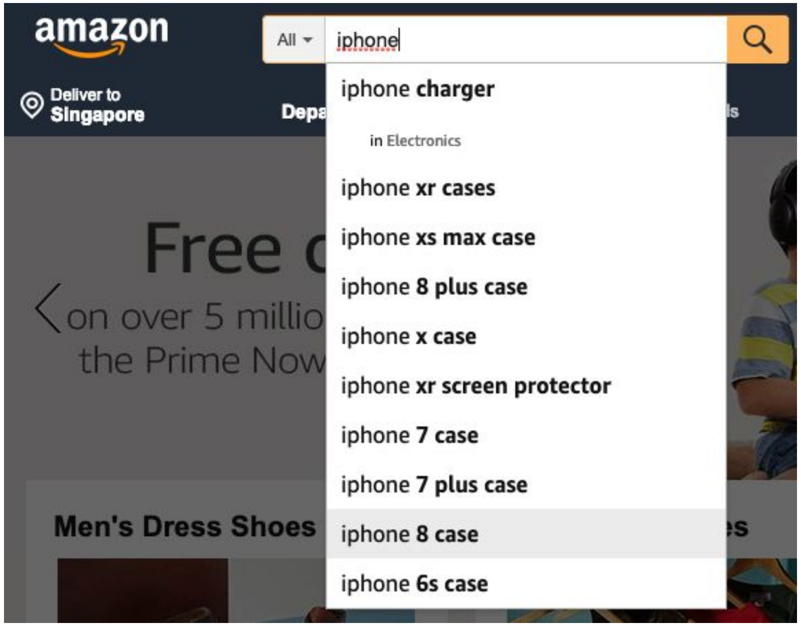
- **Sensitivity/Recall:** Correctly predicted as positives compared to total number of positives
  = TP/(TP + FN) = 100/(100+5) = 0.95

Note: Mostly we have to pick one over other, it's almost impossible to have both high Precision and Recall.

- **Specificity:** Correctly predicted as negatives compared to total number of negatives **= TN/(TN + FP)** = 50/(50+10) = 0.83

Understanding Precision and Recall

Think about the search box on Amazon home page.

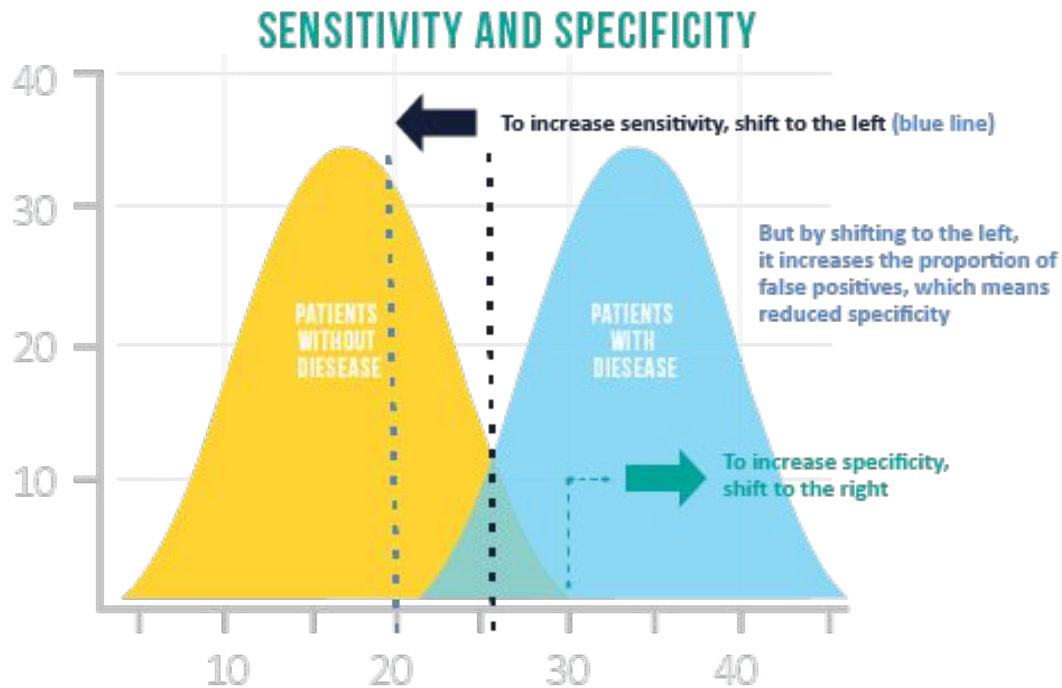


The precision is the proportion of relevant results(correctly predicted yes) in the list of all returned search results(total predicted yes).

The recall is the ratio of the relevant results(correctly predicted yes) returned by the search engine to the total number of the relevant results that could have been returned (total actual yes).

Choosing between Sensitivity and Specificity

Often, the sensitivity and specificity of a test are inversely related. Selecting the optimal balance of sensitivity and specificity depends on the objective of the problem that needs to be solved.

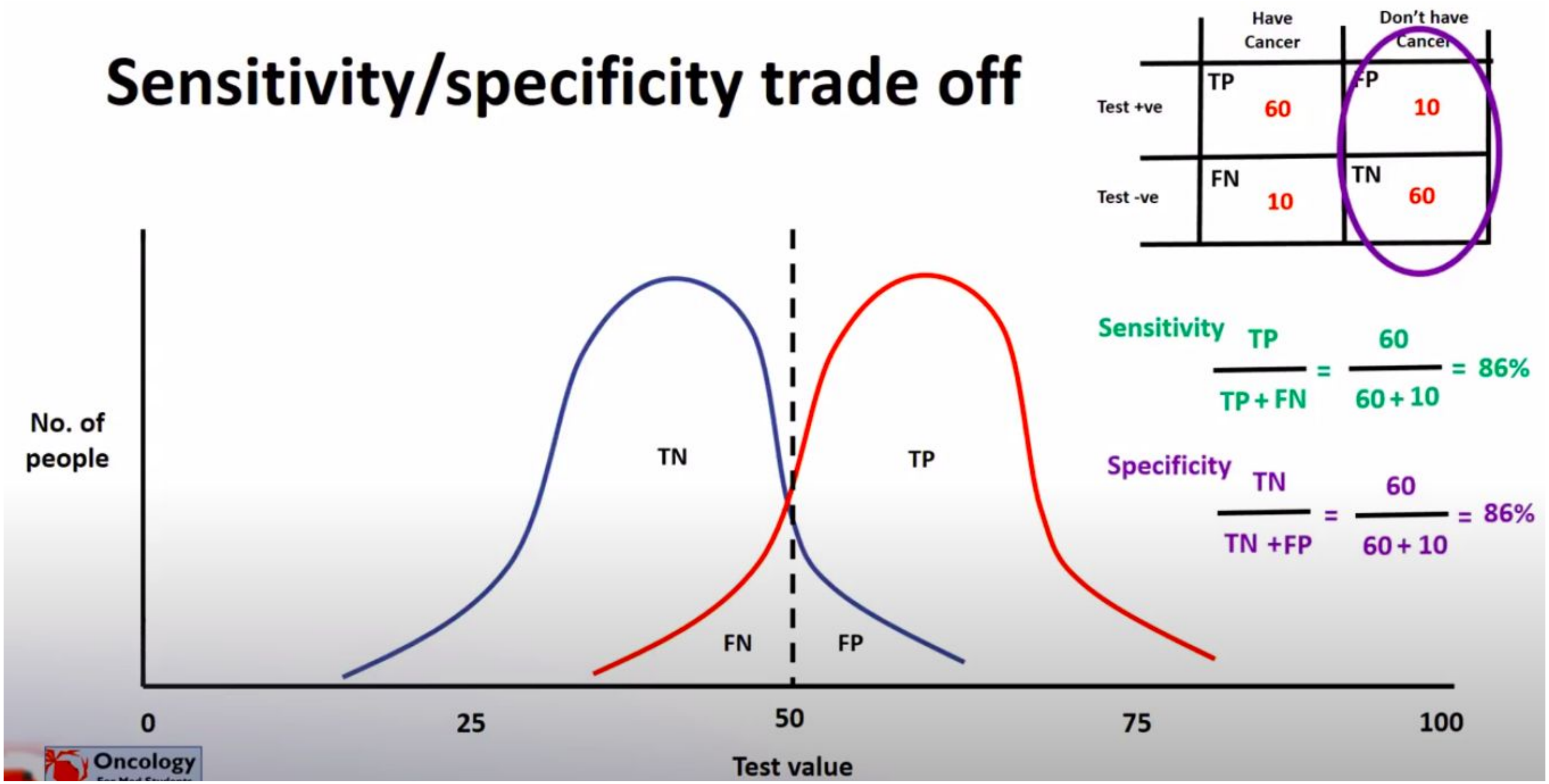


If correctly identifying positive class is important for us, then we should choose a model with higher Sensitivity. However, if correctly identifying negative class is more important, then we should choose specificity as the measurement metric.

Sensitivity or Specificity - an example

Let's say we are predicting if a patient has cancer or not. The default probability threshold is kept at 0.5 i.e

Class 0 (No cancer) – Below 0.5
Class 1 (Cancer) – Above 0.5

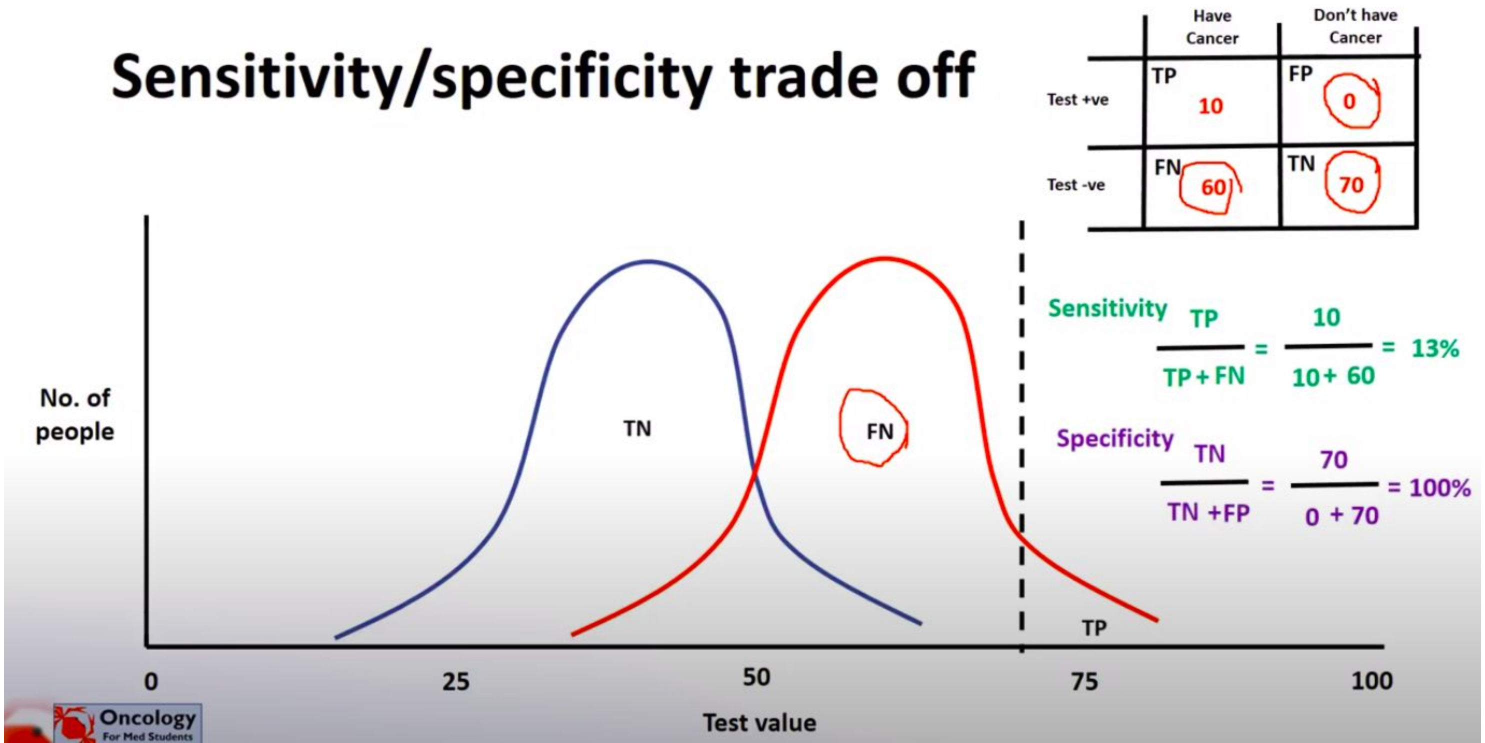


Case 1: Higher Specificity

Suppose we want to predict Class 1 (Ci.e patient has cancer) only if we are VERY confident. (To avoid giving the patient a shock and to avoid unnecessary treatment)

We can instead change this threshold to 0.7. Thus, we'll tell someone they have cancer only if we think they have greater than or equal to 70% chance of having a cancer.

Look at the graph below. SInce the threshold has shifted to the right, so the number of people correctly guessed as having cancer have increased. Thus, the specificity has increased. (We are being very specific with declaring patients with cancer).

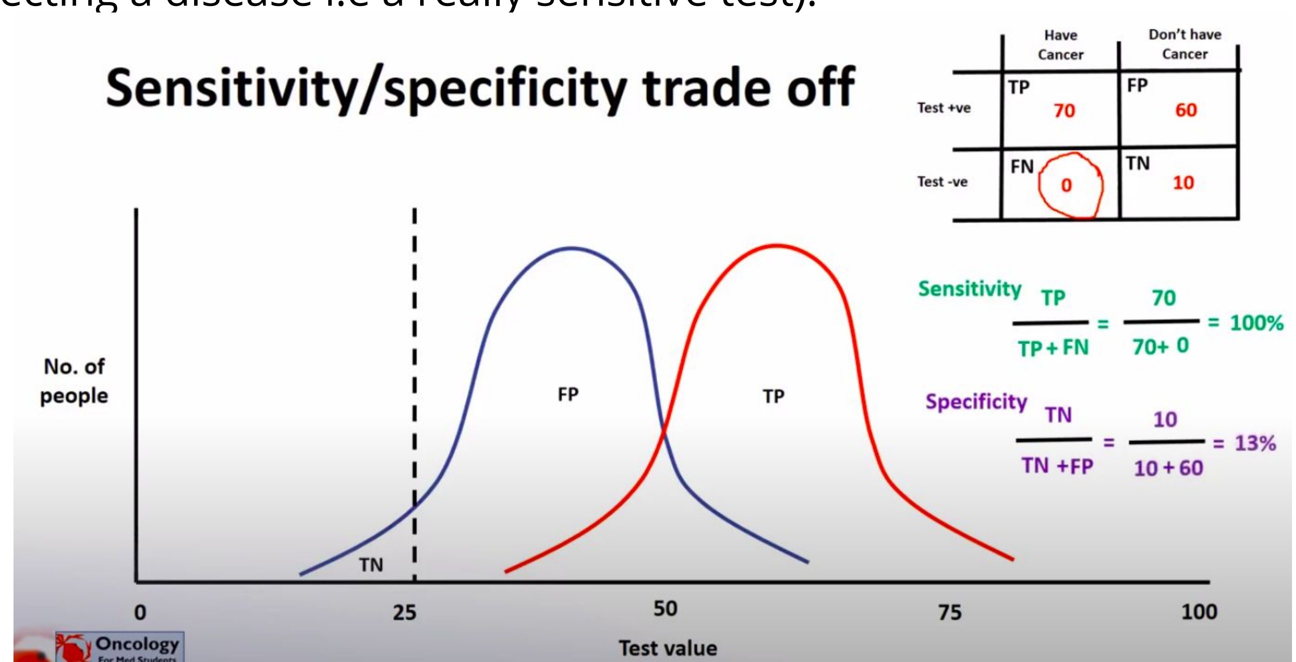


Case 2: Higher Sensitivity

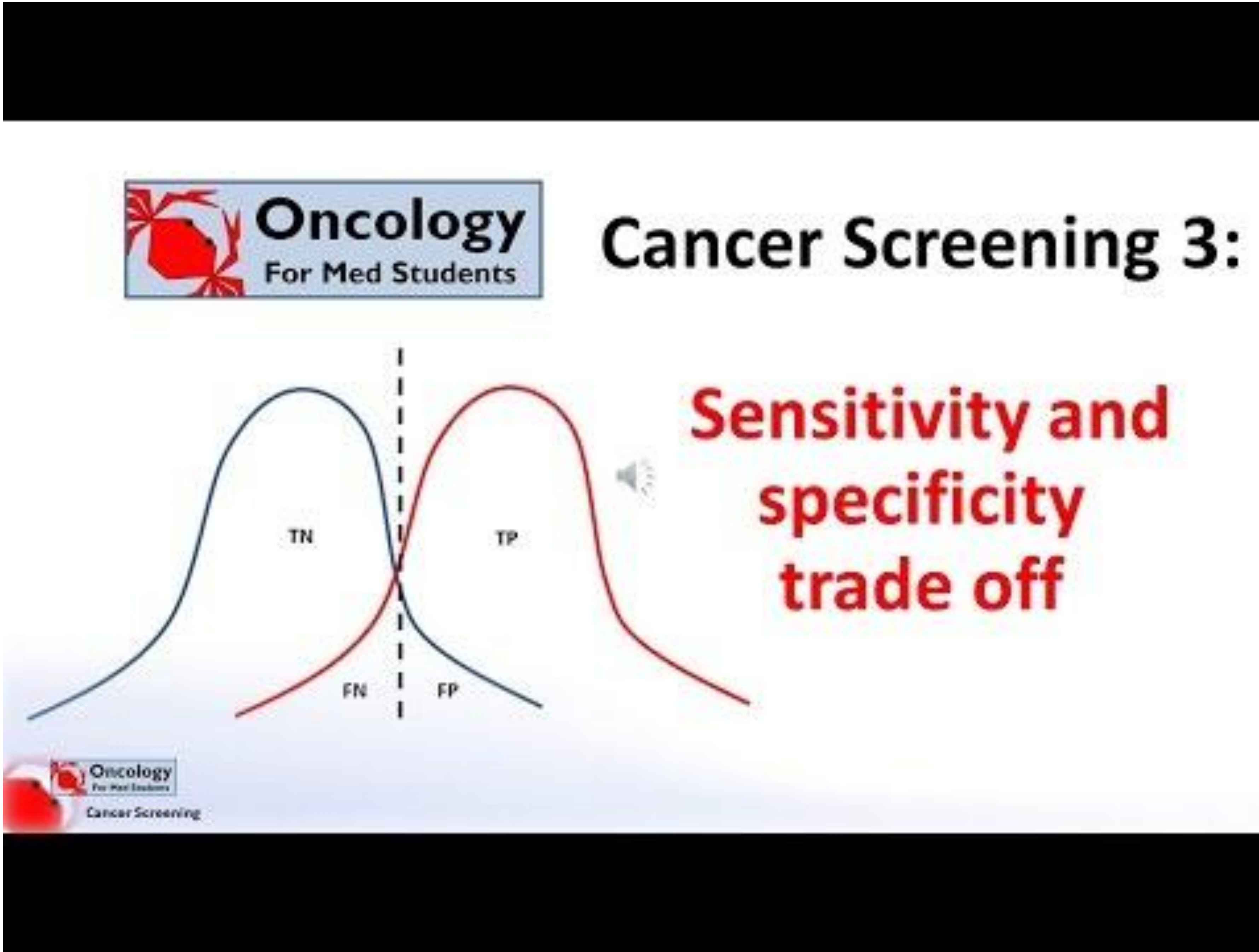
Suppose we want to avoid missing too many cases of cancer (avoid false negatives). If a person with cancer is told that he's well, it can cause a delay in treatment and affect the health badly).

In this case we can set a lower threshold, say 0.25. Even if a patient has 25% chance of having cancer, we'll inform him/her.

Looking at the graph you can see that the threshold has shifted to the left. Most of the people with cancer will be detected in advance in this case. We have completely (or almost) eliminated False Negatives. It will thus result in higher Sensitivity/ Recall. (We are being sensitive in detecting a disease i.e a really sensitive test).



You can watch this video from 00:58 to 5:32 explaining the Sensitivity and Specificity trade off



Confusion Matrix

Talking about accuracy, our favourite metric!

Accuracy is defined as the ratio of correctly predicted examples by the total examples.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- **Accuracy:** Overall, how often is the classifier correct?
  = (TP+TN)/total = (100+50)/165 = 0.91

- Remember, accuracy is a **very useful metric when all the classes are equally important.**

- But this might not be the case if we are predicting if a patient has cancer. In this example, we can probably tolerate FPs but not FNs.

- If a cancerous patient is wrongly reported as being fine, it can result in delaying of treatment. Which is not good!

Slide Download Link

- You can download the slides here: https://docs.google.com/presentation/d/1_3HAVN1BooheCPzn0K8VCGyOk61jxP88NH1_p5DS9OE/edit?usp=sharing

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack