Classification Performance Metrics

Part One: Confusion Matrix Basics

Ever heard of terms:

“false positive” or “false negative” or “accuracy”?

Say we've developed model to detect presence of a virus infection in a person based on some biological feature.

Assume this is a Logistic Regression, predicting:

- 0 - Not Infected (Tests Negative)
- 1 - Infected (Tests Positive)

Unlikely that our model will perform perfectly. This means there are 4 possible outcomes:

- Infected person tests positive.
- Healthy person tests negative.
  - *Note, these are the outcomes we want! But it is unlikely our test is perfect...*

- Infected person tests positive.
- Healthy person tests negative.
- Infected person tests negative.
- Healthy person tests positive.

Based off these 4 possibilities, there are many error metrics we can calculate.

Confusion Matrix

	ACTUAL	
	INFECTED	HEALTHY

Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		

Confusion Matrix

ACTUAL

		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		

Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		TRUE NEGATIVE

Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
	HEALTHY		TRUE NEGATIVE

Confusion Matrix

		ACTUAL: INFECTED	ACTUAL: HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
PREDICTED	HEALTHY	FALSE NEGATIVE	TRUE NEGATIVE

- Imagine a test group of 100 people
- 5 are infected. 95 are healthy.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		

We tested all of them and got these results:

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Accuracy?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

- Calculating accuracy:

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Is this a good value for accuracy?

The accuracy paradox...

		ACTUAL: INFECTED	ACTUAL: HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Imagine we **always** report back "healthy"

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Imagine we **always** report back “healthy”

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(0+95)/100 = 95% Accuracy

- Accuracy:
  - How often is the model correct?

95% accuracy for a model that always returns “healthy”!

This is the accuracy paradox!

- Classifiers dealing with **imbalanced** classes has to confront the issue of the accuracy paradox.
- **Imbalanced** classes will always result in a distorted accuracy reflecting better performance than what is truly warranted.

Imbalanced classes are often found in real world data sets.

- Medical conditions can affect small portions of the population.
- Fraud is not common (e.g. Real vs. Fraud credit card usage).

- If a class is only a small percentage (**n%**), then a classifier that always predicts the majority class will always have an accuracy of (1-n).
- In our previous example we saw infected were only 5% of the data.
- Allowing the accuracy to be 95%.

This means we shouldn’t solely rely on accuracy as a metric!

This is where precision,recall, and f1-score will come in.

Classification Performance Metrics

Part Two: Precision and Recall

- We already know how to calculate accuracy and its associated paradox.
- Let’s explore three more metrics that can help give a clearer picture of performance:
    - Recall (a.k.a. sensitivity)
    - Precision
    - F1-Score

Let’s begin with recall.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall =
(TP)/Total Actual Positives

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall =
(4)/5

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall = 0.8

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

What’s the recall if we always classify as “healthy”?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

Recall =
(TP)/Total Actual Positives

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

- What’s the recall if we always classify as “healthy”?
- A recall of 0 alerts you the model isn’t catching cases!

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

Recall =
(0)/5 !

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision =
(TP)/6

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision =
(4)/6 = 0.666

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

What's the **precision** if we always classify as "healthy"?

ACTUAL

PREDICTED		INFECTED	HEALTHY
	INFECTED	0	0
	HEALTHY	5	95

Precision =
(TP)/Total Predicted Positives
= 0/0

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

- Recall and Precision can help illuminate our performance specifically in regards to the relevant or positive case.
- Depending on the model, there is typically a trade-off between precision and recall, which we will explore later on with the ROC curve.

Since precision and recall are related to each other through the numerator (TP), we also report the F1-Score, which is the harmonic mean of precision and recall.

The harmonic mean (instead of the normal mean) allows the entire harmonic mean to go to zero if **either** precision or recall ends up being zero.

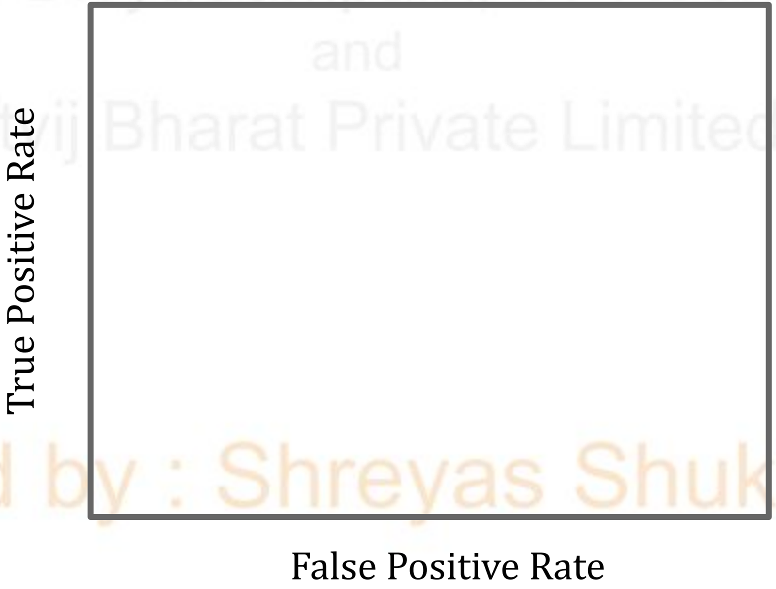
$$F = \frac{2 \times precision \times recall}{precision + recall}$$

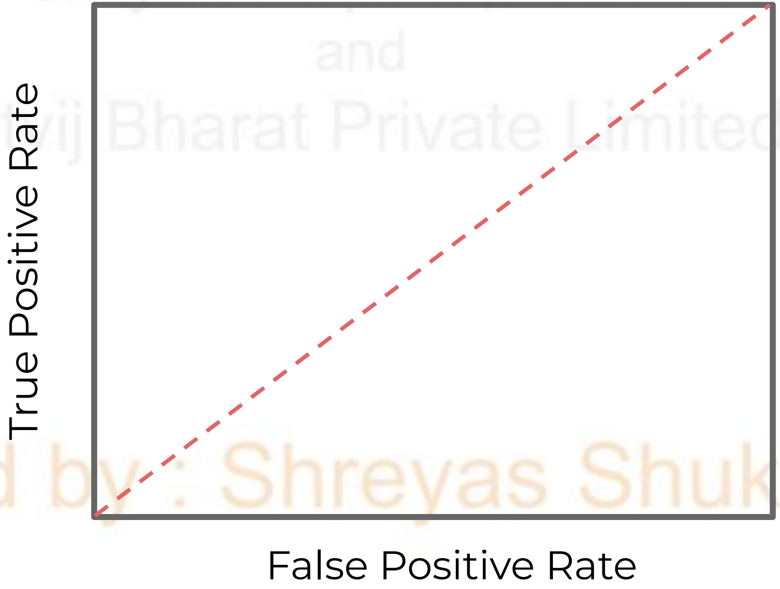
		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

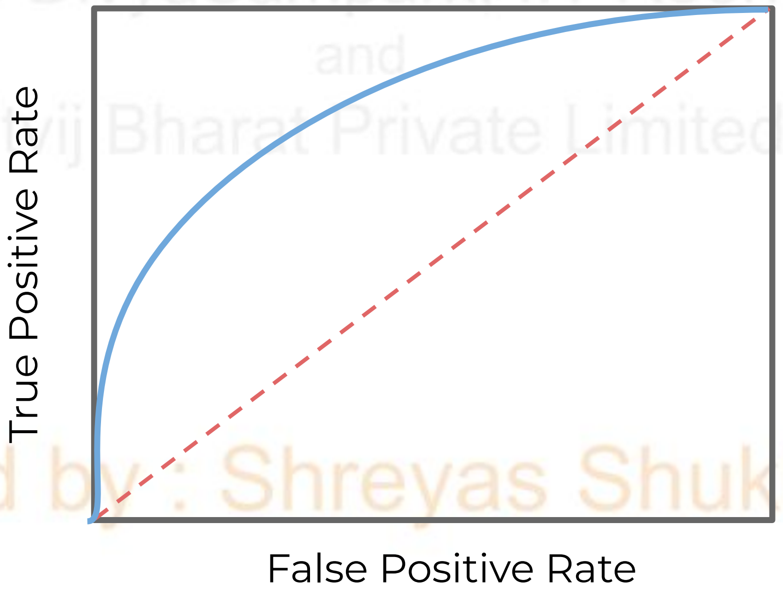
Finally, let's explore a way to visualize the relationships between metrics such as precision and recall with curves.

Classification Performance Metrics

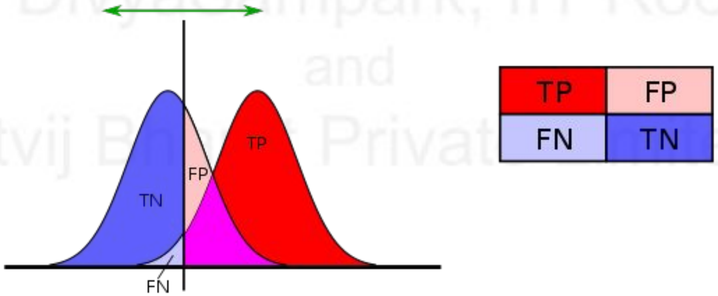
Part Three: ROC Curves

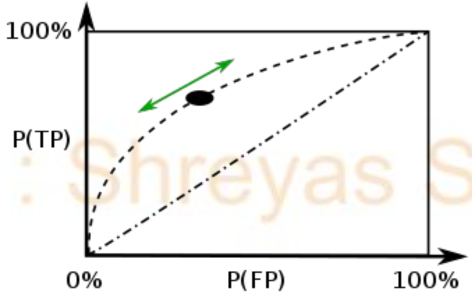

True Positive Rate
False Positive Rate


True Positive Rate
False Positive Rate

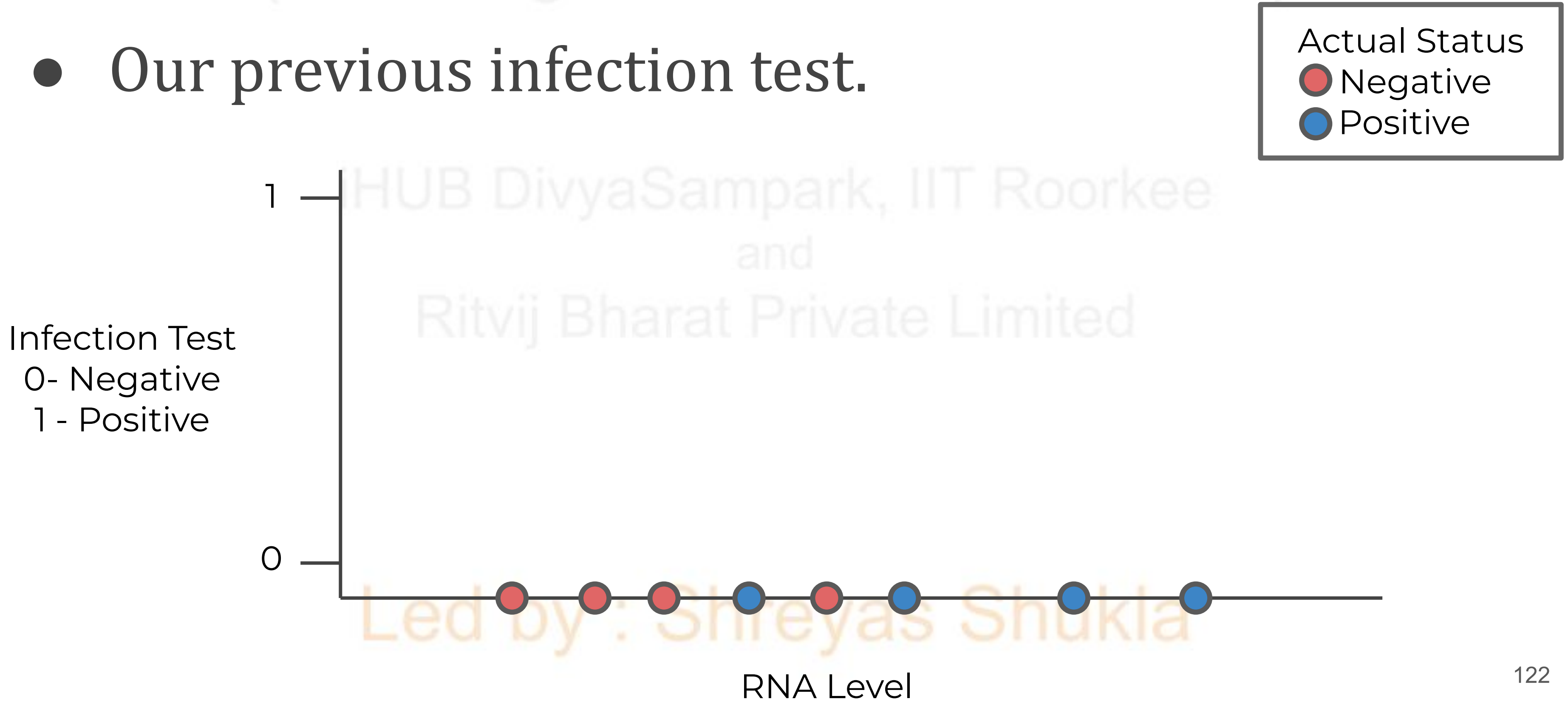

True Positive Rate
False Positive Rate

- There can be a trade-off between True Positives and False Positives.

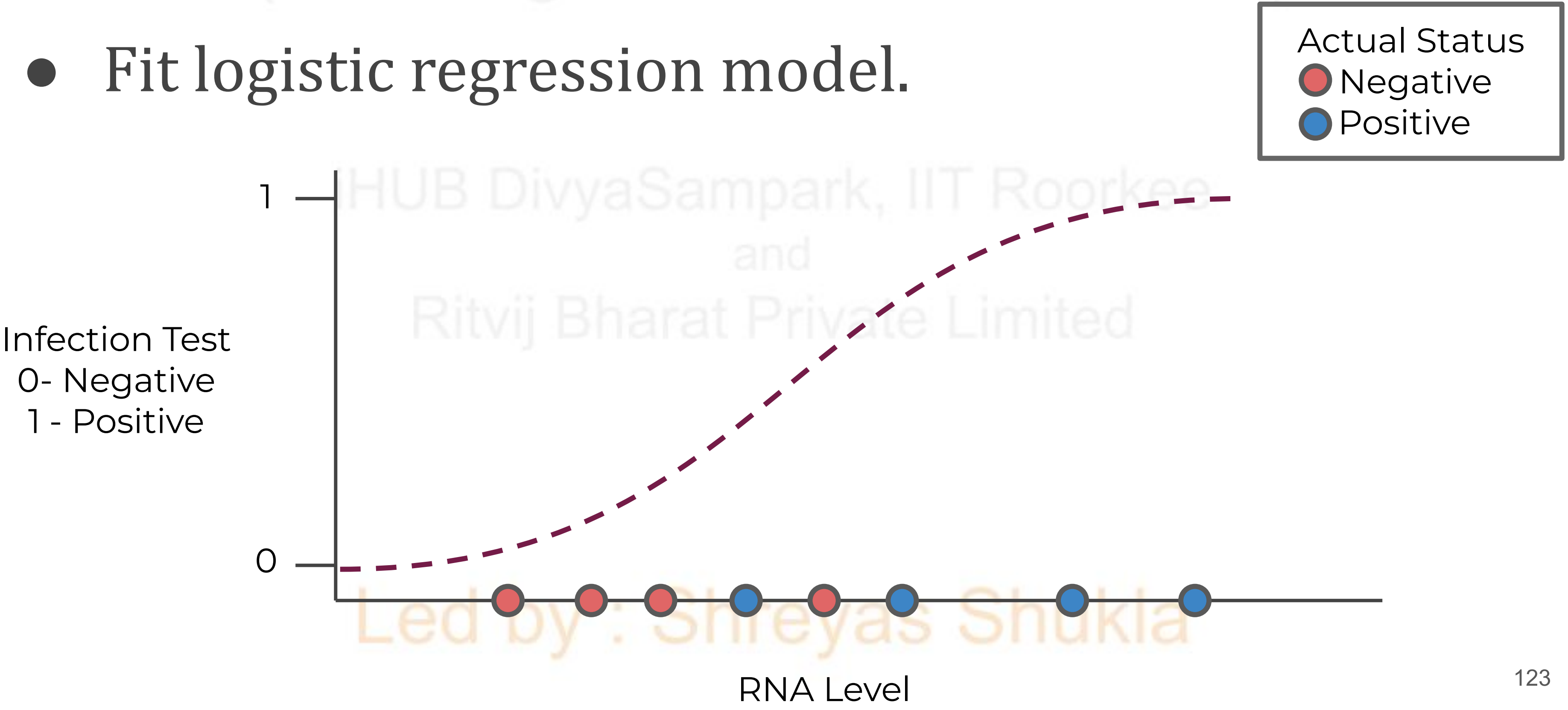




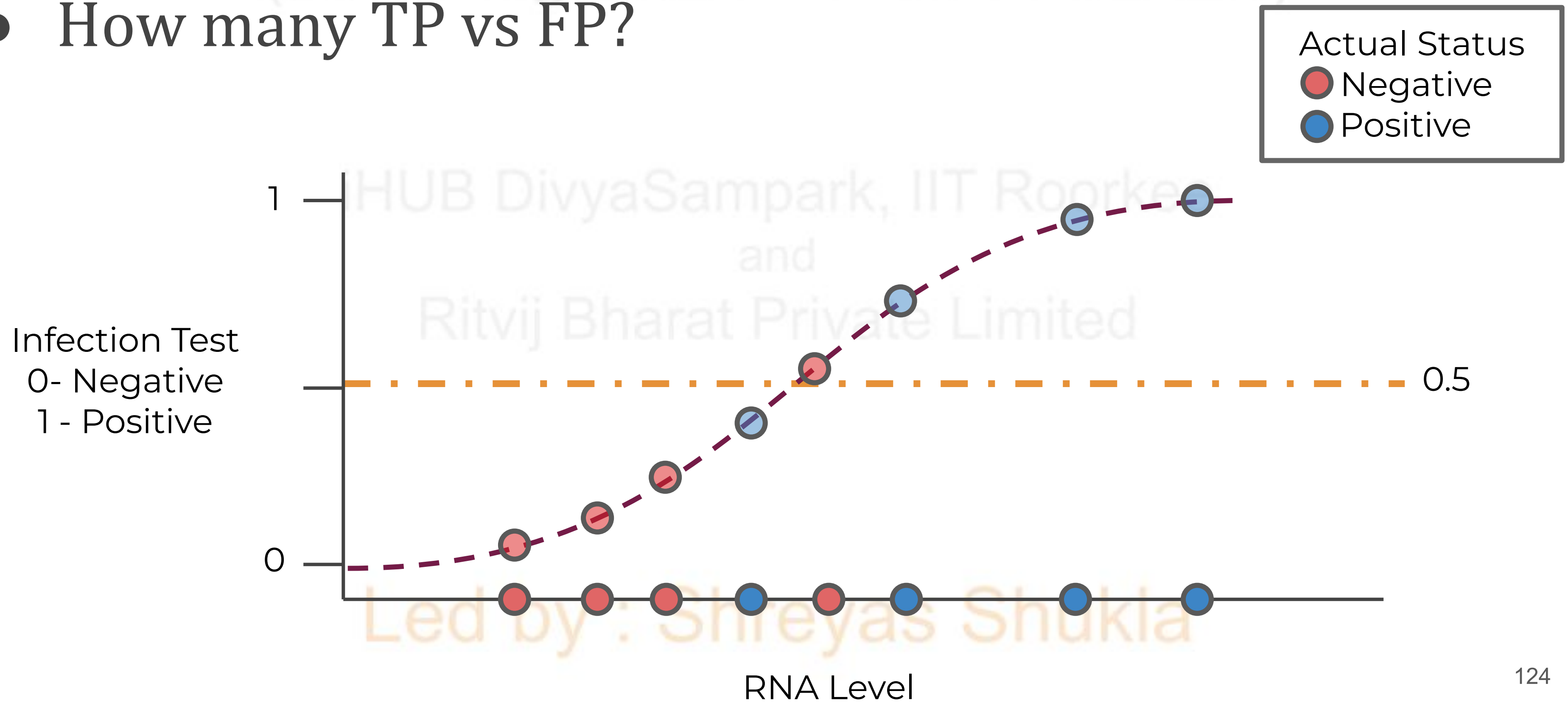
- Our previous infection test.



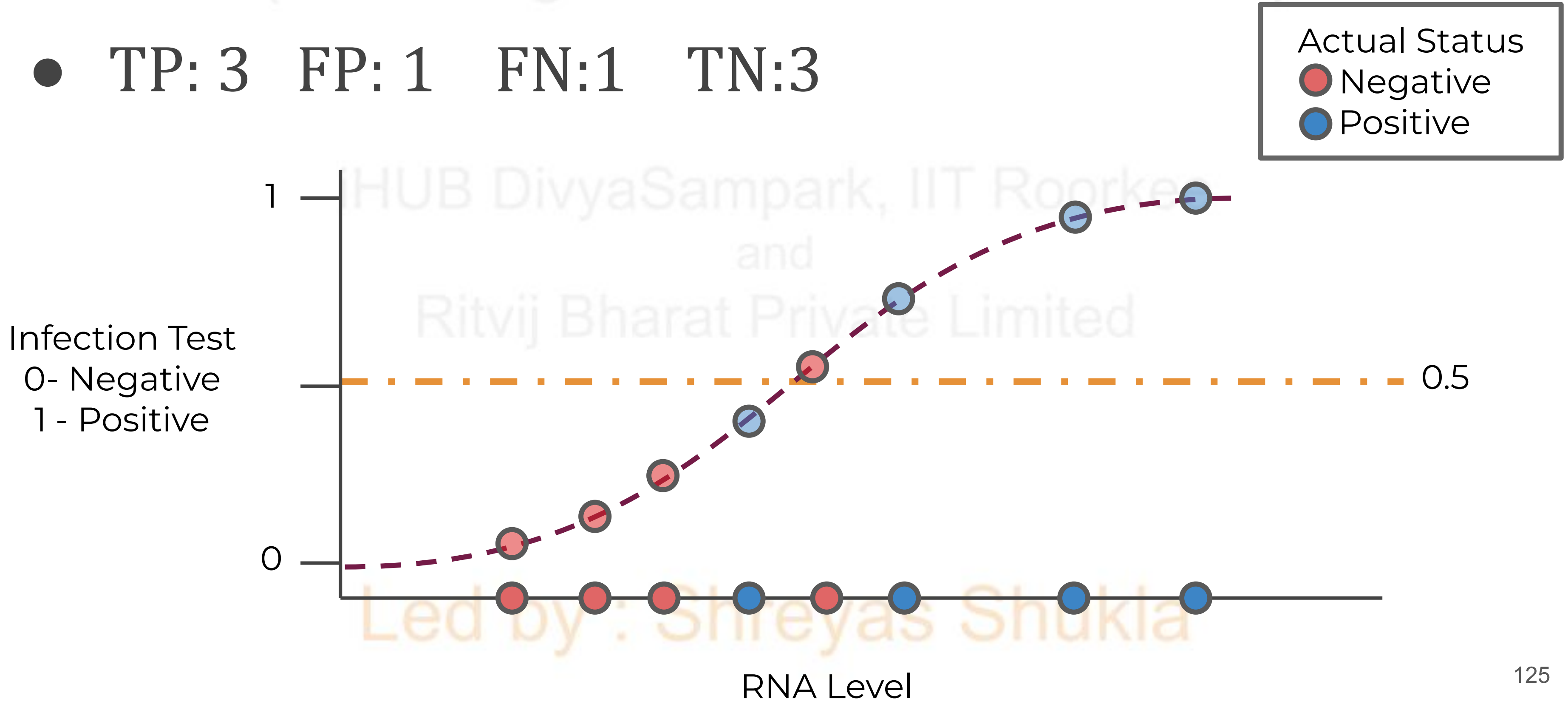
- Fit logistic regression model.



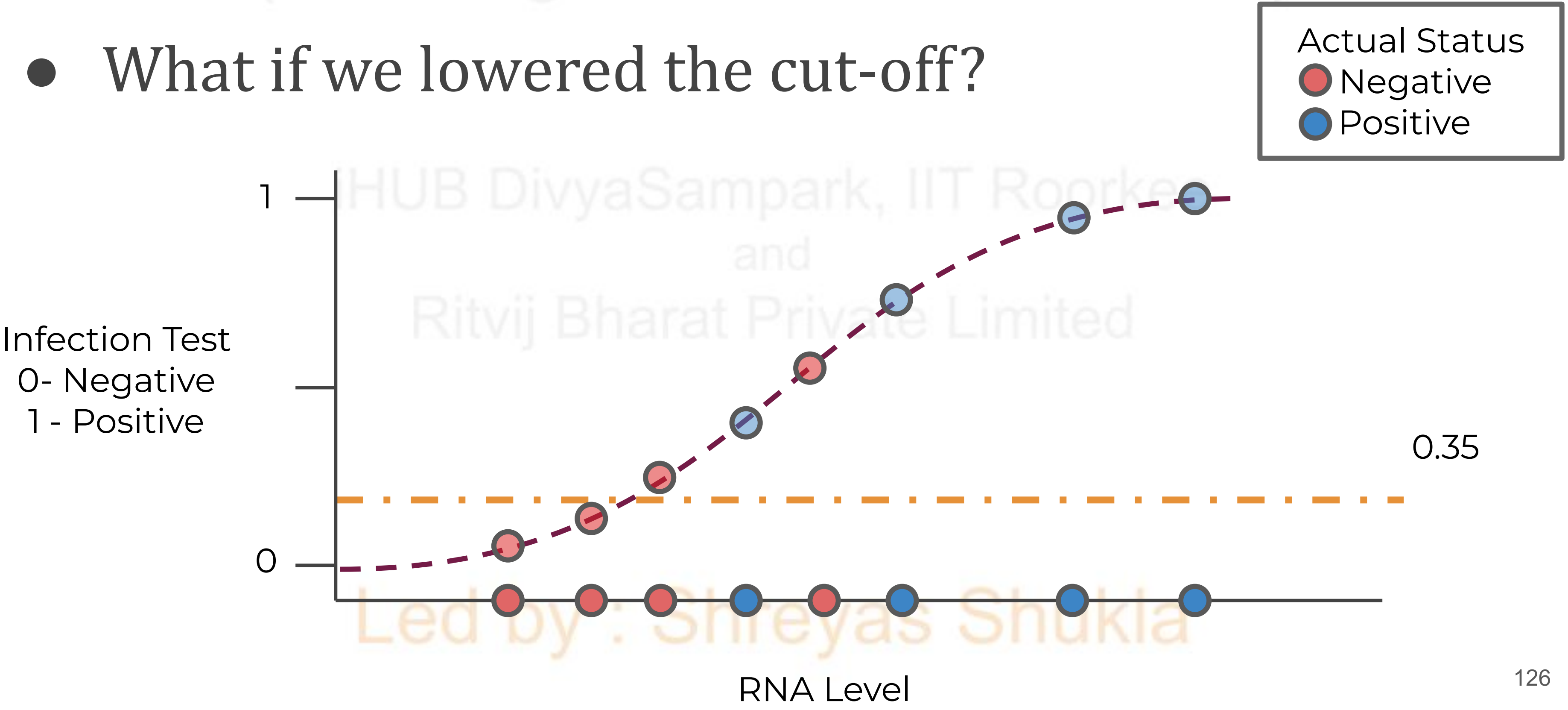
- Given X we predict 0 or 1.
- Default is to choose 0.5 as cut-off.
- How many TP vs FP?



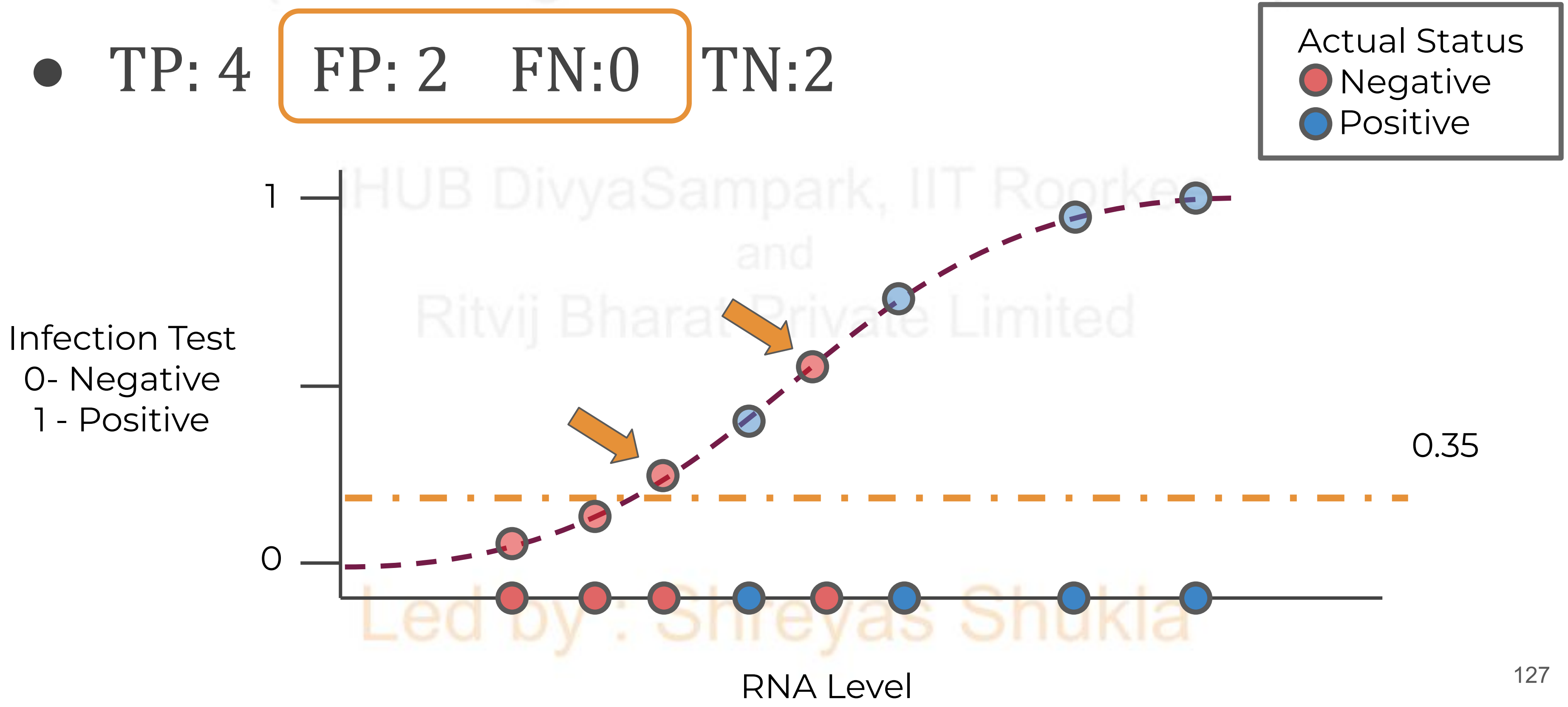
- TP: 3   FP: 1   FN:1   TN:3



- What if we lowered the cut-off?



- TP: 4 FP: 2 FN:0 TN:2



- In certain situations, we accept more false positives to reduce false negatives.
- Imagine a dangerous virus test, we would much rather produce false positives and later do more stringent examination than accidentally release a false negative!

TP: 3 FP: 2 FN:0 TN:3

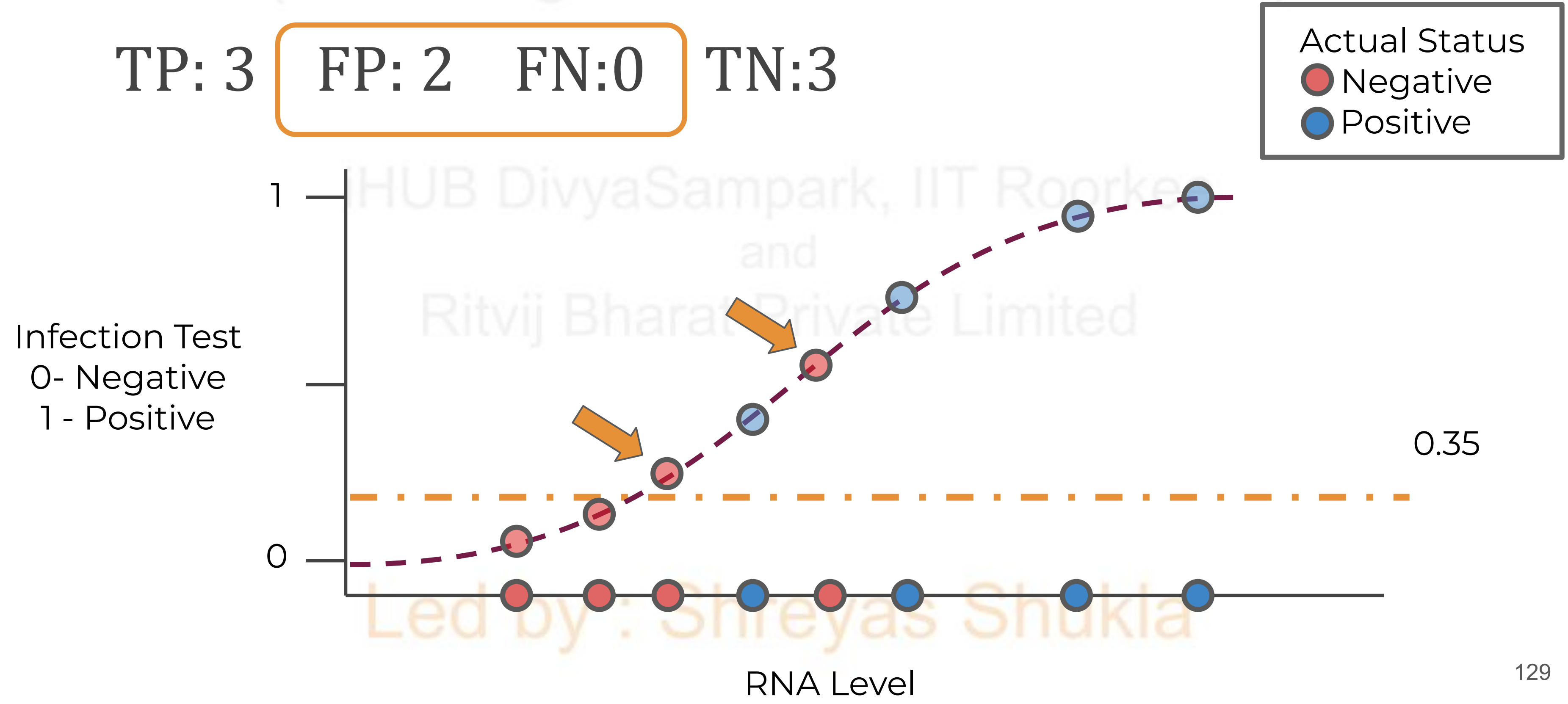
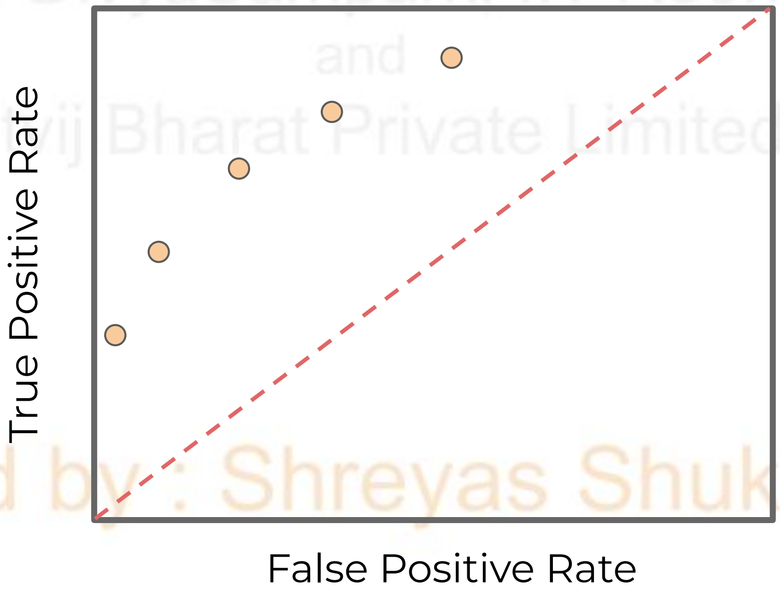
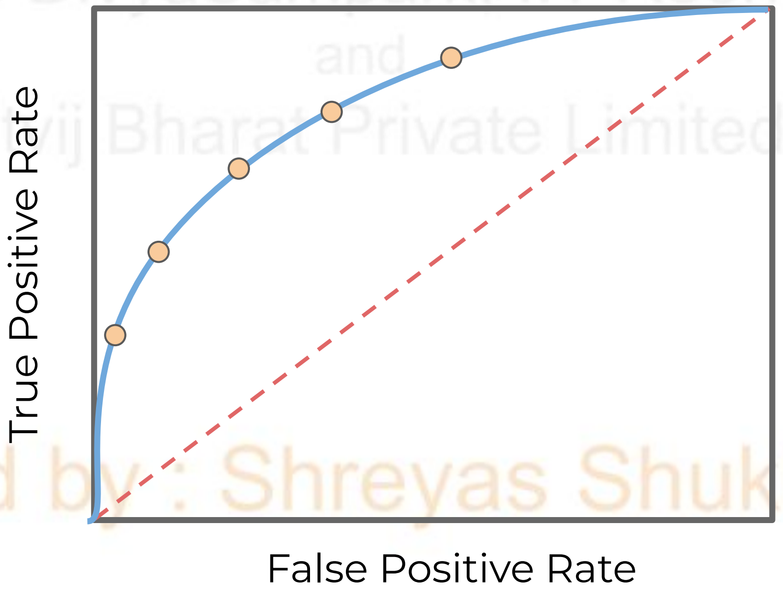


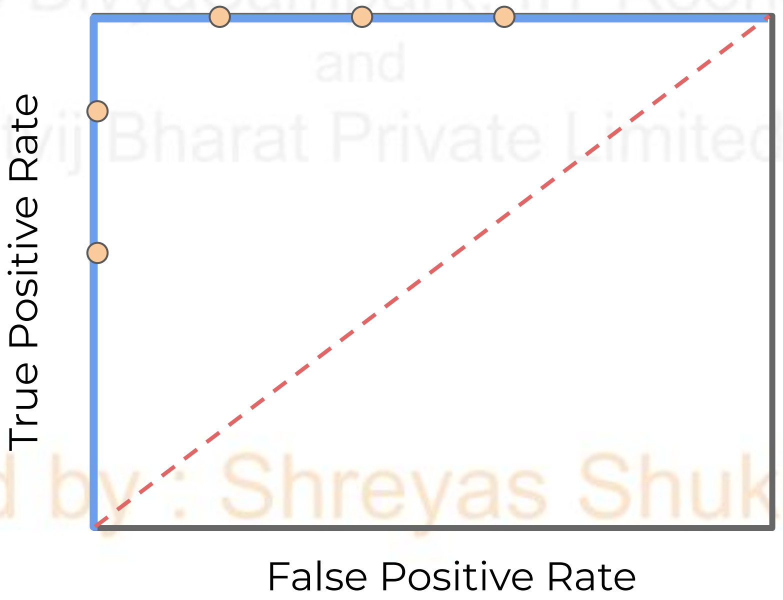
Chart the True vs. False positives for various cut-offs for the ROC curve.



By changing the cut-off limit, we can adjust our True vs. False Positives!

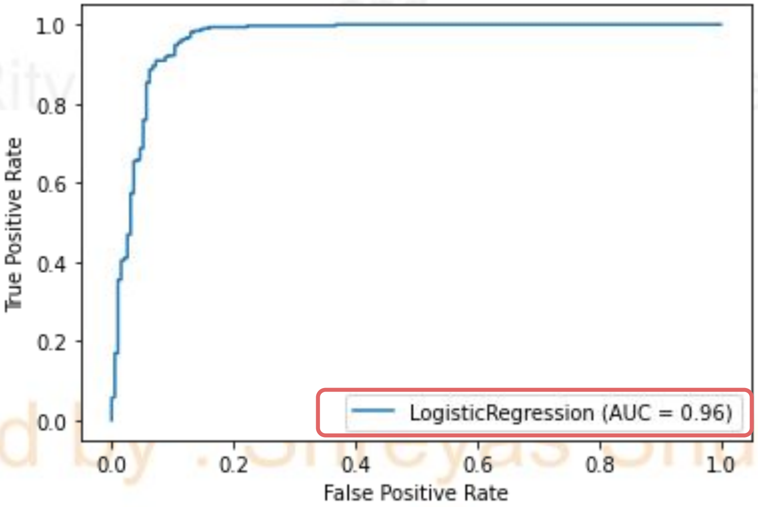


A perfect model would have a zero FPR.
Random guessing is the red line.

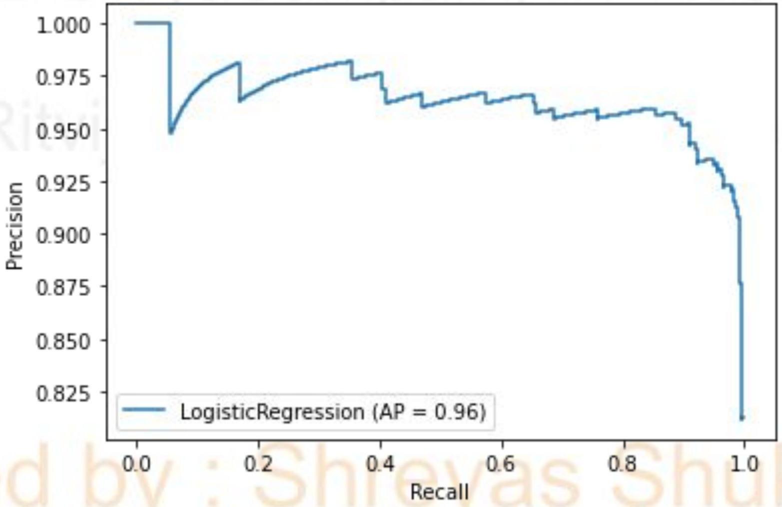


Realistically with smaller data sets the ROC curves are not as smooth.

AUC - Area Under the Curve , allows us to compare ROCs for different models.



Can also create precision vs. recall curves:



AP: Average Precision