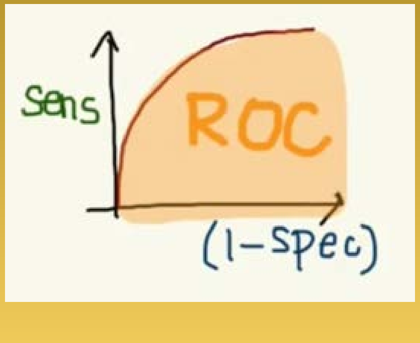
ML EVALUATION



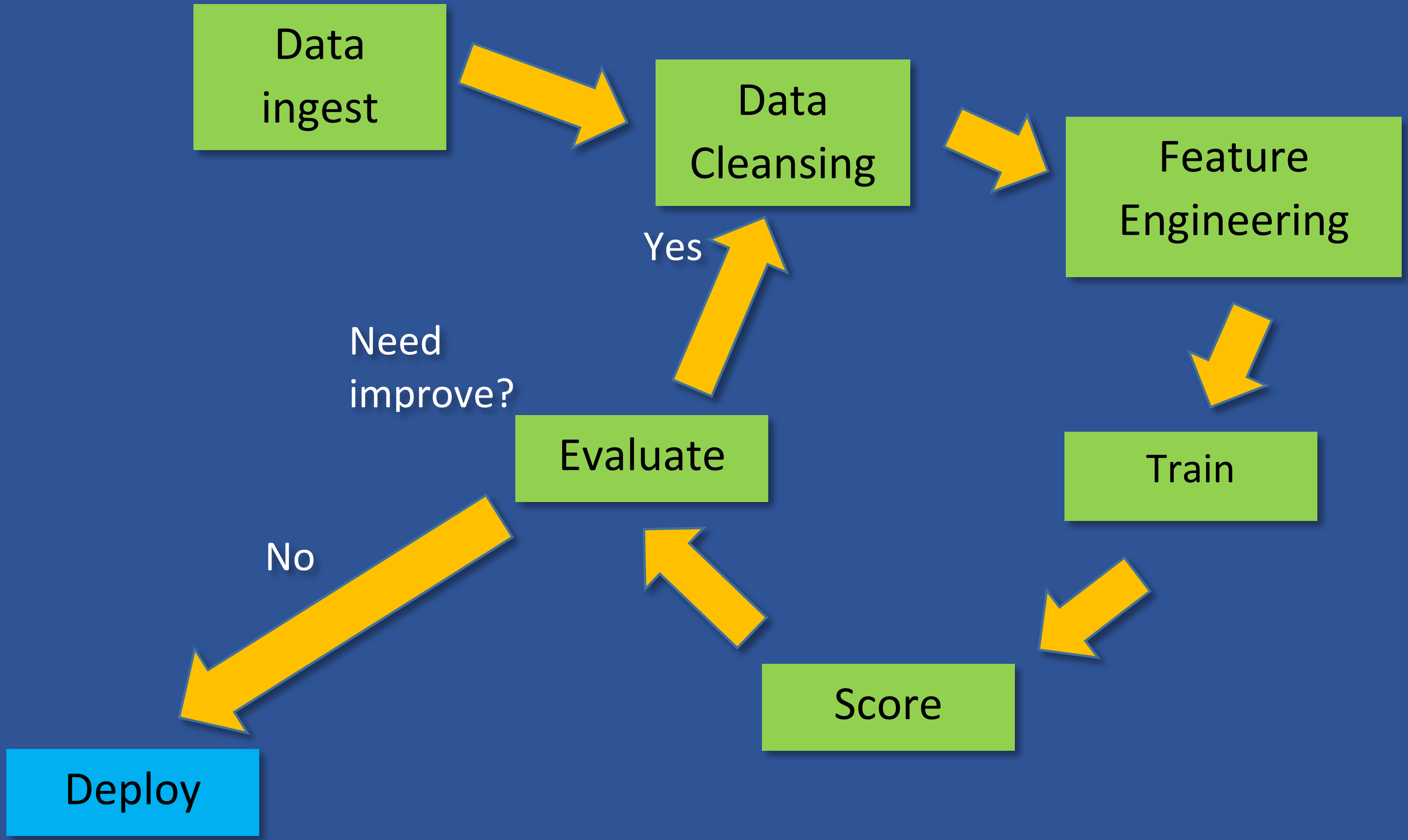
ML Evaluation

In this session

- ML train and evaluation circle
- How to read Histogram
- How to read Box Plot
- Adding Evaluate Model
- How to read ROC curve
- Area Under the Curve (AUC)
- How to read Evaluation metrics

ML Evaluation

ML evaluation circle



Data ingest

Data Cleansing

Feature Engineering

Train

Score

Evaluate

Need improve?

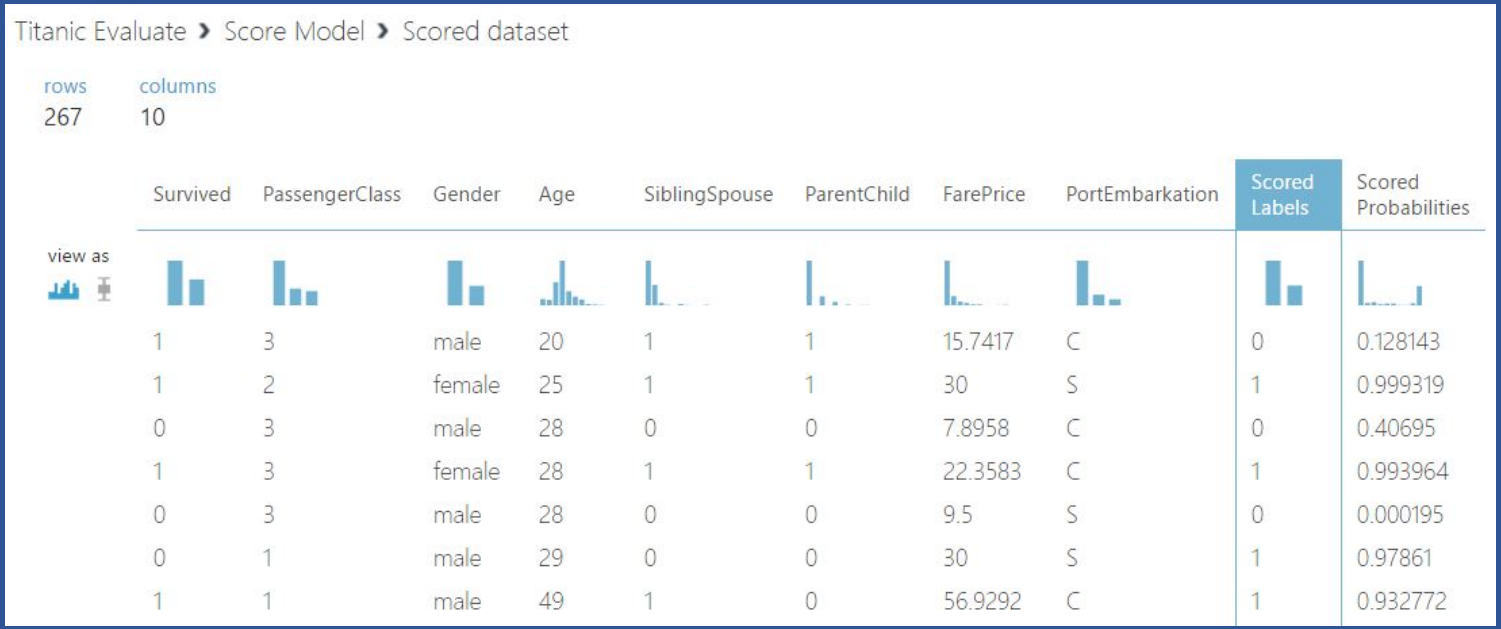
Yes

No

Deploy

ML Evaluation

How to read Scoring results



Titanic Evaluate › Score Model › Scored dataset

rows 267

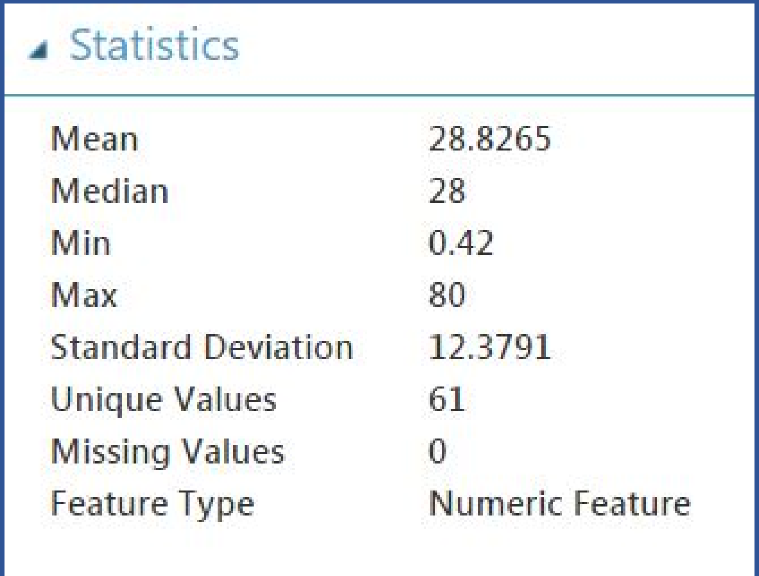
columns 10

Survived	PassengerClass	Gender	Age	SiblingSpouse	ParentChild	FarePrice	PortEmbarkation	Scored Labels	Scored Probabilities
1	3	male	20	1	1	15.7417	C	0	0.128143
1	2	female	25	1	1	30	S	1	0.999319
0	3	male	28	0	0	7.8958	C	0	0.40695
1	3	female	28	1	1	22.3583	C	1	0.993964
0	3	male	28	0	0	9.5	S	0	0.000195
0	1	male	29	0	0	30	S	1	0.97861
1	1	male	49	1	0	56.9292	C	1	0.932772

- This table = Scored dataset
- Row = 267 / Columns = 10
- Total column = 10 / Left 8 = features / Right 2 = prediction results
- Scored Label 0 = dead 1 = survived
- Scored Probabilities (SP) SP <=0.5 == dead / SP > 0.5 == survived

ML Evaluation

How to read Scoring Statistics



◢ Statistics

Mean	28.8265
Median	28
Min	0.42
Max	80
Standard Deviation	12.3791
Unique Values	61
Missing Values	0
Feature Type	Numeric Feature

Show Statistics of the Scored dataset

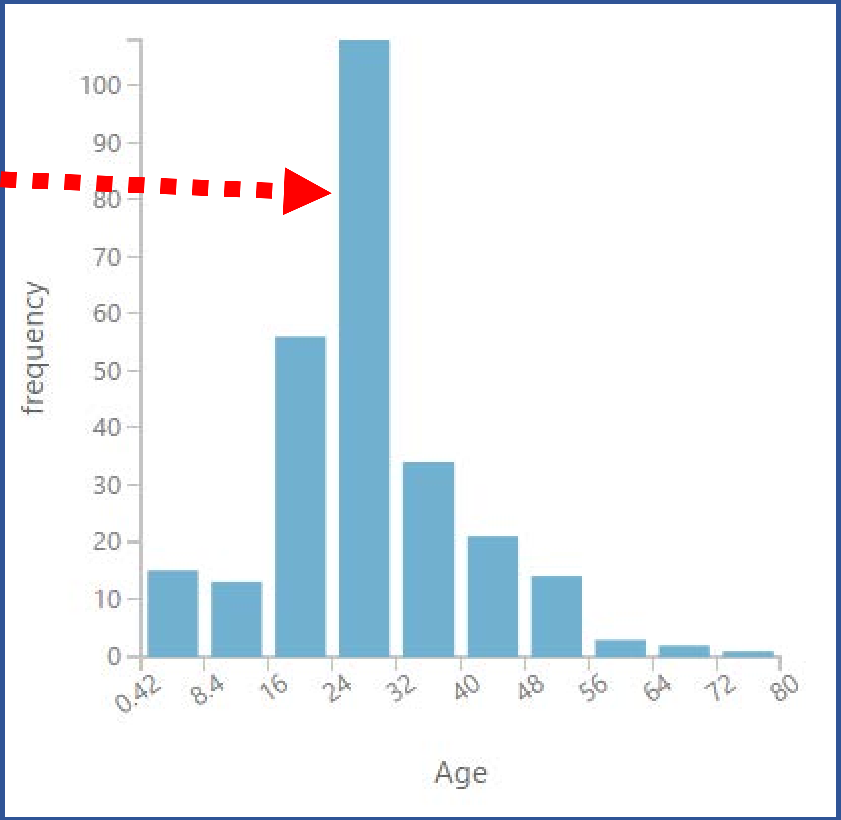
- Mean = Sum of all the values divided by the number of values
- Median = The midpoint of the data after being ranked
- Standard Deviation = The square root of the variance
- Unique Values
- Missing Value

ML Evaluation

How to read Score Histogram

Histogram

- Representation: distribution of numerical data
- Bin: series of intervals (bin)
- Count: values fall into each interval



ML Evaluation

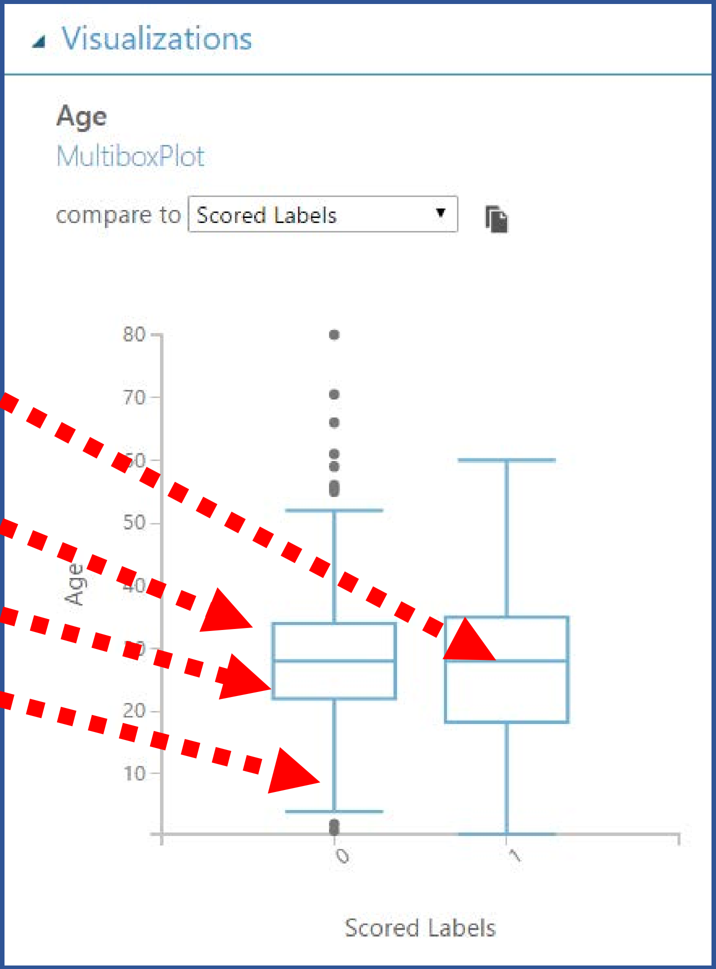
How to read Box Plot

Box Plot

Box Plot (whisker) is a standardized way of displaying the distribution of data

- Median: marks the mid-point of the data
- Box: middle 50% of scores for the group.
- Upper quartile: 75% of the scores fall below the upper quartile.
- Lower quartile: 25% of scores fall below the lower quartile.
- Whiskers: scores outside the middle 50%

0 = dead



Visualizations

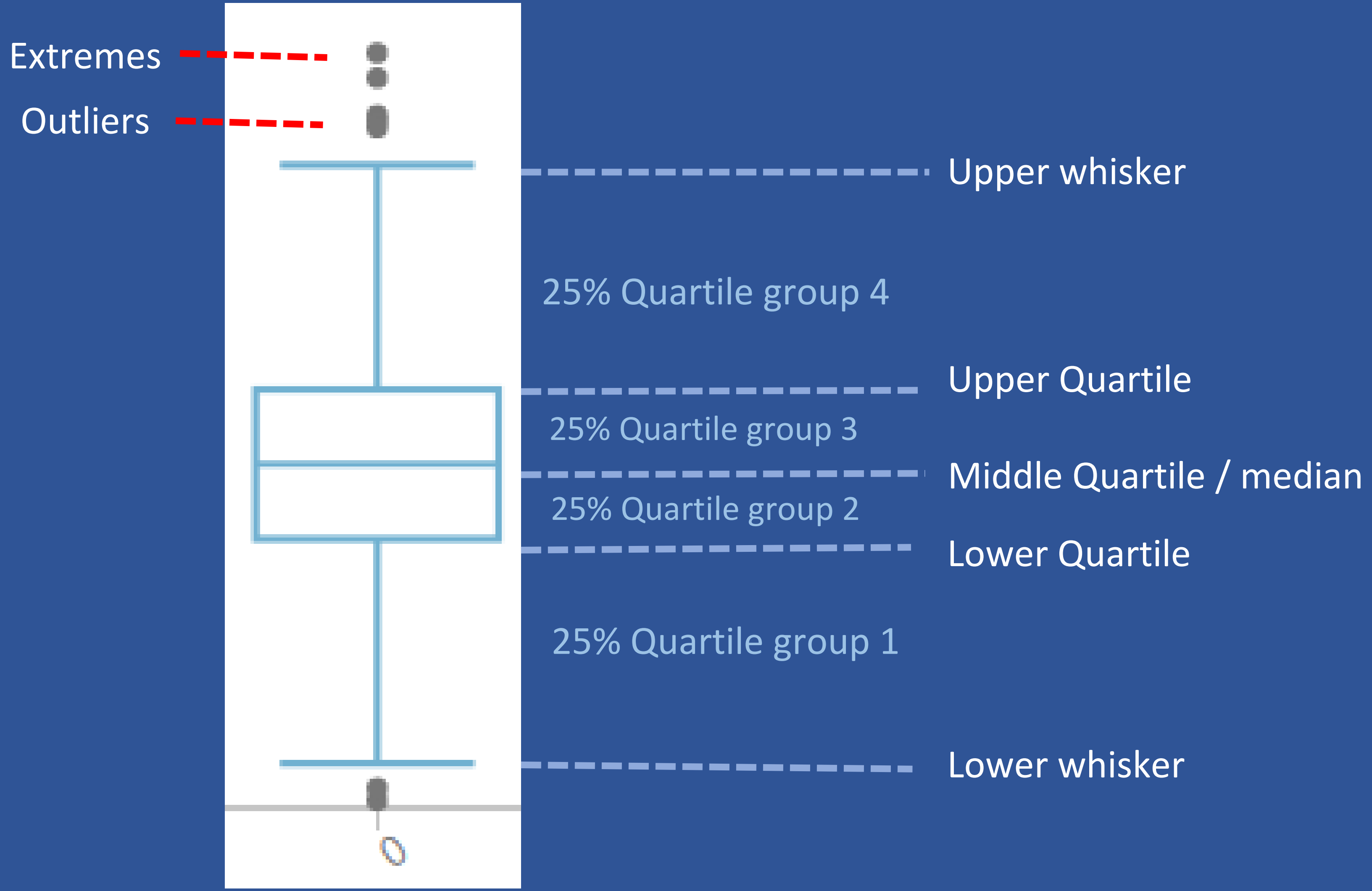
Age

MultiboxPlot

compare to Scored Labels

ML Evaluation

Box Plot Definitions

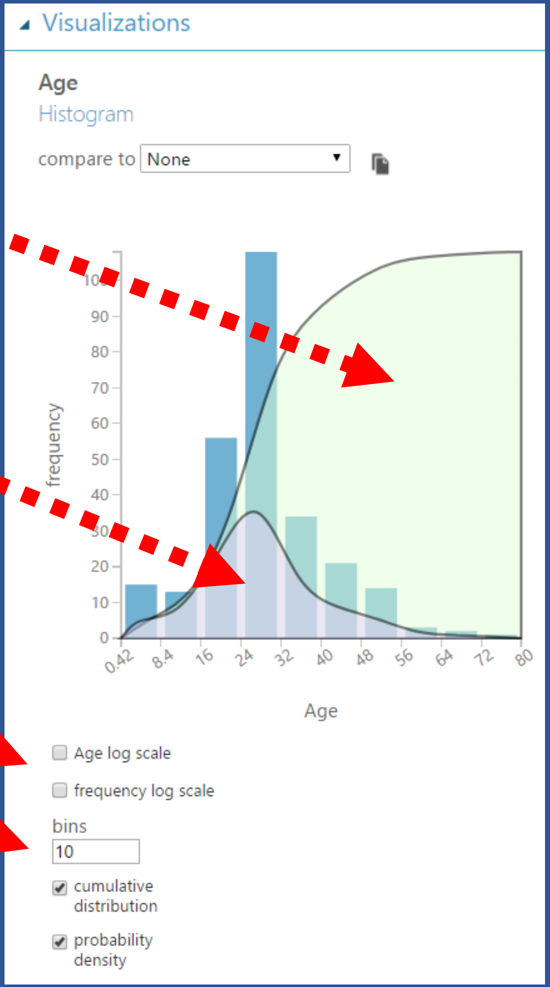


ML Evaluation

Histogram option

Histogram options

- Cumulative distribution function (cdf): shows "How common are samples that are *less than or equal* to this value?"
- Probability density function (pdf): shows "How common are samples at exactly this value?"
- Scale: scaling the distribution
- bins: number of bin



Visualizations

Age

Histogram

compare to None

□ Age log scale

□ frequency log scale

bins 10

☑ cumulative distribution

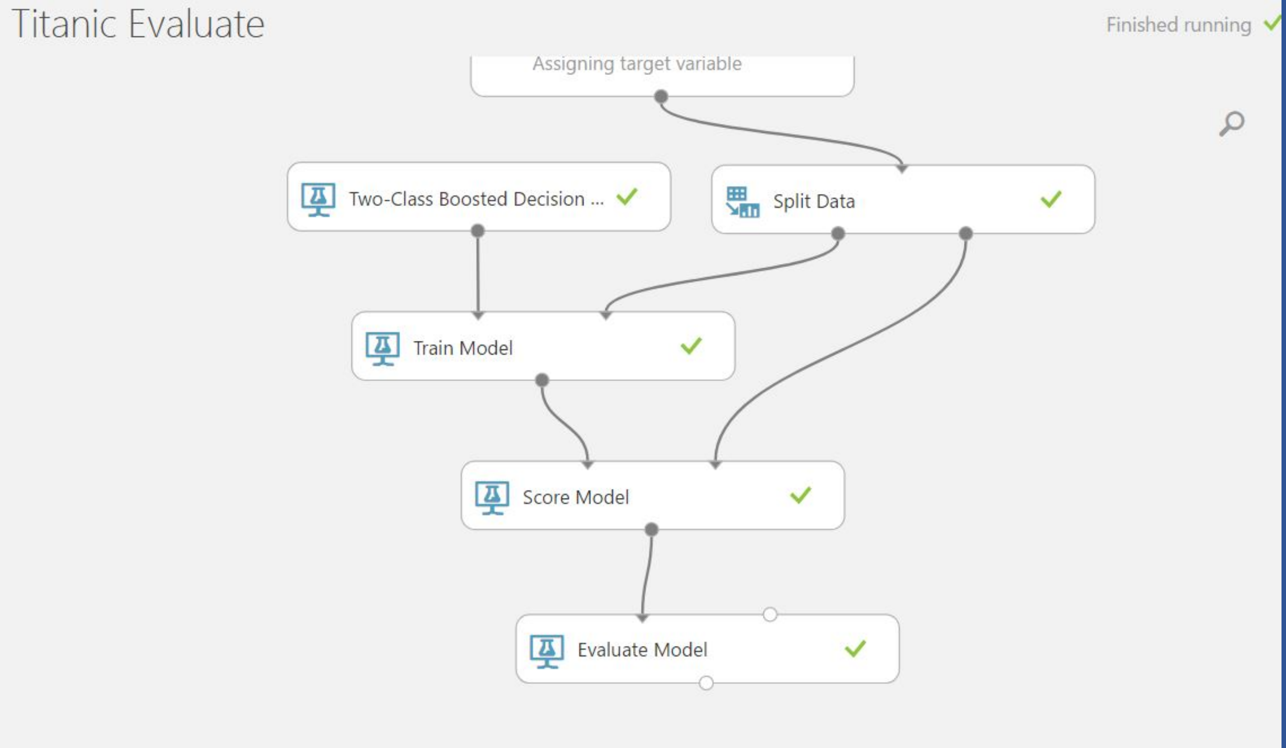
☑ probability density

ML Evaluation

Adding Evaluate Model

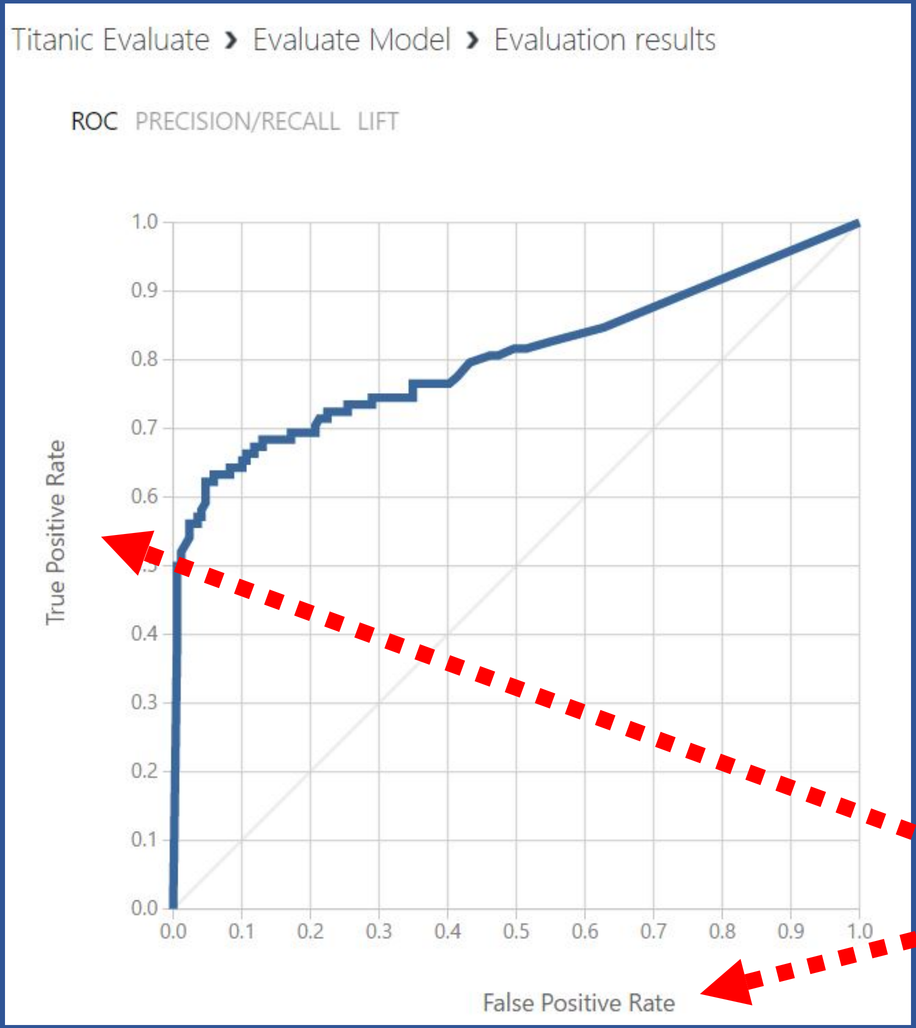
Adding Evaluate Model

1. Open Titanic 1 Experiment
2. Save as Titanic Evaluate
3. Add Evaluate Model
4. Run the Experiment

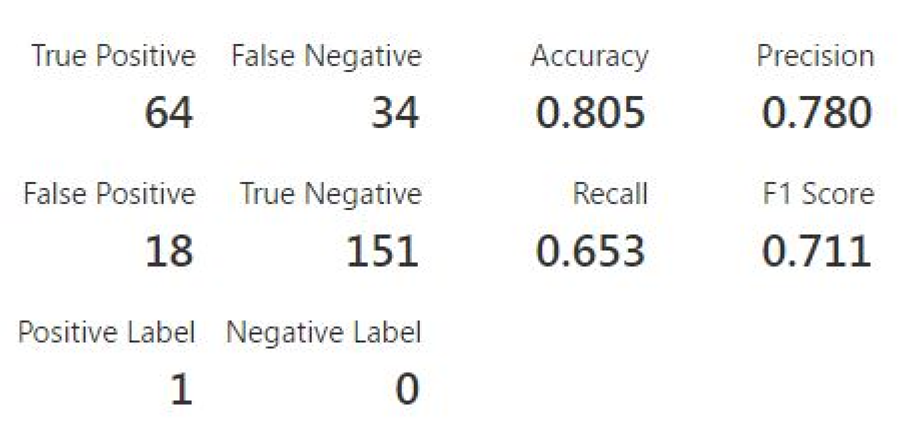


ML Evaluation

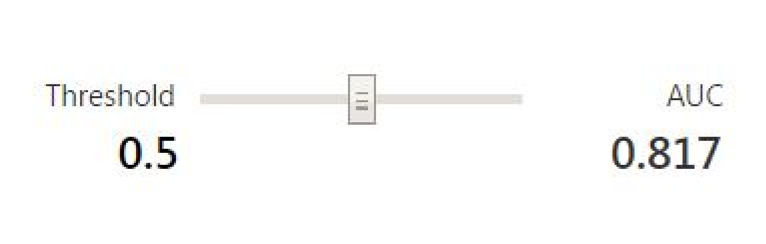
Receiver Operating Characteristic (ROC) Curve



Titanic Evaluate › Evaluate Model › Evaluation results



True Positive	False Negative	Accuracy	Precision
64	34	0.805	0.780
False Positive	True Negative	Recall	F1 Score
18	151	0.653	0.711
Positive Label	Negative Label		
1	0		



Threshold	AUC
0.5	0.817

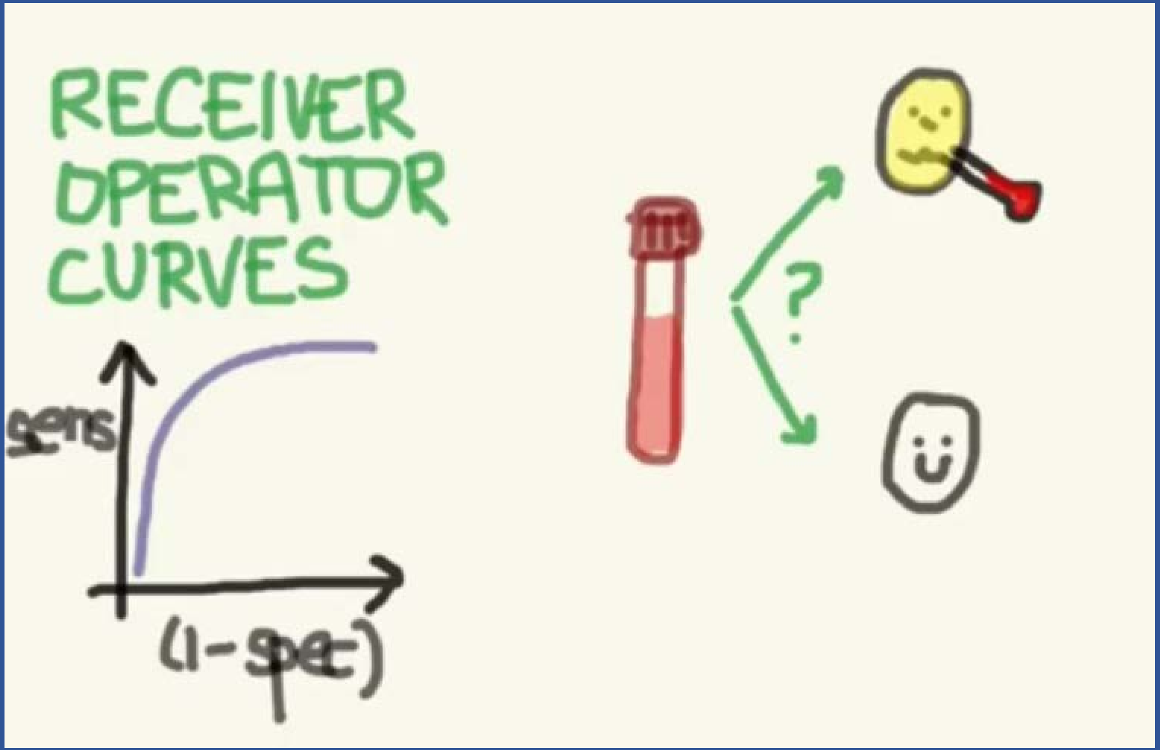
True Positive Rate (TPR)

False Positive Rate (FPR)

ML Evaluation

How to read ROC curve

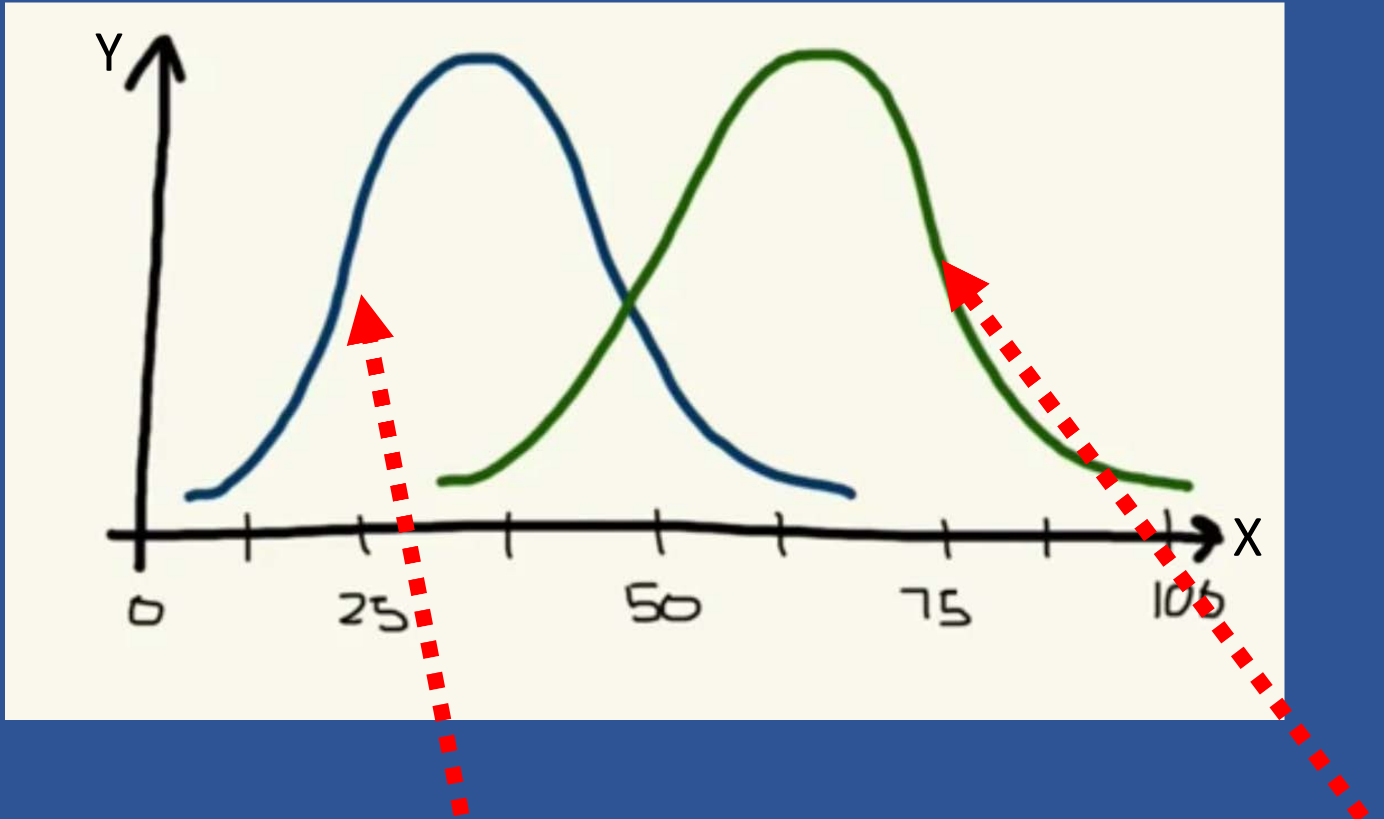
ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.



ROC curve prediction result who have disease who don't

ML Evaluation

Distribution score

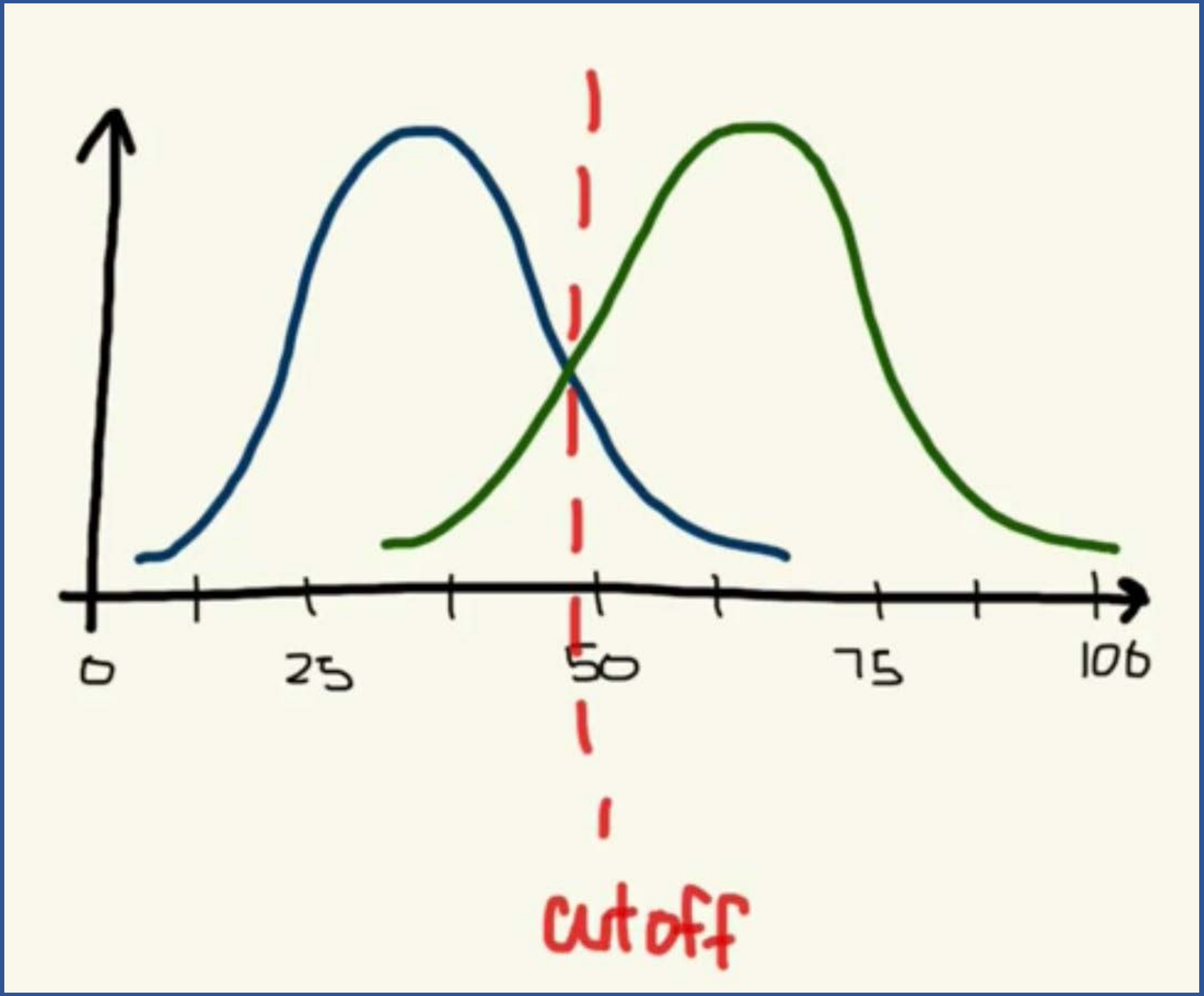


Left distribution = patient who do NOT have disease (survived) / Right = have disease (dead)

x axis = score / y axis = number of patient

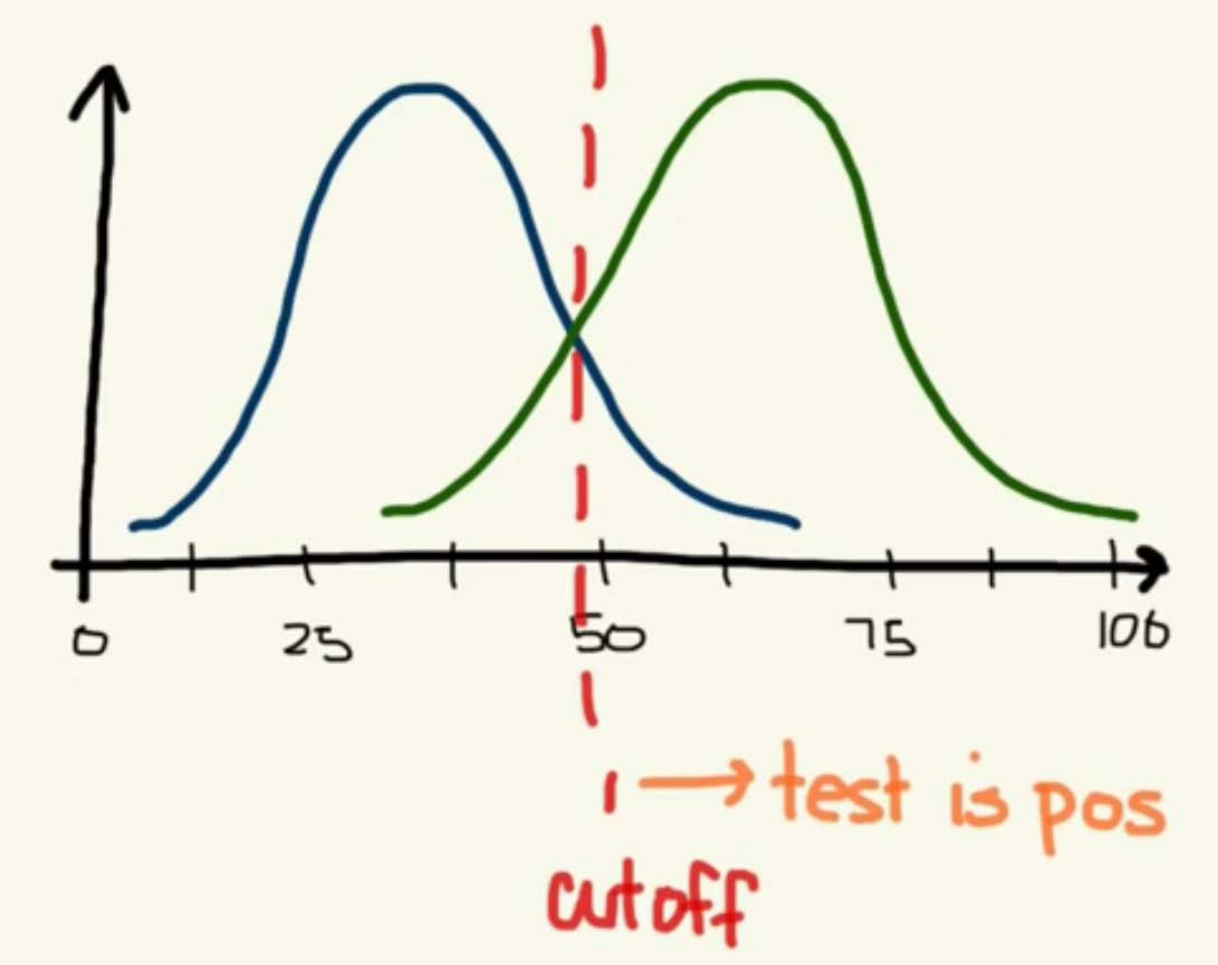
ML Evaluation

Cutoff line



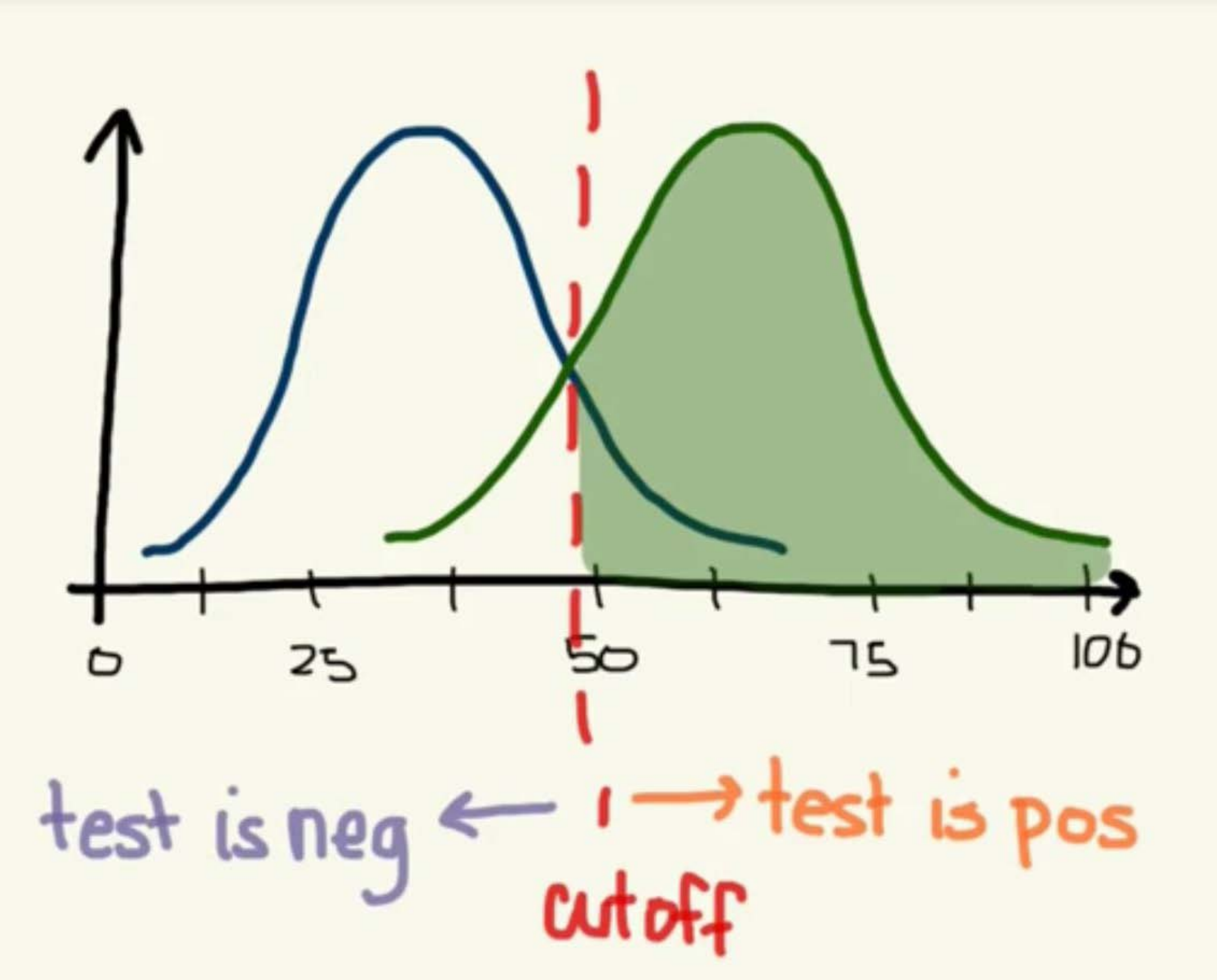
ML Evaluation

Area where the test is positive



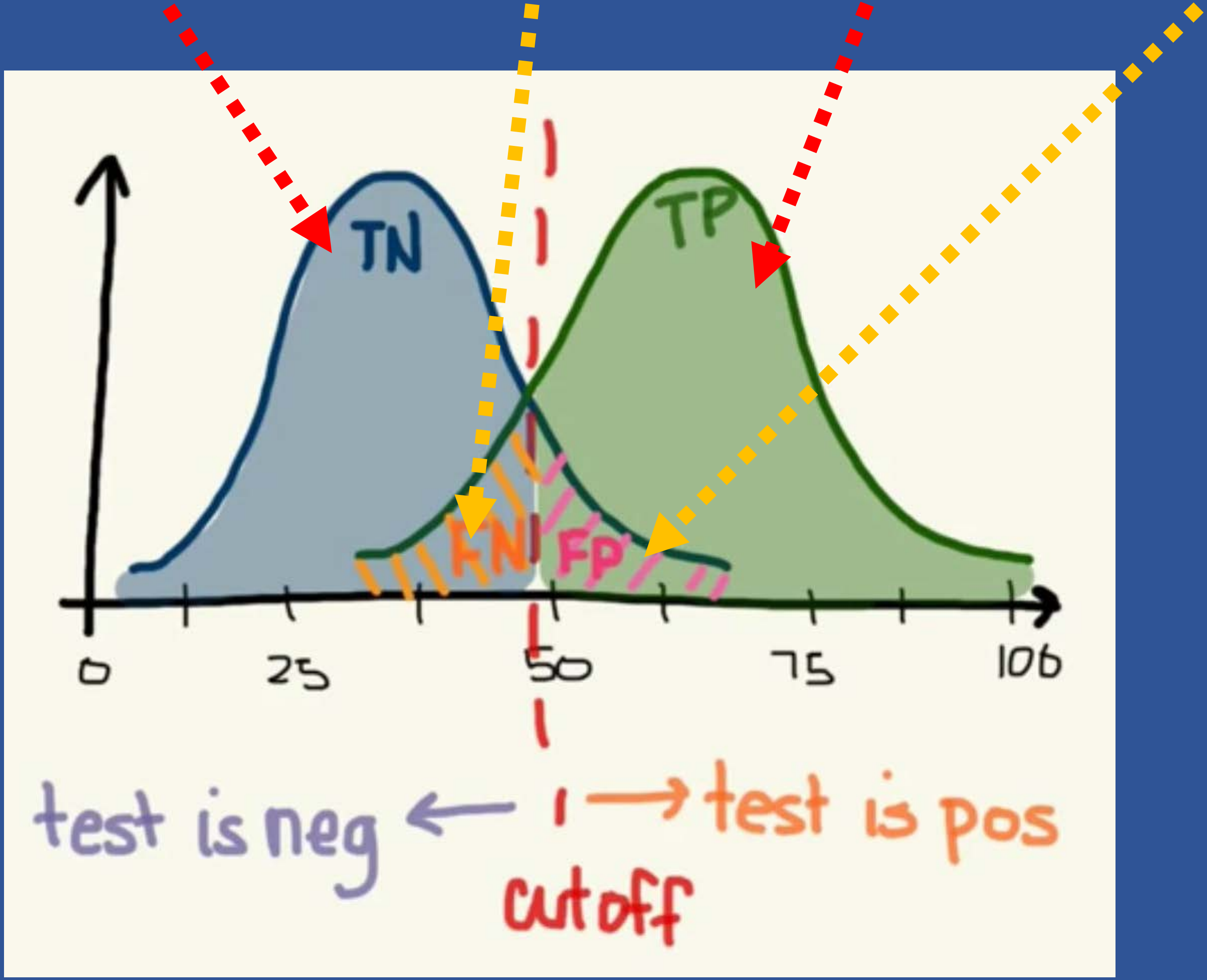
ML Evaluation

Area where the test is negative



ML Evaluation

True Negative (TN), False Negative (FN) / True Positive (TP), False Positive (FP)

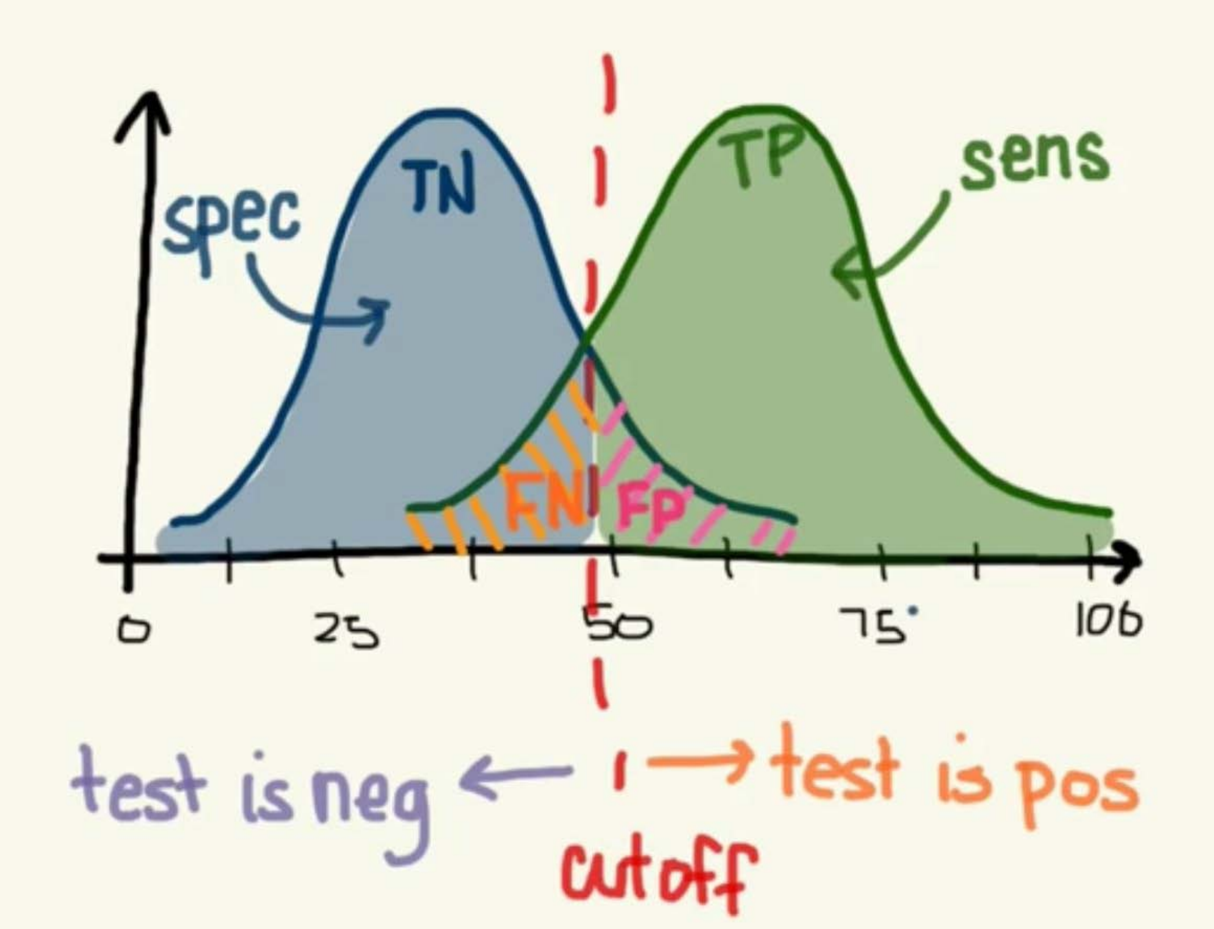


ML Evaluation

ROC Specificity / Sensitivity

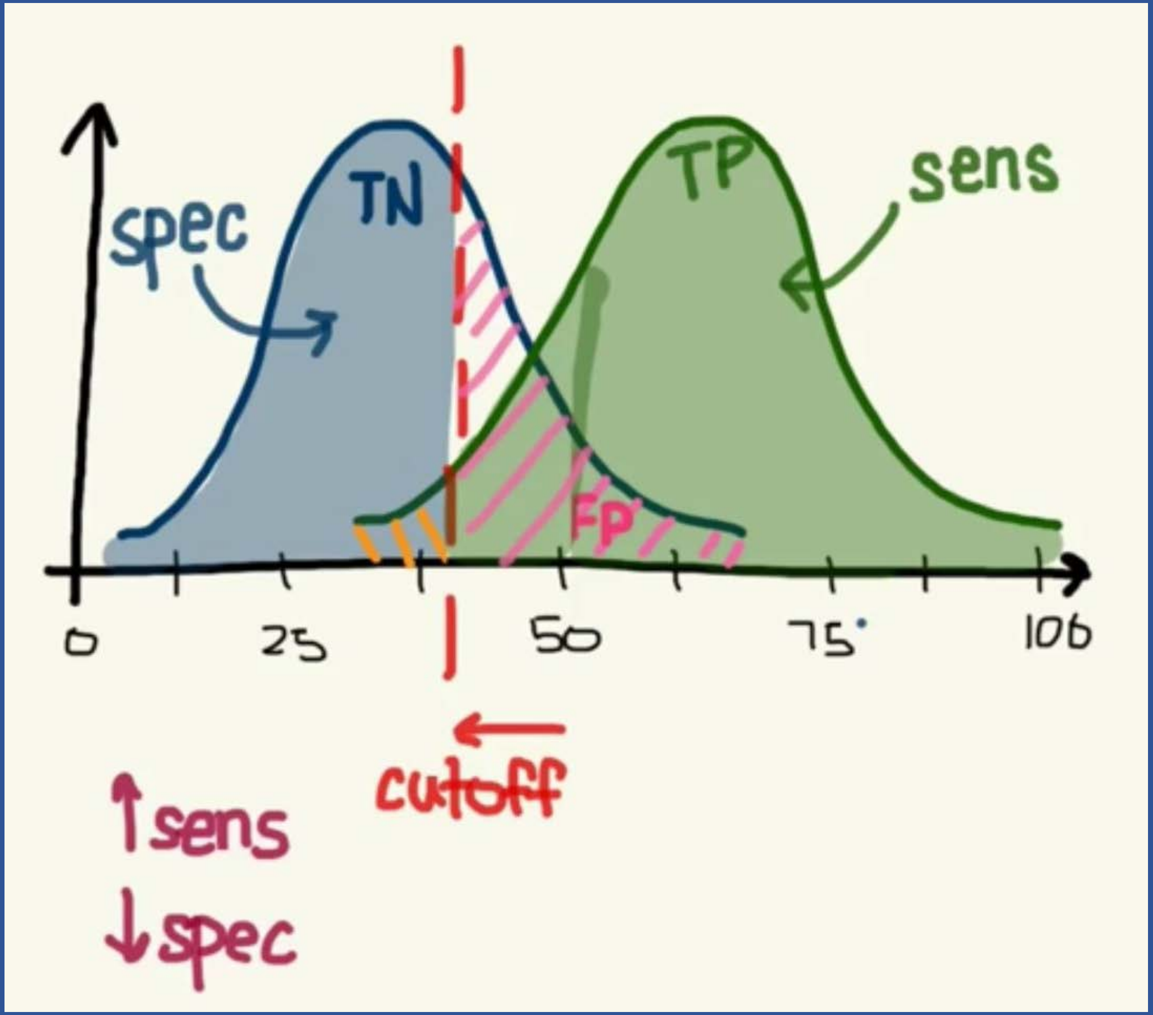
Specificity = True Negative Rate

Sensitivity (Recall) = True Positive Rate



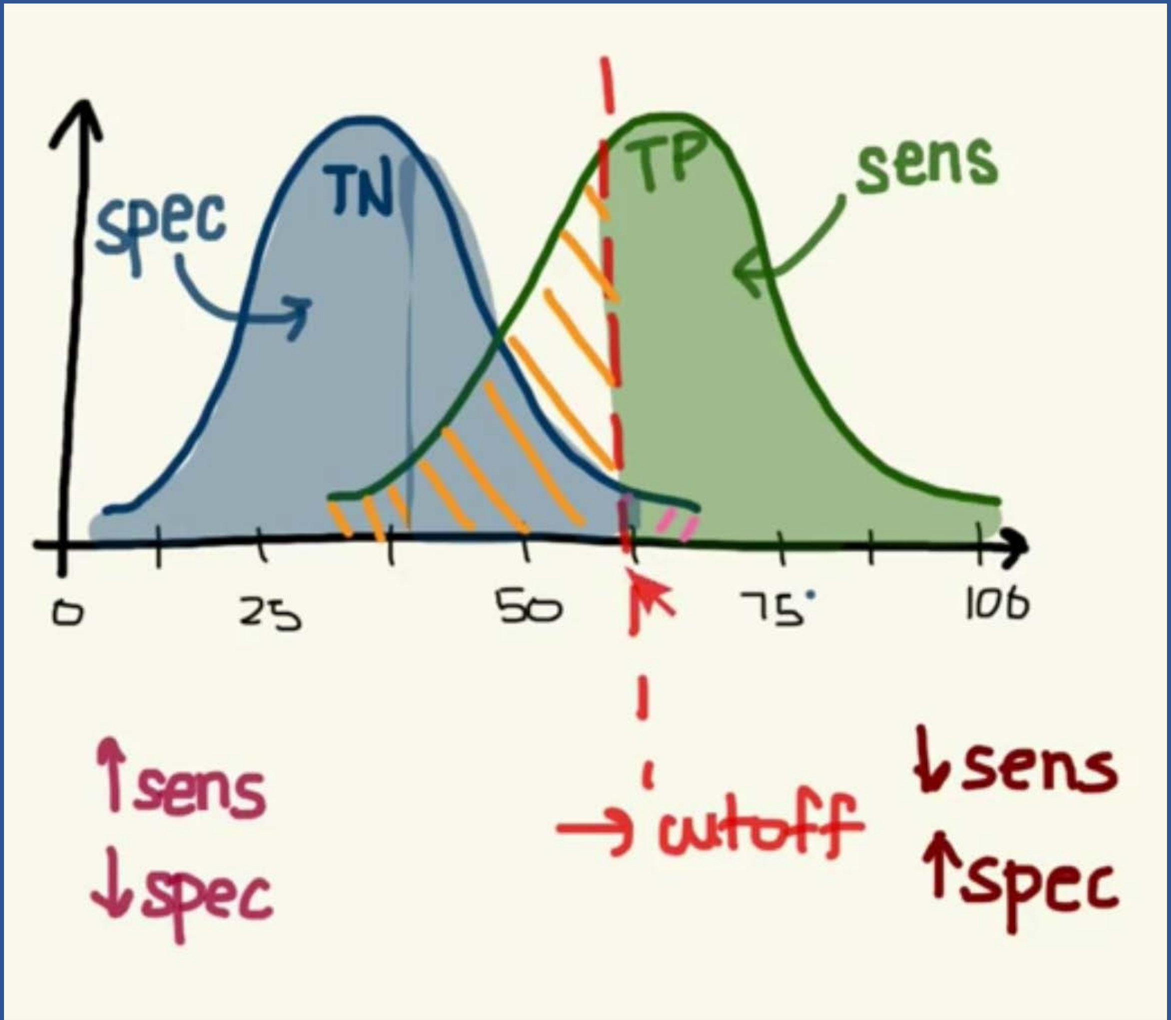
ML Evaluation

Move cutoff to the left Sens++ / Spec--



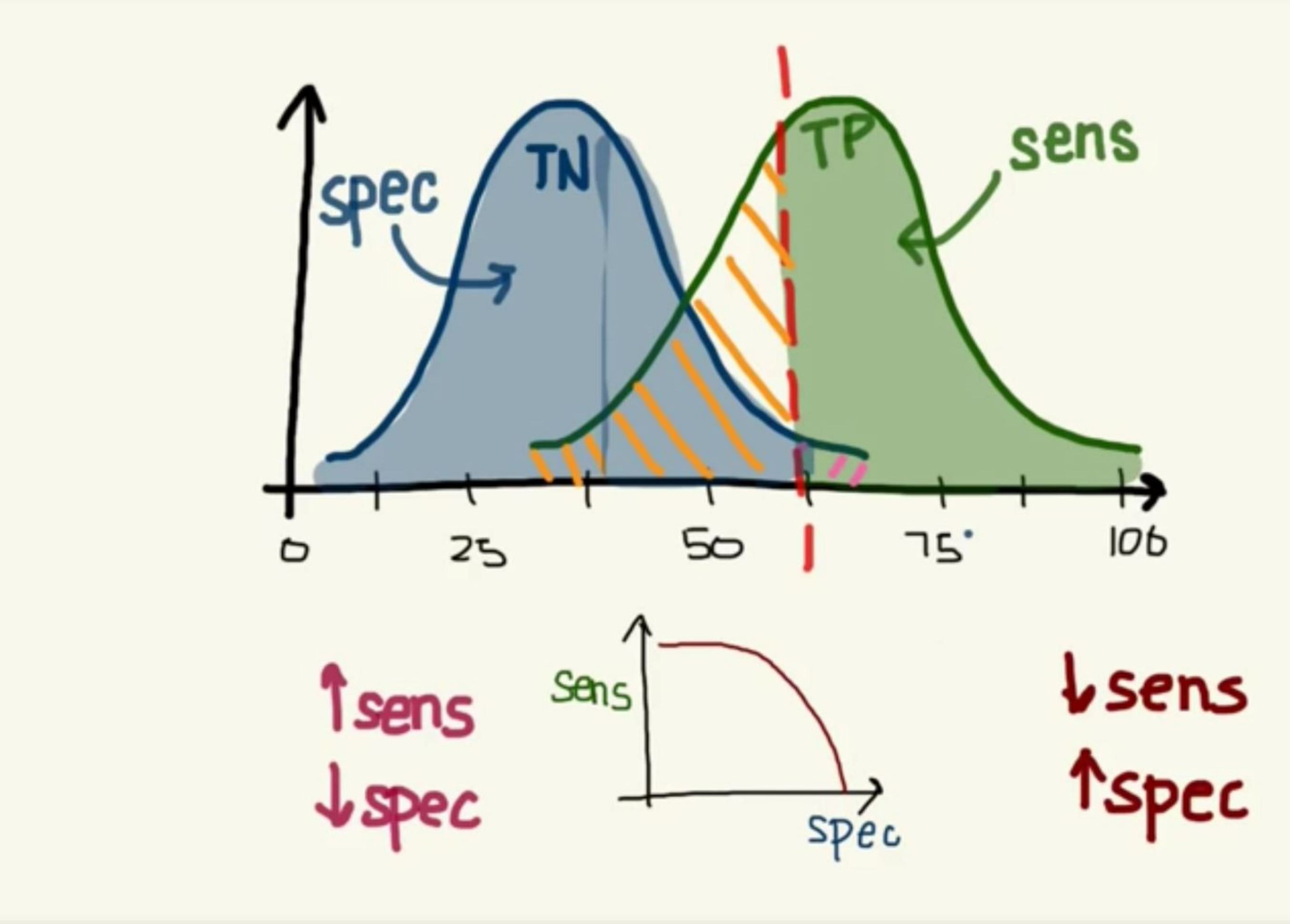
ML Evaluation

Move cutoff to right Sens-- / Spec++



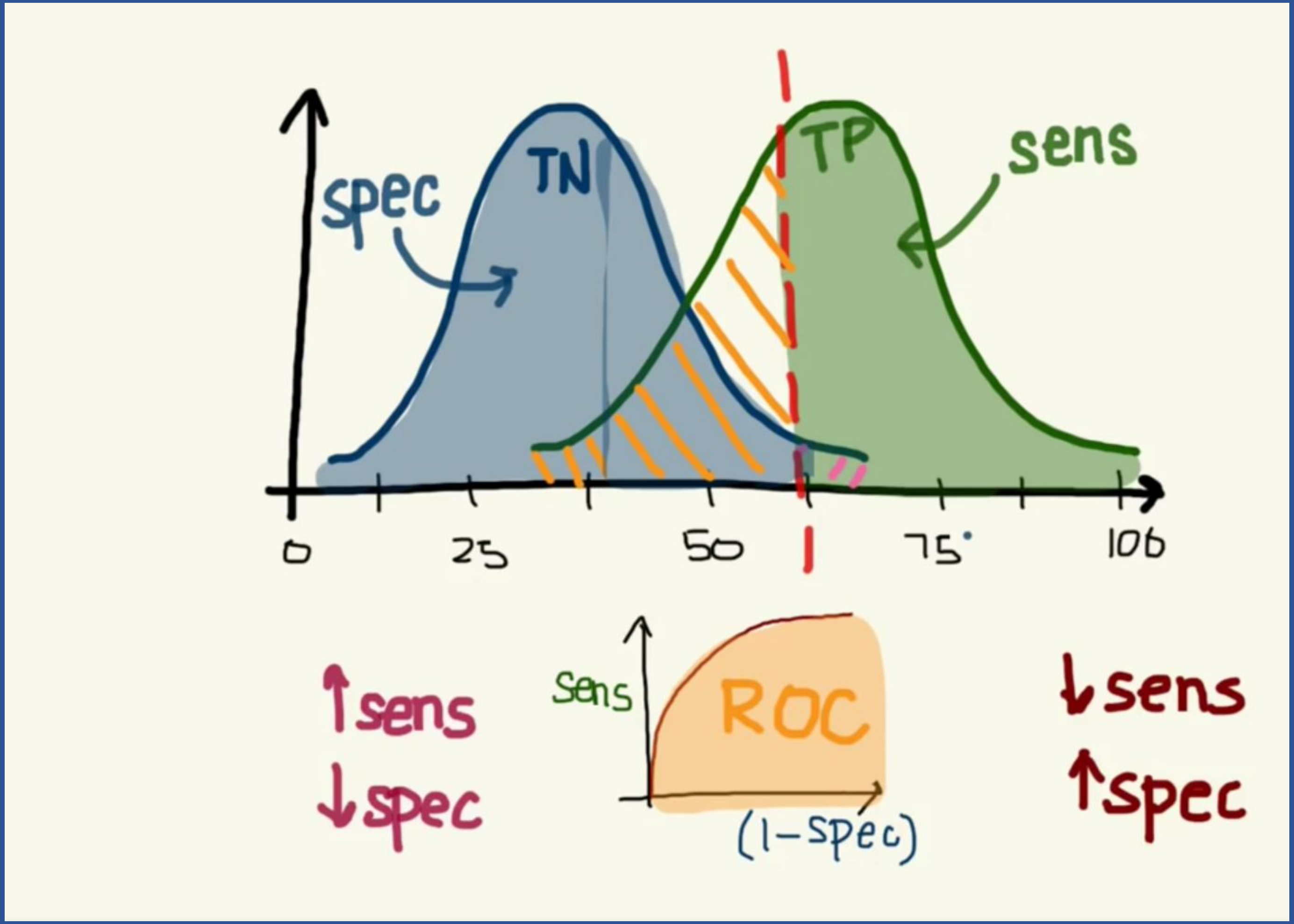
ML Evaluation

Chart proportion of Sens / Spec



ML Evaluation

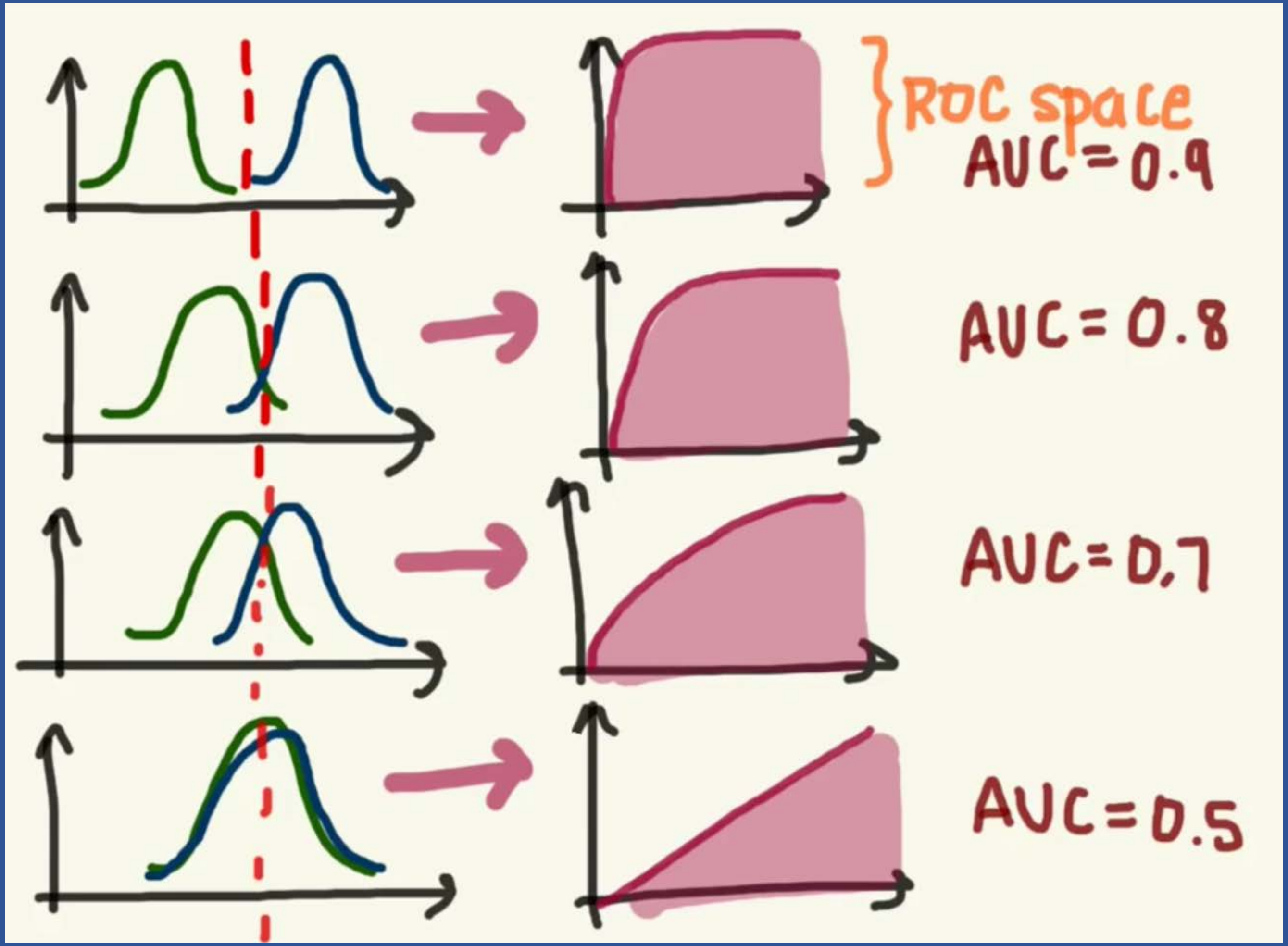
ROC curve = proportion of Sens / (1 – Spec)



ML Evaluation

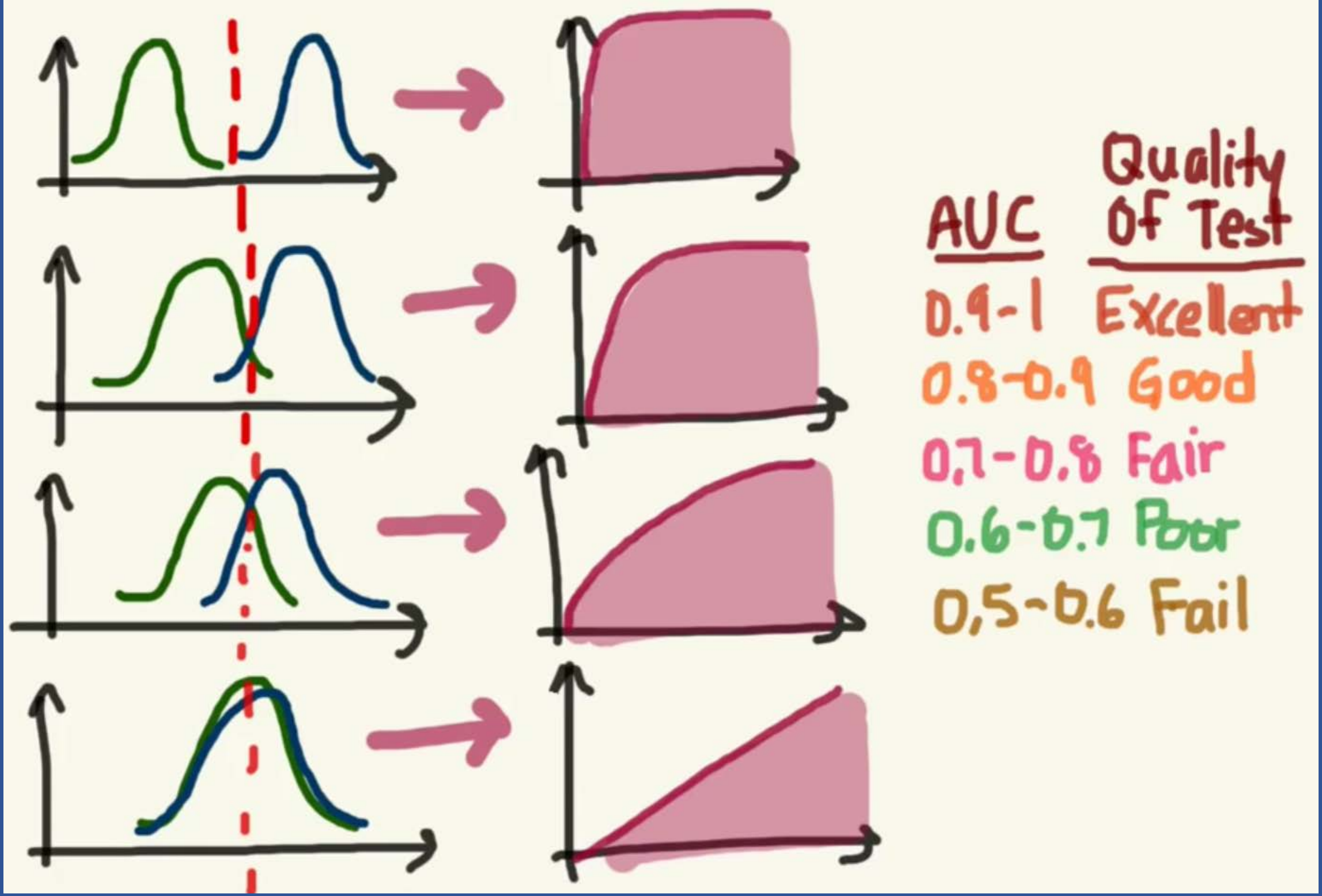
Area Under the Curve (AUC)

AUC is used to determine which of the used models predicts the classes best.



ML Evaluation

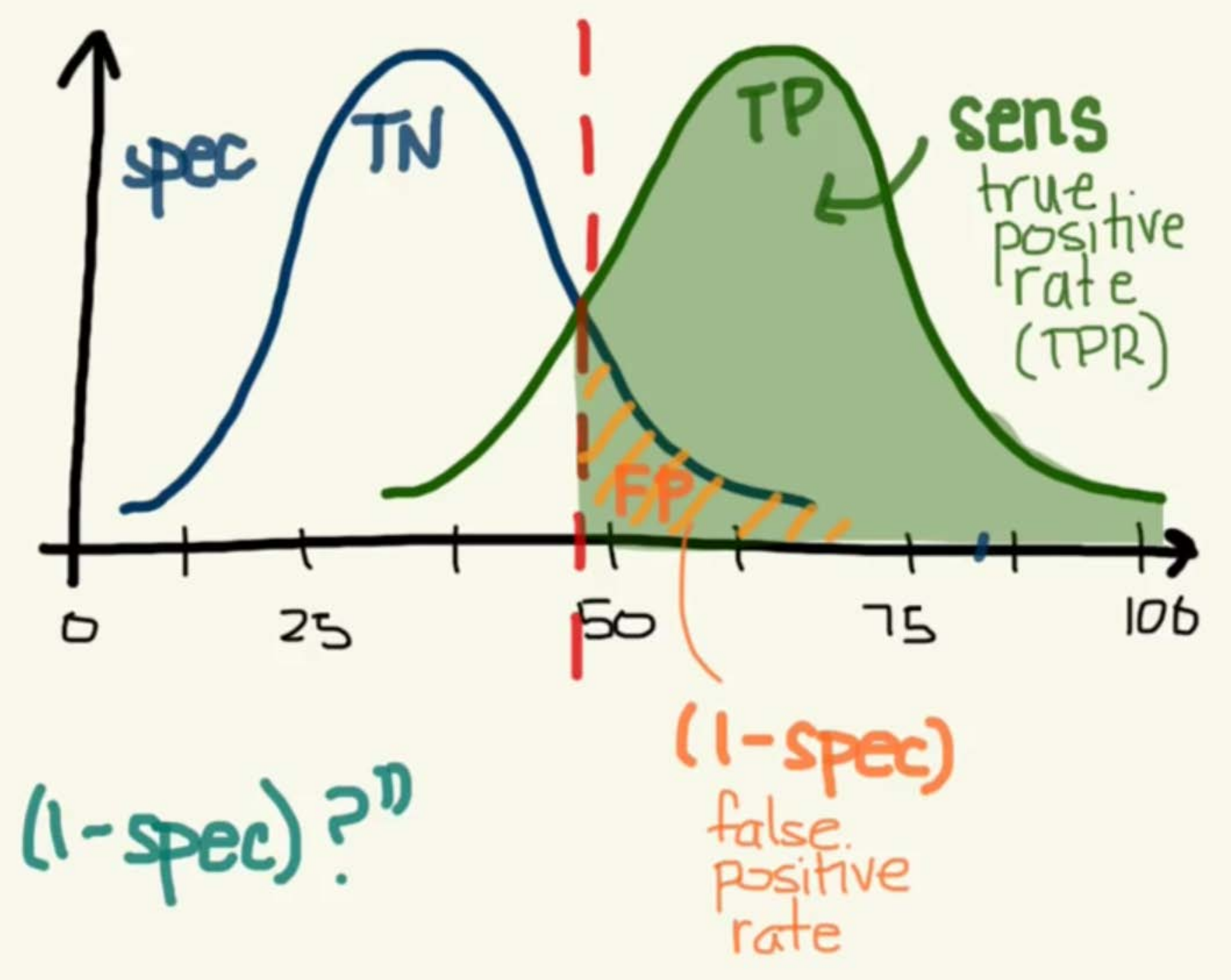
AUC score



AUC	Quality of Test
0.9-1	Excellent
0.8-0.9	Good
0,7-0.8	Fair
0.6-0.7	Poor
0,5-0.6	Fail

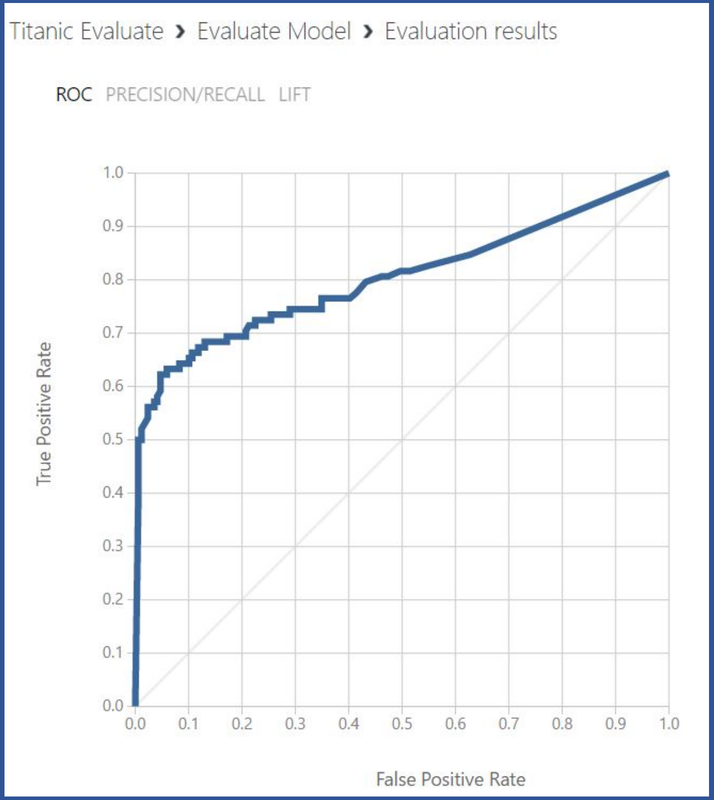
ML Evaluation

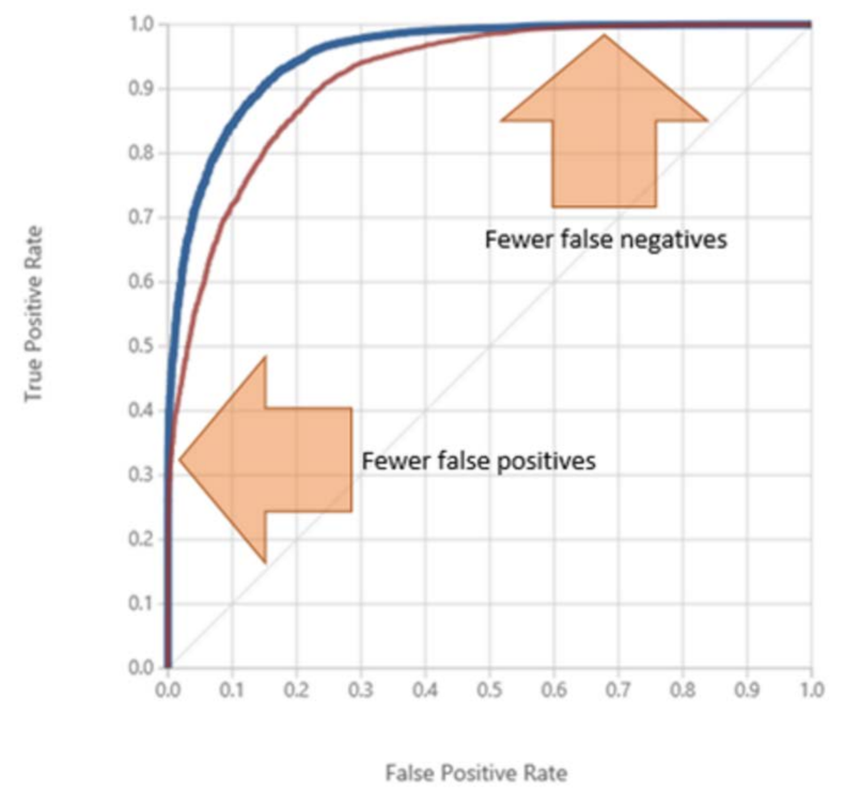
What is (1-spec)



ML Evaluation

ROC Curve

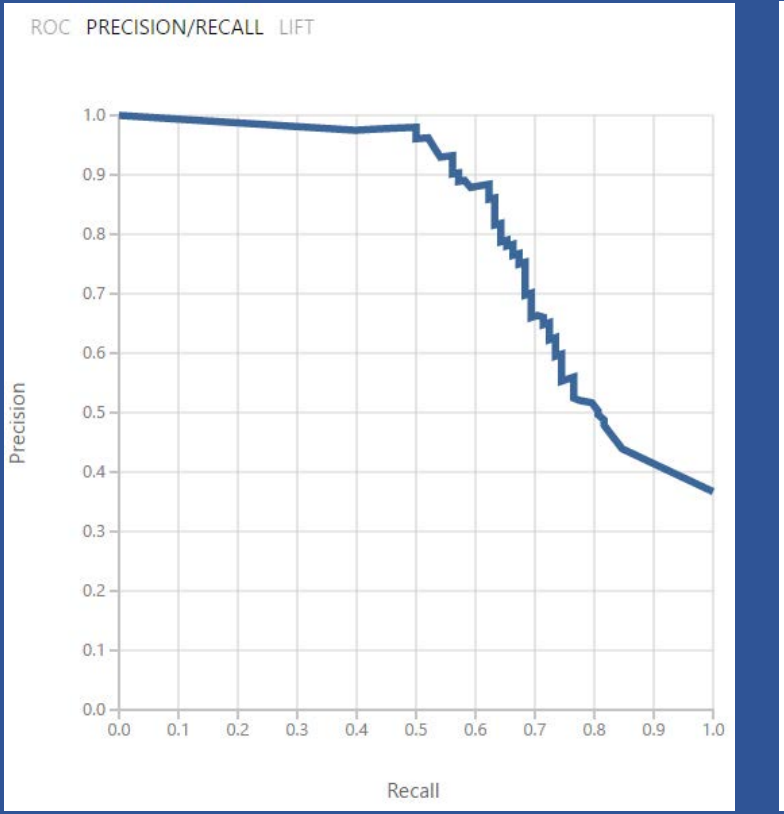

Titanic Evaluate › Evaluate Model › Evaluation results

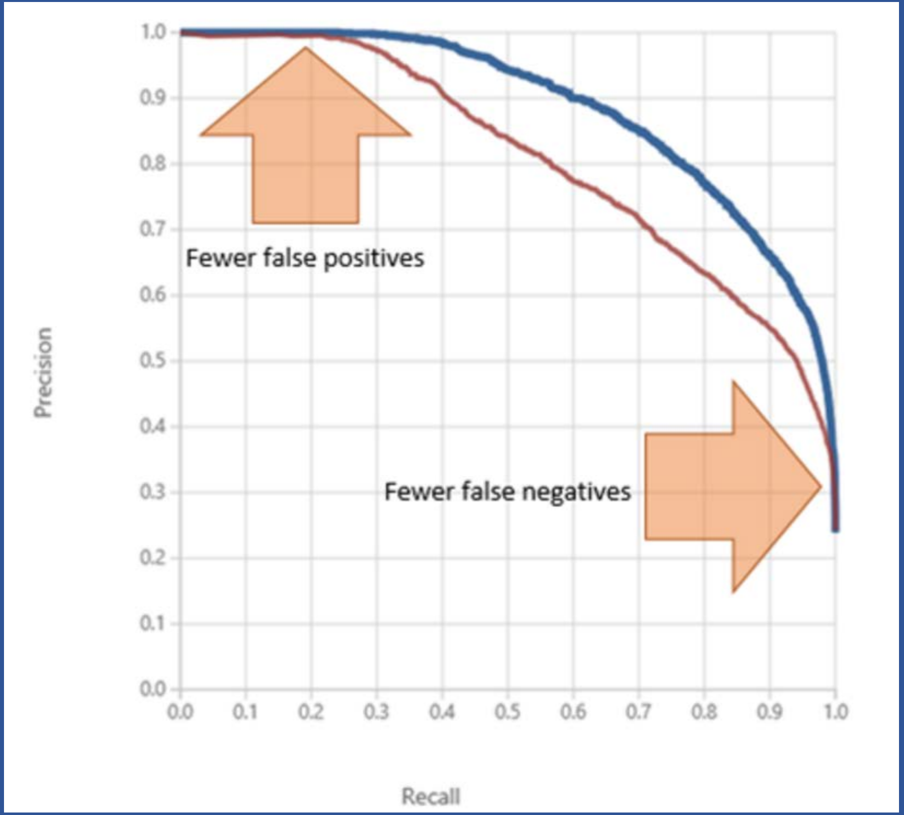


ROC curve displays the fraction of true positives out of the total actual positives. The higher and further to the left, the more accurate the model is. As you do experiments you want to see the curve move higher and to the left.

ML Evaluation

PRECISION/RECALL

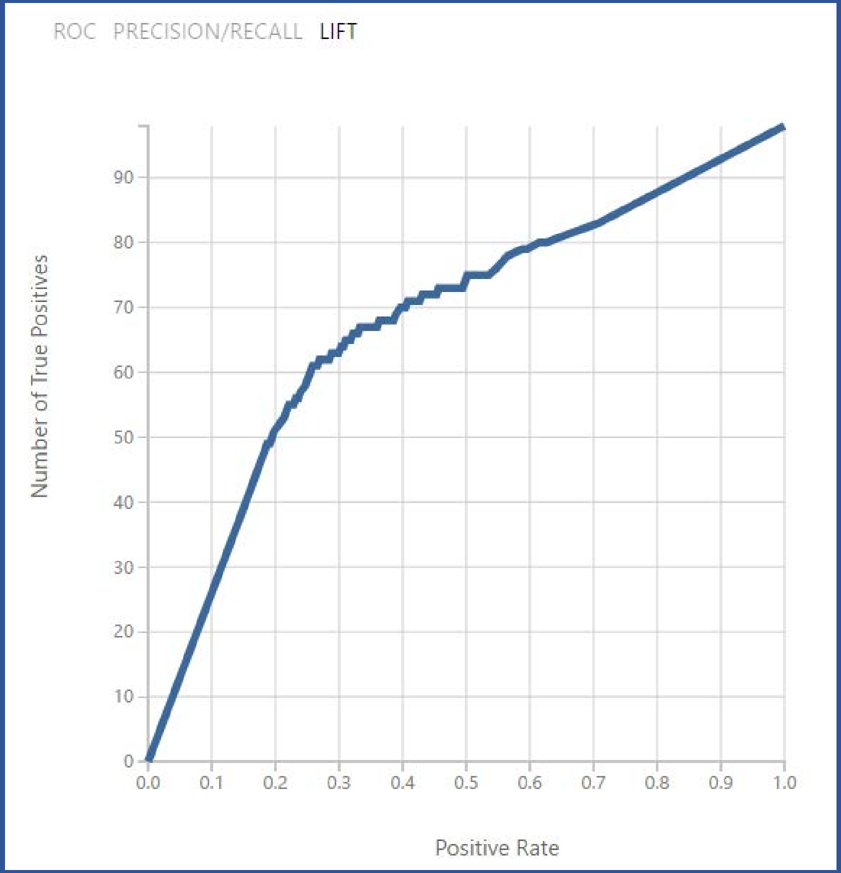




Precision represents the fraction of retrieved instances that are relevant, whereas recall represents the fraction of relevant instances that are retrieved. The “sweet spot” for the ideal model is in the upper right corner

ML Evaluation

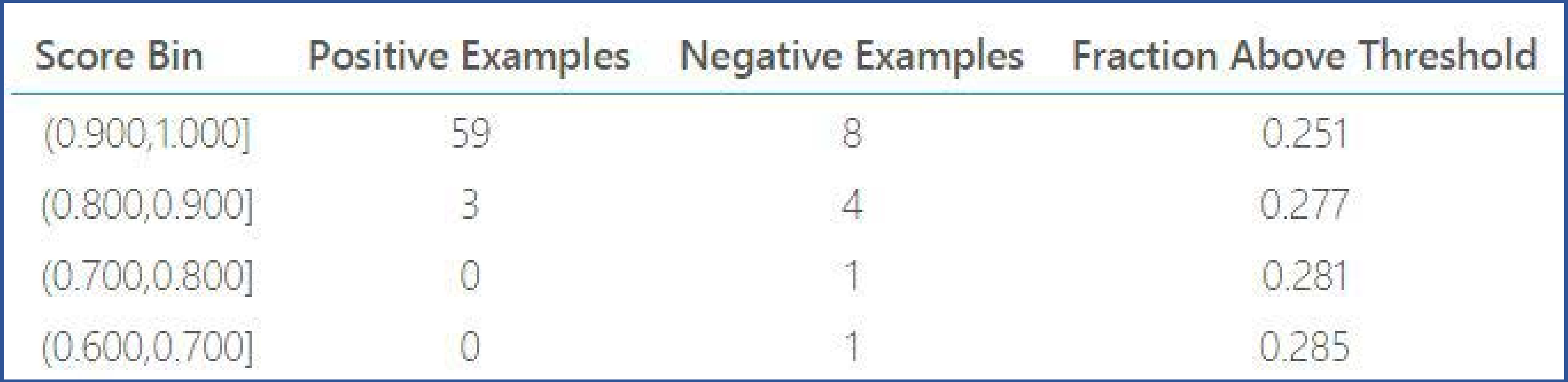
LIFT curve



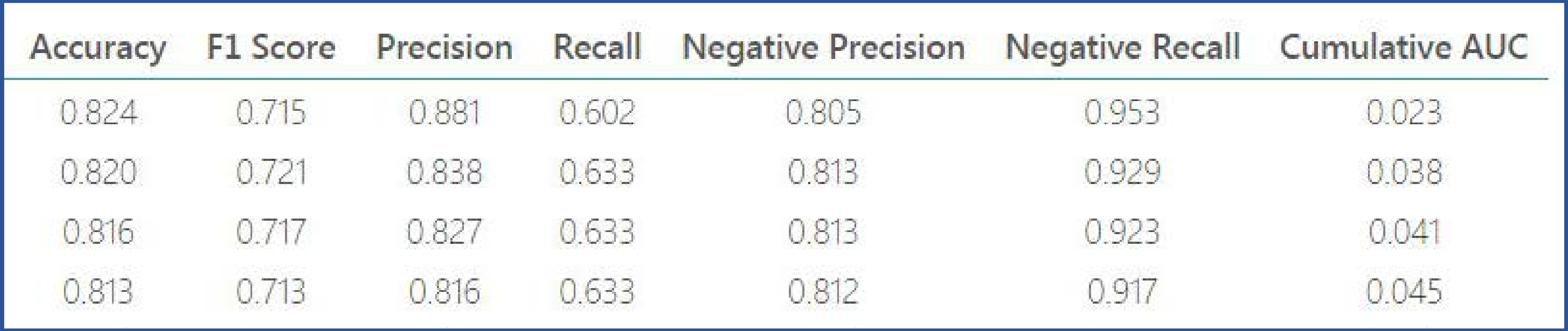
Lift curve is a variation on the ROC curve. It measures the fraction of true positives, in relation to the target response probability.

ML Evaluation

Reading Evaluation metrics



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold
(0.900,1.000]	59	8	0.251
(0.800,0.900]	3	4	0.277
(0.700,0.800]	0	1	0.281
(0.600,0.700]	0	1	0.285



Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
0.824	0.715	0.881	0.602	0.805	0.953	0.023
0.820	0.721	0.838	0.633	0.813	0.929	0.038
0.816	0.717	0.827	0.633	0.813	0.923	0.041
0.813	0.713	0.816	0.633	0.812	0.917	0.045

ML Evaluation

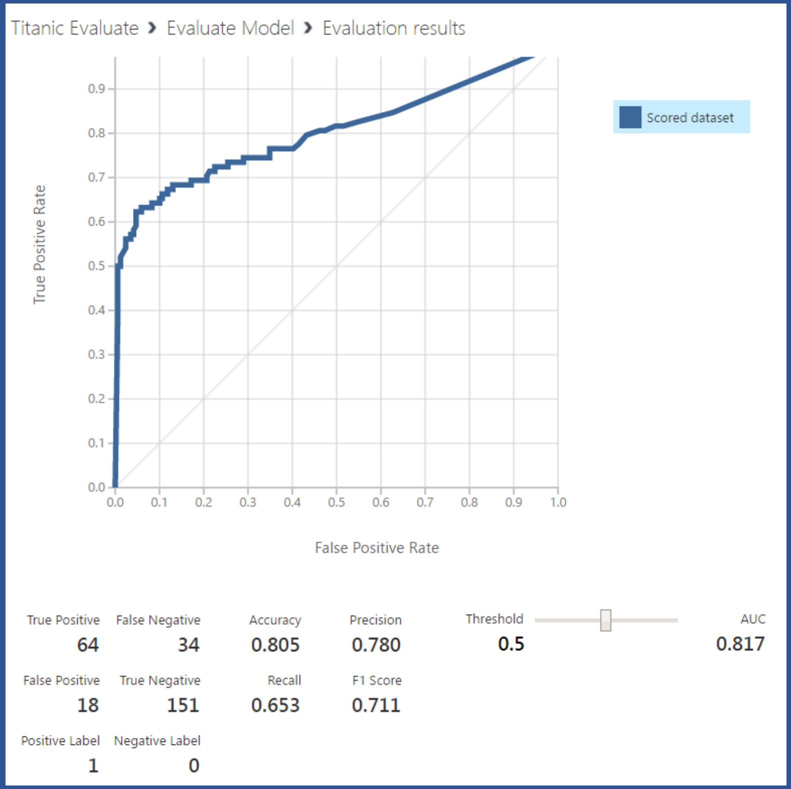
Evaluation metrics variable

- True Positive (TP): Correctly identified e.g. Sick people correctly diagnosed as sick
- False Positive (FP): Incorrectly identified e.g. healthy people incorrectly identified as sick
- True Negative (TN): Correctly rejected e.g. healthy people correctly identified as healthy
- False Negative (FN): Incorrectly rejected e.g. Sick people incorrectly identified as healthy
- Accuracy : The proportion of the total number of predictions that is correct. (TP + TN) / (TP + TN + FP + FN)
- Precision: is the proportion of positive cases that were correctly identified. TP / (TP + FP)
- Recall: Sensitivity or Recall is the proportion of actual positive cases which are correctly identified. TP / (TP + FN)
- F1 Score: is the harmonic mean of precision and Recall. 2TP / (2TP + FP + FN)
- Threshold: Threshold is the value above which it belongs to first class and all other values to the second class. E.g. if the threshold is 0.5 then any patient scored more than or equal to 0.5 is identified as sick else healthy.

ML Evaluation

Titanic evaluation results

- Positive Label: 1 = survived
- Negative Label: 0 = dead
- True Positive: correctly predict survived
- True Negative: correctly predict dead
- False Positive: incorrectly predict survived
- False Negative: incorrectly predict dead



Titanic Evaluate › Evaluate Model › Evaluation results

True Positive 64	False Negative 34	Accuracy 0.805	Precision 0.780	Threshold 0.5	AUC 0.817
False Positive 18	True Negative 151	Recall 0.653	F1 Score 0.711		
Positive Label 1	Negative Label 0				

ML Evaluation

More information

How to evaluate model performance in Azure Machine Learning
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-evaluate-model-performance

This experiment ML model

Adding Evaluation model
https://gallery.cortanaintelligence.com/Experiment/Titanic-1-2

Adding "Two-Class Decision Forest"
https://gallery.cortanaintelligence.com/Experiment/Titanic-compare-two-alogrithm