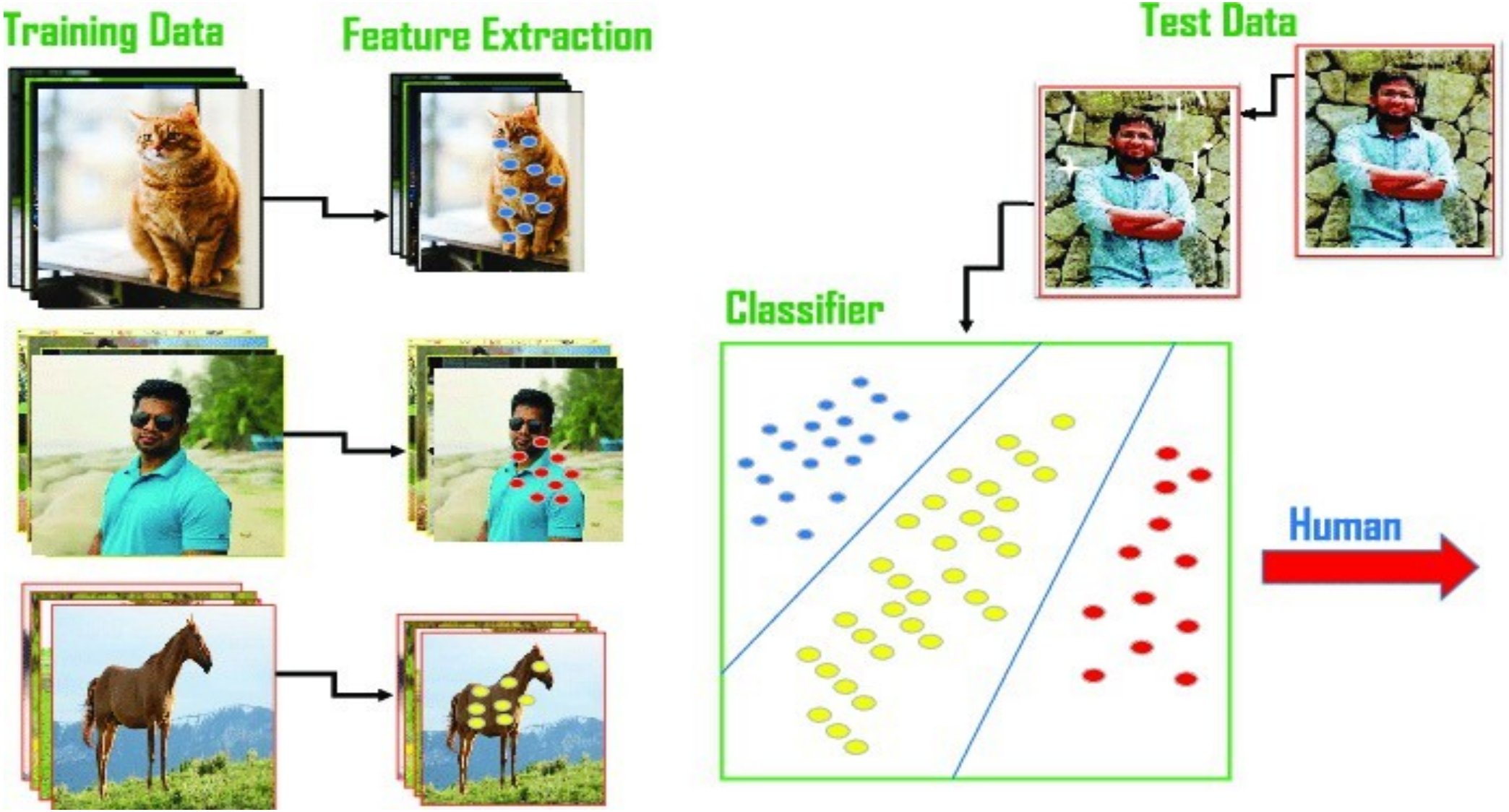
Evaluation Metrics for Classification

Classification Evaluation metrics in Machine Learning-Accuracy, Confusion Matrix-related metrics and ROC curve

Dr. Manju Venugopalan
Asst professor (Sr. Gr)
Dept of CSE
Amrita School of Computing, Bengaluru

Classification Problem



No Free Lunch Theorem in ML

- The "No Free Lunch" theorem in machine learning is named as such because it conveys the idea that there is no universally superior algorithm that performs better than all others across all possible problem domains or datasets . . In other words, there is "no free lunch" when it comes to machine learning algorithms – no one-size-fits-all solution.

  - This concept was first formalized by David Wolpert in the late 1990s. The theorem essentially states that the performance of an algorithm on a particular problem or dataset depends on the characteristics of that problem or dataset, and there is no algorithm that is superior for all possible scenarios. In practical terms, it means that different machine learning algorithms may excel in different types of problems, and the choice of algorithm should be guided by the specific nature of the problem you are trying to solve.

  - The name "No Free Lunch" is a metaphorical way of emphasizing that there are no shortcuts or guarantees of superior performance without considering the specifics of the problem at hand. It's a reminder that in machine learning, you often have to make informed choices based on your understanding of the data and problem domain, and there is no universal algorithm that will always provide the best results.

Why do we Need Evaluation Metrics?

- Evaluation metrics can help you assess your model's performance, monitor your ML system in production, and control your model to fit your business needs.

- Our goal is to create and select a model which gives high accuracy on out-of-sample data(unseen data/test data set).

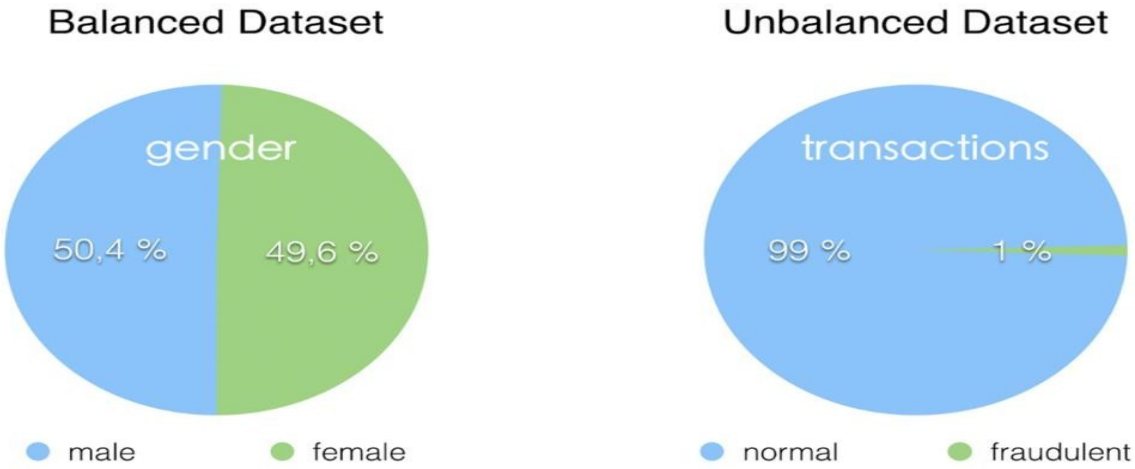
- It's very crucial to use multiple evaluation metrics to evaluate your model because a model may perform well using one measurement from one evaluation metric while may perform poorly using another measurement from another evaluation metric.

- If you choose the wrong metric to evaluate your models, you are likely to choose a poor model, or in the worst case, be misled about the expected performance of your model

Classification Accuracy

- The simplest metric for model evaluation is Accuracy. It is the ratio of the number of correct predictions to the total number of predictions made for a dataset.

Accuracy=
No of correct Predictions/ Total no of predictions

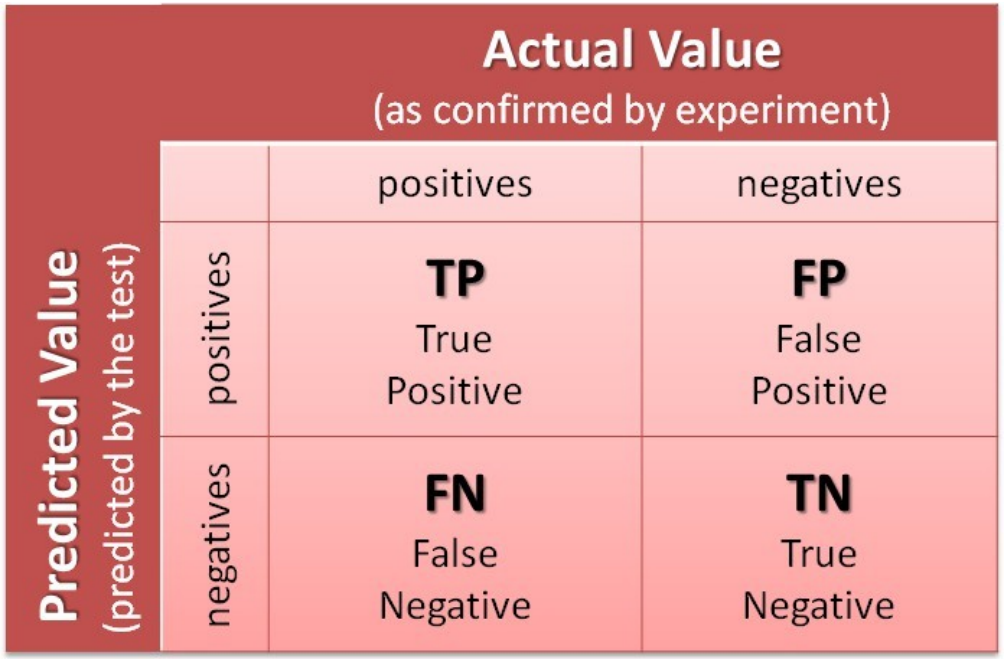


- Accuracy is useful when the target class is well balanced but is not a good choice with unbalanced classes.
- Accuracy gives us an overall picture of how much we can rely on our model's prediction.
- This metric is blind to the difference between classes and types of errors. That's why it is not good enough for imbalanced datasets.

Confusion Matrix:

- A confusion matrix or error matrix is a table that shows the number of correct and incorrect predictions made by the model compared with the actual classifications in the test set or what type of errors are being made.

- This matrix describes the performance of a classification model on test data for which true values are known. It is a n*n matrix, where n is the number of classes. This matrix can be generated after making predictions on the test data.

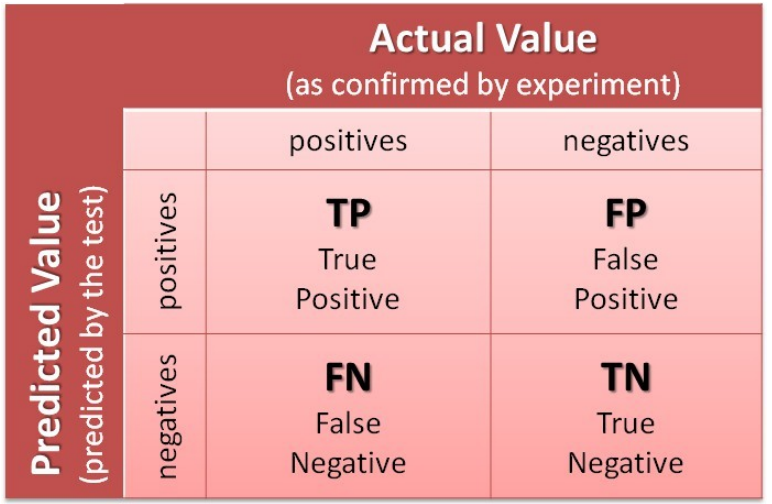


- Here, columns represent the count of actual classifications in the test data while rows represent the count of predicted classifications made by the model.

- Positive and Negatives refers to the prediction itself. True and False refers to the correctness of the prediction.

Confusion Matrix: Example

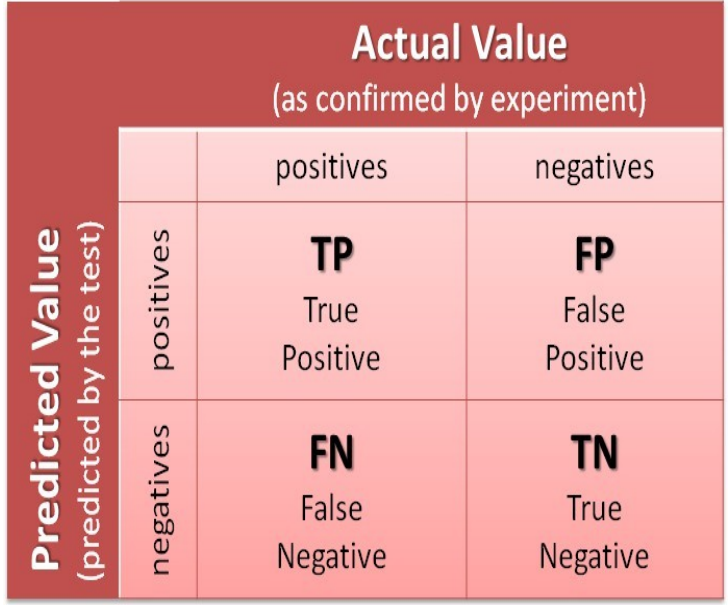
- An example of a classification problem: predicting whether a person is having diabetes or not.
- 1: A person is having diabetes
- 0: A person is not having diabetes



- Four possible outcomes could occur while performing classification predictions:
- **True Positives (TP):** Number of outcomes that are actually positive and are predicted positive.
  - **For example:** In this case, a person is actually having diabetes(1) and the model predicted that the person has diabetes(1).

- **True Negatives (TN):** Number of outcomes that are actually negative and are predicted negative.
  - **For example:** In this case, a person actually doesn't have diabetes(0) and the model predicted that the person doesn't have diabetes(0).

Confusion Matrix: Example



- **False Positives (FP):** Number of outcomes that are actually negative but predicted positive. These errors are also called Type 1 Errors. Example: The person doesn't have diabetes(0), but is mispredicted as diabetic(1)

- **False Negatives (FN):** Number of outcomes that are actually positive but predicted negative. These errors are also called Type 2 Errors.
  - **For example:** In this case, a person actually has diabetes(1) but the model predicted that the person doesn't have diabetes(0).

- The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier.
- The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

Four classification metrics from the Confusion Matrix

- It can also be calculated in terms of positives and negatives for binary classification:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment): positives	Actual Value (as confirmed by experiment): negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

- It doesn't grant us much information regarding the distribution of false positives and false negatives.
- This metric is blind to the difference between classes and types of errors. That's why it is not good enough for imbalanced datasets.

Precision

- It is the ratio of True Positives to all the positives predicted by the model.
- It is useful for the skewed and unbalanced dataset.
- The more False positives the model predicts, the lower the precision

$$Precision = \frac{TP}{TP + FP}$$

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment): positives	Actual Value (as confirmed by experiment): negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

- The precision considers when a sample is classified as Positive, but it does not care about correctly classifying all positive samples

Recall

- It is the ratio of true positives to all the positives in your dataset.
- It measures the model's ability to detect positive samples.
- The more false negatives the model predicts, the lower the recall.

$$Recall = \frac{TP}{TP + FN}$$

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment): positives	Actual Value (as confirmed by experiment): negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

- The recall cares about correctly classifying all positive samples, but it does not care if a negative sample is classified as positive.

F1-score or F-measure

- It is a single metric that combines both Precision and Recall. The higher the F1 score, the better is the performance of our model. The range for F1-score is [0,1].

- F1 score is the weighted average of precision and recall. The classifier will only get a high F-score if both precision and recall are high.

- This metric only favours classifiers that have similar precision and recall.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Specificity and Sensitivity

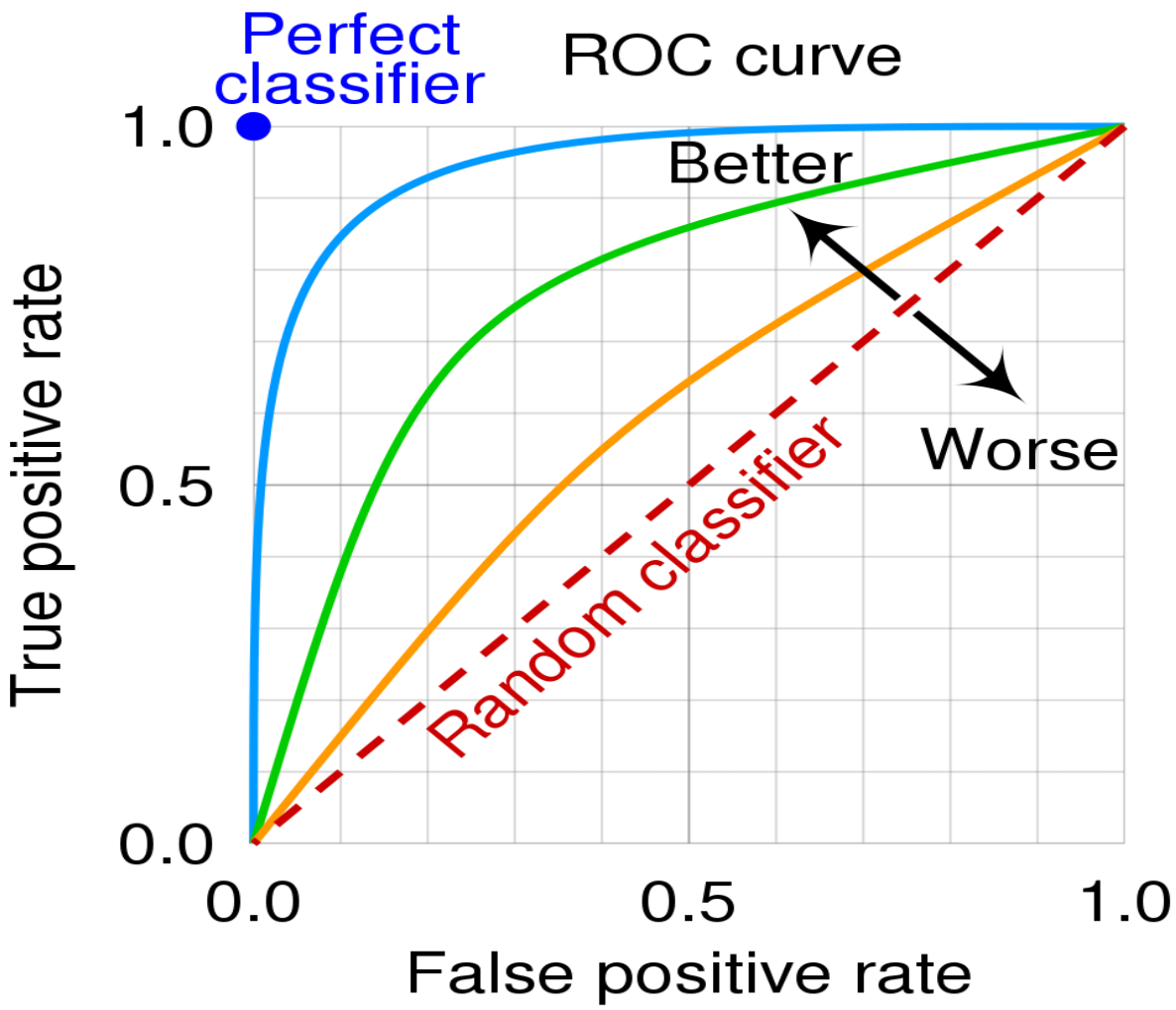
Specificity: Specificity refers to the true negative rate
Specificity=TN/(TN+FP)

Sensitivity refers to the true positive rate
Sensitivity= TP/(TP+FN)

In summary, specificity focuses on correctly identifying negatives, while sensitivity focuses on correctly identifying positives. These two measures are often inversely related; increasing sensitivity may decrease specificity, and vice versa. The choice between high sensitivity or high specificity depends on the specific goals of the test and the consequences of false positives and false negatives.

ROC Curve (Receiver Operating Characteristic Curve)

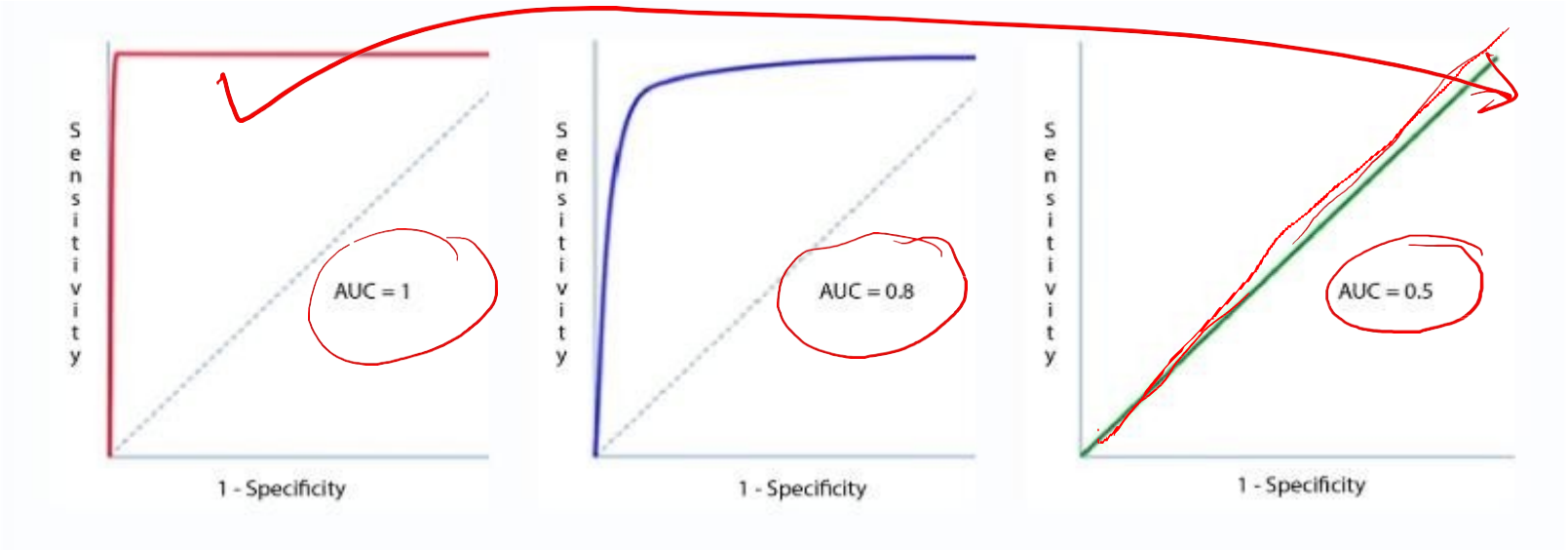
- Sensitivity(TPR) and Specificity(TNR) measures are used to plot the **ROC** curve.
- They both have values in the range of [0,1] which are computed at varying threshold values.
- **Area under the ROC curve(AUC)** is used to determine the model performance.



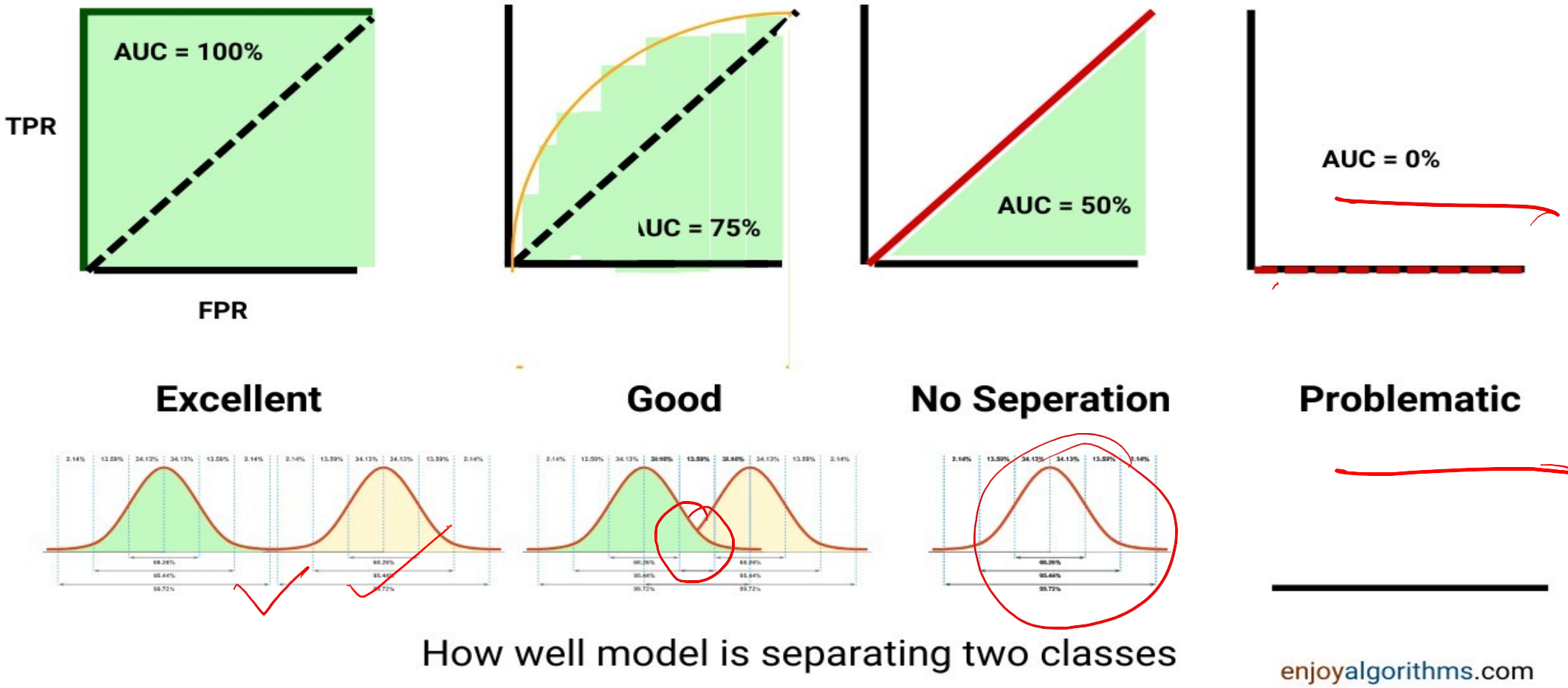
AUC (Area Under the Curve) or Area Under the ROC Curve

- AUC is a metric used to summarize a graph by using a single number. It is used for binary classification problems.

- AUC refers to the models capability to discriminate between positive and negative classes



AUC (Area Under the Curve) or Area Under the ROC Curve



Thank you!!!