Evaluation Metrics for ML Models

ML Model Evaluation

- Machine learning model needs to evaluate to measure its performance.
- How well the model generalizes on the unseen data
- By using different metrics for performance evaluation, goal is to improve the overall predictive power of our model.
- Depending only on accuracy, can lead to a problem when the respective model is deployed on unseen data and can result in poor predictions.

Model Accuracy

- There are different metrics for the tasks of classification, regression, ranking, clustering, topic modeling, etc.

- Model accuracy in terms of classification models defined as the ratio of correctly classified samples to the total number of samples:

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

Or for binary classification models, the accuracy can be defined as:

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

Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

True Positive (TP) — A true positive is an outcome where the model *correctly* predicts the positive class.

True Negative (TN)—A true negative is an outcome where the model *correctly* predicts the negative class.

False Positive (FP)—A false positive is an outcome where the model *incorrectly* predicts the positive class.

False Negative (FN)—A false negative is an outcome where the model *incorrectly* predicts the negative class.

Precision

- In a classification task, the **precision** for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

Recall

- Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

High recall means that an algorithm returned most of the relevant results

Precision and Recall

- To fully evaluate the effectiveness of a model, it's necessary to examine **both** precision and recall.
- Unfortunately, precision and recall are often in conflict. Improving precision typically reduces recall and vice versa.
- A model has a precision value of 0.5, means 50% of the time prediction correct.
- A model has a recall value of 0.11, means it correctly identifies only 11% of all samples.
- Whether high precision or high recall is preferred would depend upon the actual domain/use case.

F1 Score

- There exists another evaluation metric known as an F1 Score that helps us get the best of both metrics.
- The F1 score is the harmonic mean of the precision and recall, F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.
- ***Why harmonic mean?*** Since the harmonic mean of a list of numbers skews strongly toward the least elements of the list, it tends (compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones:

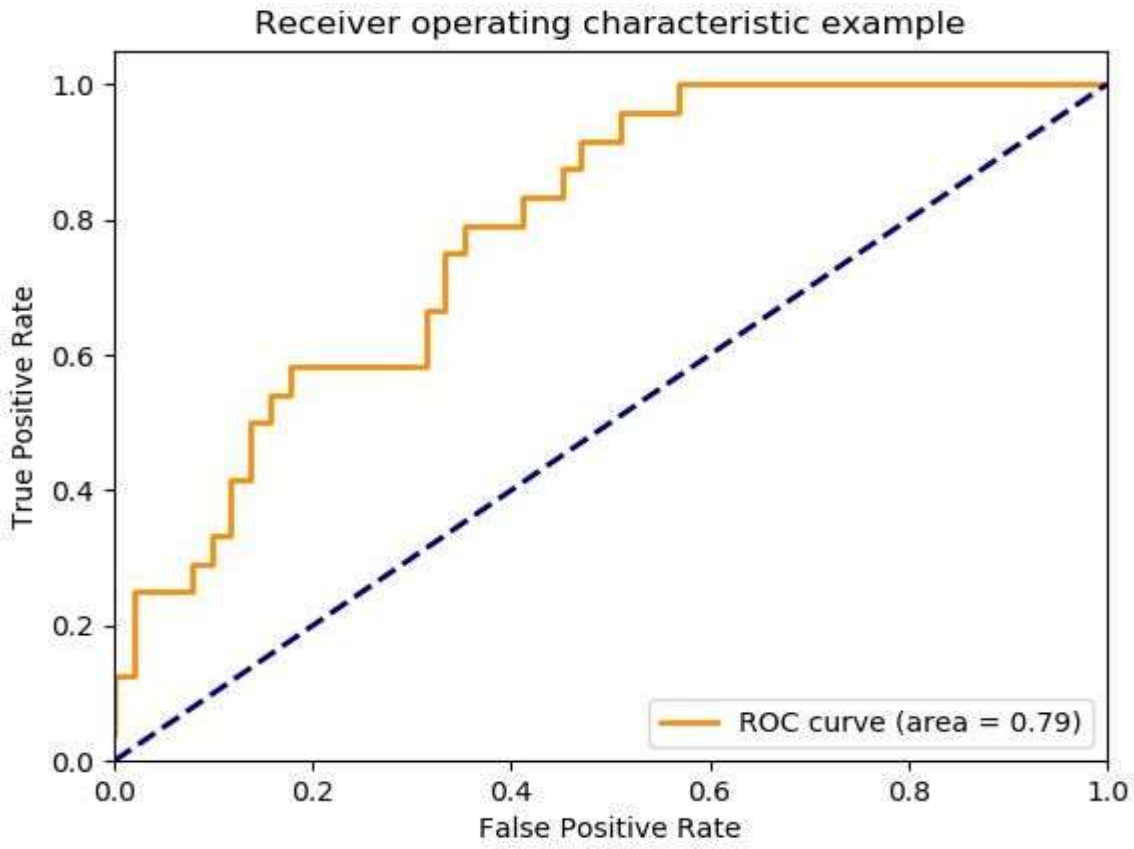
$$F_1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall}}$$

Receiver operating characteristic (ROC) Curve

- The ROC curve is created by plotting the true positive rate (TPR) also called sensitivity against the false positive rate (FPR) at various threshold settings.

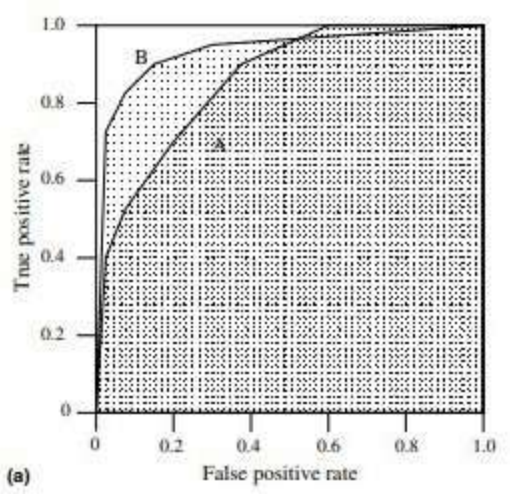
$$\text{True Positive Rate (TPR)} = \frac{TP}{(TP+FN)}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{(FP+TN)}$$



Area under the ROC Curve

- An ROC curve is a two-dimensional depiction of classifier performance.
- To compare classifiers, we calculate the area under the ROC curve, abbreviated AUC.



Classifier B has greater area and therefore better average performance

- AUC is **scale-invariant**. It measures how well predictions are ranked, rather than their absolute values.
- AUC is **a classification-threshold-invariant**. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.
- If the data is imbalanced and there's class disparity, then other methods like ROC/AUC perform better in evaluating the model performance.
- The AUC is one way to summarize the ROC curve into a single number so that it can be compared easily and automatically.