Explanation of Classification Metrics

Purpose

To help evaluate Classification Models. The metrics tell us how well our model is performing.

Example – Loan Default Prediction

- Default = Positive Class (**or also called**) Event
- Non-Default = Negative Class (**or also called**) Non-Event

I. Confusion Matrix

A table that summarizes the model's performance

	Predicted Default	Predicted Non-Default
Actual Default	True Positive (TP)	False Negative (FN)
Actual Non-Default	False Positive (FP)	True Negative (TN)

Example

- TP = 40 (correctly Predicted Default)
- FP = 10 (**in**correctly Predicted Default)
- FN = 5 (**in**correctly Predicted Non-default aka Missed Actual Defaulters)
- TN = 45 (correctly Predicted Non-default)

II. Precision

Precision : Of all Predicted Defaults, how many were Actual Defaults?

Precision = TP / (TP + FP) e.g. 40 / (40 + 10) = 40/50 = 0.80

→ 80% of Predicted defaulters, were truly defaulters.

High Precision == Few false positives (Fewer false alarms)

III. Recall (Sensitivity / True Positive Rate)

Recall : Of all actual defaults, how many did the model detect?

Recall = TP / (TP + FN) e.g. 40 / (40 + 5) = 40/45 = 0.89

→ 89% of Actual defaulters were caught by the model.

High Recall = Few false negatives (Fewer missed defaulters)

IV. F1-Score

F1-score = balances precision and recall

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

e.g. 2*(0.80 * 0.89) / (0.80 + 0.89) = 0.8426

→ When FP and FN are both important, the model scores 84%

V. Accuracy

Accuracy : Proportion of correct predictions versus all predictions

Accuracy = (TP + TN) / (TP + TN + FP + FN)

e.g. (40 + 45) / (40 + 45 + 10 + 5) = 85/100 = 0.85

→ 85% of the model predictions were correct.

This can be misleading for imbalanced datasets.

VI. AUC-ROC (Area Under the Curve of the Receiver-Operating Characteristic)

ROC : shows the models ability to separate classes, for all classification thresholds.
The area under the curve (AUC) of the ROC, is the performance metric.
In he graph the **FP** rate is along the ***x***-axis and the **TP** rate is along the ***y***-axis.

AUC = 1.0 » Best possible performance. Perfect.
AUC = 0.5 » Random performance. Average.
AUC = 0.0 » Worst possible performance. Useless.

→ AUC = 0.92 means a 92% chance of scoring a defaulter higher than a non-defaulter.

Summary of Metrics

Name	Measurement	Interpretation
i. Confusion Matrix	Counts of TP, FP, TN, FN	Raw outcome counts
ii. Precision	TP / (TP + FP)	Correct % of Positive predictions
iii. Recall	TP / (TP + FN)	Coverage of Actual Positives
iv. F1-Score	2*(ii * iii) / (ii **+** iii)	Balance of Precision and Recall
v. Accuracy	(TP+TN) / (TP + FP + TN + FN)	Overall correctness
vi. AUC-ROC	Area under the Curve of the ROC	Class separation ability

When to use which Metric

Scenario		Utilize
Avoid False Alarms	→	Precision
Catch all Actual Defaulters	→	Recall
Balanced Mistakes	→	F1-Score
Class Imbalance exists	→	AUC-ROC