Tech Bytes

ML Refresher - Classification Evaluation Metrics

by Pranali Bose

Confusion Matrix

What it is? A table used to describe the performance of a classification model.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

- **TP**: Correctly predicted positive class.
- **TN**: Correctly predicted negative class.
- **FP**: Incorrectly predicted positive class (Type I error).
- **FN**: Incorrectly predicted negative class (Type II error).

Precision

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

- **What is it?** The proportion of true positive predictions out of all positive predictions made
- **Use case:** Best for scenarios where the cost of false positives is high (e.g., spam detection).
- **Limitation:** Does not take into account false negatives, which can be significant in some contexts.

Recall

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

- **What is it?** The proportion of true positive predictions out of all actual positive instances
- **Use case:** Important in situations where missing a positive instance is costly (e.g., disease screening).
- **Limitation:** High recall can lead to low precision if many false positives are present.

Accuracy

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

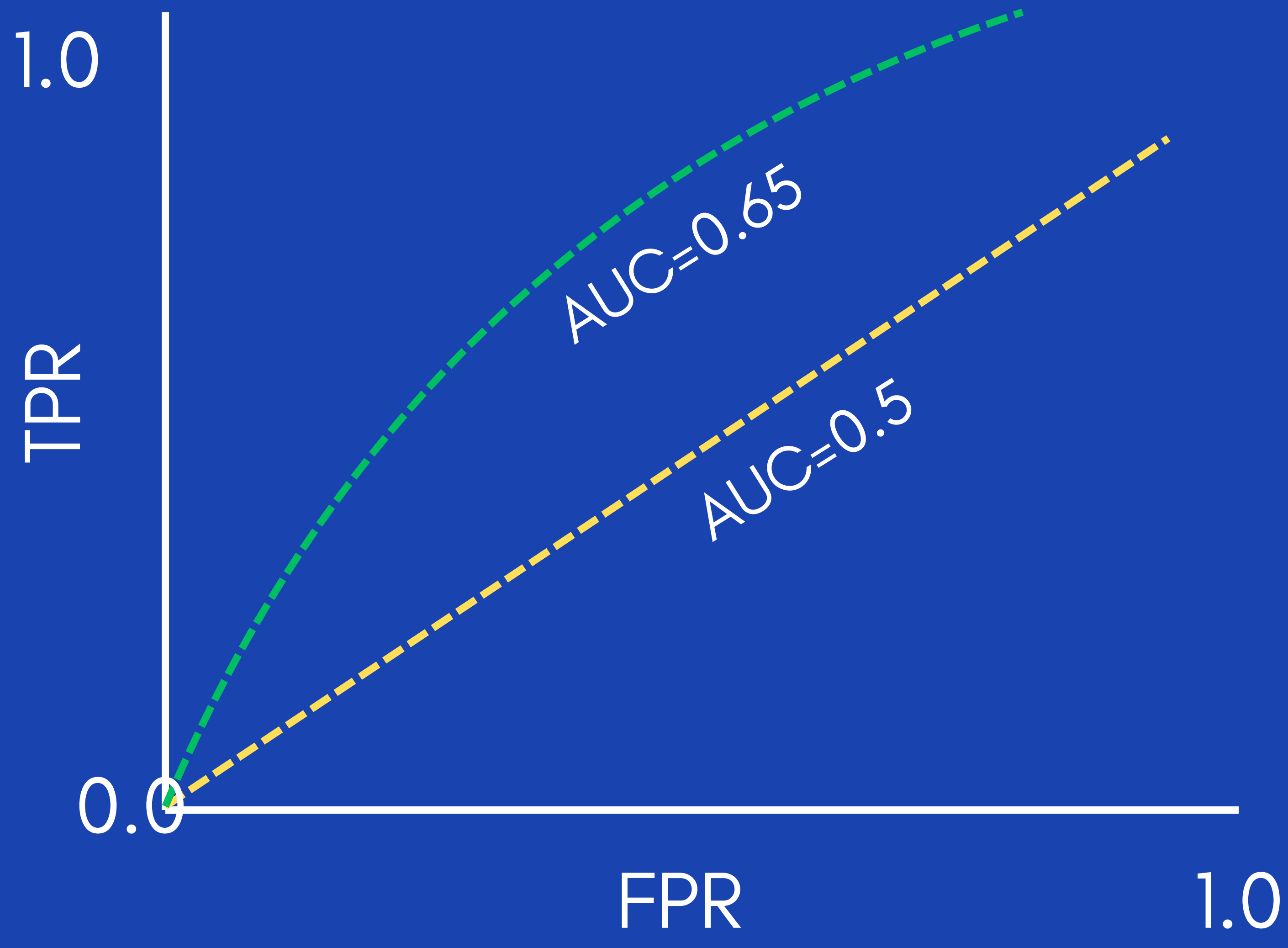
- **What is it?** The proportion of correct predictions out of all predictions.
- **Use case:** Best for balanced datasets where classes are equally distributed.
- **Limitation:** Misleading for imbalanced datasets (e.g. 95% accuracy if 95% of data belongs to one class).

F1-score

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

- **What is it?** The harmonic mean of precision and recall, providing a balance between the two
- **Use case:** Useful in cases where you need a balance between precision and recall, especially in imbalanced datasets.
- **Limitation:** Can be misleading if the underlying distribution of classes is not considered, as it averages the performance.

AUC-ROC Curve



- **What is it?** The area under the receiver operating characteristic curve, representing the trade-off between true positive rate (recall) and false positive rate across different thresholds.
- **Use case:** Good for evaluating classifiers on imbalanced datasets, as it considers all classification thresholds.
- **Limitation:** AUC-ROC can be overly optimistic for highly imbalanced datasets and may not reflect practical performance.

Ponder Upon...

- What does an AUC-ROC score of 0.5 indicate about a model's performance, and how would you interpret an AUC of 0.8 versus 0.9?
- How do you interpret the shape of an ROC curve, and what does it indicate about a model's performance at various thresholds?
- In what scenarios might a model exhibit a high F1 score while having a low overall accuracy? Provide an example.
- How would you evaluate a classification model differently in a medical diagnosis context compared to a marketing context?
- How do you interpret a precision-recall curve, and how is it different from an ROC curve in terms of evaluating model performance?
- How does the distribution of classes in your dataset influence the evaluation metrics, and how can you mitigate the effects of an imbalanced dataset?
- Why is F1-score a harmonic mean of Precision and Recall and not AM or GM?

Tech Bytes

Find this useful?

Let me know in the comments which topic would you like to see next

Follow for more...