Performance Metrics

1 Reasoning

Choosing a suitable performance metric for binary classification depends on the nature of the problem, the distribution of the classes, and the real-world implications of the predictions.

For our bank churning dataset, the main real-world considerations for choosing a performance metric are:

- We have got an imbalanced dataset as approximately 20 percent exited while 80 percent stayed
- Cost of False Positives vs. False Negatives
- Threshold Optimization - Some metrics, like AUC-ROC and AUC-PR, allow you to optimize the classification threshold rather than making a hard prediction. This is useful when the business decision involves setting a threshold (e.g., setting a high threshold to avoid false positives in high-risk cases).

2 Different Methods: Benefits and Limitations

Accuracy: Can be misleading for imbalanced datasets (e.g., predicting the majority class can give high accuracy even if the model ignores the minority class).

Precision: Suitable when the cost of false positives is high. Not a priority for our scenario as not too big of a cost if we count someone who stayed as exited.

Recall: Suitable when the cost of false negatives is high. Might be important for us as bigger cost when we believe that someone who has exitted has been counted as retained.

F1-Score: Suitable when there is a trade-off between precision and recall, especially in imbalanced datasets. F1-score is helpful when you care equally about false positives and false negatives.

ROC Curve: Plots the True Positive Rate (Recall) against the False Positive Rate (FPR = FP / (FP + TN)) for different classification thresholds. AUC-ROC: Measures the area under the ROC curve. Can be overly optimistic for highly imbalanced datasets where one class dominates.

AUC-PR: Measures the area under the precision-recall curve. Use case: More informative than AUC-ROC for imbalanced datasets, where the focus is on the minority class. The AUC-PR curve emphasizes the model's performance with respect to the positive (minority) class. Limitations: Does not capture performance on negative (majority) class.

Logarithmic Loss: Limitations: It requires probabilistic outputs and is more complex to interpret than other metrics.

Matthews Correlation Coefficient (MCC) Use case: MCC gives a balanced measure even when the classes are imbalanced. It takes into account all four confusion matrix elements (TP, TN, FP, FN). Limitations: More difficult to interpret compared to precision, recall, or accuracy.

Balanced Matthews Correlation Coefficient

Balanced Accuracy: Useful when the dataset is imbalanced, as it gives equal weight to the performance on both the positive and negative classes.