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
  False Positive: Predicting that a customer will churn when they actually will not.
  False Negative: Predicting that a customer will not churn when they actually do.
- 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: MCC is known to to mitigate the imbalance of a test set, however it still remains dependent on the prevalence of the predicted categories, this can lead to MCC being underestimated at extremely high or low positive prevalence. Balanced MCC[1] is an extension of balanced accuracy which is a performance metric that is calibrated to a test set with a positive prevalence of 50%. This means we can have an MCC metric independent of prevalence. This can help us see if our model is performing well on predicting those that stay with the bank but also on our minority class of those who churn. We are able to compute this from the confusion matrix.

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

3 Our Performance Metrics

3.1 Recall

In the context of our problem, bank churning, we believed we needed to prioritize reducing false negatives. Missing a churner (false negative) could result in revenue loss, which might outweigh the cost of reaching out to non-churners and the use of retention methods for those that would be staying anyway (false positives). From a customer satisfaction standpoint, it would be much more preferable to offer someone incentives to stay who was staying anyway, rather than failing to offer them to those who were planning on leaving. Furthermore, acquiring a new customer would be more costly than simply offering benefits such as reward programmes and engaging with customers more.

For the reasons listed one of the performance metrics we've opted to assess our models using is Recall. Recall is calculated using the following formula,

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

Since Recall for the positive class gives a measure of the proportion of positives predicted as such, we thought it would be appropriate in the context of our selected problem. A high recall value suggests that a model is correctly predicting the customers that will churn, even though it might be picking up on some FP, but as discussed this is not as big of a concern for us as having more false negatives.

3.2 F1-Score

Although Recall is a good measure of success for our models, it doesn't tell us the whole picture. For example, if one was to make a model that simply predicted that every customer will churn, then they'd get a Recall value of one, despite the model being essentially useless. Therefore, we also want to take a look at the F1-score, which accounts for both false positives and false negatives. F1-score is calculated using the following,

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where,

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

and Recall is as stated above.

F1-score gives us a more complete view of the model we've created due to the balance it incorporates between precision, which relates to correct positive predictions, and recall, which relates to identifying positive instances as positives. This works well in tandem with our Recall metric as it negates the possibility of using a model that predicts way too many customers to churn and not enough to stay, and coming out of it believing that this is a good model with which to assess the data. In addition to this, the F1-score metric is useful for imbalanced datasets because ***. Our dataset is imbalanced in favour of those not churning, with around 80% choosing to stay with the bank so a performance metric which is useful despite this is very good for us.

3.3 Balanced MCC

acquiring new customers more spemny(coin coin coin coin coin)

References

[1] Sébastien J.J. Guesné et al. "Mind your prevalence!" In: *Journal of Cheminformatics* 16.1 (Dec. 2024), pp. 1–13. ISSN: 17582946. DOI: 10.1186/S13321-024-00837-W/FIGURES/11. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00837-w%20http://creativecommons.org/publicdomain/zero/1.0/.