Performance Metrics

1 Introduction

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:

- 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).
- The imbalanced nature of the dataset, with 20% of the customers churning and 80% staying with the bank.

2 Our Performance Metrics

We explored a variety of different performance metrics for this report and we have a document with more information on these in the performance metric folder. For our project, we decided to stick with the performance metrics covered in this section.

2.1 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.

2.2 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. 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.

2.3 Balanced MCC

The balanced MCC gives a fair and interpretable evaluation of model performance on imbalanced data. Balanced MCC is calculated using the following,

$$\text{Balanced MCC} = \frac{\text{sensitivity}+\text{specificity}-1}{\sqrt{1-(\text{sensitivity}-\text{specificity})^2}}$$

This formula is derived from getting MCC as a function of prevalence, sensitivity and specificity, and then setting the prevalence to 0.5 to represent a balanced set, this is similar to how balanced accuracy can be calculated. Balanced MCC values range from -1 to +1, where +1 indicates perfect predictions, 0 indicates random guessing, and -1 means perfect misclassification. This makes MCC intuitive for interpreting how well the model separates churners from non-churners.

The F1-score is effective for imbalanced datasets because it balances precision and recall, focusing on correctly identifying the positive class. However, it only considers the positive class (churners in this case) and doesn't account for true negatives, which may be relevant if you're also interested in correctly predicting non-churners.

The balanced MCC is specifically designed to handle class imbalance and considers all four confusion matrix elements (true positives, true negatives, false positives, and false negatives). This makes it more informative when you care about both classes.

2.4 ROC-AUC

ROC-AUC is less affected by class imbalance compared to accuracy. It considers the true positive rate (sensitivity) and false positive rate, making it a more reliable metric when the positive class (e.g., churners) is much smaller than the negative class (e.g., non-churners).

The ROC curve provides a visual representation of the trade-offs between true positive and false positive rates, helping to understand model performance better. Additionally, it provides a simple but effective way to easily compare the performance of each of our different splits.

3 Conclusion

We decided to use recall as the primary metric to ensure that our model emphasized capturing true churned customers, addressing the problem's core need. The F1-score and Balanced MCC further supported our model evaluations by providing a well-rounded view of precision-recall trade-offs and accounting for imbalances. Together, these metrics aligned well with the project's goals of evaluating real-world model applicability and ensuring generalizability across different subsets of data.