Wrapup

October 28, 2024

In our project, we explored several models to determine the most appropriate approach to classify our data effectively. Beginning with logistic regression as a baseline, we sought to establish a foundational performance level. Logistic regression offered a straightforward method for binary classification. However, its limitations in handling complex, non-linear patterns prompted us to expand our focus to more sophisticated models.

We experimented with four additional models: Multi-Layer Perception (MLP), k-Nearest Neighbors (kNN), Random Forest, and Gradient Boosting. By employing cross-validation and testing on left-out data, we rigorously evaluated each model's performance with a particular focus on the recall score.

Table 1: Performance metrics

Model	Recall↑	F1-score↑	Balanced MCC↑
Baseline	0.3005	0.3987	0.3249
RandomForest	0.8054	0.5585	0.5324
KNN	0.7488	0.5366	0.4834
Gradient Boosting	**0.8202**	0.5495	0.5271
MLPClassifier	0.7586	**0.5917**	**0.5538**

As expected for a baseline model, logistic regression performs significantly poorer across all metrics, particularly recall, compared to more complex models. This serves as a reference for improvement.

In terms of recall, Gradient Boosting is our best model with a recall value of 0.8202, meaning it captures the most churners, which is crucial for our task. By maximizing recall, we prioritized minimising false negatives to ensure that customers likely to churn are correctly identified, enhancing the practical utility of our model in customer retention scenarios. Since balancing between precision and recall is a consideration, MLP is another strong model since it provides the best F1-score and Balanced MCC, suggesting it can handle classification tasks with fewer false positives, which could help stop wasting resources trying to keep customers that are not planning on churning. Random Forest offers a middle ground, combining high recall, good F1-score, and a balanced MCC, making it versatile across different metrics.

The makeup of the data with regards to the lack of variables at play within the dataset may have been a limiting factor for our investigations. Across all models, we found some features such as Gender and Geography aren't particularly helpful for classification analysis. An increase in the number of relevant variables would've been highly useful to us. We believe additional information could've led our models to learn about underlying trends the data may have missed, for example, if we had information pertaining to the time at which the data was collected, we may have been able to make more accurate predictions as the banks performance or the economic situation of the relevant countries during this time may have helped us advise our investigations better.

On the whole, all the models we created performed fairly well, vastly outperforming the baseline model and achieving strong scores in relation to our chosen performance metrics across the board, despite the challenges that an imbalanced dataset proposed.