Introduction to Predicting Bank Customer

In this project, we address Customer Churn predictions as a binary classification problem, aiming to compare the predictive performance of different models on unseen data. Predicting whether a customer will churn will is crucial for enabling the bank to implement effective retention strategies, such as targeted incentive programs.

Selecting the right performance metric is critical for evaluating model effectiveness. Accuracy, while commonly used, may be misleading when dealing with this imbalanced dataset, where the number of customers who churn is significantly smaller than those who do not. In such cases, metrics like precision, recall, F1-score, and ROC curves may offer more meaningful insights, as they take into account class imbalance and reflect the true performance of the models. We also experiment with different train/test splits and resampling techniques to evaluate model performance across varying data distributions. This ensures that our results are not overly dependent on a specific data split and provides insights into how robust the models are when exposed to different data samples.

We use logistic regression as a baseline model which is a simple linear classifier and provides us with an initial set of evaluation metrics. This allows us to assess whether the churn prediction problem is addressed well enough by a basic model or whether more complex non-linear relationships exist in the data that require more sophisticated models. We then extend our analysis by comparing this baseline to more advanced models, including k-Nearest Neighbors (KNN), Gradient Boosting, Random Forests, and Neural Networks. The goal is not only to identify the model with the highest performance but also to evaluate each model's suitability in terms of generalization to real-world scenarios. Each model is tested on a left-out portion of the dataset to ensure that its performance on new and unseen data is representative.

Through this approach, we aim to demonstrate that the "best" model is determined by more than just the one with the highest performance, but the model that generalises best to new data for the purposes of predicting bank customer churn.