ONE PAGER FOR FINAL PROJECT(LNL)

-Akshay Ramdev, Peter Ye, Steven Sun

Business Problem:

In the competitive landscape of the banking sector, the ability to effectively predict customer behaviour is crucial for driving growth and maintaining a competitive edge. The dataset offers a unique opportunity to delve into customer data and extract actionable insights, particularly in the domain of personal loan offerings.

The core business problem revolves around optimizing the personal loan approval process. Banks face the dual challenge of maximizing loan acceptance rates while minimizing the risk of default. Accurately predicting which customers are likely to accept a personal loan can lead to more targeted and efficient marketing strategies, improved customer satisfaction, and better allocation of resources.

Non-Linear Model:

Random Forest Classifier

Random Forest is an optimal choice for our dataset due to its ability to adeptly handle mixed data types and its robustness against class imbalance, which is common in loan acceptance data. This algorithm is particularly advantageous for its feature importance insights, crucial for understanding key drivers in loan acceptance. Unlike single decision trees, Random Forest mitigates overfitting through its ensemble approach, enhancing model generalizability and accuracy. Furthermore, its relative insensitivity to hyperparameter tuning makes it user-friendly and efficient for diverse datasets, making it a reliable, high-performing solution for our predictive modelling needs in the banking sector.

Assumptions:

Random Forest, an ensemble technique utilizing multiple decision trees, is valued for its robustness and minimal data assumptions. Its performance, however, hinges on several factors:

The presence of sufficient relevant features (as too many irrelevant ones introduce noise), adequate sample size (to ensure variance reduction through diverse trees), and high-quality data. While capable of modelling complex, non-linear relationships, extremely intricate patterns may require more or deeper trees, increasing computational load.
Additionally, it assumes tree independence and the ability to distinguish signal from noise, though performance can suffer in extremely noisy datasets. Finally, it operates under the assumption that training data patterns are indicative of unseen data, meaning significant distribution shifts can impact model accuracy.

Applications:

Healthcare: Disease Diagnosis and Progression
Finance: Credit Scoring and Fraud Detection
E-Commerce: Customer Segmentation and Recommendation Systems