Loan Servicing Risk Alert System: Predicting Customer Overstress in a Microfinance Setting

Financial Data Analytics – Institute of Business Administration:
Lecturer Vishal Khemani : vishal@iba.edu.pk :
Teaching Assistant Jamal Nasir : j.nasir.25260@khi.iba.edu.pk

April 17,2025

Project Description

This case study challenges participants to develop a predictive model for identifying overstressed microfinance clients using a real-world financial dataset. The context is a loan servicing environment where early warning systems can play a critical role in helping financial institutions flag vulnerable clients before default occurs. The task involves binary classification, where the goal is to predict whether a customer is overstressed (label = 1) or not (label = 0), using features derived from customer transactions, profile, and financial behavior.

The model aims to support risk mitigation strategies, allowing institutions to intervene early with financial counseling or repayment rescheduling. This case thus mirrors how data science can be integrated into microfinance operations to improve financial inclusion and client well-being while reducing loan default risks. Participants are expected to treat the project with both technical rigor and a sense of ethical responsibility, recognizing the implications of false positives and false negatives in the financial lives of low-income customers.

Submission and Evaluation Procedure

All participants are required to join the Kaggle competition link `https://www.kaggle.com/t/a3b362d2943d4d7bb770629880d0f57e`. It is essential to register using your full name as your Kaggle display name for identification purposes. Submissions using ambiguous usernames may not be evaluated. Students are allowed to make up to ten submissions per day and are strongly encouraged to begin early, experiment widely, and iterate con-

tinuously.

A Google Form must be filled out after every submission to log the approach used. The link to this submission logging form is `https://forms.gle/VzUsT2anT9a8TySZ8`. Each entry should detail the classification algorithm implemented, preprocessing and transformation strategies applied, how missing values were handled, and any parameter tuning done. These records are important for evaluating your experimentation process and model-building logic.

Unlike traditional contests, the evaluation for this assignment will prioritize the breadth of your experimentation and the learning process rather than the final accuracy score alone. Students are expected to try multiple models and analyze them critically. The number of submissions and variety in methods explored will carry greater weight than leaderboard ranking.

At the end of the competition, each student must submit two deliverables on the LMS. The deadline is April 30th (Note that this can't be extended since the kaggle doesn't allow this so request at the end won't be considered)The first is a well-commented Jupyter notebook that includes the code for your best-performing model and reproduces your Kaggle result upon execution. The notebook must also include markdown cells that explain your thought process, performance summaries, and key insights gained. The second deliverable is a PowerPoint presentation with one slide for each classification algorithm attempted. Each slide should briefly describe the model used, its performance, and the advantages or disadvantages observed.

Modeling Requirements and Dataset Overview

Participants must experiment with a diverse set of classification models, including but not limited to Decision Trees, Naive Bayes, K-Nearest Neighbors, Random Forests, Gradient Boosting, AdaBoost, LightGBM, XGBoost, CatBoost, and ExtraTreesClassifier. Use of stacking or ensemble methods is encouraged where appropriate. Students are also expected to perform appropriate data cleaning, transformation, feature engineering, and tuning as part of the modeling process.

The dataset consists of three files: a training dataset containing both features and the binary target variable Y, a test dataset containing only the features, and a sample submission file that defines the structure for submitting predictions to Kaggle. The primary evaluation metrics for this classification task are precision, recall, and AUC-ROC. Given the potential consequences of misclassification, especially false negatives (i.e., missing an overstressed customer), special attention should be paid to recall. AUC-ROC will be used to assess the balance between sensitivity and specificity, particularly in the context

of imbalanced data.

Conclusion

This assignment offers a practical opportunity to apply classification techniques to a high-stakes problem in financial services. It aims to develop your technical competencies while highlighting the societal impact of machine learning. Consistency, reflection, and breadth of experimentation will be the most valued aspects of your participation. We look forward to your innovative solutions and thoughtful analysis.