Mark Axelus Gerry Peng Manya Goel Christina Mcneice

1. Overview

This project aims to build and compare machine learning models in the financial domain that predict borrower default risk using the “Credit Risk Benchmark” dataset from Kaggle. The dataset contains approximately 15,000 borrowers' data, as well as their demographics, financial, and credit-related features.

1.1 The Scope

- **Dataset:** https://www.kaggle.com/datasets/adilshamim8/credit-risk-benchmark-dataset
- **Goal:** With the above dataset, we aim to develop, analyze, and validate a proof-of-concept pipeline that:
    1. Predicts a two-year default risk (0=No default, 1=default)
    2. Quantifies feature impact to automate underwriting

1.2 Preliminary Methods

- Logistic regression (binary classification)
    - Trained via batch gradient descent

1.3 Expected Outcomes

By the end of the term, we will have:

- A cleaned and processed dataset.
- A baseline logistic regression model trained via batch gradient descent.
- Performance metrics for the model.
- A high-level discussion of which feature(s) seem important.
- A list of next steps for a deeper analysis.