datasci 207 Group Project Milestone

David Griffin (david_griffin@berkeley.edu)
Anthony Li (anthonyleeqy@berkeley.edu)
Sabrina Sun (yolosun4488@berkeley.edu)
Duy Vo (duyvo@berkeley.edu)

https://github.com/duygithub/berkeley_datasci_207_proj

Motivation

Peer-to-Peer lending platforms like LendingClub became popular after the Great Recession when credit tightened. Due to several factors, LendingClub transitioned away from a P2P model towards institutional lenders (Jagtiani and Lemieux 2018), similar to those now involved in private credit. There is interest in private credit for investors seeking diversified, higher-yielding income sources because US household debt is at a record 18.59 trillion (New York Federal Reserve Federal Reserve Bank of New York 2025). We want to learn from LendingClub's data to help inform investment decisions in the a future credit-tightening cycle.

Prior work by (Chang, S. et al. 2015) focused on predicting defaults. We belive that there is much greater utility for an investor to know how much they can make rather than a loan that defaults. In our EDA, we can see that approximately 6.8% of loans that defaulted still have a positive return for investors because the interest rate is sufficiently high and the loans defaulted later in the term.

Data Description

The dataset we pulled from Kaggle derives from LendingClub, a U.S.-based firm that operated a peer-to-peer lending platform. The dataset covers all loans originated between 2007 and 2015 and includes borrower information, loan characteristics, and final loan outcomes.

The raw data contains 2,260,668 observations with 145 features (Singh 2021). We are only keeping finalized loans—those with a `"Default"`, `"Charged Off"`, or `"Fully Paid"` status—and the input features `grade`, `sub_grade`, `term`, `emp_length`, and `annual_inc`. We added two