datasci 207 Group Project Proposal

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

Introduction

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.

Data

Our data comes 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 (Barret 2010) and includes both borrower characteristics and final loan outcomes, enabling analysis of credit risk and loan performance.

The dataset contains approximately 890,000 observations and 75 features. We will use loan status as the outcome variable, with credit indicators and loan attributes as predictors.

The data is publicly available through Kaggle: https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv/data (Singh 2021).

Related Work

(Trinh, L. T. 2024) used a similar peer-to-peer lending data and modeled credit risk with 9 different statistical and machine learning models. They concluded that Gradient Boosting Decision Tree is the best model for most of the evaluation metrics.

(Chang, S. et al. 2015) used LendingClub data and fit Logistic Regression, SVM, and Naive Bayes. They concluded that Naive Bayes with Gaussian performs the best at predicting default rate. Furthermore, by applying the best model, the investment return could grow by 50%.

(Souadda, L. I. et al. 2025) evaluated hyperparameter optimization methods for machine learning models with similar lending data, and Bayesian methods (Hyperopt, Optuna) achieved 75.7 times faster runtime than Grid Search while maintaining similar accuracy.

Methodology

Target: Loan status with a binary default indicator (Charged Off vs Fully Paid/Current), more multi-class labels for optional granular analysis. Dataset split: loans with observed outcomes are retained, and a time-aware train/validation/test split is applied to prevent data leakage. Preprocessing: handling missing values, consolidating rare categories, encoding categorical variables, standardization. Class weighting or stratified sampling to address class imbalance. Modeling: Logistic regression as baseline. Tree-based models (random forest, gradient boosting) as comparisons. Evaluation: AUC-ROC, precision, recall, F1, and calibration metrics. Hyperparameters are tuned via cross-validation, and feature importance is analyzed to interpret default risk drivers. Robustness is assessed across loan grades and time periods.

References

Barret, Victoria. 2010. “Making Personal Loans for Fun and Profit.” Forbes. https://www.forbes.com/forbes/2010/1220/investing-lending-club-credit-cards-personal-loans-for-fun.html.

Chang, S. et al. 2015. “Stanford CS229 Project: ”Predicting Default Risk of Lending Club Loans.” https://cs229.stanford.edu/proj2015/199_report.pdf.

Federal Reserve Bank of New York. 2025. “Quarterly Report on Household Debt and Credit (2025:Q3).” Center for Microeconomic Data. https://www.newyorkfed.org/microeconomics/hhdc.

Jagtiani, Julapa, and Catharine Lemieux. 2018. “The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform.” Working Paper 18-15R. Federal Reserve Bank of Philadelphia. https://doi.org/10.21799/frbp.wp.2018.15.

Singh, Adarsh. 2021. "Lending Club Loan Data." https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv/data.

Souadda, L. I. et al. 2025. "Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning." https://www.mdpi.com/2571-9394/7/3/35.

Trinh, L. T. 2024. "A Comparative Analysis of Consumer Credit Risk Models in Peer-to-Peer Lending." https://www.emerald.com/jefas/article/29/58/346/1212619/A-comparative-analysis-of-consumer-credit-risk.