Predicting off-charge loans for risk management in P2P lending market

Thu Ha MAI
Aouatef HAOUFADI
Maria Yamina TEHAR

Advanced Methods in Big Data
Final Project Pitch

Supervisors
E. Gallic, S. Hué, P. Michel

13th December 2022

Introduction

01 Context

- Lending Club is a peer-to-peer lending company
- Headquartered in San Francisco, California
- Online lending platform

---

02 Study

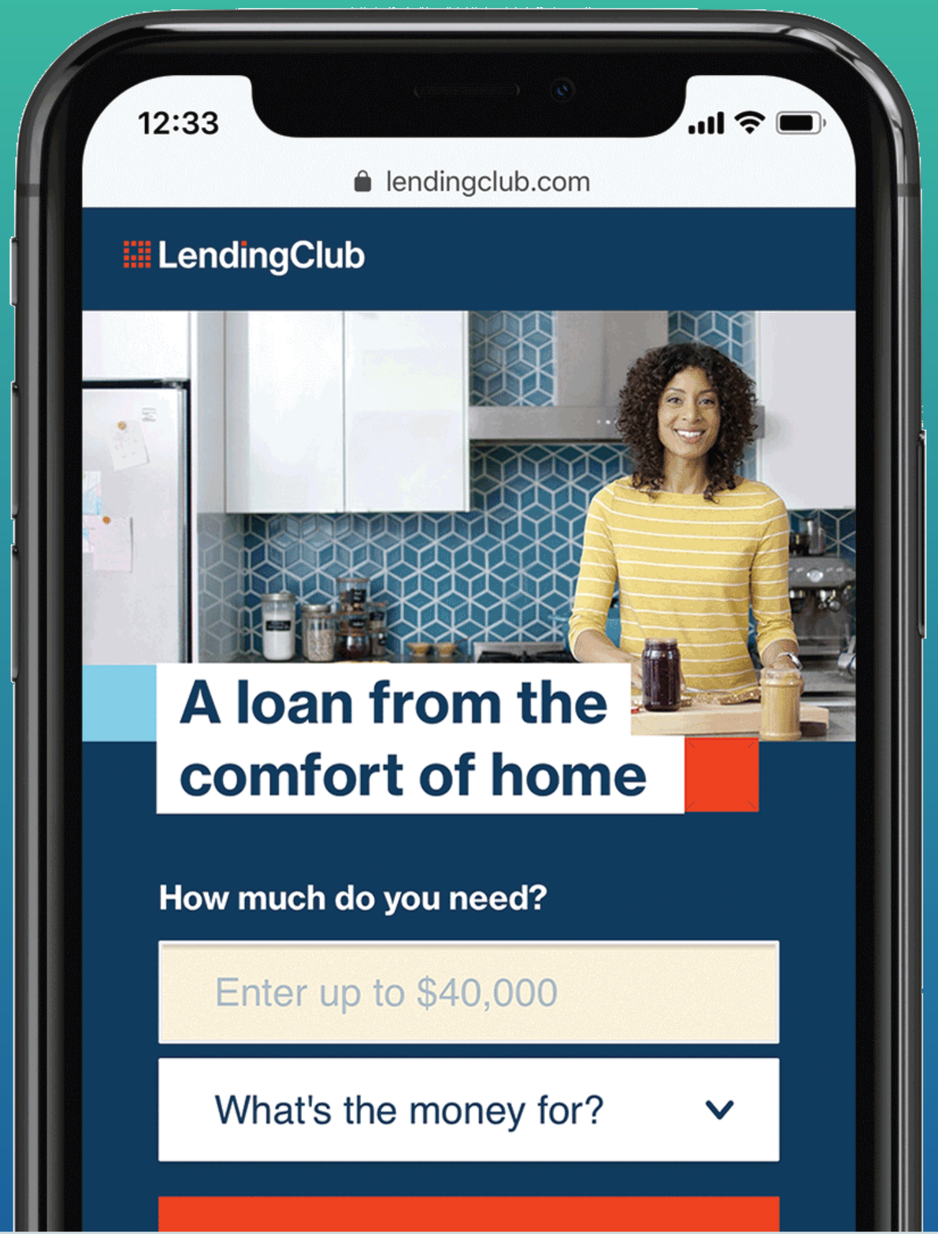
- Build classification models to predict loan defaulters based on the applicants' profiles

---

03 Objective

- Understand the driving factors behind loan default, i.e. the variables which are strong indicators of default

Source: lendingclub.com



Data Overview

Data Description

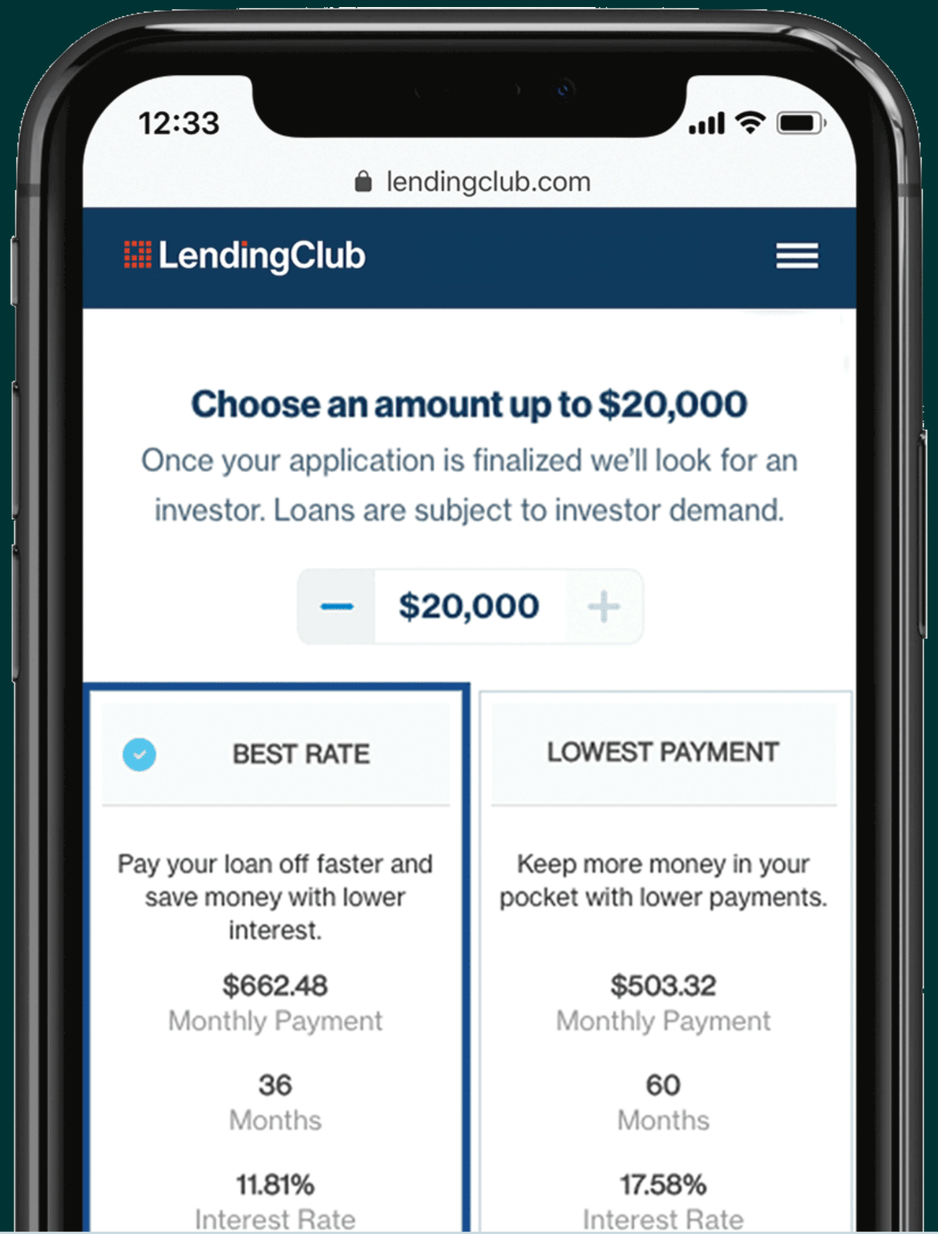
- Source of dataset: Kaggle[1]
- The dataset contains
  - 396 030 observations
  - 27 features
    - Features related to the applicant (e.g., income)
    - Features related to loan characteristics (e.g., loan amount)

Exploratory Data Analysis

- 01 target variable: **Loan Status**
  - Fully paid: 80% # observations
  - Charged off: 20%
- Multicollinearity problem

[1] https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/data

Source: lendingclub.com



Next Steps

We are currently in the stage of Data Exploratory Analysis & Data Preprocessing. Hence, the below methods & techniques are suggestions and will be adapted if necessary.

Any recommendations from the Professors and colleagues are more than welcome!

Data Preprocessing

- Outlier Detection & Treatment:
  - Inter Quantile Range
  - Removing or Capping
- Imbalanced dataset Treatment:
  - Oversampling method: SMOTE-NC

Information Reduction

- Dealing with **Multicollinearity** problem
- Handling **high-dimensional** setup
  - Dimension reduction: Principal Component Analysis
  - Shrinkage: Lasso

Classification Models

- Statistical Model
  - Logistic Regression
- Machine Learning Models
  - Random Forest
  - Neural Networks
- Explainability with SHAP

Thank you!