Team 28
Corban Chiu (cpc75), Tyler Carreja (tcc86), Caroline Ryu (jr894)

Background

LendingClub was established in 2006 as a peer-to-peer lending platform as a Facebook application to facilitate the connection between borrowers and investors who fund their loans. Its primary aim was to enable the matching of investors and borrowers, thereby eliminating traditional banks from the process. While LendingClub assesses and services loans that are approved, the decision to fund a loan is made by investors based on loan details and the debtor's credit history. LendingClub benefits the borrowers by allowing them to obtain loan rates without affecting their credit score, while also offering fixed rates and transparent terms. The company prided itself on offering reduced rates for borrowers and increasing returns for investors.

Despite its shift towards the neo-bank sector and discontinuation of its peer-to-peer platform due to increasing regulatory scrutiny, LendingClub's dataset from loans serviced between 2007 and 2018 remains a valuable resource for analyzing the risk of loan default for new loan applications. The dataset comprises financial and demographic information that LendingClub used to assess the creditworthiness of potential borrowers and determine whether to approve or reject loan applications. A detailed methodology for analyzing the data could provide valuable insights into the lending process and aid in identifying factors that contribute to loan default.

Problem Statement

Housing loans were a major factor in the 2008 market crash, and have been the subject of much debate over the past decade and a half. At what point does a loan become too risky? What are factors that contribute to creating risk in a loan? We can identify these factors that determine the strength of loans using LendingClub's datasets on loan applications.

To this end, we will use learning techniques to model the loan application process. There are two models that we want to create. The first model will predict whether a loan application will be accepted or rejected. The second model will predict the loan grade of the accepted loan applications. This will give us insight into the loan tolerance of lenders and factors that contribute to the risk of the loans.

Data

We are using two datasets, accepted_2007_to_2018Q4.csv and rejected_2007_to_2018Q4.csv, which are both linked on our GitHub repo. The accepted dataset has detailed information on over 2,000,000 accepted applications including LendingClub's assigned grade and the borrower's credit and debt history. The rejected dataset has significantly less detailed information on over 27,000,000 rejected applications mostly including the loan's risk score and the borrower's debt to income ratio. The primary data types included in both datasets are floats, strings, and datetimes.