Lendingclub's loan default prediction

Rafiq Islam

2023-02-23

Table of contents

Notebook

Project Overview

LendingClub is a U.S.-based financial services company that initially began as a peer-to-peer (P2P) lending platform, allowing individual investors to lend directly to individual borrowers through a marketplace. Founded in 2006, it grew quickly to become one of the largest and most popular P2P lending platforms, helping connect borrowers in need of personal loans with investors looking for alternative investment opportunities. LendingClub has been known for its transparent data-sharing practices, making anonymized loan data available to researchers, data scientists, and investors. This data is widely used in financial research, especially for predictive modeling of loan risk and borrower behavior. The aim of this data science project is to build a machine learning model to predict the likelihood of a loan default.

More information about LendingClub

1. **P2P Lending Model (Early Focus):**

- Borrowers could apply for loans, typically unsecured personal loans, through LendingClub's platform.
- Loans were then funded by individual investors who could review borrowers' profiles, risk grades, and other financial information before committing funds.
- The model offered borrowers a way to access loans outside traditional banks, often at lower interest rates, and allowed investors to diversify by spreading investments across multiple loans.

2. **Risk and Return:**
   - LendingClub assigned credit grades (A–G) to each loan based on creditworthiness, which affected interest rates. Higher risk meant potentially higher returns for investors but also higher chances of default.
   - Investors bore the risk if a borrower defaulted, which was a notable risk factor compared to FDIC-insured deposits.
3. **Shift to a Bank Model:**
   - Over time, LendingClub transitioned away from P2P lending and restructured as a more traditional bank, obtaining a bank charter in 2021.
   - It now offers banking products, such as high-yield savings accounts, and operates more like a digital bank while still focusing on lending products.
4. **Borrower and Loan Profiles:**
   - LendingClub primarily focuses on personal loans for debt consolidation, credit card refinancing, home improvement, and other purposes.
   - Borrowers' profiles typically include information on income, credit score, debt-to-income ratio, and loan purpose, which is used for assessing risk.

Dataset

The dataset is a publicly available data from kaggle.com. It originally contains 396030 entries, with 100.4 MB. However, for easy github push, I reduce the dataset slightly in order to have size smaller than 100 MB. So, in the reduced form, it has 395900 entries with the following columns:

- `loan_amnt:` The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.

- `term:` The number of payments on the loan. Values are in months and can be either 36 or 60.

- `int_rate:` Interest Rate on the loan installment The monthly payment owed by the borrower if the loan originates.

- `grade`: LC assigned loan grade
- `sub_grade`: LC assigned loan subgrade
- `emp_title`: The job title supplied by the Borrower when applying for the loan.
- `emp_length`: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- `home_ownership`: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
- `annual_inc`: The self-reported annual income provided by the borrower during registration.
- `verification_status`: Indicates if income was verified by LC, not verified, or if the income source was verified
- `issue_d`: The month which the loan was funded
- `loan_status`: Current status of the loan
- `purpose`: A category provided by the borrower for the loan request.
- `title`: The loan title provided by the borrower
- `zip_code`: The first 3 numbers of the zip code provided by the borrower in the loan application.
- `addr_state`: The state provided by the borrower in the loan application.
- `dti`: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- `earliest_cr_line`: The month the borrower's earliest reported credit line was opened.
- `open_acc`: The number of open credit lines in the borrower's credit file.
- `pub_rec`: Number of derogatory public records.
- `revol_bal`: Total credit revolving balance.

- `revol_util:` Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

- `total_acc:` The total number of credit lines currently in the borrower's credit file
- `initial_list_status:` The initial listing status of the loan. Possible values are – W, F
- `application_type:` Indicates whether the loan is an individual application or a joint application with two co-borrowers.

- `mort_acc:` Number of mortgage accounts.

- `pub_rec_bankruptcies:` Number of public record bankruptcies

Here `loan_status` is the predictive or dependent variable and the rest are predicting or independent variables aka features. We want to predict if the loan will be `Fully Paid` or `Charged Off` given the feature values.

Stakeholders

Key Performance Indicators (KPIs) of LendingClub

Modeling

Results and Outcome

Future Directions