Case study Phase 1

Introduction and Objectives

Samiya Islam

2023-10-31

Introduction and Objectives

1. **(i) As an investor, what are the decisions you would need to make? (ii) Which of those decisions can you make using the available data from LendingClub and which one(s) would require additional resources?**
    a. Credit history
        i. fico score, column “last_fico_range_high/low”
    b. Current employment/income
        i. “emp_title”
        ii. ”empt_length”
        iii. ”annual_inc”
    c. Intended use of the loan
        i. “purpose”
        ii. ”title”
    d. Other current active loans
        i. “open_acc”
        ii. “open_act_il”
    e. Current bank account balance
        i. may need additional resources
    f. Credit payment history
        i. “Dti”
    g. Selecting loans to invest in
        i. (e.g., based on loan attributes like "loanAmnt" and "intRate").
    h. Determining investment amounts
        i. (e.g., based on available funds and loan characteristics).
    i. Assessing credit risk
        i. (e.g., using fields like "grade" and "dti").
    j. Timing of investments
        i. (e.g., considering the listing date "listD").

2. (i) What is your objective when making those decisions in Q1? (ii) Explain how you would be able to distinguish “better” decisions from “worse” ones using the data?

When making the decisions from Q1, our main goal is to decrease risks while also trying to maximize return on investment. By looking at the dataset and the relevant columns, we can try to differentiate “better" decisions from "worse" decisions by analyzing credit history, current active loans, dti, etc. For example, a “better decision” would be that borrowers with higher credit scores may be more likely to pay off their loans and would make the investor more willing to loan to them.

3. Note that loans are temporal entities (36 or 60 months-long term). Different loans could default at different times; some will default soon after approval, some much later. Some, on the other hand, might be repaid early, before their term ends. Would these facts affect your *downstream analysis* and decision-making? How/Why?

The Case study dictionary Excel file column "browseNotesFile"contains loans with different term lengths like 36 or 60 months. Some loans may default soon after approval, while others might default much later.

The temporal nature of a loan has many implications on our analysis. This is because loans have multiple factors such as Interest payments, Time value of money, and risk factors which are all impacted by term length. This is true for both borrowers and investors. For example, longer term loans will have a greater impact on borrowers due to interest rates. Similarly, as an investor we may want a faster return so we would choose loans with a shorter term. Oftentimes shorter terms present less of a risk since there is less time for a borrower to default. Longer term loans also tend to make the investment more sensitive to change. Though this is true there is still more of an incentive for Investors to use long term loans since the higher interest rate equals a higher return. Because of these reasons the default time of an event or even the early repayment affects the expected return for investors, adding another layer of analysis and decision-making, we as investors must account for. This is due to the ever-evolving economic landscape and the context of the loans lended.

4. Based on the discussions thus far, do you think historical data would be helpful? In which ways could you use such data to help make the decisions of your interest?

I think historical data could prove to be helpful because it allows you to see trends and patterns over time that can provide insight into current or future activities and gain guidance into whether going through with an investment is a good idea. For example, by analyzing the trends in loan performance and payment over time, we can develop predictive models to assess credit risk and gain a sense of measure of whether certain decisions are “better” or “worse.” Another way we can use the data to make helpful decisions is by analyzing the credit payments of a user, and look at the frequency of how often they make their payments on time, how often their payments are late, and the ups and downs of their credit scores. But we do feel like we shouldn’t over rely on the historical data as an investor because I think while past performance can offer valuable insights, it should serve as a guide rather than a sole determinant of investment decisions. We can’t be too pessimistic about unprecedented events. For example I feel like COVID-19 pandemic, which has shown that relying solely on historical patterns can lead to significant risks and losses. So essentially for us to navigate the complexities of today's markets we as a team think it is crucial to stay flexible, consider a range of factors, including current market conditions and maintain a diversified portfolio that can better withstand unexpected disruptions.

5. Next you will take a look at the data.

(i) Write down a high-level description of the different features—that is, the variables describing the loans. How would you categorize these features? (Note that there may be multiple ways of categorizing the features; think in terms of the source of the measurements, the type, and temporal characteristics.)

We carefully took a look at the datasets, specifically “LoanStats_securev1_2019Q4.csv" and the Case study dictionary Excel file. The excel file helped us understand what each column represented. High level description about loan information includes loan_amnt that is the amount requested, funded_amnt is the total amount funded, funded_amnt_inv is the total amount funded by investors, term is basically the term of the loan for example 36 months or 60 months, int_rate is the interest rate of the loan and then installment is the monthly payment amount.

We found some borrower information for example, grade and sub_grade which is LendingClub's grading system for loan risk. Then emp_title is the employment title of the borrower, emp_length is employment length in years, home_ownership is the type of home ownership like RENT, MORTGAGE. Then we have annual_inc which is basically the annual income of the borrower and verification_status which is verification status of the borrower's income. Further in our group work, we will integrate these values to depict a clearer context of how one variable affects another, and in turn, how they influence an investor’s decision.

Moving on to credit history there's dti which is debt-to-income ratio then delinq_2yrs: which is the number of 30+ days delinquencies in the last 2 years, earliest_cr_line is the month the borrower's earliest reported credit line was opened, fico_range_low and fico_range_high is the FICO credit score range. Then we have inq_last_6mths which is the number of credit inquiries in the last 6 months. We also found variables for analyzing the loan performance. For example loan_status is the current status of the loan like Current, Late, Fully Paid. Then there's information we have about total_pymnt is the total payment received to date, total_rec_prncp is the total principal received to date, total_rec_int is the total interest received to date and lastly total_rec_late_fee is the total late fees received.

The way we thought of categorizing it is based on the type. There are three types we noticed numerical features such as loan amount, interest rate, annual income, and credit scores.
Then we noticed various categorical features like loan grade, employment title, home ownership, and loan purpose. We think temporal features could be another category related to the loan's term and credit history, like the month of the earliest credit line opened.

(ii) Just based on the feature descriptions, give an example to features that are likely to be (strongly) related if plotted on a scatter plot.

Just based on the feature descriptions an example of strongly related features in the dataset could be "intRate" and "grade." because higher-grade loans tend to have lower interest rates, creating a strong relationship and annual_inc (Annual Income) and loan_amnt (Loan Amount) for borrowers with higher annual income since they may be more likely to request larger loan amounts.

(iii) Which do you think are most valuable to an investor like yourself?

In our opinion and based on our research, as investors, we believe that the most valuable features in the dataset include the loan grade (grade) and interest rate (int_rate) since, in our view, they are critical for assessing credit risk and potential returns. The loan amount (loan_amnt) is, in our opinion, essential to determine the size of our investment, while, from our research, the borrower's annual income (annual_inc) provides insights into their financial stability. Loan purpose (purpose) is valuable, we think, for tailoring investments to our specific preferences and risk tolerance, as it indicates the intended use of the loan. Additionally, from our perspective, the debt-to-income ratio (dti) is crucial for evaluating the borrower's financial health and ability to manage debt. Lastly, as we've observed from our research, the borrower's creditworthiness, represented by the FICO score range (fico_range_low and fico_range_high), is a key factor in assessing risk. These features, in our view, collectively allow us to make informed decisions and construct a well-balanced investment portfolio.