Individual Project Report Data analytics for Loan Club investment

February 18, 2021

In this case, we follow Triss Merigold, a young professional looking to diversify her investment portfolio.[[1]](#footnote-2) Triss graduated from a Masters in Data Science program, and after four successful years as a product manager in a tech company, she had managed to save a sizable amount of money. She now wants to start diversifying her savings portfolio. So far, she has focused on traditional investments (stocks, bonds, etc.) and she now wants to look further afield.

One asset class she is particularly interested in is peer-to-peer loans issued on online platforms. The high returns advertised by these platforms seem to be an attractive value proposition, and Triss is especially excited by the large amount of data these platforms make publicly available. With her data science background, she is hoping to use data analytics tools on this data to come up with lucrative investment strategies. In this case, we follow Triss as she tackles this problem.

# Background on Peer-to-Peer Lending

Peer-to-peer lending refers to the practice of lending money to individuals (or small businesses) via online services that match anonymous lenders with borrowers. Lenders can typically earn higher returns relative to savings and investment products offered by banking institutions. However, there is of course the risk that the borrower defaults on his or her loan.

Interest rates are usually set by an intermediary platform on the basis of analyzing the borrower’s credit (using features such as FICO score, em- ployment status, annual income, debt-to-income ratio, number of open credit lines). The intermediary platform generates revenue by collecting a one-time fee on funded loans (from borrowers) and by charging a loan servicing fee to investors.

The peer-to-peer lending industry in the U.S. started in February 2006 with the launch of Prosper,2 followed by LendingClub.3 In 2008, the Securi- ties and Exchange Commission (SEC) required that peer-to-peer companies register their offerings as securities, pursuant to the Securities Act of 1933. Both Prosper and LendingClub gained approval from the SEC to offer in- vestors notes backed by payments received on the loans.

By June 2012, LendingClub was the largest peer-to-peer lender in the

U.S. based on issued loan volume and revenue, followed by Prosper.4 In December 2015, LendingClub reported that $15.98 billion in loans had been originated through its platform. With very high year-over-year growth, peer- to-peer lending has been one of the fastest growing investments. According to InvestmentZen, as of May 2017, the interest rates range from 6.7%-22.8%, depending on the loan term and the rating of the borrower, and default rates vary between 1.3% and 10.6%.5

LendingClub issues loans between $1,000 to $40,000 for a duration of either 36 months or 60 months. As mentioned, the interest rates for borrowers are determined based on personal information such as credit score and annual income. A screenshot of the LendingClub homepage is shown in Figure 1. In addition, LendingClub categorizes its loans using a grading scheme (grades A, B, C, D, E, and F, where grade A corresponds to the loans judged to be “safest” by LendingClub). Individual investors can browse loan listings online before deciding which loans(s) to invest in (see Figure 2). Each loan is split into multiples of $25, called notes (for example for a $2,000 loan, there will be 80 notes of $25 each). Investors can obtain more detailed information on each loan by clicking on the loan – Figure 3 shows an example of the additional information available for a given loan. Investors can then purchase these notes in a similar fashion to “shares” of a stock in an equity market.

2https://www.prosper.com/

3https://www.lendingclub.com/

4https://en.wikipedia.org/wiki/LendingClub

5http://www.investmentzen.com/peer-to-peer-lending-for-investors/lendingclub (date

accessed: June 2018)

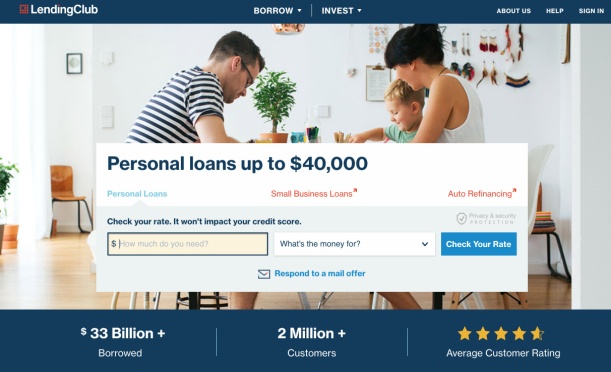


Figure 1: Screenshot of the LendingClub homepage.

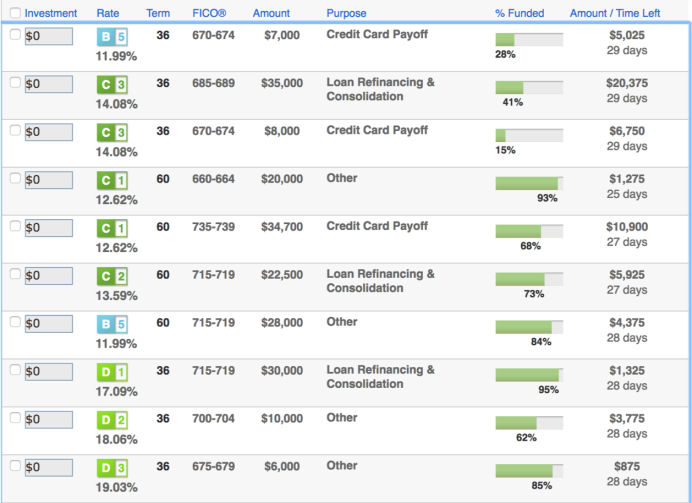


Figure 2: Example of loan listings (source: LendingClub website, date ac- cessed: May 2018).

Of course, the safer the loan the lower the interest rate, and so investors have to balance risk and return when deciding which loans to invest in.

One of the interesting features of the peer-to-peer lending market is the richness of the historical data available. The two largest U.S. platforms (LendingClub and Prosper) have chosen to give free access to their data to potential investors. This then raises a whole host of questions for investors like Triss:

ˆ Is this data valuable when selecting loans to invest in?

ˆ How could an investor use this data to develop machine learning tools to guide investment decisions?

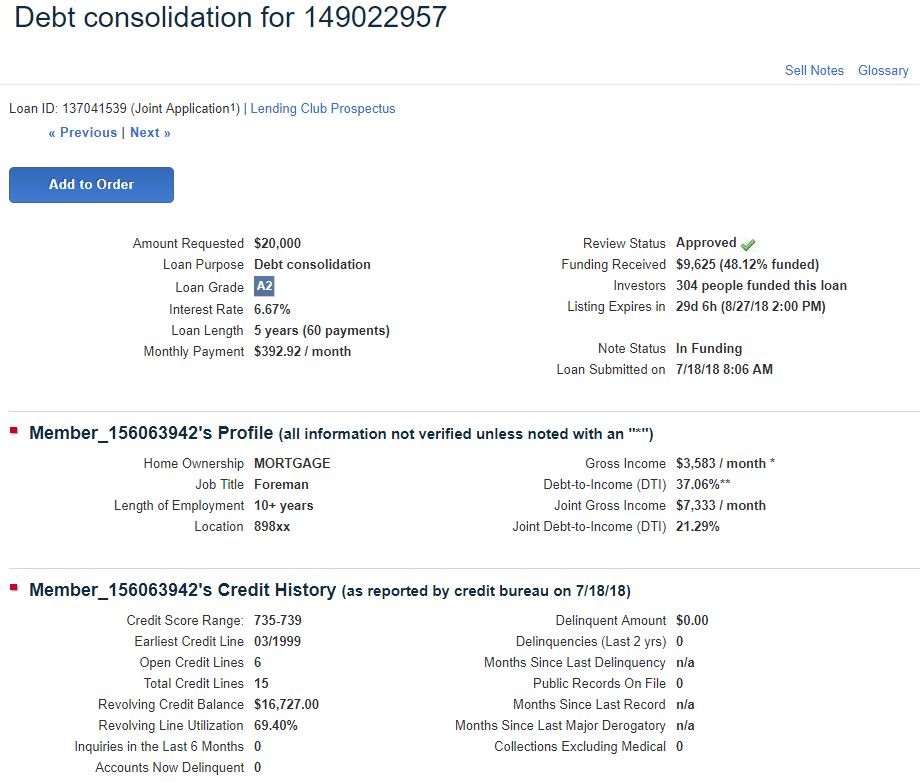


Figure 3: Example of a detailed loan listing, for a grade A loan. Information available to investors include the length of the borrower’s employment, the borrower’s credit score range, and their gross income, among others (source: LendingClub website, date accessed: July 2018).

The goal of this case study is to explore the answers to the questions above.

# Datasets and Descriptive Statistics

As mentioned, the datasets from LendingClub (and Prosper) are publicly available online.6 The dataset contains comprehensive information on all loans issued between 2007 and the fourth quarter of 2018 (a new updated dataset is uploaded every quarter). The data records hundreds of features including the following, for each loan:

* 1. Interest rate, 2.Loan amount,

1. Monthly installment amount,
2. Loan status (e.g., fully-paid, default, charged-off),
3. Several additional attributes related to the borrower such as type of house ownership, annual income, monthly FICO score, debt-to-income ratio, and number of open credit lines.

The dataset we will examine in this case study contains a sub-sample of 3,000 loan listings. The original dataset contains over 750,000 loan listings with a total value exceeding $10.7 billion. In this dataset, 99.8% of the loans were fully funded (at LendingClub, partially funded loans are issued only if the borrower agrees to receive a partial loan). Note that there is a significantly larger number of listings starting from 2016 relative to previous years.

The definition of each loan status is summarized in Table 1. *Current* refers to a loan that is still being reimbursed in a timely manner. *Late* corresponds to a loan on which a payment is between 16 and 120 days overdue. If the payment is delayed by more than 121 days, the loan is considered as being in *Default*. If LendingClub has decided that the loan will not be paid off, then it is given the status of *Charged-Off*.7

6The discussion in this case study will focus on the LendingClub data. However, a

similar analysis can be conducted using Prosper data.

7Note that sometimes the \Charged-O\_" status will occur before \Default" if/when

the borrower has \_led bankruptcy or has noti\_ed the intermediary platform.

Number of Days Past Due Status

0 Current

16-120 Late

121-150 Default

150+ Charged-Off

Table 1: Loan statuses in LendingClub.

These dynamics imply that five months after the term of each loan has ended, every loan ends in one of two LendingClub states – fully paid or charged-off8. To conform with the common meaning of the word, we call these two statuses *fully paid* and *defaulted* respectively, and we refer to a loan that has reached one of these statuses as *expired*.

One way to simplify the problem is to only consider loans that have expired at the time of analysis. For example, for an analysis carried out in April 2018, this implies looking at all 36-month loans issued on or prior to October 31st 2014 and all 60-month loans issued on or prior to October 31st 2012.

As illustrated in Figure 4, a significant portion (13.5%) of loans ended in *Default* status; depending on how much of the loan was paid back, these loans might have resulted in a significant loss to investors who had invested in them. The remainder were *Fully Paid* – the borrower fully reimbursed the loan’s outstanding balance with interest, and the investor earned a positive return on his or her investment. Therefore, to avoid unsuccessful investments, the goal is to estimate which loans are more (resp. less) likely to default and which will yield low (resp. high) returns. To address this question, one can use several data analytic tools on historical data to construct informed investment strategies.

8For example, if a borrower defaults on a loan in the last month of a 36-month loan, it

would take another five months for the loan to be charged-off.

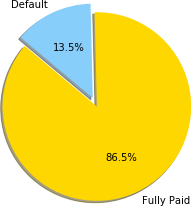


Figure 4: Proportion of *Fully Paid* versus *Default* for terminated loans (by November 2017).

# Questions

Read the case carefully. Download the data file “Loan club subsample”. Then, consider the following questions:

1. Fundamentally, what decisions will Triss need to make?
2. What is Triss’s objective when making these decisions? How will she be able to distinguish ‘better’ decisions from ‘worse’ ones?
3. Why would we even think past data would be helpful here? How could Triss use past data to help make these decisions?
4. Take a look at the data. Write a high-level description of the different “attributes”—the variables describing the loans. How would you cate- gorize these attributes (e.g. categorical/numeric etc.)? Which do you think are most important to an investor like Triss?
5. When looking through the data, you might have noticed that some of these variables seem related. For example, the total payment variable is likely to be strongly correlated to the loan status. (Why?) Does that affect the usability of the variables?
6. Based on your answers in the previous questions, what data analytic task(s) and methodologies could help Triss? What would the target variable be?
7. Describe the different tasks Triss would have to perform during the various stages of the data mining process that we discussed in class.

E.g. during the data preparation step, what tasks might Triss perform?

1. How can Triss choose between different models for the same task?
2. Triss was acutely aware of the fact the data she was using to train her

models dated from as far back as 2009, whereas she was hoping to apply it going forward. She wanted, therefore, to investigate the stability of her models over time. How might she do this?

After you consider the questions, write a report where you discuss your answers. The report you submit should be up to 2,500 words in length.

You are allowed to consult any sources (e.g. books, research papers, reports), but please make sure you provide proper references and adhere to academic integrity, authorship and plagiarism guidelines.

Please note that, you are not required to per- form any of the analyses discussed here, but you can experiment with the data in any way you see suitable.

1. The story used in this case is \_ctitious and was chosen to illustrate a real-world

   situation. Consequently, any details herein bearing resemblance to real people or events

   is purely coincidental. [↑](#footnote-ref-2)