Data-Driven Investment Strategies for Peer-to-Peer Lending

Student Handout

Maxime C. Cohen
NYU Stern School of Business, New York, NY 10012, maxcohen@nyu.edu

C. Daniel Guetta
Columbia Business School, New York, NY 10027, guetta@gsb.columbia.edu

Kevin Jiao
NYU Stern School of Business, New York, NY 10012, jjiao@stern.nyu.edu

Foster Provost
NYU Stern School of Business, New York, NY 10012, fprovost@stern.nyu.edu

---

In this case, we follow Jasmin Gonzales, a young professional looking to diversify her investment portfolio.[1] Jasmin 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 Jasmin is especially excited by the large amount of data these platforms make publicly available. With her data science background, she is hoping to use machine learning tools on this data to come up with lucrative investment strategies. In this case, we follow Jasmin as she develops such an investment strategy.

1. 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

[1] The story used in this case is fictitious and was chosen to illustrate a real-world situation. Consequently, any details herein bearing resemblance to real people or events is purely coincidental. Furthermore, the content of this article reflects only one approach to the problem. The investment strategies and results obtained are by no means the only way to solve the problem at hand, with the goal of providing material for educational purposes.

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, employment 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 Securities 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 investors 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. 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

[2] `https://www.prosper.com/`

[3] `https://www.lendingclub.com/`

[4] `https://en.wikipedia.org/wiki/Lending_Club`

[5] `http://www.investmentzen.com/peer-to-peer-lending-for-investors/lendingclub` (date accessed: June 2018)

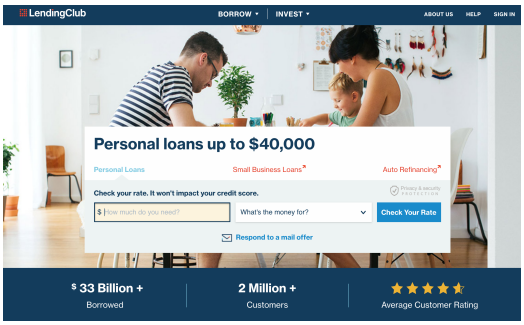
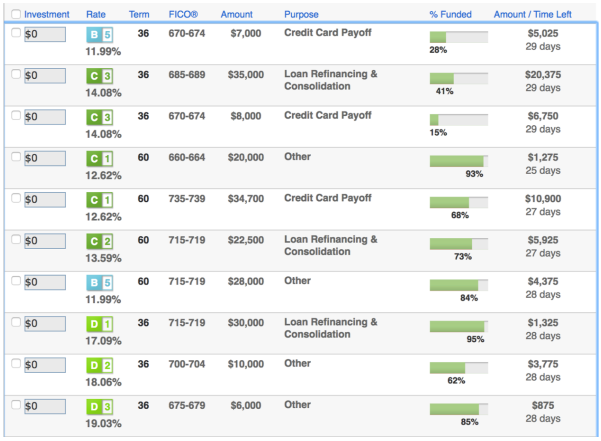


Figure 1: Screenshot of the LendingClub homepage.



Investment	Rate	Term	FICO®	Amount	Purpose	% Funded	Amount / Time Left
$0	B 5 11.99%	36	670-674	$7,000	Credit Card Payoff	28%	$5,025 29 days
$0	C 3 14.08%	36	685-689	$35,000	Loan Refinancing & Consolidation	41%	$20,375 29 days
$0	C 3 14.08%	36	670-674	$8,000	Credit Card Payoff	15%	$6,750 29 days
$0	C 1 12.62%	60	660-664	$20,000	Other	93%	$1,275 25 days
$0	C 1 12.62%	60	735-739	$34,700	Credit Card Payoff	68%	$10,900 27 days
$0	C 2 13.59%	60	715-719	$22,500	Loan Refinancing & Consolidation	73%	$5,925 27 days
$0	B 5 11.99%	60	715-719	$28,000	Other	84%	$4,375 28 days
$0	D 1 17.09%	36	715-719	$30,000	Loan Refinancing & Consolidation	95%	$1,325 28 days
$0	D 2 18.06%	36	700-704	$10,000	Other	62%	$3,775 28 days
$0	D 3 19.03%	36	675-679	$6,000	Other	85%	$875 28 days

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

access to their data to potential investors. This then raises a whole host of questions for investors like Jasmin:

- Is this data valuable when selecting loans to invest in?
- How could an investor use this data to develop machine learning tools to guide investment decisions?
- What is the impact of using data-driven tools on the portfolio performance relative to ad-hoc investment strategies?
- What average returns can an investor expect from informed investments in peer-to-peer loans?

The goal of this case study is to provide answers to the questions above. In particular, we investigate how data analytics and machine learning tools can be used in the context of peer-to-peer lending investments. We will use the historical data from loans that were issued on LendingClub between January 2009 and November 2017.

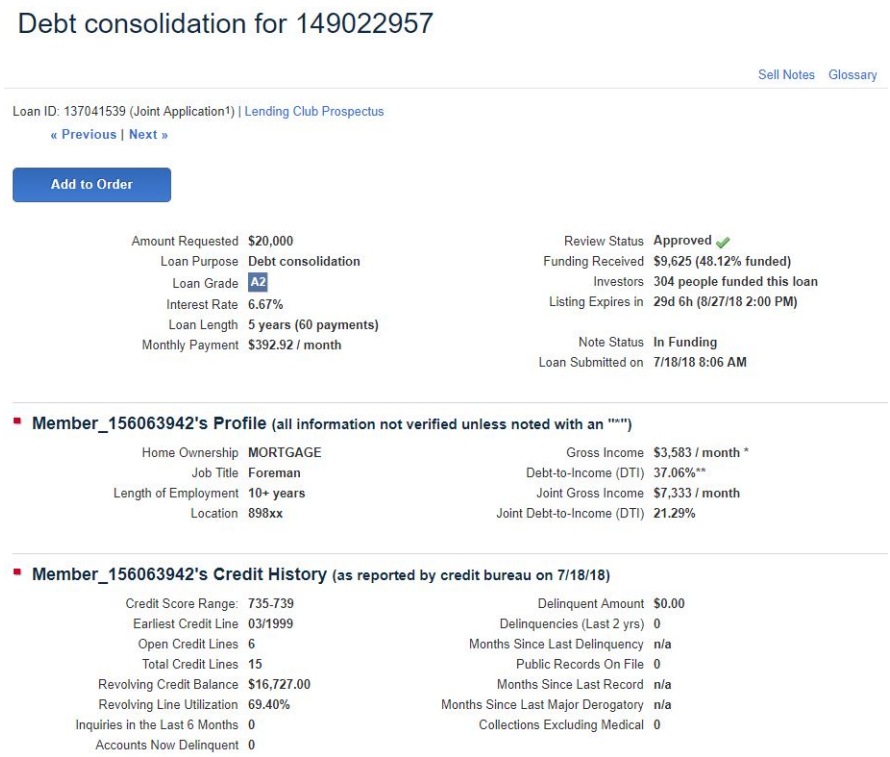


Figure 3: Example of a detailed loan listing, for a grade A loan. Information available to investors include the length of the borrower's employment, the borrower's credit score range, and their gross income, among others (source: LendingClub website, date accessed: July 2018).

2. Datasets and Descriptive Statistics

As mentioned, the datasets from LendingClub (and Prosper) are publicly available online.[6] These datasets contain comprehensive information on all loans issued between 2007 and the third quarter of 2017 (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,
3. Monthly installment amount,
4. Loan status (e.g., fully-paid, default, charged-off),
5. 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 be using in this case study 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.

[6] The analysis in this case study will focus on the LendingClub data. However, a similar analysis can be conducted using Prosper data.

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]

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-off[8]. 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, our 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, we investigate several machine learning tools and show how one can use historical data to construct informed investment strategies.

[7] Note that sometimes the "Charged-Off" status will occur before "Default" if/when the borrower has filed bankruptcy or has notified the intermediary platform.

[8] For 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.

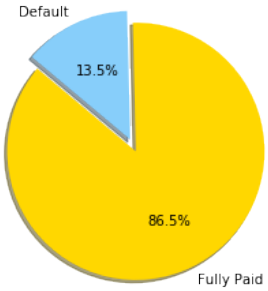


Figure 4: Proportion of *Fully Paid* versus *Default* for terminated loans (by November 2017).

3. Investment Strategies and Portfolio Construction

Making predictions and constructing a portfolio in the context of online peer-to-peer lending can be challenging. The volume of data available provides an opportunity to develop sophisticated data-driven methods. In practice, an investor like Jasmin would seek to construct a portfolio with the highest possible return, subject to constraints imposed by her risk tolerance, budget, diversification requirements (e.g., no more than 25% of loans with grades E or F). In this case study, we investigate the extent to which using appropriate predictive models can increase portfolio performance.

On important thing that Jasmin will encounter, as in many real applications of predictive analytics, is that it is far from trivial to progress from building a predictive model to using the model to make intelligent decisions. In her prior classes, Jasmin's exercises often ended with estimating the predictive ability of models on out-of-sample data. She will do that here as well—but then she will have to figure out how to estimate the return to expect from an investment. She will find that even with a seemingly good predictive model in hand, estimating the return of an investment requires additional analysis.

References

Optimization G (2014) Inc.,gurobi optimizer reference manual, 2015, url: http://www. gurobi.com.

Pre-Class Questions

Consider the following questions after reading the case:

1. Fundamentally, what decisions will Jasmin need to make?
2. What is Jasmin'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 Jasmin use past data to help make these decisions?

Downloading and Exploring the Data

Together with this case, you should have received a copy of LendingClub's data. Download it.

1. You will notice the data is provided in the form of many individual files, each spanning a certain time period. Combine the different files into a single dataset.
2. Take a look at the data. Write a high-level description of the different "attributes"—the variables describing the loans. How would you categorize these attributes? Which do you think are most important to an investor like Jasmin?
3. When looking through the data, you might have noticed that some of these variables seem related. For example, the `total_pymnt` variable is likely to be strongly correlated to the loan status. (Why?) Why would this matter, and how would you check?
4. Based on the variable names in the data, it's unclear whether the values of these variables are current as of the date the loan was issued, or as of the date the data was downloaded. For example, suppose you download the data in December 2017, and consider the `fico_range_low` variable for a loan that was issued in January 2015. It is unclear whether the score listed was the score in January 2015, or the score in December 2017. Why would this matter, and how might you check? (Hint: together with the case, you should have been given a copy of the data downloaded in 2017; use that).
5. Remove all instances (in our case, rows in the data table) representing loans that are still current (i.e., that are not in status 'Fully Paid', 'Charged-Off', or 'Default'), and all loans that were issued before January 1st 2009. Discuss the appropriateness of these filtering steps.
6. Visualize each of the attributes in the file. Are there any outliers? If yes, remove these.
7. Save the resulting dataset in a pickle. For the sake of this case, restrict yourself to the following attributes: `id`, `loan_amnt`, `funded_amnt`, `term`, `int_rate`, `grade`, `emp_length`, `home_ownership`, `annual_inc`, `verification_status`, `issue_d`, `loan_status`, `purpose`, `dti`, `delinq_2yrs`,
   `earliest_cr_line`, `open_acc`, `pub_rec`, `fico_range_high`, `fico_range_low`, `revol_bal`, `revol_util`, `total_pymnt`, and `recoveries`.

Initial Analyses

1. The most important data we will need in determining the return of each loan is the total payments that were received on each loan. There are two variables related to this information—`total_pymnt` and `recoveries`. Investigate these two variables, and for each loan, determine the total payment made on each loan.
2. A key measure we will need in working out an investment strategy is the return on each loan, defaulted or otherwise. How might you calculate this return? Add this new variable to the data.
3. As discussed in the case, LendingClub assigns a grade to each loan, from A through G. How many loans are in each grade? What is the default rate in each grade? What is the average interest rate in each grade? What about the average percentage (annual) return? Do these numbers surprise you? If you had to invest in one grade only, which loans would you invest in?

Probability of Default

1. Using the data provided, implement models to predict the probability each loan defaults. You may want to try the following models: decision tree, random forest (RF), logistic regression (ℓ_1 and ℓ_2 penalized), naive Bayes, and multi-layer perceptron. Carefully explain how you selected optimal model parameters and how you evaluated each model/modeling procedure.
2. After learning and evaluating these models, Jasmin realized that the attributes she used in her models were not all underlying facts about the loan applicants, but possibly statistics calculated by LendingClub using their own models. She wanted to assess whether the predictive power of her models came simply from LendingClub's own models. Carry out this investigation—what are your conclusions?
3. After modifying her model to ensure she did not include data calculated by LendingClub or leakage affecting the target variable, Jasmin wanted to assess the extent to which her scores agreed with the grades assigned by LendingClub. How might she do that?
4. Finally, Jasmin had one last concern. She 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?
5. Go back to the original data (before cleaning and attribute selection) and fit a model to predict the default probability using *all* attributes. (For the sake of simplicity, it will be sufficient to limit yourself to the following attributes: `id`, `loan_amnt`, `funded_amnt`, `funded_amnt_inv`, `term`, `int_rate`, `installment`, `grade`, `sub_grade`, `emp_title`, `emp_length`, `home_ownership`, `annual_inc`, `verification_status`, `issue_d`, `loan_status`, `purpose`, `title`, `zip_code`, `addr_state`, `dti`, `total_pymnt`, `delinq_2yrs`, `earliest_cr_line`, `open_acc`, `pub_rec`, `last_pymnt_d`, `last_pymnt_amnt`, `fico_range_high`, `fico_range_low`, `last_fico_range_high`, `last_fico_range_low`, `application_type`, `revol_bal`, `revol_util`, `recoveries`). Does anything surprise you about the performance of this model (out-of-sample) compared to the other models you have fit in this section?

Investment Strategies

1. What investment strategies might Jasmine use to decide which loans to invest in?
2. First, consider the three regression models described above (regressing against all returns, regressing against returns for defaulted loans, and regressing against returns for non-defaulted loans). In each case, try ℓ_1 and ℓ_2 penalized linear regression, random forest regression, and multi-layer perceptron regression.
3. Now, implement each of the investment strategies you suggested using the best performing regressor. In particular, suppose Jasmin were to invest in 1,000 loans using each of the four strategies; what would her returns be?
4. How might Jasmin test the stability of these results?
5. The strategies above were devised by investing in 1,000 loans. Jasmin is worried, however, that the strategy is not scalable—in other words, that if she wanted to increase the number of loans she wanted to invest in, she would eventually 'run out' of good loans to invest in. Test this hypothesis using the best strategy above.
6. Now investigate ways Jasmin might improve her investment strategy using optimization. Can you formulate the investment problem as an optimization model? How does your model perform compared to the best method above? To solve the optimization problem, you can use a solver such as Gurobi or CPLEX (free academic licenses are available). The reference manual for Gurobi is available (Optimization 2014).